Statistical Inference as Severe Testing

How to Get Beyond the Statistics Wars

Deborah G. Mayo

Virginia Tech

CAMBRIDGE
UNIVERSITY PRESS

Itinerary

Excursion 5 Power and Severity

Excursion 6 (Probabilist) Foundations Lost, (Probative) Foundations Found

Preface

The Statistics Wars

Today's "statistics wars" are fascinating: They are at once ancient and up to the minute. They reflect disagreements on one of the deepest, oldest, philosophical questions: How do humans learn about the world despite threats of error due to incomplete and variable data? At the same time, they are the engine behind current controversies surrounding high-profile failures of replication in the social and biological sciences. How should the integrity of science be restored? Experts do not agree. This book pulls back the curtain on why.

Methods of statistical inference become relevant primarily when effects are neither totally swamped by noise, nor so clear cut that formal assessment of errors is relatively unimportant. Should probability enter to capture degrees of belief about claims? To measure variability? Or to ensure we won't reach mistaken interpretations of data too often in the long run of experience? Modern statistical methods grew out of attempts to systematize doing all of these. The field has been marked by disagreements between competing tribes of frequentists and Bayesians that have been so contentious – likened in some quarters to religious and political debates – that everyone wants to believe we are long past them. We now enjoy unifications and reconciliations between rival schools, it will be said, and practitioners are eclectic, prepared to use whatever method "works." The truth is, long-standing battles still simmer below the surface in questions about scientific trustworthiness and the relationships between Big Data-driven models and theory. The reconciliations and unifications have been revealed to have serious problems, and there's little agreement on which to use or how to interpret them. As for eclecticism, it's often not clear what is even meant by "works." The presumption that all we need is an agreement on numbers – never mind if they're measuring different things – leads to pandemonium. Let's brush the dust off the pivotal debates, walk into the museums where we can see and hear such founders as Fisher, Neyman, Pearson, Savage, and many others. This is to simultaneously zero in on the arguments between metaresearchers – those doing research on research – charged with statistical reforms.

Statistical Inference as Severe Testing

Why are some arguing in today's world of high-powered computer searches that statistical findings are mostly false? The problem is that high-powered methods can make it easy to uncover impressive-looking findings even if they are false: spurious correlations and other errors have not been severely probed. We set sail with a simple tool: If little or nothing has been done to rule out flaws in inferring a claim, then it has not passed a *severe test*. In the severe testing view, probability arises in scientific contexts to assess and control how capable methods are at uncovering and avoiding erroneous interpretations of data. That's what it means to *view statistical inference as severe testing*. A claim is severely tested to the extent it has been subjected to and passes a test that probably would have found flaws, were they present. You may be surprised to learn that many methods advocated by experts do not stand up to severe scrutiny, are even in tension with successful strategies for blocking or accounting for cherry picking and selective reporting!

The severe testing perspective substantiates, using modern statistics, the idea Karl Popper promoted, but never cashed out. The goal of *highly well-tested* claims differs sufficiently from *highly probable* ones that you can have your cake and eat it too: retaining both for different contexts. Claims may be "probable" (in whatever sense you choose) but terribly tested by these data. In saying we may view statistical inference as severe testing, I'm not saying statistical inference is always about formal statistical testing. The testing metaphor grows out of the idea that before we have evidence for a claim, it must have passed an analysis that could have found it flawed. The probability that a method commits an erroneous interpretation of data is an *error probability*. Statistical methods based on error probabilities I call *error statistics*. The value of error probabilities, I argue, is not merely to control error in the long run, but because of what they teach us about the source of the data in front of us. The concept of severe testing is sufficiently general to apply to any of the methods now in use, whether for exploration, estimation, or prediction.

Getting Beyond the Statistics Wars

Thomas Kuhn's remark that only in the face of crisis "do scientists behave like philosophers" (1970), holds some truth in the current statistical crisis in science. Leaders of today's programs to restore scientific integrity have their own preconceptions about the nature of evidence and inference, and about "what we really want" in learning from data. Philosophy of science can also alleviate such conceptual discomforts. Fortunately, you needn't accept the severe testing view in order to employ it as a tool for bringing into focus the crux of all these issues. It's a tool for excavation, and for keeping us afloat in the marshes and quicksand that often mark today's controversies. Nevertheless, important consequences will follow once this tool is used. First there will be a reformulation of existing tools (tests, confidence intervals, and others) so as to avoid misinterpretations and abuses. The debates on statistical inference generally concern inference after a statistical model and data statements are in place, when in fact the most interesting work involves the local inferences needed to get to that point. A primary asset of error statistical methods is their contributions to designing, collecting, modeling, and learning from data. The severe testing view provides the much-needed link between a test's error probabilities and what's required for a warranted inference in the case at hand. Second, instead of rehearsing the same criticisms over and over again, challengers on all sides should now begin by grappling with the arguments we trace within. Kneejerk assumptions about the superiority of one or another method will not do. Although we'll be excavating the actual history, it's the methods themselves that matter; they're too important to be limited by what someone 50, 60, or 90 years ago thought, or to what today's discussants *think* they thought.

Who is the Reader of This Book?

This book is intended for a wide-ranging audience of readers. It's directed to consumers and practitioners of statistics and data science, and anyone interested in the methodology, philosophy, or history of statistical inference, or the controversies surrounding widely used statistical methods across the physical, social, and biological sciences. You might be a researcher or science writer befuddled by the competing recommendations offered by large groups ("megateams") of researchers (should *P*-values be set at 0.05 or 0.005, or not set at all?). By viewing a contentious battle in terms of a difference in goals – finding highly probable versus highly well-probed hypotheses – readers can see why leaders of rival tribes often talk right past each other. A fair-minded assessment may finally be possible. You may have a skeptical bent, keen to hold the experts accountable. Without awareness of the assumptions behind proposed reforms you can't scrutinize consequences that will affect you, be it in medical advice, economics, or psychology.

Your interest may be in improving statistical pedagogy, which requires, to begin with, recognizing that no matter how sophisticated the technology has become, the nature and meaning of basic statistical concepts are more unsettled than ever. You could be teaching a methods course in psychology wishing to intersperse philosophy of science in a way that is both serious and connected to immediate issues of practice. You might be an introspective statistician, focused on applications, but wanting your arguments to be on surer philosophical grounds.

Viewing statistical inference as severe testing will offer philosophers of science new avenues to employ statistical ideas to solve philosophical problems of induction, falsification, and demarcating science from pseudoscience. Philosophers of experiment should find insight into how statistical modeling bridges gaps between scientific theories and data. Scientists often question the relevance of philosophy of science to scientific practice. Through a series of excursions, tours, and exhibits, tools from the philosophy and history of statistics will be put directly to work to illuminate and solve problems of practice. I hope to galvanize philosophers of science and experimental philosophers to further engage with the burgeoning field of data science and reproducibility research.

Fittingly, the deepest debates over statistical foundations revolve around very simple examples, and I stick to those. This allows getting to the nitty-gritty logical issues with minimal technical complexity. If there's disagreement even there, there's little hope with more complex problems. (I try to use the notation of discussants, leading to some variation.) The book would serve as a one-semester course, or as a companion to courses on research methodology, philosophy of science, or interdisciplinary research in science and society. Each tour gives a small set of central works from statistics or philosophy, but since the field is immense, I reserve many important references for further reading on the CUP-hosted webpage for this book, www.cambridge.org/mayo.

Relation to Previous Work

While (1) philosophy of science provides important resources to tackle foundational problems of statistical practice, at the same time, (2) the statistical method offers tools for solving philosophical problems of evidence and inference. My earlier work, such as *Error and the Growth of Experimental Knowledge* (1996), falls under the umbrella of (2), using statistical science for philosophy of science: to model scientific inference, solve problems about evidence (problem of induction), and evaluate methodological rules (does more weight accrue to a hypothesis if it is prespecified?). *Error and Inference* (2010), with its joint work and exchanges with philosophers and statisticians, aimed to bridge the two-way street of (1) and (2). This work, by contrast, falls under goal (1): tackling foundational problems of statistical practice. While doing so will constantly find us entwined with philosophical problems of inference, it is the arguments and debates currently engaging practitioners that take the lead for our journey.

Join me, then, on a series of six excursions and 16 tours, during which we will visit three leading museums of statistical science and philosophy of science, and engage with a host of tribes marked by family quarrels, peace treaties, and shifting alliances.[1]

[1] A bit of travel trivia for those who not only read to the end of prefaces, but check its footnotes: two museums will be visited twice, one excursion will have no museums. With one exception, we engage current work through interaction with tribes, not museums. There's no extra cost for the 26 souvenirs: A–Z.

Tour I Beyond Probabilism and Performance

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

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

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

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

Getting Philosophical

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

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

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

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

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

Do I hear a protest? "There is nothing philosophical about our criticism of statistical significance tests (someone might say). The problem is that a small *P*-value is invariably, and erroneously, interpreted as giving a small probability to the null hypothesis.” Really? *P*-values are not intended to be used this way; presupposing they ought to be so interpreted grows out of a specific conception of the role of probability in statistical inference. *That conception is philosophical.* Methods characterized through the lens of over-simple epistemological orthodoxies are methods misapplied and mischaracterized. This may lead one to lie, however unwittingly, about the nature and goals of statistical inference, when what we want is to tell what’s true about them.

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

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

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

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

A Scandal Involving Personalized Medicine

A recent scandal offers an example. Over 100 patients signed up for the chance to participate in the Duke University (2007–10) clinical trials that promised a custom-tailored cancer treatment. A cutting-edge prediction model developed by Anil Potti and Joseph Nevins purported to predict your response to one or another chemotherapy based on large data sets correlating properties of various tumors and positive responses to different regimens (Potti et al. 2006). Gross errors and data manipulation eventually forced the trials to be halted. It was revealed in 2014 that a whistleblower – a student – had expressed concerns that

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

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

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

Do We Always Want to Find Things Out?

The severity requirement gives a minimal principle based on the fact that highly insevere tests yield bad evidence, no tests (BENT). We can all agree on this much, I think. We will explore how much mileage we can get from it. It applies at a number of junctures in collecting and modeling data, in linking data to statistical inference, and to substantive questions and claims. This will be our linchpin for understanding what's true about statistical inference. In addition to our minimal principle for evidence, one more thing is needed, at least during the time we are engaged in this project: *the goal of finding things out*.

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

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

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

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

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

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

A Philosophical Excursion

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

The job of the philosopher is to clarify but also to provoke reflection and scrutiny precisely in those areas that go unchallenged in ordinary practice. My focus will be on the issues having the most influence, and being most liable to obfuscation. Fortunately, that doesn't require an abundance of technicalities, but you can opt out of any daytrip that appears too technical: an idea not caught in one place should be illuminated in another. Our philosophical excursion may well land us in positions that are provocative to all existing sides of the debate about probability and statistics in scientific inquiry.

Methodology and Meta-methodology

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

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

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

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

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

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

To sum up this first tour so far, statistical inference uses data to reach claims about aspects of processes and mechanisms producing them, accompanied by an assessment of the properties of the inference methods: their capabilities to control and alert us to erroneous interpretations. We need to report if the method has satisfied the most minimal requirement for solving such a problem. Has anything been tested with a modicum of severity, or not? The severe tester also requires reporting of what has been poorly probed, and highlights the need to “bend over backwards,” as Feynman puts it, to admit where weaknesses lie. In formal statistical testing, the crude dichotomy of “pass/fail” or “significant or not” will scarcely do. We must determine the magnitudes (and directions) of any statistical discrepancies warranted, and the limits to any substantive claims you may be entitled to infer from the statistical ones. Using just our minimal principle of evidence, and a sturdy pair of shoes, join me on a tour of statistical inference, back to the leading museums of statistics, and forward to current offshoots and statistical tribes.

Why We Must Get Beyond the Statistics Wars

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

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

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

nor because

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

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

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

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

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, ***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 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: Lik(θ)= $\theta^s(1 - \theta)^f$, $0 < \theta < 1$, 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

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

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

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* 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 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. ($\sigma_{\overline{x}} = \sigma/\surd n$) 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:

$$H_0\text{: } \mu = 0 \text{ against } H_1\text{: } \mu \neq 0.$$

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 ***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($\overline{x}$, $1/n$).)

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: $\Pr(|\mathrm{d}(\boldsymbol{X})| > 1.96) = 0.05$ "is a statement about $\mathrm{d}(\boldsymbol{X})$ before it is observed. After it is observed, the event $\{\mathrm{d}(\boldsymbol{X}) > 1.96\}$ 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 $\bar{x} = 0.53$ 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 ~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.

Excursion 2 Taboos of Induction and Falsification

Itinerary

Tour I Induction and Confirmation

> Cox: [I]n some fields foundations do not seem very important, but we both think that foundations of statistical inference are important; why do you think that is?
>
> Mayo: I think because they ask about fundamental questions of evidence, inference, and probability . . . we invariably cross into philosophical questions about empirical knowledge and inductive inference. (Cox and Mayo 2011, p. 103)

Contemporary philosophy of science presents us with some taboos: Thou shalt not try to find solutions to problems of induction, falsification, and demarcating science from pseudoscience. It's impossible to understand rival statistical accounts, let alone get beyond the statistics wars, without first exploring how these came to be "lost causes." I am not talking of ancient history here: these problems were alive and well when I set out to do philosophy in the 1980s. I think we gave up on them too easily, and by the end of Excursion 2 you'll see why. Excursion 2 takes us into the land of "Statistical Science and Philosophy of Science" (StatSci/PhilSci). Our Museum Guide gives a terse thumbnail sketch of Tour I. Here's a useful excerpt:

> Once the Problem of Induction was deemed to admit of no satisfactory, non-circular solutions (~1970s), philosophers of science turned to building formal logics of induction using the deductive calculus of probabilities, often called Confirmation Logics or Theories. A leader of this Confirmation Theory movement was Rudolf Carnap. A distinct program, led by Karl Popper, denies there is a logic of induction, and focuses on Testing and Falsification of theories by data. At best a theory may be accepted or corroborated if it fails to be falsified by a severe test. The two programs have analogues to distinct methodologies in statistics: Confirmation theory is to Bayesianism as Testing and Falsification are to Fisher and Neyman–Pearson.

Tour I begins with the traditional Problem of Induction, then moves to Carnapian confirmation and takes a brief look at contemporary formal epistemology. Tour II visits Popper, falsification, and demarcation, moving into Fisherian tests and the replication crisis. Redolent of Frank Lloyd Wright's Guggenheim Museum in New York City, the StatSci/PhilSci Museum is arranged in concentric sloping oval floors that narrow as you go up. It's as if we're in a three-dimensional Normal curve. We begin in a large exposition on the ground floor. Those who start on the upper floors forfeit a central Rosetta Stone to decipher today's statistical debates.

2.1 The Traditional Problem of Induction

Start with the *asymmetry of falsification and confirmation*. One black swan falsifies the universal claim that *C*: all swans are white. Observing a single white swan, while a *positive instance* of *C*, wouldn't allow inferring generalization *C*, unless there was only one swan in the entire population. If the generalization refers to an infinite number of cases, as most people would say about scientific theories and laws, then no matter how many positive instances observed, you couldn't infer it with certainty. It's always possible there's a black swan out there, a *negative instance*, and it would only take one to falsify *C*. But surely we think enough positive instances of the right kind might warrant an argument for inferring *C*. Enter the problem of induction. First, a bit about arguments.

Soundness versus Validity

An *argument* is a group of statements, one of which is said to follow from or be supported by the others. The others are premises, the one inferred, the conclusion. A deductively *valid* argument is one where if its premises are all true, then its conclusion must be true. Falsification of "all swans are white" follows a deductively valid argument. Let ~*C* be the denial of claim *C*.

(1) *C*: All swans are white.
x is a swan but is black.
Therefore, ~*C*.

We can also infer, validly, what follows if a generalization *C* is true.

(2) *C*: All swans are white.
x is a swan.
Therefore, *x* is white.

However, validity is not the same thing as *soundness*. Here's a case of argument form (2):

(3) All philosophers can fly.
Mayo is a philosopher.
Therefore, Mayo can fly.

Validity is a matter of form. Since (3) has a valid form, it is a valid argument. But its conclusion is false! That's because it is *unsound*: at least one of its premises is false (the first). No one can stop you from applying deductively valid arguments, regardless of your statistical account. Don't assume you will get truth thereby. Bayes' Theorem can occur in a valid argument, within a formal system of probability:

> (4) If $\Pr(H_1), \ldots, \Pr(H_n)$ are the prior probabilities of an exhaustive set of hypotheses, and $\Pr(\boldsymbol{x}|H_i)$ the corresponding likelihoods.
>
> Data $\boldsymbol{x}$ are given, and $\Pr(H_1|\boldsymbol{x})$ is defined.
>
> Therefore, $\Pr(H_1|\boldsymbol{x}) = p$.[1]

The conclusion is the posterior probability $\Pr(H_1|\boldsymbol{x})$. It can be inferred only if the argument is *sound*: all the givens must hold (at least approximately). To deny that all of statistical inference is reducible to Bayes' Theorem is not to preclude your using this or any other deductive argument. What you need to be concerned about is their soundness. So, you will still need a way to vouchsafe the premises.

Now to the traditional philosophical problem of induction. What is it? Why has confusion about induction and the threat of the traditional or "logical" problem of induction made some people afraid to dare use the "I" word? The traditional problem of induction seeks to justify a type of argument: one taking a form of *enumerative induction* (EI) (or the *straight rule* of induction). Infer from past cases of A's that were B's to all or most A's will be B's:

> EI: All observed $A_1, A_2, \ldots, A_n$ have been B's.
>
> Therefore, H: all A's are B's.

It is not a deductively valid argument, because clearly its premises can all be true while its conclusion false. It's *invalid*, as is so for any inductive argument. As Hume (1739) notes, nothing changes if we place the word "probably" in front of the conclusion: it is justified to infer from past A's being B's that, *probably*, all or most A's will be B's. To "rationally" justify induction is to supply a reasoned argument for using EI. The traditional problem of induction, then, involves trying to find an argument to justify a type of argument!

Exhibit (i): Justifying Induction Is Circular. In other words, the traditional problem of induction is to justify the conclusion:

[1] i.e., $p = \dfrac{\Pr(\boldsymbol{x}|H_1)\Pr(H_1)}{\Pr(\boldsymbol{x}|H_1)\Pr(H_1) + \cdots + \Pr(\boldsymbol{x}|H_n)\Pr(H_n)}$

Conclusion: EI is rationally justified, it's a reliable rule.

We need an argument for concluding EI is reliable. Using an inductive argument to justify induction lands us in a circle. We'd be using the method we're trying to justify, or begging the question. What about a deductively valid argument? The premises would have to be things we know to be true, otherwise the argument would not be sound. We might try:

Premise 1: EI has been reliable in a set of observed cases.

Trouble is, this premise can't be used to deductively infer EI will be reliable *in general*: the known cases only refer to the past and present, not the future. Suppose we add a premise:

Premise 2: Methods that have worked in past cases will work in future cases.

Yet to assume Premise 2 is true is to use EI, and thus, again, to beg the question.

Another idea for the additional premise is in terms of assuming nature is uniform. We do not escape: to assume the *uniformity of nature* is to assume EI is a reliable method. Therefore, induction cannot be rationally justified. It is called the *logical* problem of induction because logical argument alone does not appear able to solve it. All attempts to justify EI assume past successes of a rule justify its general reliability, which is to assume EI – what we're trying to show.

I'm skimming past the rest of a large exhibition on brilliant attempts to solve induction in this form. Some argue that although an attempted justification is circular it is not *viciously* circular. (An excellent source is Skyrms 1986.)

But wait. Is inductive enumeration a rule that has been reliable even in the past? No. It is reasonable to expect that unobserved or future cases will be very different from the past, that apparent patterns are spurious, and that observed associations are not generalizable. We would only want to justify inferences of that form if we had done a good job ruling out the many ways we know we can be misled by such an inference. That's not the way confirmation theorists see it, or at least, saw it.

Exhibit (ii): Probabilistic (Statistical) Affirming the Consequent. Enter logics of confirmation. Conceding that we cannot justify the inductive method (EI), philosophers sought logics that represented apparently plausible inductive reasoning. The thinking is this: never mind trying to convince a skeptic of the inductive method, we give up on that. But we know what we mean. We need only to make sense of the habit of applying EI. True to the logical positivist spirit of the 1930s–1960s, they sought evidential relationships between statements of evidence and conclusions. I sometimes call them evidential-relation (E-R) logics. They didn't renounce enumerative induction, they sought logics that embodied it. Begin by fleshing out the full argument behind EI:

> If *H*: all A's are B's, then all observed A's (A_1, A_2, . . ., A_n) are B's.
> All observed A's (A_1, A_2, . . ., A_n) are B's.
> Therefore, *H*: all A's are B's.

The premise that we added, the first, is obviously true; the problem is that the second premise can be true while the conclusion false. The argument is deductively *invalid* – it even has a name: *affirming the consequent*. However, its probabilistic version is weaker. *Probabilistic affirming the consequent* says only that the conclusion is probable or gets a boost in confirmation or probability – a *B-boost*. It's in this sense that Bayes' Theorem is often taken to ground a plausible confirmation theory. It probabilistically justifies EI in that it embodies probabilistic affirming the consequent.

How do we obtain the probabilities? Rudolf Carnap's audacious program (1962) had been to assign probabilities of hypotheses or statements by deducing them from the logical structure of a particular (first order) language. These were called *logical probabilities*. The language would have a list of properties (e.g., "is a swan," "is white") and individuals or names (e.g., i, j, k). The task was to assign equal initial probability assignments to states of this mini world, from which we could deduce the probabilities of truth functional combinations. The degree of probability, usually understood as a rational degree of belief, would hold between two statements, one expressing a hypothesis and the other the data. $C(H,\boldsymbol{x})$ symbolizes "the confirmation of *H*, given ***x***." Once you have chosen the initial assignments to core states of the world, calculating degrees of confirmation is a formal or syntactical matter, much like deductive logic. The goal was to somehow measure the *degree of implication* or confirmation that ***x*** affords *H*. Carnap imagined the scientist coming to the inductive logician to have the rational degree of confirmation in *H* evaluated, given her evidence. (I'm serious.) Putting aside the difficulty of listing all properties of scientific interest, from where do the initial assignments come?

Carnap's first attempt at a C-function resulted in no learning! For a toy illustration, take a universe with three items, i, j, k, and a single property B. "Bk" expresses "k has property B." There are eight possibilities, each called a *state description*. Here's one: {Bi, ~Bj, ~Bk}. If each is given initial probability of ⅛, we have what Carnap called the logic $c^{\dagger}$. The degree of confirmation that j will be black given that i was white = ½, which is the same as the initial confirmation of Bi (since it occurs in four of eight state descriptions). Nothing has been learned: c† is scrapped. By apportioning initial probabilities more coarsely, one could learn, but there was an infinite continuum of inductive logics characterized by choosing the value of a parameter he called λ (λ continuum). λ in effect determines how much uniformity and regularity to expect. To restrict the field, Carnap had to postulate what he called "inductive intuitions." As a logic student, I too found these attempts tantalizing – until I walked into my first statistics class. I was also persuaded by philosopher Wesley Salmon:

> Carnap has stated that the ultimate justification of the axioms is inductive intuition. I do not consider this answer an adequate basis for a concept of rationality. Indeed, I think that *every* attempt, including those by Jaakko Hintikka and his students, to ground the concept of rational degree of belief in logical probability suffers from the same unacceptable *apriorism*. (Salmon 1988, p. 13).

This program, still in its heyday in the 1980s, was part of a general logical positivist attempt to reduce science to observables plus logic (no metaphysics). Had this reductionist goal been realized, which it wasn't, the idea of scientific inference being reduced to particular predicted observations might have succeeded. Even with that observable restriction, the worry remained: what does a highly probable claim, according to a particular inductive logic, have to do with the real world? How can it provide "a guide to life?" (e.g., Kyburg 2003, Salmon 1966). The epistemology is restricted to inner coherence and consistency. However much contemporary philosophers have gotten beyond logical positivism, the hankering for an inductive logic remains. You could say it's behind the appeal of the default (non-subjective) Bayesianism of Harold Jeffreys, and other attempts to view probability theory as extending deductive logic.

Exhibit (iii): A Faulty Analogy Between Deduction and Induction. When we heard Hacking announce (Section 1.4): "there is no such thing as a logic of statistical inference" (1980, p. 145), it wasn't only the failed attempts to build one, but the recognition that the project is "founded on a false analogy with deductive logic" (ibid.). The issue here is subtle, and we'll revisit it through our journey. I agree with Hacking, who is agreeing with C. S. Peirce:

> In the case of analytic [deductive] inference we know the probability of our conclusion (if the premises are true), but in the case of synthetic [inductive] inferences we only know the degree of trustworthiness of our proceeding. (Peirce 2.693)

In getting new knowledge, in ampliative or inductive reasoning, the conclusion should go beyond the premises; probability enters to qualify the overall "trustworthiness" of the method. Hacking not only retracts his Law of Likelihood (LL), but also his earlier denial that Neyman–Pearson statistics is inferential. "I now believe that Neyman, Peirce, and Braithwaite were on the right lines to follow in the analysis of inductive arguments" (Hacking 1980, p. 141). Let's adapt some of Hacking's excellent discussion.

When we speak of an inference, it could mean the entire argument including premises and conclusion. Or it could mean just the conclusion, or statement inferred. Let's use "inference" to mean the latter – the claim detached from the premises or data. A statistical procedure of *inferring* refers to a method for reaching a statistical inference about some aspect of the source of the data, together with its probabilistic properties: in particular, its capacities to avoid erroneous (and ensure non-erroneous) interpretations of data. These are the method's error probabilities. My argument from coincidence to weight gain (Section 1.3) inferred *H*: I've gained at least 4 pounds. The inference is qualified by the detailed data (group of weighings), and information on how capable the method is at blocking erroneous pronouncements of my weight. I argue that, very probably, my scales would not produce the weight data they do (e.g., on objects with known weight) were *H* false. What is being qualified probabilistically is the inferring or testing process.

By contrast, in a probability or confirmation logic, what is generally detached is the probability of *H*, given data. It is a *probabilism*. Hacking's diagnosis in 1980 is that this grows out of an abiding logical positivism, with which he admits to having been afflicted. There's this much analogy with deduction: In a deductively valid argument: if the premises are true then, necessarily, the conclusion is true. But we don't attach the "necessarily" to the conclusion. Instead it qualifies the entire argument. So mimicking deduction, why isn't the inductive task to qualify the method in some sense, for example, report that it would probably lead to true or approximately true conclusions? That would be to show the reliable performance of an inference method. If that's what an inductive method requires, then Neyman–Pearson tests, which afford good performance, are inductive.

My main difference from Hacking here is that I don't argue, as he seems to, that the warrant for the inference is that it stems from a method that very probably gets it right (so I may hope it is right this time). It's not that the method's reliability "rubs off" on this particular claim. I say inference *C* may be detached as *indicated* or *warranted*, having passed a severe test (a test that *C* probably would have failed, if false in a specified manner). This is the central point of Souvenir D. The logician's "semantic entailment" symbol, the double turnstile: "|=", could be used to abbreviate "entails severely":

$$\text{Data + capacities of scales} \mid=_{\text{SEV}} \text{I've gained at least } k \text{ pounds.}$$

(The premises are on the left side of |=.) However, I won't use this notation.

Keeping to a deductive logic of probability, we never detach an inference. This is in sync with a probabilist such as Bruno de Finetti:

> *The calculus of probability can say absolutely nothing about reality* . . . As with the logic of certainty, the logic of the probable adds nothing of its own: it merely helps one to see the implications contained in what has gone before. (de Finetti 1974, p. 215)

These are some of the first clues we'll be collecting on a wide difference between statistical inference as a deductive logic of probability, and an inductive testing account sought by the error statistician. When it comes to inductive learning, we want our inferences to go beyond the data: we want lift-off. To my knowledge, Fisher is the only other writer on statistical inference, aside from Peirce, to emphasize this distinction.

> In deductive reasoning all knowledge obtainable is already latent in the postulates. Rigour is needed to prevent the successive inferences growing less and less accurate as we proceed. The conclusions are never more accurate than the data. In inductive reasoning we are performing part of the process by which new knowledge is created. The conclusions normally grow more and more accurate as more data are included. It should never be true, though it is still often *said*, that the conclusions are no more accurate than the data on which they are based. (Fisher 1935b, p. 54)

2.2 Is Probability a Good Measure of Confirmation?

> It is often assumed that the degree of confirmation of x by y must be the same as the (relative) probability of x given y, i.e., that $C(x, y) = \Pr(x, y)$. My first task is to show the inadequacy of this view. (Popper 1959, p. 396; substituting Pr for P)

If your suitcase rings the alarm at an airport, this might slightly increase the probability of its containing a weapon, and slightly decrease the probability that it's clean. But the probability it contains a weapon is so small that the probability it's clean remains high, even if it makes the alarm go off. These facts illustrate a tension between two ways a probabilist might use probability to measure confirmation. A test of a philosophical confirmation theory is whether it elucidates or is even in sync with intuitive methodological principles about evidence or testing. Which, if either, fits with intuitions?

The most familiar interpretation is that H is confirmed by $\boldsymbol{x}$ if $\boldsymbol{x}$ gives a boost to the probability of *H, incremental* confirmation. The components of $C(H,\boldsymbol{x})$ are allowed to be any statements, and, in identifying C with Pr, no reference to a probability model is required. There is typically a background variable k, so that $\boldsymbol{x}$ confirms H relative to k: to the extent that $\Pr(H|\boldsymbol{x} \text{ and } k) > \Pr(H \text{ and } k)$. However, for readability, I will drop the explicit inclusion of k. More generally, if H entails $\boldsymbol{x}$, then assuming $\Pr(\boldsymbol{x}) \neq 1$ and $\Pr(H) \neq 0$, we have $\Pr(H|\boldsymbol{x}) > \Pr(H)$.

This is an instance of probabilistic affirming the consequent. (Note: if $\Pr(H|\boldsymbol{x}) > \Pr(H)$ then $\Pr(\boldsymbol{x}|H) > \Pr(\boldsymbol{x})$.)

(1) *Incremental* (B-boost): H is confirmed by $\boldsymbol{x}$ iff $\Pr(H|\boldsymbol{x}) > \Pr(H)$,
H is disconfirmed iff $\Pr(H|\boldsymbol{x}) < \Pr(H)$.

("iff" denotes if and only if.) Also plausible is an *absolute* interpretation:

(2) *Absolute*: H is confirmed by $\boldsymbol{x}$ iff $\Pr(H|\boldsymbol{x})$ is high, at least greater than $\Pr(\sim H|\boldsymbol{x})$.

Since $\Pr(\sim H|\boldsymbol{x}) = 1 - \Pr(H|\boldsymbol{x})$, (2) is the same as defining $\boldsymbol{x}$ confirms H: $\Pr(H|\boldsymbol{x}) > 0.5$. From (1), $\boldsymbol{x}$ (the alarm) *dis*confirms the hypothesis H: the bag is clean, because its probability has gone down, however slightly. Yet from (2) $\boldsymbol{x}$ confirms H: bag is clean, because $\Pr(H)$ is high to begin with.

There's a conflict. Thus, if (1) seems plausible, then probability, $\Pr(H|\boldsymbol{x})$, isn't a satisfactory way to define confirmation. At the very least, we must distinguish between an incremental and an absolute measure of confirmation for H. No surprise there. From the start Carnap recognized that "the verb 'to confirm' is ambiguous"; Carnap and most others choose the "making firmer" or incremental connotation as better capturing what is meant than that of "making firm" (Carnap 1962, p. xviii). Incremental confirmation is generally used in current Bayesian epistemology. Confirmation is a B-boost.

The first point Popper's making in the epigraph is this: to identify confirmation and probability "$C = \Pr$" leads to this type of conflict. His example is a single toss of a homogeneous die: The data $\boldsymbol{x}$: an even number occurs; hypothesis H: a 6 will occur. It's given that $\Pr(H) = 1/6$, $\Pr(\boldsymbol{x}) = 1/2$. The probability of H is increased by data $\boldsymbol{x}$, while $\sim H$ is undermined by $\boldsymbol{x}$ (its probability goes from 5/6 to 4/6). If we identify probability with degree of confirmation, $\boldsymbol{x}$ confirms H and disconfirms $\sim H$. However, $\Pr(H|\boldsymbol{x}) < \Pr(\sim H|\boldsymbol{x})$. So H is less well confirmed given $\boldsymbol{x}$ than is $\sim H$, in the sense of (2). Here's how Popper puts it, addressing Carnap: How can we say H is confirmed by $\boldsymbol{x}$, while $\sim H$ is not; but at the same time $\sim H$ is confirmed to a higher degree with $\boldsymbol{x}$ than is H? (Popper 1959, p. 390).[2]

[2] Let HJ be (H & J). To show: If there is a case where $\boldsymbol{x}$ confirms HJ more than $\boldsymbol{x}$ confirms J, then degree of probability cannot equal degree of confirmation.

(i) $C(HJ, \boldsymbol{x}) > C(J, \boldsymbol{x})$ is given.
(ii) $J = \sim HJ$ or HJ by logical equivalence.
(iii) $C(HJ, \boldsymbol{x}) > C(\sim HJ$ or $HJ, \boldsymbol{x})$ by substituting (ii) in line (i).

Since $\sim HJ$ and HJ are mutually exclusive, we have from the special addition rule for probability:

(iv) $\Pr(HJ, \boldsymbol{x}) \leq \Pr(\sim HJ$ or $HJ, \boldsymbol{x})$.

So if $\Pr = C$, (iii) and (iv) yield a contradiction. (Adapting Popper 1959, p. 391)

Moreover, Popper continues, confirmation theorists don't use $\Pr(H|\boldsymbol{x})$ alone (as they would if $C = \Pr$), but myriad functions of probability to capture how much $\boldsymbol{x}$ has firmed up H. A number of measures offer themselves for the job. A simple B-boost would report the ratio R: $\Pr(H|\boldsymbol{x})/\Pr(H)$, which in Popper's example is 2. Or we can use the likelihood ratio of H compared to $\sim H$. Since I used LR in Excursion 1, where the two hypotheses are not exhaustive, let's write [LR] to denote

$$[\mathrm{LR}]\text{: } \frac{\Pr(\boldsymbol{x}|H)}{\Pr(\boldsymbol{x}|{\sim}H)} = (1/0.4) = 2.5.$$

Many other ways of measuring the increase in confirmation that $\boldsymbol{x}$ affords H could do as well. (For some excellent lists see Popper 1959 and Fitelson 2002.)

What shall we say about the numbers like 2, 2.5? Do they mean the same thing in different contexts? Then there's the question of computing $\Pr(\boldsymbol{x}|{\sim}H)$, the *catchall factor*. It doesn't offer problems in this case because $\sim H$, the *catchall hypothesis*, is just an event statement. It's far more problematic once we move to genuine statistical hypotheses. Recall how Royall's Likelihoodist avoids the composite catchall factor by restricting his likelihood ratios to two simple statistical hypotheses.

Popper's second point is that "the probability of a statement . . . simply does not express an appraisal of the severity of the tests a theory has passed, or of the manner in which it has passed these tests" (pp. 394–5). Ultimately, Popper denies that severity can be completely formalized by any C function. Is there nothing in between a pure formal-syntactical approach and leaving terms at a vague level? I say there is.

Consider for a moment philosopher Peter Achinstein – a Carnap student. Achinstein (2000, 2001) declared that scientists should not take seriously philosophical accounts of confirmation because they make it too easy to confirm. Furthermore, scientists look to empirical grounds for confirmation, whereas philosophical accounts give us formal (non-empirical) a priori measures. (I call it Achinstein's "Dean's problem" because he made the confession to a Dean asking about the relevance of philosophy – not usually the best way to keep funding for philosophy.) Achinstein rejects confirmation as increased firmness, denying it is either necessary or sufficient for evidence (rejects (1)).[3] He requires for H to be confirmed by $\boldsymbol{x}$ that the posterior of H given $\boldsymbol{x}$ be rather high, a version of (2): $\Pr(H|\boldsymbol{x}) \gg \Pr({\sim}H|\boldsymbol{x})$, but that's not all. He requires that, before we apply

[3] Why is a B-boost not necessary for Achinstein? Suppose you know x: the newspaper says Harry won, and it's never wrong. Then a radio, also assumed 100% reliable, announces y: Harry won. Statement y, Achinstein thinks, should still count as evidence for H: he won. I agree.

confirmation measures, the components have an appropriate explanatory relationship to each other. Yet this requires an adequate way to make explanatory inferences before getting started. It's not clear how the formalism helped. He still considers himself a Bayesian epistemologist – a term that has replaced confirmation theorist – but the probabilistic representation threatens to be mostly a kind of bookkeeping for inferential work done in some other way.

Achinstein is right to object that (1) incremental confirmation makes it too easy to have evidence. After all, *J*: Mike drowns in the Pacific Ocean, entails $\boldsymbol{x}$: there is a Pacific Ocean; yet $\boldsymbol{x}$ does not seem to be evidence for *J*. Still the generally favored position is to view confirmation as (1) a B-boost.

Exhibit (iv): Paradox of Irrelevant Conjunctions. Consider a famous argument due to Glymour (1980). If we allow that $\boldsymbol{x}$ confirms *H* so long as $\Pr(H|\boldsymbol{x}) > \Pr(H)$, it seems everything confirms everything, so long as one thing is confirmed!

The first piece of the argument is the problem of irrelevant conjunctions – also called the "tacking paradox." If $\boldsymbol{x}$ confirms *H*, then $\boldsymbol{x}$ also confirms $(H \,\&\, J)$, even if hypothesis *J* is just "tacked on" to *H*. As with most of these chestnuts, there is a long history (e.g., Earman 1992, Rosenkrantz 1977) but I consider a leading contemporary representative, Branden Fitelson. Fitelson (2002) and Hawthorne and Fitelson (2004) define the statement "*J* is an *irrelevant conjunct* to *H*, with respect to evidence $\boldsymbol{x}$" as meaning $\Pr(\boldsymbol{x}|J) = \Pr(\boldsymbol{x}|J \,\&\, H)$. For instance, $\boldsymbol{x}$ might be radioastronomic data in support of

> *H*: the General Theory of Relativity (GTR) deflection of light effect is 1.75″ and
>
> *J*: the radioactivity of the Fukushima water being dumped in the Pacific Ocean is within acceptable levels.

(A) If $\boldsymbol{x}$ confirms *H*, then $\boldsymbol{x}$ confirms $(H \,\&\, J)$, where $\Pr(\boldsymbol{x}|H \,\&\, J) = \Pr(\boldsymbol{x}|H)$ for any *J* consistent with *H*.

The reasoning is as follows:

> (i) $\Pr(\boldsymbol{x}|H)/\Pr(\boldsymbol{x}) > 1$ ($\boldsymbol{x}$ *Bayesian-confirms H*).
>
> (ii) $\Pr(\boldsymbol{x}|H \,\&\, J) = \Pr(\boldsymbol{x}|H)$ (*J*'s irrelevance is given).

Substituting (ii) into (i) gives $\Pr(\boldsymbol{x}|H \,\&\, J)/\Pr(\boldsymbol{x}) > 1$.

> Therefore $\boldsymbol{x}$ Bayesian-confirms $(H \,\&\, J)$.[4]

[4] To expand the reasoning, first observe that $\Pr(H|\boldsymbol{x})/\Pr(H) = \Pr(\boldsymbol{x}|H)/\Pr(\boldsymbol{x})$ and $\Pr(H \,\&\, J|\boldsymbol{x})/\Pr(H \,\&\, J) = \Pr(\boldsymbol{x}|H \,\&\, J)/\Pr(\boldsymbol{x})$, both by Bayes' Theorem. So, when $\Pr(H|\boldsymbol{x})/\Pr(H) > 1$, we also have $\Pr(\boldsymbol{x}|H)/\Pr(\boldsymbol{x}) > 1$. This, together with $\Pr(\boldsymbol{x}|H \,\&\, J) = \Pr(\boldsymbol{x}|H)$ (given), yields $\Pr(\boldsymbol{x}|H \,\&\, \mathrm{J})/\Pr(\boldsymbol{x}) > 1$. Thus, we also have $\Pr(H \,\&\, J|\boldsymbol{x})/\Pr(H \,\&\, J)>1$.

However, it is also plausible to hold what philosophers call the "special consequence" condition: If $\boldsymbol{x}$ confirms a claim W, and W entails J, then $\boldsymbol{x}$ confirms J. In particular:

(B) If $\boldsymbol{x}$ confirms (H & J), then $\boldsymbol{x}$ confirms J.

(B) gives the second piece of the argument. From (A) and (B) we have, if $\boldsymbol{x}$ confirms H, then $\boldsymbol{x}$ confirms J for any irrelevant J consistent with H (neither H nor J have probabilities 0 or 1).

It follows that if $\boldsymbol{x}$ confirms any H, then $\boldsymbol{x}$ confirms any J.

This absurd result, however, assumed (B) (special consequence) and most Bayesian epistemologists reject it. This is the gist of Fitelson's solution to tacking, updated in Hawthorne and Fitelson (2004). It is granted that $\boldsymbol{x}$ confirms the conjunction (H & J), while denying $\boldsymbol{x}$ confirms the irrelevant conjunct J. Aren't they uncomfortable with (A), allowing (H & J) to be confirmed by $\boldsymbol{x}$?

I'm inclined to agree with Glymour that we are not too happy with an account of evidence that tells us deflection of light data confirms the conjunction of the GTR deflection and the radioactivity of the Fukushima water is within acceptable levels, while assuring us that $\boldsymbol{x}$ does not confirm the conjunct, that the Fukushima water has acceptable levels of radiation (1980, p. 31). Moreover, suppose we measure the confirmation boost by

R: $\Pr(H|\boldsymbol{x})/\Pr(\boldsymbol{x})$.

Then, Fitelson points out, the conjunction (H & J) is just as well confirmed by $\boldsymbol{x}$ as is H!

However, granting confirmation is an incremental B-boost doesn't commit you to measuring it by R. The conjunction (H & J) gets less of a confirmation boost than does H if we use, instead of R, the likelihood ratio [LR] of H against $\sim H$:

[LR]: $\Pr(\boldsymbol{x}|H)/\Pr(\boldsymbol{x}|\sim H)$.[5]

This avoids the counterintuitive result, or so it is claimed. (Note: $\Pr(H|\boldsymbol{x}) > \Pr(H)$ iff $\Pr(\boldsymbol{x}|H) > \Pr(\boldsymbol{x})$, but measuring the boost by R differs from measuring it with [LR].)

[5] Recall that Royall restricts the likelihood ratio to non-composite hypotheses, whereas here $\sim H$ is the Bayesian catchall.

What Does the Severity Account Say?

