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

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)

Souvenir C: A Severe Tester's Translation Guide

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

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

> "the test procedure would have yielded"

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

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

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

We then reason as follows:

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

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

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

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

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

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

Souvenir D: Why We Are So New

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Souvenir E: An Array of Questions, Problems, Models

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

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

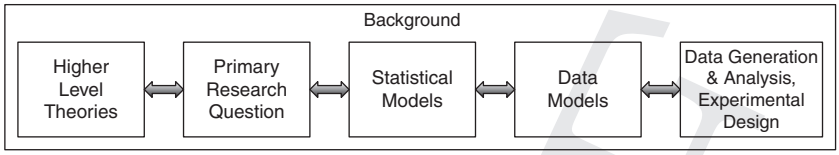

Figure 2.1 Array of questions, problems, models.

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

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

Souvenir F: Getting Free of Popperian Constraints on Language

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

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

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

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

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

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

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

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

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.

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 $N(\mu, \sigma^2)$. Nowadays, the GTR value for $\lambda = \mu$ is set at 1, and the test might be of H_0: $\mu \leq 1$ 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($\boldsymbol{x}$) $\geq c$, for some value of c. Thanks to the requirement in (B), we can calculate the probability that {d($\boldsymbol{X}$) $\geq c$} 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.

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, N(μ, σ^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 d($\boldsymbol{X}$) under H_0: d($\boldsymbol{X}$) ~ N(0,1).

Evaluating the Type II error probability (and power) requires the distribution of 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, N(μ, σ^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 d($\boldsymbol{X}$) under H_0: d($\boldsymbol{X}$) ~ 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 d($\boldsymbol{X}$) under $H_1[\mu = \mu_1]$: d($\boldsymbol{X}$) ~ St(δ_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: $\gamma := \mu_1 - \mu_2 = \gamma_0$ against H_1: $\gamma_1 \neq \gamma_0$.

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

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

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

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

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 $X \sim \text{N}(150, 10^2)$. 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 $X \sim \text{N}(\mu = 150, 10^2)$ 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 $\overline{X} \sim \text{N}(\mu = 150, 1)$.

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

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

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.

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.

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.

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.

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 $\sim H$. By contrast we wish to clearly distinguish between having poor evidence for H and having good evidence for $\sim H$. Our way of dealing with bad evidence, no test (BENT) allows us to do that. Both SEV(H) and SEV($\sim H$) 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.

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.

The last sentence would need to read "low error probabilities relevant for satisfying severity," since low error probabilities won't suffice for a good test. My problem with the general epistemological project of giving necessary and sufficient conditions for knowledge or justified belief or the like is that it does not cash out terms such as "reliability" by alluding to actual methods. The project is one of definition. That doesn't mean it's not of interest to try and link to the more traditional epistemological project to see where it leads. In so doing, Staley and Cobb are right to note that the error-statistician will not hold a strictly externalist view of justification. The trouble with "externalism" is that it makes it appear that a claim (or "belief" as many prefer), is justified so long as a severity relationship SEV holds between data, hypotheses, and a test. It needn't be able to be shown or known. The internalist view, like the appeal to inner coherence in subjective Bayesianism, has a problem in showing how internally justified claims link up to truth. The analytical internal/external distinction isn't especially clear, but from the perspective of that project, Staley and Cobb are right to view ES as a "hybrid" view. In the ES view, the reliability of a method is independent of what anybody knows, but the knower or group of knowers must be able to respond to skeptical challenges such as: you're overlooking flaws, you haven't taken precautions to block errors and so on. They must display the ability to put to rest reasonable skeptical challenges. (Not just any skeptical doubts count, as discussed in solving induction in Section 2.7.) This is an integral part of being an adequate scientific researcher in a domain. (We can sidestep the worry epistemologists might voice that this precludes toddlers from having knowledge; even toddlers can non-verbally display their know-how.) Without showing a claim has been well probed, it has not been well corroborated. Warranting purported severity claims is the task of auditing.

There are interesting attempts to locate objectivity in science in terms of the diversity and clout of the members of the social groups doing the assessing (Longino 2002). Having the stipulated characteristics might even correlate with producing good assessments, but it seems to get the order wrong (Miller 2008). It's necessary to first identify the appropriate requirements for objective criticism. What matters are methods whose statistical properties may be shown in relation to probes on real experiments and data.

Souvenir P: Transparency and Informativeness

There are those who would replace objectivity with the fashionable term "transparency." Being transparent about what was done and how one got

from the raw data to the statistical inferences certainly promotes objectivity, provided I can use that information to critically appraise the inference. For example, being told about stopping rules, cherry picking, altered endpoints, and changed variables is useful in auditing your error probabilities. Simmons, Nelson, and Simonsohn (2012) beg researchers to "just say it," if you didn't p-hack or commit other QRPs. They offer a "21 word solution" that researchers can add to a Methods section: "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study (p. 4)." If your account doesn't use error probabilities, however, it's unclear how to use reports of what would alter error probabilities.

You can't make your inference objective merely announcing your choices and your reasons; there needs to be a system in place to critically evaluate that information. It should not be assumed the scientist is automatically to be trusted. Leading experts might arrive at rival statistical inferences, each being transparent as to their choices of a host of priors and models. What then? It's likely to descend into a battle of the experts. Salesmanship, popularity, and persuasiveness are already too much a part of what passes for knowledge. On the other hand, if well-understood techniques are provided for critical appraisal of the elements of the statistical inference, then transparency could have real force.

One last thing. Viewing statistical inference as severe testing doesn't mean our sole goal is severity. "Shun error" is not a terribly interesting rule to follow. To merely state tautologies is to state objectively true claims, but they are vacuous. We are after the dual aims of severity and informativeness. Recalling Popper, we're interested in "improbable" claims – claims with high information content that can be subjected to more stringent tests, rather than low content claims. Fisher had said that in testing causal claims you should "make [your] theories elaborate by which he meant . . . [draw out] implications" for many different phenomena, increasing the chance of locating any flaw (Mayo and Cox 2006, p. 264). As I see it, the goals of stringent testing and informative theories are mutually reinforcing. Let me explain.

To attain stringent tests, we seek strong arguments from coincidence, and "corroborative tendrils" in order to triangulate results. In so doing, we seek to make our theories more general as Fisher said. A more general claim not only has more content, opening itself up to more chances of failing, it enables cross-checks to ensure that a mistake not caught in one place is likely to ramify somewhere else. A hypothesis H^* with greater depth or scope than another H may be said to be at a "higher level" than H in my horizontal "hierarchy" (Figure 2.1). For instance, the full GTR is at a higher level than the individual

hypothesis about light deflection; and current theories about prion diseases are at a higher level than Prusiner's initial hypotheses limited to kuru. If a higher level theory H^* is subjected to tests with good capacity (high probability) of finding errors, it would be necessary to check and rule out more diverse phenomena than the more limited lower level hypothesis H. Were H^* to nevertheless pass tests, then it does so with higher severity than does H.

its foundation, we still need reference analyses with weakly informative priors to alert us to how much our prior probabilities are driving our posterior probabilities" (2013, p. 76). They rightly point out that, in some circles, giving weight to the null can be the outgrowth of some ill-grounded metaphysics about "simplicity." Or it may be seen as an assumption akin to a presumption of innocence in law. So the question turns on the appropriate prior on the null.

Look what has happened! The problem was simply to express "I'm not impressed" with a result reaching a *P*-value of 0.16: Differences even larger than 1 standard error are not so very infrequent – they occur 16% of the time – even if there's zero effect. So I'm not convinced of the reality of the effect, based on this result. *P*-values did their job, reflecting as they do the severity requirement. H_1 has passed a lousy test. That's that. No prior probability assignment to H_0 is needed. Problem solved.

