Tour I What Ever Happened to Bayesian Foundations?

> By and large, Statistics is a prosperous and happy country, but it is not a completely peaceful one. Two contending philosophical parties, the Bayesians and the frequentists, have been vying for supremacy over the past two-and-a-half centuries. . . . *Unlike most philosophical arguments, this one has important practical consequences.* The two philosophies represent competing visions of how science progresses. (Efron 2013, pp. 130; emphasis added)

Surveying the statistical landscape from a hot-air balloon this morning, a bird's-flight view of the past 100 of those 250 years unfolds before us. Except for the occasional whooshing sound of the balloon burner, it's quiet enough to actually hear some of the warring statistical tribes as well as peace offerings and reconciliations – at least with a special sound amplifier they've supplied. It's today's perspective I mainly want to show you from here. Arrayed before us is a most impressive smorgasbord of technical methods, as statistics expands over increasing territory. Many professional statisticians are eclecticists; foundational discussions are often in very much of a unificationist spirit. If you observe the territories undergoing recent statistical crises, you can see pockets, growing in number over the past decade or two, who are engaged in refighting old battles. Unsurprisingly, the methods most often used are the ones most often blamed for abuses. Perverse incentives, we hear, led to backsliding, to slothful, cronyist uses of significance tests that have been deplored for donkey's years. Big Data may have foisted statistics upon fields unfamiliar with the pitfalls stemming from vast numbers of correlations and multiple testing. A pow-wow of leading statisticians from different tribes was called by the American Statistical Association in 2015. We've seen the ASA 2016 Guide on how not to use *P*-values, but some of the "other approaches" also call for scrutiny:

> In view of the prevalent misuses of and misconceptions concerning *p*-values, some statisticians prefer to supplement or even replace *p*-values with other approaches. . . . confidence, credibility, or prediction intervals; Bayesian methods; . . . likelihood ratios or Bayes Factors; and other approaches such as decision-theoretic modeling and false discovery rates. (Wasserstein and Lazar 2016, p. 132)

Suppose you're appraising a recommendation that frequentist methods should or can be replaced by a Bayesian method. Your first question should

be: Which type of Bayesian interpretation? The choices are basically three: subjectivist, default, or frequentist. The problem isn't just choosing amongst them but trying to pin down the multiple meanings being given to each! Classical subjective Bayesianism is home to a full-bodied statistical philosophy, but the most popular Bayesians live among rival tribes who favor one or another default or non-subjective prior probabilities. These are conventions chosen to ensure the data dominate the inference in some sense. By and large, these tribes do not see the growth of Bayesian methods as support for the classical subjective Bayesian philosophy, but rather as a set of technical tools that "work." Their leaders herald frequentist–Bayesian unifications as the way to serve multiple Gods. Zeus is throwing a thunderbolt!

Navigating the reforms requires a roadmap. In Tour I we'll visit the gallimaufry of very different notions of probability in current Bayesian discussions. Concerned that today's practice isn't captured by either traditional Bayesian or frequentist philosophies, new foundations are being sought – that's where we'll travel in Tour II.

Strange bedfellows: the classical subjective Bayesian and the classical frequentist tribes are at one in challenging non-subjective, default Bayesians. The small, but strong tribes of subjective Bayesians, we may imagine, ask them:

> How can you declare scientists want highly probable hypotheses (or comparatively highly probable ones) if your probabilities aren't measuring reasonable beliefs or plausibility (or the like)?

Frequentist error statisticians concur, but also, we may imagine, inquire:

> What's so good about high posterior probabilities if a method frequently assigns them to poorly tested claims?

Let's look back at Souvenir D, where Reid and Cox (2015, p. 295) press the weak repeated sampling requirement on non-frequentist assessments of uncertainty.

> The role of calibration seems essential: even if an empirical frequency-based view of probability is not used directly as a basis for inference; it is unacceptable if a procedure yielding regions of high probability in the sense of representing uncertain knowledge would, if used repeatedly, give systematically misleading conclusions.

Frequentist performance is a necessary, though not a sufficient, condition for severe testing. Even those who deny an interest in performance might not want to run afoul of the minimal requirement for severity. The onus on those who declare what we really want in statistical inference are probabilities on hypotheses is to show, for existing ways of obtaining them, *why*?

Notice that the largest statistical territory is inhabited by practitioners who identify as eclecticists, using a toolbox of various and sundry methods. Some of the fastest growing counties of machine learners and data scientists point to spell checkers and self-driving cars that learn by conjecture and refutation algorithms, at times sidestepping probability models altogether. The year 2013 was dubbed *The International Year of Statistics* partly to underscore the importance of statistics to the Big Data revolution. The best AI algorithms appear to lack a human's grasp of deception based on common sense. That little skirmish is ongoing. Eclecticism gives all the more reason to clearly distinguish the meanings of numbers that stem from methods evaluating different things. This is especially so when it comes to promoting scientific integrity, reproducibility, and in the waves of methodological reforms from journals and reports. Efron has it right: "Unlike most philosophical arguments, this one has important practical consequences" (2013, p. 130). Let's land this balloon, we're heading back to the Museum of Statistics. If you've saved your stub from Excursion 1, it's free.

6.1 Bayesian Ways: From Classical to Default

Let's begin Excursion 6 on the museum floor devoted to classical, philosophical, subjective Bayesianism (which I'm not distinguishing from personalism). This will give us a thumbnail of the position that contemporary non-subjective Bayesians generally reject as a description of what they do. An excellent starting point that is not ancient history, and also has the advantage of contemporary responses, is Dennis Lindley's (2000) "Philosophy of Statistics." We merely click on the names, and authentic-looking figures light up and speak. Here's Lindley:

> The suggestion here is that statistics is the study of uncertainty (Savage 1977): that statisticians are experts in handling uncertainty . . . (p. 294)
>
> [Consider] any event, or proposition, which can either happen or not, be true or false. It is proposed to measure your uncertainty associated with the event . . . If you think that the event is just as uncertain as the random drawing of a red ball from an urn containing N balls, of which R are red, then the event has uncertainty R/N for you. (p. 295)
>
> Historically, uncertainty has been associated with games of chance and gambling. Hence one way of measuring uncertainty is through the gambles that depend on it. (p. 297)
>
> Consider before you an urn containing a known number N of balls that are as nearly identical as modern engineering can make them. Suppose that one ball is drawn at random from the urn . . . it is needful to define randomness. Imagine that the balls are

numbered consecutively from 1 to *N* and suppose that, at no cost to you, you were offered a prize if ball 57 were drawn . . . [and] the same prize if ball 12 were drawn. If you are indifferent between the two propositions and, in extension, between any two numbers between 1 and *N*, then, for you, the ball is drawn at random. Notice that the definition of randomness is subjective; it depends on you. (p. 295)

It is immediate from [Bayes' Theorem] that the only contribution that the data make to inference is through the likelihood function for the observed *x*. This is the likelihood principle that values of *x*, other than that observed, play no role in inference. (pp. 309–10)

[U]nlike the frequency paradigm with its extensive collection of specialized methods, the coherent view provides a constructive method of formulating and solving any and every uncertainty problem of yours. (p. 333)[1]

This is so clear, clean, and neat. The severe tester, by contrast, doesn't object that "specialized" methods are required to apply formal statistics. Satisfying the requirements of severe testing demands it, and that's unity enough. But let's see what some of Lindley's critical responders said in 2000. Press the buttons under their names. I'll group by topic:

