Stats Wars: Return of the Bayesians

Rafael Fuchs (Rafael.Fuchs@campus.lmu.de)[1,2]
[1]Graduate School of Systemic Neurosciences, LMU Munich
[2]Munich Center for Mathematical Philosophy, LMU Munich

Abstract

In view of the ongoing replication crisis and various frustrations about classical null-hypothesis significance testing, calls for a statistical reform have increased. Often, it has been argued that the Bayesian approach offers a promising alternative to classical statistics, but at the same time, both approaches are sometimes hard to compare. This paper offers some basic steps for establishing a more decisive common ground and take steps towards resolving some fundamental conceptual issues, by introducing the frameworks of epistemic accuracy and efficient experimentation. Efficiency (i.e. the ratio between the accuracy of conclusions, and the experimental resources needed) is of central importance to good scientific inquiry. This paper shows how considerations of efficiency and accuracy motivate a broadly Bayesian approach, which however can also make room for frequentist tenets, and thereby push us towards a more unified and powerful framework for scientific inference.

Keywords:

Replication crisis; Optimal Experimental Design; Objective Bayesian Statistics; Epistemic Accuracy; Efficiency; Information Theory;

Introduction

Turmoil has engulfed the scientific community. Advocates of two competing statistical paradigms—the Bayesians and the frequentists—are engaged in a perennial dispute about the correct approach to statistical inference. Deborah Mayo (2018) has called this never-ending conflict the 'statistics wars'[1]. For a long time, the frequentist (or 'classical') school, in the form of null hypothesis significance testing (NHST), was the dominant paradigm for statistical analyses in cognitive science and psychology (as in other areas of science, too). More recently, however, calls for a 'Bayesian revolution' – which we may also call a 'return of the Bayesians'[2] – have grown louder, in response to frustrations about p-values (including notorious misinterpretations and misuses) and the recent replication crisis in psychology (Beaumont and Rannala, 2004; Colling and Szűcs, 2021; Hahn, 2014; Shultz, 2007). The Bayesian approach indeed promises to solve many of the eminent problems that emerged in the wake of these crises (Wagenmakers, 2007) – yet meanwhile, the foundational dispute still rages on.

In this paper, I seek to establish a decisive common ground and take steps towards resolving some of the most fundamental issues by introducing the frameworks of epistemic accuracy and efficient experimentation. The next section briefly motivates efficient experimentation and accuracy in scientific inference as a central performance metric for comparing statistical methods, followed by a (light, tutorial-like) demonstration of Bayesian efficiency in a generic (simple but quite common) statistical setting. This provides the basis for a short, conceptual discussion of a generalised framework (based on the novel concept of E-variables) that extends the advantages of the Bayesian approach to more general settings (in particular, the general possibility of optional stopping, which is of special relevance to psychological and medical research). After this, taking one step back, I will make the notion of epistemic accuracy more precise, in order to develop a maximally general outlook on optimal probability updating, inference, and efficient experiment selection. Finally, I will use the accuracy-based framework to discuss a central (alleged) point of theoretical conflict between Bayesians and frequentists – the likelihood principle – and then conclude with a brief, final outlook.

Efficiency in Scientific Inquiry

Intuitively, efficiency in scientific inquiry can be understood as maximising the ratio between (i) the accuracy of the conclusions drawn from given data (more on this later) and (ii) the cost of collecting those data. The importance of efficiency is obvious, especially when data collection is expensive or opportunities are limited. To the extent that choosing a specific statistical method can have an influence on overall efficiency (e.g. because fewer data are needed to reach the same level of accuracy), we should prefer the method that yields higher efficiency. Despite countless battles having been fought in the 'statistics wars' (which often reveal that both approaches are ultimately hard to compare, see e.g. Colling and Szűcs, 2021), the efficiency of both approaches can be fruitfully compared in several settings, and this may work as a tie-breaker in those settings. As we will see later, treating the two approaches as comparable in this way can also illuminate conceptual bridges between them, potentially helping us move beyond the conflict and toward a stronger, unified framework.

[1]I will instead propose 'stats wars' – for more dramatic effect (I thank my dear colleague Klee Schöppl for this suggestion).