But there's a predilection for changing the problem (if you're a probabilist). Greenland and Poole feel they're helping us to live with *P*-values without misinterpretation. By choosing the prior so that the *P*-value matches the posterior on H_0, they supply us "with correct interpretations" (ibid., p. 77) where "correct interpretations" are those where the misinterpretation (of a *P*-value as a posterior in the null) is not a misinterpretation. To a severe tester, this results in completely changing the problem from an assessment of how well tested the reality of the effect is, with the given data, to what odds I would give in betting, or the like. We land in the same uncharted waters as with other attempts to fix *P*-values, when we could have stayed on the cruise ship, interpreting *P*-values as intended.

Souvenir Q: Have We Drifted From Testing Country? (Notes From an Intermission)

Before continuing, let's pull back for a moment, and take a coffee break at a place called Spike and Smear. Souvenir Q records our notes. We've been exploring the research program that appears to show, quite convincingly, that significance levels exaggerate the evidence against a null hypothesis, based on evidential assessments endorsed by various Bayesian and Likelihoodist accounts. We suspended the impulse to deny it can make sense to use a rival inference school to critique significance tests. We sought to explore if there's something to the cases they bring as ammunition to this conflict. The Bayesians say the disagreement between their numbers and *P*-values is relevant for impugning *P*-values, so we try to go along with them.

Reflect just on the first argument, pertaining to the case of two-sided Normal testing H_0: $\mu = 0$ vs. H_0: $\mu \neq 0$, which was the most impressive, particularly with $n \geq 50$. It showed that a statistically significant difference from a test hypothesis

at familiar levels, 0.05 or 0.025, can correspond to a result that a Bayesian takes as evidence *for* H_0. The prior for this case is the spike and smear, where the smear will be of the sort leading to J. Berger and Sellke's results, or similar. The test procedure is to move from a statistically significant result at the 0.025 level, say, and infer the posterior for H_0.

Now our minimal requirement for data $\boldsymbol{x}$ to provide evidence for a claim H is that

> (S-1) H accords with (agrees with) $\boldsymbol{x}$, and
> (S-2) there's a reasonable, preferably a high, probability that the procedure would have produced disagreement with H, if in fact H were false.

So let's apply these severity requirements to the data taken as evidence for H_0 here.

Consider (S-1). Is a result that is 1.96 or 2 standard errors away from 0 in good accord with 0? Well, 0 is excluded from the corresponding 95% confidence interval. That does not seem to be in accord with 0 at all. Still, they have provided measures whereby $\boldsymbol{x}$ does accord with H_0, the likelihood ratio or posterior probability on H_0. So, in keeping with the most useful and most generous way to use severity, let's grant (S-1) holds.

What about (S-2)? Has anything been done to probe the falsity of H_0? Let's allow that H_0 is not a precise point, but some very small set of values around 0. This is their example, and we're trying to give it as much credibility as possible. Did the falsity of H_0 have a good chance of showing itself? The falsity of H_0 here is H_1: $\mu \neq 0$. What's troubling is that we found the probability of failing to pick up on population discrepancies as much as 1 standard error in excess of 0 is rather high (0.84) with $n = 100$. Larger sample sizes yield even less capability. Nor are they merely announcing "no discrepancy from 0" in this case. They're finding evidence for 0!

So how did the Bayesian get the bump in posterior probability on the null? It was based on a spiked prior of 0.5 to H_0. All the other points get minuscule priors having to share the remaining 0.5 probability. What was the warrant for the 0.5 prior to H_0? J. Berger and Sellke are quite upfront about it: if they allowed the prior spike to be low, then a rejection of the null would merely be showing an improbable hypothesis got more improbable. "[W]ho, after all, would be convinced," recall their asking: if "my conclusion is that H_0 has posterior probability 0.05 and should be rejected" since it previously had probability, say 0.1 (1987, p. 115). A slight lowering of probability won't cut it. Moving from a low prior to a slightly higher one also lacks punch.

This explains their high prior (at least 0.5) on H_0, but is it evidence for it? Clearly not, nor does it purport to be. We needn't deny there are cases where a theoretical parameter value has passed severely (we saw this in the case of GTR in Excursion 3). But that's not what's happening here. Here they intend for the 0.5 prior to show, *in general*, that statistically significant results problematically exaggerate evidence.[8]

A tester would be worried when the rationale for a spike is to avoid looking foolish when rejecting with a small drop; she'd be worried too by a report: "I don't take observing a mean temperature of 152 in your 100 water samples as indicating it's hotter than 150, because I give a whopping spike to our coolants being in compliance." That is why Casella and R. Berger describe J. Berger and Sellke's spike and smear as maximally biased toward the null (1987a, p. 111). Don't forget the powerful role played by the choice of how to smear the 0.5 over the alternative! Bayesians might reassure us that the high Bayes factor for a point null doesn't depend on the priors given to H_0 and H_1, when what they mean is that it depends only on the priors given to discrepancies under H_1. It was the diffuse prior to the effect size that gave rise to the Jeffreys–Lindley Paradox. It affords huge latitude in what gets supported.

We thought we were traveling in testing territory; now it seems we've drifted off to a different place. It shouldn't be easy to take data as evidence for a claim when that claim is false; but here it is easy (the claim here being H_0). How can this be one of a handful of main ways to criticize significance tests as exaggerating evidence? Bring in a navigator from a Popperian testing tribe before we all feel ourselves at sea:

> Mere supporting instances are as a rule too cheap to be worth having . . . any support capable of carrying weight can only rest upon ingenious tests, undertaken with the aim of refuting our hypothesis, if it can be refuted. (Popper 1983, p. 130)

The high spike and smear tactic can't be take as a basis from which to launch a critique of significance tests because it fails rather glaringly a minimum requirement for evidence, let alone a test. We met Bayesians who don't approve of these tests either, and I've heard it said that Bayesian testing is still a work in progress (Bernardo). Yet a related strategy is at the heart of some recommended statistical reforms.

[8] In the special case, where there's appreciable evidence for a special parameter, Senn argues that Jeffreys only required H_1's posterior probability to be greater than 0.5. One has, so to speak, used up the prior belief by using the spiked prior (Senn 2015a).

There are other interpretations of *P* values that are controversial, in that whether a categorical "No!" is warranted depends on one's philosophy of statistics and the precise meaning given to the terms involved. The disputed claims deserve recognition if one wishes to avoid such controversy. . . .

For example, it has been argued that *P* values overstate evidence against test hypotheses, based on directly comparing *P* values against certain quantities (likelihood ratios and Bayes factors) that play a central role as evidence measures in Bayesian analysis . . . Nonetheless, many other statisticians do not accept these quantities as gold standards, and instead point out that *P* values summarize crucial evidence needed to gauge the error rates of decisions based on statistical tests (even though they are far from sufficient for making those decisions). Thus, from this frequentist perspective, *P* values do not overstate evidence and may even be considered as measuring one aspect of evidence . . . with $1 - P$ measuring evidence against the model used to compute the *P* value. (p. 342)

It's erroneous to fault one statistical philosophy from the perspective of a philosophy with a different and incompatible conception of evidence or inference. The severity principle always evaluates a claim as against its denial within the framework set. In N-P tests, the frame is within a model, and the hypotheses exhaust the parameter space. Part of the problem may stem from supposing N-P tests infer a point alternative, and then seeking that point. Whether you agree with the error statistical form of inference, you can use the severity principle to get beyond this particular statistics battle.