1. Subjectivity. *Peter Armitage*: "The great merit of the Fisherian revolution, apart from the sheer richness of the applicable methods, was the ability to summarize, and to draw conclusions from, experimental and observational data without reference to prior beliefs. An experimental scientist needs to report his or her findings, and to state a range of possible hypotheses with which these findings are consistent. The scientist will undoubtedly have prejudices and hunches, but the reporting of these should not be a primary aim of the investigation.. . . There were indeed important uncertainties, about possible biases . . . [and the] existence of confounding factors. But the way to deal with them was . . . by scrupulous argument rather than by assigning probabilities . . ." (ibid., pp. 319–20)

David Cox "It seems to be a fundamental assumption of the personalistic theory that all probabilities are comparable. Moreover, so far as I understand it, we are not allowed to attach measures of precision to probabilities. They are as they are . . . I understand Dennis Lindley's irritation at the cry 'where did the prior come from?' I hope that it is clear that my objection is rather different: why should I be interested in someone else's prior and why should anyone else be interested in mine? (ibid. p. 323) . . . [I]n my view the personalistic probability is virtually worthless for reasoned discussion unless it is based on

[1] "Frequency, however, is not adequate because there is ordinarily no repetition of parameters; they have unique unknown values . . . with the result that it has been necessary for them to develop incoherent concepts like confidence intervals." (p. 311) There are, however, repetitions of types of methods.

information, often directly or indirectly of a broadly frequentist kind. . . . For example, how often have very broadly comparable laboratory studies been misleading as regards human health? How distant are the laboratory studies from a direct process affecting health?" (ibid., p. 322)

2. Non-ampliative. *David Sprott*: "This paper relegates statistical and scientific inference to a branch (probability) of pure mathematics, where inferences are deductive statements of implication: if H_I then H_2. This can say nothing about whether there is reproducible objective empirical evidence for H_I or H_2, as is required by a scientific inference." (ibid., p. 331)

3. Science is Open-Ended. *John Nelder*: "Statistical science is not just about the study of uncertainty, but rather deals with inferences about scientific theories from uncertain data. . . . [Theories] are essentially open ended; at any time someone may come along and produce a new theory outside the current set. This contrasts with probability, where to calculate a specific probability it is necessary to have a bounded universe of possibilities over which the probabilities are defined. When there is intrinsic open-endedness it is not enough to have a residual class of all the theories that I have not thought of yet [the catchall]." (ibid., p. 324)

David Sprott: "Bayes's Theorem (1) requires that *all* possibilities $H_I, H_2, \ldots, H_k$ be specified in advance, along with their prior probabilities. Any new, hitherto unthought of hypothesis or concept H will necessarily have zero prior probability. From Bayes's Theorem, H will then always have zero posterior probability no matter how strong the empirical evidence in favour of H." (ibid., p. 331)

4. Likelihood Principle. *Brad Efron*: "The likelihood principle seems to be one of those ideas that is rigorously verifiable and yet wrong." (Efron 2000, p. 330)[2]

There are also supporters of course, notably, O' Hagan and Dawid, whose remarks we take up elsewhere. The fact that classical Bayesianism reduces statistical inference to probability theory – the very reason many take it as a respite from the chaos of frequentism – could also, Dawid observes, be thought to make it boring: "What is the principal distinction between Bayesian and classical statistics? It is that Bayesian statistics is fundamentally boring. There is so little to do: just specify the model and the prior, and turn the Bayesian handle." (ibid., p. 326). He's teasing I'm sure, but let's step back.

[2] I have argued (e.g., Mayo 2014) the alleged verifications are circular. Efron, in private communication, said that he tried to argue against the result, but gave up; he was glad I did not.

The error statistician agrees with all these criticisms. In her view, statistics is collecting, modeling, and using data to make inferences about aspects of what produced them. Inferences, being error prone, are qualified by reports of the error probing capacities of the inferring method. There is a cluster of error types, real versus spurious effect, wrong magnitude for a parameter, violated statistical assumptions, and flaws in connecting formal statistical inference to substantive claims. It splits problems off piecemeal; there's no need for an exhaustive list of hypotheses that could explain data. Being able to *directly* pick up on gambits like cherry picking and optional stopping is essential for an account to be up to the epistemological task of determining if claims are poorly tested. While for Lindley this leads to incoherence (violations of the likelihood principle), for us it is the key to assessing if your tool is capable of deceptions. According to Efron: "The two philosophies, Bayesian and frequentist, are more orthogonal than antithetical" (Efron 2013, p. 145). Given the radical difference in goals between classical Bayesians and classical frequentists, he might be right. Vive la difference!

But Now Things Have Changed

What should we say now that the landscape has changed? That's what we'll explore in Excursion 6. We'll drop in on some sites we only visited briefly or passed up the first time around. We attempt to disinter the statistical philosophy practiced by the most popular of Bayesian tribes, those using non-subjective or default priors, picking up on Section 1.3, "The Current State of Play". Around 20 years ago, it began to be conceded that: "non-informative priors do not exist" (Bernardo 1997). In effect, they couldn't transcend the problems of "the principle of indifference" wherein lacking a reason to distinguish the probability of different values of θ is taken to render them all equally probable. The definitive review of default methods in statistics is Kass and Wasserman (1996). The default/non-subjective Bayesian focuses on priors that, in some sense, give heaviest weight to data. Impressive technical complexities notwithstanding, there's a multiplicity of incompatible ways to go about this job, none obviously superior. The problem is redolent of Carnap's problem of being faced with a continuum of inductive logics (Section 2.1). (A few are maximum entropy, invariance, maximizing the missing information, coverage matching.) Even for simple problems, recommended default Bayesian procedures differ.

> If the proponents of this view thought their choice of a canonical prior were intellectually compelling, they would not feel attracted to a call for an internationally agreed convention on the subject, as have Berger and Bernardo (1992, p. 57) and Jeffreys (1955, p. 277). (Kadane 2011, p. 445–6)

No such convention has been held.

Default/non-subjective Bayesianism is often offered as a way to unify Bayesian and frequentist approaches.[3] It gives frequentist error statisticians a clearer and less contentious (re)entry into statistical foundations than when Bayesian "personalists" reigned (e.g., Lindley, Savage). At an earlier time, as Cox tells it, confronted with the position that "arguments for this personalistic theory were so persuasive that anything to any extent inconsistent with that theory should be discarded" (Cox 2006a, p. 196), frequentists might have felt alienated when it came to foundations. The discourse was snarky and divisive. Nowadays, Bayesians are more diffident. It's not that unusual to hear Bayesians admit that the older appeal to ideals of rationality were hyped. Listen to passages from Gelman (2011), Kass (2011), and Spiegelhalter (2004):

> Frequentists just took subjective Bayesians at their word and quite naturally concluded that Bayesians had achieved the goal of coherence only by abandoning scientific objectivity. Every time a prominent Bayesian published an article on the unsoundness of p-values, this became confirming evidence of the hypothesis that Bayesian inference operated in a subjective zone bounded by the prior distribution. (Gelman 2011, p. 71)

> [T]he introduction of prior distributions may not have been the central bothersome issue it was made out to be. Instead, it seems to me, the really troubling point for frequentists has been the Bayesian claim to a philosophical high ground, where compelling inferences could be delivered at negligible logical cost. (Kass 2011, p. 6)

> The general statistical community, who are not stupid, have justifiably found somewhat tiresome the tone of hectoring self-righteousness that has often come from the Bayesian lobby. Fortunately that period seems to be coming to a close, and with luck the time has come for the appropriate use of Bayesian thinking to be pragmatically established. (Spiegelhalter 2004, p. 172)