Our account of inference disembarked way back at (1): that $\boldsymbol{x}$ confirms H so long as $\Pr(H|\boldsymbol{x}) > \Pr(H)$. That is, we reject probabilistic affirming the consequent. In the simplest case, H entails $\boldsymbol{x}$, and $\boldsymbol{x}$ is observed. (We assume the probabilities are well defined, and H doesn't already have probability 1.) H gets a B-boost, but there are many other "explanations" of $\boldsymbol{x}$. It's the same reason we reject the Law of Likelihood (LL). Unless stringent probing has occurred, finding an H that fits $\boldsymbol{x}$ is not difficult to achieve even if H is false. H hasn't passed severely. Now severely passing is obviously stronger than merely finding some evidence for H, and the confirmation theorist is only saying a B-boost suffices for some evidence. To us, to have *any* evidence, or even the weaker notion of an "indication," requires a minimal threshold of severity be met.

How about tacking? As always, the error statistician needs to know the relevant properties of the test procedure or rule, and just handing me the H's, $\boldsymbol{x}$'s, and relative probabilities will not suffice. The process of tacking, at least one form, is this – once you have an incrementally confirmed H with data $\boldsymbol{x}$, tack on any consistent J and announce "$\boldsymbol{x}$ confirms $(H\ \&\ J)$." Let's allow that $(H\ \&\ J)$ fits or accords with $\boldsymbol{x}$ (since GTR entails or renders probable the deflection data $\boldsymbol{x}$). However, the very claim: "$(H\ \&\ J)$ is confirmed by $\boldsymbol{x}$" has been subjected to a radically non-risky test. Nothing has been done to measure the radioactivity of the Fukushima water being dumped into the ocean. B-boosters might reply, "We're admitting J is irrelevant and gets no confirmation," but our testing intuitions tell us then it's crazy to regard $(H\ \&\ J)$ as confirmed. They will point out other examples where this doesn't seem crazy. But what matters is that it's being permitted in general.

We should *punish* a claim to have evidence for H with a tacked-on J, when nothing has been done to refute J. Imagine the chaos. Are we to allow positive trial data on diabetes patients given drug D to confirm the claim that D improves survival of diabetes patients *and* Roche's artificial knee is effective, when there's only evidence for one? If the confirmation theorist simply stipulates that (1) defines confirmation, then it's within your rights to deny it captures ordinary notions of evidence. On the other hand, if you do accept (1), then why are you bothered at all by tacking? Many are not.

Patrick Maher (2004) argues that if B-boosting is confirmation, then there is nothing counterintuitive about data confirming irrelevant conjunctions; Fitelson should not even be conceding "he bites the bullet." It makes sense that $(H\ \&\ J)$ increases the probability assignment to $\boldsymbol{x}$ just as much as does H, for J the irrelevant conjunct. The supposition that this is problematic and that therefore one must move away from R: $\Pr(\boldsymbol{x}|H)/\Pr(\boldsymbol{x})$ sits uneasily with the fact that $R > 1$ is just what confirmation boost means. Rather than "solve" the problem by saying we can measure boost so that $(H \& J)$ gets less confirmation than H, using [LR], why not see it as what's meant by an irrelevant conjunct J: J doesn't improve the ability to predict $\boldsymbol{x}$. Other philosophers working in this arena, Crupi and Tentori (2010), notice that [LR] is not without problems. In particular, if $\boldsymbol{x}$ *dis*confirms hypothesis Q, then $(Q \& J)$ isn't as badly disconfirmed as Q is, for irrelevant conjunct J. Just as $(H \& J)$ gets less of a B-boost than does H, $(Q \& J)$ gets less disconfirmation in the case where $\boldsymbol{x}$ disconfirms J. This too makes sense on the [LR] measure, though I will spare the details. Their intuitions are that this is worse than the irrelevant conjunction case, and is not solved by the use of [LR]. Interesting new measures are offered. Again, this seems to our tester to reflect the tension between Bayes boosts and good tests.

What They Call Confirmation We Call Mere "Fit" or "Accordance"

> In opposition to [the] inductivist attitude, I assert that $C(H,\boldsymbol{x})$ must not be interpreted as the degree of corroboration of H by $\boldsymbol{x}$, unless $\boldsymbol{x}$ reports the results of *our sincere efforts to overthrow H*. The requirement of sincerity cannot be formalized – no more than the inductivist requirement that $\boldsymbol{x}$ must represent our total observational knowledge. (Popper 1959, p. 418, substituting H for h; $\boldsymbol{x}$ for e)

Sincerity! Popper never held that severe tests turned on a psychological notion, but he was at a loss to formalize severity. A fuller passage from Popper (1959) is worth reading if you get a chance.[6] All the measures of confirmation, be it R or LR, or one of the others, count merely as "fit" or "accordance" measures to Popper and to the severe tester. They may each be relevant for different problems – that there are different dimensions for fit is to be expected. These measures do not capture what's needed to determine if much (or anything) has been done to find H is flawed. What we need to add are the associated error probabilities. Error probabilities do not enter into these standard confirmation theories – which isn't to say they couldn't. If R is used and observed to be r, we want to compute $\Pr(R > r; \sim(H \& J))$. Here, the probability of getting $R > 1$ is maximal (since $(H \& J)$ entails $\boldsymbol{x}$), even if $\sim(H \& J)$ is true. So $\boldsymbol{x}$ is "bad evidence,

[6] "I must insist that $C(h, e)$ can be interpreted as degree of corroboration only if e is *a report on the severest tests we have been able to design*. It is this point that marks the difference between the attitude of the inductivist, or verificationist, and my own attitude. The inductivist or verificationist wants *affirmation* for his hypothesis. He hopes to make it *firmer* by his evidence e and he looks out for '*firmness*' – for '*confirmation*.' . . . Yet if e is *not* a report about the results of our sincere attempts to overthrow h, then we shall simply deceive ourselves if we think we can interpret $C(h,e)$ as degree of corroboration, or anything like it." (Popper 1959, p. 418).

no test" (BENT) for the conjunction.[7] It's not a psychological "sincerity" being captured; nor is it purely context free. Popper couldn't capture it as he never made the error probability turn.

Time prevents us from entering multiple other rooms displaying paradoxes of confirmation theory, where we'd meet up with such wonderful zombies as the white shoe confirming all ravens are black, and the "grue" paradox, which my editor banished from my 1996 book. (See Skyrms 1986.) Enough tears have been shed. Yet they shouldn't be dismissed too readily; they very often contain a puzzle of deep relevance for statistical practice. There are two reasons the tacking paradox above is of relevance to us. The first concerns a problem that arises for both Popperians and Bayesians. There is a large-scale theory *T* that predicts $\boldsymbol{x}$, and we want to discern which portion of *T* to credit. Severity says: do not credit those portions that could not have been found false, even if they're false. They are poorly tested. This may not be evident until long after the experiment. We don't want to say there is evidence for a large-scale theory such as GTR just because one part was well tested. On the other hand, it may well be that all relativistic theories with certain properties have passed severely.

Second, the question of whether to measure support with a Bayes boost or with posterior probability arises in Bayesian statistical inference as well. When you hear that what you want is some version of probabilism, be sure to ask if it's a boost (and if so which kind) or a posterior probability, a likelihood ratio, or something else. Now statisticians might rightly say, we don't go around tacking on hypotheses like this. True, the Bayesian epistemologist invites trouble by not clearly spelling out corresponding statistical models. They seek a formal logic, holding for statements about radiation, deflection, fish, or whatnot. I think this is a mistake. That doesn't preclude a general account for statistical inference; it just won't be purely formal.

Statistical Foundations Need Philosophers of Statistics

> The idea of putting probabilities over hypotheses delivered to philosophy a godsend, an entire package of superficiality. (Glymour 2010, p. 334)

Given a formal epistemology, the next step is to use it to represent or justify intuitive principles of evidence. The problem to which Glymour is alluding is this: you can start with the principle you want your confirmation logic to reflect, and then *reconstruct* it using probability. The task, for the formal epistemologist, becomes the problem of assigning priors and likelihoods that mesh with the principle you want to defend. Here's an example. Some think

[7] The real problem is that $\Pr(x; H \,\&\, J) = \Pr(x; H \,\&\, {\sim}J)$.

that GTR got more confirmation than a rival theory (e.g., Brans-Dicke theory) because the latter is made to fit the data thanks to adjustable parameters (Jefferys and Berger 1992). Others think the fact it had adjustable parameters does not alter the confirmation (Earman 1992). They too can reconstruct the episode so that Brans-Dicke pays no penalty. The historical episode can be "rationally reconstructed" to accord with either philosophical standpoint.

Although the problem of statistical inference is only a small part of what today goes under the umbrella of formal epistemology, progress in the statistics wars would advance more surely if philosophers regularly adopted the language of statistics. Not only would we be better at the job of clarifying the conceptual discomforts among practitioners of statistics and modeling, some of the classic problems of confirmation could be scotched using the language of random variables and their distributions.[8] Philosophy of statistics had long been ahead of its time, in the sense of involving genuinely interdisciplinary work with statisticians, scientists, and philosophers of science. We need to return to that. There are many exceptions, of course; yet to try to list them would surely make me guilty of leaving several out.

[8] For a discussion and justification of the use of "random variables," see Mayo (1996).

Tour II Falsification, Pseudoscience, Induction

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

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

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

2.3 Popper, Severity, and Methodological Probability

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

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

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

There are four key, interrelated themes:

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

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

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

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

(3) Observations Are Not Given. Popper rejected the time-honored empiricist assumption that observations are known relatively unproblematically. If they are at the "foundation," it is only because there are apt methods for testing their validity. We dub claims observable *because* or to the extent that they are open to stringent checks. (Popper: "anyone who has learned the relevant technique can test it" (1959, p. 99).) Accounts of hypothesis appraisal that start with "evidence $\boldsymbol{x}$," as in confirmation logics, vastly oversimplify the role of data in learning.

(4) Corroboration Not Confirmation, Severity Not Probabilism. Last, there is his radical view on the role of probability in scientific inference. Rejecting probabilism, Popper not only rejects Carnap-style logics of confirmation, he denies scientists are interested in highly probable hypotheses (in any sense). They seek bold, informative, interesting conjectures and ingenious and severe attempts to refute them. If one uses a logical notion of probability, as philosophers (including Popper) did at the time, the high content theories are highly improbable; in fact, Popper said universal theories have 0 probability. (Popper also talked of statistical probabilities as propensities.)

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

Demarcation and Investigating Bad Science

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

Following Laudan, philosophers tend to shy away from saying anything general about science versus pseudoscience – the predominant view is that there is no such thing. Some say that there's at most a kind of "family resemblance" amongst domains people tend to consider scientific (Dupré 1993, Pigliucci 2010, 2013). One gets the impression that the demarcation task is being left to committees investigating allegations of poor science or fraud. They are forced to articulate what to count as fraud, as bad statistics, or as mere questionable research practices (QRPs). People's careers depend on their ruling: they have "skin in the game," as Nassim Nicholas Taleb might say (2018). The best one I know – the committee investigating fraudster Diederik Stapel – advises making philosophy of science a requirement for researchers (Levelt Committee, Noort Committee, and Drenth Committee 2012). So let's not tell them philosophers haven given up on it.

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

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

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

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

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

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

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

A valuable idea to take from Popper is that probability in learning attaches to a method of conjecture and refutation, that is to testing: it is *methodological probability*. An error probability is a special case of a methodological probability. We want methods with a high probability of teaching us (and machines) how to distinguish approximately correct and incorrect interpretations of data, even leaving murky cases in the middle, and how to advance knowledge of detectable, while strictly unobservable, effects.

The choices for probability that we are commonly offered are stark: "in here" (beliefs ascertained by introspection) or "out there" (frequencies in long runs, or chance mechanisms). This is the "epistemology" versus "variability" shoehorn we reject (Souvenir D). To qualify the method by which *H* was tested, frequentist performance is necessary, but it's not sufficient. The assessment must be relevant to ensuring that claims have been put to severe tests. You can talk of a test having a type of *propensity* or capability to have discerned flaws, as Popper did at times. A highly explanatory, high-content theory, with interconnected tentacles, has a higher probability of having flaws discerned than low-content theories that do not rule out as much. Thus, when the bolder, higher content, theory stands up to testing, it may earn higher overall severity than the one with measly content. That a theory is plausible is of little interest, in and of itself; what matters is that it is *im*plausible for it to have passed these tests were it false or incapable of adequately solving its set of problems. It is the fuller, unifying, theory developed in the course of solving interconnected problems that enables severe tests.

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

Falsification Is Rarely Deductive. It is rare for any interesting scientific hypotheses to be logically falsifiable. This might seem surprising given all the applause heaped on falsifiability. For a scientific hypothesis *H* to be deductively falsified, it would be required that some observable result taken together with *H* yields a logical contradiction (A & ~A). But the only theories that deductively prohibit observations are of the sort one mainly finds in philosophy books: All swans are white is falsified by a single non-white swan. There are some statistical claims and contexts, I will argue, where it's possible to achieve or come close to deductive falsification: claims such as, these data are independent and identically distributed (IID). Going beyond a mere denial to replacing them requires more work.

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

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

H: All infectious agents have nucleic acid.

Any infectious agent free of nucleic acid would be *anomalous* for *H* – meaning it goes against what *H* claims. A separate step is required to decide when *H*'s anomalies should count as falsifying *H*. There needn't be a cut-off so much as a standpoint as to when continuing to defend *H* becomes bad science. Prion researchers weren't looking to test the central dogma of biology, but to understand kuru and related diseases. The anomaly erupted only because kuru appeared to be transmitted by a protein alone, by changing a normal protein shape into an abnormal fold. Stanley Prusiner called the infectious protein a prion – for which he received much grief. He thought, at first, he'd made a mistake "and was puzzled when the data kept telling me that our preparations contained protein but not nucleic acid" (Prusiner 1997). The anomalous results would not go away and, eventually, were demonstrated via experimental transmission to animals. The discovery of prions led to a "revolution" in molecular biology, and Prusiner received a Nobel prize in 1997. It is *logically* possible that nucleic acid is somehow involved. But continuing to block the falsification of *H* (i.e., block the "protein only" hypothesis) precludes learning more about prion diseases, which now include Alzheimer's. (See Mayo 2014a.)

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

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

The Popperian (Methodological) Falsificationist Is an Error Statistician

When is a statistical hypothesis to count as falsified? Although extremely rare events may occur, Popper notes:

> such occurrences would not be physical effects, because, on account of their immense improbability, *they are not reproducible at will* . . . If, however, we find *reproducible*

deviations from a macro effect . . . deduced from a probability estimate . . . then we must assume that the probability estimate is *falsified.* (Popper 1959, p. 203)

In the same vein, we heard Fisher deny that an "isolated record" of statistically significant results suffices to warrant a reproducible or genuine effect (Fisher 1935a, p. 14). Early on, Popper (1959) bases his statistical falsifying rules on Fisher, though citations are rare. Even where a scientific hypothesis is thought to be deterministic, inaccuracies and knowledge gaps involve error-laden predictions; so our methodological rules typically involve inferring a statistical hypothesis. Popper calls it a *falsifying hypothesis*. It's a hypothesis inferred in order to falsify some other claim. A first step is often to infer an anomaly is real, by falsifying a "due to chance" hypothesis.

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

This is hugely problematic for Popper because being a strict Popperian means never having to justify a claim as true or a method as reliable. After all, this was part of Popper's escape from induction. The problem is this: Popper's account rests on severe tests, tests that would probably falsify claims if false, but he cannot warrant saying a method is probative or severe, because that would mean it was reliable, which makes Popperians squeamish. It would appear to concede to his critics that Popper has a "whiff of induction" after all. But it's not inductive enumeration. Error statistical methods (whether from statistics or informal) can supply the severe tests Popper sought. This leads us to Pierre Duhem, physicist and philosopher of science.

Duhemian Problems of Falsification

Consider the simplest form of deductive falsification: If H entails observation O, and we observe $\sim O$, then we infer $\sim H$. To infer $\sim H$ is to infer H is false, or there is some discrepancy in what H claims about the phenomenon in question. As with any argument, in order to *detach* its conclusion (without which there is no *inference*), the premises must be true or approximately true. But *O* is derived only with the help of various additional claims. In statistical contexts, we may group these under two umbrellas: auxiliary factors linking substantive and statistical claims, A_1 & ⋯ & A_n, and assumptions of the statistical model E_1 & ⋯ & E_k. You are to imagine a great big long conjunction of factors, in the following argument:

1. If *H* & (A_1 & ⋯ & A_n) & (E_1 & ⋯ & E_k), then *O*.
2. ~*O*.
3. Therefore, either ~*H* or ~A_1 or . . . or ~A_n or ~E_1 or . . . or ~E_k.

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

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

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

How can Popper, if he is really a deductivist, solve Duhem in order to falsify? At best he’d subject each of the conjuncts to as stringent a test as possible, and falsify accordingly. This still leaves, Popper admits, a disjunction of non-falsified hypotheses (he thought infinitely many)! Popperian philosophers of science advise you to choose a suitable overall package of hypotheses, assumptions, auxiliaries, on a set of criteria: simplicity, explanatory power, unification and so on. There's no agreement on which, nor how to define them. On this view, you can't really solve Duhem, you accept or "prefer" (as Popper said) the large-scale research program or paradigm as a whole. It's intended to be an advance over *bon sens* in blocking certain types of tinkering (see Section 2.4). There's a remark in the Popper museum display I only recently came across:

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

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

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

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

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

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

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

Souvenir E: An Array of Questions, Problems, Models

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

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

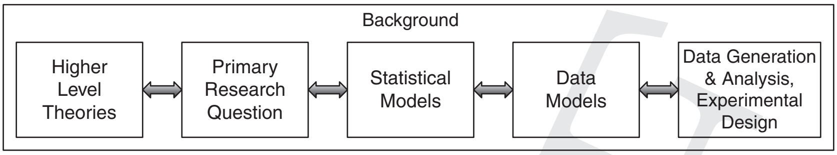
Background
Higher Level Theories
Primary Research Question
Statistical Models
Data Models
Data Generation & Analysis, Experimental Design

Figure 2.1 Array of questions, problems, models.

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

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

Souvenir F: Getting Free of Popperian Constraints on Language

Popper allows that anyone who wants to define induction as the procedure of corroborating by severe testing is free to do so; and I do. Free of the bogeyman that induction must take the form of a probabilism, let's get rid of some linguistic peculiarities inherited by current-day Popperians (critical rationalists). They say things such as: it is *warranted* to infer (prefer or believe) *H* (because *H* has passed a severe test), but there is no *justification* for *H* (because "justifying" *H* would mean *H* was true or highly probable). In our language, if *H* passes a severe test, you can say it is warranted, corroborated, justified – along with whatever qualification is appropriate. I tend to use "warranted." The Popperian "hypothesis *H* is corroborated by data ***x***" is such a tidy abbreviation of "*H* has passed a severe test with ***x***" that we may use the two interchangeably. I've already co-opted Popper's description of science as *problem solving*. A hypothesis can be seen as a potential solution to a problem (Laudan 1978). For example, the theory of protein folding purports to solve the problem of how pathological prions are transmitted. The problem might be to explain, to predict, to unify, to suggest new problems, etc. When we severely probe, it's not for falsity per se, but to investigate if a problem has been adequately solved by a model, method, or theory.

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

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

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

Traveler 2: I've been thinking about this. On your first point, Popper confuses things by making it sound as if he's asking: *When is a theory unscientific?* What he is actually asking or should be asking is: *When is an inquiry into a theory, or an appraisal of claim H, unscientific*? We want to distinguish meritorious modes of inquiry from those that are BENT. If the test methods enable ad hoc maneuvering, sneaky face-saving devices, then the inquiry – the handling and use of data – is unscientific. Despite being logically falsifiable, theories can be rendered immune from falsification by means of cavalier methods for their testing. Adhering to a falsified theory no matter what is poor science. Some areas have so much noise and/or flexibility that they can't or won't distinguish warranted from unwarranted explanations of failed predictions. Rivals may find flaws in one another's inquiry or model, but the criticism is not constrained by what's actually responsible. This is another way inquiries can become unscientific.[1]

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

She continues:

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

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

> *A scientific inquiry or test: must be able to embark on a reliable probe to pinpoint blame for anomalies (and use the results to replace falsified claims and build a repertoire of errors).*

The parenthetical remark isn't absolutely required, but is a feature that greatly strengthens scientific credentials. Without solving, not merely embarking on, some Duhemian problems there are no interesting falsifications. The ability or inability to pin down the source of failed replications – a familiar occupation these days – speaks to the scientific credentials of an inquiry. At any given time, even in good sciences there are anomalies whose sources haven't been traced – unsolved Duhemian problems – generally at "higher" levels of the theory-data array. Embarking on solving these is the impetus for new conjectures. Checking test assumptions is part of working through the Duhemian maze. The reliability requirement is: infer claims just to the extent that they pass severe tests. There's no sharp line for demarcation, but when these requirements are absent, an inquiry veers into the realm of questionable science or pseudoscience. Some physicists worry that highly theoretical realms can't be expected to be constrained by empirical data. Theoretical constraints are also important. We'll flesh out these ideas in future tours.

2.4 Novelty and Severity

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

This "something else that must come out right" is often called a "novel" predictive success. Whether or not novel predictive success is required is a very old battle that parallels debates between frequentists and inductive logicians, in both statistics and philosophy of science, for example, between Mill and Peirce and Popper and Keynes. Walking up the ramp from the ground floor to the gallery of Statistics, Science, and Pseudoscience, the novelty debate is used to intermix Popper and statistical testing.

When Popper denied we can capture severity formally, he was reflecting an astute insight: there is a tension between the drive for a logic of confirmation and our strictures against practices that lead to poor tests and ad hoc hypotheses. Adhering to the former downplays or blocks the ability to capture the latter, which demands we go beyond the data and hypotheses. Imre Lakatos would say we need to know something about the *history* of the hypothesis: how was it developed? Was it the result of deliberate and ad hoc attempts to spare one's theory from refutation? Did the researcher continue to adjust her theory in the face of an anomaly or apparent discorroborating result? (He called these "exception incorporations".) By contrast, the confirmation theorist asks: why should it matter how the hypothesis inferred was arrived at, or whether data-dependent selection effects were operative? When holders of the Likelihood Principle (LP) wonder why data can't speak for themselves, they're echoing the logical empiricist (Section 1.4). Here's Popperian philosopher Alan Musgrave:

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

John Maynard Keynes likewise held that the ". . . question as to whether a particular hypothesis happens to be propounded before or after examination of [its instances] is quite irrelevant (Keynes 1921/1952, p. 305). Logics of confirmation ran into problems because they insisted on purely formal or syntactical criteria of confirmation that, like deductive logic, "should contain no reference to the specific subject-matter" (Hempel 1945, p. 9) in question. The Popper–Lakatos school attempts to avoid these shortcomings by means of novelty requirements:

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

For decades Popperians squabbled over how to define novel predictive success. There's (1) *temporal novelty* – the data were not already available before the hypothesis was erected (Popper, early); (2) *theoretical novelty* – the data were not already predicted by an existing hypothesis (Popper, Lakatos); and (3) *use-*

novelty – the data were not used to construct or select the hypothesis. Defining novel success is intimately linked to defining Popperian severity.

Temporal novelty is untenable: known data (e.g., the perihelion of Mercury, anomalous for Newton) are often strong evidence for theories (e.g., GTR). Popper ultimately favored theoretical novelty: *H* passes a severe test with ***x***, when *H* entails ***x***, and ***x*** is theoretically novel – according to a letter he sent me. That, of course, freed me to consider my own notion as distinct. (We replace "entails" with something like "accords with.") However, as philosopher John Worrall (1978, pp. 330–1) shows, to require theoretical novelty prevents passing *H* with severity, so long as there's already a hypothesis that predicts the data or phenomenon ***x*** (it's not clear which). Why should the first hypothesis that explains ***x*** be better tested?

I take the most promising notion of novelty to be a version of use-novelty: *H* passes a test with data ***x*** severely, so long as ***x*** was not used to construct *H* (Worrall 1989). Data can be known, so long as they weren't used in building *H*, presumably to ensure *H* accords with ***x***. While the idea is in sync with the error statistical admonishment against "peeking at the data" and finding your hypothesis in the data – it's far too vague as it stands. Watching this debate unfold in philosophy, I realized none of the notions of novelty were either sufficient or necessary for a good test (Mayo 1991).

You will notice that statistical researchers go out of their way to state a prediction at the start of a paper, presenting it as temporally novel, and if *H* is temporally novel, it also satisfies use-novelty. If *H* came first, the data could not have been used to arrive at *H*. This stricture is desirable, but to suppose it suffices for a good test grows out of a discredited empiricist account where the data are *given* rather than the product of much massaging and interpretation. There is as much opportunity for bias to arise in interpreting or selectively reporting results, with a known hypothesis, as there is in starting with data and artfully creating a hypothesis. Nor is violating use-novelty a matter of the implausibility of *H*. On the contrary, popular psychology thrives by seeking to explain results by means of hypotheses expected to meet with approval, at least in a given political tribe. Preregistration of the detailed protocol is supposed to cure this. We come back to this.

Should use-novelty be *necessary* for a good test? Is it ever okay to use data to arrive at a hypothesis *H* as well as to test *H* – even if the data use ensures agreement or disagreement with *H*? The answers, I say, are no and yes, respectively. Violations of use-novelty need not be pejorative. A trivial example: count all the people in the room and use it to fix the parameter of the number in the room. Or, less trivially, think of confidence intervals: we use the data to form the interval estimate. The estimate is really a hypothesis about the value of the parameter. The same data warrant the hypothesis constructed! Likewise, using the same data to arrive at and test assumptions of statistical models can be entirely reliable. What matters is not novelty, in any of the senses, but severity in the error statistical sense. Even where our intuition is to prohibit use-novelty violations, the requirement is murky. We should instead consider specific ways that severity can be violated. Let's define:

> *Biasing Selection Effects*: when data or hypotheses are selected or generated (or a test criterion is specified), in such a way that the minimal severity requirement is violated, seriously altered, or incapable of being assessed.[2]

Despite using this subdued label, it's too irresistible to banish entirely a cluster of colorful terms for related gambits – double-counting, cherry picking, fishing, hunting, significance seeking, searching for the pony, trying and trying again, data dredging, monster barring, look elsewhere effect, and many others besides – unless we're rushing. New terms such as *P*-hacking are popular, but don't forget that these crooked practices are very old.[3]

To follow the Popper–Lakatos school (although entailment is too strong):

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

This describes corroborating a claim, it is "strong" severity. Weak severity denies *H* is warranted if the test method would probably have passed *H* even if false. While severity got its start in this Popperian context, in future excursions, we will need more specifics to describe both clauses (S-1) and (S-2).

2.5 Fallacies of Rejection and an Animal Called NHST

One of Popper's prime examples of non-falsifiable sciences was Freudian and Adlerian psychology, which gave psychologist Paul Meehl conniptions

[2] As noted earlier, I follow Ioannidis in using bias this way, in speaking of selections.

[3] For a discussion of novelty and severity in philosophy of science, see Chapter 8 of Mayo (1996). Worrall and I have engaged in a battle over this in numerous places (Mayo 2010d, Worrall 1989, 2010). Related exchanges include Mayo 2008, Hitchcock and Sober 2004.

because he was a Freudian as well as a Popperian. Meehl castigates Fisherian significance tests for providing a sciency aura to experimental psychology, when they seem to violate Popperian strictures: "[T]he almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas . . . is basically unsound, poor scientific strategy . . ." (Meehl 1978, p. 817). Reading Meehl, Lakatos wrote, "one wonders whether the function of statistical techniques in the social sciences is not primarily to provide a machinery for producing phoney corroborations and . . . an increase in pseudo-intellectual garbage" (Lakatos 1978, pp. 88–9, note 4).

Now Meehl is a giant when it comes to criticizing statistical practice in psychology, and a good deal of what contemporary critics are on about was said long ago by him. He's wrong, though, to pin the blame on "Sir Ronald" (Fisher). Corroborating substantive theories merely by means of refuting the null? Meehl may be describing what is taught and permitted in the "soft sciences," but the practice of moving from statistical to substantive theory violates the testing methodologies of both Fisher and Neyman–Pearson. I am glad to see Gerd Gigerenzer setting the record straight on this point, given how hard he, too, often is on Fisher:

> It should be recognized that, according to Fisher, rejecting the null hypothesis is not equivalent to accepting the efficacy of the cause in question. The latter cannot be established on the basis of one single experiment, but requires obtaining more significant results when the experiment, or an improvement of it, is repeated at other laboratories or under other conditions. (Gigerenzer et al. 1989, pp. 95–6)

According to Gigerenzer et al., "careless writing on Fisher's part, combined with selective reading of his early writings has led to the identification of the two, and has encouraged the practice of demonstrating a phenomenon on the basis of a single statistically significant result" (ibid., p. 96). I don't think Fisher can be accused of carelessness here; he made two crucial clarifications, and the museum display case bears me out. The first is that "[W]e need, not an isolated record, but a reliable method of procedure" (Fisher 1935a, p. 14), from Excursion 1. The second is Fisher's requirement that even a genuine statistical effect *H* fails to warrant a substantive research hypothesis H^*. Using "$\not\Rightarrow$" to abbreviate "does not imply": $H \not\Rightarrow H^*$.

Here's David Cox defining significance tests over 40 years ago:

> . . . we mean by a significance test a procedure for measuring the consistency of data with a null hypothesis . . . there is a function $\mathrm{d} = \mathrm{d}(\boldsymbol{y})$ of the observations, called a test statistic, and such that the larger is $\mathrm{d}(\boldsymbol{y})$ the stronger is the inconsistency of $\boldsymbol{y}$ with H_0, in the respect under study . . . we need to be able to compute, at least approximately,

> $$p_{\text{obs}} = \Pr(\mathrm{d} \geq \mathrm{d}(\boldsymbol{y}_{\text{obs}}); H_0)$$
>
> called the observed level of significance [or *P*-value].
>
> Application of the significance test consists of computing approximately the value of p_{obs} and using it as a summary measure of the degree of consistency with H_0, in the respect under study. (Cox 1977, p. 50; replacing t with d)

Statistical test requirements follow non-statistical tests, Cox emphasizes, though at most H_0 entails some results with high probability. Say 99% of the time the test would yield $\{d < d_0\}$, if H_0 adequately describes the data-generating mechanism where d_0 abbreviates $\mathrm{d}(\boldsymbol{x}_0)$. Observing $\{d \geq d_0\}$ indicates inconsistency with H_0 in the respect tested. (Implicit alternatives, Cox says, "lurk in the undergrowth," given by the test statistic.) So significance tests reflect statistical *modus tollens*, and its reasoning follows severe testing – BUT, an isolated low *P*-value won't suffice to infer a genuine effect, let alone a research claim. Here's a list of *fallacies of rejection.*

1. The reported (nominal) statistical significance result is *spurious* (it's not even an actual *P*-value). This can happen in two ways: biasing selection effects, or violated assumptions of the model.
2. The reported statistically significant result is genuine, but it's an isolated effect not yet indicative of a genuine experimental phenomenon. (Isolated low *P*-value $\nRightarrow$ *H*: statistical effect.)
3. There's evidence of a genuine statistical phenomenon but either (i) the magnitude of the effect is less than purported, call this a *magnitude error*,[4] or (ii) the substantive interpretation is unwarranted ($H \nRightarrow H^*$).

I will call an *audit* of a *P*-value, a check of any of these concerns, generally in order, depending on the inference. That's why I place the background information for auditing throughout our "series of models" representation (Figure 2.1). Until audits are passed, the relevant statistical inference is to be reported as "unaudited." Until 2 is ruled out, it's a mere "indication," perhaps, in some settings, grounds to get more data.

Meehl's criticism is to a violation described in 3(ii). Like many criticisms of significance tests these days, it's based on an animal that goes by the acronym NHST (null hypothesis significance testing). What's wrong with NHST in relation to Fisherian significance tests? The museum label says it for me:

[4] This is the term used by Andrew Gelman.

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

We can add: if it permits biasing selection effects it's triply guilty. Too often NHST refers to a monster describing highly fallacious uses of Fisherian tests, introduced in certain social sciences. I now think it's best to drop the term NHST. Statistical tests will do, although our journey requires we employ the terms used in today's battles.

Shall we blame the wise and sagacious Meehl with selective reading of Fisher? I don't know. Meehl gave me the impression that he was irked that using significance tests seemed to place shallow areas of psychology on a firm falsification footing; whereas, more interesting, deep psychoanalytic theories were stuck in pseudoscientific limbo. He and Niels Waller give me the honor of being referred to in the same breath as Popper and Salmon:

> For the corroboration to be strong, we have to have 'Popperian risk', … 'severe test' [as in Mayo], or what philosopher Wesley Salmon called a *highly improbable coincidence* ["damn strange coincidence"]. (Meehl and Waller 2002, p. 284)

Yet we mustn't blur an argument from coincidence merely to a real effect, and one that underwrites arguing from coincidence to research hypothesis *H**. Meehl worried that, by increasing the sample size, trivial discrepancies can lead to a low *P*-value, and using NHST, evidence for *H** too readily attained. Yes, if you plan to perform an illicit inference, then whatever makes the inference easier (increasing sample size) is even more illicit. Since proper statistical tests block such interpretations, there's nothing anti-Popperian about them.

The fact that selective reporting leads to unreplicable results is an *asset* of significance tests: If you obtained your apparently impressive result by violating Fisherian strictures, preregistered tests will give you a much deserved hard time when it comes to replication. On the other hand, evidence of a statistical effect *H* does give a B-boost to *H**, since if *H** is true, a statistical effect follows (statistical affirming the consequent).

Meehl's critiques rarely mention the methodological falsificationism of Neyman and Pearson. Why is the field that cares about power – which is defined in terms of N-P tests – so hung up on simple significance tests? We'll disinter the answer later on. With N-P tests, the statistical alternative to the null hypothesis is made explicit: the null and alternative exhaust the possibilities. There can be no illicit jumping of levels from statistical to causal (from H to H^*). Fisher didn't allow jumping either, but he was less explicit. Statistically significant increased yields in Fisher's controlled trials on fertilizers, as Gigerenzer notes, are intimately linked to a causal alternative. If the fertilizer does not increase yield (H^* is false, so ~H^* is true), then no statistical increase is expected, if the test is run well.[5] Thus, finding statistical increases (rejecting H_0) is grounds to falsify ~H^* and find evidence of H^*. Unlike the typical psychology experiment, here rejecting a statistical null very nearly warrants a statistical causal claim. If you want a statistically significant effect to (statistically) warrant H^* show:

> If ~H^* is true (research claim H^* is false), then H_0 won't be rejected as inconsistent with data, at least not regularly.

Psychology should move to an enlightened reformulation of N-P and Fisher (see Section 3.3). To emphasize the Fisherian (null hypothesis only) variety, we follow the literature in calling them "simple" significance tests. They are extremely important in their own right: They are the basis for testing assumptions without which statistical methods fail scientific requirements. View them as just one member of a panoply of error statistical methods.

Statistics Can't Fix Intrinsic Latitude. The problem Popper found with Freudian and Adlerian psychology is that any observed behavior could be readily interpreted through the tunnel of either theory. Whether a man jumped in the water to save a child, or if he failed to save her, you can invoke Adlerian inferiority complexes, or Freudian theories of sublimation or Oedipal complexes (Popper 1962, p. 35). Both Freudian and Adlerian theories can explain whatever happens. This latitude has nothing to do with statistics. As we learned from Exhibit (vi), Section 2.3, we should really speak of the latitude offered by the overall inquiry: research question, auxiliaries, and interpretive rules. If it has self-sealing facets to account for any data, then it fails to probe with severity. Statistical methods cannot fix this. Applying statistical methods is just window dressing. Notice that Freud/Adler, as Popper describes them, are amongst the few cases where the latitude really is part of the theory or terminology. It's not obvious that Popper's theoretical novelty bars this, unless one of Freud/Adler is deemed first. We've arrived at the special topical installation on:

[5] Gigerenzer calls such a "no increase" hypothesis the substantive null hypothesis.

2.6 The Reproducibility Revolution (Crisis) in Psychology

> I was alone in my tastefully furnished office at the University of Groningen. . . . I opened the file with the data that I had entered and changed an unexpected 2 into a 4; then, a little further along, I changed a 3 into a 5. . . . When the results are just not quite what you'd so badly hoped for; when you know that that hope is based on a thorough analysis of the literature; . . . then, surely, you're entitled to adjust the results just a little? . . . I looked at the array of data and made a few mouse clicks to tell the computer to run the statistical analyses. When I saw the results, the world had become logical again. (Stapel 2014, p. 103)

This is Diederik Stapel describing his "first time" – when he was still collecting data and not inventing them. After the Stapel affair (2011), psychologist Daniel Kahneman warned that he "saw a train wreck looming" for social psychology and called for a "daisy chain" of replication to restore credibility to some of the hardest hit areas such as priming studies (Kahneman 2012). Priming theory holds that exposure to an experience can unconsciously affect subsequent behavior. Kahneman (2012) wrote: "right or wrong, your field is now the poster child for doubts about the integrity of psychological research." One of the outgrowths of this call was the 2011–2015 Reproducibility Project, part of the Center for Open Science Initiative at the University of Virginia. In a nutshell: This is a crowd-sourced effort to systematically subject published statistically significant findings to checks of reproducibility. In 2011, 100 articles from leading psychology journals from 2008 were chosen; in August of 2015, it was announced only around 33% could be replicated (depending on how that was defined). Whatever you think of the results, it's hard not to be impressed that a field could organize such a self-critical project, obtain the resources, and galvanize serious-minded professionals to carry it out.

First, on the terminology: The American Statistical Association (2017, p. 1) calls a study "reproducible if you can take the original data and the computer code used . . . and reproduce all of the numerical findings . . ." In the case of Anil Potti, they couldn't reproduce his numbers. By contrast, replicability refers to "the act of repeating an entire study, independently of the original investigator without the use of the original data (but generally using the same methods)" (ibid.).[6] The Reproducibility Project, however, is really a replication project (as the ASA defines it). These points of terminology shouldn't affect

[6] This is a "direct replication," whereas a "conceptual replication" probes the same hypothesis but through a different phenomenon.

our discussion. The Reproducibility Project is appealing to what most people have in mind in saying a key feature of science is reproducibility, namely replicability.

So how does the Reproducibility Project proceed? A team of (self-selected) knowledgeable replicators, using a protocol that is ideally to be approved by the initial researchers, reruns the study on new subjects. A failed replication occurs when there's a non-statistically significant or *negative* result, that is, a *P*-value that is not small (say >0.05). Does a negative result mean the original result was a false positive? Or that the attempted replication was a false negative? The interpretation of negative statistical results is itself controversial, particularly as they tend to keep to Fisherian tests, and effect sizes are often fuzzy. When RCTs fail to replicate observational studies, the presumption is that, were the effect genuine, the RCTs would have found it. That is why they are taken as an indictment of the observational study. But here, you could argue, the replication of the earlier research is not obviously a study that checks its correctness. Yet that would be to overlook the strengthened features of the replications in the 2011 project: they are preregistered, and are designed to have high power (against observed effect sizes). What is more, they are free of some of the "perverse incentives" of usual research. In particular, the failed replications are guaranteed to be published in a collective report. They will not be thrown in file drawers, even if negative results ensue.

Some ironic consequences immediately enter in thinking about the project. The replication researchers in psychology are the same people who hypothesize that a large part of the blame for lack of replication may be traced to the reward structure: to incentives to publish surprising and sexy studies, coupled with an overly flexible methodology opening the door to promiscuous QRPs. Call this the *flexibility, rewards, and bias* hypothesis. Supposing this hypothesis is correct, as is quite plausible, what happens when non-replication becomes sexy and publishable? Might non-significance become the new significance? *Science* likely wouldn't have published individual failures to replicate, but they welcomed the splashy OSC report of the poor rate of replication they uncovered, as well as back-and-forth updates by critics. Brand new fields of meta-research open up for replication specialists, all ostensibly under the appealing banner of improving psychology. Some ask: should authors be prey to results conducted by a self-selected group – results that could obviously impinge rather unfavorably on them? Many say no and even liken the enterprise to a witch hunt. Kahneman (2014) called for "a new etiquette" requiring original authors to be consulted on protocols:

> . . . tension is inevitable when the replicator does not believe the original findings and intends to show that a reported effect does not exist. The relationship between replicator and author is then, at best, politely adversarial. The relationship is also radically asymmetric: the replicator is in the offense, the author plays defense. The threat is one-sided because of the strong presumption in scientific discourse that more recent news is more believable. (p. 310)

It's not hard to find potential conflicts of interest and biases on both sides. There are the replicators' attitudes – not only toward the claim under study, but toward the very methodology used to underwrite it – usually statistical significance tests. Every failed replication is seen (by some) as one more indictment of the method (never minding its use in showing irreplication). There's the replicator's freedom to stop collecting data once minimal power requirements are met, and the fact that subjects – often students, whose participation is required – are aware of the purpose of the study, revealed at the end. (They are supposed to keep it confidential over the life of the experiment, but is that plausible?) On the other hand, the door may be open too wide for the original author to blame any failed replication on lack of fidelity to nuances of the original study. Lost in the melee is the question of whether any constructive criticism is emerging.

Incidentally, here's a case where it might be argued that loss and cost functions are proper, since the outcome goes beyond statistical inference to reporting a failure to replicate Jane's study, perhaps overturning her life's research.

What Might a Real Replication Revolution in Psychology Require?

Even absent such concerns, the program seems to be missing the real issues that leap out at the average reader of the reports. The replication attempts in psychology stick to what might be called "purely statistical" issues: can we get a low *P*-value or not? Even in the absence of statistical flaws, research conclusions may be disconnected from the data used in their testing, especially when experimentally manipulated variables serve as proxies for variables of theoretical interest. A serious (and ongoing) dispute arose when a researcher challenged the team who failed to replicate her hypothesis that subjects "primed" with feelings of cleanliness, sometimes through unscrambling soap-related words, were less harsh in judging immoral such bizarre actions as whether it is acceptable to eat your dog after it has been run over. A focus on the *P*-value computation ignores the larger question of the methodological adequacy of the leap from the statistical to the substantive. Is unscrambling soap-related words an adequate proxy for the intended cleanliness variable? The less said about eating your run-over dog, the better. At this point Bayesians might argue, "We know these theories are implausible, we avoid the inferences by invoking our disbeliefs." That can work in some cases, except that the researchers find them plausible, and, more than that, can point to an entire literature on related studies, say, connecting physical and moral purity or impurity (part of "embodied cognition" e.g., Schnall et al. 2008~~)~~. The severe tester shifts the unbelievability assignment. What's unbelievable is supposing their experimental method provides evidence for purported effects! Some philosophers look to these experiments on cleanliness and morality, and many others, to appraise their philosophical theories "experimentally."[7] Whether or not this is an advance over philosophical argument, philosophers should be taking the lead in critically evaluating the methodology, in psychology and, now, in philosophy.