Souvenir R: The Severity Interpretation of Rejection (SIR)

In Tour II you have visited the tribes who lament that *P*-values are sensitive to sample size (Section 4.3), and they exaggerate the evidence against a null hypothesis (Sections 4.4, 4.5). We've seen that significance tests take into account sample size in order to critique the discrepancies indicated objectively. A researcher may choose to decrease the *P*-value as *n* increases, but there's no problem in understanding that the same *P*-value reached with a larger sample size indicates fishing with a finer mesh. Surely we should not commit the fallacy exposed over 50 years ago.

Here's a summary of the severe tester's interpretation (of a rejection) putting it in terms that seem most clear:

> ***SIR: The Severity Interpretation of a Rejection in test T+:*** *(small P-value)*
>
> (i): [*Some* discrepancy is indicated]: $d(\boldsymbol{x}_0)$ is a good indication of $\mu > \mu_1 = \mu_0 + \gamma$ if there is a high probability of observing a *less* statistically significant difference than $d(\boldsymbol{x}_0)$ if $\mu = \mu_0 + \gamma$.

N-P and Fisher tests officially give the case with $\gamma = 0$. In that case, what does a small *P*-value mean? It means the test very probably $(1 - P)$ would have produced a result more in accord with H_0, were H_0 an adequate description of the data-generating process. So it indicates a discrepancy from H_0, especially if I can bring it about fairly reliably. To avoid making mountains out of molehills, it's good to give a second claim about the discrepancies that are *not* indicated:

> (ii): [I'm not *that* impressed]: $d(\boldsymbol{x}_0)$ is a poor indication of $\mu > \mu_1 = \mu_0 + \gamma$ if there is a high probability of an even more statistically significant difference than $d(\boldsymbol{x}_0)$ even if $\mu = \mu_0 + \gamma$.

As for the exaggeration allegation, merely finding a single statistically significant difference, even if audited, is indeed weak: it's an indication of *some* discrepancy from a null, a first step in a task of identifying a genuine effect. But, a legitimate significance tester would never condone rejecting H_0 in favor of alternatives that correspond to a low severity or confidence level such as 0.5. Stephen Senn sums it up: "Certainly there is much more to statistical analysis than P-values but they should be left alone rather than being deformed . . . to become second class Bayesian posterior probabilities" (Senn 2015a). Reformers should not be deformers.

There is an urgency here. Not only do some reforms run afoul of the minimal severity requirement, to suppose things are fixed by lowering *P*-values ignores or downplays the main causes of non-replicability. According to Johnson:

> [I]t is important to note that this high rate of nonreproducibility is not the result of scientific misconduct, publication bias, file drawer biases, or flawed statistical designs; it is simply the consequence of using evidence thresholds that do not represent sufficiently strong evidence in favor of hypothesized effects. (2013a, p. 19316)

This sanguine perspective sidesteps the worry about the key sources of spurious statistical inferences: biasing selection effects and violated assumptions, at all levels. (Fortunately, recent reforms admit this; Benjamin et al. 2017.) Catching such misdemeanors requires *auditing*, the topic of Tours III and IV of this Excursion.

shared by all members of the sample] every independent property has an actual ratio of exactly one-half in the total population. (ibid., p. 376)[8]

The bottom line is, showing how you can distort error probabilities through the efforts of finagling shows the *value* of these methods. It's hard to see how accounts that claim error probabilities are irrelevant can supply such direct protection, although they may *indirectly* block the same fallacies. This remains to be shown.

Souvenir S: Preregistration and Error Probabilities

"One of the best-publicized approaches to boosting reproducibility is preregistration . . . to prevent cherry picking statistically significant results" (Baker 2016, p. 454). It shouldn't be described as too onerous to carry out. Selection effects alter the outcomes in the sample space, showing up in altered error probabilities. If the sample space (and so error probabilities) is deemed irrelevant post-data, the direct rationale for preregistration goes missing. Worse, in the interest of promoting a methodology that downplays error probabilities, researchers who most deserve lambasting are thrown a handy line of defense. Granted it is often presupposed that error probabilities are relevant only for long-run performance goals. I've been disabusing you of that notion. Perhaps some of the "never error probabilities" tribe will shift their stance now: 'But Mayo, using error probabilities for severity, differs from the official line, which is all about performance.' One didn't feel too guilty denying a concern with error probabilities before. If viewing statistical inference as severe tests yields such a concession, I will consider my project a success. Actually, my immediate goal is less ambitious: to show that looking through the severity tunnel lets you unearth the crux of major statistical battles. In the meantime, no fair critic of error statistics should proclaim error control is all about hidden intentions that a researcher can't be held responsible for. They should be.

4.7 Randomization

> The purpose of randomisation . . . is to guarantee the validity of the test of significance, this test being based on an estimate of error made possible by replication. (Fisher [1935b]1951, p. 26)
>
> The problem of analysing the idea of randomization is more acute, and at present more baffling, for subjectivists than for objectivists, more baffling because an ideal subjectivist would not need randomization at all. He would

[8] For a miniature example, if $U = 6$ (there are 6 A's in the population) and $n = 2$, there are 15 possible pairs. Each pair is given a property and so is one additional member.

corrections due to site and data collection can be made later. But the reverse isn't true. As Fisher said, "To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination:. . . to say what the experiment died of" (Fisher 1938, p. 17). Nevertheless, in an attempt to fix the batch effect driven spurious associations,

> [A] whole class of post-experiment statistical methods has emerged . . . These methods . . . represent a palliative, not a cure. . . . the GWAS research community has too often accommodated bad experimental design with automated post-experiment cleanup. . . . experimental designs for large-scale hypothesis testing have produced so many outliers that the field has made it standard practice to automate discarding outlying data. (Lambert and Black 2012, pp. 196–7)

By contrast, with proper design they find that post-experiment automatic filters are unneeded. In other words, the introduction of randomized design *frees them to deliberate* over the handful of extreme values to see if they are real or artifacts. (This contrasts with the worry raised by Cartwright and Hardie (2012).)

Souvenir T: Even Big Data Calls for Theory and Falsification

> Historically, epidemiology has focused on minimizing Type II error (missing a relationship in the data), often ignoring multiple testing considerations, while traditional statistical study has focused on minimizing Type I error (incorrectly attributing a relationship in data better explained by random chance). When traditional epidemiology met the field of GWAS, a flurry of papers reported findings which eventually became viewed as nonreplicable. (Lambert and Black 2012, p. 199)

This is from Christophe Lambert and Laura Black's important paper "Learning from our GWAS Mistakes: From Experimental Design to Scientific Method"; it directly connects genome-wide association studies (GWAS) to philosophical themes from Meehl, Popper and falsification. In an attempt to staunch the non-replication, they explain, adjusted genome-wide thresholds of significance were required as well as replication in an independent sample (Section 4.6).

However, the intended goal is often thwarted by how this is carried out. "[R]esearchers commonly take forward, say, 20–40 nominally significant signals" that did not meet the stricter significance levels, "then run association tests for those signals in a second study, concluding that all the signals with a p-value $\leq .05$ have replicated (no Bonferroni adjustment). Frequently 1 or 2 associations replicate – which is also the number expected by random chance" (ibid.). Next these "replicated" cases are combined with the original data "to compute p-values considered genome-wide significant. This method has been

propagated in publications, leading us to wonder if standard practice could become to publish random signals and tell a plausible biological story about the findings" (ibid.).

Instead of being satisfied with a post-data biological story to explain correlations, "[i]f journals were to insist that association studies also suggest possible experiments that could falsify a putative theory of causation based on association, the quality and durability of association studies could increase" (ibid., p. 201). At the very least, the severe tester argues, we should strive to falsify methods of inquiry and analysis. This might at least scotch the tendency Lambert and Black observe, for others to propagate a flawed methodology once seen in respected journals: "[W]ithout a clear falsifiable stance – one that has implications for the theory – associations do not necessarily contribute deeply to science" (ibid., p. 199).

distribution is strictly inadequate, and a far more complex distribution would be required for answering certain questions. He knows it's strictly false. Yet it suffices to show why the first attempt failed, and it's adequate to solving his immediate problem in pest control.