Bayesian empathy with objections to subjective foundations – "we feel your pain" – is a big deal, and still rather new to this traveler's ears. What's the new game all about? There's an important thread that needs to be woven into any answer. Not so long after the retreat from classical subjective Bayes, though it's impossible to give dates (early 2000s?), we saw the rise of irreproducible results and the statistical crisis in science. A new landscape of statistical conflict followed, but grew largely divorced from the older Bayesian–frequentist battles. "Younger readers . . . may not be fully aware of the passionate battles over Bayesian inference among statisticians in the last half of the twentieth century" (Gelman and Robert 2013, p. 1). Opening with Lindley's statistical philosophy

[3] Since we'll be talking a lot about default Bayesians in this tour, I'll use "default/non-subjective" lest I be seen as taking away an appealing name.

lets us launch into newer battles. Finding traditional Bayesian foundations ripped open, coupled with invitations for Bayesian panaceas to the reproducibility crisis, we are swept into a dizzying whirlpool where deeper and more enigmatic puzzles swirl. Do you still have that quicksand stick? (Section 3.6) Grab it and join me on some default, pragmatic, and eclectic Bayesian pathways.

(Note: Since we are discussing existing frequentist–Bayesian arguments, I'll usually use "frequentism" in this excursion, rather than our preferred error statistics.)

6.2 What Are Bayesian Priors? A Gallimaufry

The prevalent Bayesian position might be said to be: there are shortcomings or worse in standard frequentist methods, but classical subjective Bayesianism is, well, too subjective, so default Bayesianism should be used. Yet when you enter default Bayesian territory you'll need to juggle a plethora of competing meanings given to Bayesian priors, and consequently to posteriors.

To show you what I mean, look at a text by Ghosh, Delampady, and Samanta (2010): They say they will stress "objective" (default) priors, "because it still seems difficult to elicit fully subjective priors . . . If a fully subjective prior is available we would indeed use it" (p. 36). Can we slip in and out of non-subjective and subjective priors so easily? Several contemporary Bayesian texts say yes. How should a default prior be construed? Ghosh et al. say that "it represents a shared belief or shared convention," while on the same page it is "to represent small or no information" (p. 30). Maybe it can be all three. The seminal points to keep in mind are spelled out by Bernardo:

> By definition, 'non-subjective' prior distributions are *not* intended to describe personal beliefs, and in most cases, they are *not even proper* probability distributions in that they often do not integrate [to] one. Technically they are *only* positive functions to be formally used in Bayes' theorem to obtain 'non-subjective posteriors' . . . (Bernardo 1997, pp. 159–60)

Bernardo depicts them as a convention chosen "to make precise the type of prior knowledge which" for a given inference problem within a model "would make the data dominant" (ibid, p. 163). Can you just hear Fisher reply (as he did about washing out of priors), "we may well ask what [the prior] is doing in our reasoning at all" (1934b, p. 287). Bernardo might retort: They are merely formal tools "which, for a *given model*, are supposed to describe whatever the data 'have to say' about some *particular quantity*" (1997, p. 160). The desire for an inductive logic of probabilism is familiar to us. Statistician Christian Robert

echoes this sentiment: "Having a prior attached to [a parameter θ] has nothing to do with 'reality,' it is a reference measure that is necessary for making probability statements" (2011, pp. 317–18). How then do we interpret the posterior, Cox asks? "If the prior is only a formal device and not to be interpreted as a probability, what interpretation is justified for the posterior as an adequate summary of information?" (2006a, p. 77)[4]

A Bayesian text by Gelman et al. (2014), to its credit, doesn't blithely assume that because probability works to express uncertainty about events in games of chance, we may assume it is relevant in making inferences about parameters. They aim to show the overall usefulness of the approach. What about meanings of priors?

> We consider two basic interpretations that can be given to prior distributions. In the *population* interpretation, the prior distribution represents a population of possible parameter values, from which the θ of current interest has been drawn. In the more subjective *state of knowledge* interpretation, the guiding principle is that we must express our knowledge (and uncertainty) about θ as if its value could be thought of as a random realization from the prior distribution. (p. 34)

An example from Ghosh et al. (2010) lends itself to the "population interpretation," which to me sounds like a frequentist prior.

Exhibit (i): Blood Pressure and Historical Aside. "Let $X_1, X_2, \ldots, X_n$ be IID $N(\mu, \sigma^2)$ and assume for simplicity σ^2 is known. . . . μ may be the expected reduction of blood pressure due to a new drug. You want to test H_0: $\mu \leq \mu_0$ vs. H_1: $\mu > \mu_0$, where μ_0 corresponds with a standard drug already in the market" (Ghosh et al. 2010, p. 34; their Example 2.4).

Here, μ can be viewed as a random variable that takes on values with different probabilities. The drug of interest may be regarded as a random selection from a population of drugs, each with its expected reductions in blood pressure, i.e., various values of μ. Neyman and Pearson would not have objected; here's a historical aside:

"I began as a quasi-Bayesian": Neyman

> I began as a quasi-Bayesian. My assumption was that the estimated parameter (just one!) is a particular value of a random variable having an unknown prior distribution. (Neyman 1977, p. 128)

Finding them so rarely available, he sought interval estimators with a probability of covering the true value being independent of the prior distribution.

[4] "When the parameter space is finite it [Bernardo reference priors] produces the maximum entropy prior of E. T. Jaynes and, for a one-dimensional parameter, the Jeffreys prior" (Cox 2006b, p. 6).

[My student Churchill Eisenhart in the 1930s] attended my lectures at the University College, London, and witnessed my introducing a prior distribution ... and then making efforts to produce an interval estimator, the properties of which would be independent of the prior. Once, Eisenhart's comment was that the whole theory would look nicer if it were built from the start without any reference to Bayesianism and priors. That remark proved inspiring. (ibid.)

Even the famous 1933 paper considered the Bayesian possibility. E. Pearson had been fully convinced by Fisher's non-Bayesian stance before Neyman, never mind the clash with his (Bayesian-leaning) father. It's one thing to forgo marriage with the woman you love because dad disapproves (as K. Pearson did); it's quite another to follow his view of probability (Section 3.2). Neyman was still exploring. He thought it "important to show that even if a statistician started from the point of view of inverse probabilities he would be led to the same" tests as those he and Pearson recommended (C. Reid 1998, p. 83). Neyman begged Pearson to sign on to a paper that included inverse probability solely for this purpose, but he would not. Pearson worried "they would find themselves involved in a disagreement with Fisher, who had come out decisively against [inverse probability]" (ibid., p. 84) and he never signed on. For more on this episode see C. Reid.

The kind of frequentist prior Neyman allowed were those in genetics. One might consider the probability a person is born with a trait as the effect of a combination of environmental and genetic factors that combine to produce the trait. In an example very like Exhibit (i), Neyman worries that we only know of a finite number of drugs, and we at best have estimates of their average pressure-lowering ability. However, Neyman (1977, p. 115) welcomes the "brilliant idea ... due to Herbert Robbins (1956)" launching "a novel chapter of frequentist mathematical statistics": Empirical Bayes Theory. There may be a sufficient stockpile of information of drugs (or, for that matter, black holes or pulsars) deemed similar to the one in question to arrive at frequentist priors, important for prediction. Some develop "enthusiastic priors" to be contrasted to "skeptical ones" in recommending policy (Spiegelhalter et al. 1994). The severe tester questions if even a fully warranted frequentist posterior gives a report of well-testedness, in and of itself. In any event, most cases aren't like this.

