Tour II Pragmatic and Error Statistical Bayesians

6.5 Pragmatic Bayesians

> The protracted battle for the foundations of statistics, joined vociferously by Fisher, Jeffreys, Neyman, Savage and many disciples, has been deeply illuminating, but it has left statistics without a philosophy that matches contemporary attitudes. (Robert Kass 2011, p. 1)

Is there a philosophy that "matches contemporary attitudes"? Worried that "our textbook explanations have not caught up with the eclecticism of statistical practice," Robert Kass puts forward a statistical pragmatism "as a foundation for inference" (ibid.) as now practiced. It reflects the current disinclination to embrace subjective Bayes, skirts the rigid behavioral decision model of frequentist inference, and hangs on to a kind of frequentist calibration. "Subjectivism was never satisfying as a logical framework: an important purpose of the scientific enterprise is to go beyond personal decision-making" (ibid., p. 6). Nevertheless, Kass thinks "it became clear, especially from the arguments of Savage . . . the only solid foundation for Bayesianism is subjective" (ibid., p. 7). Statistical pragmatism pulls us out of that "solipsistic quagmire" (ibid.). Statistical pragmatism, says Kass, holds that confidence intervals, statistical significance, and posterior probability are all valuable tools. So long as we recognize that our statistical models exist in an abstract theoretical world, Kass avers, we can safely employ methods from either tribe, retaining unobjectionable common denominators: frequencies without long runs and Bayesianism without the subjectivity.

> The advantage of the pragmatic framework is that it considers frequentist and Bayesian inference to be equally respectable and allows us to have a consistent interpretation, without feeling as if we must have split personalities in order to be competent statisticians. (ibid., p. 7)

Kass offers a valuable analysis of the conundrums of members of today's eclectic and non-subjective/default Bayesian tribes. We want to expose some hidden layers only visible after peeling away a more familiar mindset. Do we escape split personality? Or has he, perhaps inadvertently, explained the necessity for a degree of schizophrenia in current electic practice? That's where today's journey begins.

Kass and the Pragmatists

Bayes–frequentist agreement is typically closer with confidence intervals rather than tests. Kass (2011) uses a "paradigm case of confidence and posterior intervals for a Normal mean based on a sample of size n [49], with the standard deviation [1] being known" (ibid., pp. 3–4). Both the frequentist and the Bayesian arrive at the same interval, but he'll supply "mildly altered interpretations of frequentist and Bayesian inference" (ibid., p. 3).

We assume $X_1, X_2, \ldots, X_{49}$ are IID random variables from $N(\mu, 1)$. The 0.95 two-sided confidence interval estimator, we know is:

$$\mu = \overline{X} \pm Z_{0.025}(\sigma/\sqrt{n}).$$

Following Kass, take $Z_{0.025}$ to be 2 (instead of the more accurate 1.96):

$$\mu = (\overline{X} - 2/7, \overline{X} + 2/7).$$

We observe $\overline{x} = 10.2$. Plugging in 1/7 = 0.14, 2/7 = 0.28, the particular interval estimate I is

$$(10.2 - 2/7, 10.2 + 2/7) = [9.92, 10.48].$$

He contrasts the usual frequentist interpretation with the pragmatic one he recommends:

FREQUENTIST INTERPRETATION... Under the assumptions above, if we were to draw infinitely many random samples from a $N(\mu,1)$ distribution, 95% of the corresponding confidence intervals $(\overline{X} - 2/7, \ \overline{X} + 2/7)$ would cover μ. (ibid., p. 4)

PRAGMATIC INTERPRETION... If we were to draw a random sample according to the assumptions above, the resulting confidence interval $(\overline{X} - 2/7, \ \overline{X} + 2/7)$ would have probability 0.95 of covering μ. Because the random sample lives in the theoretical world, this is a theoretical statement. Nonetheless, substituting $\overline{X} = \overline{x}$ ... we obtain the interval *I*, and are able to draw useful conclusions as long as our theoretical world is aligned well with the real world that produced the data. (ibid.)

Kass's pragmatic Bayesian treatment runs parallel to his construal of the frequentist, except that the Bayesian also assumes a prior distribution for parameter μ. It is given its own mean μ_0 and variance τ^2.