Souvenir U: Severity in Terms of Problem-Solving

The aim of inquiry is finding things out. To find things out we need to solve problems that arise due to limited, partial, noisy, and error-prone information. Statistical models are at best approximations of aspects of the data-generating process. Reasserting this fact is not informative about the case at hand. These models work because they need only capture rather coarse properties of the phenomena: the error probabilities of the test method are approximately and conservatively related to actual ones. A problem beset by variability is turned into one where the variability is known at least approximately. Far from wanting true (or even "truer") models, we need models whose deliberate falsity enables finding things out.

Our threadbare array of models and questions is just a starter home to grow the nooks and crannies between data and what you want to know (Souvenir E, Figure 2.1). In learning about the large-scale theories of sexy science, intermediate statistical models house two "would-be" claims. Let me explain. The theory of GTR does not directly say anything about an experiment we could perform. Splitting off some partial question, say about the deflection effect, we get a prediction about what *would be* expected were the deflection effect approximately equal to the Einstein value, 1.75". Raw data from actual experiments, cleaned and massaged, afford inferences about intermediate (astrometric) models; inferences as to what it would be like were we taking measurements at the limb of the sun. The two counterfactual inferences – from the theory down, and the data up – meet in the intermediate statistical models. We don't seek a probabilist assignment to a hypothesis or model. We want to know what the data say about a conjectured solution to a problem: What erroneous interpretations have been well ruled out? Which have not even been probed? The warrant for these claims is afforded by the method's capabilities to have informed us of mistaken interpretations. *Statistical methods are useful for testing solutions to problems when this capability/incapability is captured by the relative frequency with which the method avoids misinterpretations.*

If you want to avoid speaking of "truth" you can put the severity requirement in terms of solving a problem. A claim *H* asserts a proposed solution S to an inferential problem is adequate in some respects. It could be a model for prediction, or anything besides.

H: S is adequate for a problem

To reject *H* means "infer *S* is inadequate for a problem." If none of the possible outcomes lead to reject *H* even if *H* is false – the test is incapable of finding inadequacies in *S* – then "do not reject *H*" is BENT evidence that *H* is true. We move from no capability, to some, to high:

> If the test procedure (which generally alludes to a cluster of tests) very rarely rejects *H*, if *H* is true, then "reject *H*" provides evidence for falsifying *H* in the respect indicated.

You could say, a particular inadequacy is corroborated. It's still an inferential question: what's warranted to infer. We start, not with hypotheses, but questions and problems. We want to appraise hypothesized answers severely.

I'll meet you in the ship's library for a reenactment of George Box (1983) issuing "An Apology for Ecumenism in Statistics."

4.9 For Model-Checking, They Come Back to Significance Tests

> Why can't all criticism be done using Bayes posterior analysis . . .? The difficulty with this approach is that by supposing all possible sets of assumptions known *a priori*, it discredits the possibility of new discovery. But new discovery is, after all, the most important object of the scientific process. (George Box 1983, p. 73)

Why the apology for ecumenism? Unlike most Bayesians, Box does not view induction as probabilism in the form of probabilistic updating (posterior probabilism), or any form of probabilism. Rather, it requires critically testing whether a model M_i is "consonant" with data, and this, he argues, demands frequentist significance testing. Our ability "to find patterns in discrepancies $M_i - y_d$ between the data and what might be expected if some tentative model were true is of great importance in the search for explanations of data and of discrepant events" (Box 1983, p. 57). But the dangers of apophenia raise their head.

> However, some check is needed on [the brain's] pattern seeking ability, for common experience shows that some pattern or other can be seen in almost any set of data or facts. This is the object of diagnostic checks and tests of fit which, I will argue, require frequentist theory significance tests for their formal justification. (ibid.)

Once you have inductively arrived at an appropriate model, the move, on his view, "is entirely *deductive* and will be called *estimation*" (ibid., p. 56). The

$$y_t = 17 + 0.2t + 1.5y_{t-1} - 0.6y_{t-2} + \varepsilon_t.$$

The Secret Variable Revealed. At this point, Spanos revealed that x_t was the number of pairs of shoes owned by his grandmother over the observation period! She lives in the mountains of Cyprus, and at last count continues to add to her shoe collection. You will say this is a quirky made-up example, sure. It serves as a "canonical exemplar" for a type of erroneous inference. Some of the best known spurious correlations can be explained by trending means. For live exhibits, check out an entire website by Tyler Vigen devoted to exposing them. I don't know who collects statistics on the correlation between death by getting tangled in bed sheets and the consumption of cheese, but it's exposed as nonsense by the trending means. An example from philosophy that is similarly scotched is the case of sea levels in Venice and the price of bread in Britain (Sober 2001), as shown by Spanos (2010d, p. 366). In some cases, x is a variable that theory suggests is doing real work; discovering the misspecification effectively falsifies the theory from which the statistical model is derived.

I've omitted many of the tests, parametric and non-parametric, single assumption and joint (several assumptions), used in a full application of the same ideas, and mentioned only the bare graphs for simplicity. As you add questions you might wish to pose, they become your new primary inferences. The first primary statistical inference might indicate an effect of a certain magnitude passes with severity, and then background information might enter to tell if it's substantively important. At yet another level, the question might be to test a new model with variables to account for the trending mean of an earlier stage, but this gets beyond our planned M-S testing itinerary. That won't stop us from picking up souvenirs.

Souvenir V: Two More Points on M-S Tests and an Overview of Excursion 4

M-S Tests versus Model Selection: Fit Is Not Enough. M-S tests are distinct from model selection techniques that are so popular. Model selection begins with a family of models to be ranked by one or another criterion. Perhaps the most surprising implication of statistical inadequacy is to call into question the most widely used criterion of model selection: the goodness-of-fit/prediction measures. Such criteria rely on the "smallness" of the residuals. Mathematical fit isn't the same as what's needed for statistical inference. The residuals can be "small" while systematically different from white noise.

Members of the model selection tribe view the problem differently. Model selection techniques reflect the worry of overfitting: that if you add enough factors (e.g., $n - 1$ for sample size n), the fit can be made as good as desired,

even if the model is inadequate for future prediction. (In our examples the factors took the form of trends or lags.) Thus, model selection techniques make you pay a penalty for the number of factors. We share this concern – it's too easy to attain fit without arriving at an adequate model. The trouble is that it remains easy to jump through the model selector's hoops, and still not achieve model adequacy, in the sense of adequately capturing the systematic information in the data. The goodness-of-fit measures already assume the likelihood function, when that's what the M-S tester is probing.

Take the Akaike Information Criterion (AIC) developed by Akaike in the 1970s (Akaike 1973). (There are updated versions, but nothing in our discussion depends on this.) The best known defenders of this account in philosophy are Elliott Sober and Malcolm Forster (Sober 2008, Forster and Sober 1994). An influential text in ecology is by Burnham and Anderson (2002). Model selection begins with a family of models such as the LRM: $y_t = \beta_0 + \beta_1 x_t + u_t$. They ask: Do you get a better fit – smaller residual – if you add $x^2{}_t$? What about adding both $x^2{}_t$ and $x^3{}_t$ terms? And so on. Each time you add a factor, the fit improves, but Akaike kicks you in the shins and handicaps you by 1 for the additional parameter. The result is a preference ranking of models by AIC score.[4] For the granny shoe data above, the model that AIC likes best is

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x^2{}_t + \beta_3 x^3{}_t + u_t.$$

Moreover, it's selection is within this family. But we know that all these LRMs are statistically inadequate! As with other purely comparative measures, there's no falsification of models.