We sometimes hear: But the claim $\{\theta = \theta'\}$ and a claim $\{X = x\}$ are both statements, as if to say, if you can apply probability in one case, why not the other. There's a huge epistemic difference in assessing the probabilities of these different statements. There needs to be a well-defined model assigning probabilities to event statements – just what's missing when we are loath to assign probabilities to parameters. On the other hand, if we are limited to the

legitimate frequentist priors of Neyman, there's no difference between what the Bayesian and frequentist can do, if they wanted to. Donald Fraser (2011) says only these frequentist priors should be called "objective" (p. 313) but, like Fisher, denies this is "Bayesian" inference, because it's just a deductive application of conditional probability, and "whether to include the prior becomes a modeling issue" (ibid., p. 302). But then it's not clear how much of the current Bayesian revolution is obviously Bayesian. Lindley himself famously said that there's "no one less Bayesian than an Empirical Bayesian . . . because he has to consider a sequence of similar problems" (1969, p. 421). Non-frequentist Bayesians switch the role of probability (compared to the frequentist) in a dramatic enough way to be a gestalt change of perspective.

A Dramatic Switch: Flipping the Role of Probability

"A Bayesian takes the view that all unknown quantities, namely the unknown parameter and data before observation, have a probability distribution" (Ghosh et al. 2010, p. 30). By contrast, frequentists don't assign probability to parameters (excepting the special cases noted), and data retain probabilities even after they are observed. This assertion, or close rewordings of it, while legion in Bayesian texts, is jarring to the frequentist ear because it flips the role of probability. Statisticians David Draper and David Madigan put it clearly:

> When we reason in a frequentist way, . . . we view the data as random and the unknown as fixed. When we are thinking Bayesianly, we hold constant things we know, including the data values, . . . – the data are fixed and the unknowns are random. (1997, p. 18)

That's why, when the Higgs researchers spoke of the probability the results are mere statistical flukes, they appeared to be assigning probability to a hypothesis. There was nothing left as random, given the data – at least to a Bayesian. If known data $\boldsymbol{x}$ are given probability 1, we are led to the "old (or known) evidence" problem (Section 1.5) where no Bayes boost is forthcoming. Some further consequences will arise in this tour.[5]

Even where parameters are regarded as fixed, we may assign them probabilities to express uncertainty in them. Where do I get a probability on θ if fixed but unknown? The classic subjective way, we saw, is to find an event with known probability, and build a subjective prior by considering $\{\theta < \theta'\}$ for different values of parameter θ, now regarded as a random variable. If you locate an event E, with known frequentist probability k, such that you're indifferent to bets on $\{\theta < \theta'\}$ and E, then the former gets probability k. A non-

[5] The default Bayesian needn't give probability 1 to data, but it's unclear how they proceed with Bayes' Rule or other computations with a probability on the data and assumptions. Rejecting this possibility, Box and others use frequentist methods for model testing.

subjective/default approach can avoid this, and in some cases arrive at the same test as the frequentist in Exhibit (i) by setting a mathematically convenient conjugate, or an uninformative prior, say by viewing θ itself as Normally distributed $N(\eta, \tau^2)$. Instead of reporting the significance level of 0.05, this allows reporting that the posterior probability of H_0 is 0.05.

> $\Pr(\theta = \theta_0) = 0.95$ is meaningless unless θ is a random variable. . . . this expression signifies that we are ready to bet that θ is equal to θ_0 with a 95/5 odds ratio, or, in other words, that the uncertainty about the value of θ is reduced to a 5% zone. (Robert 2007, p. 25; Pr for *P*)

Would we want to equate error probabilities to readiness to bet? As always it's most useful to look at cases of poor or weak evidence. Suppose you arrive at statistical significance of 0.2. We would be entitled to say we're as ready to bet on $\theta > \theta'$ as on the occurrence of an event with probability 0.8. I don't think we'd want to be so entitled. The default Bayesian replies, this just means the default prior doesn't reflect my beliefs. OK, but recall the question at the outset of this tour: *why assume we want a posterior probability on statistical hypotheses*, in any of the ways now available? The default Bayesian was to supply the (ideally) unique prior to use, not send us back to subjective priors.

The Bayesian treats the blood-pressure example very differently if the null is a point such as $\theta = 0$, whereas there's no difference for a frequentist. The spike and smear priors surveyed in Excursion 4 are common. Greenland and Poole (2013) suggest:

> [A] null spike represents an assertion that, with prior probability q, we have background data that prove $\theta_t = 0$ with absolute certainty; $q = ½$ thus represents a 50–50 bet that there is decisive information literally proving the null. [Otherwise a] spike at the null is an example of 'spinning knowledge out of ignorance.' (p. 66)

This is an interesting construal. Instead of how strongly you believe the null, it's how strongly you believe in a proof of it. That decisive information exists (their second clause) is weaker than actually having it (their first clause), but both are stronger than presuming they arise from a noncommittal "equipoise." Of course, the severe tester wants to know how strong the existing demonstration of H is, not how strong your belief in such a demonstration is.

Some subjective Bayesians would chafe at the idea of betting on scientific hypotheses or theoretical quantities. For one thing, it's hard to imagine people would be indifferent between a bet they know will be settled and one that is unlikely to be – as in the case of most scientific hypotheses. No one's going to put their money down now (unless they get interest). Still, cashing out Bayesian uncertainty with betting seems the most promising way to

"operationalize it." Other types of scoring functions may be used, but still, there's a nagging feeling they leave us in the dark about what's really meant.

> For both subjectivist and objectivist [default] Bayesians, probability models including both parameter priors and sampling models do not model the data-generating process, but rather represent plausibility or belief from a certain point of view. (Gelman and Hennig 2017, pp. 990–1)

Yet Gelman et al. (above) suggested expressing uncertainty as if a parameter's "value could be thought of as a random realization from the prior distribution" (2014, p. 34). If this is bending your brain, then you're getting it. Claims like it's "the knowledge [of fixed but unknown parameters] that Bayesians model as random" (Gelman and Robert 2013, p. 4) feel as if they ought to make perfect sense, but the more you think about them, the more they're liable to slip from grasp. For our purposes, let's understand claims that unknown quantities *have* probability distributions in terms of a person or persons who are doing the having – by assigning different degrees of belief (or other weights) to different parameter values.

The Probabilities of Events

Many Bayesian texts open with a focus on probabilities of simple events, or statements of events, like the "heads" on the toss of a coin. By focusing on probabilities of events which even frequentists condone, the reader may wonder what all the fuss is about. Problem is, the central point of contention between Bayesians and frequentists is whether to place probabilities on parameters in a statistical model. It isn't that frequentists don't assign probabilities to events, and any statistics based on them. It is that they recognize the need to infer a statistical model, and hypotheses about its parameters, in order to get those probabilities. How does that inference occur? It rests on probabilistic properties of a test method, which is very different from the deductive assignment of probability to hypotheses.

The severe tester uses probabilities assigned to events like {test T yields $d(\boldsymbol{X}) > d(\boldsymbol{x})$} to detach statistical inferences. She might argue: If $\Pr\{d(\boldsymbol{X}) > d(\boldsymbol{x}); H'\}$ is not very small, infer there's a poor indication of a discrepancy H'. Computing $\Pr\{d(\boldsymbol{X}) > d(\boldsymbol{x}); \theta\}$ for varying θ tells me the capability of the test to detect various discrepancies from a reference value of interest. This does not give a posterior probability to the hypothesis, but it allows making statistical inferences which are qualified by how well or poorly tested claims are.