Our skepticism is not a denial that we may often use statistical tests to infer a phenomenon quite disparate from the experimental manipulations. Even an artificial lab setting can teach us about a substantive phenomenon "in the wild" so long as there are *testable implications* for the statistically modeled experiment. The famous experiments by Harlow, showing that monkeys prefer a cuddly mom to a wire mesh mom that supplies food (Harlow 1958), are perfectly capable of letting us argue from coincidence to what matters to actual monkeys. Experiments in social psychology are rarely like that.

The "replication revolution in psychology" won't be nearly revolutionary enough until they subject to testing the methods and measurements intended to link statistics with what they really want to know. If you are an ordinary skeptical reader, outside psychology, you're probably flummoxed that researchers blithely assume that role playing by students, unscrambling of words, and those long-standing 5, 7, or 10 point questionnaires are really measuring the intended psychological attributes. Perhaps it's taboo to express this. Imagine that Stapel had not simply fabricated his data, and he'd found that students given a mug of M&M's emblazoned with the word "capitalism" ate statistically significantly more candy than those with a scrambled word on their mug– as one of his make-believe studies proposed (Stapel 2014, pp. 127–8). Would you think you were seeing the effects of greed in action?

Psychometrician Joel Michell castigates psychology for having bought the operationalist Stevens' (1946, p. 667) "famous definition of measurement as

[7] The experimental philosophy movement should be distinguished from the New Experimentalism in philosophy.

‘the assignment of numerals to objects or events according to rules’”, a gambit he considers a deliberate and “pathological” ploy to deflect “attention from the issue of whether psychological attributes are quantitative” to begin with (Michell 2008, p. 9). It’s easy enough to have a rule for assigning numbers on a Likert questionnaire, say on degrees of moral opprobrium (never OK, sometimes OK, don’t know, always OK) if it’s not required to have an independent source of its validity. (Are the distances between units really equal, as statistical analysis requires?) I prefer not to revisit studies against which it’s easy to take pot shots. Here’s a plausible phenomenon, confined, fortunately, to certain types of people.

Macho Men: Falsifying Inquiries

I have no doubts that certain types of men feel threatened by the success of their female partners, wives, or girlfriends – more so than the other way around. I’ve even known a few. Some of my female students, over the years, confide that their boyfriends were angered when they got better grades than they did! I advise them to drop the bum immediately if not sooner. The phenomenon is backed up by field statistics (e.g., on divorce and salary differentials where a woman earns more than a male spouse, Thaler 2013[8]). As we used H (the statistical hypothesis), and H^* (a corresponding causal claim), let’s write this more general phenomenon as $\mathcal{H}$. Can this be studied in the lab? Ratliff and Oishi (2013) “examined the influence of a romantic partner’s success or failure on one’s own implicit and explicit self-esteem” (p. 688). Their statistical studies show that

> *H:* “men’s implicit self-esteem is lower when a partner succeeds than when a partner fails.” (ibid.)

To take the weakest construal, H is the statistical alternative to a “no effect” null H_0. The “treatment” is to think and write about a time their partner succeeded at something or failed at something. The effect will be a measure of “self-esteem,” obtained either explicitly, by asking: “How do you feel about yourself?” or implicitly, based on psychological tests of positive word associations (with “me” versus “other”). Subjects are randomly assigned to five “treatments”: think, write about a time your partner (i) succeeded, (ii) failed, (iii) succeeded when you failed, (iv) failed when you succeeded, or (v) a typical day (control) (ibid., p. 695). Here are a few of the several statistical null hypotheses

[8] There are some fairly strong statistics, too, of correlations between wives earning more than their husbands and divorce or marital dissatisfaction – although it is likely the disgruntlement comes from both sides.

(as no significant results are found among women, these allude to males thinking about female partners):

(a) The average implicit self-esteem is no different when subjects think about their partner succeeding (or failing) as opposed to an ordinary day.
(b) The average implicit self-esteem is no different when subjects think about their partner succeeding while the subject fails ("she does better than me").
(c) The average implicit self-esteem is no different when subjects think about their partner succeeding as opposed to failing ("she succeeds at something").
(d) The average explicit self-esteem is no different under any of the five conditions.

These statistical null hypotheses are claims about the distributions from which participants are sampled, limited to populations of experimental subjects – generally students who receive course credit. They merely assert the treated/non-treateds can be seen to come from the same populations as regards the average effect in question.

None of these nulls are able to be statistically rejected except (c)! Each negative result is anomalous for H. Should they take the research hypothesis as disconfirmed? Or as casting doubt on their test? Or should they focus on the null hypotheses that were rejected, in particular null (c). They opt for the third, viewing their results as "demonstrating that men who thought about their romantic partner's success had lower implicit self-esteem than men who thought about their romantic partner's failure (ibid., p. 698). This is a highly careful wording. It refers only to a statistical effect, restricted to the experimental subjects. That's why I write it as H. Of course they really want to infer a causal claim – the self-esteem of males studied is negatively influenced, on average, by female partner success of some sort H^*. More than that, they'd like the results to be evidence that H^* holds in the population of men in general, and speaks to the higher level theory $\mathcal{H}$.

On the face of it, it's a jumble. We do not know if these negative results reflect negatively on a research causal hypothesis – even limited to the experimental population – or whether the implicit self-esteem measure is actually picking up on something else, or whether the artificial writing assignment is insufficiently relevant to the phenomenon of interest. The auxiliaries linking the statistical and the substantive, the audit of the P-values and the statistical assumptions – all are potential sources of blame as we cast about solving the Duhemian challenge. Things aren't clear enough to say researchers *should* have regarded their research hypothesis as disconfirmed much less falsified. This is the nub of the problem.

What Might a Severe Tester Say?

I'll let her speak:

It appears from failing to reject (a) that our "treatment" has no bearing on the phenomenon of interest. It was somewhat of a stretch to suppose that thinking about her "success" (examples given are dancing, cooking, solving an algebra problem) could really be anything like the day Ann got a raise while Mark got fired. Take null hypothesis (b). It was expected that "she beat me in X" would have a greater negative impact on self-esteem than merely, "she succeeded at X." Remember these are completely different groups of men, thinking about whatever it is they chose to. That the macho man's partner bowled well one day should have been less deflating than her getting a superior score. We confess that the non-significant difference in (b) casts a shadow on whether the intended phenomenon is being picked up at all. We could have interpreted it as supporting our research hypothesis. We could view it as lending "some support to the idea that men interpret 'my partner is successful' as 'my partner is more successful than me'" (ibid., p. 698). We could have reasoned, the two conditions show no difference because any success of hers is always construed by macho man as "she showed me up." This skirts too close to viewing the data through the theory, to a *self-sealing fallacy*. Our results lead us to question that this study is latching onto the phenomenon of interest. In fact, insofar as the general phenomenon $\mathcal{H}$ (males taking umbrage at a partner's superior performance) is plausible, it would imply no effect would be found in this artificial experiment. Thus spake the severe tester.

I want to be clear that I'm not criticizing the authors for not proceeding with the severe tester's critique; it would doubtless be considered outlandish and probably would not be accepted for publication. I deliberately looked at one of the better inquiries that also had a plausible research hypothesis. View this as a futuristic self-critical researcher.

While we're at it, are these implicit self-esteem tests off the table? Why? The authors admit that *explicit* self-esteem was unaffected (in men and women). Surely if explicit self-esteem *had* shown a significant difference, they would have reported it as support for their research hypothesis. Many psychology measurements not only lack a firm, independent check on their validity; if they disagree with more direct measurements, it is easily explained away or even taken as a point in their favor. Why do no differences show up on explicit measures of self-esteem? Available reasons: Men do not want to admit their self-esteem goes down when their partner succeeds, or they might be unaware of it. Maybe so, but this assumes what hasn't been vouchsafed. Why not revisit the subjects at a later date to compare their scores on implicit self-esteem? If we find no difference from their score under the experimental manipulation, we'd have some grounds to deny it was validly measuring the effect of the treatment.

Here's an incentive: They're finding that the replication revolution has not made top psychology journals more inclined to publish non-replications – even of effects they have published. The journals want new, sexy effects. Here's sexy: stringently test (and perhaps falsify) some of the seminal measurements or types of inquiry used in psychology. In many cases we may be able to falsify given studies. If that's not exciting enough, imagine showing some of the areas now studied admit of no robust, generalizable effects. You might say it would be ruinous to set out to question basic methodology. Huge literatures on the "well established" Macbeth effect, and many others besides, might come in for question. I said it would be revolutionary for psychology. Psychometricians are quite sophisticated, but their work appears separate from replication research. Who would want to undermine their own field? Already we hear of psychology's new "spirit of self-flaggelation" (Dominus 2017). It might be an apt job for philosophers of science, with suitable expertise, especially now that these studies are being borrowed in philosophy.[9]

A hypothesis to be considered must always be: the results point to the inability of the study to severely probe the phenomenon of interest. The goal would be to build up a body of knowledge on closing existing loopholes when conducting a type of inquiry. How do you give evidence of "sincerely trying (to find flaws)?" Show that you would find the studies poorly run, if the flaws were present. When authors point to other studies as offering replication, they should anticipate potential criticisms rather than showing "once again I can interpret my data through my favored theory." The scientific status of an inquiry is questionable if it cannot or will not distinguish the correctness of inferences from problems stemming from a poorly run study. What must be subjected to grave risk are assumptions that the experiment was well run. This should apply as well to replication projects, now under way. If the producer of the report is not sufficiently self-skeptical, then we the users must be.

Live Exhibit (vii): Macho Men. Entertainment on this excursion is mostly home grown. A reenactment of this experiment will do. Perhaps hand questionnaires to some of the males after they lose to their partners in shuffle board

[9] One of the failed replications was the finding that reading a passage against free will contributes to a proclivity for cheating. Both the manipulation and the measured effects are shaky – never mind any statistical issues.

or ping pong. But be sure to include the most interesting information unreported in the study on self-esteem and partner success. Possibly it was never even looked at: What did the subjects write about? What kind of question would Mr. "My-self-esteem-goes-down-when-she-succeeds" elect to think and write about? Consider some questions that would force you to reinterpret even the statistically significant results.

Exhibit (viii): The Multiverse. Gelman and Loken (2014) call attention to the fact that even without explicitly cherry picking, there is often enough leeway in the "forking paths" between data and inference so that a researcher may be led to a desired inference. Growing out of this recognition is the idea of presenting the results of applying the same statistical procedure, but with different choices along the way (Steegen et al. 2016). They call it a *multiverse analysis*. One lists the different choices thought to be plausible at each stage of data processing. The multiverse displays "… which constellation of choices corresponds to which statistical result" (p. 707).

They consider an example from 2012 purporting to show that single women prefer Obama to Romney when they are highly fertile; the reverse when they're at low fertility.

> In two studies with relatively large and diverse samples of women, we found that ovulation had different effects on women's religious and political orientation depending on whether women were single or in committed relationships. Ovulation led single women to become more socially liberal, less religious, and more likely to vote for Barack Obama. (Durante et al. 2013, p. 1013)

Unlike the Macho Men study, this one's not intuitively plausible. In fact, it was pummeled so vehemently by the public that it had to be pulled off CNN.[10] Should elaborate statistical criticism be applied to such studies? I had considered them only human interest stories. But Gelman rightly finds in them some general lessons.

One of the choice points in the ovulation study would be where to draw the line at "highly fertile" based on days in a woman's cycle. It wasn't based on any hormone check but on an online questionnaire asking when they'd had their last period. There's latitude in using such information to decide whether to place someone in a low or high fertility group (Steegen et al. 2016, p. 705, find five sets of data that could have been used). It turned out that under the other five choice points, many of the results were not statistically significant. Each of the different consistent combinations of choice points could count as a distinct

[10] "Last week CNN pulled a story about a study purporting to demonstrate a link between a woman's ovulation and how she votes… The story was savaged online as 'silly,' 'stupid,' 'sexist,' and 'offensive.'" (Bartlett, 2012b)

hypothesis, and you can then consider how many of them are statistically insignificant.

> If no strong arguments can be made for certain choices, we are left with many branches of the multiverse that have large *P* values. In these cases, the only reasonable conclusion on the effect of fertility is that there is considerable scientific uncertainty. One should reserve judgment . . . (ibid., p. 708)

Reserve judgment? If we're to apply our severe testing norms on such examples, and not dismiss them as entertainment only, then we'd go further. Here's another reasonable conclusion: The core presumptions are falsified (or would be with little effort). Say each person with high fertility in the first study is tested for candidate preference next month when they are in the low fertility stage. If they have the same voting preferences, the test is falsified. The spirit of their multiverse analysis is a quintessentially error statistical gambit. Anything that increases the flabbiness in uncovering flaws lowers the severity of the test that has passed (we'll visit *P*-value adjustments later on). But the onus isn't on us to give them a pass. As we turn to impressive statistical meta-critiques, what can be overlooked is whether the entire inquiry makes any sense. Readers will have many other tomatoes to toss at the ovulation research. Unless the overall program is falsified, the literature will only grow. We don't have to destroy statistical significance tests when what we really want is to show that a lot of studies constitute pseudoscience.

Souvenir G: The Current State of Play in Psychology

Failed replications, we hear, are creating a "cold war between those who built up modern psychology and those" tearing it down with failed replications (Letzter 2016). The severe tester is free to throw some fuel on both fires.

The widespread growth of preregistered studies is all to the good; it's too early to see if better science will result. Still, credit is due to those sticking their necks out to upend the status quo. I say it makes no sense to favor preregistration and also deny the relevance to evidence of optional stopping and outcomes other than the one observed. That your appraisal of the evidence is altered when you actually see the history supplied by the registered report is equivalent to worrying about biasing selection effects when they're not written down; your statistical method should pick up on them.

> By reviewing the hypotheses and analysis plans in advance, RRs (registered reports) should also help neutralize P-hacking and HARKing (hypothesizing after the results are known) by authors, and CARKing (critiquing after the results are known) by reviewers

with their own investments in the research outcomes, although empirical evidence will be required to confirm that this is the case. (Munafò et al. 2017, p. 5)

The papers are provisionally accepted before the results are in. To the severe tester, that requires the author to explain how she will pinpoint blame for negative results. I see nothing in preregistration, in and of itself, to require that. It would be wrong-headed to condemn CARKing: post-data criticism of assumptions and inquiries into hidden biases might be altogether warranted. For instance, one might ask about the attitude toward the finding conveyed by the professor: what did the students know and when did they know it? Of course, they must not be ad hoc saves of the finding.

The field of meta-research is bursting at the seams: distinct research into changing incentives is underway. The severe tester may be jaundiced to raise qualms, but she doesn't automatically assume that research into incentivizing researchers to behave in a fashion correlated with good science – data sharing, preregistration – is itself likely to improve the original field. Not without thinking through what would be needed to link statistics up with the substantive research problem. In some fields, one wonders if they would be better off ignoring statistical experiments and writing about plausible conjectures about human motivations, prejudices, or attitudes, perhaps backed by interesting field studies. It's when researchers try to test them using sciency methods that the project becomes pseudosciency.

2.7 How to Solve the Problem of Induction Now

Viewing inductive inference as severe testing, the problem of induction is transformed into the problem of showing the existence of severe tests and methods for identifying insevere ones. The trick isn't to have a formal, context-free method that you can show is reliable – as with the traditional problem of induction; the trick is to have methods that alert us when an application is shaky. As a relaxing end to a long tour, our evening speaker on ship, a severe tester, will hold forth on statistics and induction.

Guest Speaker: A Severe Tester on Solving Induction Now

Here's his talk:

For a severe tester like me, the current and future problem of induction is to identify fields and inquiries where inference problems are solved efficiently, and ascertain how obstacles are overcome – or not. You've already assembled the ingredients for this final leg of Tour II, including: lift-off, convergent arguments (from coincidence), pinpointing blame (Duhem's problem), and falsification. Essentially, the updated problem is to show that there exist methods for controlling and assessing error probabilities. Does that seem too easy? The problem has always been rather minimalist: to show at least some reliable methods exist; the idea being that they could then be built upon. Just find me one. They thought enumerative induction was the one, but it's not. I will examine four questions: 1. What warrants inferring a hypothesis that stands up to severe tests? 2. What enables induction (as severe testing) to work? 3. What is Neyman's quarrel with Carnap? and 4. What is Neyman's empirical justification for using statistical models?

1. What Warrants Inferring a Hypothesis that Passes Severe Tests? Suppose it is agreed that induction is severe testing. What warrants moving from *H* passing a severe test to warranting *H*? Even with a strong argument from coincidence akin to my weight gain showing up on myriad calibrated scales, there is no logical inconsistency with invoking a hypothesis from *conspiracy*: all these instruments conspire to produce results as if *H* were true but in fact *H* is false. The ultra-skeptic may invent a *rigged* hypothesis *R*:

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

without actually saying what this something else is. That is, we're imagining the extreme position of someone who simply asserts, *H* is actually false, although everything is as if it's true. Weak severity alone can block inferring a generic rigged hypothesis *R* as a way to discount a severely tested *H*. It can't prevent you from stopping there and never allowing a hypothesis is warranted. (Weak severity merely blocks inferring claims when little if anything has been done to probe them.) Nevertheless, if someone is bound to discount a strong argument for *H* by rigging, then she will be adopting a highly unreliable method. Why? Because a conspiracy hypothesis can always be adduced! Even with claims that are true, or where problems are solved correctly, you would have no chance of finding this out. I began with the stipulation that we wish to learn. Inquiry that blocks learning is pathological. Thus, because I am a severe tester, I hold both strong and weak severity:

> Data from test T are an indication of, or evidence for, *H* just to the extent that *H* has severely passed test T.

"Just to the extent that" indicates the "if then" goes in both directions: a claim that passes with low severity is unwarranted; one that passes with high severity is warranted. The phrase "to the extent that" refers to degrees of severity. That said, evidence requires a decent threshold be met, low severity is lousy evidence. It's still useful to point out in our travels when only weak severity suffices.

2. What Enables Induction (as Severe Testing) to Work: Informal and Quasi-formal Severity. You visited briefly the Exhibit (v) on prions and the deadly brain disease kuru. I'm going to use it as an example of a quasi-formal inquiry. Prions were found to contain a single protein dubbed PrP. Much to their surprise, researchers found PrP in normal cells too – it doesn't always cause disease. Our hero, Prusiner, again worries he'd "made a terrible mistake" (and prions had nothing to do with it). There are four strategies:

(a) Can we trick the phenomenon into telling us what it would be like if it really was a mere artifact (H_0)? Transgenic mice with PrP deliberately knocked out. Were H_0 true, they'd be expected to be infected as much as normal mice – the test hypothesis H_0 would not be rejected. H_0 asserts an *implicationary assumption* – one assumed just for testing. Abbreviate it as an *i-assumption*. It turns out that without PrP, none could be infected. Once PrP is replaced, they can again be infected. They argue, there's evidence to reject the artifact error H_0 because a procedure that would have revealed it fails to do so, and instead consistently finds departures from H_0.

(b) Over a period of more than 30 years, Prusiner and other researchers probed a series of local hypotheses. The levels of our hierarchy of models distinguishes various questions – even though I sketch it horizontally to save space (Figure 2.1). Comparativists deny we can proceed with a single hypothesis, but we do. Each question may be regarded as asking: would such and such be an erroneous interpretation of the data? Say the primary question is protein only or not. The alternatives do not include for the moment other "higher level" explanations about the mechanism of prion infectivity or the like. Given this localization, if *H* has been severely tested – by which I mean it has passed a severe test – then its denial has passed with low severity. That follows by definition of severity.

(c) Another surprise: the disease-causing form, call it pD, has the same exact amino acids as the normal type, call it pN. What's going on? Notice that a method that precluded exceptions to the central dogma (only nucleic acid directs replication of pathogens) would be incapable of identifying the culprit of prion transmission: the misfolding protein. Prusiner's prion hypothesis H^* is that prions target normal PrP, pinning and flattening their spirals to flip from their usual pN shape into pD, akin to a "deadly Virginia reel in the brain," adding newly formed pD's to the ends each time (Prusiner Labs 2004). When the helix is long enough, it ruptures, sending more pD seeds to convert normal prions. Another i-assumption to subject to the test of experiment. Trouble is, the process is so slow it can take years to develop. Not long ago, they found a way to deceive the natural state of affairs, while not altering what they want to learn: artificially rupture (with ultrasound or other means) the pathogenic prion. It's called protein misfolding cyclical amplification, PMCA. They get huge amounts of pD starting with a minute quantity, even a single molecule, so long as there's lots of normal PrP ready to be infected. All normal prions are converted into diseased prions in vitro. They could infer, with severity, that *H** gives a correct understanding of prion propagation, as well as corroborate the new research tool: They corroborated both at once, not instantly of course but over a period of a few years.

(d) Knowing the exponential rates of amplification associated with a method, researchers can infer, statistically, back to the amount of initial infectivity present – something they couldn't discern before, given the very low concentration of pD in accessible bodily fluids. Constantly improved and even automated, pD can now be detected in living animals for the first time.

What are some key elements? Honest self-criticism of how one may be wrong, deliberate deception to get counterfactual knowledge, conjecturing i-assumptions whose rejection leads to finding something out, and so on. Even researchers who hold different theories about the mechanism of transmission do not dispute PMCA – they can't if they want to learn more in the domain. I'm leaving out the political and personality feuds, but there's a good story there (see Prusiner 2014). I also didn't discuss statistically modeled aspects of prion research, but controlling the mean number of days for incubation allowed a stringent causal argument. I want to turn to statistical induction at a more rudimentary entry point.

3. Neyman's Quarrel with Carnap. Statistics is the *sine qua non* for extending our powers to severely probe. Jerzy Neyman, with his penchant for inductive behavior and performance rather than inductive inference, is often seen as a villain in the statistics battles. So take a look at a paper of his with the tantalizing title: "The Problem of Inductive Inference" (Neyman 1955). Neyman takes umbrage with the way confirmation philosophers, in particular Carnap, view frequentist inference:

> . . . when Professor Carnap criticizes some attitudes which he represents as consistent with my ("frequentist") point of view, I readily join him in his criticism without, however, accepting the responsibility for the criticized paragraphs. (p. 13)

In effect, Neyman says I'd never infer from observing that 150 out of 1000 throws with this die landed on six, "nothing else being known," that future throws will result in around 0.15 sixes, as Carnap alleges I would. This is a version of enumerative induction (or Carnap's straight rule). You need a statistical model! Carnap should view "Statistics as the Frequentist Theory of Induction," says Neyman in a section with this title, here the Binomial model. The Binomial distribution builds on n Bernoulli trials, the success–failure trials (visited in Section 1.4). It just adds up all the ways that number of successes could occur:

$$\Pr(k \text{ out of } n \text{ successes}) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

Carnapians could have formed the straight rule for the Binomial experiment, and argued:

> If an experiment can be generated and modeled Binomially, then sample means can be used to reliably estimate population means.
> An experiment can be modeled Binomially.
> Therefore, we can reliably estimate population means in those contexts.

The reliability comes from controlling the method's error probabilities.

4. What Is Neyman's Empirical Justification for Using Statistical Models?

Neyman pays a lot of attention to the empirical justification for using statistical models. Take his introductory text (Neyman 1952). The models are not invented from thin air. In the beginning there are records of different results and stable relative frequencies with which they occurred. These may be called empirical frequency distributions. There are real experiments that "even if carried out repeatedly with the utmost care to keep conditions constant, yield varying results" (ibid., p. 25). These are real, not hypothetical, experiments, he stresses. Examples he gives are roulette wheels (electrically regulated), tossing coins with a special machine (that gives a constant initial velocity to the coin), the number of disintegrations per minute in a quantity of radioactive matter, and the tendency for properties of organisms to vary despite homogeneous breeding. Even though we are unable to predict the outcome of such experiments, a certain stable pattern of regularity emerges rather quickly, even in moderately long series of trials; usually around 30 or 40 trials suffices. The pattern of regularity is in the relative frequency with which specified results occur.

Neyman takes a toy example: toss a die twice and record the frequency of sixes: 0, 1, or 2. Call this a *paired* trial. Now do this 1000 times. You'll have 1000 paired trials. Put these to one side for a moment. Just consider the entire set of

2000 tosses – *first order* trials Neyman calls these. Compute the relative frequency of sixes out of 2000. It may not be 1/6, due to the structure of the die or the throwing. Whatever it is, call it *f*. Now go back to the paired trials. Record the relative frequency of six found in paired trial 1, maybe it's 0, the relative frequency of six found in paired trial 2, all the way through your 1000 paired trials. We can then ask: what proportion of the 1000 paired trials had no sixes, what proportion had 1 six, what proportion 2 sixes? We find "the proportions of pairs with 0, 1 and 2 sixes will be, approximately,

$$(1-f)^2,\ 2f(1-f),\ \text{and}\ f^2.\text{''}$$

Instead of pairs of trials, consider *n*-fold trials: each trial has *n* throws of the die. Compute *f* as before: it is the relative frequency of six in the 1000*n* first order trials. Then, turn to the 1000 *n*-fold trials, and compute the proportion where six occurs *k* times (for $k < n$). It will be very nearly equal to

$$\binom{n}{k} f^k (1-f)^{n-k}.$$

"In other words, the relative frequency" of *k* out of *n* successes in the *n*-fold trials "is connected with the relative frequency of the first order experiments in very nearly the same way as the probability" of *k* out of *n* successes in a Binomial trial is related to the probability of success at each trial, θ (Neyman 1952, p. 26).

> The above fact, which has been found empirically many times . . . may be called the empirical law of large numbers. I want to emphasize that this law applies not only to the simple case connected with the binomial formula . . . but also to other cases. In fact, this law seems to be perfectly general . . . Whenever the law fails, we explain the failure by suspecting a "lack of randomness" in the first order trials. (ibid., p. 27)

Now consider, not just 1000 repetitions of all *n*-fold trials, but all. Here, *f*, the relative frequency of heads is θ in the Binomial probability model with *n* trials. It is this universe of hypothetical repetitions that our one *n*-fold sample is a random member of. Figure 2.2 shows the frequency distribution if we chose $n = 100$ and $\theta = 1/6$.

The Law of Large Numbers (LLN) shows we can use the probability derived from the probability model of the experiment to approximate the relative frequencies of outcomes in a series of *n*-fold trials. The LLN is both an empirical law and a mathematical law. The proofs are based on idealized random samples, but there are certain actual experiments that are well

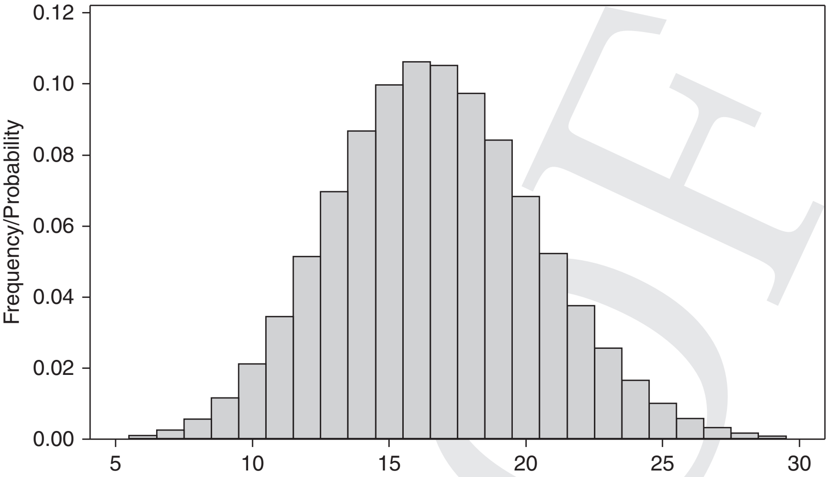

Figure 2.2 Binomial distribution for $n = 100$, $\theta = 1/6$.

approximated by the mathematical law – something we can empirically test (von Mises 1957).

You may bristle at this talk of random experiments, but, as Neyman repeatedly reminds us, these are merely "picturesque" shorthands for results squarely linked up with empirical tests (Neyman 1952, p. 23). We keep to them in order to explicate the issues at the focus of our journey. The justification for applying what is strictly an abstraction is no different from other cases of applied mathematics. We are not barred from fruitfully applying geometry because a geometric point is an abstraction.

"Whenever we succeed in arranging" the data generation such that the relative frequencies adequately approximate the mathematical probabilities in the sense of the LLN, we can say that the probabilistic model "adequately represents the method of carrying out the experiment" (ibid., p. 19). In those cases we are warranted in describing the results of real experiments as random samples from the population given by the probability model. You can reverse direction and ask about f or θ when unknown. Notice that we are modeling something we do, we may do it well or badly. All we need is that mysterious supernatural powers keep their hands off our attempts to carry out inquiry properly, to take one of Peirce's brilliant insights: "the supernal powers withhold their hands and let me alone, and that no mysterious uniformity . . . interferes with the action of chance" (2.749) in order to justify induction. *End of talk*.

I wonder if Carnap ever responded to Neyman's grumblings. Why didn't philosophers replace a vague phrase like "if these *k* out of *n* successes are all I know about the die" and refer to the Binomial model?, I asked Wesley Salmon in the 1980s. Because, he said, we didn't think the Binomial model could be justified without getting into a circle. But it can be tested empirically. By varying a known Binomial process to violate one of the assumptions deliberately, we develop tests that would very probably detect such violations should they occur. This is the key to justifying induction as severe testing: it corrects its assumptions. Testing the assumption of randomness is independent of estimating θ given that it's random. Salmon and I met weekly to discuss statistical tests of assumptions when I visited the Center for Philosophy of Science at Pittsburgh in 1989. I think I convinced him of this much (or so he said): the confirmation theorists were too hasty in discounting the possibility of warranting statistical model assumptions.

Souvenir H: Solving Induction Is Showing Methods with Error Control

How is the problem of induction transformed if induction is viewed as severe testing? Essentially, it becomes a matter of showing that there exist methods with good error probabilities. The specific task becomes examining the fields or inquiries that are – and are not – capable of assessing and controlling severity. Nowadays many people abjure teaching the different distributions, preferring instead to generate frequency distributions by resampling a given random sample (Section 4.6). It vividly demonstrates what really matters in appealing to probability models for inference, as distinct from modeling phenomena more generally: Frequentist error probabilities are of relevance when frequencies represent the capabilities of inquiries to discern and discriminate various flaws and biases. Where Popper couldn't say that methods probably would have found *H* false, if it is false, error statistical methods let us go further.

The severity account puts forward a statistical philosophy associated with statistical methods. To see what I mean, recall the Likelihoodist. It's reasonable to suppose that we favor, among pairs of hypotheses, the one that predicts or makes probable the data – proposes the Likelihoodist. The formal Law of Likelihood (LL) is to capture this, and we appraise it according to how well it succeeds, and how well it satisfies the goals of statistical practice. Likewise, the severe tester proposes, there is a pre-statistical plausibility to infer hypotheses to the extent that they have passed stringent tests. The error statistical methodology is the frequentist theory of induction. Here too the statistical philosophy is to be appraised according to how well it captures and supplies rationales for inductive-statistical inference. The rest of our journey will bear this out. Enjoy the concert in the Captain's Central Limit Lounge while the breezes are still gentle, we set out on Excursion 3 in the morn.

Excursion 3 Statistical Tests and Scientific Inference

Itinerary

Tour I Ingenious and Severe Tests

> [T]he impressive thing about [the 1919 tests of Einstein's theory of gravity] is the *risk* involved in a prediction of this kind. If observation shows that the predicted effect is definitely absent, then the theory is simply refuted. The theory is *incompatible with certain possible results of observation* – in fact with results which everybody before Einstein would have expected. This is quite different from the situation I have previously described, [where] . . . it was practically impossible to describe any human behavior that might not be claimed to be a verification of these [psychological] theories. (Popper 1962, p. 36)

The 1919 eclipse experiments opened Popper's eyes to what made Einstein's theory so different from other revolutionary theories of the day: Einstein was prepared to subject his theory to risky tests.[1] Einstein was eager to galvanize scientists to test his theory of gravity, knowing the solar eclipse was coming up on May 29, 1919. Leading the expedition to test GTR was a perfect opportunity for Sir Arthur Eddington, a devout follower of Einstein as well as a devout Quaker and conscientious objector. Fearing "a scandal if one of its young stars went to jail as a conscientious objector," officials at Cambridge argued that Eddington couldn't very well be allowed to go off to war when the country needed him to prepare the journey to test Einstein's predicted light deflection (Kaku 2005, p. 113).

The museum ramps up from Popper through a gallery on "Data Analysis in the 1919 Eclipse" (Section 3.1) which then leads to the main gallery on origins of statistical tests (Section 3.2). Here's our Museum Guide:

> According to Einstein's theory of gravitation, to an observer on earth, light passing near the sun is deflected by an angle, λ, reaching its maximum of 1.75″ for light just grazing the sun, but the light deflection would be undetectable on earth with the instruments available in 1919. Although the light deflection of stars near the sun (approximately 1 second of arc) *would* be detectable, the sun's glare renders such stars invisible, save during a total eclipse, which "by strange good

[1] You will recognize the above as echoing Popperian "theoretical novelty" – Popper developed it to fit the Einstein test.

> fortune" would occur on May 29, 1919 (Eddington [1920] 1987, p. 113).
>
> There were three hypotheses for which "it was especially desired to discriminate between" (Dyson et al. 1920 p. 291). Each is a statement about a parameter, the deflection of light at the limb of the sun (in arc seconds): $\lambda = 0''$ (no deflection), $\lambda = 0.87''$ (Newton), $\lambda = 1.75''$ (Einstein). The Newtonian predicted deflection stems from assuming light has mass and follows Newton's Law of Gravity.
>
> The difference in statistical prediction masks the deep theoretical differences in how each explains gravitational phenomena. Newtonian gravitation describes a force of attraction between two bodies; while for Einstein gravitational effects are actually the result of the curvature of spacetime. A gravitating body like the sun distorts its surrounding spacetime, and other bodies are reacting to those distortions.

Where Are Some of the Members of Our Statistical Cast of Characters in 1919? In 1919, Fisher had just accepted a job as a statistician at Rothamsted Experimental Station. He preferred this temporary slot to a more secure offer by Karl Pearson (KP), which had so many strings attached – requiring KP to approve everything Fisher taught or published – that Joan Fisher Box writes: After years during which Fisher "had been rather consistently snubbed" by KP, "It seemed that the lover was at last to be admitted to his lady's court – on conditions that he first submit to castration" (J. Box 1978, p. 61). Fisher had already challenged the old guard. Whereas KP, after working on the problem for over 20 years, had only approximated "the first two moments of the sample correlation coefficient; Fisher derived the relevant distribution, not just the first two moments" in 1915 (Spanos 2013a). Unable to fight in WWI due to poor eyesight, Fisher felt that becoming a subsistence farmer during the war, making food coupons unnecessary, was the best way for him to exercise his patriotic duty.

In 1919, Neyman is living a hardscrabble life in a land alternately part of Russia or Poland, while the civil war between Reds and Whites is raging. "It was in the course of selling matches for food" (C. Reid 1998, p. 31) that Neyman was first imprisoned (for a few days) in 1919. Describing life amongst "roaming bands of anarchists, epidemics" (ibid., p. 32), Neyman tells us, "existence" was the primary concern (ibid., p. 31). With little academic work in statistics, and "since no one in Poland was able to gauge the importance of his statistical work (he was 'sui generis,' as he later described himself)" (Lehmann 1994, p. 398), Polish authorities sent him to University College in

London in 1925/1926 to get the great Karl Pearson's assessment. Neyman and E. Pearson begin work together in 1926.

Egon Pearson, son of Karl, gets his B.A. in 1919, and begins studies at Cambridge the next year, including a course by Eddington on the theory of errors. Egon is shy and intimidated, reticent and diffident, living in the shadow of his eminent father, whom he gradually starts to question after Fisher's criticisms. He describes the psychological crisis he's going through at the time Neyman arrives in London: "I was torn between conflicting emotions: a. finding it difficult to understand R.A.F., b. hating [Fisher] for his attacks on my paternal 'god,' c. realizing that in some things at least he was right" (C. Reid 1998, p. 56). As far as appearances amongst the statistical cast: there are the two Pearsons: tall, Edwardian, genteel; there's hardscrabble Neyman with his strong Polish accent and small, toothbrush mustache; and Fisher: short, bearded, very thick glasses, pipe, and eight children.

Let's go back to 1919, which saw Albert Einstein go from being a little known German scientist to becoming an international celebrity.

3.1 Statistical Inference and Sexy Science: The 1919 Eclipse Test

The famous 1919 eclipse expeditions purported to test Einstein's new account of gravity against the long-reigning Newtonian theory. I get the impression that statisticians consider there to be a world of difference between statistical inference and appraising large-scale theories in "glamorous" or "sexy science." The way it actually unfolds, which may not be what you find in philosophical accounts of theory change, revolves around local data analysis and statistical inference. Even large-scale, sexy theories are made to connect with actual data only by intermediate hypotheses and models. To falsify, or even provide anomalies, for a large-scale theory like Newton's, we saw, is to infer "falsifying hypotheses," which are statistical in nature.

Notably, from a general theory we do not deduce observable data, but at most a general phenomenon such as the Einstein deflection effect due to the sun's gravitational field (Bogen and Woodward 1988). The problem that requires the most ingenuity is finding or inventing a phenomenon, detector, or probe that will serve as a meeting ground between data that can actually be collected and a substantive or theoretical effect of interest. This meeting ground is typically statistical. Our array in Souvenir E provides homes within which relevant stages of inquiry can live. Theories and laws give constraints but the problem at the experimental frontier has much in common with research in fields where there is at most a vague phenomenon and no real theories to speak of.

There are two key stages of inquiry corresponding to two questions within the broad umbrella of *auditing an inquiry*:

(i) is there a deflection effect of the amount predicted by Einstein as against Newton (the "Einstein effect")?
(ii) is it attributable to the sun's gravitational field as described in Einstein's hypothesis?

A distinct third question, "higher" in our hierarchy, in the sense of being more theoretical and more general, is: is GTR an adequate account of gravity as a whole? These three are often run together in discussions, but it is important to keep them apart.

The first is most directly statistical. For one thing, there's the fact that they don't observe stars just grazing the sun but stars whose distance from the sun is at least two times the solar radius, where the predicted deflection is only around 1″ of arc. They infer statistically what the deflection would have been for starlight near the sun. Second, they don't observe a deflection, but (at best) photographs of the positions of certain stars at the time of the eclipse. To "observe" the deflection, if any, requires inferring what the positions of these same stars would have been were the sun's effect absent, a "control" as it were. Eddington remarks:

> The bugbear of possible systematic error affects all investigations of this kind. How do you know that there is not something in your apparatus responsible for this apparent deflection? . . . To meet this criticism, a different field of stars was photographed . . . at the same altitude as the eclipse field. If the deflection were really instrumental, stars on these plates should show relative displacements of a similar kind to those on the eclipse plates. But on measuring these check-plates no appreciable displacements were found. That seems to be satisfactory evidence that the displacement . . . is not due to differences in instrumental conditions. ([1920] 1987, p. 116)

If the check plates can serve as this kind of a control, the researchers are able to use a combination of theory, controls, and data to transform the original observations into an approximate linear relationship between two observable variables and use least squares to estimate the deflection. The position of each star photographed at the eclipse (the eclipse plate) is compared to its normal position photographed at night (months before or after the eclipse), when the effect of the sun is absent (the night plate). Placing the eclipse and night plates together allows the tiny distances to be measured in the *x* and *y* directions (Figure 3.1). The estimation had to take account of how the two plates are

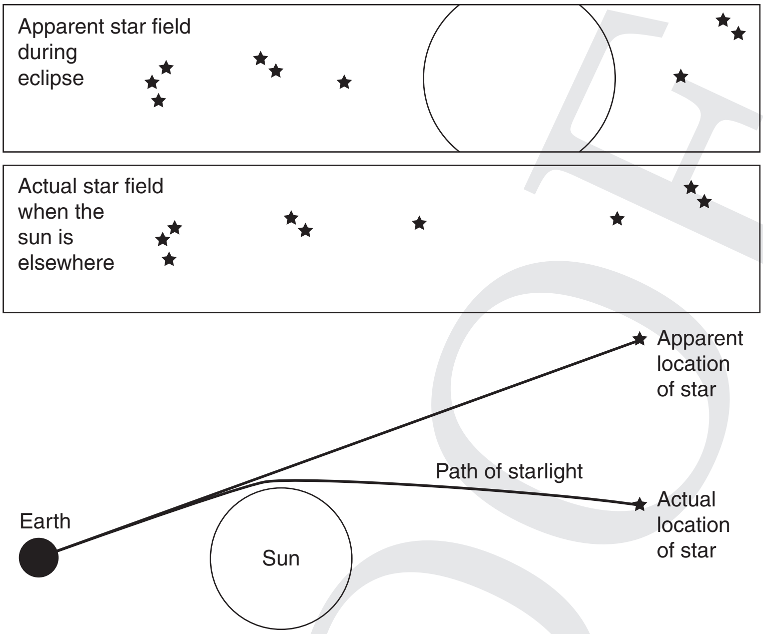
Apparent star field during eclipse
Actual star field when the sun is elsewhere
Apparent location of star
Path of starlight
Actual location of star
Earth
Sun

Figure 3.1 Light deflection.

accidentally clamped together, possible changes in the scale – due mainly to differences in the focus between the exposure of the eclipse and the night plates – on a set of other plate parameters, and, finally, on the light deflection.

The general technique was known to astronomers from determining the angle of stellar parallax, “for which much greater accuracy is required” (ibid., pp. 115–16). (The relation between a star position and the sun changes as the earth moves around the sun, and the angle formed is its parallax.) Somewhat like the situation with Big Data, scientists already had a great deal of data on star positions and now there’s a highly theoretical question that can be probed with a known method. Still, the eclipse poses unique problems of data analysis, not to mention the precariousness of bringing telescopes on expeditions to Sobral in North Brazil and Principe in the Gulf of Guinea (West Africa).

The problem in (i) is reduced to a statistical one: the observed mean deflections (from sets of photographs) are Normally distributed around the predicted mean deflection μ. The proper way to frame this as a statistical test is to choose one of the values as H_0 and define composite H_1 to include alternative values of interest. For instance, the Newtonian “half deflection” can specify the H_0: $\mu \leq 0.87$, and the H_1: $\mu > 0.87$ includes the Einsteinian value of

1.75. Hypothesis H_0 also includes the third value of potential interest, $\mu = 0$: no deflection.[2] After a good deal of data analysis, the two eclipse results from Sobral and Principe were, with their standard errors,

Sobral: the eclipse deflection = 1.98″ ± 0.18″.

Principe: the eclipse deflection = 1.61″ ± 0.45″.

