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

applications (*performance*), and (b) to assign degrees of support, confirmation, or plausibility to hypotheses (*probabilism*). We set sail on our journey with an informal tool for telling what's true about statistical inference: If little if anything has been done to rule out flaws in taking data as evidence for a claim, then that claim has not passed a *severe test*. From this minimal severe-testing requirement, we develop a statistical philosophy that goes beyond probabilism and performance. The goals of the severe tester (*probativism*) arise in contexts sufficiently different from those of probabilism that you are free to hold both, for distinct aims (Section 1.2). For statistical inference in science, it is severity we seek. A claim passes with severity only to the extent that it is subjected to, and passes, a test that it probably would have failed, if false. Viewing statistical inference as severe testing alters long-held conceptions of what's required for an adequate account of statistical inference in science. In this view, a *normative statistical epistemology* – an account of what's warranted to infer – must be:

- directly altered by biasing selection effects
- able to falsify claims statistically
- able to test statistical model assumptions
- able to block inferences that violate minimal severity

These overlapping and interrelated requirements are disinterred over the course of our travels. This final keepsake collects a cluster of familiar criticisms of error statistical methods. They are not intended to replace the detailed arguments, pro and con, within; here we cut to the chase, generally keeping to the language of critics. Given our conception of evidence, we retain testing language even when the statistical inference is an estimation, prediction, or proposed answer to a question. The concept of severe testing is sufficiently general to apply to any of the methods now in use. It follows that a variety of statistical methods can serve to advance the severity goal, and that they can, in principle, find their foundations in an error statistical philosophy. However, each requires supplements and reformulations to be relevant to real-world learning. Good science does not turn on adopting any formal tool, and yet the statistics wars often focus on whether to use one type of test (or estimation, or model selection) or another. Meta-researchers charged with instigating reforms do not agree, but the foundational basis for the disagreement is left unattended. It is no wonder some see the statistics wars as proxy wars between competing tribe leaders, each keen to advance one or another tool, rather than about how to do better science. Leading minds are drawn into inconsequential battles, e.g., whether to use a pre-specified cut-off of 0.025 or 0.0025 – when in fact good inference is not about cut-offs altogether but about a series of small-scale steps in collecting, modeling and analyzing data that work together to

find things out. Still, we need to get beyond the statistics wars in their present form. By viewing a contentious battle in terms of a difference in goals – finding highly probable versus highly well probed hypotheses – readers can see why leaders of rival tribes often talk past each other. To be clear, the standpoints underlying the following criticisms are open to debate; we're far from claiming to do away with them. What should be done away with is rehearsing the same criticisms ad nauseum. Only then can we hear the voices of those calling for an honest standpoint about responsible science.

**1. NHST Licenses Abuses.** First, there's the cluster of criticisms directed at an abusive NHST animal: NHSTs infer from a single *P*-value below an arbitrary cut-off to evidence for a research claim, and they encourage *P*-hacking, fishing, and other selection effects. The reply: this ignores crucial requirements set by Fisher and other founders: isolated significant results are poor evidence of a genuine effect and statistical significance doesn't warrant substantive, (e.g., causal) inferences. Moreover, selective reporting invalidates error probabilities. Some argue significance tests are un-Popperian because the higher the sample size, the easier to infer one's research hypothesis. It's true that with a sufficiently high sample size any discrepancy from a null hypothesis has a high probability of being detected, but statistical significance does not license inferring a research claim *H*. Unless *H*'s errors have been well probed by merely finding a small P-value, *H* passes an extremely insevere test. No mountains out of molehills (Sections 4.3 and 5.1). Enlightened users of statistical tests have rejected the cookbook, dichotomous NHST, long lampooned: such criticisms are behind the times. When well-intentioned aims of replication research are linked to these retreads, it only hurts the cause. One doesn't need a sharp dichotomy to identify rather lousy tests – a main goal for a severe tester. Granted, policy-making contexts may require cut-offs, as do behavioristic setups. But in those contexts, a test's error probabilities measure overall error control, and are not generally used to assess well-testedness. Even there, users need not fall into the NHST traps (Section 2.5). While attention to banning terms is the least productive aspect of the statistics wars, since NHST is not used by Fisher or N-P, let's give the caricature its due and drop the NHST acronym; "statistical tests" or "error statistical tests" will do. Simple significance tests are a small part of a conglomeration of error statistical methods.