BAYESIAN ASSUMPTIONS: Suppose $X_1, X_2, \ldots X_n$ form a random sample from a $N(\mu, 1)$ distribution and the prior distribution of μ is $N(\mu_0, \tau^2)$, with $\tau^2 \gg 1/49$ and $49\tau^2 \gg |\mu_0|$. (ibid.)

This conjugate prior might be invoked as what Berger called a "prior completion strategy" or simply a choice of a prior enabling the data to sufficiently dominate. The results:

$$\text{Posterior mean: } \overline{\mu} = \frac{\tau^2}{1/49 + \tau^2} 10.2 + \frac{1/49}{1/49 + \tau^2} \mu_0,$$

$$\text{Posterior variance: } v = \left(49 + \frac{1}{\tau^2}\right)^{-1}.$$

Given the stipulations that $\tau^2 \gg 1/49$ and $49\tau^2 \gg |\mu_0|$, we get approximately the same interval as the frequentist, only it gets a posterior probability not just a coverage performance of 0.95.

> BAYESIAN INTERPRETATION. . . Under the assumptions above, the probability that μ is in the interval I is 0.95.
>
> PRAGMATIC INTERPRETATION. . . If the data were a random sample [as described and $\overline{x} = 10.2$], and if the assumptions above were to hold, then the probability that μ is in the interval I would be 0.95. This refers to a hypothetical value $\overline{x}$ of the random variable $\overline{X}$, and because $\overline{X}$ lives in the theoretical world the statement remains theoretical. Nonetheless, we are able to draw useful conclusions from the data as long as our theoretical world is aligned well with the real world that produced the data. (ibid., p. 4)

We get an agreement on numbers and we can cross over the Bayesian–frequentist bridge with aplomb. Assuming, that is, we don't notice the crocodiles in the water waiting to bite us!

Analysis. First off, the frequentist readily agrees with Kass that "long-run frequencies may be regarded as consequences of the Law of Large Numbers rather than as part of the definition of probability or confidence" (p. 2). The probability is methodological: the proportion of correct estimates produced in hypothetical applications is 0.95. The long run need only be long enough to see the pattern emerge and is hypothetical. One can simulate the statistical mechanism associated with the model to produce realizations of the process on a computer, today, not in a long run (Spanos 2013c).

According to Kass (2011), "the commonality between frequentist and Bayesian inferences is the use of theoretical assumptions . . ." (p. 4). The prior becomes part of the model in the Bayesian case. For *both* cases, "When we use a statistical model to make a statistical inference we implicitly assert that the variation exhibited by data is captured reasonably well by the statistical model" (p. 2). Everything is left as a subjunctive conditional: if the "theoretical world is aligned well with the real world that produced the data"

(p. 4) then such and such follows. "Perhaps we might even say that most practitioners are subjunctivists" (p. 7).

Would either frequentists or Bayesians be content with this? The frequentist error statistician would not, because there's no detachment of an inference. It remains a subjunctive or 'would be' claim that is entirely deductive, not ampliative. Moreover, she insists on checking the adequacy of the model. I doubt the Bayesian would be satisfied with life as a subjunctivist either. The payoff for the extra complexity in positing a prior is the ability to detach the probability that μ is in (10.2 − 2/7, 10.2+ 2/7) is 0.95. Thus, locating the common frequentist/Bayesian ground in assuming the "theoretical world is aligned well with the real world that produced the data" doesn't get us very far, and it keeps key differences in goals under wraps. We can hear some Bayesian tribe members grumbling at the very assumption that they're modeling "the real world that produced the data" (p. 4).

A Deeper Analysis. We have developed enough muscle from our workouts to peel back some layers here. It only comes near the end of his paper, where Kass tells us what he says to his class (citing Brown and Kass 2009). I'll number his points, the highlight being (3):

> (1) I explicitly distinguish the use of probability to describe variation and to express knowledge. . . . [These] are sometimes considered to involve two different kinds of probability . . . 'aleatory probability' and 'epistemic probability' [the latter in the sense of] quantified belief . . . .
>
> (2) Bayesians merge these, applying the laws of probability to go from quantitative description to quantified belief (ibid., p. 5).
>
> (3) But in every form of statistical inference aleatory probability is used, somehow, to make epistemic statements (ibid., pp. 5–6).