What if we start with the adequate model that the PR arrived at, the autoregressive model with a trend? In that case, the AIC ranks at the very top of the model with the wrong number of trends. That is, it ranks a statistically inadequate model higher than the statistically adequate one. Moreover, the Akaike method for ranking isn't assured of having decent error probabilities. When the Akaike ranking is translated into a N-P test comparing this pair of models, the Type I error probability is around 0.18, and

[4] For each y_t form the residual squared. The sum of the squared residuals:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \hat{\beta}_0 - \sum_{j=1}^{3}\hat{\beta}_j x_t^j\right)^2$$

gives an estimate of σ^2 for the model.

The AIC score for each contender in the case of the LRM, with sample size n, is $\log(\hat{\sigma}^2) + 2K/n$, where K is the number of parameters in model i. The models are then ranked with the smallest being preferred. The log-likelihood is the goodness-of-fit measure which is traded against simplicity, but if the statistical model is misspecified, one is using the wrong measure of fit.

For a comparison of the AIC using these data, and a number of related model-selection measures, see Spanos (2010a). None of these points change using the unbiased variant of AIC.

no warning of the laxity is given. As noted, model selection methods don't hone in on models outside the initial family. By contrast, building the model through our M-S approach is intended to accomplish both tasks – building and checking – in one fell swoop.

Leading proponents of AIC, Burnham and Anderson (2014, p. 627), are quite critical of error probabilities, declaring "P values are not proper evidence as they violate the likelihood principle (Royall 1997)." This tells us their own account forfeits control of error probabilities. "Burnham and Anderson (2002) in their textbook on likelihood methods for assessing models warn against data dredging . . . But there is nothing in the evidential measures recommended by Burnham and Anderson" to pick up on this (Dienes 2008, p. 144). See also Spanos (2014).

M-S Tests and Predesignation. Don't statistical M-S tests go against the error statistician's much-ballyhooed requirement that hypotheses be predesignated? The philosopher of science Rosenkrantz says yes:

> [O]rthodox tests . . . show how to test underlying assumptions of randomness, independence and stationarity, where none of these was the predesignated object of the test (the "tested hypothesis"). And yet, astoundingly in the face of all this, orthodox statisticians are one in their condemnation of "shopping for significance," picking out significant correlations in data post hoc, or "hunting for trends. . .". It is little wonder that orthodox tests tend to be highly ambivalent on the matter of predesignation. (Rosenkrantz 1977, 204–5)

Are we hoisted by our own petards? No. This is another case where failing to disentangle a rule's *raison d'être* leads to confusion. The aim of predesignation, as with the preference for novel data, is to avoid biasing selection effects in your primary statistical inference (see Tour III). The data are remodeled to ask a different question. Strictly speaking our model assumptions are predesignated as soon as we propose a given model for statistical inference. These are the pigeonholes in the PR menu. It has never been a matter of the time – of who knew what, when – but a matter of avoiding erroneous interpretations of the data at hand. M-S tests in the error statistical methodology are deliberately designed to be independent of (or orthogonal to) the primary question at hand. The model assumptions, singly or in groups, arise as argumentative assumptions, ready to be falsified by criticism. In many cases, the inference is as close to a deductive falsification as to be wished.

Parametric tests of assumptions may themselves have assumptions, which is why judicious combinations of varied tests are called upon to ensure their overall error probabilities. Order matters: Tests of the distribution, e.g.,

Normal, Binomial, or Poisson, assume IID, so one doesn't start there. The inference in the case of an M-S test of assumptions is not a statistical inference to a *generalization*: It's explaining given data, as with explaining a "known effect," only keeping to the statistical categories of distribution, independence/dependence, and homogeneity/heterogeneity (Section 4.6). Rosenkrantz's concerns pertain to the kind of pejorative hunting for variables to include in a substantive model. That's always kept distinct from the task of M-S testing, including respecifying.

Our argument for a respecified model is a *convergent* argument: questionable conjectures along the way don't bring down the tower (section 1.2). Instead, problems ramify so that the specification finally deemed adequate has been sufficiently severely tested for the task at hand. The trends and perhaps the lags that are required to render the statistical model adequate generally cry out for a substantive explanation. It may well be that different statistical models are adequate for probing different questions.[5] Violated assumptions are responsible for a good deal of non-replication, and yet it has gone largely unattended in current replication research.

Take-away of Excursion 4. For a severe tester, a crucial part of a statistical method's objectivity (Tour I) is registering how test specifications such as sample size (Tour II) and biasing selection effects (Tour III) alter its error-probing capacities. Testing assumptions (Tour IV) is also crucial to auditing. If a probabilist measure such as a Bayes factor is taken as a gold standard for critiquing error statistical tests, significance levels and other error probabilities appear to overstate evidence – at least on certain choices of priors. From the perspective of the severe tester, it can be just the reverse. Preregistered reports are promoted to advance replication by blocking selective reporting. Thus there is a tension between preregistration and probabilist accounts that downplay error probabilities, that declare them only relevant for long runs, or tantamount to considering hidden intentions. Moreover, in the interest of promoting Bayes factors, researchers who most deserve censure are thrown a handy life preserver. Violating the LP, using the sampling distribution for inferences with the data at hand, and the importance of error probabilities form an interconnected web of severe testing. They are necessary for every one of the requirements for objectivity.

[5] When two different models capture the data adequately, they are called *reparameterizations* of each other.

Why object to applying the severity analysis by changing the null hypothesis, and doing a simple *P*-value computation? *P*-values, especially if plucked from thin air this way, are themselves in need of justification. That's a major goal of this journey. It's only by imagining we have either a best or good test or corresponding distance measure (let alone assuming we don't have to deal with lots of nuisance parameters) that substituting different null hypotheses works out.

Pre-data, we need a test with good error probabilities (as discussed in Section 3.2). That assures we avoid some worst case. Post-data we go further.

For a claim H to pass with severity requires not just that (S-1) the data accord with H, but also that (S-2) the test probably would have produced a worse fit, if H were false in specified ways. We often let the measure of accordance (in (S-1)) vary and train our critical focus on (S-2), but here it's a best test. Consider statistically insignificant results from test T+. The result "accords with" H_0, so we have (S-1), but we're wondering about (S-2): how probable is it that test T+ would have produced a result that accords *less* well with H_0 than $\boldsymbol{x_0}$ does, were H_0 false? An equivalent but perhaps more natural phrase for "a result that accords *less* well with H_0" is "a result *more discordant.*" Your choice.

Souvenir W: The Severity Interpretation of Negative Results (SIN) for Test T+

Applying our general abbreviation: SEV(test T+, outcome $\boldsymbol{x}$, inference H), we get "the severity with which $\mu \leq \mu_1$ passes test T+, with data $\boldsymbol{x}_0$":

$$\text{SEV}(\text{T+}, \text{d}(\boldsymbol{x}_0), \mu \leq \mu_1),$$

where $\mu_1 = (\mu_0 + \gamma)$, for some $\gamma \geq 0$. If it's clear which test we're discussing, we use our abbreviation: SEV($\mu \leq \mu_1$). We obtain a companion to the severity interpretation of rejection (SIR), Section 4.4, Souvenir R:

SIN *(Severity Interpretation for Negative Results)*

(a) If there is a very *low* probability that $\text{d}(\boldsymbol{x_0})$ would have been larger than it is, even if $\mu > \mu_1$, then $\mu \leq \mu_1$ passes with *low* severity: SEV($\mu \leq \mu_1$) is low.
(b) If there is a very *high* probability that $\text{d}(\boldsymbol{x}_0)$ would have been larger than it is, were $\mu > \mu_1$, then $\mu \leq \mu_1$ passes the test with *high* severity: SEV($\mu \leq \mu_1$) is high.