The actual report was in probable errors in use at the time, 0.12 and 0.30 respectively, where 1 probable error equals 0.68 standard errors. A sample mean differs from a Normal population mean by one or more probable errors (in either direction) 50% of the time.

> It is usual to allow a margin of safety of about twice the probable error on either side of the mean. The evidence of the Principe plates is thus just about sufficient to rule out the possibility of the 'half deflection,' and the Sobral plates exclude it with practical certainty. (Eddington [1920]1987, p. 118)

The idea of reporting the "probable error" is of interest to us. There is no probability assignment to the interval, it's an error probability *of the method*. To infer μ = observed mean ± 1 probable error is to use a method that 50% of the time correctly covers μ. Two probable errors wouldn't be considered much of a margin of safety these days, being only ~1.4 standard errors. Using the term "probable error" might be thought to encourage misinterpretation – and it does – but it's not so different from the current use of "margin of error."

A text by Ghosh et al. (2010, p. 48) presents the Eddington results as a two-sided Normal test of H_0: $\mu = 1.75$ (the Einstein value) vs. H_1: $\mu \neq 1.75$, with a lump of prior probability given to the point null. If any theoretical prediction were to get a lump at this stage, it is Newton's. The vast majority of Newtonians, understandably, regarded Newton as far more plausible, never mind the small well-known anomalies, such as being slightly off in its prediction of the orbit of the planet Mercury. Few could even understand Einstein's radically different conception of space and time.

Interestingly, the (default) Bayesian statistician Harold Jeffreys was involved in the eclipse experiment in 1919. He lauded the eclipse results as finally putting the Einstein law on firm experimental footing – despite his low Bayesian prior in GTR (Jeffreys 1919). Actually, even the experimental footing did not emerge until the 1960s (Will 1986). The eclipse tests, not just those of 1919, but all eclipse tests of the deflection effect, failed to give very precise results. Nothing like a stringent estimate of the deflection effect

[2] "A ray of light nicking the edge of the sun, for example, would bend a minuscule 1.75 arcseconds – the angle made by a right triangle 1 inch high and 1.9 *miles* long" (Buchen 2009).

emerged until the field was rescued by radioastronomical data from quasars (quasi-stellar radio sources). This allowed testing the deflection using radio waves instead of light waves, and without waiting for an eclipse.

Some Popperian Confusions About Falsification and Severe Tests

Popper lauds GTR as sticking its neck out, bravely being ready to admit its falsity were the deflection effect not found (1962, pp. 36–7). Even if no deflection effect had been found in the 1919 experiments, it would have been blamed on the sheer difficulty in discerning so small an effect. This would have been entirely correct. Yet many Popperians, perhaps Popper himself, get this wrong. Listen to Popperian Meehl:

> [T]he stipulation beforehand that one will be pleased about substantive theory *T* when the numerical results come out as forecast, but will not necessarily abandon it when they do not, seems on the face of it to be about as blatant a violation of the Popperian commandment as you could commit. For the investigator, in a way, is doing . . . what astrologers and Marxists and psychoanalysts allegedly do, playing 'heads I win, tails you lose.' (Meehl 1978, p. 821)

There is a confusion here, and it's rather common. A successful result may rightly be taken as evidence for a real effect *H*, even though failing to find the effect would not, and should not, be taken to refute the effect, or as evidence against *H*. This makes perfect sense if one keeps in mind that a test might have had little chance to detect the effect, even if it exists.

One set of eclipse plates from Sobral (the astrographic plates) was sufficiently blurred by a change of focus in the telescope as to preclude any decent estimate of the standard error (more on this case later). Even if all the 1919 eclipse results were blurred, this would at most show no deflection had been found. This is not automatically evidence there's no deflection effect.[3] To suppose it is would violate our minimal principle of evidence: the probability of failing to detect the tiny effect with the crude 1919 instruments is high – even if the deflection effect exists.

Here's how the severity requirement cashes this out: Let H_0 assert the Einstein effect is absent or smaller than the predicted amount, and H_1 that the deflection exists. An observed failure to detect a deflection "accords with" H_0, so the first severity requirement holds. But there's a high probability of this occurring even if H_0 is false and H_1 true (whether as explained in GTR or other theory). The point really reflects the asymmetry of falsification and corroboration (Section 2.1): if the deflection effect passes an audit, then it is a genuine

[3] To grasp this, consider that a single black swan proves the hypothesis *H*: some swans are not white, even though a white swan would not be taken as strong evidence for *H*'s denial. *H*'s denial would be that all swans are white.

anomaly for Newton's half deflection – only one is needed. Yet not finding an anomaly in 1919 isn't grounds for supposing no deflection anomalies exist. Alternatively, you can see this as an unsound but valid deductive argument (*modus tollens*):

> If GTR, then the deflection effect is observed in the 1919 eclipse tests.
> No deflection is observed in the 1919 eclipse tests.
> Therefore ~GTR (or evidence against GTR).

Because the first premise of this valid argument is false, the argument is unsound. By contrast, once instruments were available to powerfully detect any deflection effects, a no-show would have to be taken against its existence, and thus against GTR. In fact, however, a deflection was observed in 1919, although the accuracy was only 30%. Either way, Popperian requirements are upheld, even if some Popperians get this wrong.

George Barnard on the Eclipse Tests

The first time I met George Barnard in 1985, the topics of the 1919 eclipse episode and the N-P vs. Fisher battles were front and center. The focus of his work on the eclipse was twofold: First, "to draw attention to a reasonably accessible instance . . . where the inferential processes can be seen at work – and in the mind of someone who, (unlike so many physicists!) had taken the trouble to familiarise himself thoroughly with mathematical statistics" (Barnard 1971, p. 294). He is alluding to Eddington. Of course that was many years ago. Barnard's second reason is to issue a rebuke to Neyman! – or at least to a crude performance construal often associated with Neyman (ibid., p. 300). Barnard's point is that bad luck with the weather resulted in the sample size of usable photographs being very different from what could have been planned. They only used results where enough stars could be measured to apply least squares regression reliably (at least equal to the number of unknown parameters – six). Any suggestion that the standard errors "be reduced because in a repetition of the experiment" more usable images might be expected, "would be greeted with derision" (ibid., p. 295). Did Neyman say otherwise? In practice, Neyman describes cases where he rejects the data as unusable because of failed assumptions (e.g., Neyman 1977, discussing a failed randomization in a cloud seeding experiment).

Clearly, Barnard took Fisher's side in the N-P vs. Fisher disputes; he wanted me to know he was the one responsible for telling Fisher that Neyman had converted "his" significance tests into tools for acceptance sampling, where only long-run performance matters (Pearson 1955 affirms this). Pearson was kept out of it. The set of hypothetical repetitions used in obtaining the relevant error probability, in Barnard's view, should consist of "results of reasonably similar precision" (1971, p. 300). This is a very interesting idea, and it will come up again.

Big Picture Inference: Can Other Hypotheses Explain the Observed Deflection?

Even to the extent that they had found a deflection effect, it would have been fallacious to infer the effect "attributable to the sun's gravitational field." The question (ii) must be tackled: A statistical effect is not a substantive effect. Addressing the causal attribution demands the use of the eclipse data as well as considerable background information. Here we're in the land of "big picture" inference: the inference is "given everything we know". In this sense, the observed effect is used and is "non-novel" (in the use-novel sense). Once the deflection effect was known, imprecise as it was, it had to be used. Deliberately seeking a way to explain the eclipse effect while saving Newton's Law of Gravity from falsification isn't the slightest bit pejorative – so long as each conjecture is subject to severe test. Were *any* other cause to exist that produced a considerable fraction of the deflection effect, that alone would falsify the Einstein hypothesis (which asserts that *all* of the 1.75″ are due to gravity) (Jeffreys 1919, p. 138). That was part of the riskiness of the GTR prediction.

It's Not How Plausible, but How Well Probed

One famous case was that of Sir Oliver Lodge and his proposed "ether effect." Lodge was personally invested in the Newtonian ether, as he believed it was through the ether that he was able to contact departed souls, in particular his son, Raymond. Lodge had "preregistered" in advance that if the eclipse results showed the Einstein deflection he would find a way to give a Newtonian explanation (Lodge 1919). Others, without a paranormal bent, felt a similar allegiance to Newton. "We owe it to that great man to proceed very carefully in modifying or retouching his Law of Gravitation" (Silberstein 1919, p. 397). But respect for Newton was kept out of the data analysis. They were free to try and try again with Newton-saving factors because, unlike in pejorative seeking, it would be extremely difficult for any such factor to pass if false – given the standards available and insisted on by the relevant community of scientists. Each Newton-saving hypothesis collapsed on the basis of a one-two punch: the magnitude of effect that could have been due to the conjectured factor is far too small to account for the eclipse effect; and were it large enough to account for the eclipse effect, it would have blatantly false or contradictory implications elsewhere. Could the refraction of the sun's corona be responsible (as one scientist proposed)? Were it sufficient to explain the deflection, then comets would explode when they pass near the sun, which they do not! Or take another of Lodge's ether modification hypotheses. As scientist Lindemann put it:

> Sir Oliver Lodge has suggested that the deflection of light might be explained by assuming a change in the effective dielectric constant near a gravitating body. . . . It sounds quite promising at first . . . The difficulty is that one has in each case to adopt a different constant in the law, giving the dielectric constant as a function of the gravitational field, unless some other effect intervenes. (1919, p. 114)

This would be a highly insevere way to retain Newton. These criticisms combine quantitative and qualitative severity arguments. We don't need a precise quantitative measure of how frequently we'd be wrong with such ad hoc finagling. The Newton-saving factors might have been plausible but they were unable to pass severe tests. Saving Newton this way would be bad science.

As is required under our demarcation (Section 2.3): the 1919 players were able to embark upon an inquiry to pinpoint the source for the Newton anomaly. By 1921, it was recognized that the deflection effect was real, though inaccurately measured. Further, the effects revealed (corona effect, shadow effect, lens effect) were themselves used to advance the program of experimental testing of GTR. For instance, learning about the effect of the sun's corona (corona effect) not only vouchsafed the eclipse result, but pointed to an effect that could not be ignored in dealing with radioastronomy. Time and space prevents going further, but I highly recommend you return at a later time. For discussion and references, see Mayo (1996, 2010a, e).

The result of all the analysis was merely evidence of a small piece of GTR: an Einstein-like deflection effect. The GTR "passed" the test, but clearly they couldn't infer GTR severely. Even now, only its severely tested parts are accepted, at least to probe relativistic gravity. John Earman, in criticism of me, observes:

> [W]hen high-level theoretical hypotheses are at issue, we are rarely in a position to justify a judgment to the effect that $\Pr(E|{\sim}H \;\&\; K) \ll 0.5$. If we take H to be Einstein's general theory of relativity and E to be the outcome of the eclipse test, then in 1918 and 1919 physicists were in no position to be confident that the vast and then unexplored space of possible gravitational theories denoted by ~GTR does not contain alternatives to GTR that yield that same prediction for the bending of light as GTR. (Earman 1992, p. 117)

A similar charge is echoed by Laudan (1997), Chalmers (2010), and Musgrave (2010). For the severe tester, being prohibited from regarding GTR as having passed severely – especially in 1918 and 1919 – is just what an account ought to do. (Do you see how this relates to our treatment of irrelevant conjunctions in Section 2.2?)

From the first exciting results to around 1960, GTR lay in the doldrums. This is called the period of *hibernation* or stagnation. Saying it remained uncorroborated or inseverely tested does not mean GTR was deemed scarcely true, improbable, or implausible. It hadn't failed tests, but there were too few link-ups between the highly mathematical GTR and experimental data. Uncorroborated is very different from disconfirmed. We need a standpoint that lets us express being at that stage in a problem, and viewing inference as severe testing gives us one. Soon after, things would change, leading to the Renaissance from 1960 to 1980. We'll pick this up at the end of Sections 3.2 and 3.3. To segue into statistical tests, here's a souvenir.

Souvenir I: So What Is a Statistical Test, Really?

So what's in a statistical test? First there is a question or problem, a piece of which is to be considered statistically, either because of a planned experimental design, or by embedding it in a formal statistical model. There are (A) hypotheses, and a set of possible outcomes or data; (B) a measure of accordance or discordance, fit, or misfit, d($\boldsymbol{X}$) between possible answers (hypotheses) and data; and (C) an appraisal of a relevant distribution associated with d($\boldsymbol{X}$). Since we want to tell what's true about tests now in existence, we need an apparatus to capture them, while also offering latitude to diverge from their straight and narrow paths.

(A) *Hypotheses*. A statistical hypothesis H_i is generally couched in terms of an unknown parameter θ. It is a claim about some aspect of the process that might have generated the data, $\boldsymbol{x}_0 = (x_1, \ldots, x_n)$, given in a model of that process. Statistical hypotheses assign probabilities to various outcomes $\boldsymbol{x}$ "computed under the supposition that H_i is correct (about the generating mechanism)." That is how to read $f(\boldsymbol{x}; H_i)$, or as I often write it: Pr($\boldsymbol{x}$; H_i). This is just an analytic claim about the assignment of probabilities to $\boldsymbol{x}$ stipulated in H_i.

In the GTR example, we consider n IID Normal random variables: $(X_1, \ldots, X_n)$ that are N(μ, σ^2). Nowadays, the GTR value for $\lambda = \mu$ is set at 1, and the test might be of H_0: $\mu \le 1$ vs. H: $\mu > 1$. The hypothesis of interest will typically be a claim C posed after the data, identified within the predesignated parameter spaces.

(B) *Distance function and its distribution*. A function of the sample d($\boldsymbol{X}$), the *test statistic*, reflects how well or poorly the data ($\boldsymbol{X} = \boldsymbol{x}_0$) accord with the hypothesis H_0, which serves as a reference point. The term "test statistic" is generally reserved for statistics whose distribution can be computed under the main or test hypothesis. If we just want to speak of a statistic measuring distance, we'll call it that.

It is the observed distance d($\boldsymbol{x}_0$) that is described as "significantly different" from the null hypothesis H_0. I use $\boldsymbol{x}$ to say something general about the data, whereas $\boldsymbol{x}_0$ refers to a fixed data set.

(C) *Test rule T*. Some interpretative move or methodological rule is required for an account of inference. One such rule might be to infer that $\boldsymbol{x}$ is evidence of a discrepancy δ from H_0 just when d($\boldsymbol{x}$) ≥ c, for some value of c. Thanks to the requirement in (B), we can calculate the probability that {d($\boldsymbol{X}$) ≥ c} under the assumption that H_0 is true. We want also to compute it under various discrepancies from H_0, whether or not there's an explicit specification of H_1. Therefore, we can calculate the probability of inferring evidence for discrepancies from H_0 when in fact the interpretation would be erroneous. Such an *error probability* is given by the probability distribution of d($\boldsymbol{X}$) – its *sampling distribution* – computed under one or another hypothesis.

To develop an account adequate for solving foundational problems, special stipulations and even reinterpretations of standard notions may be required. (D) and (E) reflect some of these.

(D) *A key role of the distribution* of d($\boldsymbol{X}$) will be to characterize the probative abilities of the inferential rule for the task of unearthing flaws and misinterpretations of data. In this way, error probabilities can be used to assess the severity associated with various inferences. We are able to consider outputs outside the N-P and Fisherian schools, including "report a Bayes ratio" or "infer a posterior probability" by leaving our measure of agreement or disagreement open. We can then try to compute an associated error probability and severity measure for these other accounts.

(E) *Empirical background assumptions*. Quite a lot of background knowledge goes into implementing these computations and interpretations. They are guided by the goal of assessing severity for the primary inference or problem, housed in the manifold steps from planning the inquiry, to data generation and analyses.

We've arrived at the N-P gallery, where Egon Pearson (actually a hologram) is describing his and Neyman's formulation of tests. Although obviously the museum does not show our new formulation, their apparatus is not so different.

3.2 N-P Tests: An Episode in Anglo-Polish Collaboration

> We proceed by setting up a specific hypothesis to test, H_0 in Neyman's and my terminology, the null hypothesis in R. A. Fisher's . . . in choosing the test, we take into account alternatives to H_0 which we believe possible or at any rate consider it most important to be on the look out for . . .Three steps in constructing the test may be defined:
>
> **Step 1.** We must first specify the set of results . . .
>
> **Step 2.** We then divide this set by a system of ordered boundaries . . .
>
> such that as we pass across one boundary and proceed to the next, we come to a class of results which makes us more and more inclined, on the information available, to reject the hypothesis tested in favour of alternatives which differ from it by increasing amounts.
>
> **Step 3.** We then, if possible, associate with each contour level the chance that, if H_0 is true, a result will occur in random sampling lying beyond that level . . .
>
> In our first papers [in 1928] we suggested that the likelihood ratio criterion, λ, was a very useful one . . . Thus Step 2 proceeded Step 3. In later papers [1933–1938] we started with a fixed value for the chance, ε, of Step 3 . . . However, although the mathematical procedure may put Step 3 before 2, we cannot put this into operation before we have decided, under Step 2, on the guiding principle to be used in choosing the contour system. That is why I have numbered the steps in this order. (Egon Pearson 1947, p. 173)

In addition to Pearson's 1947 paper, the museum follows his account in "The Neyman–Pearson Story: 1926–34" (Pearson 1970). The subtitle is "Historical Sidelights on an Episode in Anglo-Polish Collaboration"!

We meet Jerzy Neyman at the point he's sent to have his work sized up by Karl Pearson at University College in 1925/26. Neyman wasn't that impressed:

> Neyman found . . . [K.]Pearson himself surprisingly ignorant of modern mathematics. (The fact that Pearson did not understand the difference between independence and lack of correlation led to a misunderstanding that nearly terminated Neyman's stay . . .) (Lehmann 1988, p. 2)

Thus, instead of spending his second fellowship year in London, Neyman goes to Paris where his wife Olga ("Lola") is pursuing a career in art, and where he could attend lectures in mathematics by Lebesque and Borel. "[W]ere it not for Egon Pearson [whom I had briefly met while in London], I would have probably drifted to my earlier passion for [pure mathematics]" (Neyman quoted in Lehmann 1988, p. 3).

What pulled him back to statistics was Egon Pearson's letter in 1926. E. Pearson had been "suddenly smitten" with doubt about the justification of tests then in use, and he needed someone with a stronger mathematical background to pursue his concerns. Neyman had just returned from his fellowship years to a hectic and difficult life in Warsaw, working multiple jobs in applied statistics.

> [H]is financial situation was always precarious. The bright spot in this difficult period was his work with the younger Pearson. Trying to find a unifying, logical basis which would lead systematically to the various statistical tests that had been proposed by Student and Fisher was a 'big problem' of the kind for which he had hoped . . . (ibid., p. 3)

N-P Tests: Putting Fisherian Tests on a Logical Footing

For the Fisherian simple or "pure" significance test, alternatives to the null "lurk in the undergrowth but are not explicitly formulated probabilistically" (Mayo and Cox 2006, p. 81). Still there are constraints on a Fisherian test statistic. Criteria for the test statistic d($\boldsymbol{X}$) are

(i) it reduces the data as much as possible;
(ii) the larger d($\boldsymbol{x}_0$) the further the outcome from what's expected under H_0, with respect to the particular question;
(iii) the *P*-value can be computed $p(\boldsymbol{x}_0)=\Pr(\mathrm{d}(\boldsymbol{X}) \geq \mathrm{d}(\boldsymbol{x}_0); H_0)$.

Fisher, arch falsificationist, sought test statistics that would be *sensitive* to discrepancies from the null. Desiderata (i)–(iii) are related, as emerged clearly from N-P's work.

Fisher introduced the idea of a parametric statistical model, which may be written $\mathrm{M}_\theta(\boldsymbol{x})$. Karl Pearson and others had been prone to mixing up a parameter θ, say the mean of a population, with a sample mean $\bar{x}$. As a result, concepts that make sense for statistic $\overline{\boldsymbol{X}}$, like having a distribution, were willy-nilly placed on a fixed parameter θ. Neyman and Pearson [N-P] gave mathematical rigor to the components of Fisher's tests and estimation. The model can be represented as a pair (S, Θ) where S denotes the set of all possible values of the *sample* $\boldsymbol{X} = (X_1, \ldots, X_n)$ – one such value being the data $\boldsymbol{x}_0 = (x_1, \ldots, x_n)$ – and Θ denotes the set of all possible values of the unknown *parameter(s)* θ. In hypothesis testing, Θ is used as shorthand for the family of probability distributions or, in continuous cases, densities *indexed* by θ. Without the abbreviation, we'd write the full model as

$$\mathrm{M}_\theta(\boldsymbol{x}) \coloneqq \{\mathrm{f}(\boldsymbol{x};\, \theta),\ \theta \in \Theta\},$$

where $\mathrm{f}(\boldsymbol{x}; \theta)$, for all $\boldsymbol{x} \in \mathrm{S}$, is the distribution (or density) of the sample. We don't test all features of the model at once; it's part of the test specification to indicate which features (parameters) of the model are under test. The *generic form* of *null* and *alternative* hypotheses is

$$H_0\text{: } \theta \in \Theta_0 \text{ vs. } H_1\text{: } \theta \in \Theta_1,$$

where (Θ_0, Θ_1) constitute subsets of Θ that partition Θ. Together, Θ_0 and Θ_1 exhaust the parameter space. N-P called H_0 the *test hypothesis*, which is preferable to null hypothesis, since for them it's on par with alternative H_1; but for brevity and familiarity, I mostly call H_0 the null. I follow A. Spanos' treatment.

Lambda Criterion

What were Neyman and Pearson looking for in their joint work from 1928? They sought a criterion for choosing, as well as generating, sensible test statistics. Working purely on intuition, which they later imbued with a justification, N-P employ the likelihood ratio. Pearson found the spark of the idea from correspondence with Gosset, known as Student, but we will see that generating good tests requires much more than considering alternatives.

How can we consider the likelihood ratio of hypotheses when one or both can contain multiple values of the parameter? They consider the maximum values that the likelihood could take over ranges of the parameter space. In particular, they take the maximum likelihood over all possible values of θ in the entire parameter space Θ (not Θ_1), and compare it to the maximum over the restricted range of values in Θ_0, to form the ratio

$$\Lambda(\boldsymbol{X}) = \frac{\max_{\theta \in \Theta} \mathrm{L}(\boldsymbol{X};\theta)}{\max_{\theta \in \Theta_0} \mathrm{L}(\boldsymbol{X};\theta)}.$$

Let's look at this. The numerator is the value of θ that makes the data $\boldsymbol{x}$ most probable over the entire parameter space. It is the *maximum likelihood estimator* for θ. Write it as $\hat{\theta}$. The denominator is the value of θ that maximizes the probability of $\boldsymbol{x}$ restricted just to the members of the null Θ_0. It may be called the *restricted* likelihood. Write it as $\widetilde{\theta}$:

$$\Lambda(\boldsymbol{X}) = \frac{\mathrm{L}(\hat{\theta}\text{-unrestricted})}{\mathrm{L}(\widetilde{\theta}\text{-restricted})}.$$

Suppose that looking through the entire parameter space Θ we cannot find a θ value that makes the data more probable than if we restrict ourselves to the parameter values in Θ_0. Then the restricted likelihood in the denominator is large, making the ratio $\Lambda(\boldsymbol{X})$ small. Thus, a small $\Lambda(\boldsymbol{X})$ corresponds to H_0 being in accordance with the data (Wilks 1962, p. 404). It's a matter of convenience which way one writes the ratio. In the one we've chosen, following Aris Spanos (1986, 1999), the larger the $\Lambda(\boldsymbol{X})$, the more discordant the data are from H_0. This suggests the null would be rejected whenever

$$\Lambda(\boldsymbol{X}) \geq k_\alpha$$

for some value of k_α.

So far all of this was to form the distance measure $\Lambda(\boldsymbol{X})$. It's looking somewhat the same as the Likelihoodist account. Yet we know that the additional step 3 that error statistics demands is to compute the probability of $\Lambda(\boldsymbol{X})$ under different hypotheses. Merely reporting likelihood ratios does not produce meaningful control of errors; nor do likelihood ratios mean the same thing in different contexts. So N-P consider the probability distribution of $\Lambda(\boldsymbol{X})$, and they want to ensure the probability of the event $\{\Lambda(\boldsymbol{X}) \geq k_\alpha\}$ is sufficiently small under H_0. They set k_α so that

$$\Pr(\Lambda(\boldsymbol{X}) \geq k_\alpha; H_0) = \alpha$$

for small α. Equivalently, they want to ensure high probability of accordance with H_0 just when it adequately describes the data generation process. Note the complement:

$$\Pr(\Lambda(\boldsymbol{X}) < k_\alpha; H_0) = (1 - \alpha).$$

The event statement to the left of ";" does not reverse positions with H_0 when you form the complement, H_0 stays where it is.

The set of data points leading to $(\Lambda(\boldsymbol{X}) \geq k_\alpha)$ is what N-P call the *critical region* or *rejection region* of the test $\{\boldsymbol{x}: \Lambda(\boldsymbol{X}) \geq k_\alpha\}$ – the set of outcomes that will be taken to reject H_0 or, in our terms, to infer a discrepancy from H_0 in the direction of H_1. Specifying the test procedure, in other words, boils down to specifying the rejection (of H_0) region.

Monotonicity. Following Fisher's goal of maximizing sensitivity, N-P seek to maximize the capability of detecting discrepancies from H_0 when they exist. We need the sampling distribution of $\Lambda(\boldsymbol{X})$, but in practice, $\Lambda(\boldsymbol{X})$ is rarely in a form that one could easily derive this. $\Lambda(\boldsymbol{X})$ has to be transformed in clever ways to yield a test statistic $\mathrm{d}(\boldsymbol{X})$, a function of the sample that has a known distribution under H_0. A general trick to finding a suitable test statistic $\mathrm{d}(\boldsymbol{X})$ is to find a function $\mathrm{h}(\cdot)$ of $\Lambda(\boldsymbol{X})$ that is *monotonic* with respect to a statistic $\mathrm{d}(\boldsymbol{X})$. The greater $\mathrm{d}(\boldsymbol{X})$ is, the greater the likelihood ratio; the smaller d($\boldsymbol{X}$) is, the smaller the likelihood ratio. Having transformed $\Lambda(\boldsymbol{X})$ into the test statistic d($\boldsymbol{X}$), the rejection region becomes

$$\text{Rejection Region, RR} \coloneqq \{\boldsymbol{x}\text{: d}(\boldsymbol{x}) \geq c_\alpha\},$$

the set of data points where d($\boldsymbol{x}$) $\geq c_\alpha$. All other data points belong to the "non-rejection" or "acceptance" region, NR. At first Neyman and Pearson introduced an "undecided" region, but tests are most commonly given such that the RR and NR regions exhaust the entire sample space S. The term "acceptance," Neyman tells us, was merely shorthand: "The phrase 'do not reject *H*' is longish and cumbersome . . . My own preferred substitute for 'do not reject *H*' is 'no evidence against *H* is found'" (Neyman 1976, p. 749). That is the interpretation that should be used.

The use of the $\Lambda(\cdot)$ criterion began as E. Pearson's intuition. Neyman was initially skeptical. Only later did he show it could be the basis for good and even optimal tests.

> Having established the usefulness of the Λ-criterion, we realized that it was essential to explore more fully the sense in which it led to tests which were likely to be effective in detecting departures from the null hypothesis. So far we could only say that it seemed to appeal to intuitive requirements for a good test. (E. Pearson 1970 p. 470, I replace λ with Λ)

Many other desiderata for good tests present themselves.

We want a higher and higher value for Pr(d($\boldsymbol{X}$) $\geq c_\alpha$; θ_1) as the discrepancy ($\theta_1 - \theta_0$) increases. That is, the larger the discrepancy, the easier (more probable) it should be to detect it. This came to be known as the *power function*. Likewise, the power should increase as the sample size increases, and as the variability decreases. The point is that Neyman and Pearson did not start out with a conception of optimality. They groped for criteria that intuitively made sense and that reflected Fisher's tests and theory of estimation. There are some early papers in 1928, but the N-P classic result isn't until the paper in 1933.

Powerful Tests. Pearson describes the days when he and Neyman are struggling to compare various different test statistics – Neyman is in Poland, he is in England. Pearson found himself simulating power for different test statistics and tabling the results. He calls them "empirical power functions." Equivalently, he made tables of the complement to the empirical power function: "what was tabled was the percentage of samples for which a test at 5 percent level failed to establish significance, as the true mean shifted from μ_0 by steps of $\sigma/\surd n$ (ibid. p. 471). He's construing the test's capabilities in terms of percentage of samples. The formal probability distributions serve as shortcuts to cranking out the percentages. "While the results were crude, they show that our thoughts were turning towards the justification of tests in terms of power"(ibid.).

While Pearson is busy experimenting with simulated power functions, Neyman writes to him in 1931 of difficulties he is having in more complicated cases, saying: I found a test in which, paradoxically, "*the true hypothesis will be rejected more often than some of the false ones*. I told Lola [his wife] that we had invented such a test. She said: 'good boys!'" (ibid. p. 472). A test should have a higher probability of leading to a rejection of H_0 when H_1: $\theta \in \Theta_1$ than when H_0: $\theta \in \Theta_0$. After Lola's crack, pretty clearly, they would insist on *unbiased tests*: the probability of rejecting H_0 when it's true or adequate is always less than that of rejecting it when it's false or inadequate. There are direct parallels with properties of good estimators of θ (although we won't have time to venture into that).

Tests that violate unbiasedness are sometimes called "worse than useless" (Hacking 1965, p. 99), but when you read for example in Gigerenzer and Marewski (2015) that N-P found Fisherian tests "worse than useless" (p. 427), there is a danger of misinterpretation. N-P aren't bad-mouthing Fisher. They know he wouldn't condone this, but want to show that without making restrictions explicit, it's possible to end up with such unpalatable tests. In the case of two-sided tests, the additional criterion of unbiasedness led to uniformly most powerful (UMP) unbiased tests.

Consistent Tests. Unbiasedness by itself isn't a sufficient property for a good test; it needs to be supplemented with the property of *consistency*. This requires that, as the sample size n increases without limit, the probability of detecting any discrepancy from the null hypothesis (the power) should approach 1. Let's consider a test statistic that is unbiased yet inconsistent. Suppose we are testing the mean of a Normal distribution with σ known. The test statistic to which the Λ gives rise is

$$\mathrm{d}(\boldsymbol{X}) = \surd n(\overline{x} - \theta_0)/\sigma.$$

Say that, rather than using the sample mean $\overline{x}$, we use the average of the first and last values. This is to estimate the mean θ as $\hat{\theta} = 0.5(X_1 + X_n)$. The test statistic is then $\surd 2(\hat{\theta} - \theta_0)/\sigma$. This is an unbiased estimator of θ. The distribution of $\hat{\theta}$ is $\mathrm{N}(\theta, \sigma^2/2)$. Even though this is unbiased and enables control of the Type I error, it is inconsistent. The result of looking only at two outcomes is that the power does not increase as n increases. The power of this test is much lower than a test using the sample mean for any $n > 2$. If you come across a criticism of tests, make sure *consistency* is not being violated.

Historical Sidelight. Except for short visits and holidays, their work proceeded by mail. When Pearson visited Neyman in 1929, he was shocked at the conditions in which Neyman and other academics lived and worked in Poland. Numerous letters from Neyman describe the precarious position in his statistics lab: "You may have heard that we have in Poland a terrific crisis in everything" [1931] (C. Reid 1998, p. 99). In 1932, "I simply cannot work; the crisis and the struggle for existence takes all my time and energy" (Lehmann 2011, p. 40). Yet he managed to produce quite a lot. While at the start, the initiative for the joint work was from Pearson, it soon turned in the other direction with Neyman leading the way.

By comparison, Egon Pearson's greatest troubles at the time were personal: He had fallen in love "at first sight" with a woman engaged to his cousin George Sharpe, and she with him. She returned the ring the very next day, but Egon still gave his cousin two years to win her back (C. Reid 1998, p. 86). In 1929, buoyed by his work with Neyman, Egon finally declares his love and they are set to be married, but he let himself be intimidated by his father, Karl, deciding "that I could not go against my family's opinion that I had stolen my cousin's fiancée . . . at any rate my courage failed" (ibid., p. 94). Whenever Pearson says he was "suddenly smitten" with doubts about the justification of tests while gazing on the fruit station that his cousin directed, I can't help thinking he's also referring to this woman (ibid., p. 60). He was lovelorn for years, but refused to tell Neyman what was bothering him.

N-P Tests in Their Usual Formulation: Type I and II Error Probabilities and Power

Whether we accept or reject or remain in doubt, say N-P (1933, p. 146), it must be recognized that we can be wrong. By choosing a distance measure d($\boldsymbol{X}$) wherein the probability of different distances may be computed, if the source of the data is $H_{0,}$ we can determine the probability of an erroneous rejection of H_0 – a Type I error.

The test specification that dovetailed with the Fisherian tests in use began by ensuring the probability of a Type I error – an erroneous rejection of the null – is fixed at some small number, α, the *significance level* of the test:

Type I error probability = $\Pr(\mathrm{d}(\boldsymbol{X}) \geq c_\alpha; H_0) \leq \alpha$.

Compare the Type I error probability and the P-value:

***P*-value**: Pr(d($\boldsymbol{X}$) ≥ d($\boldsymbol{x}_0$); H_0) = p($\boldsymbol{x}_0$).

So the N-P test could easily be given in terms of the P-value:

Reject H_0 iff p($\boldsymbol{x}_0$) ≤ α.

Equivalently, the rejection (of H_0) region consists of those outcomes whose *P*-value is less than or equal to α. Reflecting the tests commonly used, N-P suggest the Type I error be viewed as the "more important" of the two. Let the relevant hypotheses be H_0: $\theta = \theta_0$ vs. H_1: $\theta > \theta_0$.

The Type II error is failing to reject the null when it is false to some degree. The test leads you to declare "no evidence of discrepancy from H_0" when H_0 is false, and a discrepancy exists. The alternative hypothesis H_1 contains more than a single value of the parameter, it is *composite*. So, abbreviate by ß(θ_1): the Type II error probability assuming $\theta = \theta_1$, for θ_1 values in the alternative region H_1:

Type II error probability (at θ_1) = Pr(d($\boldsymbol{X}$) < c_α; θ_1) = ß(θ_1),
for $\theta_1 \in \Theta_1$.

In Figure 3.2, this is the area to the left of c_α, the vertical dotted line, under the H_1 curve. The shaded area, the complement of the Type II error probability (at θ_1), is the *power* of the test (at θ_1):

Power of the test (POW) (at θ_1) = Pr(d($\boldsymbol{X}$) ≥ c_α; θ_1).

This is the area to the right of the vertical dotted line, under the H_1 curve, in Figure 3.2. Note d($\boldsymbol{x}_0$) and c_α are always approximations expressed as decimals. For continuous cases, Pr is the probability density.

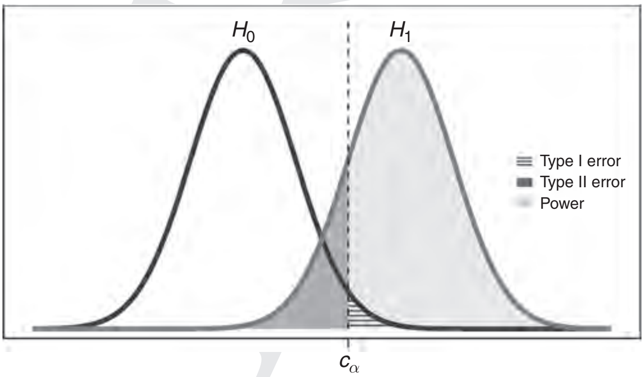

Figure 3.2 Type II error and power.

A *uniformly most powerful* (UMP) N-P test of a hypothesis at level α is one that minimizes ß(θ_1), or, equivalently, maximizes the power for all $\theta > \theta_0$. One reason alternatives are often not made explicit is the property of being a best test for any alternative. We'll explore power, an often-misunderstood creature, further in Excursion 5.

Although the manipulations needed to derive a test statistic using a monotonic mapping of the likelihood ratio can be messy, it's exhilarating to deduce them. Wilks (1938) derived a general asymptotic result, which does not require such manipulations. He showed that, under certain regularity conditions, as n goes to infinity one can define the asymptotic test, where "~" denotes "is distributed as".

$$2\ln\Lambda(\mathbf{X}) \sim \chi^2(r), \text{ under } H_0, \text{ with rejection region RR} \coloneqq \{\boldsymbol{x}\text{: } 2\ln\Lambda(\boldsymbol{x}) \geq c_\alpha\},$$

where $\chi^2(r)$ denotes the chi-square distribution with r degrees of freedom determined by the restrictions imposed by H_0.[4] The monotonicity of the likelihood ratio condition holds for familiar models including one-parameter variants of the Normal, Gamma, Beta, Binomial, Negative Binomial, Poisson (the Exponential family), the Uniform, Logistic, and others (Lehmann 1986). In a wide variety of tests, the Λ principle gives tests with all of the intuitively desirable test properties (see Spanos 2018, chapter 13).

Performance versus Severity Construals of Tests

"The work [of N-P] quite literally transformed mathematical statistics" (C. Reid 1998, p. 104). The idea that appraising statistical methods revolves around optimality (of some sort) goes viral. Some compared it "to the effect of the theory of relativity upon physics" (ibid.). Even when the optimal tests were absent, the optimal properties served as benchmarks against which the performance of methods could be gauged. They had established a new pattern for appraising methods, paving the way for Abraham Wald's decision theory, and the seminal texts by Lehmann and others. The rigorous program overshadowed the more informal Fisherian tests. This came to irk Fisher. Famous feuds between Fisher and Neyman erupted as to whose paradigm would reign supreme. Those who sided with Fisher erected examples to show that tests could satisfy predesignated criteria and long-run error control while leading to counterintuitive tests in specific cases. That was Barnard's point on the eclipse

[4] The general likelihood ratio $\Lambda(\boldsymbol{X})$ should be contrasted with the simple likelihood ratio associated with the well-known Neyman–Pearson (N-P) lemma, which assumes that the parameter space Θ includes only two values, i.e., $\Theta \coloneqq (\theta_0, \theta_1)$. In such a case no estimation is needed because one can take the simple likelihood ratio. Even though the famous lemma for UMP tests uses the highly artificial case of point against point hypotheses (θ_0, θ_1), it is erroneous to suppose the recommended tests are intended for this case. A UMP test, after all, alludes to all the possible parameter values, so just picking two and ignoring the others would not be UMP.

experiments (Section 3.1): no one would consider the class of repetitions as referring to the hoped-for 12 photos, when in fact only some smaller number were usable. We'll meet up with other classic chestnuts as we proceed.

N-P tests began to be couched as formal mapping rules taking data into "reject H_0" or "do not reject H_0" so as to ensure the probabilities of erroneous rejection and erroneous acceptance are controlled at small values, independent of the true hypothesis and regardless of prior probabilities of parameters. Lost in this *behavioristic* formulation was how the test criteria naturally grew out of the requirements of probative tests, rather than good long-run performance. Pearson underscores this in his paper (1947) in the epigraph of Section 3.2: Step 2 comes before Step 3. You must first have a sensible distance measure. Since tests that pass muster on performance grounds can simultaneously serve as probative tests, the severe tester breaks out of the behavioristic prison. Neither Neyman nor Pearson, in their applied work, was wedded to it. Where performance and probativeness conflict, probativeness takes precedent. Two decades after Fisher allegedly threw Neyman's wood models to the floor (Section 5.8), Pearson (1955) tells Fisher: "From the start we shared Professor Fisher's view that in scientific enquiry, a statistical test is 'a means of learning'" (p. 206):

> . . . it was not till after the main lines of this theory had taken shape with its necessary formalization in terms of critical regions, the class of admissible hypotheses, the two sources of error, the power function, etc., that the fact that there was a remarkable parallelism of ideas in the field of acceptance sampling became apparent. Abraham Wald's contributions to decision theory of ten to fifteen years later were perhaps strongly influenced by acceptance sampling problems, but that is another story. (ibid., pp. 204–5)

In fact, the tests as developed by Neyman–Pearson began as an attempt to obtain tests that Fisher deemed intuitively plausible, and this goal is easily interpreted as that of computing and controlling the severity with which claims are inferred.

Not only did Fisher reply encouragingly to Neyman's letters during the development of their results, it was Fisher who first informed Neyman of the split of K. Pearson's duties between himself and Egon, opening up the possibility of Neyman's leaving his difficult life in Poland and gaining a position at University College in London. Guess what else? Fisher was a referee for the all-important N–P 1933 paper, and approved of it.

> To Neyman it has always been a source of satisfaction and amusement that his and Egon's fundamental paper was presented to the Royal Society by Karl Pearson, who was hostile and skeptical of its contents, and favorably refereed by the formidable Fisher,

who was later to be highly critical of much of the Neyman–Pearson theory. (C. Reid 1998, p. 103)

Souvenir J: UMP Tests

Here are some familiar Uniformly Most Powerful (UMP) unbiased tests that fall out of the Λ criterion (letting μ be the mean):

(1) One-sided Normal test. Each X_i is NIID, $\mathrm{N}(\mu, \sigma^2)$, with σ known: H_0: $\mu \leq \mu_0$ against H_1: $\mu > \mu_0$.

$$\mathrm{d}(\boldsymbol{X}) = \sqrt{n}(\overline{X} - \mu_0)/\sigma, \;\; \mathrm{RR}(\alpha) = \{\boldsymbol{x}\colon \mathrm{d}(\boldsymbol{x}) \geq c_\alpha\}.$$

Evaluating the Type I error probability requires the distribution of $\mathrm{d}(\boldsymbol{X})$ under H_0: $\mathrm{d}(\boldsymbol{X}) \sim \mathrm{N}(0,1)$.

Evaluating the Type II error probability (and power) requires the distribution of $\mathrm{d}(\boldsymbol{X})$ under $H_1[\mu = \mu_1]$:

$$\mathrm{d}(\boldsymbol{X}) \sim \mathrm{N}(\delta_1,1), \text{ where } \delta_1 = \sqrt{n}(\mu_1 - \mu_0)/\sigma.$$

(2) One-sided Student's t test. Each X_i is NIID, $\mathrm{N}(\mu, \sigma^2)$, σ unknown: H_0: $\mu \leq \mu_0$ against H_1: $\mu > \mu_0$:

$$\mathrm{d}(\boldsymbol{X}) = \sqrt{n}(\overline{X} - \mu_0)/s, \quad \mathrm{RR}(\alpha) = \{\boldsymbol{x}\colon \mathrm{d}(\boldsymbol{x}) \geq c_\alpha\},$$

$$s^2 = \left[\frac{1}{(n-1)}\right]\sum(X_i - \overline{X})^2.$$

Two-sided Normal test of the mean H_0: $\mu = \mu_0$ against H_1: $\mu \neq \mu_0$:

$$\mathrm{d}(\boldsymbol{X}) = \sqrt{n}(\overline{X} - \mu_0)/s, \quad \mathrm{RR}(\alpha) = \{\boldsymbol{x}\colon |\mathrm{d}(\boldsymbol{x})| \geq c_\alpha\}.$$

Evaluating the Type I error probability requires the distribution of $\mathrm{d}(\boldsymbol{X})$ under H_0: $\mathrm{d}(\boldsymbol{X}) \sim \mathrm{St}(n - 1)$, the Student's t distribution with $(n - 1)$ degrees of freedom (df).