What do the frequentist and Bayesian say to (1)–(3)? I can well imagine all tribes mounting resistance. Let's assume for the moment the model–theory match entitles the detachment. Then we can revisit the frequentist and Bayesian inferences.

What does the frequentist do to get her epistemological claim according to Kass? Since the probability the estimator yields true estimates is 0.95 (performance), Kass's frequentist can assign probability 0.95 to the estimate. But this is a fallacy. Worse, it's to be understood as degree of belief (probabilism) rather than confidence (in the performance sense). If a frequentist is not to be robbed of a notion of statistical inference, "epistemic" couldn't be limited to posterior probabilism. An informal notion of probability can work, but that's still to rob probability from playing its intended frequentist role.

What about the pragmatic Bayesian? The pragmatic Bayesian infers 'I'm 95% sure that μ is in (10.2 – 2/7, 10.2 + 2/7).' The model, Kass says, gives the variability (here of μ), and "Bayesians merge these"(p. 5), variability with belief. That's what the theory–real-world match *means* for the Bayesian. One might rightly wonder if the permission to merge isn't tantamount to blessing the split personality we are to avoid.

Suppose you are a student sitting in Professor Kass's class. Probability as representing random variability is quite different from its expression as degree of belief, or uncertainty of knowledge, Professor Kass begins in (1). Nevertheless, Kass will tell you how to juggle them. If you're a pragmatic frequentist, you get your inferential claim by slyly misinterpreting the confidence coefficient as a (degree of belief) probability on the estimate. If you're a pragmatic Bayesian, however, you merge these. Aha, relief. But won't you wonder at the rationale? Kass might say 'in the cases where the numbers match, viewing probability as both variability and belief is unproblematic' for a Bayesian. What about where Bayesians and frequentist numbers don't match? Then "statistical pragmatism is agnostic" (p. 7). In such cases, Kass avers, "procedures should be judged according to their performance under . . . relevant real-world variation" (ibid.). But if the probabilistic assessment doesn't supply performance measures (as in the case of a mismatch), where do they come from?

Bayesians are likely to bristle at the idea of adjudicating disagreement by appeal to frequentist performance; whereas a frequentist like Fraser (2011) argues that it's misleading to even use "probability" if it doesn't have the performance sense of confidence. Christian Robert (commenting on Fraser) avers: "the Bayesian perspective on confidence (or credible) regions and statements does not claim 'correct coverage' from a frequentist viewpoint since it is articulated in terms of the parameters" (Robert 2011, p. 317). He suggests, "the chance identity occurring for location parameters is a coincidence" (ibid.), which raises doubts about a genuine consilience even in the case of frequentist matching. Even where they match, they mean different things.

> Frequentist error probabilities relate to the sampling distribution, where we consider hypothetically different outcomes that could have occurred in investigating this one system of interest. The Bayesian allusion to frequentist 'matching' refers to the fixed data and considers frequencies over different systems (that could be investigated by a model like the one at hand). (Cox and Mayo 2010, p. 302)

These may be entirely different hypotheses, even in different fields, in reasoning about this particular *H*. A reply might be that, when there's matching, the frequentist uses error probability$_1$; the Bayesian uses error probability$_2$ (Section 3.6).

The frequentist may welcome adjudication by performance, so she feels she's doing well by Kass. But she still must carve out a kosher epistemological construal, keeping only to frequency. Now we get to the crux of our entire journey, hinted at way back in Excursion 1 (Souvenir D), "we may avoid the need for a different version of probability . . . by assessing the performance of proposed methods under hypothetical repetition" (Reid and Cox 2015, p. 295). But how? The common answer is for the error probability of the method to rub off on a particular inference in the form of confidence. I grant this is an evidential or epistemological use of probability.[1] Yet the severe tester isn't quite happy with it. It's still too performance oriented: We detach the inference, and the performance measure qualifies it by the trustworthiness of the proceeding. Our assessment of well-testedness is different. It's the sampling distribution of the given experiment that informs us of the capabilities of the method to have unearthed erroneous interpretations of data. From here, we reason, counterfactually, to what is well and poorly warranted. What justifies this reasoning? The severity requirements, weak and strong.[2]