To break it down, in the case of a statistically insignificant result:

$$\text{SEV}(\mu \leq \mu_1) = \Pr(\text{d}(\boldsymbol{X}) > \text{d}(\boldsymbol{x}_0); \mu \leq \mu_1 \text{ false}).$$

We look at $\{d(\boldsymbol{X}) > d(\boldsymbol{x}_0)\}$ because severity directs us to consider a "worse fit" with the claim of interest. That $\mu \leq \mu_1$ is false within our model means that $\mu > \mu_1$. Thus:

$$\text{SEV}(\mu \leq \mu_1) = \Pr(d(\boldsymbol{X}) > d(\boldsymbol{x}_0); \mu > \mu_1).$$

Now $\mu > \mu_1$ is a composite hypothesis, containing all the values in excess of μ_1. How can we compute it? As with power calculations, we evaluate severity at a point $\mu_1 = (\mu_0 + \gamma)$, for some $\gamma \geq 0$, because for values $\mu \geq \mu_1$ the severity increases. So we need only to compute

$$\text{SEV}(\mu \leq \mu_1) > \Pr(d(\boldsymbol{X}) > d(\boldsymbol{x}_0); \mu = \mu_1).$$

To compute SEV we compute $\Pr(d(\boldsymbol{X}) > d(\boldsymbol{x}_0); \mu = \mu_1)$ for any μ_1 of interest. Swapping out the claims of interest (in significant and insignificant results), gives us a single criterion of a good test, severity.

Exhibit(v): Severity Curves. The severity tribes want to present severity using a standard Normal example, one where $\sigma_{\overline{X}} = 1$ (as in the water plant accident). For this illustration:

$$\text{Test T+}: H_0: \mu \leq 0 \text{ vs. } H_1: \mu > 0,\ \sigma = 10,\ n = 100, \sigma/\sqrt{n} = \sigma_{\overline{X}} = 1.$$

$$\text{If } \alpha = 0.025, \text{ we reject } H_0 \text{ iff } d(\boldsymbol{X}) \geq c_{0.025} = 1.96.$$

Suppose test T+ yields the statistically insignificant result $d(\boldsymbol{x}_0) = 1.5$. Under the alternative $d(\boldsymbol{X})$ is $N(\delta, 1)$ where $\delta = (\mu - \mu_0)/\sigma_{\overline{X}}$.

Even without identifying a discrepancy of importance ahead of time, the severity associated with various inferences can be evaluated.

The severity curves (Figure 5.6) show $d(\boldsymbol{x}_0) = 0.5$, 1, 1.5, and 1.96.

$$\text{How severely does } \mu \leq 0.5 \text{ pass the test with } \overline{X} = 1.5\ (d(\boldsymbol{x}_0) = 1.5)?$$

The easiest way to compute it is to go back to the observed $\overline{x}_0$, which would be 1.5.

$$\text{SEV}(\mu \leq 0.5) = \Pr(\overline{X} > 1.5;\ \mu = 0.5) = 0.16.$$

Here, $Z = [(1.5 - 0.5)/1] = 1$, and the area under the standard Normal distribution to the right of 1 is 0.16. Lousy. We can read it off the curve, looking at where the $d(\boldsymbol{x}) = 1.5$ curve hits the bottom-most dotted line. The severity (vertical) axis hits 0.16, and the corresponding value on the μ axis is 0.5. This could be used more generally as a discrepancy axis, as I'll show.

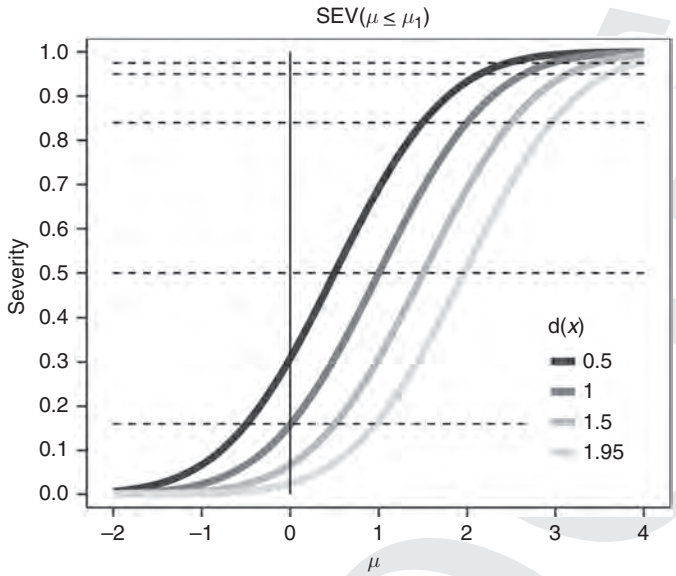

Figure 5.6 Severity curves.

We can find some discrepancy from the null that this statistically insignificant result warrants ruling out at a reasonable level – one that very probably would have produced a more significant result than was observed. The value $d(\boldsymbol{x}_0) = 1.5$ yields a severity of 0.84 for a discrepancy of 2.5. $SEV(\mu \leq 2.5) = 0.84$ with $d(\boldsymbol{x}_0) = 1.5$. Compare this to $d(\boldsymbol{x}) = 1.95$, failing to trigger the significance alarm. Now a larger upper bound is needed for severity 0.84, namely, $\mu \leq 2.95$. If we have discrepancies of interest, by setting a high power to detect them, we ensure – ahead of time – that any insignificant result entitles us to infer "it's not that high." Power against μ_1 evaluates the worst (i.e., lowest) severity for values $\mu \leq \mu_1$ for any outcome that leads to non-rejection. This test ensures any insignificant result entitles us to infer $\mu \leq 2.95$, call it $\mu \leq 3$. But we can determine discrepancies that pass with severity, post-data, without setting them at the outset. Compare four different outcomes:

$d(\boldsymbol{x}_0) = 0.5$, $SEV(\mu \leq 1.5) = 0.84$; $d(\boldsymbol{x}_0) = 1$, $SEV(\mu \leq 2) = 0.84$;

$d(\boldsymbol{x}_0) = 1.5$, $SEV(\mu \leq 2.5) = 0.84$; $d(\boldsymbol{x}_0) = 1.95$, $SEV(\mu \leq 2.95) = 0.84$.

In relation to test T+ (standard Normal): If you add $1\sigma_{\overline{X}}$ to $d(\boldsymbol{x}_0)$, the result being μ_1, then $SEV(\mu \leq \mu_1) = 0.84$.[8]

We can also use severity curves to compare the severity for a given claim, say $\mu \leq 1.5$:

$d(\boldsymbol{x}_0) = 0.5$, $SEV(\mu \leq 1.5) = 0.84$; $d(\boldsymbol{x}_0) = 1$, $SEV(\mu \leq 1.5) = 0.7$;

$d(\boldsymbol{x}_0) = 1.5$, $SEV(\mu \leq 1.5) = 0.5$; $d(\boldsymbol{x}_0) = 1.95$, $SEV(\mu \leq 1.5) = 0.3$.

Low and high benchmarks convey what is and is not licensed, and suffice for avoiding fallacies of acceptance. We can deduce SIN from the case where T+ has led to a statistically significant result, SIR. In that case, the inference that passes the test is of form $\mu > \mu_1$, where $\mu_1 = \mu_0 + \gamma$. Because $(\mu > \mu_1)$ and $(\mu \leq \mu_1)$ partition the parameter space of μ, we get $SEV(\mu > \mu_1) = 1 - SEV(\mu \leq \mu_1)$.