[2]Since historically, the school of 'inverse conditional probability' (which performs inference via Bayes' theorem) is in fact older than frequentism, which emerged only in the 20th century.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

Bayesian Efficiency and Optional Stopping

In order to start comparing the efficiency of both approaches, I will use a simple and common experimental framework. Suppose we are repeatedly observing some real-valued random variable X, where all subsequent observations are independent from previous outcomes. There are two hypotheses of interest, one null hypothesis H_0, predicting an effect size of $\delta = 0$ and some alternative(s) H_1 that predicts some effect size $\delta > 0$. This general setup is representative of many standard experimental designs, such as measuring participant's response times in two different conditions, or patient's recovery times in a medical trial with a treatment- and a control group, and so on. The goal of performing this experiment is to decide on one of the two hypotheses (i.e. either – tentatively – accept the null or the alternative), and we want to do this as efficiently as possible. That is, given a fixed target-level of accuracy (e.g. desired type I and type II error rates), the goal is to minimise the number of experiments needed, or equivalently, for a given number of experiments, maximise accuracy (minimise error rates). Let us denote the aspired type I error level as α, and the target type II error as β (the complement, $1 - \beta$ is also referred to as statistcal *power*).

The Neyman-Pearson Lemma (Neyman & Pearson, 1933) shows that for given α and sample size n, the *likelihood-ratio test* (LRT) , if it is well-defined in the given setting, maximises statistical power. That is, for a given main hypothesis H_0 and some specific alternative H_1, we accept H_1 (reject H_0) if and only if

$$\frac{f_1(x)}{f_0(x)} > \gamma, \tag{1}$$

for some constant γ (the *region of rejection*) subject to the constraint that $P(\text{reject } H_0 \mid H_0) \leq \alpha$, and where $f_i(x) = P(x \mid H_i)$ is the likelihood[3] hypothesis i assigns to the observed data x (the observed value of our random variable X). Many classical hypothesis tests are derived from or indirectly connected to this relation. If the LRT maximises power $(1-\beta)$ for given α and n, then conversely, we can derive from it the optimal sample size n that maximises efficiency for given error rates α, β. This is the idea behind power analysis and sample size calculations.

However, the problem is that this optimality result only holds if the task is to choose among two given hypotheses, with a fixed effect size. It doesn't necessarily hold if we want to efficiently determine whether there is *some* effect of unknown size. It is well-known and intuitively clear that larger effect sizes require smaller samples to be detected. So, testing H_0 against an alternative H_{min} that implies the *smallest* measurable effect size δ_{min} might be maximally inefficient, if the true effect size is large (we could have done with much less data). On the other hand, if δ is actually small, then testing against an alternative with a large effect size leads to a loss of power (i.e. higher type II error). One natural-seeming solution might be to proceed 'sequentially', that is, start with the largest reasonable effect size, and if we fail to reject, continue with a smaller effect size, and repeat until we either eventually reject H_0 or fail to reject even at H_{min} (in which case we eventually keep H_0). To make this more efficient, one might naively think that it would be good if we could include the previous data in all our subsequent analysis steps, i.e. always test with respect to the *whole* dataset that we have collected so far. But now, the classical statistician is alarmed: this is data peeking! 'Testing until significant' is considered a questionable research practice (a form of p-hacking) that leads to inflated type I error, i.e. the *nominal* α at which we are testing (our official '*target* type I error') is actually not the *effective* type I error that results from this procedure (the constraint for γ is silently violated). So, either we need to collect a new sample for every test, and throw away our old data, or *adjust* the nominal α level, in order to keep the desired type I error under control. A straightforward way of doing this is the Bonferroni correction (for the k-th test, and target-level α, define α/k as the new nominal significance level), but modern statistics has developed many different – much more sophisticated and elegant – methods for error control (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001; García & Berger, 2005; Storey, 2025; Tamhane, 1996).