We've often enough visited the severity interpretation of confidence intervals (Sections 3.7 and 4.3). Kass's 0.95 interval estimate is (9.92, 10.48). There's a good indication that $\mu > 9.92$ because if μ were 9.92 (or lower), we very frequently would have gotten a smaller value of $\overline{X}$ than we did. There's a poor indication that $\mu \geq 10.2$ because we'd frequently observe values even larger than we did in a world where $\mu < 10.2$, indeed, we'd expect this around 50% of the time. The evidence becomes indicative of $\mu < \mu'$ as we move from 10.2 toward the upper bound 10.48. What would occur in hypothetical repetitions, under various claims about parameters, is not for future assurance in using the rule, but to understand what caused the events on which this inference is based. If a method is incapable of reflecting changes in the data-generating source, its inferences are criticized on grounds of severity. It's possible that no parameter values are ruled out with reasonable severity, and we are left with the entire interval.

What then of Kass's attempt at peace? It may well describe a large tribe of "contemporary attitudes." As Larry Wasserman points out, if you ask people if a 95% CI means 95% of the time the method used gets it right, they will say yes. If you ask if 0.95 describes their belief in a resulting estimate, they may say yes too (2012c). But Kass was looking for respite from conceptual confusion.

[1] We know performance is necessary but not sufficient for severity, nor for confidence distributions or fiducial inference, but here we imagine we have got the *relevant* error probability.

[2] There's no need for the philosopher's appeal to things like closest possible worlds to use counter factuals either.

Maybe all Kass is saying is that in the special case of frequentist matching priors, the split personality scarcely registers, and no one will notice.

A philosophy limited to frequentist matching results would very much restrict the Bayesian edifice. Notice we've been keeping to the simple examples on which nearly all statistics battles are based. Here matching is at least in principle possible. Fraser (2011, p. 299) shows that when we move away from examples on "location," as with mean μ, the matching goes by the board. We cannot be accused of looking to examples where the frequentist and Bayesian numbers necessarily diverge, no "gotcha" examples from our side. Nevertheless, they still diverge because of differences in meaning and goals.

Optional Stopping and Bayesian Intervals

Disagreement on numbers can also be due to different attitudes toward gambits that alter error probabilities but not likelihoods. Way back in Section 1.5, we illustrated a two-sided Normal test i.e., H_0: $\mu = 0$ vs. H_1: $\mu \neq 0$, $\sigma = 1$, with a rule that keeps sampling until H_0 is rejected. At the 1962 Savage Forum, Armitage needled Savage that the same thing happens with Bayesian methods:

> The departure of the mean by two standard errors corresponds to the ordinary five per cent level. It also corresponds to the null hypothesis being at the five per cent point of the posterior distribution. (Armitage 1962, p. 72)

The identical point can be framed in terms of the corresponding 95% confidence interval method. I follow Berger and Wolpert (1988, pp. 80–1).

Suppose a default/non-subjective Bayesian [they call him an "objective conditionalist"] states that with a fixed sample size n, he would use the interval

$$\mu = \overline{x} \pm 1.96\sigma/\sqrt{n}$$

He "would not interpret confidence in the frequency sense" but instead would "use a posterior Bayesian viewpoint with the non-informative prior density $\pi(\theta) = 1$, which leads to a $N(\overline{x}_n, 1/\sqrt{n})$ posterior" [given the variance is 1] (ibid., p. 80).

Consider the rule: keep sampling until the 95% confidence interval excludes 0.

Berger and Wolpert concede that the Bayesian "being bound to ignore the stopping rule will still use [the above interval] as his confidence interval, but this can *never* contain zero" (ibid., p. 81). The experimenter using this stopping rule "has thus succeeded in getting the [Bayesian] conditionalist to perceive that $\theta \neq 0$, and has done so honestly" (ibid.). This is so despite the Bayesian interval assigning a probability of 0.95 to the interval estimate. "The 'misleading,' however, is solely from a frequentist viewpoint, and will not be of concern to a conditionalist" (ibid.). Why are they unconcerned?