The more devoted amongst you will want to improve and generalize my severity curves. Some of you are staying the night at Confidence Court Inn, others at Best Bet and Breakfast. We meet at the shared lounge, Calibration. Here's a souvenir of SIR, and SIN.

Souvenir X: Power and Severity Analysis

Let's record some highlights from Tour I:

First, ordinary power analysis versus severity analysis for Test T+:

Ordinary Power Analysis: If $\Pr(d(\mathbf{X}) \geq c_\alpha; \mu_1)$ = high and the result is not significant, then it's an indication or evidence that $\mu \leq \mu_1$.

Severity Analysis: If $\Pr(d(\mathbf{X}) \geq d(\boldsymbol{x}_0); \mu_1)$ = high and the result is not significant, then it's an indication or evidence that $\mu \leq \mu_1$.

It can happen that claim $\mu \leq \mu_1$ is warranted by severity analysis but not by power analysis.

[8] • If you add $k\sigma_{\overline{X}}$ to $d(\boldsymbol{x}_0)$, $k > 0$, the result being μ_1, then $SEV(\mu \leq \mu_1)$ = area to the right of $-k$ under the standard Normal (SEV > 0.5).
• If you subtract $k\sigma_{\overline{X}}$ from $d(\boldsymbol{x}_0)$, the result being μ_1, then $SEV(\mu \leq \mu_1)$ = area to the right of k under the standard Normal (SEV ≤ 0.5).

For the general case of Test T+, you'd be adding or subtracting $k\sigma_{\overline{X}}$ to $(\mu_0 + d(\boldsymbol{x}_0)\sigma_{\overline{X}})$. We know that adding $0.85\sigma_{\overline{X}}$, $1\ \sigma_{\overline{X}}$, and $1.28\sigma_{\overline{X}}$ to the cut-off for rejection in a test T+ results in μ values against which the test has 0.8, 0.84, and 0.9 power. If you treat the observed $\overline{x}$ as if it were being contemplated as the cut-off, and add $0.85\sigma_{\overline{X}}$, $1\sigma_{\overline{X}}$, and $1.28\sigma_{\overline{X}}$, you will arrive at μ_1 values such that $SEV(\mu \leq \mu_1) = 0.8$, 0.84, and 0.9, respectively. That's because severity goes in the same direction as power for non-rejection in T+. For familiar numbers of $\sigma_{\overline{X}}$'s added/subtracted to $\overline{x} = \mu_0 + d_0\sigma_{\overline{X}}$:

Claim	$(\mu \leq \overline{x} - 1\sigma_{\overline{X}})$	$(\mu \leq \overline{x})$	$(\mu \leq \overline{x} + 1\sigma_{\overline{X}})$	$(\mu \leq \overline{x} + 1.65\sigma_{\overline{X}})$	$(\mu \leq \overline{x} + 1.98\sigma_{\overline{X}})$
SEV	0.16	0.5	0.84	0.95	0.975

Now an overview of severity for test T+: Normal testing: H_0: $\mu \leq \mu_0$ vs. H_1: $\mu > \mu_0$ with σ known. The severity reinterpretation is set out using discrepancy parameter γ. We often use μ_1 where $\mu_1 = \mu_0 + \gamma$.

> Reject H_0 (with $\boldsymbol{x}_0$) licenses inferences of the form $\mu > [\mu_0 + \gamma]$, for some $\gamma \geq 0$, but with a warning as to $\mu \leq [\mu_0 + \kappa]$, for some $\kappa \geq 0$.
>
> Non-reject H_0 (with $\boldsymbol{x}_0$) licenses inferences of the form $\mu \leq [\mu_0 + \gamma]$, for some $\gamma \geq 0$, but with a warning as to values fairly well indicated $\mu > [\mu_0 + \kappa]$, for some $\kappa \geq 0$.

The severe tester reports the attained significance levels and at least two other benchmarks: claims warranted with severity, and ones that are poorly warranted.

Talking through SIN and SIR. Let $d_0 = d(\boldsymbol{x}_0)$.

SIN *(Severity Interpretation for Negative Results)*

(a) *low*: If there is a very *low* probability that d_0 would have been larger than it is, even if $\mu > \mu_1$, then $\mu \leq \mu_1$ passes with *low* severity: SEV($\mu \leq \mu_1$) is low (i.e., your test wasn't very capable of detecting discrepancy μ_1 even if it existed, so when it's not detected, it's poor evidence of its absence).

(b) *high*: If there is a very *high* probability that d_0 would have been larger than it is, were $\mu > \mu_1$, then $\mu \leq \mu_1$ passes the test with *high* severity: SEV($\mu \leq \mu_1$) is high (i.e., your test was highly capable of detecting discrepancy μ_1 if it existed, so when it's not detected, it's a good indication of its absence).

SIR *(Severity Interpretation for Significant Results)*

If the significance level is small, it's indicative of some discrepancy from H_0, we're concerned about the magnitude:

(a) *low*: If there is a fairly high probability that d_0 would have been larger than it is, even if $\mu = \mu_1$, then d_0 is not a good indication $\mu > \mu_1$: SEV($\mu > \mu_1$) is low.[9]

(b) *high*: Here are two ways, choose your preferred:

- (b-1) If there is a very high probability that d_0 would have been smaller than it is, if $\mu \leq \mu_1$, then when you observe so large a d_0, it indicates $\mu > \mu_1$: SEV($\mu > \mu_1$) is high.

[9] A good rule of thumb to ascertain if a claim C is warranted is to think of a statistical *modus tollens* argument, and find what would occur with high probability, were claim C false.

- (b-2) If there's a very low probability that so large a d_0 would have resulted, if μ were no greater than μ_1, then d_0 indicates $\mu > \mu_1$: SEV($\mu > \mu_1$) is high.[10]

[10] For a shorthand that covers both severity and FEV for Test T+ with small significance level (Section 3.1):

(FEV/SEV): If d($\boldsymbol{x}_0$) is not statistically significant, then $\mu \leq \overline{x} + k_\varepsilon \sigma/\sqrt{n}$ passes the test T+ with severity $(1 - \varepsilon)$

(FEV/SEV): If d($\boldsymbol{x}_0$) is statistically significant, then $\mu > \overline{x} - k_\varepsilon \sigma/\sqrt{n}$ passes test T+ with severity $(1 - \varepsilon)$,

where $\Pr(d(\boldsymbol{X}) > k_\varepsilon) = \varepsilon$ (Mayo and Spanos (2006), Mayo and Cox (2006).)

in $\underline{H}$ get a high posterior, even if one or more are adequate. Perhaps by suitably restricting the space ("small worlds") this can work, but the idea of inference as continually updating goes by the board.

The open-endedness of science is essential – as pointed out by Nelder and Sprott. The severe tester agrees. Posterior probabilism, with its single probability pie, is inimical to scientific discovery. Barnard's point at the Savage Forum was, why not settle for comparative likelihoods? I think he has a point, but for error control, that limited us to predesignated hypotheses. Nelder was a Likelihoodist and there's a lot of new work that goes beyond Royall's Likelihoodism – suitable for future journeys. The error statistician still seeks an account of severe testing, and it's hard to see that comparativism can ever give that. Despite science's open-endedness, hypotheses can pass tests with high severity. Accompanying reports of poorly tested claims point the way to novel theories. Remember Neyman's modeling the variation in larvae hatched from moth eggs (Section 4.8)? As Donald Gillies (2001) stresses, "Neyman did not consider any hypotheses other than that of the Poisson distribution" (p. 366) until it was refuted by statistical tests, which stimulated developing alternatives.