Any of these corrections involves a trade-off between either increasing type I or type II error. The Bonferroni controls type I error very strictly, but this means that type II error increases massively (so, much larger samples are needed to regain power). For other methods, the trade-off between type I and type II error can be made more flexibly, which is good in the case of performing *simultaneous* multiple comparisons (i.e. testing for effects in different, logically independent dimensions based on the same dataset[4]). However, in the case of our sequential analysis, the goal is to efficiently detect whether there there is *one* effect (of unknown size), and we set out doing this with a given target-level of accuracy in mind, i.e. *fixed* type I and type II error rates. So, the only way to guarantee these error rates in a classical framework is to crank up the sample size (which may become a costly enterprise or not possible at all – at least not at the desired level of accuracy).

Not all hope is lost, however. In fact, there *is* a 'correction' that can do with fewer samples, and that demonstratively improves upon the original approach via the Neyman-Pearson lemma: a 'correction' that involves Bayesian updating. This approach was already explored and demonstrated to improve upon classical testing (in terms of efficiency) by Schönbrodt et al. (2017), but let us try to retrace and understand the mechanics of this procedure within our current setting. Going back to the Neyman-Pearson lemma and LRT (eq. (1)), we find a close connection to Bayes' theorem, which can be expressed

[3] Strictly speaking, frequentist hypothesis tests involve the *cumulative* density $f_i(X \geq x)$ and/or $f_j(X \leq x)$, i.e. the conditional probability of observing an outcome that is *at least* as extreme (depending on which tails of the distribution are tested), but for our purposes we can leave this aside.

[4] For example, from fixed set of patient data, analyse the positive effects of a drug, but also various potential side-effects.

as a relation between prior- and posterior odds, mediated by the likelihood ratio (LR):

$$\frac{P(H_1 \mid x)}{P(H_0 \mid x)} = \frac{P(H_1)}{P(H_0)} \cdot \frac{f_1(x)}{f_0(x)}, \tag{2}$$

where $P(H_i)$ is the prior probability assigned to hypothesis i, $P(H_i \mid X)$ is the posterior probability after having observed x, and $f_i(x) = P(x \mid H_i)$ are the likelihoods. Note that if the prior is *uniform*, the right-hand side of equation (2) reduces to the LR, which also figures in Neyman-Pearson's LRT. Furthermore, by relying on Bayes' rule, prior odds can be updated iteratively via equation (2). First, let $Odds_t(H_1 \mid x) = \frac{P_t(H_1|x)}{P_t(H_0|x)}$ and initialise $Odds_t(H_1)$ for $t = 0$ (the initial prior odds before having observed any data), e.g. with the uniform prior. Whenever we observe x at $t \geq 1$, we update via $Odds_t(H_1) = Odds_{t-1}(H_1 \mid x)$. Since starting from a uniform prior makes the posterior depend only on the LR, this means that we can decompose n observations (a product of observed LRs, $\prod_i \frac{f_1(x_i)}{f_0(x_i)}$) into observations made *before* and *after* some time $t < n$, by storing observations before t in the updated prior odds $Odds_t(H_1)$. Now, as in the Neyman-Pearson lemma, we can also define a decision threshold, which the product of LRs must surpass, in order to reach a final decision as to whether we prefer H_1 or H_0. The important difference, however, is that we don't just have *one* decision threshold γ, but *two*: either $Odds_t(H_1) \geq \gamma_1$ (*upper* threshold) or $Odds_t(H_1) \leq \gamma_0$ (*lower* threshold). We can make these thresholds symmetric (corresponding to a single stopping posterior $P(H_i) \geq \tau$ for either hypothesis, e.g. with $\tau = 0.95$) or *asymmetric*: for example, stop if either $Odds_t(H_1) \geq \frac{1-\beta}{\alpha}$ ($\rightarrow$ prefer H_1) or $Odds_t(H_1) \leq \frac{\beta}{1-\alpha}$ ($\rightarrow$ prefer H_0), where α, β correspond to the desired type I and type II error rates. Since the uniform prior makes the decision only dependent on the observed likelihoods, the resulting decisions will be calibrated to the respective target error rates α, β (i.e. the probability of preferring H_1 while H_0 is true will be at most α or slightly lower, and analogously the probability of erroneously preferring H_0 while H_1 is true will be at most β, or slightly lower). Now, suppose we fix H_1 as H_{min} (i.e. the minimum detectable effect size). What if the true effect size is larger (i.e., assuming a one-sided test, $\delta^* > \delta_{min}$)? This reveals one central, decisive advantage of the updating- and decision procedure just described. If we just stick with H_0 vs. H_{min} (minimal positive effect), but the true effect size is larger, the product of the likelihood ratios will converge to the upper threshold much more quickly than they usually would if the true effect size was *exactly* δ_{min}. It can even be shown (in line with the experiments performed in Schönbrodt et al., 2017) that this procedure is no less efficient on average than if we had 'luckily' chosen the true H^* as our first alternative to H_0 and just conducted the classical LRT. Moreover, if we compare this Bayesian-inspired sequential procedure to the 'classical' one described above (involving *classical* corrections for multiple testing), the efficiency of the Bayesian approach is much higher, especially for moderate-to-small effect sizes (i.e. if the 'classical' procedure means starting high, and then working downward). Conversely, compared to a standard LRT assuming H_{min} as the alternative, the Bayesian approach gains much more efficiency if the true effect size is even just moderately high (since the LRs will drift more quickly, as explained above). Finally, we don't need to worry about how to set up the set of alternative hypotheses in our sequential analysis (starting high and working downwards), because the LRs will converge *automatically* where the speed just depends on where on the continuum the true δ^* is (note that, even if the true effect size is *negative*, i.e. against all initial considerations, the treatment actually has a negative effect on the target variable X, we will simply converge on H_0 much more quickly, and even be able to notice that our initial idea of $\delta^* \geq 0$ was faulty[5]). In other words, the Bayesian approach allows us to do efficient and flexible *optional stopping* ('data peeking' transformed into a valid procedure), without blowing up *either* error rates *or* sample sizes, and even without having to set up an explicit sequence of alternative hypotheses with increasing effect sizes – the convergence automatically follows the continuous value of the true δ^*.