It's hard to pin down their response; they go on to other examples.[3] I take it they are unconcerned because they are not in the business of computing frequentist error probabilities. From their perspective, taking into account the stopping rule is tantamount to taking account of the experimenter's intentions (when to stop). Moreover, from the perspective of what the agent believes, Berger and Wolpert explain, he is not *really* being misled. They further suggest we should trust our intuitions about the Likelihood Principle in simple situations "rather than in extremely complex situations such as" with our stopping rule (ibid., p. 83).

As they surmise, this won't "satisfy a frequentist's violated intuition" (ibid.). It kills her linchpin for critically interpreting the data.[4]

It isn't that the stopping rule problem is such a big deal; but it's routinely given as an exemplar *in favor* of ignoring the sampling distribution, and Berger and Wolpert (1988) is the standard to which Bayesian texts refer. Ironically, as Roderick Little observes: "This example is cited as a counterexample by both Bayesians and frequentists! If we statisticians can't agree which theory this example is counter to, what is a clinician to make of this debate?" (Little 2006, p. 215). But by 2006, Berger embraces default priors that are model dependent, "leading to violations of basic principles, such as the likelihood principle and the stopping rule principle" (Berger 2006, p. 394). This is all part of the abandonment of Bayesian foundations. Still, he admits to having "trouble with saying that good frequentist coverage is a necessary requirement" (ibid., p. 463). Conditioning on the data, he says, is more important. "Since the calibrated Bayes inference is Bayesian, there are no penalties for peeking. . ." (Little 2006, p. 220). The disparate attitudes in default/non-subjective Bayesian texts are common – even in the same book.

For example, at the opening of Ghosh et al. (2010) we hear of the "stopping rule paradox in classical inference" (p. 38), happily avoided by the Bayesian. Then later on it's granted: "Inference based on objective [default] priors does violate the stopping rule principle, which is closely related to the likelihood principle" (p. 148). Arguably, any texts touting the stopping rule principle, while using default priors that violate it, should do likewise. Yet this doesn't bring conceptual clarity to their readers. Why are they upending their core principle?[5]

[3] They allow the possibility that the knowledge that optional stopping will be used alters their prior for 0. I take it they recognize this is at odds with the presumption that "optional stopping is no sin," and they don't press it. See Section 1.5 where we first took this up.

[4] In observing that "informative stopping rules occur only rarely in practice" (p. 90), Berger and Wolpert make the insightful point that disagreement on this is "due to the misconception that an informative stopping rule is one for which *N* carries information about θ."

[5] One explanation is in Bernardo's appeal to a decision theory that considers the sampling distribution in computing utilities.

Readers are assured that the violations of the Likelihood Principle are “minor,” so the authors’ hearts aren’t in it. An error statistician wants *major* violations of a principle that denies the relevance of error probabilities. If the violation is merely an unfortunate quirk growing from the desire for a post-data probabilism, but with default priors, it may seem they gave up on those splendid foundations too readily. The non-subjective Bayesian might be seen as left in the worst of both frequentist and Bayesian worlds. They violate dearly held principles and relinquish the simplicity of classical Bayesianism, while coming up short on error statistical performance properties.

6.6 Error Statistical Bayesians: Falsificationist Bayesians

A final view, which can be seen as either more middle of the road or more extreme, is that of Andrew Gelman (writing both on his own and with others). It’s the former for an error statistician, the latter for a Bayesian. I focus on what Gelman calls “falsificationist Bayesianism” as developed in Gelman and Shalizi (2013). You could see it as the outgrowth of an error-statistical-Bayesian powwow: Shalizi being an error statistician and Gelman a Bayesian. But it’s consistent with Gelman’s general (Bayesian) research program both before and after, and I don’t think anyone can come away from it without recognizing how much statistical foundations are currently in flux. I begin by listing three striking points in their work (Mayo 2013b).