Evaluating the Type II error probability (and power) requires the distribution of $\mathrm{d}(\boldsymbol{X})$ under $H_1[\mu = \mu_1]$: $\mathrm{d}(\boldsymbol{X}) \sim \mathrm{St}(\delta_1, n - 1)$, where $\delta_1 = \sqrt{n}(\mu_1 - \mu_0)/\sigma$ is the non-centrality parameter.

This is the UMP, unbiased test.

(3) The difference between two means (where it is assumed the variances are equal):

$H_0\colon \gamma := \mu_1 - \mu_2 = \gamma_0$ against $H_1\colon \gamma_1 \neq \gamma_0$.

A Uniformly Most Powerful Unbiased (UMPU) test is defined by

$$\tau(\mathbf{Z}) = \frac{\sqrt{n}\left[(\overline{X}_n - \overline{Y}_n) - \gamma_0\right]}{s\sqrt{2}}, \mathrm{RR} = \left\{\mathbf{z}\colon |\tau(\mathbf{z})| \geq c_\alpha\right\}.$$

Under H_0: $\tau(\mathbf{Z}) = \dfrac{\sqrt{n}\left[(\overline{X}_n - \overline{Y}_n) - \gamma_0\right]}{s\sqrt{2}} \sim \mathrm{St}(2n{-}2)$,

under $H_1[\gamma = \gamma_1]$: $\tau(\mathbf{Z}) \sim \mathrm{St}(\delta_1; 2n{-}2)$, $\delta_1 = \dfrac{\sqrt{n}\,(\gamma_1 - \gamma_0)}{\sigma\sqrt{2}}$, for $\gamma_1 \neq \gamma_0$.

Many excellent sources of types of tests exist, so I'll stop with these.

Exhibit (i): N-P Methods as Severe Tests: First Look (Water Plant Accident). There's been an accident at a water plant where our ship is docked, and the cooling system had to be repaired. It is meant to ensure that the mean temperature of discharged water stays below the temperature that threatens the ecosystem, perhaps not much beyond 150 degrees Fahrenheit. There were 100 water measurements taken at randomly selected times and the sample mean $\overline{x}$ computed, each with a known standard deviation $\sigma = 10$. When the cooling system is effective, each measurement is like observing $X \sim \mathrm{N}(150, 10^2)$. Because of this variability, we expect different 100-fold water samples to lead to different values of $\overline{X}$, but we can deduce its distribution. If each $X \sim \mathrm{N}(\mu = 150, 10^2)$ then $\overline{X}$ is also Normal with $\mu = 150$, but the standard deviation of $\overline{X}$ is only $\sigma/\sqrt{n} = 10/\sqrt{100} = 1$. So $\overline{X} \sim \mathrm{N}(\mu = 150, 1)$.

It is the distribution of $\overline{X}$ that is the relevant sampling distribution here. Because it's a large random sample, the sampling distribution of $\overline{X}$ is Normal or approximately so, thanks to the Central Limit Theorem. Note the mean of the sampling distribution of $\overline{X}$ is the same as the underlying mean, both are μ. The frequency link was *created* by randomly selecting the sample, and we assume for the moment it was successful. Suppose they are testing

$$H_0\colon \mu \leq 150 \text{ vs. } H_1\colon \mu > 150.$$

The test rule for $\alpha = 0.025$ is

$$\text{Reject } H_0\text{: iff } \overline{X} \geq 150 + c_\alpha\sigma/\sqrt{100} = 150 + 1.96(1) = 151.96,$$

$$\text{since } c_\alpha = 1.96.$$

For simplicity, let's go to the 2-standard error cut-off for rejection:

$$\text{Reject } H_0 (\text{infer there's an indication that } \mu > 150) \text{ iff } \overline{X} \geq 152.$$

The test statistic d($\boldsymbol{x}$) is a standard Normal variable: $Z = \sqrt{100}(\overline{X} - 150)/10 = \overline{X} - 150$, which, for $\overline{x} = 152$, is 2. The area to the right of 2 under the standard Normal is around 0.025.

Now we begin to move beyond the strict N-P interpretation. Say $\overline{x}$ is just significant at the 0.025 level ($\overline{x} = 152$). What warrants taking the data as indicating $\mu > 150$ is not that they'd rarely be wrong in repeated trials on cooling systems by acting this way – even though that's true. There's a good indication that it's not in compliance right now. Why? *The severity rationale*: Were the mean temperature no higher than 150, then over 97% of the time their method would have resulted in a lower mean temperature than observed. Were it clearly in the safe zone, say $\mu = 149$ degrees, a lower observed mean would be even more probable. Thus, $\overline{x} = 152$ indicates *some* positive discrepancy from H_0 (though we don't consider it rejected by a single outcome). They're going to take another round of measurements before acting. In the context of a policy action, to which this indication might lead, some type of loss function would be introduced. We're just considering the evidence, based on these measurements; all for illustrative purposes.

Severity Function

I will abbreviate "the severity with which claim C passes test T with data $\boldsymbol{x}$":

$$\text{SEV(test T, outcome } \boldsymbol{x}\text{, claim } C).$$

Reject/Do Not Reject will be interpreted inferentially, in this case as an indication or evidence of the presence or absence of discrepancies of interest.

Let us suppose we are interested in assessing the severity of C: $\mu > 153$. I imagine this would be a full-on emergency for the ecosystem!

Reject $\boldsymbol{H_0}$. Suppose the observed mean is $\overline{x} = 152$, just at the cut-off for rejecting H_0:

$$\mathrm{d}(\boldsymbol{x}_0) = \sqrt{100}(152 - 150)/10 = 2.$$

The data reject H_0 at level 0.025. We want to compute

$$\mathrm{SEV}(\mathrm{T}, \overline{x} = 152, C\text{: } \mu > 153).$$

We may say: "the data accord with C: $\mu > 153$," that is, severity condition (S-1) is satisfied; but severity requires there to be at least a reasonable probability of a worse fit with C if C is false (S-2). Here, "worse fit with C" means $\overline{x} \leq 152$ (i.e., $\mathrm{d}(\boldsymbol{x}_0) \leq 2$). Given it's continuous, as with all the following examples, < or ≤ give the same result. The context indicates which is more useful. This probability must be high for C to pass severely; if it's low, it's BENT.

Table 3.1 Reject in test T+: H_0: $\mu \leq 150$ vs. H_1: $\mu > 150$ with $\bar{x} = 152$

Claim $\mu > \mu_1$	Severity $\Pr(\overline{X} \leq 152;\, \mu = \mu_1)$
$\mu > 149$	0.999
$\mu > 150$	0.97
$\mu > 151$	0.84
$\mu > 152$	0.5
$\mu > 153$	0.16

We need $\Pr(\overline{X} \leq 152; \mu > 153$ is false). To say $\mu > 153$ is false is to say $\mu \leq 153$. So we want $\Pr(\overline{X} \leq 152; \mu \leq 153)$. But we need only evaluate severity at the point $\mu = 153$, because this probability is even greater for $\mu < 153$:

$$\Pr(\overline{X} \leq 152; \mu = 153) = \Pr(Z \leq -1) = 0.16.$$

Here, $Z = \sqrt{100}(152 - 153)/10 = -1$. Thus $\text{SEV}(\text{T}, \bar{x} = 152, C : \mu > 153) = 0.16$. Very low. Our minimal severity principle blocks $\mu > 153$ because it's fairly probable (84% of the time) that the test would yield an even larger mean temperature than we got, if the water samples came from a body of water whose mean temperature is 153. Table 3.1 gives the severity values associated with different claims, given $\bar{x} = 152$. Call tests of this form T+

In each case, we are making inferences of form $\mu > \mu_1 = 150 + \gamma$, for different values of γ. To merely infer $\mu > 150$, the severity is 0.97 since $\Pr(\overline{X} \leq 152; \mu = 150) = \Pr(Z \leq 2) = 0.97$. While the data give an indication of non-compliance, $\mu > 150$, to infer C: $\mu > 153$ would be making mountains out of molehills. In this case, the observed difference just hit the cut-off for rejection. N-P tests leave things at that coarse level in computing power and the probability of a Type II error, but severity will take into account the actual outcome. Table 3.2 gives the severity values associated with different claims, given $\bar{x} = 153$.

If "the major criticism of the Neyman–Pearson frequentist approach" is that it fails to provide "error probabilities fully varying with the data," as J. Berger alleges (2003, p. 6), then we've answered the major criticism.

Non-rejection. Now suppose $\bar{x} = 151$, so the test does not reject H_0. The standard formulation of N-P (as well as Fisherian) tests stops there. But we want to be alert to a fallacious interpretation of a "negative" result: inferring there's no positive discrepancy from $\mu = 150$. No (statistical) evidence of non-compliance isn't evidence of compliance; here's why. We have (S-1): the data

Table 3.2 Reject in test T+: H_0: $\mu \le 150$ vs. H_1: $\mu > 150$ with $\bar{x} = 153$

Claim $\mu > \mu_1$	**Severity (with $\overline{X}$ = 153)** $\Pr(\overline{X} \le 153; \mu = \mu_1)$
$\mu > 149$	~1
$\mu > 150$	0.999
$\mu > 151$	0.97
$\mu > 152$	0.84
$\mu > 153$	0.5

Table 3.3 Non-reject in test T+: H_0: $\mu \le 150$ vs. H_1: $\mu > 150$ with $\bar{x} = 151$

Claim $\mu \le \mu_1$	**Severity** $\Pr(\overline{X} > 151; \mu = \mu_1)$
$\mu \le 150$	0.16
$\mu \le 150.5$	0.3
$\mu \le 151$	0.5
$\mu \le 152$	0.84
$\mu \le 153$	0.97

"accord with" H_0, but what if the test had little capacity to have alerted us to discrepancies from 150? The alert comes by way of "a worse fit" with H_0 – namely, a mean $\bar{x} = 151$. Condition (S-2) requires us to consider $\Pr(\overline{X} > 151; \mu = 150)$, which is only 0.16. To get this, standardize $\overline{X}$ to obtain a standard Normal variate: $Z = \sqrt{100}(151 - 150)/10 = 1$; and $\Pr(\overline{X} > 151; \mu = 150) = 0.16$. Thus, $\text{SEV}(\text{T+}, \bar{x} = 151, C\text{: } \mu \le 150) = \text{low}(0.16)$. Table 3.3 gives the severity values associated with different inferences of form $\mu \le \mu_1 = 150 + \gamma$, given $\bar{x} = 151$.

Can they at least say that $\bar{x} = 151$ is a good indication that $\mu \le 150.5$? No, $\text{SEV}(\text{T+}, \bar{x} = 151, C\text{: } \mu \le 150.5) \simeq 0.3$, [$Z = 151 - 150.5 = 0.5$]. But $\bar{x} = 151$ is a good indication that $\mu \le 152$ and $\mu \le 153$ (with severity indications of 0.84 and 0.97, respectively).

You might say, assessing severity is no different from what we would do with a judicious use of existing error probabilities. That's what the severe tester says. Formally speaking, it may be seen merely as a good rule of thumb to avoid fallacious interpretations. What's new is the statistical philosophy behind it.

We no longer seek either probabilism or performance, but rather using relevant error probabilities to assess and control severity.[5]

3.3 How to Do All N-P Tests Do (and More) While a Member of the Fisherian Tribe

When Karl Pearson retired in 1933, he refused to let his chair go to Fisher, so they split the department into two: Fisher becomes Galton Chair and Head of Eugenics, while Egon Pearson becomes Head of Applied Statistics. They are one floor removed (Fisher on top)! The Common Room had to be "carefully shared," as Constance Reid puts it: "Pearson's group had afternoon tea at 4; and at 4:30, when they were safely out of the way, Fisher and his group trouped in" (C. Reid 1998, p. 114). Fisher writes to Neyman in summer of 1933 (cited in Lehmann 2011, p. 58):

> You will be interested to hear that the Dept. of Statistics has now been separated officially from the Galton Laboratory. I think Egon Pearson is designated as Reader in Statistics. This arrangement will be much laughed at, but it will be rather a poor joke . . . I shall not lecture on statistics, but probably on 'the logic of experimentation'.

Finally E. Pearson was able to offer Neyman a position at University College, and Neyman, greatly relieved to depart Poland, joins E. Pearson's department in 1934.[6]

Neyman doesn't stay long. He leaves London for Berkeley in 1938, and develops the department into a hothouse of statistics until his death in 1981. His first Ph.D. student is Erich Lehmann in 1946. Lehmann's *Testing Statistical Hypotheses*, 1959, the canonical N-P text, developed N-P methods very much in the mode of the N-P-Wald, behavioral-decision language. I find it interesting that even Neyman's arch opponent, subjective Bayesian Bruno de Finetti, recognized that "inductive behavior . . . that was for Neyman simply a slogan underlining and explaining the difference between his own, the Bayesian and the Fisherian formulations" became, with Wald's work,

[5] Initial developments of the severity idea were Mayo (1983, 1988, 1991, 1996). In Mayo and Spanos (2006, 2011), it was developed much further.

[6] "By the fall of 1932 there appeared to be several reasons why Neyman might never become a professor in Poland. One was his subject matter, which was not generally recognized as an academic specialty. Another was the fact that he was married to a Russian – and an independent, outspoken Russian who lived on occasion apart from her husband, worked and painted in Paris, traveled on a freighter as a nurse for the adventure of it, and sometimes led tourist excursions into the Soviet Union." (C. Reid 1998, p. 105).

"something much more substantial." De Finetti called this "the involuntarily destructive aspect of Wald's work" (1972, p. 176). Cox remarks:

> [T]here is a distinction between the Neyman–Pearson formulation of testing regarded as clarifying the meaning of statistical significance via hypothetical repetitions and that same theory regarded as in effect an instruction on how to implement the ideas by choosing a suitable α in advance and reaching different decisions accordingly. The interpretation to be attached to accepting or rejecting a hypothesis is strongly context-dependent . . . (Cox 2006a, p. 36)

If N-P long-run performance concepts serve to clarify the meaning of statistical significance tests, yet are not to be applied literally, but rather in some inferential manner – call this the *meaning* vs. *application distinction* – the question remains – how?

My answer, in terms of severity, may be used whether you prefer the N-P tribe (tests or confidence intervals) or the Fisherian tribe. What would that most eminent Fisherian, Sir David Cox, say? In 2004, in a session we were in on statistical philosophy, at the semi-annual Lehmann conference, we asked: Was it possible to view "Frequentist Statistics as a Theory of Inductive Inference"? If this sounds familiar it's because it echoes a section from Neyman's quarrel with Carnap (Section 2.7), but how does a Fisherian answer it? We began "with the core elements of significance testing in a version very strongly related to but in some respects different from both Fisherian and Neyman–Pearson approaches, at least as usually formulated" (Mayo and Cox 2006, p. 80). First, there is no suggestion that the significance test would typically be the only analysis reported. Further, we agree that "the justification for tests will not be limited to appeals to long-run behavior but will instead identify an inferential or evidential rationale" (ibid., p. 81).

With N-P results available, it became easier to understand why intuitively useful tests worked for Fisher. N-P and Fisherian tests, while starting from different places, "lead to the same destination" (with few exceptions) (Cox 2006a, p. 25). Fisher begins with seeking a test statistic that reduces the data as much as possible, and this leads him to a *sufficient* statistic. Let's take a side tour to sufficiency.

Exhibit (ii): Side Tour of Sufficient Statistic. Consider n independent trials $\boldsymbol{X} := (X_1, X_2, \ldots, X_n)$ each with a binary outcome (0 or 1), where the probability of success is an unknown constant θ associated with Bernoulli trials. The number of successes in n trials, $Y = X_1 + X_2 + \cdots + X_n$ is Binomially distributed with parameters θ and n. The sample mean, which is just $\overline{X} = Y/n$, is a natural estimator of θ with a highly desirable property: it is *sufficient*, i.e., it is a function of the *sufficient* statistic Y. Intuitively, a sufficient statistic reduces the n-dimensional sample $\boldsymbol{X}$ into a statistic of much smaller dimensionality without losing any relevant information for inference purposes. Y reduces the n-fold outcome $\boldsymbol{x}$ to one dimension: the number of successes in n trials. The parameter of the Binomial model θ also has one dimension (the probability of success on each trial).

Formally, a statistic Y is said to be sufficient for θ when the distribution of the sample is no longer a function of θ when conditioned on Y, i.e., $\mathrm{f}(x \mid y)$ does not depend on θ,

$$\mathrm{f}(x; \theta) = \mathrm{f}(y; \theta)\ \mathrm{f}(x|y).$$

Knowing the distribution of the sufficient statistic Y suffices to compute the probability of any data set $\boldsymbol{x}$. The test statistic $\mathrm{d}(\boldsymbol{X})$ in the Binomial case is $\sqrt{}n(\overline{X} - \theta_0)/\sigma$, $\sigma = \sqrt{[\theta(1-\theta)]}$ and, as required, gets larger as $\overline{X}$ deviates from θ_0. Thanks to $\overline{X}$ being a function of the sufficient statistic Y, it is the basis for a test statistic with maximal sensitivity to inconsistencies with the null hypothesis.

The Binomial experiment is equivalent to having been given the data $\boldsymbol{x}_0 = (x_1, x_2, \ldots, x_n)$ in two stages (Cox and Mayo 2010, p. 285):

First, you're told the value of Y, the number of successes out of n Bernoulli trials. Then an inference can be drawn about θ using the sampling distribution of Y.

Second, you learn the value of the specific data, e.g., the first k trials are successes, the rest failure. The second stage is equivalent to observing a realization of the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Y} = \boldsymbol{y}$. If the model is appropriate then "the second phase is equivalent to a random draw from a totally known distribution." All permutations of the sequence of successes and failures are equally probable (ibid., pp 284–5).

"Because this conditional distribution is totally known, it can be used to assess the validity of the assumed model." (ibid.) Notice that for a given $\boldsymbol{x}$ *within* a given Binomial experiment, the ratio of likelihoods at two different values of θ depends on the data only through Y. This is called the *weak likelihood principle* in contrast to the general (or strong) LP in Section 1.5.

Principle of Frequentist Evidence, FEV

Returning to our topic, "Frequentist Statistics as a Theory of Inductive Inference," let me weave together three threads: (1) the Frequentist Principle of Evidence (Mayo and Cox 2006), (2) the divergent interpretations growing out of Cox's taxonomy of test hypotheses, and (3) the links to statistical inference as severe tests. As a starting point, we identified a general principle that we dubbed the Frequentist Principle of Evidence, FEV:

> *FEV(i)*: $\boldsymbol{x}$ is … evidence against H_0 (i.e., evidence of a discrepancy from H_0), if and only if, were H_0 a correct description of the mechanism generating $\boldsymbol{x}$, then, with high probability, this would have resulted in a less discordant result than is exemplified by $\boldsymbol{x}$. (Mayo and Cox 2006, p. 82; substituting $\boldsymbol{x}$ for $\boldsymbol{y}$)

This sounds wordy and complicated. It's much simpler in terms of a quantitative difference as in significance tests. Putting FEV(i) in terms of formal *P*-values, or test statistic d (abbreviating d($\boldsymbol{X}$)):

> *FEV(i)*: $\boldsymbol{x}$ is evidence against H_0 (i.e., evidence of discrepancy from H_0), if and only if the *P*-value $\Pr(d \geq d_0; H_0)$ is very low (equivalently, $\Pr(d < d_0; H_0) = 1 - P$ is very high).

(We used "strong evidence", although I would call it a mere "indication" until an appropriate audit was passed.) Our minimalist claim about bad evidence, no test (BENT) can be put in terms of a corollary (from contraposing FEV(i)):

> *FEV(ia)*: $\boldsymbol{x}$ is poor evidence against H_0 (poor evidence of discrepancy from H_0), if there's a high probability the test would yield a more discordant result, if H_0 is correct.

Note the one-directional 'if' claim in FEV(ia). We wouldn't want to say this is the only way $\boldsymbol{x}$ can be BENT.

Since we wanted to move away from setting a particular small *P*-value, we refer to "*P*-small" (such as 0.05, 0.01) and "*P*-moderate", or "not small" (such as 0.3 or greater). We need another principle in dealing with non-rejections or insignificant results. They are often imbued with two very different false interpretations: one is that (a) non-significance indicates the truth of the null, the other is that (b) non-significance is entirely uninformative.

The difficulty with (a), regarding a modest *P*-value as evidence in favor of H_0, is that accordance between H_0 and $\boldsymbol{x}$ may occur even if rivals to H_0 seriously different from H_0 are true. This issue is particularly acute when the capacity to have detected discrepancies is low. However, as against (b), null results have an important role ranging from the scrutiny of substantive theory – setting bounds to parameters to scrutinizing the capability of a method for finding something out. In sync with our "arguing from error" (Excursion 1), we may infer a discrepancy from H_0 is absent if our test very probably would have alerted us to its presence (by means of a more significant P-value).

FEV(ii): A moderate P-value is evidence of the absence of a discrepancy δ from H_0, only if there is a high probability the test would have given a worse fit with H_0 (i.e., smaller P-value) were a discrepancy δ to exist (ibid., pp. 83–4).

This again is an "if-then" or conditional claim. These are canonical pieces of statistical reasoning, in their naked form as it were. To dress them up to connect with actual questions and problems of inquiry requires context-dependent, background information.

FIRST Interpretations: Fairly Intimately Related to the Statistical Test – Cox's Taxonomy

> In the statistical analysis of scientific and technological data, there is virtually always external information that should enter in reaching conclusions about what the data indicate with respect to the primary question of interest. Typically, these background considerations enter not by a probability assignment but by identifying the question to be asked, designing the study, interpreting the statistical results and relating those inferences to primary scientific ones . . . (Mayo and Cox 2006, p. 84)

David Cox calls for an interpretive guide between a test's mathematical formulation and substantive applications: "I think that more attention to these rather vague general issues is required if statistical ideas are to be used in the most fruitful and wide-ranging way" (Cox 1977, p. 62). There are aspects of the context that go beyond the mathematics but which are Fairly Intimately Related to the Statistical Test (FIRST) interpretations. I'm distinguishing these FIRST interpretations from wider substantive inferences, not that there's a strict red line difference.

While warning that "it is very bad practice to summarise an important investigation solely by a value of P" (1982, p. 327), Cox gives a rich taxonomy of null hypotheses that recognizes how significance tests can function as part of complex and context-dependent inquiries (1977, pp. 51–2). Pure or simple Fisherian significance tests (with no explicit alternative) are housed within the taxonomy, not separated out as some radically distinct entity. If his taxonomy had been incorporated into the routine exposition of tests, we could have avoided much of the confusion we are still suffering with. The proper way to view significance tests acknowledges a variety of problem situations:

- Are we testing parameter values within some overriding model? (fully embedded)
- Are we merely checking if a simplification will do? (nested alternative)
- Do we merely seek the direction of an effect already presumed to exist? (dividing)
- Would a model pass an audit of its assumptions? (test of assumption)
- Should we worry about data that appear anomalous for a theory that has already passed severe tests? (substantive)

Although Fisher, strictly speaking, had only the null hypothesis, and context directed an appropriate test statistic, the result of such a selection is that the test is sensitive to a type of discrepancy. Even if they only become explicit after identifying a test statistic – which some regard as more basic (e.g., Senn) – we may still regard them as alternatives.

Sensitivity Achieved or Attained

For a Fisherian like Cox, a test's power only has relevance pre-data, in planning tests, but, like Fisher, he can measure "sensitivity":

> In the Neyman–Pearson theory of tests, the sensitivity of a test is assessed by the notion of *power*, defined as the probability of reaching a preset level of significance . . . for various alternative hypotheses. In the approach adopted here the assessment is via the distribution of the random variable P, again considered for various alternatives. (Cox 2006a, p. 25)

This is the key: Cox will measure sensitivity by a function we may abbreviate as $\Pi(\gamma)$. Computing $\Pi(\gamma)$ may be regarded as viewing the P-value as a statistic. That is:

$$\Pi(\gamma) = \Pr(P \leq p_{\text{obs}}; \mu_0 + \gamma).$$

The alternative is $\mu_1 = \mu_0 + \gamma$. Using the P-value distribution has a long history and is part of many approaches. Given the P-value inverts the distance, it is clearer and less confusing to formulate $\Pi(\gamma)$ in terms of the test statistic d. $\Pi(\gamma)$ is very similar to *power* in relation to alternative μ_1, except that $\Pi(\gamma)$ considers the observed difference rather than the N-P cut-off c_α:

$$\Pi(\gamma) = \Pr(d \geq d_0; \mu_0 + \gamma),$$

$$\text{POW}(\gamma) = \Pr(d \geq c_\alpha; \mu_0 + \gamma).$$

Π may be called a "sensitivity function," or we might think of $\Pi(\gamma)$ as the "attained power" to detect discrepancy γ (Section 5.3). The nice thing about power is that it's always in relation to an observed difference from a test or null hypothesis, which gives it a reference. Let's agree that Π will always relate to an observed difference from a designated test hypothesis H_0.

Aspects of Cox's Taxonomy

I won't try to cover Cox's full taxonomy, which has taken different forms. I propose that the main delineating features are, first, whether the null and alternative exhaust the answers or parameter space for the given question, and, second, whether the null hypothesis is considered a viable basis for a substantive research claim, or merely as a reference for exploring the way in which it is false. None of these are hard and fast distinctions, but you'll soon see why they are useful. I will adhere closely to what Cox has said about the taxonomy and the applications of FEV; all I add is a proposed synthesis. I restrict myself now to a single parameter. We assume the *P*-value has passed an audit (except where noted).

1. **Fully embedded.** Here we have exhaustive parametric hypotheses governed by a parameter θ, such as the mean deflection of light at the 1919 eclipse, or the mean temperature. H_0: $\mu = \mu_0$ vs. H_1: $\mu > \mu_0$ is a typical N-P setting. Strictly speaking, we may have $\theta = (\mu,k)$ with additional parameters k to be estimated. This formulation, Cox notes, "will suggest the most sensitive test statistic, essentially equivalent to the best estimate of μ" (Cox 2006a, p. 37).

A. P-value is modest (not small): Since the data accord with the null hypothesis, FEV directs us to examine the probability of observing a result more discordant from H_0 if $\mu = \mu_0 + \gamma$: $\Pi(\gamma) = \Pr(d \geq d_0; \mu_0 + \gamma)$.

If that probability is very high, following FEV(ii), the data indicate that $\mu < \mu_0 + \gamma$.

Here $\Pi(\gamma)$ gives the severity with which the test has probed the discrepancy γ. So we don't merely report "no evidence against the null," we report a discrepancy that can be ruled out with severity. "This avoids unwarranted interpretations of consistency with H_0 with insensitive tests . . . [and] is more relevant to specific data than is the notion of power" (Mayo and Cox 2006, p. 89).

B. P-value is small: From FEV(i), a small *P*-value indicates evidence of *some* discrepancy $\mu > \mu_0$ since $\Pr(d < d_0; H_0) = 1 - P$ is large. This is the basis for ordinary (statistical) falsification.

However, we add, "if a test is so sensitive that a *P*-value as small as or even smaller than the one observed is probable even when $\mu \leq \mu_0 + \gamma$, then a small value of *P*" is poor evidence of a discrepancy from H_0 in excess of γ (ibid.). That is, from FEV(ia), if $\Pi(\gamma) = \Pr(d \geq d_0; \mu_0 + \gamma)$ = moderately high (greater than 0.3, 0.4, 0.5), then there's poor grounds for inferring $\mu > \mu_0 + \gamma$. This is equivalent to saying the SEV($\mu > \mu_0 + \gamma$) is poor.

There's no need to set a sharp line between significance or not in this construal – extreme cases generally suffice. FEV leads to an inference as to both what's indicated, and what's not indicated. Both are required by a severe tester. Go back to our accident at the water plant. The non-significant result, $\bar{x} = 151$ in testing $\mu \leq 150$ vs. $\mu > 150$, only attains a P-value of 0.16. SEV(C: $\mu > 150.5$) is 0.7 (Table 3.3). Not terribly high, but if that discrepancy was of interest, it wouldn't be ignorable. A reminder: we are not making inferences about point values, even though we need only compute Π at a point. In this first parametric pigeonhole, confidence intervals can be formed, though we wouldn't limit them to the typical 0.95 or 0.99 levels.[7]

2. **Nested alternative** (non-exhaustive). In a second pigeonhole an alternative statistical hypothesis H_1 is considered not "as a possible base for ultimate interpretation but as a device for determining a suitable test statistic" (Mayo and Cox 2006, p. 85). Erecting H_1 may be only a convenient way to detect small departures from a given statistical model. For instance, one may use a quadratic model H_1 to test the adequacy of a linear relation. Even though polynomial regressions are a poor base for final analysis, they are very convenient and interpretable for detecting small departures from linearity. (ibid.)

Failing to reject the null (moderate P-value) might be taken to indicate the simplifying assumption is adequate; whereas rejecting the null (small P-value) is not evidence for alternative H_1. That's because there are lots of non-linear models not probed by this test. The H_0 and H_1 do not exhaust the space.

A. P-value is modest (not small): At best it indicates adequacy of the model in the respects well probed; that is, it indicates the absence of discrepancies that, very probably, would have resulted in a smaller P-value.

B. P-value small: This indicates discrepancy from the null in the direction of the alternative, but it is unwarranted to infer the particular H_1 insofar as other non-linear models could be responsible.

We are still employing the FEV principle, even where it is qualitative.

[7] "A significance test is defined by a set of [critical] regions [w_α] satisfying the following essential requirements. First,

$$w_{\alpha_1} \subset \mathrm{w}_{\alpha_2} \text{ if } \alpha_1 < \alpha_2;$$

this is to avoid such nonsense as saying that data depart significantly from H_0 at the 1% level but not at the 5% level." Next "we require that, for all α, $\Pr(Y \in w_\alpha; H_0) = \alpha$." (Cox and Hinkley 1974, pp. 90–1)

3. **Dividing nulls:** H_0: $\mu = \mu_0$ vs. H_1: $\mu > \mu_0$ and H_0: $\mu = \mu_0$ vs. H_1: $\mu < \mu_0$. In this pigeonhole, we may already know or suspect the null is false and discrepancies exist: but which? Suppose the interest is comparing two or more treatments. For example, compared with a standard, a new drug may increase or may decrease survival rates.

The null hypothesis of zero difference *divides* the possible situations into two qualitatively different regions with respect to the feature tested. To look at both directions, one combines two tests, the first to examine the possibility that $\mu > \mu_0$, say, the second for $\mu < \mu_0$. The overall significance level is twice the smaller *P*-value, because of a "selection effect." One may be wrong in two ways. It is standard to report the observed direction and double the initial *P*-value (if it's a two-sided test).

While a small *P*-value indicates a direction of departure (e.g., which of two treatments is superior), failing to get a small *P*-value here merely tells us the data do not provide adequate evidence even of the direction of any difference. Formally, the statistical test may look identical to the fully embedded case, but the nature of the problem, and your background knowledge, yields a more relevant construal. These interpretations are still FIRST. You can still report the upper bound ruled out with severity, bringing this case closer to the fully embedded case (Table 3.4).

4. **Null hypotheses of model adequacy.** In "auditing" a *P*-value, a key question is: how can I check the model assumptions hold adequately for the data in hand? We distinguish two types of tests of assumptions (Mayo and Cox 2006, p. 89): (i) omnibus and (ii) focused.

(i) With a general *omnibus* test, a group of violations is checked all at once. For example: H_0: IID (independent and identical distribution) assumptions hold vs. H_1: IID is violated. The null and its denial exhaust the possibilities, for the question being asked. However, sensitivity can be so low that failure to reject may be uninformative. On the other hand, a small *P*-value indicates H_1: there's a departure *somewhere*. The very fact of its low sensitivity indicates that when the alarm goes off something's there. But where? Duhemian problems loom. A subsequent task would be to pin this down.

(ii) A *focused* test is sensitive to a specific kind of model inadequacy, such as a lack of independence. This lands us in a situation analogous to the non-exhaustive case in "nested alternatives." Why? Suppose you erect an alternative H_1 describing a particular type of non-independence, e.g., Markov. While a small *P*-value indicates some departure, you cannot infer H_1 so long as various alternative models, not probed by this test, could account for it.

Table 3.4 FIRST Interpretations

Taxon	Remarks	Small *P*-value	*P*-value Not Small
1. Fully embedded exhaustive	H_1 may be the basis for a substantive interpretation	Indicates $\mu > \mu_0 + \gamma$ iff $\Pi(\gamma)$ is low	If $\Pi(\gamma)$ is high, there's poor indication of $\mu > \mu_0 + \gamma$
2. Nested alternatives non-exhaustive	H_1 is set out to search departures from H_0	Indicates discrepancy from H_0 but not grounds to infer H_1	Indicates H_0 is adequate in respect probed
3. Dividing exhaustive	$\mu \leq \mu_0$ vs. $\mu > \mu_0$; a discrepancy is presumed, but in which direction?	Indicates direction of discrepancy If $\Pi(\gamma)$ low, $\mu > \mu_0 + \gamma$ is indicated	Data aren't adequate even to indicate direction of departure
4. Model assumptions (i) omnibus exhaustive	e.g., non-parametric runs test for IID (may have low power)	Indicates departure from assumptions probed, but not specific violation	Indicates the absence of violations the test is capable of detecting
Model assumptions (ii) focused non-exhaustive	e.g., parametric test for specific type of dependence	Indicates departure from assumptions in direction of H_1 but can't infer H_1	Indicates the absence of violations the test is capable of detecting

It may only give suggestions for alternative models to try. The interest may be in the effect of violated assumptions on the primary (statistical) inference if any. We might ask: Are the assumptions sufficiently questionable to preclude using the data for the primary inference? After a lunch break at Einstein's Cafe, we'll return to the museum for an example of that.

Scotching a Famous Controversy

At a session on the replication crisis at a 2015 meeting of the Society for Philosophy and Psychology, philosopher Edouard Machery remarked as to how, even in so heralded a case as the eclipse tests of GTR, one of the results didn't replicate the other two. The third result pointed, not to Einstein's prediction, but as Eddington ([1920]1987) declared, "with all too good agreement to the 'half-deflection,' that is to say, the Newtonian value" (p. 117). He was alluding to a famous controversy that has grown up surrounding the allegation that Eddington selectively ruled out data that supported the Newtonian "half-value" against the Einsteinian one. Earman and Glymour (1980), among others, alleged that Dyson and Eddington threw out the results unwelcome for GTR for political purposes ("... one of the chief benefits to be derived from the eclipse results was a rapprochement between German and British scientists and an end to talk of boycotting German science" (p. 83)).[8] Failed replication may indeed be found across the sciences, but this particular allegation is mistaken. The museum's display on "Data Analysis in the 1919 Eclipse" shows a copy of the actual notes penned on the Sobral expedition *before* any data analysis:

> May 30, 3 a.m., four of the astrographic plates were developed ... It was found that there had been a serious change of focus ... This change of focus can only be attributed to the unequal expansion of the mirror through the sun's heat ... It seems doubtful whether much can be got from these plates. (Dyson et al. 1920, p. 309)

Although a fair amount of (unplanned) data analysis was required, it was concluded that there was no computing a usable standard error of the estimate. The hypothesis:

> The data $\boldsymbol{x}_0$ (from Sobral astrographic plates) were due to systematic distortion by the sun's heat, not to the deflection of light,

passes with severity. An even weaker claim is all that's needed: we can't compute a valid estimate of error. And notice how very weak the claim to be corroborated is!

[8] Barnard was surprised when I showed their paper to him, claiming it was a good example of why scientists tended not to take philosophers seriously. But in this case even the physicists were sufficiently worried to reanalyze the experiment.

The mirror distortion hypothesis hadn't been predesignated, but it is altogether justified to raise it in auditing the data: It could have been chewing gum or spilled coffee that spoilt the results. Not only that, the same data hinting at the mirror distortion are to be used in testing the mirror distortion hypothesis (though differently modeled)! That sufficed to falsify the requirement that there was no serious change of focus (scale effect) between the eclipse and night plates. Even small systematic errors are crucial because the resulting scale effect from an altered focus quickly becomes as large as the Einstein predicted effect. Besides, the many staunch Newtonian defenders would scarcely have agreed to discount an apparently pro-Newtonian result.

The case was discussed and soon settled in the journals of the time: the brouhaha came later. It turns out that, if these data points are deemed usable, the results actually point to the Einsteinian value, not the Newtonian value. A reanalysis in 1979 supports this reading (Kennefick 2009). Yes, in 1979 the director of the Royal Greenwich Observatory took out the 1919 Sobral plates and used a modern instrument to measure the star positions, analyzing the data by computer.

> [T]he reanalysis provides after-the-fact justification for the view that the real problem with the Sobral astrographic data was the difficulty . . . of separating the scale change from the light deflection. (Kennefick 2009, p. 42)

What was the result of this herculean effort to redo the data analysis from 60 years before?

> Ironically, however, the 1979 paper had no impact on the emerging story that something was fishy about the 1919 experiment . . . so far as I can tell, the paper has never been cited by anyone except for a brief, vague reference in Stephen Hawking's *A Brief History of Time* [which actually gets it wrong and was corrected]. (ibid.) [9]

The bottom line is, there was no failed replication; there was one set of eclipse data that was unusable.

5. **Substantively based hypotheses.** We know it's fallacious to take a statistically significant result as evidence in affirming a substantive theory, even if that theory predicts the significant result. A qualitative version of FEV, or, equivalently, an appeal to severity, underwrites this. Can failing to reject statistical null hypotheses ever inform about substantive claims? Yes. First consider how, in the midst of auditing, there's a concern to test a claim: is an apparently anomalous result real or spurious?

[9] Data from ESA's Gaia mission should enable light deflection to be measured with an accuracy of 2×10^{-6} (Mignard and Klioner 2009, p. 308).

Finding cancer clusters is sometimes compared to our Texas Marksman drawing a bull's-eye around the shots after they were fired into the barn. They often turn out to be spurious. Physical theory, let's suppose, suggests that because the quantum of energy in non-ionizing electromagnetic fields, such as those from high-voltage transmission lines, is much less than is required to break a molecular bond, there should be no carcinogenic effect from exposure to such fields. Yet a cancer association was reported in 1979 (Wertheimer and Leeper 1979). Was it real? In a randomized experiment where two groups of mice are under identical conditions except that one group is exposed to such a field, the null hypothesis that the cancer incidence rates in the two groups are identical may well be true. Testing this null is a way to ask: was the observed cancer cluster really an anomaly for the theory? Were the apparently anomalous results for the theory genuine, it is expected that H_0 would be rejected, so if it's not, it cuts against the reality of the anomaly. Cox gives this as one of the few contexts where a reported small *P*-value alone might suffice.

This wouldn't entirely settle the issue, and our knowledge of such things is always growing. Nor does it, in and of itself, show the flaw in any studies purporting to find an association. But several of these pieces taken together can discount the apparent effect with severity. It turns out that the initial researchers in the 1979 study did not actually measure magnetic fields from power lines; when they were measured no association was found. Instead they used the wiring code in a home as a proxy. All they really showed, it may be argued, was that people who live in the homes with poor wiring code tend to be poorer than the control (Gurney et al. 1996). The study was biased. Twenty years of study continued to find negative results (Kheifets et al. 1999). The point just now is not when to stop testing – more of a policy decision – or even whether to infer, as they did, that there's no evidence of a risk, and no theoretical explanation of how there could be. It is rather the role played by a negative statistical result, given the background information that, if the effects were real, these tests were highly capable of finding them. It amounts to a failed replication (of the observed cluster), but with a more controlled method. If a well-controlled experiment fails to replicate an apparent anomaly for an independently severely tested theory, it indicates the observed anomaly is spurious. The indicated severity and potential gaps are recorded; the case may well be reopened. Replication researchers might take note.

Another important category of tests that Cox develops, is what he calls testing *discrete families of models*, where there's no nesting. In a nutshell, each model is taken in turn to assess if the data are compatible with one, both, or neither of the possibilities (Cox 1977, p. 59). Each gets its own severity assessment.

Who Says You Can't Falsify Alternatives in a Significance Test?

Does the Cox–Mayo formulation of tests change the logic of significance tests in any way? I don't think so and neither does Cox. But it's different from some of the common readings. Nothing turns on whether you wish to view it as a revised account. SEV goes a bit further than FEV, and I do not saddle Cox with it. The important thing is how you get a nuanced interpretation, and we have barely begun our travels! Note the consequences for a familiar bugaboo about falsifying alternatives to significance tests. Burnham and Anderson (2014) make a nice link with Popper:

> While the exact definition of the so-called "scientific method" might be controversial, nearly everyone agrees that the concept of "falsifiability" is a central tenant [sic] of empirical science (Popper 1959). It is critical to understand that historical statistical approaches (i.e., *P* values) leave no way to "test" the alternative hypothesis. The alternative hypothesis is never tested, hence cannot be rejected or falsified! . . . Surely this fact alone makes the use of significance tests and *P* values bogus. Lacking a valid methodology to reject/falsify the alternative science hypotheses seems almost a scandal. (p. 629)

I think we *should* be scandalized. But not for the reasons alleged. Fisher emphasized that, faced with a non-significant result, a researcher's attitude wouldn't be full acceptance of H_0 but, depending on the context, more like the following:

> The possible deviation from truth of my working hypothesis, to examine which test is appropriate, seems not to be of sufficient magnitude to warrant any immediate modification.
>
> Or, . . . the body of data available so far is not by itself sufficient to demonstrate their [the deviations] reality. (Fisher 1955, p. 73)

Our treatment cashes out these claims, by either indicating the magnitudes ruled out statistically, or inferring that the observed difference is sufficiently common, even if spurious.

If you work through the logic, you'll see that in each case of the taxonomy the alternative may indeed be falsified. Perhaps the most difficult one is ruling out model violations, but this is also the one that requires a less severe test, at least with a reasonably robust method of inference. So what do those who repeat this charge have in mind? Maybe they mean: you cannot falsify an alternative, if you don't specify it. But specifying a directional or type of alternative is an outgrowth of specifying a test statistic. Thus we still have the implicit alternatives in Table 3.4, all of which are open to being falsified with severity. It's a key part of test specification to indicate which claims or features of a model are being tested. The charge might stand if a point null is known to be false, for in those cases we can't say μ is precisely 0, say. In that case you wouldn't want to infer it. One can still set upper bounds for how far off an adequate hypothesis can be. Moreover, there are many cases in science where a point null *is* inferred severely.