True, when the frequentist assigns a probability to an event, it is seen as a general type, whereas a Bayesian can assign subjective probability to a unique event on November 8, 2016! Or so it is averred. But when they appeal to bets by

reference to events with known probabilities, aren't they viewing it as a type? ("That's the kind of thing I'd bet 0.9 on.")

Cox points out that even subjectivists must think their probabilities have a frequentist interpretation. Consider n events/hypotheses:

> ... all judged by You to have the same probability p and not to be strongly dependent ... It follows from the Weak Law of Large Numbers obeyed by personalistic probability that Your belief that about a proportion p of the events are true has probability close to 1. (Cox 2006a, p. 79)

This suggests, Cox continues, that to elicit Your probability for H you try to find events or hypotheses that you judge for good reason to have the same probability as H, and then find out what proportion of this set is true. This proportion would yield Your subjective probability for H. Echoes of the screening model of tests (Section 5.6). Here the (hypothetical or actual) urn contains hypotheses that you thus far judge to be as probable as the H of interest. If the proportion of hypotheses in this urn that turned out true was, say, 80%, then H would get probability 0.8. It would be rare to know the truth rates of the hypotheses in this urn – would it be the proportion now assigned probability 1 by the subjectivist? Perhaps the proportion not yet falsified could be used.

Still, this would be a crazy way to actually go about evaluating evidence and hypotheses! But what if you considered H as if it were randomly selected from an urn of hypotheses that had passed severe tests, perhaps made up of claims in the same field. You check the relative frequency that are true or have held up so far. You'd still need to show why you're putting H in the high severity urn. In other words, you would have circled right back to the initial assignment of severity. All you'd be doing is reporting how often severely corroborated claims are true, or continue to solve their empirical problems (predicting or explaining). There would be nothing added by the imaginary urn.

A different attempt to assign a frequentist probability to a hypothesis H might try to consider how probable it is that the universe is such that H is true, considering fundamental laws, other worlds, multiverses, or what have you. One might consider the rarity of possible worlds that would have such a law. Even if we could somehow compute this, how would it be relevant to assessing hypotheses about this world? Here's C. S. Peirce:

> [The present account] does not propose to look through all the possible universes, and say in what proportion of them a certain uniformity occurs; such a proceeding, were it possible, would be quite idle. The theory here presented only says how frequently, in this universe, the special form of induction or hypothesis would lead us right. The probability given by this theory is in every way different – in meaning, numerical

value, and form – from that of those who would apply to ampliative inference the doctrine of inverse chances. (Peirce 2.748)

This objection, I take it, is different from trying to determine, on theoretical principles, how "fine tuned" this world would have to be for various parameters to be as we estimate them. Those pursuits, whose validity I'm in no position to judge, are aimed at deciding whether we should fiddle with theoretical assumptions so that this universe is not so "unnatural."

6.3 Unification or Schizophrenia: Bayesian Family Feuds

> COX: There's a lot of talk about what used to be called inverse probability and is now called Bayesian theory. That represents at least two extremely different approaches. How do you see the two? Do you see them as part of a single whole? Or as very different?
> MAYO: It's hard to give a single answer, because of a degree of schizophrenia among many Bayesians. On paper at least, the subjective Bayesian and the so-called default Bayesians . . . are wildly different. For the former the prior represents your beliefs apart from the data, . . . Default Bayesians, by contrast, look up 'reference' priors that do not represent beliefs and might not even be probabilities, . . . Yet in reality default Bayesians seem to want it both ways. (Cox and Mayo 2011, p. 104)

If you want to tell what's true about today's Bayesian debates, you should consider what they say in talking amongst themselves. I began to sense a shifting of sands in the foundations of statistics landscape with an invitation to comment on Jim Berger (2003). The trickle of discontent from family feuds issuing from Bayesian forums pulls back the curtain on how Bayesian–frequentist debates have metamorphosed. To show you what I mean, let's watch the proceedings of a conference at Carnegie Mellon, published in *Bayesian Analysis* (vol. 1, no. 3, 2006) in the museum library. Unlike J. Berger's (2003) attempted amalgam of Jeffreys, Neyman, and Fisher (Section 3.6), here it's Berger smoking the peace pipe, making "The Case for Objective Bayesianism" to his subjective compatriots (Section 4.1). The forum gives us a look at the inner sanctum, with Berger presenting a tough love approach: If we insist on subjectivity, we're out. "[T]hey come to statistics in large part because they wish it to provide objective validation of their science" (J. Berger 2006, p. 388).

Four Philosophical Positions

Admitting there is no unanimity as to either the definition or goal of "objective" (default) Bayesianism, Berger (2006, p. 386) outlines "four philosophical positions" that default Bayesianism might be seen to provide:

1. A complete coherent objective Bayesian methodology for learning from data.
2. The best method for objectively synthesizing and communicating the uncertainties that arise in a specific scenario, but is not necessarily coherent.
3. A convention we should adopt in scenarios in which a subjective analysis is not tenable.
4. A collection of ad hoc but useful methodologies for learning from data.

Berger regards (1) as unattainable; (2) as often attainable and should be done if possible, but concedes that often the best we can hope for is (3), or maybe (4). Lindley would have gone with (1).

Is a collection of ad hoc but useful methodologies good enough? There is a fascinating philosophical tension in Berger's work: while in his heart of hearts he holds "the (arguably correct) view that science should embrace subjective statistics", he realizes this "falls on deaf ears" (ibid., p. 388). When scientists demur: "I do not want to do a subjective analysis, and hence I will not use Bayesian methodology," Berger convincingly argues they can have it both ways (p. 389).

Among the advantages to adopting a default Bayesian methodology is avoiding a subjective elicitation of experts. Berger finds elicitation does not work out too well. Far from providing a route within which to describe background knowledge in terms of prior probabilities, he finds elicitation foibles are common even with statistically sophisticated practitioners. "[V]irtually never would different experts give prior distributions that even overlapped; there would be massive confusions over statistical definitions (e.g., what does a positive correlation mean?)" coupled with the difficulty of eliciting priors when, as is typical, "the expert has already seen the data" (ibid., p. 392). But if the prior is determined post-data, one wonders how it can be seen to reflect information independent of the data. I come back to this. In his own experience Berger found:

> . . . for the many parameters for which there was data . . . all of the expert time was used to assist model building. It was necessary to consider many different models, and expert insight was key to obtaining good models; there simply was no extra available expert time for prior elicitation. (ibid.)

He argues that the default choices have the advantage over trying to elicit a subjective prior:

> The problem is that, to elicit all features of a subjective prior $\pi(\theta)$, one must infinitely accurately specify a (typically) infinite number of things. In practice, only a modest number of (never fully accurate) subjective elicitations are possible, so practical

Bayesian analysis must somehow construct the entire prior distribution $\pi(\theta)$ from these elicitations. (ibid., p. 397)

A standard way to turn elicitations into full prior distributions is to use mathematically convenient priors (as with default priors). The trouble is this leads to Bayesian incoherence, in violation of the Likelihood Principle (LP). Why? Because "depending on the experiment designed to study θ, the subjective Bayesian following this 'prior completion' strategy would be constructing different priors for the same θ, clearly incoherent" (ibid.). Ironically, this LP violation is not directly driven by the need to compute the sampling distribution to obtain frequentist error probabilities: it is a way to try to capture a reasonably non-informative prior – as this is thought to depend on the experiment to be performed.[6] Any good error properties are touted as a nice bonus, not the deliberate aim, except for the special case of error probability matching priors.