(1) Methodology is ineluctably bound up with philosophy. If nothing else, “strictures derived from philosophy can inhibit research progress” (Gelman and Shalizi 2013, p. 11), as when Bayesians are reluctant to test their models either because they assume they’re subjective, or that checking involved non-Bayesian methods (Section 4.9).
(2) Bayesian methods need a new foundation. Although the subjective Bayesian philosophy, “strongly influenced by Savage (1954), is widespread and influential in the philosophy of science (especially in the form of Bayesian confirmation theory. . .),” and while many practitioners see the “rising use of Bayesian methods in applied statistical work” (ibid., p. 9) as supporting this Bayesian philosophy, the authors flatly declare that “most of the standard philosophy of Bayes is wrong” (p. 10, n. 2). While granting that “a statistical method can be useful even if its common philosophical justification is in error” (ibid.), their stance will rightly challenge many a Bayesian. This is especially so in considering their third thesis.
(3) The new foundation uses error statistical ideas. While at first professing that their “perspective is not new,” but rather follows many other

statisticians in regarding "the value of Bayesian inference as an approach for obtaining statistical methods with good frequency properties"(p. 10), they admit they are "going beyond the evaluation of Bayesian methods based on their frequency properties – as recommended by Rubin (1984), Wasserman (2006), among others" (p. 21). "Indeed, crucial parts of Bayesian data analysis, such as model checking, can be understood as 'error probes' in Mayo's sense" (p. 10), which might be seen as using modern statistics to implement the Popperian criteria of severe tests.

Testing in Their Data-Analysis Cycle

Testing in their "data-analysis cycle" involves a "non-Bayesian checking of Bayesian models." This is akin to Box's eclecticism, except that their statistical analysis is used "not for computing the posterior probability that any particular model was true – we never actually did that" (p. 13), but rather "to fit rich enough models" and upon discerning that aspects of the model "did not fit our data", to build a more complex and better fitting model, which in turn called for alteration when faced with new data. They look to "pure significance testing" (p. 20) with just a null hypothesis, where no specific alternative models are considered. In testing for misspecifications, we saw, the null hypothesis asserts that a given model is adequate, and a relevant test statistic is sought whose distribution may be computed, at least under the null hypothesis (Section 4.9).

They describe their *P*-values as "generalizations of classical *p*-values, replacing point estimates of parameters θ with averages over the posterior distribution…" (p. 18). If a pivotal characteristic is available, their approach matches the usual significance test. The difference is that where the error statistician estimates the "nuisance" parameters, they supply them with priors. It may require a whole hierarchy of priors so that, once integrated out, you are left assigning probabilities to a distance measure $d(\boldsymbol{X})$. This allows complex modeling, and the distance measures aren't limited to those independent of unknowns, but if it seems hard to picture the distribution of μ, try to picture the distribution of a whole iteration of parameters. These will typically be default priors, with their various methods and meanings as we've seen. The approach is a variant on the "Bayesian *P*-value" research program, developed by Gelman, Meng, and Stern (1996) (Section 4.8). The role of the sampling distribution is played by what they call the *posterior predictive distribution*. The usual idea of the sampling distribution now refers to the probability of $d(\boldsymbol{X}) > d(\boldsymbol{x})$ in future replications, averaging over all the priors. These are generally computed by simulation. Their *P*-value is a kind of error probability. Like the severe tester, I take it the concern is well-testedness, not ensuring good long-run performance decisions about misspecification. So, to put it in severe testing terms, if the model, which now includes a series of priors, was highly incapable of generating results so extreme, they infer a statistical indication of inconsistency. Some claim that, at least for large sample sizes, Gelman's approach leads essentially to "rediscovering" frequentist *P*-values (Ghosh et al. 2010, pp. 181–2; Bayarri and Berger 2004), which may well be the reason we so often agree. Moreover, as Gelman and Shalizi (2013, p. 18, n. 11) observe, all participants in the Bayesian *P*-value program implicitly "*disagree* with the standard inductive view" of Bayesianism – at least insofar as they are engaged in model checking (ibid.).