Nordtvedt Effect: Do the Earth and Moon Fall With the Same Acceleration?

We left off Section 3.1 with GTR going through a period of "stagnation" or "hibernation" after the early eclipse results. No one knew how to link it up with experiment. Discoveries around 1959–1960 sparked the "golden era" or "renaissance" of GTR, thanks to quantum mechanics, semiconductors, lasers, computers, and pulsars (Will 1986, p. 14). The stage was set for new confrontations between GTR's experiments; from 1960 to 1980, a veritable "zoo" of rivals to GTR was erected, all of which could be constrained to fit the existing test results.

Not only would there have been too many alternatives to report a pairwise comparison of GTR, the testing had to manage without having full-blown alternative theories of gravity. They could still ask, as they did in 1960: How could it be a mistake to regard the existing evidence as good evidence for GTR (or even for the deflection effect)?

They set out a scheme of parameters, the Parameterized Post Newtonian (PPN) framework, that allowed experimental relativists to describe violations to GTR's hypotheses – discrepancies with what it said about specific gravitational phenomena. One parameter is λ – the curvature of spacetime. An explicit goal was to prevent researchers from being biased toward accepting GTR prematurely (Will 1993, p. 10). These alternatives, by the physicist's own admission, were set up largely as straw men to either set firmer constraints on estimates of parameters, or, more interestingly, find violations. They could test 10 or 20 or 50 rivals without having to develop them! The work involved local statistical testing and estimation of parameters describing curved space.

Interestingly, these were non-novel hypotheses set up after the data were known. However rival theories had to be *viable*; they had to (1) account for experimental results already severely passed and (2) be able to show the relevance of the data for gravitational phenomena. They would have to be able to analyze and explore data about as well as GTR. They needed to permit stringent probing to learn more about gravity. (For an explicit list of requirements for a viable theory, see Will 1993, pp. 18–21.[10])

[10] While a viable theory can't just postulate the results ad hoc, "this does not preclude 'arbitrary parameters' being required for gravitational theories to accord with experimental results" (Mayo 2010a, p. 48).

All the viable members of the zoo of GTR rivals held the *equivalence principle (EP)*, roughly the claim that bodies of different composition fall with the same accelerations in a gravitational field. This principle was inferred with severity by passing a series of null hypotheses (examples include the Eötvös experiments) that assert a zero difference in the accelerations of two differently composed bodies. Because these null hypotheses passed with high precision, it was warranted to infer that: "gravity is a phenomenon of curved spacetime," that is, it must be described by a "metric theory of gravity" (ibid., p. 10). Those who deny we can falsify non-nulls take note: inferring that an adequate theory must be relativistic (even if not necessarily GTR) was based on inferring a point null with severity! What about the earth and moon, examples of self-gravitating bodies? Do they also fall at the same rate?

While long corroborated for solar system tests, the equivalence principle (later the weak equivalence principle, WEP) was untested for such massive self-gravitating bodies (which requires the *strong equivalence principle*). Kenneth Nordtvedt discovered in the 1960s that in one of the most promising GTR rivals, the Brans–Dicke theory, the moon and earth fell at different rates, whereas for GTR there would be no difference. Clifford Will, the experimental physicist I've been quoting, tells how in 1968 Nordtvedt finds himself on the same plane as Robert Dicke. "Escape for the beleaguered Dicke was unfeasible at this point. Here was a total stranger telling him that his theory violated the principle of equivalence!" (1986 pp. 139–40). To Dicke's credit, he helped Nordtvedt design the experiment. A new parameter to describe the Nordtvedt effect was added to the PPN framework, i.e., η. For GTR, $\eta = 0$, so the statistical or substantive null hypothesis tested is that $\eta = 0$ as against $\eta \neq 0$ for rivals.

How can they determine the rates at which the earth and moon are falling? Thank the space program. It turns out that measurements of the round trip travel times between the earth and moon (between 1969 and 1975) enable the existence of such an anomaly for GTR to be probed severely (and the measurements continue today). Because the tests were sufficiently sensitive, these measurements provided good evidence that the Nordtvedt effect is absent, set upper bounds to the possible violations, and provided evidence for the correctness of what GTR says with respect to this effect.

So the old saw that we cannot falsify $\eta \neq 0$ is just that, an old saw. Critics take Fisher's correct claim, that failure to reject a null isn't automatically evidence for its correctness, as claiming we never have such evidence. Even he says it lends some weight to the null (Fisher 1955). With the N-P test, the null and alternative needn't be treated asymmetrically. In testing H_0: $\mu \geq \mu_0$ vs. H_1: $\mu < \mu_0$, a rejection falsifies a claimed increase.[11] Nordtvedt's null result added weight to GTR, not in rendering it more probable, but in extending the domain for which GTR gives a satisfactory explanation. It's still provisional in the sense that gravitational phenomena in unexplored domains could introduce certain couplings that, strictly speaking, violate the strong equivalence principle. The error statistical standpoint describes the state of information at any one time, with indications of where theoretical uncertainties remain.

You might discover that critics of a significance test's falsifying ability are themselves in favor of methods that preclude falsification altogether! Burnham and Anderson raised the scandal, yet their own account provides only a comparative appraisal of fit in model selection. No falsification there.

Souvenir K: Probativism

> [A] fundamental tenet of the conception of inductive learning most at home with the frequentist philosophy is that inductive inference requires building up incisive arguments and inferences by putting together several different piece-meal results . . . the payoff is an account that approaches the kind of full-bodied arguments that scientists build up in order to obtain reliable knowledge and understanding of a field. (Mayo and Cox 2006, p. 82)

The error statistician begins with a substantive problem or question. She jumps in and out of piecemeal statistical tests both formal and quasi-formal. The pieces are integrated in building up arguments from coincidence, informing background theory, self-correcting via blatant deceptions, in an iterative movement. The inference is qualified by using error probabilities to determine not "how probable," but rather, "how well-probed" claims are, and what has been poorly probed. What's wanted are ways to measure how far off what a given theory says about a phenomenon can be from what a "correct" theory would need to say about it by setting bounds on the possible violations.

An account of testing or confirmation might entitle you to confirm, support, or rationally accept a large-scale theory such as GTR. One is free to reconstruct episodes this way – after the fact – but as a forward-looking account, they fall far short. Even if somehow magically it was known in 1960 that GTR was true, it wouldn't snap experimental relativists out of their doldrums because they still couldn't be said to have understood gravity, how it behaves, or how to use one severely affirmed piece to opportunistically probe entirely distinct areas.

[11] Some recommend "equivalence testing" where H_0: $\mu \geq \mu_0$ or $\mu \leq -\mu_0$ and rejecting both sets bounds on μ. One might worry about low-powered tests, but it isn't essentially different from setting upper bounds for a more usual null. (For discussion see Lakens 2017, Senn 2001a, 2014, R. Berger and Hsu 1996, R. Berger 2014, Wellek 2010).

Learning from evidence turns not on appraising or probabilifying large-scale theories but on piecemeal tasks of data analysis: estimating backgrounds, modeling data, and discriminating signals from noise. Statistical inference is not radically different from, but is illuminated by, sexy science, which increasingly depends on it. Fisherian and N-P tests become parts of a cluster of error statistical methods that arise in full-bodied science. In Tour II, I'll take you to see the (unwarranted) carnage that results from supposing they belong to radically different philosophies.

Tour II It's The Methods, Stupid

> There is perhaps in current literature a tendency to speak of the Neyman–Pearson contributions as some static system, rather than as part of the historical process of development of thought on statistical theory which is and will always go on. (Pearson 1962, p. 276)

This goes for Fisherian contributions as well. Unlike museums, we won't remain static.

The lesson from Tour I of this Excursion is that Fisherian and Neyman–Pearsonian tests may be seen as offering clusters of methods appropriate for different contexts within the large taxonomy of statistical inquiries. There is an overarching pattern:

> Just as with the use of measuring instruments, applied to the specific case, we employ the performance features to make inferences about aspects of the particular thing that is measured, aspects that the measuring tool is appropriately capable of revealing. (Mayo and Cox 2006, p. 84)

This information is used to ascertain what claims have, and have not, passed severely, post-data. Any such proposed inferential use of error probabilities gives considerable fodder for criticism from various tribes of Fisherians, Neyman–Pearsonians, and Bayesians. We can hear them now:

- N-P theorists can only report the preset error probabilities, and can't use *P*-values post-data.
- A Fisherian wouldn't dream of using something that skirts so close to power as does the "sensitivity function" $\Pi(\gamma)$.
- Your account cannot be evidential because it doesn't supply posterior probabilities to hypotheses.
- N-P and Fisherian methods preclude any kind of inference since they use "the sample space" (violating the LP).

How can we reply? To begin, we need to uncover how the charges originate in traditional philosophies long associated with error statistical tools. That's the focus of Tour II.

Only then do we have a shot at decoupling traditional philosophies from those tools in order to use them appropriately today. This is especially so when the traditional foundations stand on such wobbly grounds, grounds largely rejected by founders of the tools. There is a philosophical disagreement between Fisher and Neyman, but it differs importantly from the ones that you're presented with and which are widely accepted and repeated in scholarly and popular treatises on significance tests. Neo-Fisherians and N-P theorists, keeping to their tribes, forfeit notions that would improve their methods (e.g., for Fisherians: explicit alternatives, with corresponding notions of sensitivity, and distinguishing statistical and substantive hypotheses; for N-P theorists, making error probabilities relevant for inference in the case at hand).

The spadework on this tour will be almost entirely conceptual: we won't be arguing for or against any one view. We begin in Section 3.4 by unearthing the basis for some classic counterintuitive inferences thought to be licensed by either Fisherian or N-P tests. That many are humorous doesn't mean disentangling their puzzles is straightforward; a medium to heavy shovel is recommended. We can switch to a light to medium shovel in Section 3.5: excavations of the evidential versus behavioral divide between Fisher and N-P turn out to be mostly built on sand. As David Cox observes, Fisher is often more performance-oriented in practice, but not in theory, while the reverse is true for Neyman and Pearson. At times, Neyman exaggerates the behavioristic conception just to accentuate how much Fisher's tests need reining in. Likewise, Fisher can be spotted running away from his earlier behavioristic positions just to derogate the new N-P movement, whose popularity threatened to eclipse the statistics program that was, after all, his baby. Taking the polemics of Fisher and Neyman at face value, many are unaware how much they are based on personality and professional disputes. Hearing the actual voices of Fisher, Neyman, and Pearson (F and N-P), you don't have to accept the gospel of "what the founders really thought." Still, there's an entrenched history and philosophy of F and N-P: A thick-skinned jacket is recommended. On our third stop (Section 3.6) we witness a bit of magic. The very concept of an error probability gets redefined and, hey presto!, a reconciliation between Jeffreys, Fisher, and Neyman is forged. Wear easily removed shoes and take a stiff walking stick. The Unificationist tribes tend to live near underground springs and lakeshore bounds; in the heady magic, visitors have been known to accidentally fall into a pool of quicksand.

3.4 Some Howlers and Chestnuts of Statistical Tests

> The well-known definition of a statistician as someone whose aim in life is to be wrong in exactly 5 per cent of everything they do misses its target. (Sir David Cox 2006a, p. 197)

Showing that a method's stipulations could countenance absurd or counterintuitive results is a perfectly legitimate mode of criticism. I reserve the term "howler" for common criticisms based on logical fallacies or conceptual misunderstandings. Other cases are better seen as chestnuts – puzzles that the founders of statistical tests never cleared up explicitly. Whether you choose to see my "howler" as a "chestnut" is up to you. Under each exhibit is the purported basis for the joke.

Exhibit (iii): Armchair Science. *Did you hear the one about the statistical hypothesis tester* . . . who claimed that observing "heads" on a biased coin that lands heads with probability 0.05 is evidence of a statistically significant improvement over the standard treatment of diabetes, on the grounds that such an event occurs with low probability (0.05)?

The "armchair" enters because diabetes research is being conducted solely by flipping a coin. The joke is a spin-off from Kadane (2011):

> Flip a biased coin that comes up heads with probability 0.95, and tails with probability 0.05. If the coin comes up tails reject the null hypothesis. Since the probability of rejecting the null hypothesis if it is true is 0.05, this is a valid 5 percent level test. It is also very robust against data errors; indeed it does not depend on the data at all. It is also nonsense, of course, but nonsense allowed by the rules of significance testing. (p. 439)

Basis for the joke: Fisherian test requirements are (allegedly) satisfied by any method that rarely rejects the null hypothesis.

But are they satisfied? I say no. The null hypothesis in Kadane's example can be in any field, diabetes, or the mean deflection of light. (Yes, Kadane affirms this.) He knows the test entirely ignores the data, but avers that "it has the property that Fisher proposes" (Kadane 2016, p. 1). Here's my take: in significance tests and in scientific hypotheses testing more generally, data can disagree with *H* only by being counter to what would be expected under the assumption that *H* is correct. An improbable series of coin tosses or plane crashes does not count as a disagreement from hypotheses about diabetes or light deflection. In Kadane's example, there is accordance so long as a head occurs – but this is a nonsensical distance measure. Were someone to tell you that any old improbable event (three plane crashes in one week) tests a hypothesis about light deflection, you would say that person didn't understand the meaning of testing in science or in ordinary life. You'd be right (for some great examples, see David Hand 2014).

Kadane knows it's nonsense, but thinks the only complaint a significance tester can have is its low power. What's the power of this "test" against any alternative? It's just the same as the probability it rejects, period, namely, 0.05. So an N-P tester could at least complain. Now I agree that bad tests may still be tests; but I'm saying Kadane's is no test at all. If you want to insist Fisher permits this test, fine, but I don't think that's a very generous interpretation. As egregious as is this howler, it is instructive because it shows like nothing else the absurdity of a crass performance view that claims: reject the null and infer evidence of a genuine effect, so long as it is done rarely. Crass performance is bad enough, but this howler commits a further misdemeanor: It overlooks the fact that a test statistic d($\boldsymbol{x}$) must track discrepancies from H_0, becoming bigger (or smaller) as discrepancies increase (I list it as (ii) in Section 3.2). With any sensible distance measure, a misfit with H_0 must be *because* of the falsity of H_0. The probability of "heads" under a hypothesis about light deflection isn't even defined, because deflection hypotheses do not assign probabilities to coin-tossing trials. Fisher wanted test statistics to reduce the data from the generating mechanism, and here it's not even from the mechanism.

Kadane regards this example as "perhaps the most damaging critique" of significance tests (2016, p. 1). Well, Fisher can get around this easily enough.

Exhibit (iv): Limb-sawing Logic. *Did you hear the one about significance testers sawing off their own limbs?*

> As soon as they reject the null hypothesis H_0 based on a small *P*-value, they no longer can justify the rejection because the *P*-value was computed under the assumption that H_0 holds, and now it doesn't.

Basis for the joke: If a test assumes *H*, then as soon as *H* is rejected, the grounds for its rejection disappear!

This joke, and I swear it is widely repeated but I won't name names, reflects a serious misunderstanding about ordinary conditional claims. The assumption we use in testing a hypothesis *H*, statistical or other, is an *implicationary* or *i-assumption*. We have a conditional, say: If *H* then expect $\boldsymbol{x}$, with *H* the antecedent. The entailment from *H* to $\boldsymbol{x}$, whether it is statistical or deductive, does not get sawed off after the hypothesis or model *H* is rejected when the prediction is not borne out. A related criticism is that statistical tests assume the truth of their test or null hypotheses. No, once again, they may serve only as i-assumptions for drawing out implications. The howler occurs when a test hypothesis that serves merely as an i-assumption is purported to be an actual assumption, needed for the inference to go through. A little logic goes a long way toward exposing many of these howlers. As the point is general, we use *H*.

This next challenge is by Harold Jeffreys. I won't call it a howler because it hasn't, to my knowledge, been excised by testers: it's an old chestnut, and a very revealing one.

Exhibit (v): Jeffreys' Tail Area Criticism. *Did you hear the one about statistical hypothesis testers rejecting H_0 because of outcomes it failed to predict?*

> What's unusual about that?
>
> What's unusual is that they do so even when these unpredicted outcomes haven't occurred!

Actually, one can't improve upon the clever statement given by Jeffreys himself. Using *P*-values, he avers, implies that "*a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred*" (1939/1961 p. 385).

Basis for the joke: The *P*-value, $\Pr(d \geq d_0; H_0)$, uses the "tail area" of the curve under H_0. d_0 is the observed difference, but $\{d \geq d_0\}$ includes differences even further from H_0 than d_0.

This has become the number one joke in comical statistical repertoires. Before debunking it, let me say that Jeffreys shows a lot of admiration for Fisher: "I have in fact been struck repeatedly in my own work . . . to find that Fisher had already grasped the essentials by some brilliant piece of common sense, and that his results would either be identical with mine or would differ only in cases where we should both be very doubtful" (ibid., p. 393). The famous quip is funny because it seems true, yet paradoxical. Why consider more extreme outcomes that didn't occur? The non-occurrence of more deviant results, Jeffreys goes on to say, "might more reasonably be taken as evidence for the law [in this case, H_0], not against it" (ibid., p. 385). The implication is that considering outcomes beyond d_0 is to unfairly discredit H_0, in the sense of finding more evidence against it than if only the actual outcome d_0 is considered.

The opposite is true.

Considering the tail area makes it harder, not easier, to find an outcome statistically significant (although this isn't the only function of the tail area). Why? Because it requires not merely that $\Pr(d = d_0; H_0)$ be small, but that $\Pr(d \geq d_0; H_0)$ be small. This alone squashes the only sense in which this could be taken as a serious criticism of tests. Still, there's a legitimate question about why the tail area probability is relevant. Jeffreys himself goes on to give it a rationale: "If mere improbability of the observations, given the hypothesis, was the criterion, any hypothesis whatever would be rejected. Everybody rejects the conclusion" (ibid., p. 385), so some other criterion is needed. Looking at the tail area supplies one, another would be a prior, which is Jeffreys' preference.

It's worth reiterating Jeffreys' correctly pointing out that "everybody rejects" the idea that the improbability of data under *H* suffices for evidence against *H*.

Shall we choose priors or tail areas? Jeffreys chooses default priors. Interestingly, as Jeffreys recognizes, for Normal distributions "the tail area represents the probability, given the data" that the actual discrepancy is in the direction opposite to that observed – d_0 is the wrong "sign" (ibid., p. 387). (This relies on a uniform prior probability for the parameter.) This connection between *P*-values and posterior probabilities is often taken as a way to "reconcile" them, at least for one-sided tests (Sections 4.4, 4.5). This was not one of Fisher's given rationales.

Note that the joke talks about outcomes the null does not predict – just what we wouldn't know without an assumed test statistic or alternative. One reason to evoke the tail area in Fisherian tests is to determine what H_0 "has not predicted," that is, to identify a sensible test statistic d($\boldsymbol{x}$). Fisher, strictly speaking, has only the null distribution, with an implicit interest in tests with sensitivity of a given type. Fisher discusses this point in relation to the lady tasting tea (1935a, pp. 14–15). Suppose I take an observed difference d_0 as grounds to reject H_0 on account of it's being improbable under H_0, when in fact larger differences (larger d values) are even more probable under H_0. Then, as Fisher rightly notes, the improbability of the observed difference would be a poor indication of underlying discrepancy. (In N-P terms, it would be a biased test.) Looking at the tail area would reveal this fallacy; whereas it would be readily committed, Fisher notes, in accounts that only look at the improbability of the observed outcome d_0 under H_0.

When E. Pearson (1970) takes up Jeffreys' question: "Why did we use tail-area probabilities . . .?", his reply is that "this interpretation was not part of our approach" (p. 464). Tail areas simply fall out of the N-P desiderata of good tests. Given the lambda criterion one needed to decide at what point H_0

> should be regarded as no longer tenable, that is where should one choose to bound the rejection region? To help in reaching this decision it appeared that the probability of falling into the region chosen, if H_0 were true, was one necessary piece of information. (ibid.)

So looking at the tail area could be seen as the result of formulating a sensible distance measure (for Fisher), or erecting a good critical region (for Neyman and Pearson).

Pearson's reply doesn't go far enough; it does not by itself explain why reporting the probability of falling into the rejection region is relevant for *inference*. It points to a purely performance-oriented justification that I know Pearson shied away from: It ensures data fall in a critical region rarely under H_0 and sufficiently often under alternatives in H_1 – but this tends to be left as a pre-data, performance goal (recall Birnbaum's Conf, Souvenir D). It is often alleged the N-P tester only reports whether or not $\boldsymbol{x}$ falls in the rejection region. Why are N-P collapsing all outcomes in this region?

In my reading, the error statistician does not collapse the result beyond what the minimal sufficient statistic requires for the question at hand. From our Translation Guide, Souvenir C, considering $(\mathrm{d}(\boldsymbol{X}) \geq \mathrm{d}(\boldsymbol{x}_0))$ signals that we're interested in the method, and we insert "the test procedure would have yielded" before $\mathrm{d}(\boldsymbol{X})$. We report what was observed $\boldsymbol{x}_0$ and the corresponding $\mathrm{d}(\boldsymbol{x}_0)$ – or d_0 – but we require the methodological probability, via the sampling distribution of $\mathrm{d}(\boldsymbol{X})$ – abbreviated as *d*. This could mean looking at other stopping points, other endpoints, and other variables. We require that with high probability our test would have warned us if the result could easily have come about in a universe where the test hypothesis is true, that is $\Pr(\mathrm{d}(\boldsymbol{X}) < \mathrm{d}(\boldsymbol{x}_0); H_0)$ is high. Besides, we couldn't throw away the detailed data, since they're needed to audit model assumptions.

To conclude this exhibit, considering the tail area does not make it easier to reject H_0 but harder. Harder because it's not enough that the outcome be improbable under the null, outcomes even greater must be improbable under the null. $\Pr(d(\boldsymbol{X}) = d(\boldsymbol{x}_0); H_0)$ could be small while $\Pr(d(\boldsymbol{X}) \geq d(\boldsymbol{x}_0); H_0)$ not small. This leads to blocking a rejection when it should be because it means the test could readily produce even larger differences under H_0. Considering other possible outcomes that could have arisen is essential for assessing the test's capabilities. To understand the properties of our inferential tool is to understand what it would do under different outcomes, under different conjectures about what's producing the data. (Yes, the sample space matters post-data.) I admit that neither Fisher nor N-P adequately pinned down an inferential justification for tail areas, but now we have.

A bit of foreshadowing of a later shore excursion: some argue that looking at $d(\boldsymbol{X}) \geq d(\boldsymbol{x}_0)$ actually *does* make it easier to find evidence against H_0. How can that be? Treating $(1 - \beta)/\alpha$ as a kind of likelihood ratio in favor of an alternative over the null, then fed into a Likelihoodist or Bayesian algorithm, it can appear that way. Stay tuned.

Exhibit (vi): Two Measuring Instruments of Different Precisions. *Did you hear about the frequentist who, knowing she used a scale that's right only half the time, claimed her method of weighing is right 75% of the time*?

> She says, "I flipped a coin to decide whether to use a scale that's right 100% of the time, or one that's right only half the time, so, overall, I'm right 75% of the time." (She wants credit because she could have used a better scale, even knowing she used a lousy one.)

Basis for the joke: An N-P test bases error probabilities on all possible outcomes or measurements that could have occurred in repetitions, but did not.

As with many infamous pathological examples, often presented as knockdown criticisms of all of frequentist statistics, this was invented by a frequentist, Cox (1958). It was a way to highlight what could go wrong in the case at hand, if one embraced an unthinking behavioral-performance view. Yes, error probabilities are taken over hypothetical repetitions of a process, but not just any repetitions will do. Here's the statistical formulation.

We flip a fair coin to decide which of two instruments, E_1 or E_2, to use in observing a Normally distributed random sample $\boldsymbol{Z}$ to make inferences about mean θ. E_1 has variance of 1, while that of E_2 is 10^6. Any randomizing device used to choose which instrument to use will do, so long as it is irrelevant to θ. This is called a *mixture* experiment. The full data would report both the result of the coin flip and the measurement made with that instrument. We can write the report as having two parts: First, which experiment was run and second the measurement: $(E_i, \boldsymbol{z})$, $i = 1$ or 2.

In testing a null hypothesis such as $\theta = 0$, the same $\boldsymbol{z}$ measurement would correspond to a much smaller *P*-value were it to have come from E_1 rather than from E_2: denote them as $p_1(\boldsymbol{z})$ and $p_2(\boldsymbol{z})$, respectively. The overall significance level of the mixture: $[p_1(\boldsymbol{z}) + p_2(\boldsymbol{z})]/2$, would give a misleading report of the precision of the actual experimental measurement. The claim is that N-P statistics would report the average *P*-value rather than the one corresponding to the scale you actually used! These are often called the unconditional and the conditional test, respectively. The claim is that the frequentist statistician must use the unconditional test.

Suppose that we know we have observed a measurement from E_2 with its much larger variance:

> The unconditional test says that we can assign this a higher level of significance than we ordinarily do, because if we were to repeat the experiment, we might sample some quite different distribution. But this fact seems irrelevant to the interpretation of an observation which we know came from a distribution [with the larger variance]. (Cox 1958, p. 361)

Once it is known which E_i has produced $\boldsymbol{z}$, the *P*-value or other inferential assessment should be made with reference to the experiment actually run. As we say in Cox and Mayo (2010):

> The point essentially is that the marginal distribution of a *P*-value averaged over the two possible configurations is misleading for a particular set of data. It would mean that an individual fortunate in obtaining the use of a precise instrument in effect sacrifices some of that information in order to rescue an investigator who has been unfortunate enough to have the randomizer choose a far less precise tool. From the perspective of interpreting the specific data that are actually available, this makes no sense. (p. 296)

To scotch his famous example, Cox (1958) introduces a principle: weak conditionality.

Weak Conditionality Principle (WCP)**:** If a mixture experiment (of the aforementioned type) is performed, then, if it is known which experiment produced the data, inferences about θ *are appropriately drawn in terms of the sampling behavior* in the experiment known to have been performed (Cox and Mayo 2010, p. 296).

It is called weak conditionality because there are more general principles of conditioning that go beyond the special case of mixtures of measuring instruments.

While conditioning on the instrument actually used seems obviously correct, nothing precludes the N-P theory from choosing the procedure "which is best on the average over both experiments" (Lehmann and Romano 2005, p. 394), and it's even possible that the average or unconditional power is better than the conditional. In the case of such a conflict, Lehmann says relevant conditioning takes precedence over average power (1993b). He allows that in some cases of acceptance sampling, the average behavior may be relevant, but in scientific contexts the conditional result would be the appropriate one (see Lehmann 1993b, p. 1246). Context matters. Did Neyman and Pearson ever weigh in on this? Not to my knowledge, but I'm sure they'd concur with N-P tribe leader Lehmann. Admittedly, if your goal in life is to attain a precise α level, then when discrete distributions preclude this, a solution would be to flip a coin to decide the borderline cases! (See also Example 4.6, Cox and Hinkley 1974, pp. 95–6; Birnbaum 1962 p. 491.)

Is There a Catch?

The "two measuring instruments" example occupies a famous spot in the pantheon of statistical foundations, regarded by some as causing "a subtle earthquake" in statistical foundations. Analogous examples are made out in terms of confidence interval estimation methods (Tour III, Exhibit (viii)). It is a warning to the most behavioristic accounts of testing from which we have already distinguished the present approach. Yet justification for the conditioning (WCP) is fully within the frequentist error statistical philosophy, for contexts of scientific inference. There is no suggestion, for example, that only the particular data set be considered. That would entail abandoning the sampling distribution as the basis for inference, and with it the severity goal. Yet we are told that "there is a catch" and that WCP leads to the Likelihood Principle (LP)!

It is not uncommon to see statistics texts argue that in frequentist theory one is faced with the following dilemma: either to deny the appropriateness of conditioning on the precision of the tool chosen by the toss of a coin, or else to embrace the strong likelihood principle, which entails that frequentist sampling distributions are irrelevant to inference once the data are obtained. This is a false dilemma. Conditioning is warranted to achieve objective frequentist goals, and the [weak] conditionality principle coupled with sufficiency does not entail the strong likelihood principle. The 'dilemma' argument is therefore an illusion. (Cox and Mayo 2010, p. 298)

There is a large literature surrounding the argument for the Likelihood Principle, made famous by Birnbaum (1962). Birnbaum hankered for something in between radical behaviorism and throwing error probabilities out the window. Yet he himself had apparently proved there is no middle ground (if you accept WCP)! Even people who thought there was something fishy about Birnbaum's "proof" were discomfited by the lack of resolution to the paradox. It is time for post-LP philosophies of inference. So long as the Birnbaum argument, which Savage and many others deemed important enough to dub a "breakthrough in statistics," went unanswered, the frequentist was thought to be boxed into the pathological examples. She is not.

In fact, I show there is a flaw in his venerable argument (Mayo 2010b, 2013a, 2014b). That's a relief. Now some of you will howl, "Mayo, not everyone agrees with your disproof! Some say the issue is not settled." Fine, please explain where my refutation breaks down. It's an ideal brainbuster to work on along the promenade after a long day's tour. Don't be dismayed by the fact that it has been accepted for so long. But I won't revisit it here.

3.5 P-values Aren't Error Probabilities Because Fisher Rejected Neyman's Performance Philosophy

> Both Neyman–Pearson and Fisher would give at most lukewarm support to standard significance levels such as 5% or 1%. Fisher, although originally recommending the use of such levels, later strongly attacked any standard choice. (Lehmann 1993b, p. 1248)
>
> Thus, Fisher rather incongruously appears to be attacking his own past position rather than that of Neyman and Pearson. (Lehmann 2011, p. 55)

By and large, when critics allege that Fisherian *P*-values are not error probabilities, what they mean is that Fisher wanted to interpret them in an evidential manner, not along the lines of Neyman's long-run behavior. I'm not denying there is an important difference between using error probabilities inferentially and behavioristically. The truth is that N-P and Fisher used

P-values and other error probabilities in both ways.[1] What they didn't give us is a clear account of the former. A big problem with figuring out the "he said/they said" between Fisher and Neyman–Pearson is that "after 1935 so much of it was polemics" (Kempthorne 1976) reflecting a blow-up which had to do with professional rivalry rather than underlying philosophy. Juicy details later on.

We need to be clear on the meaning of an error probability. A method of statistical inference moves from data to some inference about the source of the data as modeled. Associated error probabilities refer to the probability the method outputs an erroneous interpretation of the data. Choice of test rule pins down the particular error; for example, it licenses inferring there's a genuine discrepancy when there isn't (perhaps of a given magnitude). The test method is given in terms of a test statistic $\mathrm{d}(\boldsymbol{X})$, so the error probabilities refer to the probability distribution of $\mathrm{d}(\boldsymbol{X})$, the sampling distribution, computed under an appropriate hypothesis. Since we need to highlight subtle changes in meaning, call these ordinary "frequentist" error probabilities. (I can't very well call them error statistical error probabilities, but that's what I mean.)[2] We'll shortly require subscripts, so let this be error probability$_1$. Formal error probabilities have almost universally been associated with N-P statistics, and those with long-run performance goals. I have been disabusing you of such a straightjacketed view; they are vital in assessing how well probed the claim in front of me is. Yet my reinterpretation of error probabilities does not change their mathematical nature.

We can attach a frequentist performance assessment to any inference method. Post-data, these same error probabilities can, though they need not, serve to quantify the severity associated with an inference. Looking at the mathematics, it's easy to see the *P*-value as an error probability. Take Cox and Hinkley (1974):

> For given observations **y** we calculate $t = t_{\mathrm{obs}} = t(\mathbf{y})$, say, and the *level of significance* p_{obs} by $p_{\mathrm{obs}} = \Pr(T \geq t_{\mathrm{obs}}; H_0)$.
>
> . . . Hence p_{obs} is the probability that we would mistakenly declare there to be evidence against H_0, were we to regard the data under analysis as just decisive against H_0. (p. 66)

Thus p_{obs} would be the Type I error probability associated with the test procedure consisting of finding evidence against H_0 when reaching p_{obs}.[3]

[1] Neyman (1976) said he was "not aware of a conceptual difference between a 'test of a statistical hypothesis' and a 'test of significance' and uses these terms interchangeably" (p. 737). We will too, with qualifications as needed.

[2] Thanks to the interpretation being fairly intimately related to the test, we get the error probabilities (formal or informal) attached to the interpretation.

[3] Note that p_{obs} and t_{obs} are the same as our p_0 and d_0. (or $\mathrm{d}(\boldsymbol{x}_0)$)

Thus the *P*-value equals the corresponding Type I error probability. [I've been using upper case *P*, but it's impossible to unify the literature.] Listen to Lehmann, speaking for the N-P camp:

> [I]t is good practice to determine not only whether the hypothesis is accepted or rejected at the given significance level, but also to determine the smallest significance level ... at which the hypothesis would be rejected for the given observation. This number, the so-called *P-value* gives an idea of how strongly the data contradict the hypothesis. It also enables others to reach a verdict based on the significance level of their choice. (Lehmann and Romano 2005, pp. 63–4)

N-P theorists have no compunctions in talking about N-P tests using attained significance levels or *P*-values. Bayesians Gibbons and Pratt (1975) echo this view:

> The *P*-value can then be interpreted as the smallest level of significance, that is, the 'borderline level', since the outcome observed would ... not [be] significant at any smaller levels. Thus it is sometimes called the 'level attained' by the sample ... Reporting a *P*-value ... permits each individual to choose his own ... maximum tolerable probability for a Type I error. (p. 21)

Is all this just a sign of texts embodying an inconsistent hybrid? I say no, and you should too.

A certain tribe of statisticians professes to be horrified by the remarks of Cox and Hinkley, Lehmann and Romano, Gibbons and Pratt and many others. That these remarks come from leading statisticians, members of this tribe aver, just shows the depth of a dangerous "confusion over the evidential content (and mixing) of *p*'s and *α*'s" (Hubbard and Bayarri 2003, p. 175). On their view, we mustn't mix what they call "evidence and error": F and N-P are incompatible. For the rest of this tour, we'll alternate between the museum and engaging the Incompatibilist tribes themselves. When viewed through the tunnel of the Incompatibilist statistical philosophy, these statistical founders appear confused.

> The distinction between evidence (*p*'s) and error (*α*'s) is not trivial ... it reflects the fundamental differences between Fisher's ideas on significance testing and inductive inference, and [N-P's] views on hypothesis testing and inductive behavior. (Hubbard and Bayarri 2003, p. 171)

What's fascinating is that the Incompatibilists admit it's the philosophical difference they're on about, not a mathematical one. The paper that has become *the* centerpiece for the position in this subsection is Berger and Sellke (1987). They ask:

> Can *P* values be justified on the basis of how they perform in repeated use? We doubt it. For one thing, how would one measure the performance of *P* values? With

> significance tests and confidence intervals, they are either right or wrong, so it is possible to talk about error rates. If one introduces a decision rule into the situation by saying that H_0 is rejected when the P value < 0.05, then of course the classical error rate is 0.05. (p. 136)

Good. Then we can agree a *P*-value is, mathematically, an error probability. Berger and Sellke are merely opining that Fisher wouldn't have *justified* their use on grounds of error rate performance. That's different. Besides, are we so sure Fisher wouldn't sully himself with crass error probabilities, and dichotomous tests? Early on at least, Fisher appears as a behaviorist par excellence. That he is later found "attacking his own position," as Lehmann puts it, is something else.

Mirror Mirror on the Wall, Who's the More Behavioral of Them All?

N-P were striving to emulate the dichotomous interpretation they found in Fisher:

> It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result. It is obvious that an experiment would be useless of which no possible result would satisfy him. . . . It is usual and convenient for the experimenters to take 5 per cent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results. (Fisher 1935a, pp. 13–14)

Fisher's remark can be taken to justify the tendency to ignore negative results or stuff them in file drawers, somewhat at odds with his next lines, the ones that I specifically championed in Excursion 1: "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. . ." (1935a, p. 14).[4] This would require us to keep the negative results around for a while. How else could we see if we are rarely failing, or often succeeding?

What I mainly want to call your attention to now are the key phrases "willing to admit," "satisfy him," "deciding to ignore." What are these, Neyman asks, but actions or behaviors? He'd learned from R. A. Fisher! So, while many take

[4] Fisher, in a 1926 paper, gives another nice rendering: "A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance. The very high odds sometimes claimed for experimental results should usually be discounted, for inaccurate methods of estimating error have far more influence than has the particular standard of significance chosen" (pp. 504–5).

the dichotomous “up-down” spirit of tests as foreign to Fisher, it is not foreign at all. Again from Fisher (1935a):

> Our examination of the possible results of the experiment has therefore led us to a statistical test of significance, by which these results are divided into two classes with opposed interpretations ... those which show a significant discrepancy from a certain hypothesis; ... and on the other hand, results which show no significant discrepancy from this hypothesis. (pp. 15–16)

No wonder Neyman could counter Fisher’s accusations that he’d turned his tests into tools for inductive behavior by saying, in effect, look in the mirror (for instance, in the acrimonious exchange of 1955–6, 20 years after the blow-up): Pearson and I were only systematizing your practices for how to interpret data, taking explicit care to prevent untoward results that you only managed to avoid on intuitive grounds!

Fixing Significance Levels. What about the claim that N-P tests fix the Type I error probability in advance, whereas *P*-values are post-data? Doesn’t *that* prevent a *P*-value from being an error probability? First, we must distinguish between fixing the significance level for a test prior to data collection, and fixing a threshold to be used across one’s testing career. Fixing α and power is part of specifying a test with reasonable capabilities of answering the question of interest. Having done so, there’s nothing illicit about reporting the *achieved* or *attained* significance level, and it is even recommended by Lehmann. As for setting a threshold for habitual practice, that’s actually more Fisher than N-P.

Lehmann is flummoxed by the association of fixed levels of significance with N-P since “[U]nlike Fisher, Neyman and Pearson (1933, p. 296) did not recommend a standard level but suggested that ‘how the balance [between the two kinds of error] should be struck must be left to the investigator’” (Lehmann 1993b, p. 1244). From their earliest papers, Neyman and Pearson stressed that the tests were to be “used with discretion and understanding” depending on the context (Neyman and Pearson 1928, p. 58). In a famous passage, Fisher (1956) raises the criticism – but without naming names:

> A man who ‘rejects’ a hypothesis provisionally, as a matter of habitual practice, when the significance is at the 1% level or higher, will certainly be mistaken in not more than 1% of such decisions. For when the hypothesis is correct he will be mistaken in just 1% of these cases, and when it is incorrect he will never be mistaken in rejection ... However, the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. (pp. 44–5)

It is assumed Fisher is speaking of N-P, or at least Neyman. But N-P do not recommend such habitual practice.

Long Runs Are Hypothetical. What about the allegation that N-P error probabilities allude to actual long-run repetitions, while the *P*-value is a *hypothetical* distribution? N-P error probabilities are also about hypothetical would-be's. Each sample of size *n* gives a single value of the test statistic d(***X***). Our inference is based on this one sample. The third requirement (Pearson's "Step 3") for tests is that we be able to compute the distribution of d(***X***), under the assumption that the world is approximately like H_0, and under discrepancies from H_0. Different outcomes would yield different d(***X***) values, and we consider the frequency distribution of d(***X***) over hypothetical repetitions.

At the risk of overkill, the sampling distribution is all about hypotheticals: the relative frequency of outcomes under one or another hypothesis. These also equal the relative frequencies assuming you really did keep taking samples in a long run, tiring yourself out in the process. It doesn't follow that the value of the hypothetical frequencies depends on referring to, much less actually carrying out, that long run. A statistical hypothesis has implications for some hypothetical long run in terms of how frequently this or that would occur. A statistical test uses the data to check how well the predictions are met. The sampling distribution is the testable meeting-ground between the two.

The same pattern of reasoning is behind resampling from the one and only sample in order to generate a sampling distribution. (We meet with resampling in Section 4.10.) The only gap is to say why such a hypothetical (or counterfactual) is relevant for inference in the case at hand. Merely proposing that error probabilities give a vague "strength of evidence" to an inference won't do. Our answer is that they capture the capacities of tests, which in turn tell us how severely tested various claims may be said to be.

It's Time to Get Beyond the "Inconsistent Hybrid" Charge

Gerd Gigerenzer is a wonderful source of how Fisherian and N-P methods led to a statistical revolution in psychology. He is famous for, among much else, arguing that the neat and tidy accounts of statistical testing in social science texts are really an inconsistent hybrid of elements from N-P's behavioristic philosophy and Fisher's more evidential approach (Gigerenzer 2002, p. 279). His tribe is an offshoot of the Incompatibilists, but with a Freudian analogy to illuminate the resulting tension and anxiety that a researcher is seen to face.

N-P testing, he says, "functions as the Superego of the hybrid logic" (ibid., p. 280). It requires alternatives, significance levels, and power to be prespecified, while strictly outlawing evidential or inferential interpretations about the truth of a particular hypothesis. The Fisherian "Ego gets things done . . . and gets papers published" (ibid.). Power is ignored, and the level of significance is found after the experiment, cleverly hidden by rounding up to the nearest standard level. "The Ego avoids . . . exact predictions of the alternative hypothesis, but claims support for it by rejecting a null hypothesis" and in the end is "left with feelings of guilt and shame for having violated the rules" (ibid.). Somewhere in the background lurks his Bayesian Id, driven by wishful thinking into misinterpreting error probabilities as degrees of belief.

As with most good caricatures, there is a large grain of truth in Gigerenzer's Freudian metaphor – at least as the received view of these methods. I say it's time to retire the "inconsistent hybrid" allegation. Reporting the attained significance level is entirely legitimate and is recommended in N-P tests, so long as one is not guilty of other post-data selections causing *actual P*-values to differ from *reported* or nominal ones. By failing to explore the inferential basis for the stipulations, there's enormous unclarity as to what's being disallowed and why, and what's mere ritual or compulsive hand washing (as he might put it (ibid., p. 283)). Gigerenzer's Ego might well *deserve* to feel guilty if he has chosen the hypothesis, or characteristic to be tested, based on the data, or if he claims support for a research hypothesis by merely rejecting a null hypothesis – the illicit NHST animal. A post-data choice of test statistic may be problematic, but not an attained significance level.

Gigerenzer recommends that statistics texts teach the conflict and stop trying "to solve the conflict between its parents by denying its parents" (2002, p. 281). I, on the other hand, think we should take responsibility for interpreting the tools according to their capabilities. Polemics between Neyman and Fisher, however lively, taken at face value, are a highly unreliable source; we should avoid chiseling into even deeper stone the hackneyed assignments of statistical philosophy – "he's inferential, he's an acceptance sampler." The consequences of the "inconsistent hybrid" allegation are dire: both schools are caricatures, robbed of features that belong in an adequate account.