**2. Against Error Probabilities: Inference Should Obey the LP.** A common criticism is that error statistical methods use error probabilities post-data. Facets of the same argument take the form of criticizing methods that take account of outcomes other than the one observed, the sampling distribution, the sample space, and researcher "intentions" in optional stopping. It will also be charged that they violate the Likelihood Principle (LP), and are incoherent

(Sections 1.5, 4.6, and 6.6). From the perspective of a logic of induction, considering what other outputs might have resulted seems irrelevant. If there's anything we learn from the consequences of biasing selection effects it is that such logics come up short: data do not speak for themselves. To regard the sampling distribution irrelevant is to render error probabilities irrelevant, and error probability control is necessary (though not sufficient) for severe testing. The problem with cherry picking, hunting for significance, and a host of biasing selection effects – the main source of handwringing behind the statistics crisis in science – is they wreak havoc with a method's error probabilities. It becomes easy to arrive at findings that have not been severely tested. Ask yourself: what bothers you when cherry pickers selectively report favorable findings, and then claim to have good evidence of an effect? You're not concerned that making a habit out of this would yield poor long-run performance. What bothers you, and rightly so, is they haven't done a good job in ruling out spurious findings in the case at hand. The severity requirement explains this evidential standpoint. You can't count on being rescued by the implausibility of cherry-picked claims. It's essential to be able to say that, a claim is plausible but horribly tested by these data.

There is a tension between popular calls for preregistration – arguably, one of the most promising ways to boost replication – and accounts that downplay error probabilities (Souvenir G, Section 4.6). The critical reader of a registered report, post-data, looks at the probability that one or another hypothesis, stopping point, choice of grouping variables, and so on, could have led to a false positive–in effect, she looks at the sampling distribution even without a formal error probability computation. We obtain a rationale never made clear by users of significance tests or confidence intervals as to the relevance of error probabilities in the case at hand. Ironically, those who promote methodologies that reject error probabilities are forced to beat around the bush rather than directly upbraid researchers for committing QRPs that damage error probabilities. They give the guilty party a life raft (Section 4.6). If you're in the market for a method that directly registers flexibilities, p-hacking, outcome-switching and all the rest, then you want one that picks up on a method's error probing capacities.

Granted the rejection of error probabilities is often tied to presupposing they only serve for behavioristic or performance goals. The severe tester breaks out of the behavioristic prison from which this charge arises. Error probabilities are used to assess and control how severely tested claims are.

**3. Fisher and N-P Form an Inconsistent Hybrid.** We debunk a related charge: that Fisherian and N-P methods are an incompatible hybrid and

should be kept segregated. They offer distinct tools under the overarching umbrella of error statistics, along with other methods that employ a method's sampling distribution for inference (confidence intervals, N-P and Fisherian tests, resampling, randomization). While in some quarters the incompatibilist charge is viewed as merely calling attention to the very real, and generally pathological, in-fighting between Fisher and Neyman, it's not innocuous (Section 5.8). Incompatibilist or segregationist positions keep people adhering to caricatures of both approaches where Fisherians can't use power, and N-P testers can't use *P*-values. Some charge that Fisherian *P*-values are not error probabilities because Fisher wanted an evidential, not a performance, interpretation. In fact, N-P and Fisher used *P*-values in both ways. Reporting the actual *P*-value is recommended by N-P, Lehmann, and others (Section 3.5). It is a post-data error probability. To paraphrase Cox and Hinkley (1974, p. 66), it's the probability we'd mistakenly report evidence against $H_0$ were we to regard the data as just decisive for issuing such a report. The charge that N-P tests preclude distinguishing highly significant from just significant results is challenged on historical, statistical, and philosophical grounds. Most importantly, even if their founders were die-hard behaviorists it doesn't stop us from giving them an inferential construal (Section 3.3). Personality labels should be dropped. It's time we took responsibility for interpreting tests. It's the methods, stupid.

The most consequential variant under the banner of "*P*-values aren't error probabilities" goes further, and redefines error probabilities to refer to one or another variant on a posterior probability of hypotheses. It is no wonder the disputants so often talk past each other. Once we pull back the curtain on this equivocal use of "error probability," with the help of subscripts, it is apparent that all these arguments must be revisited (Sections 3.5 and 3.6). Even where it may be argued the critics haven't left the frequentist station, the train takes us to probabilities on hypotheses – requiring priors.