Berger maintains that the default, at best, achieves "a readily understandable communication of the information in the observed data, as communicated through a statistical model, for any scientific question that is posed" (ibid., p. 388). We've seen that there's considerable latitude open to the default Bayesian – the source of arguments that *P*-values overstate the evidence. It's hard to view those spiked priors as merely conveying what the data say (especially when they use a two-sided test). Another issue is that it often distinguishes parameters of "interest" from additional "nuisance" parameters, each of which must be ordered. In Bernardo's system of reference priors, there are as many reference priors as possible parameters of interest (1997, p. 169). That's because what counts as data dominance (he calls it "maximizing the missing information") will differ for different parameters. Each ordering of parameters will yield different posteriors. Despite Berger's own misgivings in avoiding elicitation bias:

A common and reasonable practice is to develop subjective priors for the important parameters or quantities of interest in a problem, with the unimportant or 'nuisance' parameters being given objective priors. (ibid., p. 393)

Here again we see the default Bayesian inviting both types of priors: if you have information, put it in the elicitation; if not, keep it out and choose one of the

[6] One way to link the LP violation with Bayesian incoherence is to show that the posterior depends on the order of two independent experiments for the same parameter. We know the Binomial and Negative Binomial experiments have different sample spaces (Section 4.9), and yet are not distinguished on the LP. Default priors, however, are sample-space dependent. If the first experiment is Binomial and the second Negative Binomial, both for inferences about the probability of success on each trial, a different posterior results depending on the order that the default rule is applied. Excellent discussions are in Seidenfeld (1979) and Kass and Wasserman (1996, p. 1359). Note that some consider that coherence only concerns the assignment of the prior; a violation of Bayes' Rule is called a failure of Bayesian conditionalization.

conventional priors. You might say there's nothing really schizophrenic in this, even subjectivists argue that default priors are kosher as approximations to what they would have arrived at in cases of minimal information (O'Hagan, this forum). It's just faster. Should they be so sanguine? The tasks are quite different.

[O]bjective priors can vary depending on the goal of the analysis for a given model. For instance, in a normal model, the reference prior will be different if inference is desired for the mean μ or if inference is desired for μ/σ. This, of course, does not happen with subjective Bayesianism. (Berger 2006, p. 394)

Trying to describe your beliefs is different from trying to make the data dominant relative to a given model and ordering of parameters. Subjectivists hold that prior beliefs in H shouldn't change according to the experiment to be performed. However, if they incorporate default priors, when required by complex problems, this changes. Since priors of different sorts are then combined in a posterior, how do you tell what's what? If nothing else, the simplicity that led Dawid to joke that Bayesianism is boring disappears.

Ironic and Bad Faith Bayesianism

A major impetus for developing default Bayesian methods, for Berger, is to combat what he calls "casual Bayesianism" or pseudo-Bayesianism.

One of the mysteries of modern Bayesianism is the lip service that is often paid to subjective Bayesian analysis as opposed to objective Bayesian analysis, but then the practical analysis actually uses a very adhoc version of objective Bayes, including use of constant priors, vague proper priors, choosing priors to 'span' the range of the likelihood, and choosing priors with tuning parameters that are adjusted until the answer 'looks nice.' I call such analyses *pseudo-Bayes* because, while they utilize Bayesian machinery, they do not carry with them any of the guarantees of good performance that come with either true subjective analysis (with a very extensive elicitation effort) or (well-studied) objective Bayesian analysis. (Berger 2006, pp. 397–8)

Berger stops short of prohibiting casual Bayesianism, but warns that it "must be validated by some other route" (ibid.), left open. One thing to keep in mind: "good performance guarantees" mean disparate things to Bayesians and to frequentist error statisticians. Remember those subscripts. "In general reference priors have some good frequentist properties but except in one-dimensional problems it is unclear that they have any special merit in that regard" (Cox 2006b, p. 6). Judging from the ensuing discussion, Berger's concern here is with resulting improper posteriors that can remain hidden in the use of computer packages. Improper priors are often not problematic, but posteriors that are not probabilities (because they don't add to 1) are a disaster.

Interestingly, Lindley came to his subjective Bayesian stance after he was shown that conventional priors can lead to improper posteriors and thus to violations of probability theory (Dawid, Stone, and Zidek 1973). A remark that is especially puzzling or revealing, depending on your take:

> Too often I see people pretending to be subjectivists, and then using 'weakly informative' priors that the objective Bayesian community knows are terrible and will give ridiculous answers; subjectivism is then being used as a shield to hide ignorance . . . In my own more provocative moments, I claim that the only true subjectivists are the objective Bayesians, because they refuse to use subjectivism as a shield against criticism of sloppy pseudo-Bayesian practice. (Berger 2006, pp. 462–3)

How shall we deconstruct this fantastic piece of apparent doublespeak? I take him to mean that a subjectivist who properly recognizes her limits and biases and opts to be responsible for her priors would accept the constraints of default priors. A pseudo-Bayesian uses priors as if these really reflected properly elicited subjective judgments. In doing so, she (thinks that she) doesn't have to justify them – she claims that they reflect subjective judgments (and so who can argue with them?).

Although most Bayesians these days disavow classic subjective Bayesian foundations, even the most hard-nosed, "we're not squishy" Bayesians retain the view that a prior distribution is an important if not the best way to bring in background information. Here's Christian Robert:

> The importance of the prior distribution in a Bayesian statistical analysis is not at all that the parameter of interest θ can (or cannot) be perceived as generated from [prior distribution π] . . . but rather that the use of a prior distribution is the best way to summarize the available information (or even the lack of information) about this parameter. (Robert 2007, p. 10)

But is it? To suppose it is pulls in the opposite direction from the goal of the default prior which is to reflect just the data.

Grace and Amen Bayesians

> I edit an applied statistics journal. Perhaps one quarter of the papers employs Bayes' theorem, and most of these do *not* begin with genuine prior information. (Efron 2013, p. 134)

Stephen Senn wrote a paper "You Might Believe You Are a Bayesian But You Are Probably Wrong." More than a clever play on words, Senn's title highlights the common claim of researchers to have carried out a (subjective) Bayesian analysis when they have actually done something very different. They start and end with thanking the (subjective?) Bayesian account for housing all their

uncertainties within prior probability distributions; in between, the analysis immediately turns to default priors, coupled with ordinary statistical modeling considerations that may well enter without being put in probabilistic form. "It is this sort of author who believes that he or she is Bayesian but in practice is wrong" (Senn 2011, p. 58). In one example Senn cities Lambert et al. (2005, p. 2402):

> [T]he authors make various introductory statements about Bayesian inference. For example, 'In addition to the philosophical advantages of the Bayesian approach, the use of these methods has led to increasingly complex, but realistic, models being fitted,' and 'an advantage of the Bayesian approach is that the uncertainty in all parameter estimates is taken into account' . . . but whereas one can neither deny that more complex models are being fitted than had been the case until fairly recently, nor that the sort of investigations presented in this paper are of interest, these claims are clearly misleading. . . (Senn 2011, p. 62)

While the authors "considered thirteen different Bayesian approaches to the estimation of the so-called random effects variance in meta-analysis . . ." – techniques fully available to the frequentist, "[n]one of the thirteen prior distributions considered can possibly reflect what the authors believe about the random effect variance" (ibid., pp. 62–3).