> Yet it is difficult to see how all these changes in degrees of belief by Bayesian conditionalisation could have produced the solution to the problem, . . . The Bayesian mechanism seems capable of doing no more than change the statistician's degree of belief in particular values of λ [in the Poisson distribution]. (Gillies 2001, p. 367)

At the stage of inventing new models, Box had said, the Bayesian should call in frequentist tests. This is also how GTR and HEP scientists set out to extend their theories into new domains. In describing the goal of "efficient tests of hypotheses," Pearson said, if a researcher is going to have to abandon his hypothesis, he would like to do so quickly. The Bayesian, Gillies observes, might have to wait a very long time or never discover the problem (ibid., p. 368). By contrast, "The classical statisticians do not need to indulge in such toil. They can begin with any assumption (or conjecture) they like, provided only they obey the golden rule of testing it severely" (ibid., p. 376).

Souvenir Y: Axioms Are to Be Tested by You (Not Vice Versa)

Axiomatic Challenge. What do you say if you're confronted with a very authoritative-sounding challenge like this: To question classic subjective Bayesian tenets (e.g., your beliefs are captured by probability, must be betting coherent, and updated via Bayes' Rule) comes up against accepted mathematical axioms. First, recall a point from Section 2.1: You're free to use any formal deductive system, the issue will be soundness. Axioms can't run up against

empirical claims: they are formal stipulations of a system that gets meaning, and thus truth value, by interpretations. Carefully cashed out, the axioms they have in mind subtly assume your beliefs are well represented by probability, and usually that belief change follows Bayes' Theorem. If this captures your intuitions, fine, but there's no non-circular proof of this.

Empirical Studies. We skipped over a wing of the museum that is at least worth mentioning: there have been empirical studies over many years that refute the claim that people are intuitive Bayesians: "we need not pursue this debate any further, for there is now overwhelming empirical evidence that no Bayesian model fits the thoughts or actions of real scientists" (Giere 1988, p. 149). The empirical studies refer to experiments conducted since the 1960s to assess how well people obey Bayes' Theorem. These experiments, such as those performed by Daniel Kahneman, Paul Slovic, and Amos Tversky (1982), reveal substantial deviations from the Bayesian model even in simple cases where the prior probabilities are given, and even with statistically sophisticated subjects. Some of the errors may result from terminology, such as the common understanding of probability as the likelihood. I don't know if anyone has debunked the famous "Linda paradox" this way, but given the data, it's more likely that Linda's a feminist and a bank teller than that she's a bank teller, in the technical sense of "likely." Gerd Gigerenzer (1991) gives a thorough analysis showing that rephrasing the most popular probability violations frequentistly has them disappear.

> What is called in the heuristics and biases literature the "normative theory of probability" or the like is in fact a very narrow kind of neo-Bayesian view . . . (p. 86)

> . . . Since "cognitive illusions" tend to disappear in frequency judgments, it is tempting to think of the intuitive statistics of the mind as frequentist statistics. (ibid., p. 104)

While interesting in their own right, I don't regard these studies as severe tests of whether Bayesian models are a good representation for scientific inference. Why? Because in these experiments the problem is set up to be one in which the task is calculating probabilities; the test-taker is right to assume they are answerable by probabilities.

Normative Epistemology. We have been querying the supposition that what we really want for statistical inference is a probabilism. What might appear as a direct way to represent beliefs may not at all be a direct way to use probability for a normative epistemology, to determine claims that are and are not evidentially warranted. An adequate account must be able to falsify claims statistically, and in so doing it's always from demonstrated effects to hypotheses, theories, or models. Neither a posterior probability nor a Bayes

factor falsifies. Even to corroborate a real effect depends on falsifying "no effect" hypotheses. Granted, showing that you have a genuine effect is just a first step in the big picture of scientific inference. You need also to show you've correctly pinpointed causes, that you can triangulate with other ways of measuring the same quantity, and, more strongly still, that you understand a phenomenon well enough to exploit it to probe new domains. These abilities are what demarcate science and non-science (Section 2.3). Formal statistics hardly makes these assessments automatic, but we want piecemeal methods ready to serve these ends. If our language had kept to the root of probability, *probare*, to demonstrate or show how well you can put a claim to the test, and have it survive, we'd find it more natural to speak of claims being well probed rather than highly probable. Severity is not to be considered the goal of science or a sum-up of the growth of knowledge, but it has a crucial role in statistical inference.

Someone is bound to ask: Can a severity assessment be made to obey the probability axioms? If the severity for the statistical hypothesis H is high, then little problem arises in having a high degree of belief in H. But we know the axioms don't hold. Consider H: Humans will be cloned by 2030. Both H and $\sim H$ are poorly tested on current evidence. This always happens unless one of H, $\sim H$ is corroborated. Moreover, passing with low severity isn't akin to having a little bit of evidence but rather no evidence to speak of, or a poor test. What if we omitted cases of low severity due to failed audits (from violated assumptions or selection effects)? I still say no, but committed Bayesians might want to try. Since it would require the assessments to make use of sampling distributions and all that error statistics requires, it could at most be seen as a kind of probabilistic bookkeeping of inferences done in an entirely different way.

Nearly all tribes are becoming aware that today's practice isn't captured by tenets of classical probabilism. Even some subjective Bayesians, we saw, question updating by Bayes' Rule. Temporal incoherence can require a do-over. The most appealing aspects of non-subjective/default Bayesianism – a way to put in background information while allowing the data to dominate – are in tension with each other, and with updating. The gallimaufry of priors alone is an obstacle to scrutinizing the offerings. There are a few tribes where brand new foundations are being sought – that's our last port of call.

Bayesianism, in current formulations, do not falsify, although they can undergo prior redos or shifts. The Bayesian probabilist regards error probabilities as indirect because they seek a posterior; for the Bayesian falsificationist, like the severe tester, the shoe is on the other foot.

Souvenir Z: Understanding Tribal Warfare

We began this tour asking: Is there an overarching philosophy that "matches contemporary attitudes"? More important is changing attitudes. Not to encourage a switch of tribes, or even a tribal truce, but something more modest and actually achievable: to understand and get beyond the tribal warfare. To understand them, at minimum, requires grasping how the goals of probabilism differ from those of probativeness. This leads to a way of changing contemporary attitudes that is bolder and more challenging. Snapshots from the error statistical lens let you see how frequentist methods supply tools for controlling and assessing how well or poorly warranted claims are. All of the links, from data generation to modeling, to statistical inference and from there to substantive research claims, fall into place within this statistical philosophy. If this is close to being a useful way to interpret a cluster of methods, then the change in contemporary attitudes is radical: it has never been explicitly unveiled. Our journey was restricted to simple examples because those are the ones fought over in decades of statistical battles. Much more work is needed. Those grappling with applied problems are best suited to develop these ideas, and see where they may lead. I never promised, when you bought your ticket for this passage, to go beyond showing that viewing statistics as severe testing will let you get beyond the statistics wars.

6.7 Farewell Keepsake

Despite the eclecticism of statistical practice, conflicting views about the roles of probability and the nature of statistical inference – holdovers from long-standing frequentist–Bayesian battles – still simmer below the surface of today's debates. Reluctance to reopen wounds from old battles has allowed them to fester. To assume all we need is an agreement on numbers – even if they're measuring different things – leads to statistical schizophrenia. Rival conceptions of the nature of statistical inference show up unannounced in the problems of scientific integrity, irreproducibility, and questionable research practices, and in proposed methodological reforms. If you don't understand the assumptions behind proposed reforms, their ramifications for statistical practice remain hidden from you.

Rival standards reflect a tension between using probability (a) to constrain the probability that a method avoids erroneously interpreting data in a series of

Souvenirs