As mentioned, these observations are not really new. Apart from the aforementioned work, related things have been pointed out in the literature (Rouder, 2014; Wagenmakers, 2007; Wagenmakers et al., 2012) , in particular concerning the possibility of optional stopping, and the fact that resulting 'decisions' are correctly calibrated for frequentist error rates (Edwards et al., 1963; I. Good, 1991; Lindley, 1957), at least for the kind of scenario presented here. Specifically, Schönbrodt et al. (2017) found in their experiments (which even involved more complex setups and prior distributions than the one presented here) that the average sample size of the 'sequential Bayes factor design' (i.e. the Bayesian optional stopping approach presented here) is up to 75 % smaller than the sample size required for a classical design with corresponding error rates.

So, this is good news. Importantly, however, this ability to perform optional stopping (at the desired calibration level) may break down in more complex settings involving composite hypotheses (De Heide & Grünwald, 2021; Sanborn & Hills, 2014). Contemporary research in statistics aims to overcome the limitations posed by these more complex settings and to develop a generalised framework that can control optional stopping in general. This is the framework of '*safe tests*' based on E-values, which we will discuss in the next section. We will see that, while this framework indeed extends the standard Bayesian approach (with precise probabilities) to optional stopping, and establishes further connections to classical hypothesis testing, it is actually part of an *even more general* Bayesian framework that allows for *imprecise* prob-

[5]Relatedly, based on the results of the experiment, we can seamlessly proceed with a Bayesian *estimation* of the true effect size, i.e. starting from an extended prior distribution over the whole hypothesis space, determine the posterior expected value, a credible interval, or the maximum-posterior estimate given the data. If in fact $\delta^* < 0$, we will see this in the resulting estimate.

abilities, and offers a more nuanced perspective on optional stopping, and the more fundamental issue of efficiency, in general.

E-Variables: a Generalised Framework

Recent work (Grünwald et al., 2024; Ramdas et al., 2023; Turner, 2023) employs the novel concept of *E-variables* to build a framework for *Safe Anytime Valid Inference* (SAVI), and E-tests for parametric and non-parametric statistics (Vovk & Wang, 2024). Given a *composite* hypothesis $\mathcal{H}_0$ (with the special case of a single point hypothesis H_0), an E-variable is a random variable E such that the expected value is at most one, i.e. $\mathbb{E}_P(E) \leq 1$, for all $P \in \mathcal{H}_0$. E-variables are closely related to betting- and martingale theory (Doob, 1971), and can be related to classical p-values: via Markov's inequality, a conservative p-value can be derived from an observed E-value as $p = 1/E$.