Hubbard and Bayarri (2003) are a good example of this; they proclaim an N-P tester is forbidden – forbidden! – from reporting the observed *P*-value. They eventually concede that an N-P test "could be defined equivalently in terms of the p value . . . the null hypothesis should be rejected if the observed $p < \alpha$, and accepted otherwise" (p. 175). But they aver "no matter how small the p value is, the appropriate report is that the procedure guarantees a $100\alpha\%$ false rejection of the null on repeated use" (ibid.). An N-P tester must robotically obey the reading that has grown out of the Incompatibilist tribe to which they belong. A user must round up to the predesignated α. This type of prohibition gives a valid guilt trip to Gigerenzer's Ego; yet the hang-up stems from the Freudian metaphor, not from Neyman and Pearson, who say:

> it is doubtful whether the knowledge that P_z [the *P*-value associated with test statistic *z*] was really 0.03 (or 0.06) rather than 0.05, . . . would in fact ever modify our judgment . . . regarding the origin of a single sample. (Neyman and Pearson 1928, p. 27)

But isn't it true that rejection frequencies needn't be indicative of the evidence against a null? Yes. Kadane's example, if allowed, shows how to get a small rejection frequency with no evidence. But this was to be a problem for Fisher, solved by N-P (even if Kadane is not fond of them either). Granted, even in tests not so easily dismissed, crude rejection frequencies differ from an evidential assessment, especially when some of the outcomes leading to rejection vary considerably in their evidential force. This is the lesson of Cox's famous "two machines with different precisions." Some put this in terms of selecting the relevant reference set which "need not correspond to all possible repetitions of the experiment" (Kalbfleisch and Sprott 1976, p. 272). We've already seen that relevant conditioning is open to a N-P tester. Others prefer to see it as a matter of adequate model specification. So once again it's not a matter of Fisher vs. N-P.

I'm prepared to admit Neyman's behavioristic talk. Mayo (1996, Chapter 11) discusses: "Why Pearson rejected the (behavioristic) N-P theory" (p. 361). Pearson does famously declare that "the behavioristic conception is Neyman's not mine" (1955, p. 207). Furthermore, Pearson explicitly addresses "the situation where statistical tools are applied to an isolated investigation of considerable importance . . ." (1947, p. 170).

> In other and, no doubt, more numerous cases there is no repetition of the same type of trial or experiment, but all the same we can and many of us do use the same test rules . . . Why do we do this? . . . Is it because the formulation of the case in terms of hypothetical repetition helps to that clarity of view needed for sound judgment?
>
> Or is it because we are content that the application of a rule, now in this investigation, now in that, should result in a long-run frequency of errors in judgment which we control at a low figure? (ibid., p. 172)

While tantalizingly leaving the answer dangling, it's clear that for Pearson: "the formulation of the case in terms of hypothetical repetition helps to that clarity of view needed for sound judgment" (ibid.) in learning about the particular case at hand. He gives an example from his statistical work in World War II:

> Two types of heavy armour-piercing naval shell of the same caliber are under consideration; they may be of different design or made by different firms . . . Twelve

shells of one kind and eight of the other have been fired; two of the former and five of the latter failed to perforate the plate . . . (Pearson 1947, 171)

> Starting from the basis that individual shells will never be identical in armour-piercing qualities, . . . he has to consider how much of the difference between (i) two failures out of twelve and (ii) five failures out of eight is likely to be due to this inevitable variability. (ibid.)

He considers what other outcomes could have occurred, and how readily, in order to learn what variability alone is capable of producing.[5] Pearson opened the door to the evidential interpretation, as I note in 1996, and now I go further.

Having looked more carefully at the history before the famous diatribes, and especially at Neyman's applied work, I now hold that Neyman largely rejected it as well! Most of the time, anyhow. But that's not the main thing. Even if we couldn't point to quotes and applications that break out of the strict "evidential versus behavioral" split: *we* should be the ones to interpret the methods for inference, and supply the statistical philosophy that directs their right use.

Souvenir L: Beyond Incompatibilist Tunnels

What people take away from the historical debates is Fisher (1955) accusing N-P, or mostly Neyman, of converting his tests into acceptance sampling rules more appropriate for five-year plans in Russia, or making money in the USA, than for science. Still, it couldn't have been too obvious that N-P distorted his tests, since Fisher tells us only in 1955 that it was Barnard who explained that, despite agreeing mathematically in very large part, there is this distinct philosophical position. Neyman suggests that his terminology was to distinguish what he (and Fisher!) were doing from the attempts to define a unified rational measure of belief on hypotheses. N-P both denied there was such a thing. Given Fisher's vehement disavowal of subjective Bayesian probability, N-P thought nothing of crediting Fisherian tests as a step in the development of "inductive behavior" (in their 1933 paper).

The myth of the radical difference in either methods or philosophy is a myth. Yet, as we'll see, the hold it has over people continues to influence the use and discussion of tests. It's based almost entirely on sniping between Fisher and Neyman from 1935 until Neyman leaves for the USA in 1938. Fisher didn't engage much with statistical developments during World War II. Barnard describes Fisher as cut off "by some mysterious personal or political agency. Fisher's isolation occurred, I think, at a particularly critical

[5] Pearson said that a statistician has an α and a β side, the former alludes to what they say in theory, the latter to what they do in practice. In practice, even Neyman, so often portrayed as performance-oriented, was as inferential as Pearson.

time, when opportunities existed for a fruitful fusion of ideas stemming from Neyman and Pearson and from Fisher" (Barnard 1985, p. 2). Lehmann observes that Fisher kept to his resolve not to engage in controversy with Neyman until the highly polemical exchange of 1955 at age 65. Fisher alters some of the lines of earlier editions of his books. For instance, Fisher's disinterest in the attained *P*-value was made clear in *Statistical Methods for Research Workers* (SMRW) (1934a, p. 80):

> . . . in practice we do not want to know the exact value of P for any observed value of [the test statistic], but, in the first place, whether or not the observed value is open to suspicion.
>
> If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05.

Lehmann explains that it was only "fairly late in life, Fisher's attitude had changed" (Lehmann 2011, p. 52). In the 13th edition of SMRW, Fisher changed his last sentence to:

> The actual value of P obtainable . . . indicates the strength of the evidence against the hypothesis. [Such a value] is seldom to be disregarded. (p. 80)

Even so, this at most suggests how the methodological (error) probability is thought to provide a measure of evidential strength – it doesn't abandon error probabilities. There's a deeper reason for this backtracking by Fisher; I'll save it for Excursion 5. One other thing to note: F and N-P were creatures of their time. Their verbiage reflects the concern with "operationalism" and "behaviorism," growing out of positivistic and verificationist philosophy. I don't deny the value of tracing out the thrust and parry between Fisher and Neyman in these excursions. None of the founders solved the problem of an inferential interpretation of error probabilities – though they each offered tidbits. Their name-calling: "you're too mechanical," "no *you* are," at most shows, as Gigerenzer and Marewski observe, that they all rejected mechanical statistics (2015, p. 422).

The danger is when one group's interpretation is the basis for a historically and philosophically "sanctioned" reinterpretation of one or another method. Suddenly, rigid rules that the founders never endorsed are imposed. Through the Incompatibilist philosophical tunnel, as we are about to see, these reconstruals may serve as an effective way to dismiss the entire methodology – both F and N-P. After completing this journey, you shouldn't have to retrace this "he said/they said" dispute again. It's the methods, stupid.

3.6 Hocus-Pocus: *P*-values Are Not Error Probabilities, Are Not Even Frequentist!

> Fisher saw the *p* value as a measure of evidence, not as a frequentist evaluation. Unfortunately, as a measure of evidence it is very misleading. (Hubbard and Bayarri 2003, p. 181)

This entire tour, as you know, is to disentangle a jungle of conceptual issues, not to defend or criticize any given statistical school. In sailing forward to scrutinize Incompatibilist tribes who protest against mixing *p*'s and *α*'s, we need to navigate around a pool of quicksand. They begin by saying *P*-values are for evidence and inference, unlike error probabilities. N-P error probabilities are too performance oriented to be measures of evidence. In the next breath we're told *P*-values aren't good measures of evidence either. A good measure of evidence, it's assumed, should be probabilist, in some way, and *P*-values disagree with probabilist measures, be they likelihood ratios, Bayes factors, or posteriors. If you reinterpret error probabilities, they promise, you can make peace with all tribes. Whether we get on firmer ground or sink in a marshy swamp will have to be explored.

Berger's Unification of Jeffreys, Neyman, and Fisher

With "reconciliation" and "unification" in the air, Jim Berger, a statistician deeply influential in statistical foundations, sets out to see if he can get Fisher, Neyman, and (non-subjective) Bayesian Jeffreys to agree on testing (2003). A compromise awaits, if we nip and tuck the meaning of "error probability" (Section 3.5). If you're an N-P theorist and like your error probability$_1$, you can keep it he promises, but he thinks you will want to reinterpret it. It then becomes possible to say that a *P*-value is not an error probability (full stop), meaning it's not the newly defined error probability$_{2.}$ What's error probability$_2$? It's a type of posterior probability in a null hypothesis, conditional on the outcome, given a prior. It may still be frequentist in some sense. On this reinterpretation, *P*-values are not error probabilities. Neither are N-P Type I and II, α and β. Following the philosopher's clarifying move via subscripts, there is error probability$_1$ – the usual frequentist notion – and error probability$_2$ – notions from probabilism that had never been called error probabilities before.

In commenting on Berger (2003), I noted my surprise at his redefinition (Mayo 2003b). His reply: "Why should the frequentist school have exclusive right to the term 'error probability?' It is not difficult to simply add the designation 'frequentist' (or Type I or Type II) or 'Bayesian' to the term to differentiate between the schools" (Berger 2003, p. 30). That would work splendidly. So let error probability$_2$ = Bayesian error probability. Frankly, I didn't think Bayeslans would want the term. In a minute, however, Berger will claim they alone are the true frequentist error probabilities! If you feel yourself sinking in a swamp of sliding meanings, remove your shoes, flip onto your back atop your walking stick and you'll stop sinking. Then, you need only to pull yourself to firm land. (See Souvenir M.)

The Bayes Factor. In 1987, Berger and Sellke said that in order to consider *P*-values as error probabilities we need to introduce a decision or test rule. Berger (2003) proposes such a rule and error probability$_2$ is born. In trying to merge different methodologies, there's always a danger of being biased in favor of one, begging the question against the others. From the severe tester's perspective, this is what happens here, but so deftly that you might miss it if you blink.[6]

His example involves $X_1, \ldots, X_n$ IID data from $N(\theta, \sigma^2)$, with σ^2 known, and the test is of two simple hypotheses H_0: $\theta = \theta_0$ and H_1: $\theta = \theta_1$. Consider now their two *P*-values: "for $i = 0, 1$, let p_i be the *p*-value in testing H_i against the other hypothesis" (ibid., p. 6). Then reject H_0 when $p_0 \leq p_1$, and accept H_0 otherwise. If you reject H_0 you next compute the posterior probability of H_0 using one of Jeffreys' default priors giving 0.5 to each hypothesis. The computation rests on the *Bayes factor* or likelihood ratio $B(\boldsymbol{x}) = \Pr(\boldsymbol{x}|H_0)/\Pr(\boldsymbol{x}|H_1)$:

$$\Pr(H_0|\boldsymbol{x}) = B(\boldsymbol{x})/[1 + B(\boldsymbol{x})].$$

The priors drop out, being 0.5. As before, $\boldsymbol{x}$ refers to a generic value for $\boldsymbol{X}$.

This was supposed to be something Fisher would like, so what happened to *P*-values? They have a slight walk-on part: the rejected hypothesis is the one that has the lower *P*-value. Its value is irrelevant, but it directs you to which posterior to compute. We might understand his Bayesian error probabilities this way: If I've rejected H_0, I'd be wrong if H_0 were true, so $\Pr(H_0|\boldsymbol{x})$ is a probability of being wrong about H_0. It's the *Bayesian Type I error probability$_2$*. If instead you reject H_1, then you'd be wrong if H_1 were true. So in that case you report the Bayesian Type II error probability$_2$, which would be $\Pr(H_1|\boldsymbol{x}) = 1/[1 + B(\boldsymbol{x})]$. Whatever you think of these, they're quite different from error probability$_1$, which does not use priors in H_i.

Sleight of Hand? Surprisingly, Berger claims to give a "dramatic illustration of the nonfrequentist nature of *P*-values" (ibid., p. 3). Wait a second, how did they become *non-frequentist*? What he means is that the *P*-value can be shown to disagree with the special posterior probability for H_0, defined as error

[6] We are forced to spend more time on *P*-values than one would wish simply because so many of the criticisms and proposed reforms are in terms of them.

probability$_2$. They're not called Bayesian error probabilities any more but frequentist conditional error probabilities (CEPs). Presto! A brilliant sleight of hand.

This 0.5 prior is not supposed to represent degree of belief, but it is Berger's "objective" default Bayesian prior. Why does he call it frequentist? He directs us to an applet showing if we imagine randomly selecting our test hypothesis from a population of null hypotheses, 50% of which are true, the rest false, and then compute the relative frequency of true nulls conditional on its having been rejected at significance level p, we get a number that is larger than p. This violates what he calls the frequentist principle (not to be confused with FEV):

> *Berger's frequentist principle*: $\Pr(H_0$ true $\mid H_0$ rejected at level $p)$ should equal p.

This is very different from what a P-value gives us, namely, $\Pr(P \leq p; H_0) = p$ (or $\Pr(d(\boldsymbol{X}) \geq d(\boldsymbol{x}_0); H_0) = p$).

He actually states the frequentist principle more vaguely; namely, that the reported error probability should equal the actual one, but the computation is to error probability$_2$. If I'm not being as clear as possible, it's because Berger isn't, and I don't want to prematurely saddle him with one of at least two interpretations he moves between. For instance, Berger says the urn of nulls applet is just a heuristic, showing how it could happen. So suppose the null was randomly selected from an urn of nulls 50% of which are true. Wouldn't 0.5 be its frequentist prior? One has to be careful. First consider a legitimate frequentist prior. Suppose I selected the hypothesis H_0: that the mean temperature in the water, θ, is 150 degrees (Section 3.2). I can see this value resulting from various features of the lake and cooling apparatus, and identify the relative frequency that θ takes different values. $\{\Theta = \theta\}$ is an event associated with random variable Θ. Call this an *empirical* or *frequentist* prior just to fix the notion. What's imagined in Berger's applet is very different. Here the analogy is with diagnostic screening for disease, so I will call it that (Section 5.6). We select one null from an urn of nulls, which might include all hypotheses from a given journal, a given year, or lots of other things.[7] If 50% of the nulls in this urn are true, the experiment of

[7] It is ironic that it's in the midst of countering a common charge that he requires repeated sampling from the same population that Neyman (1977) talks about a series of distinct scientific inquiries (presumably independent) with Type I and Type II error probabilities (for specified alternatives) $\alpha_1, \alpha_2, \ldots, \alpha_n, \ldots$ and $\beta_1, \beta_2, \ldots, \beta_n, \ldots$

> I frequently hear a particular regrettable remark … that the frequency interpretation of either the level of significance α or of power $(1 - \beta)$ is only possible when one deals many times WITH THE SAME HYPOTHESIS H, TESTED AGAINST THE SAME ALTERNATIVE. (Neyman 1977, 109, his use of capitals)

randomly selecting a null from the urn could be seen as a Bernoulli trial with two outcomes: a null that is true or false. The probability of selecting a null that has the property "true" is 0.5. Suppose I happen to select H_0: $\theta = 150$, the hypothesis from the accident at the water plant. It would be incorrect to say 0.5 was the relative frequency that $\theta = 150$ would emerge with the empirical prior. So there's a frequentist computation, but it differs from what Neyman's empirical Bayesian would assign it. I'll come back to this later (Excursion 6).

Suppose instead we keep to the default Bayesian construal that Berger favors. The priors come from one or another conventional assignment. On this reading, his frequentist principle is: the *P*-value should equal the default posterior on H_0. That is, a reported *P*-value should equal error probability$_2$. By dropping the designation "Bayesian" that he himself recommended "to differentiate between the schools" (p. 30), it's easy to see how confusion ensues.

Berger emphasizes that the confusion he is on about "is different from the confusion between a *P*-value and the posterior probability of the null hypothesis" (p. 4). What confusion? That of thinking *P*-values are frequentist error probabilities$_2$ – but he has just introduced the shift of meaning! But the only way error probability$_2$ inherits a frequentist meaning is by reference to the heuristic (where the prior is the proportion of true nulls in a hypothetical urn of nulls), giving a diagnostic screening posterior probability. The subscripts are a lifesaver for telling what's true when definitions shift about throughout an argument. The frequentist had only ever wanted error probabilities$_1$ – the ones based solely on the sampling distribution of d($\boldsymbol{X}$). Yet now he declares that error probability$_2$ – Bayesian error probability – is the only real or relevant frequentist error probability! If this is the requirement, preset α, β aren't error probabilities either.

It might be retorted, however, that this was to be a compromise position. We can't dismiss it out of hand because it requires Neyman and Fisher to become default Bayesians. To smoke the peace pipe, everyone has to give a little. According to Berger, "Neyman criticized p-values for violating the frequentist principle." (p. 3) With Berger's construal, it is not violated. So it appears Neyman gets something. Does he? We know N-P used *P*-values, and never saw them as non-frequentist; and surely Neyman wouldn't be criticizing a *P*-value for not being equal to a default (or other) posterior probability. Hence Nancy Reid's quip: "the Fisher/Jeffreys agreement is essentially to have Fisher"

From the Central Limit Theorem, Neyman remarks:

> The relative frequency of the first kind of errors will be close to the arithmetic mean of numbers $\alpha_1, \alpha_2, \ldots, \alpha_n, \ldots$ Also the relative frequency of detecting the falsehood of the hypotheses tested, when false . . . will differ but little from the average of [the corresponding powers, for specified alternatives].

kowtow to Jeffreys (N. Reid 2003). The surest sign that we've swapped out meanings are the selling points.

Consider the Selling Points

"Teaching statistics suddenly becomes easier . . . it is considerably less important to disabuse students of the notion that a frequentist error probability is the probability that the hypothesis is true, given the data" (Berger 2003, p. 8), since his error probability$_2$ actually has that interpretation. We are also free of having to take into account the stopping rule used in sequential tests (ibid.). As Berger dangles his tests in front of you with the labels "frequentist," "error probabilities," and "objectivity," there's one thing you know: if the methods enjoy the simplicity and freedom of paying no price for optional stopping, you'll want to ask if they're also controlling error probabilities$_1$. When that handwringing disappears, unfortunately, so does our assurance that we block inferences that have passed with poor severity.

Whatever you think of default Bayesian tests, Berger's error probability$_2$ differs from N-P's error probability$_1$. N-P requires controlling the Type I and II error probabilities at low values regardless of prior probability assignments. The scrutiny here is not of Berger's recommended tests – that comes later. The scrutiny here is merely to shine a light on the type of shifting meanings that our journey calls for. Always carry your walking stick – it serves as a metaphorical subscript to keep you afloat.

Souvenir M: Quicksand Takeaway

The howlers and chestnuts of Section 3.4 call attention to: the need for an adequate test statistic, the difference between an i-assumption and an actual assumption, and that tail areas serve to raise, and not lower, the bar for rejecting a null hypothesis. The stop in Section 3.5 pulls back the curtain on one front of typical depictions of the N-P vs. Fisher battle, and Section 3.6 disinters equivocal terms in a popular peace treaty between the N-P, Fisher, and Jeffreys tribes. Of these three stops, I admit that the last may still be murky. One strategy we used to clarify are subscripts to distinguish slippery terms. Probabilities of Type I and Type II errors, as well as *P*-values, are defined exclusively in terms of the sampling distribution of d($\boldsymbol{X}$), under a statistical hypothesis of interest. That's error probability$_1$. Error probability$_2$, in addition to requiring priors, involves conditioning on the particular outcome, with the hypothesis varying. There's no consideration of the sampling distribution of d($\boldsymbol{X}$), if you've conditioned on the actual outcome. A second strategy is to consider the selling points of the new "compromise" construal, to gauge what it's asking you to buy.

Here's from our guidebook:

> You're going to need to be patient. Depending on how much quicksand is around you, it could take several minutes or even hours to slowly, methodically get yourself out . . .
>
> *Relax*. Quicksand usually isn't more than a couple feet deep . . . If you panic you can sink further, but if you relax, your body's buoyancy will cause you to float.
>
> Breathe deeply . . . It is impossible to "go under" if your lungs are full of air (WikiHow 2017).

In later excursions, I promise, you'll get close enough to the edge of the quicksand to roll easily to hard ground. More specifically, all of the terms and arguments of Section 3.6 will be excavated.

Tour III Capability and Severity: Deeper Concepts

From the itinerary: A long-standing family feud among frequentists is between hypotheses tests and confidence intervals (CIs), but in fact there's a clear duality between the two. The dual mission of the first stop (Section 3.7) of this tour is to illuminate both CIs and severity by means of this duality. A key idea is arguing from the capabilities of methods to what may be inferred. The severity analysis seamlessly blends testing and estimation. A typical inquiry first tests for the existence of a genuine effect and then estimates magnitudes of discrepancies, or inquires if theoretical parameter values are contained within a confidence interval. At the second stop (Section 3.8) we reopen a highly controversial matter of interpretation that is often taken as settled. It relates to statistics and the discovery of the Higgs particle – displayed in a recently opened gallery on the "Statistical Inference in Theory Testing" level of today's museum.

3.7 Severity, Capability, and Confidence Intervals (CIs)

It was shortly before Egon offered him a faculty position at University College starting 1934 that Neyman gave a paper at the Royal Statistical Society (RSS) which included a portion on confidence intervals, intending to generalize Fisher's fiducial intervals. With K. Pearson retired (he's still editing *Biometrika* but across campus with his assistant Florence David), the tension is between E. Pearson, along with remnants of K.P.'s assistants, and Fisher on the second and third floors, respectively. Egon hoped Neyman's coming on board would melt some of the ice.

Neyman's opinion was that "Fisher's work was not really understood by many statisticians . . . mainly due to Fisher's very condensed form of explaining his ideas" (C. Reid 1998, p. 115). Neyman sees himself as championing Fisher's goals by means of an approach that gets around these expository obstacles. So Neyman presents his first paper to the Royal Statistical Society (June, 1934), which includes a discussion of confidence intervals, and, as usual, comments (later published) follow. Arthur Bowley (1934), a curmudgeon on the K.P. side of the aisle, rose to thank the speaker. Rubbing his hands together in gleeful anticipation of a blow against Neyman by Fisher, he declares: "I am very glad Professor Fisher is present, as it is his work that Dr Neyman has accepted and incorporated. . . . I am not at all sure that the 'confidence' is not a confidence trick" (p.132). Bowley was to be disappointed. When it was Fisher's turn, he was full of praise. "Dr Neyman . . . claimed to have generalized the argument of fiducial probability, and he had every reason to be proud of the line of argument he had developed for its perfect clarity" (Fisher 1934c, p.138). Caveats were to come later (Section 5.7). For now, Egon was relieved:

> Fisher had on the whole approved of what Neyman had said. If the impetuous Pole had not been able to make peace between the second and third floors of University College, he had managed at least to maintain a friendly foot on each! (C. Reid 1998, p. 119)

CIs, Tests, and Severity. I'm always mystified when people say they find *P*-values utterly perplexing while they regularly consume polling results in terms of confidence limits. You could substitute one for the other.

Suppose that 60% of 100 voters randomly selected from a population U claim to favor candidate Fisher. An estimate of the proportion of the population who favor Fisher, θ, at least at this point in time, is typically given by means of confidence limits. A 95% confidence interval for θ is $\overline{x} \pm 1.96\sigma_{\overline{X}}$ where $\overline{x}$ is the observed proportion and we estimate $\sigma_{\overline{X}}$ by plugging $\overline{x}$ in for θ to get $\sigma_{\overline{X}} = \sqrt{[0.60\ (0.40)/100]} = 0.048$. The 95% CI limits for $\theta = 0.6 \pm 0.09$ using the Normal approximation. The lower limit is 0.51 and the upper limit is 0.69. Often, 0.09 is reported as the *margin of error*. We could just as well have asked, having observed $\overline{x} = 0.6$,

> what value of θ would 0.6 be statistically significantly greater than at the 0.025 level, and what value of θ would 0.6 be statistically significantly less than at the 0.025 level?

The two answers would yield 0.51 and 0.69, respectively. So infer $\theta > 0.51$ and infer $\theta < 0.69$ (against their denials), each at level 0.025, for a combined error probability of 0.05.

Not only is there a duality between confidence interval estimation and tests, they were developed by Jerzy Neyman at the same time he was developing tests! The 1934 paper in the opening to this tour builds on Fisher's fiducial intervals dated in 1930, but he'd been lecturing on it in Warsaw for a few years already. Providing upper and lower confidence limits shows the range of plausible values for the parameter and avoids an "up/down" dichotomous tendency of some users of tests. Yet, for some reason, CIs are still often used in a dichotomous manner: rejecting μ values excluded from the interval, accepting (as plausible or the like) those included. There's the tendency, as well, to fix the confidence level at a single $1 - \alpha$, usually 0.9, 0.95, or 0.99. Finally, there's the adherence to a performance rationale: the estimation method will cover the true θ 95% of the time in a series of uses. We will want a much more nuanced, inferential construal of CIs. We take some first steps toward remedying these shortcomings by relating confidence limits to tests and to severity.

To simply make these connections, return to our test T+, an IID sample from a Normal distribution, H_0: $\mu \leq \mu_0$ against H_1: $\mu > \mu_0$. In a CI estimation procedure, an observed statistic is used to set an upper or lower (one-sided) bound, or both upper and lower (two-sided) bounds for parameter μ. Good and best properties of tests go over into good or best properties of corresponding confidence intervals. In particular, the uniformly most powerful (UMP) test T+ corresponds to a uniformly most accurate lower confidence bound (see Lehmann and Romano 2005, p. 72). The $(1 - \alpha)$ uniformly most accurate (UMA) lower confidence bound for μ, which I write as $\hat{\mu}_{1-\alpha}(\overline{X})$, corresponding to test T+ is

$$\mu > \overline{X} - c_\alpha(\sigma/\surd n),$$

where $\overline{X}$ is the sample mean, and the area to the right of c_α under the standard Normal distribution is α. That is $\Pr(Z \geq c_\alpha) = \alpha$ where Z is the standard Normal statistic. Here are some useful approximate values for c_α:

α	0.5	0.16	0.05	0.025	0.02	0.005	0.001
c_α	0	1	1.65	1.96	2	2.5	3

The Duality

"Infer: $\mu > \overline{X} - 2.5(\sigma/\surd n)$" alludes to the rule for inferring; it is the CI *estimator*. Substituting $\overline{x}$ for $\overline{X}$ yields an *estimate*. Here are some abbreviations, alluding throughout to our example of a UMA estimator:

> A *generic* $1 - \alpha$ lower confidence interval estimator is $\mu > \hat{\mu}_{1-\alpha}(\overline{X}) = \mu > \overline{X} - c_\alpha(\sigma/\surd n)$.
>
> A *specific* $1 - \alpha$ lower confidence interval estimate is $\mu > \hat{\mu}_{1-\alpha}(\overline{x}) = \mu > \overline{x} - c_\alpha(\sigma/\surd n)$.

The corresponding value for α is close enough to 0.005 to allow $c_{0.005} = 2.5$ (it's actually closer to 0.006). The impressive thing is that, regardless of the true value of μ, these rules have high coverage probability. If, for any observed $\overline{x}$, in our example, you shout out

$$\mu > \overline{X} - 2.5(\sigma/\surd n),$$

your assertions will be correct 99.5% of the time. The specific inference results from plugging in $\overline{x}$ for $\overline{X}$. The specific 0.995 lower limit = $\hat{\mu}_{0.995}(\overline{x}) = \overline{x} - 2.5(\sigma/\surd n)$, and the specific 0.995 estimate is $\mu > \hat{\mu}_{0.995}(\overline{x})$. This inference is qualified by the error probability of the method, namely the confidence level 0.995. But the upshot of this qualification is often misunderstood. Let's have a new example to show the duality between the lower confidence interval estimator $\mu > \hat{\mu}_{1-\alpha}(\overline{X})$ and the *generic* (α level) test T+ of form: H_0: $\mu \le \mu_0$ against H_1: $\mu > \mu_0$. The "accident at the water plant" has a nice standard error of 1, but that can mislead about the role of sample size n. Let $\sigma = 1$, $n = 25$, $\sigma_{\overline{X}} = (\sigma/\surd n) = 0.2$. (Even though we'd actually have to estimate σ, the logic is the same and it's clearer.) I use $\sigma/\surd n$ rather than $\sigma_{\overline{X}}$ when a reminder of sample size seems needed.

Work backwards. Suppose we've collected the 25 samples and observed sample mean $\overline{x} = 0.6$. (The 0.6 has nothing to do with the polling example at the outset.) For what value of μ_0 would $\overline{x} = 0.6$ exceed μ_0 by $2.5\sigma_{\overline{X}}$? Since $2.5\sigma_{\overline{X}} = 0.5$, the answer is $\mu = 0.1$. If we were testing H_0: $\mu \le 0.1$ vs. H_1: $\mu > 0.1$ at level 0.005, we'd reject with this outcome. The corresponding 0.995 lower estimate would be

$$\mu > 0.1.$$

(see Note 1).

Now for the duality. $\overline{X}$ is not statistically significantly greater than any μ value larger than 0.1 (e.g., 0.15, 0.2, etc.) at the 0.005 level. A test of form T+ would fail to reject each of the values in the CI interval at the 0.005 level, with $\overline{x} = 0.6$. Since this is continuous, it does not matter if the cut-off is at 0.1 or greater than or equal to 0.1.[1] By contrast, if we were testing μ_0 values 0.1 or less (T+: H_0: $\mu \le 0.1$ against H_1: $\mu > 0.1$), these nulls *would* be rejected by $\overline{x} = 0.6$ at the 0.005 level (or even lower for values less than 0.1). That is, under the supposition that the data were generated from a world where H_0: $\mu \le 0.1$, at least 99.5% of the time a *smaller* $\overline{X}$ than what was observed (0.6) would occur:

$$\Pr(\overline{X} < 0.6; \mu = 0.1) = 0.995.$$

The probability of observing $\overline{X} \ge 0.6$ would be low, 0.005.

Severity Fact (for test T+): Taking an outcome $\overline{x}$ that just reaches the α level of significance $(\overline{x}_\alpha)$as warranting H_1: $\mu > \mu_0$ with severity $(1 - \alpha)$

[1] To avoid confusion, note the duality is altered accordingly. If we set out the test rule for T + H_0: $\mu \le \mu_0$ vs H_1: $\mu > \mu_0$ as reject H_0: iff $\overline{X} \ge \mu_0 + c_\alpha(\sigma/\surd n)$, then we do not reject H_0 iff $\overline{X} < \mu_0 + c_\alpha(\sigma/\surd n)$. This is the same as $\mu_0 > \overline{X} - c_\alpha(\sigma/\surd n)$, the corresponding lower CI bound. If the test rule is $\overline{X} > \mu_0 + c_\alpha(\sigma/\surd n)$, the corresponding lower bound is $\mu_0 \ge \overline{X} - c_\alpha(\sigma/\surd n)$.

is mathematically the same as inferring $\mu > \overline{x} - c_\alpha(\sigma/\surd n)$ at level $(1 - \alpha)$.

Hence, there's an intimate mathematical relationship between severity and confidence limits. However, severity will break out of the fixed $(1 - \alpha)$ level, and will supply a non-behavioristic rationale that is now absent from confidence intervals.[2]

Severity and Capabilities of Methods

Begin with an instance of our "Fact": To take an outcome that just reaches the 0.005 significance level as warranting H_1 with severity 0.995, is the same as taking the observed $\overline{x}$ and inferring μ just exceeds the 99.5 and lower confidence bound: $\mu > 0.1$. My justification for inferring $\mu > 0.1$ (with $\overline{x} = 0.6$) is this. Suppose my inference is false. Take the smallest value that renders it false, namely $\mu = 0.1$. Were $\mu = 0.1$, then the test very probably would have resulted in a smaller observed $\overline{X}$ than I got (0.6). That is, 99.5% of the time it would have produced a result *less discordant* with claim $\mu > 0.1$ than what I observed. (For μ values less than 0.1 this probability is increased.) Given that the method was highly *in*capable of having produced a value of $\overline{X}$ as large as 0.6, if $\mu \leq 0.1$, we argue that there is an indication at least (if not full blown evidence) that $\mu > 0.1$. The severity with which $\mu > 0.1$ "passes" (or is indicated by) this test is approximately 0.995.

Some caveats: First, throughout this exercise, we are assuming these values are "audited," and the assumptions of the model permit the computations to be licit. Second, we recognize full well that we merely have a single case, and inferring a genuine experimental effect requires being able to produce such impressive results somewhat regularly. That's why I'm using the word "indication" rather than evidence. Interestingly though, you don't see the same admonition against "isolated" CIs as with tests. (Rather than repeating these auditing qualifications, I will assume the context directs the interpretation.)

Severity versus Performance. The severity interpretation differs from both the construals that are now standard in confidence interval theory: The first is the *coverage probability* construal, and the second I'm calling *rubbing-off*. The coverage probability rationale is straightforwardly performance oriented. The rationale for the rule: infer

[2] For the computations, in test T+: H_0: $\mu \leq \mu_0$ against H_1: $\mu > \mu_0$. Suppose the observed $\overline{x}$ just reaches the c_α cut-off: $\overline{x} = \mu_0 + c_\alpha\sigma_{\overline{X}}$. The $(1 - \alpha)$ CI lower bound, $\mathrm{CI_L}$, is $\mu > \overline{X} - c_\alpha\sigma_{\overline{X}}$. So $\Pr(\text{test T+ does not reject } H_0; \mu = \mathrm{CI_L}) = \Pr(\overline{X} < \mu_0 + c_\alpha\sigma_{\overline{X}}; \mu = \mu_0)$. Standardize $\overline{X}$ to get Z: $Z = [(\mu_0 + c_\alpha\sigma_{\overline{X}}) - \mu_0](1/\sigma_{\overline{X}}) = c_\alpha$. So the severity for $\mu > \mu_0$ = Pr(test T+ does not reject H_0; $\mu = \mathrm{CI_L}$) = $\Pr(Z < c_\alpha) = (1 - \alpha)$.

$$\mu > \overline{X} - 2.5\sigma/\surd n,$$

is simply that you will correctly cover the true value at least 99.5% of the time in repeated use (we can allow the repetitions to be actual or hypothetical):

$$\Pr(\mu > (\overline{X} - 2.5\sigma/\surd n); \mu) = 0.995.$$

Aside: The equation above is not treating μ as a random variable, although it might look that way. $\overline{X}$ is the random variable. It's the same as asserting $\Pr(\overline{X} \geq \mu + 2.5(\sigma/\surd n); \mu) = 0.005$. Is this performance-oriented interpretation really all you can say? The severe tester says no. Here's where different interpretive philosophies enter.

Cox and Hinkley (1974) do not adhere to a single choice of 1 – α. Rather, to assert a 0.995 CI estimate, they say, is to follow:

> . . . a procedure that would be wrong only in a proportion α of cases, in hypothetical repeated applications, whatever may be the true value μ. Note that this is a hypothetical statement that gives an empirical meaning, which in principle can be checked by experiment, rather than a prescription for using confidence limits. In particular, we do not recommend or intend that a fixed value α_0 should be chosen in advance and the information in the data summarized in the single assertion [$\mu > \hat{\mu}_{1-\alpha}$]. (p. 209, μ is substituted for their θ)

We have the *meaning versus application* gap again, which severity strives to close. "[W]e define procedures for assessing evidence that are calibrated by how they would perform were they used repeatedly. In that sense they do not differ from other measuring instruments" (Cox 2006a, p. 8). Yet this performance is not the immediate justification for the measurement in the case at hand. What I mean is, it's not merely that if you often use a telescope with good precision, your measurements will have a good track record – no more than with my scales (in Section 1.1). Rather, the thinking is, knowing how they would perform lets us infer how they're performing now. Good long-run properties "rub-off" in some sense on the case at hand (provided at least they are the relevant ones).

It's not so clear what's being rubbed off. You can't say the probability it's correct *in this case* is 0.995, since either it's right or not. That's why "confidence" is introduced. Some people say from the fact that the procedure is rarely wrong we may assign a low probability to its being wrong in the case at hand. First, this is dangerously equivocal, since the probability properly attaches to the method of inferring. Some espouse it as an informal use of "probability" outside of statistics, for instance, that confidence is "the degree of belief of a rational person that the confidence interval covers the parameter" (Schweder and Hjort 2016, p. 11). They call this "epistemic probability." My main gripe is that neither epistemic probability, whatever it is, nor performance gives a report of well-testedness associated with the claim at hand.

By providing several limits at different values, we get a more informative assessment, sometimes called a confidence distribution (CD). An early reference is Cox (1958). "The set of all confidence intervals at different levels of probability. . . [yields a] confidence distribution" (Cox 1958, p. 363). We'll visit others later. The severe tester still wants to nudge the CD idea; whether it's a large or small nudge is unclear because members of CD tribes are unclear. By and large, they're either a tad bit too performance oriented or too close to a form of probabilism for a severe tester. Recall I've said I don't see the severity construal out there, so I don't wish to saddle anyone with it. If that is what some CD tribes intend, great.

The severity logic is the counterfactual reasoning: Were μ less than the 0.995 lower limit, then it is very probable (> 0.995) that our procedure would yield a smaller sample mean than 0.6. This probability gives the severity. To echo Popper, $\mu > \hat{\mu}_{1-\alpha}$ is corroborated (at level 0.995) because it may be presented as a *failed attempt to falsify* it statistically. The severe testing philosophy hypothesizes that this is how humans reason. It underwrites formal error statistics as well as day-to-day reasoning.

Exhibit (vii): Capability. Let's see how severity is computed for the CI claim $(\mu > \hat{\mu}_{0.995})$ with $\overline{x} = 0.6$:

1. The particular assertion h is $\mu > 0.1(\hat{\mu}_{0.995} = 0.1)$.
2. $\overline{x} = 0.6$ accords with h, an assertion about a positive discrepancy from 0.1.
3. Values of $\overline{X}$ less than 0.6 accord less well with h. So we want to compute the probability $(\overline{X} < 0.6)$ just at the point that makes h false: $\mu = 0.1$.
   Pr(method would yield $\overline{X} < 0.6; 0.1) = 0.995$.
4. From (3), SEV(μ > 0.1) = 0.995 (or we could write ≥, but our convention will be to write =).

Although we are moving between values of the parameter and values of $\overline{X}$, so long as we are careful, there is no illegitimacy. We can see that CI limits follow severity reasoning. For general lower $1 - \alpha$ limits, with small level α:

The inference of interest is h: $\mu > \hat{\mu}_{1-\alpha}$.
Since Pr(method would yield $\overline{X} < \overline{x}; \mu = \hat{\mu}_{1-\alpha}) = (1 - \alpha)$,
it follows that SEV(h) = $(1 - \alpha)$.

(Lower case h emphasizes these are typically members of the full alternative in a test.) Table 3.5 gives several examples.

Perhaps "capability or incapability" of the method can serve to get at what's rubbing off. The specific moral I've been leading up to can be read right off the Table, as we vary the value for α (from 0.001 to 0.84) and form the corresponding lower confidence bound from $\hat{\mu}_{0.999}$ to $\hat{\mu}_{0.16}$.

> The higher the test's capability to produce such large (or even larger) differences as we observe, under the assumption $\mu = \hat{\mu}$, the *less* severely tested is assertion $\mu > \hat{\mu}$. (See Figure 3.3.)

The third column of Table 3.5 gives the complement to the severity assessment: the capability of a more extreme result, which in this case is α: $\Pr(\overline{X} > \overline{x}; \mu = \hat{\mu}_{1-\alpha}) = \alpha$. This is the Π function – the attained sensitivity in relation to μ: $\Pi(\gamma)$ (section 3.3) – but there may be too many moving parts to see this simply right away. You can return to it later.

We do not report a single, but rather several confidence limits, and the corresponding inferences of form h_1. Take the third row. The 0.975 lower limit that would be formed from $\overline{x} = 0.6$, $\hat{\mu}_{0.975}$, is $\mu = 0.2$. The estimate takes the form $\mu > 0.2$. Moreover, the observed mean, 0.6, is statistically significantly greater than 0.2 at level 0.025. Since $\mu = 0.2$ would very probably produce $\overline{X} < 0.6$, the severe tester takes the outcome as a good indication of $\mu \geq 0.2$. I want to draw your attention to the fact that the probability of producing an $\overline{X} \geq 0.6$ ranges from 0.005 to 0.5 for values of μ between 0.1 and the observed $\overline{x} = 0.6$. It never exceeds 0.5. To see this compute $\Pr(\overline{X} \geq 0.6; \mu = \mu')$ letting μ' range from 0.1 to 0.6. We standardize $\overline{X}$ to get $Z = (\overline{X} - \mu')/(\sigma/\sqrt{n})$ which is N(0,1). To find $\Pr(\overline{X} \geq 0.6; \mu = \mu')$, compute $Z = (0.6 - \mu')/0.2$ and use the areas under the standard Normal curve to get $\Pr(Z \geq z_0)$, μ' ranging from 0.1 to 0.6.

Do you notice it is only for negative z values that the area to the right of z exceeds 0.5? The test only begins to have more than 50% capability of generating observed means as large as 0.6, when μ is larger than 0.6. An important benchmark enters. The lower 0.5 bound $\hat{\mu}_{0.5}$ is 0.6. Since a result even larger than observed is brought about 50% of the time when $\mu = 0.6$, we rightly *block* the inference to $\mu > 0.6$.