Ironically, Senn says, a person who takes into account the specifics of the case in their statistical modeling is "being more Bayesian in the de Finetti sense" (ibid) than the default/non-subjective Bayesian. By focusing on how to dress the case into ill-fitting probabilistic clothing, Senn is insinuating, the Bayesians may miss context-dependent details solely because they were not framed probabilistically. Leo Breiman, an early leader in machine learning, needled Bayesians:

> The Bayesian claim that priors are the only (or best) way to incorporate domain knowledge into the algorithms is simply not true. Domain knowledge is often incorporated into the structure of the method used. . . . In handwritten digit recognition, one of the most accurate algorithms uses nearest-neighbor classification with a distance that is locally invariant to things such as rotations, translations, and thickness. (Breiman 1997, p. 22)

Nor need context-dependent information of a repertoire of mistakes and pitfalls be cashed out in terms of priors. They'd surely be reflected in a post-data assessment of severity, which would be open to model builders from any camp.

Finally, the "the lip service that is often paid to subjective Bayesian analysis as opposed to objective Bayesian analysis," far from being the "modern mystery," Berger (2006, p. 397) dubs it, might reflect the degree of schizophrenia of

default Bayesianism. After all, Berger[7] says that the default prior "is used to describe an individual's (or group's) 'degree of belief'" (ibid., p. 385), while ensuring the influence of subjective belief is minimal. Moreover, in using default priors, he maintains, you're getting closer to the subjective Bayesian ideal (absent a full elicitation). So there should be no surprise when a default Bayesian says she's being a good subjective Bayesian. The default Bayesians attain an aura of subjective foundations for philosophical appeal, and non-subjective foundations for scientific appeal. If you come face to face with a default posterior probability, you need to ask which default method was used, the ordering of parameters, the mixture of subjective and default priors and so on. Even a transparent description of all that may not help you appraise whether a high default posterior in H indicates warranted grounds for H.

6.4 What Happened to Updating by Bayes' Rule?

In striving to understand how today's Bayesians view their foundations, we find even some true-blue subjective Bayesians reject some principles thought to be important, such as Dutch book arguments. If it is agreed that we have degrees of belief in any and all propositions, then it is argued that if your beliefs do not conform to the probability calculus you are being incoherent. We can grant that if we had degrees of belief, and were required to take any bets on them, that, given we prefer not to lose, we do not agree to a series of bets that ensures losing. This is just a tautologous claim and entails nothing about degree of belief assignments. "That an agent ought not to accept a set of wagers according to which she loses come what may, if she would prefer not to lose, is a matter of deductive logic and not a property of beliefs" (Bacchus, Kyburg, and Thalos 1990, pp. 504–5).

The dynamic Dutch book argument was to show that the rational agent, upon learning some data E, would update by Bayes' Rule, else be guilty of irrationality. Confronted with counterexamples in which violating Bayes' Rule seems perfectly rational on intuitive grounds, many if not most Bayesian philosophers dismiss threats of being Dutch-booked as irrelevant. "It is the entirely rational claim that I may be induced to act irrationally that the dynamic Dutch book argument, absurdly, would condemn as incoherent" (Howson 1997a, p. 287). Howson declares it was absurd all along to consider it irrational to be induced to act irrationally. It's insisting on updating by Bayes' Rule that is irrational. "I am not inconsistent in planning … to entertain

[7] I've no objection to Berger's viewing "probability as a primitive concept" (p. 385). Theoretical concepts may arise in models and receive their explication through applications. It's problematic when the subsequent meanings shift, as happens with default probabilities.

a degree of belief [that is inconsistent with what I now hold], I have merely changed my mind" (ibid.). One thought the job of Bayesian updating was to show *how* to change one's mind reasonably.

Counterexamples to Bayes' Rule often take the following form: While an agent assigns probability 1 to event S at time t, i.e., $\Pr(S) = 1$, he also believes that at some time in the future, say t', he may assign a low probability, say 0.1, to S, i.e., $\Pr'(S) = 0.1$, where P' is the agent's belief function at later time t'.

> Let E be the assertion: $P'(S) = 0.1$.
> So at time t, $\Pr(E) > 0$.
> But $\Pr(S|E) = 1$ since $P(S) = 1$.

Now, Bayesian updating says:

> If $\Pr(E) > 0$, then $\Pr'(.) = \Pr(.\,|E)$.
> But at t' we have, $\Pr'(S) = 0.1$,

which contradicts $\Pr'(S) = \Pr(S|\ \Pr'(S) = 0.1) = 1$ obtained by Bayesian updating. It is assumed, by the way, that learning E does not change any of the other degree of belief assignments held at t – never mind how one knows this.

The kind of example at the heart of this version of the counterexample was given by William Talbott (1991, p. 139). In one of his examples: S is "Mayo ate spaghetti at 6 p.m., April 6, 2016". $\Pr(S) = 1$, where Pr is my degree of belief in S now (time t), and E is "$\Pr'(S) = r$", where r is the proportion of times Mayo eats spaghetti (over an appropriate time period); say $r = 0.1$. As vivid as eating spaghetti is today, April 6, 2016, as Talbott explains, I believe, rationally, that next year at this time I will have forgotten, and will (rationally) turn to the relative frequency with which I eat spaghetti to obtain $\Pr'$. Variations on the counterexample involve current beliefs about impairment at t' through alcohol or drugs. This is temporal incoherency.

It may seem surprising for a subjective Bayesian like Howson to reject Bayes' Rule: Typically, it's the subjectivist who recoils in finding the default/non-subjective tribes living in conflict with it. Jon Williamson, a non-subjective Bayesian philosopher in a Carnap-maximum entropy mold,[8] identifies the problem in these examples as stemming from two sources of probabilistic information (Williamson 2010). Relative frequency information tells

[8] The noteworthy Carnapian part is his relativization to first order languages, rather than to statistical models.

us Pr′(S) = 0.1, but also, since this is known, Pr′(Pr′(S) = 0.1) = 1. Bayes' Rule holds, he allows, just when it holds. When there's a conflict with Bayes' Rule, default Bayesian "updating" takes place to reassign priors.

The position of Howson and Williamson is altogether plausible if one is forced with the given assignments. Ian Hacking, who is not a Bayesian, sympathizes, and blames universal Bayesianism. Universal Bayesianism, Hacking (1965, p. 223) remarks, forces Savage (Savage 1962, p. 16) to hold "if you come to doubt [e.g., the Normality of a sample], you must always have had a little doubt". To Hacking, it's plausible to be completely certain of something, betting the whole house and more, and later come to doubt it. All the more reason we should be loath to assign probability 1 to "known" data, while seeking a posterior probabilism.

Bayesian statisticians, at least of the default/non-subjective variety, follow suit, though for different reasons: "Betting incoherency thus seems to be too strong a condition to apply to communication of information" (J. Berger 2006, p. 395). Berger avers that even subjective Bayesianism is not coherent in practice, "except for trivial versions such as always estimate $\theta \in (0, \infty)$ by 17.35426 (a coherent rule, but not one particularly attractive in practice)" (pp. 396–7). His point appears to be that, while incoherence is part and parcel of default/non-subjective Bayesian accounts, in practice, idealizations lead the subjectivist to be incoherent as well. It gets worse: "in practice, subjective Bayesians will virtually always experience what could be called practical marginalization paradoxes" (p. 397), where posteriors don't sum to 1. If this is so, it's very hard to see how they can be happy using any kind of probability logic.

There are a great many complex twists and turns to the discussions of Dutch books; too many to do justice with a sample list.