The big advantage of E-variables (as is advertised in the label 'safe *anytime valid* inference') is that they allow for optional stopping and continuous monitoring of data, as new samples come in, without jeopardising target error rates. One instance of such an E-variable is the error correction via observed likelihood ratios, as demonstrated in the previous section. Moreover, E-values still work in cases where composite null- or alternative hypotheses are compared against other – simple or composite – hypotheses. In some of these cases, it has been pointed out that the standard Bayesian approach cannot handle optional stopping while guaranteeing desired error rates. In particular, in the case of composite hypotheses[6] $\mathcal{H}$, the standard Bayesian approach requires an exact prior over $\mathcal{H}$, in order to define the likelihood $P(X \mid \mathcal{H})$, and these priors (e.g. in the case of *some* continuous 'improper priors') may not have the properties required to deal with optional stopping adequately (see also discussion in De Heide and Grünwald, 2021). Within SAVI, however, likelihoods can be defined e.g. by taking the lower bound of the hypothesis space, or select based on loss- and growth criteria (e.g. minimise worst-case expected loss). Approaches based on E-variables use more general concepts from martingale theory to derive growth rates that represent how evidence against the null hypothesis accumulates over time. In order to define growth rates for evidence accumulation, often 'hyperparameters' are needed that are somewhat analogous to Bayesian priors. However, these *learning rates* will standardly only affect the *efficiency* of the test (i.e. the data needed to reach a decision at a desired accuracy level), but not the *error rates*.

Indeed, this development seems very promising for making research and statistical data analysis more efficient. Consequently, there are some important conceptual observations and connections to be made, in order to push the development and potential applications even further. First of all, we may note that the general setup involving E-variables is in fact an *imprecise probability* model. That is, instead of defining a single, precise probability- or likelihood distribution over outcome- and hypothesis spaces, we allow for *sets* of probability distributions. In particular, in the case of a composite hypothesis, we are really dealing with a (dense) set of possible statistical models (probability distributions) without defining a precise prior over them (i.e. we may entertain any *set* of priors). Such instances *can* be handled within a *more general* Bayesian framework that allows for imprecise priors and likelihoods, where every single probability model within the set is updated by Bayes' theorem. However, this can easily become computationally intractable. Therefore, the E-variable framework has practical advantages insofar as it offers more *computational* efficiency. Choices of learning parameters and growth criteria (e.g. based on bounding worst-case expected loss) correspond to focussing on specific subsets of interest, within a larger imprecise probability model. In the next section, I will introduce a more general, accuracy-based framework for selecting optimal (prior and updated) probability distributions, which can give additional underpinnings to 'focused' imprecise probability models like SAVI.

In this context, it is also interesting to see that E-variables and related concepts may be another important step towards a (long-sought-after) convergence of classical and Bayesian approaches. Since the presentation of E-variables is partly framed within a classical framework (see relation to p-values), we may also say that classical statistics is now taking steps towards incorporating more Bayesian ideas.

However, as a final point, it is also important to stress that optional stopping, even if it can be controlled, is not necessarily always desirable in terms of *efficiency*. While it is certainly advantageous to have methods that can correctly 'handle' optional stopping whenever the necessity arises (e.g. budgets or time constraints forcing research to be interrupted), there are cases where having a more specific and elaborate sampling plan is substantially more efficient (in the sense of optimising the ratio between experimental resources needed and information gained), and thus preferable. Thus, in order to develop a more general perspective on these issues, the next section introduces the framework of epistemic accuracy.