Go to the next to last row: using a lower confidence limit at level 0.31! Now nobody goes around forming confidence bounds at level 0.5, let alone 0.31, but they might not always realize that's what they're doing! We could give a performance-oriented justification: the inference to $\mu > 0.7$ from $\overline{x} = 0.6$ is an instance of a rule that errs 31% of the time. Or we could use counterfactual,

Table 3.5 Lower confidence limit with $\overline{x} = 0.6$, α ranging from 0.001 to 0.84 in T+: $\sigma = 1$, $n = 25$, $\sigma_{\overline{X}} = (\sigma/\sqrt{\mathrm{n}}) = 0.2, \overline{x} = 0.6$

α	c_α	$\hat{\mu}_{1-\alpha}$	$h_1{:}\mu{>}\hat{\mu}_{1-\alpha}$	$\Pr(\overline{X} \geq 0.6;\ \mu{=}\hat{\mu}_{1-\alpha}){=}\alpha$	SEV(h_1)
0.001	3	0	$(\mu > 0)$	0.001	0.999
0.005	2.5	0.1	$(\mu > 0.1)$	0.005	0.995
0.025	2	0.2	$(\mu > 0.2)$	0.025	0.975
0.07	1.5	0.3	$(\mu > 0.3)$	0.07	0.93
0.16	1	0.4	$(\mu > 0.4)$	0.16	0.84
0.3	0.5	0.5	$(\mu > 0.5)$	0.3	0.7
0.5	0	0.6	$(\mu > 0.6)$	0.5	0.5
0.69	−0.5	0.7	$(\mu > 0.7)$	0.69	0.31
0.84	−1	0.8	$(\mu > 0.8)$	0.84	0.16

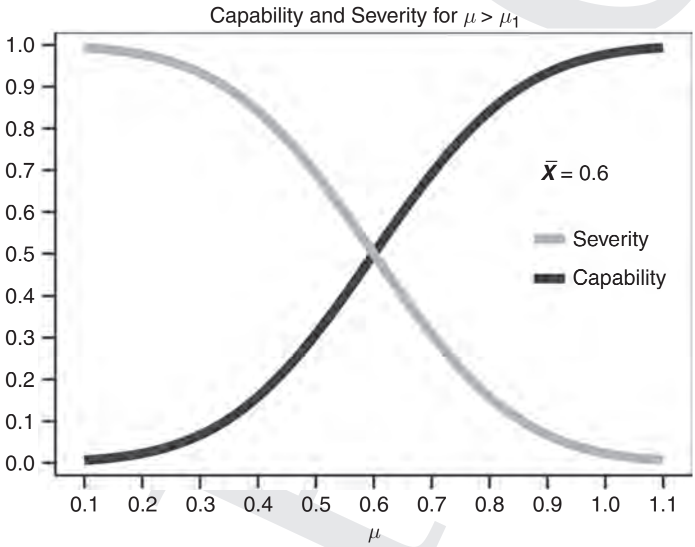

Figure 3.3 Severity for $\mu > \hat{\mu}$.

severity reasoning: even if μ were only 0.7, we'd get a larger $\overline{X}$ than we observed a whopping 69% of the time. Our observed $\overline{X}$ is terrible grounds to suppose μ must exceed 0.7. If anything, we're starting to get an indication that $\mu < 0.7$! Observe that, with larger α, the argument is more forceful by emphasizing >, rather than ≥, but it's entirely correct either way, as it is continuous.

In grasping the duality between tests and confidence limits we consider the *general form* of the test in question, here we considered T+. Given the general form, we imagine the test hypotheses varying, with a fixed outcome $\overline{x}$. Considering other instances of the general test T+ is a heuristic aid in interpreting confidence limits using the idea of statistical inference as severe testing. We will often allude to confidence limits to this end. However, the way the severe tester will actually use the duality is best seen as a post-data way to ask about various discrepancies indicated. For instance, in testing H_0: $\mu \leq 0$ vs. H_1: $\mu > 0$, we may wish to ask, post-data, about a discrepancy such as $\mu > 0.2$. That is, we ask, for each of the inferences, how severely passed it is.

Granted this interval estimator has a nice pivot. If I thought the nice cases weren't the subject of so much misinterpretation, I would not start there. But there's no chance of seeing one's way into more complex cases if we are still hamstrung by the simple ones. In fact, the vast majority of criticism and proposed reforms revolve around our test T+ and two-sided variants. If you grasp the small cluster of the cases that show up in the debates, you'll be able to extend the results. The severity interpretation enables confidence intervals to get around some of their current problems. Let's visit a few of them now. (See also Excursion 4 Tour II, Excursion 5 Tour II.)

Exhibit (viii): Vacuous and Empty Confidence Intervals: Howlers and Chestnuts. *Did you hear the one about the frequentist who reports a confidence level of 0.95 despite knowing the interval must contain the true parameter value?*

Basis for the joke: it's possible that CIs wind up being vacuously true: including all possible parameter values. "Why call it a 95 percent CI if it's known to be true?" the critics ask. The obvious, performance-based, answer is that the confidence level refers to the probability the method outputs true intervals; it's not an assignment of probability to the specific interval. It's thought to be problematic only by insisting on a probabilist interpretation of the confidence level. Jose Bernardo thinks that the CI user "should be subject to some re-education using well-known, standard counterexamples. . . . conventional 0.95-confidence regions may actually consist of the whole real line" (Bernardo 2008, p. 453). Not so.

Cox and Hinkley (1974, p.226) proposed interpreting confidence intervals, or their corresponding confidence limits (lower or upper), as the set of parameter values consistent at the confidence level.

> This interpretation of confidence intervals also scotches criticisms of examples where, due to given restrictions, it can happen that a $(1 - \alpha)$ estimate contains all possible

parameter values. Although such an inference is 'trivially true,' it is scarcely vacuous in our construal. That all parameter values are consistent with the data is an informative statement about the limitations of the data to detect discrepancies at the particular level. (Cox and Mayo 2010, p. 291)

Likewise it can happen that all possible parameter points are inconsistent with the data at the $(1 - \alpha)$ level. Criticisms of "vacuous" and empty confidence intervals stem from a probabilist construal of $(1 - \alpha)$ as the degree of support, belief, or probability attached to the particular interval; but this construal isn't endorsed by CI interval methodology. There is another qualification to add: the error probability computed must be relevant. It must result from the relevant sampling distribution.

Pathological Confidence Set. Here's a famous chestnut that is redolent of Exhibit (vi) in Section 3.4 (Cox's 1958 two measuring instruments with different precisions). It is usually put in terms of a "confidence set" with $n = 2$. It could also be put in the form of a test. Either way, it is taken to question the relevance of error statistical assessments in the case at hand (e.g., Berger and Wolpert 1988, Berger 2003, p. 6). Two independent and identically distributed observations are to be made represented by random variables X_1, X_2. Each X can take either value $\psi - 1$ or $\psi + 1$ with probability of 0.5, where ψ is the unknown parameter to be estimated using the data. The data can result in both outcomes being the same, or both different.

Consider the second case: With both different, we know they will differ by 2. A possibility might be $\langle 9, 11 \rangle$. Right away, we know ψ must be 10. What luck! We know we're right to infer ψ is 10. To depict this case more generally, the two outcomes are $\boldsymbol{x}_1 = x' - 1$ and $\boldsymbol{x}_2 = x' + 1$, for some value x'.

Consider now that the first case obtains. We are not so lucky. The two outcomes are the same: $x_1 = x_2$ (maybe they're both 9 or whatever). What should we infer about the value of parameter ψ? We know ψ is either $x_1 - 1$ or $x_1 + 1$ (e.g., 8 or 10); each accords equally well with the data. Say we infer ψ is $x_1 - 1$. The method is correct with probability 0.5. Averaging over the two possibilities, the probability of an erroneous inference is 0.25. Now suppose I was lucky and observed two different outcomes. Then I know the value of ψ so it makes no sense to infer "ψ is $(x_1 + x_2)/2$" while attaching a confidence coefficient of 0.75.

You see the pattern. The example is designed so that some outcomes yield much more information than others. As with Cox's "two measuring instruments," the data have two parts: First, an indication of whether the two outcomes are the same or different; second, the observed result. Let A be an indicator of the first part: $A = 1$ if both are the same (unlucky); $A = 2$ if the sample values differ by 2 (lucky!). The full data may be represented as $(A, \boldsymbol{x})$. The distribution of A is fixed independently of the parameter of interest: $\Pr(A = 1) = \Pr(A = 2) = 0.5$. It is an example of an *ancillary* statistic. However, learning whether $A = 1$ or $A = 2$ is very informative as to the precision achieved by the inference. Thus the relevant properties associated with the particular inference would be conditional on the value of A.

The tip-off that we're dealing with a problem case is this: The sufficient statistic S has two parts $(A, \boldsymbol{X})$, that is it has *dimension* 2. But there's only one parameter ψ. Without getting into the underlying theory, this alone indicates that S has a property known as being *incomplete*, opening the door to different P-values or confidence levels when calculated conditionally on the value of A. In particular, the marginal distribution of a P-value averaged over the two possibilities $(0.5(0) + 0.5(0.5) = 0.25)$ would be misleading for any particular set of data. Instead we condition on the value of A obtained. David Cox calls this process "*technical conditioning to induce relevance* of the frequentist probability to the inference at hand" (Cox and Mayo 2010, pp. 297–8).

Such examples have other noteworthy features: the ancillary part A gives a sneaky way of assigning a probability to "being correct" in the subset of cases given by the value of A. It's an example of what Fisher called "recognizable subsets." By careful artifice, the event that "a random variable A takes a given value a" is equivalent to "the data were generated by a hypothesized parameter value." So the probability of $A = a$ gives the probability a hypothesis is true. Aris Spanos considers these examples "rigged" for this reason, and he discusses these and several other famous pathological examples (Spanos 2012).

Even putting pathologies aside, is there any reason the frequentist wouldn't do the sensible thing and report on how well probed the inference is once A is known? No. Certainly a severe testing theorist would.

Live Exhibit (ix). What Should We Say When Severity Is Not Calculable? In developing a system like severity, at times a conventional decision must be made. However, the reader can choose a different path and still work within this system.

What if the test or interval estimation procedure does not pass the audit? Consider for the moment that there has been optional stopping, or cherry picking, or multiple testing. Where these selection effects are well understood, we may adjust the error probabilities so that they do pass the audit. But what if the moves are so tortuous that we can't reliably make the adjustment? Or perhaps we don't feel secure enough in the assumptions? Should the severity for $\mu > \mu_0$ be low or undefined?

You are free to choose either. The severe tester says SEV($\mu > \mu_0$) is low. As she sees it, having evidence requires a minimum threshold for severity, even without setting a precise number. If it's close to 0.5, it's quite awful. But if it cannot be computed, it's also awful, since the onus on the researcher is to satisfy the minimal requirement for evidence. I'll follow her: If we cannot compute the severity even approximately (which is all we care about), I'll say it's low, along with an explanation as to why: It's low because we don't have a clue how to compute it!

A probabilist, working with a single "probability pie" as it were, would take a low probability for *H* as giving a high probability to ~*H*. By contrast we wish to clearly distinguish between having poor evidence for *H* and having good evidence for ~*H*. Our way of dealing with bad evidence, no test (BENT) allows us to do that. Both SEV(*H*) and SEV(~*H*) can be low enough to be considered lousy, even when both are computable.

Souvenir N: Rule of Thumb for SEV

> Can we assume that if SEV($\mu > \mu_0$) is a high value, $1 - \alpha$, then SEV($\mu \leq \mu_0$) is α?

Because the claims $\mu > \mu_0$ and $\mu \leq \mu_0$ form a partition of the parameter space, and because we are assuming our test has passed (or would pass) an audit, else these computations go out the window, the answer is yes.

> If SEV($\mu > \mu_0$) is high, then SEV($\mu \leq \mu_0$) is low.

The converse need not hold – given the convention we just saw in Exhibit (ix). At the very least, "low" would not exceed 0.5.

A rule of thumb (for test T+ or its dual CI):

- If we are pondering a claim that an observed difference from the null seems *large* enough to indicate $\mu > \mu'$, we want to be sure the test was highly capable of producing *less* impressive results, were $\mu = \mu'$.
- If, by contrast, the test was highly capable of producing *more* impressive results than we observed, even in a world where $\mu = \mu'$, then we block an inference to $\mu > \mu'$ (following weak severity).

This rule will be at odds with some common interpretations of tests. Bear with me. I maintain those interpretations are viewing tests through "probabilist-colored" glasses, while the correct error-statistical view is this one.

3.8 The Probability Our Results Are Statistical Fluctuations: Higgs' Discovery

One of the biggest science events of 2012–13 was the announcement on July 4, 2012 of evidence for the discovery of a Higgs-like particle based on a "5-sigma observed effect." With the March 2013 data analysis, the 5-sigma difference grew to 7 sigmas, and some of the apparent anomalies evaporated. In October 2013, the Nobel Prize in Physics was awarded jointly to François Englert and Peter W. Higgs for the "theoretical discovery of a mechanism" behind the particle experimentally discovered by the collaboration of thousands of scientists (on the ATLAS and CMS teams) at CERN's Large Hadron Collider in Switzerland. Yet before the dust had settled, the very nature and rationale of the 5-sigma discovery criterion began to be challenged among scientists and in the popular press. Because the 5-sigma standard refers to a benchmark from frequentist significance testing, the discovery was immediately imbued with controversies that, at bottom, concern statistical philosophy.

Why a 5-sigma standard? Do significance tests in high-energy particle (HEP) physics escape the misuses of *P*-values found in social and other sciences? Of course the main concern wasn't all about philosophy: they were concerned that their people were being left out of an exciting, lucrative, many-years project. But unpacking these issues is philosophical, and that is the purpose of this last stop of Excursion 3. I'm an outsider to HEP physics, but that, apart from being fascinated by it, is precisely why I have chosen to discuss it. Anyone who has come on our journey should be able to decipher the more public controversies about using *P*-values.

I'm also an outsider to the International Society of Bayesian Analysis (ISBA), but a letter was leaked to me a few days after the July 4, 2012 announcement, prompted by some grumblings raised by a leading subjective Bayesian, Dennis Lindley. The letter itself was sent around to the ISBA list by statistician Tony O'Hagan. "Dear Bayesians," the letter began. "We've heard a lot about the Higgs boson."

> Why such an extreme evidence requirement? We know from a Bayesian perspective that this only makes sense if (a) the existence of the Higgs boson . . . has extremely small prior probability and/or (b) the consequences of erroneously announcing its discovery are dire in the extreme. (O'Hagan 2012)

Neither of these seemed to be the case in his opinion: "[Is] the particle physics community completely wedded to frequentist analysis? If so, has anyone tried to explain what bad science that is?" (ibid.).

Bad science? Isn't that a little hasty? HEP physicists are sophisticated with their statistical methodology: they'd seen too many bumps disappear. They want to ensure that before announcing a new particle has been discovered that, at the very least, the results being spurious is given a run for its money. Significance tests, followed by confidence intervals, are methods of choice here for good reason. You already know that I favor moving away from traditional interpretations of statistical tests and confidence limits. But some of the criticisms, and the corresponding "reforms," reflect misunderstandings, and the knottiest of them all concerns the very meaning of the phrase (in the title of Section 3.8): "the probability our results are merely statistical fluctuations." Failing to clarify it may well impinge on the nature of future big science inquiry based on statistical models. The problem is a bit delicate, and my solution is likely to be provocative. You may reject my construal, but you'll see what it's like to switch from wearing probabilist, to severe testing, glasses.

The Higgs Results

Here's a quick sketch of the Higgs statistics. (I follow the exposition by physicist Robert Cousins (2017). See also Staley (2017). There is a general model of the detector within which researchers define a "global signal strength" parameter μ "such that H_0: $\mu = 0$ corresponds to the background-only hypothesis and $\mu = 1$ corresponds to the [Standard Model] SM Higgs boson signal in addition to the background" (ATLAS collaboration 2012c). The statistical test may be framed as a one-sided test:

$$H_0\colon \mu = 0 \text{ vs. } H_1\colon \mu > 0.$$

The test statistic d($\boldsymbol{X}$) data records how many *excess events* of a given type are "observed" (from trillions of collisions) in comparison to what would be expected from background alone, given in standard deviation or sigma units. Such excess events give a "signal-like" result in the form of bumps off a smooth curve representing the "background" alone.

The improbability of the different d($\boldsymbol{X}$) values, its sampling distribution, is based on simulating what it would be like under H_0 fortified with much cross-checking of results. These are converted to corresponding probabilities under a standard Normal distribution. The probability of observing results as extreme as or more extreme than 5 sigmas, under H_0, is approximately 1 in 3,500,000! Alternatively, it is said that the probability that the results were just a statistical fluke (or fluctuation) is 1 in 3,500,000.

Why such an extreme evidence requirement, Lindley asked. Given how often bumps disappear, the rule for interpretation, which physicists never intended to be rigid, is something like: if d($\boldsymbol{X}$) ≥ 5 sigma, infer discovery, if d($\boldsymbol{X}$) ≥ 2 sigma, get more data.

Now "deciding to announce" the results to the world, or "get more data" are actions all right, but each corresponds to an evidential standpoint or inference: infer there's evidence of a genuine particle, and infer that spurious bumps had not been ruled out with high severity, respectively.

What "the Results" Really Are

You know from the Translation Guide (Souvenir C) that Pr(d($\boldsymbol{X}$) ≥ 5; H_0) is to be read Pr (the test procedure would yield d($\boldsymbol{X}$) ≥ 5; H_0). Where do we record Fisher's warning that we can only use *P*-values to legitimately indicate a genuine effect by demonstrating an *experimental phenomenon*. In good sciences and strong uses of statistics, "the results" may include demonstrating the "know-how" to generate results that rarely fail to be significant. Also important is showing the test passes an audit (it isn't guilty of selection biases, or violations of statistical model assumptions). "The results of test T" incorporates the entire display of know-how and soundness. That's what the severe tester means by Pr(test T would produce $d(\boldsymbol{X}) \geq d(x_0)$; H_0). So we get:

> *Fisher's Testing Principle*: To the extent that you know how to bring about results that rarely fail to be statistically significant, there's evidence of a genuine experimental effect.

There are essentially two stages of analysis. The first stage is to test for a genuine Higgs-like particle, the second, to determine its properties (production mechanism, decay mechanisms, angular distributions, etc.). Even though the SM Higgs sets the signal parameter to 1, the test is going to be used to learn about the value of any discrepancy from 0. Once the null is rejected at the first stage, the second stage essentially shifts to learning the particle's properties, and using them to seek discrepancies from a new null hypothesis: the SM Higgs.

The *P*-Value Police

The July 2012 announcement gave rise to a flood of buoyant, if simplified, reports heralding the good news. This gave ample grist for the mills of *P*-value critics. Statistician Larry Wasserman playfully calls them the "P-Value Police"(2012a) such as Sir David Spiegelhalter (2012), a professor of the Public's Understanding of Risk at the University of Cambridge. Their job was to examine if reports by journalists and scientists could be seen to be misinterpreting the sigma levels as posterior probability assignments to the various models and claims. Thumbs up or thumbs down! Thumbs up went to the ATLAS group report:

> A statistical combination of these channels and others puts the significance of the signal at 5 sigma, meaning that *only one experiment in 3 million would see an apparent signal this strong in a universe without a Higgs.* (2012a, emphasis added)

Now HEP physicists have a term for an apparent signal that is actually produced due to chance variability alone: a *statistical fluctuation* or *fluke*. Only one experiment in 3 million would produce so strong a background fluctuation. ATLAS (2012b) calls it the "background fluctuation probability." By contrast, Spiegethalter gave a thumbs down to:

> There is less than a one in 3 million chance that their results are a statistical fluctuation.

If they had written "would be" instead of "is" it would get thumbs up. Spiegelhalter's ratings are generally echoed by other Bayesian statisticians. According to them, the thumbs down reports are guilty of misinterpreting the P-value as a posterior probability on H_0.

A careful look shows this is not so. H_0 does not say the observed results are due to background alone; H_0 does not say the result is a fluke. It is just H_0: $\mu = 0$. Although if H_0 were true it *follows* that various results would occur with specified probabilities. In particular, it entails (along with the rest of the background) that large bumps are improbable.

It may in fact be seen as an ordinary error probability:

$$(1)\ \Pr(\text{test T would produce } \mathrm{d}(\boldsymbol{X}) \geq 5;\ H_0) \leq 0.0000003.$$

The portion within the parentheses is how HEP physicists understand "a 5-sigma fluctuation." Note (1) is not a conditional probability, which involves a prior probability assignment to the null. It is not

$$\Pr(\text{test T would produce } \mathrm{d}(\boldsymbol{X}) \geq 5 \text{ and } H_0)/\Pr(H_0).$$

Only random variables or their values are conditioned upon. This may seem to be nit-picking, and one needn't take a hard line on the use of "conditional." I mention it because it may explain part of the confusion here. The relationship between the null hypothesis and the test results is intimate: the assignment of probabilities to test outcomes or values of d($\boldsymbol{X}$) "under the null" may be seen as a tautologous statement.

Since it's not just a single result, but also a dynamic test display, we might even want to emphasize a fortified version:

> (1)* Pr(test T would display d($\boldsymbol{X}$) ≥ 5; H_0) ≤ 0.0000003.

Critics may still object that (1), even fortified as (1)*, only entitles saying:

> There is less than a one in 3 million chance of a fluctuation (at least as strong as in their results).

It does not entitle one to say:

> There is less than a one in 3 million chance that *their results* are a statistical fluctuation.

Let's compare three "ups" and three "downs" to get a sense of the distinction that leads to the brouhaha:

Ups

U-1. The probability of the background alone fluctuating up by this amount or more is about one in 3 million. (CMS 2012)
U-2. Only one experiment in 3 million would see an apparent signal this strong in a universe described in H_0.
U-3. The probability that their signal would result by a chance fluctuation was less than one chance in 3 million.

Downs

D-1. The probability their results were due to the background fluctuating up by this amount or more is about one in 3 million.
D-2. One in 3 million is the probability the signal is a false positive – a fluke produced by random statistical fluctuation.
D-3. The probability that their signal was the result of a statistical fluctuation was less than one chance in 3 million.

The difference is that the thumbs down allude to "this" signal or "these" data are due to chance or is a fluctuation. Critics might say the objection to "this" is that the *P*-value refers to a difference as great or greater – a tail area. But if the probability of {d($\boldsymbol{X}$) ≥ d($\boldsymbol{x}$)} is low under H_0, then Pr (d($\boldsymbol{X}$) = d($\boldsymbol{x}$); H_0) is even lower. We've dealt with this back with Jeffreys' quip (Section 3.4). No statistical account recommends going from improbability of a point result on a continuum under *H* to rejecting *H*. The Bayesian looks to the prior probability in H and its alternatives. The error statistician looks to the general procedure. The notation $\{d(\boldsymbol{X}) \geq d(\boldsymbol{x})\}$ is used to signal the latter.

But if we're talking about the procedure, the critic rightly points out, we are not assigning probability to these particular data or signal. True, but that's the way frequentists always give probabilities to general events, whether they have occurred, or we are contemplating a hypothetical excess of 5 sigma that might occur. It's always treated as a generic type of event. We are never considering the probability "the background fluctuates up this much on Wednesday July 4, 2012," except as that is construed as a type of collision result at a type of detector, and so on. It's illuminating to note, at this point:

> [t]he key distinction between Bayesian and sampling theory statistics is the issue of what is to be regarded as random and what is to be regarded as fixed. To a Bayesian, parameters are random and data, once observed, are fixed... (Kadane 2011, p. 437)

Kadane's point is that "[t]o a sampling theorist, data are random even after being observed, but parameters are fixed" (ibid.). When an error statistician speaks of the probability that the results standing before us are a mere statistical fluctuation, she is referring to a methodological probability: the probability the method used would produce data displays (e.g., bumps) as impressive as these, under the assumption of H_0. If you're a Bayesian probabilist D-1 through D-3 appear to be assigning a probability to a hypothesis (about the parameter) because, since the data are known, only the parameter remains unknown. But they're to be scrutinizing a non-Bayesian procedure here. Whichever approach you favor, my point is that they're talking past each other. To get beyond this particular battle, this has to be recognized.

The Real Problem with D-1 through D-3. The error probabilities in U-1 through U-3 are straightforward. In the Higgs experiment, the needed computations are based on simulating relative frequencies of events where H_0: $\mu = 0$ (given a detector model). In terms of the corresponding P-value:

(1) Pr(test T would produce a P-value ≤ 0.0000003; H_0) ≤ 0.0000003.

D-1, 2, 3 are just slightly imprecise ways of expressing U-1, 2, 3. So what's the objection to D-1, 2, 3? It's the danger some find in moving from such claims to their complements. If I say there's a 0.0000003 probability their results are due to chance, some infer there's a 0.999999 (or whatever) probability their results are not due to chance – are not a false positive, are not a fluctuation. And those claims are wrong. If Pr(A; H_0) = p, for some assertion A, the probability of the complement is Pr(not-A; H_0) = $1 - p$. In particular:

(1) Pr(test T would *not* display a P-value ≤ 0.0000003; H_0) > 0.9999993.

There's no transposing! That is, the hypothesis after the ";" does not switch places with the event to the left of ";"! But despite how the error statistician hears D-1 through D-3, I'm prepared to grant the corresponding U claims are safer. I assure you that my destination is not merely refining statistical language, but when critics convince practitioners that they've been speaking Bayesian prose without knowing it (as in Molière), the consequences are nontrivial. I'm about to get to them.

Detaching Inferences Uses an Implicit Severity Principle

Phrases such as "the probability our results are a statistical fluctuation (or fluke) is very low" are common enough in HEP – although physicists tell me it's the science writers who reword their correct U-claims as slippery D-claims. Maybe so. But if you follow the physicist's claims through the process of experimenting and modeling, you find they are alluding to proper error probabilities. You may think they really mean an illicit posterior probability assignment to "real effect" or H_1 if you think that statistical inference takes the form of probabilism. In fact, if you're a Bayesian probabilist, and assume the statistical inference must have a posterior probability, or a ratio of posterior probabilities, you will regard U-1 through U-3 as legitimate but irrelevant to inference; and D-1 through D-3 as relevant only by misinterpreting P-values as giving a probability to the null hypothesis H_0.

If you are an error statistician (whether you favor a behavioral performance or a severe probing interpretation), even the correct claims U-1 through U-3 are not statistical inferences! They are the (statistical) justifications associated with implicit statistical inferences, and even though HEP practitioners are well aware of them, they should be made explicit. Such inferences can take many forms, such as those I place in brackets:

U-1. The probability of the background alone fluctuating up by this amount or more is about one in 3 million.

[Thus, our results are not due to background fluctuations.]

U-2. Only one experiment in 3 million would see an apparent signal this strong in a universe [where H_0 is adequate].

[Thus H_0 is not adequate.]

U-3. The probability that their signal would result by a chance fluctuation was less than one in 3.5 million.

[Thus the signal was not due to chance.]

The formal statistics moves from

(1) Pr(test T produces $d(\boldsymbol{X}) \geq 5; H_0) < 0.0000003$

to

(2) there is strong evidence for
(first) (2a) a genuine (non-fluke) discrepancy from H_0;
(later) (2b) H^*: a Higgs (or a Higgs-like) particle.

They move in stages from indications, to evidence, to discovery. Admittedly, moving from (1) to inferring (2) relies on the implicit assumption of error statistical testing, the severity principle. I deliberately phrase it in many ways. Here's yet another, in a Popperian spirit:

> *Severity Principle* (from low *P*-value) Data provide evidence for a genuine discrepancy from H_0 (just) to the extent that H_0 would (very probably) have survived, were H_0 a reasonably adequate description of the process generating the data.

What *is* the probability that H_0 would have "survived" (and not been falsified) at the 5-sigma level? It is the probability of the complement of the event $\{d(\boldsymbol{X}) \geq 5\}$, namely, $\{d(\boldsymbol{X}) < 5\}$ under H_0. Its probability is correspondingly 1 – 0.0000003. So the overall argument starting from a fortified premise goes like this:

> (1)* With probability 0.9999997, the bumps would be smaller, would behave like statistical fluctuations: disappear with more data, wouldn't be produced at both CMS and ATLAS, in a world adequately modeled by H_0.

They did not disappear, they grew (from 5 to 7 sigma). So,

> (2a) infer there's evidence of H_1: non-fluke, or (2b) infer H^*: a Higgs (or a Higgs-like) particle.

There's always the error statistical qualification of the inference in (2), given by the relevant methodological probability. Here it is a report of the stringency or severity of the test that the claim has passed, as given in (1)*: 0.9999997. We might even dub it the severity coefficient. Without making the underlying principle of testing explicit, some critics assume the argument is all about the reported *P*-value. It's a mere stepping stone to an inductive inference that is detached.

Members of a strict (N-P) behavioristic tribe might reason as follows: If you follow the rule of behavior: Interpret 5-sigma bumps as a real effect (a discrepancy from 0), you'd erroneously interpret data with probability less than 0.0000003 – a very low *error probability.* Doubtless, HEP physicists are keen to avoid repeating such mistakes as apparently finding particles that move faster than light, only to discover some problem with the electrical wiring (Reich 2012). I claim the specific evidential warrant for the 5-sigma Higgs inferences aren't low long-run errors, but being able to detach an inference based on a stringent test or a *strong* argument from coincidence.[3]

Learning How Fluctuations Behave: The Game of Bump-Hunting

Dennis Overbye (2013) wrote an article in the *New York Times*: "Chasing the Higgs," based on his interviews with spokespeople Fabiola Gianotti (ATLAS) and Guido Tonelli (CMS). It's altogether common, Tonelli explains, that the bumps they find are "random flukes" – spuriously significant results – "So 'we crosscheck everything' and 'try to kill' any anomaly that might be merely random."

> One bump on physicists' charts . . . was disappearing. But another was blooming like the shy girl at a dance. . . . nobody could remember exactly when she had come in. But she was the one who would marry the prince . . . It continued to grow over the fall until it had reached the 3-sigma level – the chances of being a fluke [spurious significance] were less than 1 in 740, enough for physicists to admit it to the realm of "evidence" of something, but not yet a discovery. (Overbye 2013)

What's one difference between HEP physics and fields where most results are claimed to be false? HEP physicists don't publish on the basis of a single, isolated (nominal) *P*-value. That doesn't mean promising effects don't disappear. "'We've made many discoveries,' Dr. Tonelli said, 'most of them false'" (ibid.).

Look Elsewhere Effect (LEE). The null hypothesis is formulated to correspond to regions where an excess or bump is found. Not knowing the mass region in advance means "the local *p*-value did not include the fact that 'pure chance' had lots of opportunities . . . to provide an unlikely occurrence" (Cousins 2017, p. 424). So here a nominal (they call it local) *P*-value is assessed at a particular, data-determined, mass. But the probability of so impressive a difference anywhere in a mass range – the global

[3] The inference to (2) is a bit stronger than merely falsifying the null because certain properties of the particle must be shown at the second stage.

P-value – would be greater than the local one. "The original concept of '5σ' in HEP was therefore mainly motivated as a (fairly crude) way to account for a multiple trials factor . . . known as the 'Look Elsewhere Effect'" (ibid. p. 425). HEP physicists often report both local and global *P*-values.

Background information enters, not via prior probabilities of the particles' existence, but as to how researchers might be led astray. "If they were flukes, more data would make them fade into the statistical background . . . If not, the bumps would grow in slow motion into a bona fide discovery" (Overbye 2013). So, they give the bump a hard time, they stress test, look at multiple decay channels, and they hide the details of the area they found it from the other team. When two independent experiments find the same particle signal at the same mass, it helps to overcome the worry of multiple testing, strengthening an argument from coincidence.

Once the null is rejected, the job shifts to testing if various parameters agree with the SM predictions.

> This null hypothesis of no Higgs (or Higgs-like) boson was definitively rejected upon the announcement of the observation of a new boson by both ATLAS and CMS on July 4, 2012. The confidence intervals for signal strength θ . . . were in reasonable agreement with the predictions for the SM Higgs boson. Subsequently, much of the focus shifted to measurements of . . . production and decay mechanisms. For measurements of continuous parameters, . . . the tests . . . use the frequentist duality . . . between interval estimation and hypothesis testing. One constructs (approximate) confidence intervals and regions for parameters . . . and checks whether the predicted values for the SM Higgs boson are within the confidence regions. (Cousins 2017, p. 414)

Now the corresponding null hypothesis, call it H_0^2, is the SM Higgs boson

$$H_0^2\text{: SM Higgs boson: } \mu = 1$$

and discrepancies from it are probed and estimated with confidence intervals. The most important role for statistical significance tests is actually when results are insignificant, or the *P*-values are not small: *negative* results. They afford a standard for blocking inferences that would be made too readily. In this episode, they arose to

(a) block precipitously declaring evidence of a new particle;
(b) rule out values of various parameters, e.g., spin values that would preclude its being "Higgs-like," and various mass ranges of the particle.

While the popular press highlighted the great success for the SM, the HEP physicists, at both stages, were vigorously, desperately seeking to uncover BSM (Beyond the Standard Model) physics.

Once again, the background knowledge of fluke behavior was central to curbing their enthusiasm about bumps that hinted at discrepancies with the new null: H_0^2: $\mu = 1$. Even though July 2012 data gave evidence of the existence of a Higgs-like particle – where calling it "Higgs-like" still kept the door open for an anomaly with the "plain vanilla" particle of the SM – they also showed some hints of such an anomaly.

Matt Strassler, who, like many, is longing to find evidence for BSM physics, was forced to concede: "The excess (in favor of BSM properties) has become a bit smaller each time … That's an unfortunate sign, if one is hoping the excess isn't just a statistical fluke" (2013a). Or they'd see the bump at ATLAS … and not CMS. *"Taking all of the LHC's data, and not cherry picking* … there's nothing here that you can call 'evidence'" for the much sought BSM (Strassler 2013b). They do not say the cherry-picked results 'give evidence, but disbelief in BSM physics lead us to discount it,' as Royall's Likelihoodist may opt to. They say: "There's nothing here that you can call evidence."

Considering the frequent statistical fluctuations, and the hot competition between the ATLAS and CMS to be first, a tool for when to "curb their enthusiasm" is exactly what was wanted. So, this negative role of significance tests is crucial for denying BSM anomalies are real, and setting upper bounds for these discrepancies with the SM Higgs. Since each test has its own test statistic, I'll use g($\boldsymbol{x}$) rather than d($\boldsymbol{x}$).

> *Severity Principle (for non-significance):* Data provide evidence to rule out a discrepancy δ^* to the extent that a larger g($\boldsymbol{x}_0$) would very probably have resulted if δ were as great as δ^*.

This can equivalently be seen as inferring confidence bounds or applying FEV. The particular value of δ^* isn't so important at this stage. What happens with negative results here is that the indicated discrepancies get smaller and smaller as do the bumps, and just vanish. These were not genuine effects, even though there's no falsification of BSM.

Negative results in HEP physics are scarcely the stuff of file drawers, a serious worry leading to publication bias in many fields. Cousins tells of the wealth of papers that begin "Search for …" (2017, p. 412). They are regarded as important and informative – if only in ruling out avenues for theory development. There's another idea for domains confronted with biases against publishing negative results.

Back to O'Hagan and a 2015/2016 Update

O'Hagan published a digest of responses a few days later. When it was clear his letter had not met with altogether enthusiastic responses, he backed off, admitting that he had only been being provocative with the earlier letter. Still, he declares, the Higgs researchers would have been better off avoiding the "ad hoc" 5 sigma by doing a proper (subjective) Bayesian analysis. "They would surely be willing to [announce SM Higgs discovery] if they were, for instance, 99.99 percent certain" [SM Higgs] existed. Wouldn't it be better to report

$$\Pr(\text{SM Higgs}|\text{data}) = 0.9999?$$

Actually, no. Not if it's taken as a formal probability rather than a chosen way to abbreviate: the reality of the SM Higgs has passed a severe test. Physicists believed in a Higgs particle before building the big billion-dollar collider. Given the perfect predictive success of the SM, and its simplicity, such beliefs would meet the familiar standards for plausibility. But that's very different from having evidence for a discovery, or information about the characteristics of the particle. Many aver they didn't expect it to have so small a mass, 125 GeV. In fact, given the unhappy consequences some find with this low mass, some researchers may well have gone back and changed their prior probabilities to arrive at something more sensible (more "natural" in the parlance of HEP). Yet, their strong argument from coincidence via significance tests prevented the effect from going away.

O'Hagan/Lindley admit that a subjective Bayesian model for the Higgs would require prior probabilities to scads of high dimensional "nuisance" parameters of the background and the signal; it would demand multivariate priors, correlations between parameters, joint priors, and the ever worrisome Bayesian catchall factor: Pr(data|not-H^*). Lindley's idea of subjectively eliciting beliefs from HEP physicists is rather unrealistic here.

Now for the update. When the collider restarted in 2015, it had far greater collider energies than before. On December 15, 2015 something exciting happened: "ATLAS and CMS both reported a small 'bump' in their data" at a much higher energy level than the Higgs: 750 GeV (compared to 125 GeV) (Cartlidge 2016). "As this unexpected bump could be the first hint of a new massive particle that is not predicted by the Standard Model of particle physics, the data generated hundreds of theory papers that attempt to explain the signal" (ibid.). I believe it was 500.

> The significance reported by CMS is still far below physicists' threshold for a discovery: 5 sigma, or a chance of around 3 in 10 million that the signal is a statistical fluke. (Castelvecchi and Gibney 2016)

We might replace "the signal" with "a signal like this" to avoid criticism. While more stringent than the usual requirement, the "we're not that impressed" stance kicks in. It's not so very rare for even more impressive results to occur by background alone. As the data come in, the significance levels will either grow or wane with the bumps:

> Physicists say that by June, or August [2016] at the latest, CMS and ATLAS should have enough data to either make a statistical fluctuation go away – if that's what the excess is – or confirm a discovery. (Castelvecchi and Gibney 2016)

Could the Bayesian model wind up in the same place? Not if Lindley/ O'Hagan's subjective model merely keeps updating beliefs in the already expected parameters. According to Savage, "The probability of 'something else' . . . is definitely very small" (Savage 1962, p. 80). It would seem to require a long string of anomalies before the catchall is made sufficiently probable to start seeking new physics. Would they come up with a particle like the one they were now in a frenzy to explain? Maybe, but it would be a far less efficient way for discovery than the simple significance tests.

I would have liked to report a more exciting ending for our tour. The promising bump or "resonance" disappeared as more data became available, drowning out the significant indications seen in April. Its reality was falsified.

Souvenir O: Interpreting Probable Flukes

There are three ways to construe a claim of the form: A small *P*-value indicates it's improbable that the results are statistical flukes.

(1) The person is using an informal notion of probability, common in English. They mean a small *P*-value gives grounds (or is evidence) of a genuine discrepancy from the null. Under this reading there is no fallacy. Having inferred *H**: Higgs particle, one may say informally, "so probably we have experimentally demonstrated the Higgs," or "probably, the Higgs exists."

"So probably" H_1 is merely qualifying the grounds upon which we assert evidence for H_1.

(2) An ordinary error probability is meant. When particle physicists associate a 5-sigma result with claims like "it's highly improbable our results are a statistical fluke," the reference for "our results" includes: the overall display of bumps, with significance growing with more and better data, along with satisfactory crosschecks. Under this reading, again, there is no fallacy.

To turn the tables on the Bayesians a bit, maybe they're illicitly sliding from what may be inferred from an entirely legitimate high probability. The reasoning is this: With probability 0.9999997, our methods would show that the bumps disappear, under the assumption the data are due to background H_0. The bumps don't disappear but grow. Thus, infer H^*: real particle with thus and so properties. Granted, unless you're careful about forming probabilistic complements, it's safer to adhere to the claims along the lines of U-1 through U-3. But why not be careful in negating D claims? An interesting phrase ATLAS sometimes uses is in terms of "the background fluctuation probability": "This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with . . . the Standard Model Higgs boson" (2012b, p.1).

(3) The person is interpreting the P-value as a posterior probability of null hypothesis H_0 based on a prior probability distribution: $p = \Pr(H_0|\boldsymbol{x})$. Under this reading there is a fallacy. Unless the P-value tester has explicitly introduced a prior, it would be "ungenerous" to twist probabilistic assertions into posterior probabilities. It would be a kind of "confirmation bias" whereby one insists on finding a sentence among many that could be misinterpreted Bayesianly.

ASA 2016 Guide: Principle 2 reminds practitioners that P-values aren't Bayesian posterior probabilities, but it slides into questioning an interpretation sometimes used by practitioners – including Higgs researchers:

> P-values do not measure (a) the probability that the studied hypothesis is true, or (b) the probability that the data were produced by random chance alone. (Wasserstein and Lazar 2016, p. 131)[4]

[4] The ASA 2016 Guide's Six Principles:

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

I insert the (a), (b), absent from the original principle 2, because, while (a) is true, phrases along the lines of (b) should not be equated to (a).

Some might allege that I'm encouraging a construal of *P*-values that physicists have bent over backwards to avoid! I admitted at the outset that "the problem is a bit delicate, and my solution is likely to be provocative." My question is whether it is legitimate to criticize frequentist measures from a perspective that assumes a very different role for probability. Let's continue with the ASA statement under principle 2:

> Researchers often wish to turn a *p*-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The *p*-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself. (Wasserstein and Lazar 2016, p. 131)

Start from the very last point: what does it mean, that it's not "about the explanation"? I think they mean it's not a posterior probability on a hypothesis, and that's correct. The *P*-value is a methodological probability that can be used to quantify "how well probed" rather than "how probable." Significance tests can be the basis for, among other things, falsifying a proposed explanation of results, such as that they're "merely a statistical fluctuation." So the statistical inference that emerges is surely a statement about the explanation. Even proclamations issued by high priests – especially where there are different axes to grind – should be taken with severe grains of salt.

As for my provocative interpretation of "probable fluctuations," physicists might aver, as does Cousins, that it's the science writers who take liberties with the physicists' careful U-type statements, turning them into D-type statements. There's evidence for that, but I think physicists may be reacting to criticisms based on how things look from Bayesian probabilists' eyes. For a Bayesian, once the data are known, they are fixed; what's

3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

These principles are of minimal help when it comes to understanding and using *P*-values. The first thing that jumps out is the absence of any mention of *P*-values as error probabilities. (Fisher-N-P Incompatibilist tribes might say "they're not!" In tension with this is the true claim (under #4) that cherry picking results in spurious *P*-values; p. 132.) The ASA effort has merit, and should be extended and deepened.

random is an agent's beliefs or uncertainties on what's unknown – namely the hypothesis. For the severe tester, considering the probability of $\{d(\boldsymbol{X}) \geq d(\boldsymbol{x}_0)\}$ is scarcely irrelevant once $d(\boldsymbol{x}_0)$ is known. It's the way to determine, following the severe testing principles, whether the null hypothesis can be falsified. ATLAS reports, on the basis of the *P*-value display, that "these results provide conclusive evidence for the discovery of a new particle with mass [approximately 125 GeV]" (ATLAS collaboration 2012b, p. 15).

Rather than seek a high probability that a suggested new particle is real; the scientist wants to find out if it disappears in a few months. As with GTR (Section 3.1), at no point does it seem we want to give a high formal posterior probability to a model or theory. We'd rather vouchsafe some portion, say the SM model with the Higgs particle, and let new data reveal, perhaps entirely unexpected, ways to extend the model further. The open-endedness of science must be captured in an adequate statistical account. Most importantly, the 5-sigma report, or corresponding *P*-value, strictly speaking, *is not the statistical inference*. Severe testing premises – or something like them – are needed to move from statistical data plus background (theoretical and empirical) to detach inferences with lift-off.