**4. *P*-values Overstate Evidence Against the Null Hypothesis.** This is a very common charge (Sections 4.5 and 4.6). What's often meant is that the *P*-value can be smaller than a posterior probability on a point null hypothesis $H_0$, based on a lump prior (often 0.5) on $H_0$. Why take the context that leads to the criticism – one where a point null value has a high prior probability – as typical? It has been questioned by Bayesians and frequentists alike (some even say all such nulls are false). Moreover, *P*-values can also agree with the posterior: in short, there is wide latitude for Bayesian assignments. Other variations on the criticism judge *P*-values on the standard of a comparative probabilism (Bayes factors, likelihood ratios). We do not discount any of these criticisms simply because they hinge on taking probabilist measures as an appropriate

standard for judging an error probability. The critics think their standard is relevant, so we go with them as far as possible, until inseverity hits. It does hit. A small *P*-value appears to exaggerate the evidence from the standpoint of probabilism, while from that of performance or severity, it can be the other way around (Sections 4.4, 4.5, 5.2, and 5.6). Without unearthing the presuppositions of rival tribes, users may operate with inconsistent recommendations. Finally, that the charge it is too easy to obtain small *P*-values is belied by how difficult it is to replicate low *P*-values – particularly with preregistration (replication paradox): the problem isn't *P*-values but selective reporting and other abuses (Section 4.6).

**5. Inference Should Be Comparative.** Statistical significance tests, it may be charged, are not real accounts of evidence because they are not comparative. A comparative assessment takes the form of hypothesis $H_1$ is comparatively better supported, believed (or otherwise favored) than $H_0$. Comparativist accounts do not say there's evidence against one hypothesis, nor for the other: neither may be warranted by the data. Nor do they statistically falsify a model or claim as required by a normative epistemology. So why be a comparativist? Comparativism is an appealing way to avoid the dreaded catchall factor required by a posterior probabilism: all hypotheses that could explain the data (Section 6.4).

What of the problems that are thought to confront the non-comparativist who is not a probabilist but a tester? (The criticism is usually limited to Fisherian tests, but N-P tests aren't comparative in the sense being used here.) The problems fall under fallacies of rejection (Section 2.5). Notably, it is assumed a Fisherian test permits an inference from reject $H_0$ to an alternative $H_1$ further from the data than is $H_0$. Such an inference is barred as having low severity (Section 4.3). We do not deny the informativeness of a comparativist measure within an overall severe testing rationale.[9] We agree with Fisher in denying there's just one way to use probability in statistical inquiry (he used likelihoods, P-values, and fiducial intervals). Our point is that the criticism of significance tests for not being comparativist is based on a straw man. Actually, it's their ability to test accordance with a single model or hypothesis that makes simple significance tests so valuable for testing assumptions, leading to (6).

**6. Accounts Should Test Model Assumptions.** Statistical tests are sometimes criticized as assuming the correctness of their statistical models. In fact, the battery of diagnostic and M-S tests are error statistical. When it comes to testing model assumptions – an important part of auditing – it's to significance

[9] We would need predesignation of hypotheses (and/or other restrictions) if there is to be error control.

tests, or the analogous graphical analysis, to which people turn (Sections 4.9, 4.11, and 6.7). Sampling distributions are the key. Also under the heading of auditing are: checking for violated assumptions in linking statistical to substantive inferences, and illicit error probabilities due to biasing selection effects. Reports about what has been poorly audited, far from admissions of weakness, should become the most interesting parts of research reports, at least if done in the severity spirit. They are what afford building a cumulative repertoire of errors, pointing to rival theories to probe. Even domains that lack full-blown theories have theories of mistakes and fallibilities. These suffice to falsify inquiries or even entire measurement procedures, long assumed valid.

**7. Inference Should Report Effect Sizes.** Pre-data error probabilities and *P*-values do not report effect sizes or discrepancies – their major weakness. We avoid this criticism by interpreting statistically significant results, or "reject $H_0$," in terms of indications of a discrepancy $\gamma$ from $H_0$. In test T+: [Normal testing: $H_0$: $\mu \leq \mu_0$ vs. $H_1$: $\mu > \mu_0$], reject $H_0$ licenses inferences of the form: $\mu > [\mu_0 + \gamma]$; non-reject $H_0$, to inferences of the form: $\mu \leq [\mu_0 + \gamma]$. A report of discrepancies poorly warranted is also given (Section 3.1). The severity assessment takes account of the particular outcome $\boldsymbol{x}_0$ (Souvenir W). In some cases, a qualitative assessment suffices, for instance, that there's no real effect.

The desire for an effect size interpretation is behind a family feud among frequentists, urging that tests be replaced by confidence intervals (CIs). In fact there's a duality between CIs and tests: the parameter values within the $(1 - \alpha)$ CI are those that are not rejectable by the corresponding test at level $\alpha$ (Section 3.7). Severity seamlessly connects tests and CIs. A core idea is arguing from the capabilities of methods to what may be inferred, much as we argue from the capabilities of a key to open a door to the shape of the key's teeth.[10] In statistical contexts, a method's capabilities are represented by its probabilities of avoiding erroneous interpretations of data (Section 2.7).