Epistemic Accuracy and Efficiency

A major goal of science and statistics is to produce accurate inferences and estimates. In the literature on forecasting, information theory, and formal epistemology, the concept of *scoring rules* is fundamental to measuring the accuracy of probabilistic estimates, forecasts, or 'credences' (Brier, 1950; Crupi et al., 2018; Gneiting & Raftery, 2007; McCutcheon, 2019; Pettigrew, 2016). A *measure of inaccuracy* (or *scoring rule*) is a non-negative real-valued function $I(X, P)$, from some set of random variables X (an outcome space) of interest, and a probability function P, which measures the inaccuracy

[6] Composite hypotheses correspond to sets of probability distributions (for example, in a binomial experiment, a point hypothesis is a single success probability, like $p = 0.51$, while a composite hypothesis specifies a region of the probability space, like $p > 0.5$, corresponding to a collection of many point hypotheses). In the generic example from the previous section, we only discussed point hypotheses.

of the prediction of $P(x)$ for different possible materialisations x of X. We will only require two very general properties for I (which will become obvious quite immediately):

1. **Truth-directedness**: $\frac{\partial I(X,P)}{\partial P(x)} < 0$ (i.e. inaccuracy decreases with increasing probability assigned to actual events)
2. **Strict Propriety**: the *expected* inaccuracy $\mathbb{E}_Q[I(X,P)]$ of P under some distribution Q is minimal if and only if $P = Q$.

Truth-directedness is self-explanatory, while strict propriety might seem mysterious at first (though it really isn't). A standard explanation given in Bayesian formal epistemology is the following. Q is the agent's credence function, which they expect to be optimal (reporting their '*true* credences', i.e. $P = Q$, minimises expected inaccuracy at their current stage). This is the internalistic perspective. However, there is also an *externalistic* perspective that is usually overlooked in Bayesian epistemology. If Q is the limiting relative frequency of an event, then adopting Q as one's probabilistic estimate ('credence') is maximally accurate on expectation.

These two fundamental properties of adequate scoring rules yield two central implications:

1. Updating prior probabilities by minimising the expected inaccuracy of the posterior subject to propositional constraints (i.e. based on observations of any set of random variables $X = x_1, \ldots, x_n$) yields updating via *Bayes' theorem* (Rosenkrantz, 1992); ***and***
2. All adequate estimates (based on proper scoring rules) converge in actual relative frequencies in the limit (whenever they are well-defined).

Whenever there are *no* relative frequencies (e.g. for scientific theories), of course, different estimates based on different scoring rules will differ. But this is not a problem, because optimal estimates are relative to the inference problem we want to solve. Different scoring rules can be interpreted as different utility/loss functions[7] that reflect *which aspects* in a given problem of inference or estimation are relatively important to get right compared to others (i.e. how are different kinds of errors weighted against each other, see also Levinstein, 2017).

Together with different decision rules (e.g. minimisation of worst-case (expected) loss, minimisation of total sum loss, some combination or something else entirely) we can also make a selection for focal (sets of) prior distributions[8]. This also coincides with and offers some further grounding for the parameter selection question in the E-variables framework.

Most importantly, however, epistemic accuracy informs efficient experiment selection (Crupi et al., 2018; Meder et al., 2022; Rainforth et al., 2024). Given a measure of information gain Inf, based on a measure of inaccuracy I, and a cost function C over possible experiments, the goal is to maximise the (expected) ratio of information gain to cost $\mathbb{E}[Inf/C]$. If we have a fixed budget to conduct a number of experiments, the goal is to select a feasible sequence of experiments that maximises Inf (or equivalently, minimises expected posterior inaccuracy). If experiments are independent (like in our initial scenario) minimising expected posterior inaccuracy straightforwardly entails optional stopping. However, there may be interdependencies between different experiments which make a more complex set of contingency plans more efficient (this is non-trivial, because it can imply that it is better in the long run to choose an experiment that *doesn't* maximise mutual information at the next stage, but the overall sequence does, as opposed to a sequence that 'myopically' maximises mutual information at every step).

The accuracy-first perspective gives us a very fundamental framework that provides updating- or inference rules (via distance minimisation) that are even more general than Bayes' theorem (see also (Eva & Hartmann, 2018)). Furthermore, priors can be defined in non-arbitrary ways, and thereby accommodate the desire for adequately constraining subjectivity in scientific inference, which is (rightfully) articulated by advocates of classical statistics. Thus, taking the accuracy-first perspective seriously might eventually help us bridging the trenches of the statistics wars once and for all.

However, when it comes to performing *inference*, there is still a perceived fundamental conflict between the two statistical paradigms, which consists in the so-called *likelihood principle*. We will discuss this in the next section, where I sketch a fundamental accuracy-based assessment of the question.