Non-significant results: They compute whether "the observed data set is the kind of thing that the fitted model produces with reasonably high probability" assuming the replicated data are of the same size and shape as $\mathbf{y}_0$, "generated under the assumption that the fitted model, prior and likelihood both, is true" (2013, pp. 17–18). If the Bayesian *P*-value is reasonably high, then the data are "unsurprising if the model is true" (ibid., p. 18). However, as the authors themselves note, "whether this is evidence *for* the usefulness of the model depends how likely it is to get such a high *p*-value when the model is false: the 'severity' of the test" (ibid.) associated with *H*: the adequacy of the combined model with prior. I'm not sure if they entertain alternatives needed for this purpose.

Significant results: A small *P*-value, on the other hand, is taken as evidence of incompatibility between model and data. The question that arises here is: what kind of incompatibility are we allowed to say this is evidence of? The particular choice of test statistic goes hand in hand with a type of discrepancy from the null. But this does not exhaust the possibilities. It would be fallacious to take a small *P*-value as directly giving evidence for a specific alternative that "explains" the effect – at least not without further work to pass the alternative with severity. (See Cox's taxonomy, Section 3.3.) Else there's a danger of a fallacy of rejection we're keen to avoid.

What lets the ordinary *P*-value work in M-S tests is that the null, a mere implicationary assumption, allows a low *P*-value to point the finger at the null – as a hypothesis about what generated this data. Can't they still see their null hypothesis as implicationary? Yes, but indicating a misfit doesn't pinpoint what's warranted to blame. The problem traces back to the split personality involved in interpreting priors. In most Bayesian model testing, it seems the prior is kept sufficiently vague (using default priors), so that the main work is finding flaws in the likelihood part of the model. They claim their methods provide ways for priors to be tested. To check if something is satisfying its role, we had better be clear on what its intended role is. Gelman and Shalizi tell us what a prior need not be: It need not be a default prior (p. 19), nor need it represent a statistician's beliefs. They suggest the model combines the prior and the likelihood "each of which represents some compromise among scientific knowledge, mathematical convenience, and computational tractability" (p. 20). It may be "a regularization device," (p. 19) to smooth the likelihood, making fitted models less sensitive to details of the data. So if the prior fails the goodness-of-fit test, it could mean it represented false beliefs, or that it was not so convenient after all, or . . .? Duhemian problems loom large; there are all kinds of things one might consider changing to make it all fit.

There is no difficulty with the prior serving different functions, so long as its particular role is pinned down for the given case at hand.[6] Since Gelman regards the test as error statistical, it might work to use the problem solving variation on severe testing (Souvenir U), the problem, I surmise, being one of prediction. For prediction, it might be that the difference between M-S testing and Gelman's tests is more of a technical issue (about the best way to deal with nuisance parameters), and less a matter of foundations, on which we often seem to agree. I don't want to downplay differences and others are better equipped to locate them.[7]

What's most provocative and welcome is that by moving away from probabilisms (posteriors and Bayes factors) and inviting error statistical falsification, we alter the conception of which uses of probability are direct, and which indirect.[8] In Gelman and Hennig (2017), "falsificationist Bayesianism" is described as:

> a philosophy that openly deviates from both objectivist and subjectivist Bayesianism, integrating Bayesian methodology with an interpretation of probability that can be seen as frequentist in a wide sense and with an error statistical approach to testing assumptions . . . (p. 991)

In their view, falsification requires something other than probabilism of any type. "Plausibility and belief models can be modified by data in ways that are specified *a priori*, but they cannot be falsified by data" (p. 991). Actually any Bayesian (or even Likelihoodist) account can become falsificationist, indirectly, by adding a falsification rule – provided it has satisfactory error probabilities. But Gelman and Hennig are right that subjective and default

[6] The error statistical account would suggest first checking the likelihood portion of the model, after which they could turn to the prior.

[7] Note, for example, that for a given parameter θ, one has presumably only selected a single θ, not the n samples of our usual M-S test. It's not clear why we should expect it to produce typical outcomes. I owe this point to Christian Hennig.

[8] A co-developer of posterior predictive checks, Xiao-Li Meng, is a leader of the "Bayes–Fiducial–Frequentist" movement.

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 γ 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 α (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?