Can You Change Your Bayesian Prior?

> As an exercise in *mathematics* [computing a posterior based on the clients prior probabilities] it is not superior to showing the client the data, eliciting a posterior distribution and then calculating the prior distribution; as an exercise in *inference* Bayesian updating does not appear to have greater claims than 'downdating' . . . (Senn 2011, p. 59)

> If you could really express your uncertainty as a prior distribution, then you could just as well observe data and directly write your subjective posterior distribution, and there would be no need for statistical analysis at all. (Gelman 2011, p. 77)

Lindley's answer is that it would be a lot harder to be coherent that way, however, he's prepared to allow: "[I]f a prior leads to an unacceptable posterior

then I modify it to cohere with properties that seem desirable in the inference" (Lindley 1971, p. 436). He resists saying the rejected prior was wrong though. Not wrong? No, just failing to cohere with desirable properties. I. J. Good (1971b) advocated his device of "imaginary results" whereby a subjective Bayesian would envisage all possible results in advance (p. 431) and choose a prior that she can live with regardless of results. Recognizing that his device is so difficult to apply that most are prepared to bend the rules, Good allowed "that it is possible after all to change a prior in the light of *actual* experimental results" (ibid.) – appealing to an informal, second-order rationality of "type II."

So can you change your Bayesian prior? I don't mean update it, but reject the one you had and replace it with another. I raised this question on my blog (June 18, 2015), hoping to learn what current practitioners think. Over 30 competing answers ensued (from over 100 comments), contributed by statisticians from different tribes. If the answer is yes you can, then how do they avoid the verification biases we are keen to block? Lindley seems to be saying it's quite open-ended. Cox says, "there is nothing intrinsically inconsistent in changing prior assessments" in the light of data, however the danger is that "even initially very surprising effects can post hoc be made to seem plausible" (Cox 2006b, p. 78). Berger had said elicitation typically takes place after "the expert has already seen the data" (2006, p. 392), a fact of life he understandably finds worrisome. If the prior is determined post-data, then it's not reflecting information *independent* of the data. All the work would have been done by the likelihoods, normalized to be in the form of a probability distribution or density. No wonder many look askance at changing priors based on the data.

> [N]o conceivable possible constellation of results can cause you to wish to change your prior distribution. If it does, you had the wrong prior distribution and this prior distribution would therefore have been wrong even for cases that did not leave you wishing to change it. (Senn 2011, p. 63)

The prior, after all, is "like already having some data, but what statistical procedure would allow you to change your data?" (Senn 2015b). As with Good's appeal to type II rationality, Senn is saying this is tantamount to admitting "the informal has to come to the rescue of the formal" (Senn 2011, p. 58), which would otherwise permit counterintuitive results. He makes the interesting point that the post-data adjustment of priors could conceivably be taken account of in the posterior: "If you see there is a problem with the placeholder model and replace it, it may be that you can somehow reflect this

'sensible cheating' in your posterior probabilities" (Senn 2015b; see also Senn 2013a). I think Senn is applying a frequentist error statistical mindset to the Bayesian analysis, wherein the posterior might be qualified by an error statistical assessment. Bayesians would need a principle to this effect.

Dawid (2015) weighs in with his "prequential" approach. "In this approach the prior is constructed, not regarded as given in advance". Maybe the idea is that the subjective Bayesian is trying to represent her psychological states; the posterior from the data indicate her first stab failed to do so, so it makes sense to change it. The main thing for Dawid is to have a coherent package; his Bayesian starts over with a better prior and a new test. But it's not obvious how you block yourself from engineering the result you want. Gelman and Hennig say "priors in the subjectivist Bayesian conception are not open to falsification . . . because by definition they have to be fixed before observation" (2017, p. 989). Now Howson's Bayesian changes his mind post-data, but admittedly this is not the same as falsification. In his comment to the blog discussion, Gelman (2015) says that "if some of the [posterior] inferences don't 'make sense,' this implies that you have additional information that has not been incorporated into the model" and it should be improved. But not making sense might just mean that more information would be necessary to get an answer, not that you rightfully have it. Shouldn't we worry that among the many ways you fix things, you choose one that protects or enhances a favored view, even if poorly probed? A reply might be that frequentists worry about data-dependent adjustments as well. There's one big difference.

In order for a methodological "worry" to be part of an inference account, it needs an explicit rationale, not generally found in contemporary Bayesianism – though Gelman is an exception. An error statistician changes her model in order to ensure the reported error probabilities are close to the actual ones (whether for performance or severe testing). There seem to be at least two situations where the default/non-subjective Bayesian may start over: The first, already noted, is when there's a conflict with Bayes' Rule. "Updating" takes place by going back to assign new prior probabilities using a chosen default prior. Philosophers Gaifman and Vasudevan (2012) describe it thus: ". . . the revision of an agent's [rational] subjective probabilities proceeds in fits and starts, with periods of conditionalization punctuated by abrupt alterations of the prior" (p. 170).[9] Why wouldn't this be taken as questioning the entire method of reaching a default prior assignment? Surely it relinquishes a key

[9] They have in mind maximal entropy methods advanced by Jaynes and also developed by Roger Rosenkrantz (1977). It is thought to work well in contexts where the current experiment may be seen as a typical instance of a known physical θ-generating process.

feature Bayesianism claims to provide: a method of accumulating and updating knowledge probabilistically. A second situation might be finding information that statistical assumptions are violated. But this brings up Duhemian problems as we'll see in Section 6.6.

The Bayesian Catchall

The key obstacle to probabilistic updating, and to viewing an evidential assessment at a given time in terms of a posterior probabilism, is the Bayesian catchall hypothesis. One is supposed to save some probability for a catchall hypothesis: "everything else," in case new hypotheses are introduced, which they certainly will be. Follow me to the gallery on the 1962 Savage Forum for a snippet from the discussion between Savage and Barnard (Savage 1962, pp. 79–84):

BARNARD: ... Professor Savage, as I understood him, said earlier that a difference between likelihoods and probabilities was that probabilities would normalize because they integrate to one, whereas likelihoods will not. Now probabilities integrate to one only if all possibilities are taken into account. This requires in its application to the probability of hypotheses that we should be in a position to enumerate all possible hypotheses which might explain a given set of data. Now I think it is just not true that we ever can enumerate all possible hypotheses. . . . If this is so we ought to allow that in addition to the hypotheses that we really consider we should allow something that we had not thought of yet, and of course as soon as we do this we lose the normalizing factor of the probability, and from that point of view probability has no advantage over likelihood. (p. 80)

SAVAGE: ... The list can, however, always be completed by tacking on a catchall 'something else.' ... In practice, the probability of a specified datum given 'something else' is likely to be particularly vague – an unpleasant reality. The probability of 'something else' is also meaningful of course, and usually, though perhaps poorly defined, it is definitely very small.

BARNARD: Professor Savage says in effect, 'add at the bottom of list $H_1, H_2, \ldots$ 'something else'.' But what is the probability that a penny comes up heads given the hypothesis 'something else.' We do not know.

Suppose a researcher makes the catchall probability small, as Savage recommends, and yet the true hypothesis is not in the set so far envisaged, call this set $\underline{H}$. Little by little, data might erode the probabilities in $\underline{H}$, but it could take a very long time until the catchall is probable enough so that a researcher begins to develop new theories. On the other hand, if a researcher suspects the existing hypothesis set $\underline{H}$ is inadequate, she might give the catchall a high prior. In that case, Barnard points out, it may be that none of the available hypotheses

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.