The Likelihood Principle

Loosely stated, the likelihood principle says that all information an experimental outcome x conveys about any statistical hypothesis H_i is contained in the likelihood function $f_i(x)$. That is, the likelihood function is the complete basis for making statistical inferences from data. More formally, it has been spelled out as follows (Berger & Wolpert, 1988; Edwards et al., 1963): given two hypotheses H_i, H_j, two experimental outcomes x, y are informationally equivalent, precisely if there is a positive constant k such that $f_i(x) = k \cdot f_j(y)$.

Based on this, it has been pointed out that the likelihood principle, thus formulated, means that *the stopping rule doesn't matter* (Fletcher, 2024). This has been perceived as a major point of conflict between (subjective) Bayesians and fre-

[7] Nevertheless, in the *specific* context of scientific inquiry (including purely theoretical research that doesn't care about practical utilities, see e.g. (Fisher, 1935)), the *logarithmic score* $I(X = x, P) = -\log(P(x))$ is a perfect default rule, because it intuitively expresses accuracy roughly as 'the minimum number of questions needed to identify the value of X', and the expected value $\mathbb{E}[I(X,P)]$ (Shannon entropy) corresponds to the expected length of a search tree (or encoding scheme) over X that is optimised for P, as demonstrated by Shannon's source coding theorem (Shannon, 1948). See also (I. J. Good, 1952; McCutcheon, 2019)

[8] For a classical information-theoretic perspective, see also (Berger et al., 2024; Bernardo, 2011), and more objective Bayesianism following the principle of Maximum Entropy: (Jaynes, 2007; Landes & Williamson, 2013; Williamson, 2010)

quentists. While the former argue that it is absurd and downright superstitious to have inferences depend on the 'intentions of the experimenter' (i.e. not only what *has been* observed, but also what '*could have*' been observed, depending on the plan), classical statisticians (especially of the error-statistical variety) counter that *what could have been observed* is precisely relevant for determining *error* probabilities, and this has nothing to do with the 'intentions' of the experimenter (if the result was already guaranteed a priori, in light of the sampling setup, the experiment is not a severe test of the hypothesis, see also Fletcher, 2024; Mayo, 1996).

To make this point on behalf of error statistics, Mayo & Kruse (2001) discuss the following example. Suppose we repeatedly and independently draw from a given population to observe some real-valued variable X of interest, and calculate the sample mean, which is normally distributed with $X \sim \mathcal{N}(\mu, 1)$. The question is: what is the true population mean μ? Suppose that our stopping rule is as follows. Terminate, when the observed sample mean is two standard deviations away from 0, i.e. $\overline{x}_n \geq 2/\sqrt{n}$ (where $\overline{x}_n$ is the sample mean after n draws). Given a non-informative (uniform) prior, the posterior calculated from $\overline{x}_n$ is $\mathcal{N}(\overline{x}_n, 1/n)$. This conclusion, which is definitely far from $\mu = 0$ (*at least* two standard deviations), even if $\mu = 0$ is actually true, can *always* be achieved, since the stopping rule will terminate almost surely (whatever μ is). Thus, more generally, it seems that a 'persistent experimenter' of the particularly shady kind can generate any conclusion they want, independently of the underlying truth. This is troubling – isn't it?

In fact, the trouble in this scenario results from ignoring the specific stopping *time* in a case where it is indeed relevant. Intuitively, if the stopping threshold is far away from the true mean, it will take much longer for the procedure to terminate than if the stopping threshold is close to the true mean.