The "CIs only" battlers have encouraged the use of CIs as supplements to tests, which is good; but there have been casualties. They often promulgate the perception that the only alternative to standard CIs is the abusive NHST animal, with cookbook, binary thinking. The most vociferous among critics in group (1) may well be waging a proxy war for replacing tests with CIs. Viewing statistical inference as severe testing leads to improvements that the CI advocate should welcome (Sections 3.7, 4.3, and 5.5): (a) instead of a fixed confidence level, usually 95%, several levels are needed, as with confidence distributions CDs. (b) We move away from the dichotomy of parameter

[10] I allude to a pin and tumbler lock.

values being inside or outside a CI estimate; points within a CI correspond to distinct claims, and get different severity assignments. (c) CIs receive an inferential rather than a performance "coverage probability" justification. (d) Fallacies and chestnuts of confidence intervals (vacuous intervals) are avoided.

**8. Inference Should Provide Posterior Probabilities**, final degrees of support, belief, probability. Wars often revolve around assuming what is wanted is a posterior probabilism of some sort. Informally, we might say we want probable claims. But when we examine each of the ways this could be attained with formal probablity, the desirability for science evanesces. The onus is on those who hold that what we want are probabilities on hypotheses to show, for existing ways of obtaining them, why.[11] (Note that it's not provided by comparative accounts, Bayes factors, likelihood ratios or model selections.) The most prevalent Bayesian accounts are default/non-subjective, but there is no agreement on suitable priors. The priors are mere formal devices for obtaining a posterior so that the data dominate in some sense. The main assets of the Bayesian picture – a coherent way to represent and update beliefs – go by the board.

Error statistical methods, deemed indirect for probabilism, are direct for severe probing and falsification. Severe testers do not view scientists as seeking highly probable hypotheses, but learning which interpretations of data are well and poorly tested. Of course we want well-warranted claims, but arriving at them does not presuppose a single probability pie with its requirements of exhaustiveness: science must be open ended. We want methods that efficiently find falsity, not ones that are based on updating values for parameters in an existing model. We want to infer local variants of theories piecemeal, falsify others, and be free to launch a probe of any hypothesis which we can subject to severe testing. If other ways to falsify satisfy error statistical requirements, then they are happily in sync with us.

**9. Severe Testing Is Not All You Do in Inquiry.** Agreed. I have used a neutral word "warranted" to mean justified, adding "with severity" when appropriate. There's a distinctive twist that goes with severely warranting claims – some prefer to say "beliefs," and you could substitute throughout if you wish. It is this twist that makes it possible to have your probabilist cake, and probativism too – each for distinct contexts. The severe testing assessment is not measuring how strong your belief in $H$ is but how well you can show why $H$ ought to be believed. It is relevant when the aim is to know *why* claims pass (or fail) the tests they do. View the error statistical notions as

[11] Some will use a $P$-value as a degree of inconsistency with a null hypothesis.

a picturesque representation of the real life, flesh and blood, capability or incapability to put to rest reasonable skeptical challenges. It's in the spirit of Fisher's requiring you know how to bring about results that would rarely fail to corroborate *H*. It's not merely knowing, but *showing* you're prepared, or would be, to tackle skeptical challenges. I'm using "you" but it could be a group or a machine. It's just not all that you do in inquiry. I admitted at the outset that we do not always want to find things out. If your goal is belief probabilism, or you're in a context where the aim is to assign direct probabilities to events (a deductive task), then you are better off recognizing the differences than trying to unify or reconcile. Let me be clear, severe testing isn't reserved for cases of strong evidence; it is operative at every stage of inquiry, but even more so in early stages – where skepticism is greatest. The severity demand is what we naturally want as consumers of statistics, namely, grounds that reports would very probably have revealed flaws of relevance when they're present. To pass tests with severity gives strong evidence, yes, but most of the time it's to learn that much less than was thought or hoped has passed. Showing (with severity!) that a study was poorly run is important in its own right, even if done semi-formally. Better still is to pinpoint a flaw that's been overlooked.

Our journey has taken you far beyond the hackneyed statistical battles that make up much of today's statistics wars. I've chosen to focus on some of them in your final "keepsake" because, if you have to refight them, you can begin from the places we've reached. These criticisms can no longer be blithely put forward as having weight without wrestling with the underlying presuppositions and challenges about evidence and inference. You might say that the criticisms have force against garden-variety treatments of error statistical methods, that I've changed things by adding an explicit severe testing philosophy. I'll happily concede this, but that is the whole reason for taking this journey. You needn't accept this statistical philosophy to use it to peel back the layers of the statistics wars; you will then be beyond them. It's time.

**Live (Final) Exhibit. What Does the Severe Tester Say About Positions 1–9?** What do you say?