Given the improper flat prior on $(-\infty, \infty)$, the resulting posterior distribution is determined by the likelihood:

$$p(\mu \mid \overline{x}_n, n) \propto \varphi(\overline{x}_n - \mu) \prod_{k=1}^{n-1} \Phi\left(\frac{2}{\sqrt{k}} - \mu\right) = p(\overline{x}_n, n \mid \mu), \quad (3)$$

where φ and Φ, respectively, are the PDF and CDF of the standard normal distribution[9]. The likelihood is the conditional joint probability $p(\overline{x}_n, n \mid \mu)$, which expresses the probability of stopping after exactly n trials *and* observing a value of $\overline{x}_n$, given that μ is the true mean (the contribution of observing $\overline{x}_n$ is the density φ, whereas the contribution of n is the product of cumulative densities $\prod \Phi$, which expresses the probability of a sequence of independent trials that terminates after exactly n draws). The fact that we also condition on n pushes the centre of the resulting posterior down, often very much below the observed $\overline{x}_n$, and in fact, close to the true mean (though this has to be simulated numerically). For example, if $n = 25$, and $\overline{x}_n = 0.8$, then the posterior mean is about -1.28, the mode (i.e the maximum posterior estimate) about -1.17, and a 95% credible interval includes $[-2.4; -0.42]$.

Thus, summing up: yes, Bayesians *can* include the stopping rule in the likelihood, no problem (despite Mayo & Kruse arguing against it vehemently). The point is that, when we have the *concrete sequence of observations*, a non-informative stopping rule (which only depends on the current data, and not the hypothesised parameter) amounts to a constant that cancels out in Bayes' theorem (see e.g. Gandenberger, 2017), and thus makes no difference to the posterior distribution. However, when we only have a *summary* of the data (like the observed sample mean that is used in the example above), the stopping rule (the concrete stopping time in this case) can make a significant difference, as just demonstrated. Thus, not only can Bayesian inference handle differences in terms of stopping rules, but *crucially*, an existing result quoted in the last section (Rosenkrantz, 1992) shows that inference via Bayes' theorem *minimises expected inaccuracy* for all proper scoring rules, given any propositional evidence. Thus, whenever the stopping rule cancels out, it is – indeed – irrelevant in terms of accuracy. Nevertheless, if we don't have the precise data sequence, it can still be strongly informative, in which case it *must* be included for a correct, inaccuracy-minimising inference. Ignoring *any* piece of information that is relevantly connected to the evaluation of the hypothesis would be ignoring the *principle of total evidence* (I. J. Good, 1966), which is firmly grounded in the fundamental principles of inaccuracy minimisation and value-of-information considerations. Apart from that, the likelihood principle itself is not really fundamental to accuracy-based Bayesian inference, and it can even be occasionally 'violated', depending on how one looks at it (e.g. in the framework of Bernardo, 2011). What *is* fundamental, however, is the perspective of epistemic accuracy, and whenever the likelihood principle is justified, it falls out as a consequence of this more fundamental framework.

Conclusion

Have the Bayesians thus returned? It seems that they are about to, even if slightly transformed, or 'in the guise' of new alternative approaches like SAVI, which creates interesting and important conceptual bridges between both paradigms. In line with that, there is hope that the accuracy-based perspective helps to further deepen these bridges, and to move towards a more general framework for efficient experimentation and optimal statistical inference. Eventually, an optimal, maximally general statistical framework will help us overcome recent crises, draw out the last bit of genuine information from any small dataset, let us efficiently aggregate many small studies into a coherent and informative overall picture, and be able to robustly handle various research practices. Thereby, even though the accuracy-based framework is still very general and theoretical, we can hope that it will soon become increasingly useful in practice.

[9] For a full posterior distribution this must be normalised by $\int_{-\infty}^{\infty} \varphi(\overline{x}_n - \mu) \prod_{k=1}^{n-1} \Phi(\frac{2}{\sqrt{k}} - \mu) d\mu$, which has no closed form solution but must be evaluated numerically.

Acknowledgements

I gratefully acknowledge support from the Graduate School of Systemic Neurosciences and from the Chair of Philosophy of Science at the Munich Center for Mathematical Philosophy, LMU Munich. I thank Klee Schöppl, Matteo Michelini, and Felix Kopecky for helpful feedback on an early draft, and Stephan Hartmann, Ulrike Hahn, and Stephan Sellmaier for their precious advice and support.

References

Beaumont, M. A., & Rannala, B. (2004). The bayesian revolution in genetics. *Nature Reviews Genetics*, *5*(4), 251–261.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, *29*(4), 1165–1188.

Berger, J. O., Bernardo, J.-m., & Sun, D. (2024). *Objective bayesian inference*. World Scientific.

Berger, J. O., & Wolpert, R. L. (1988). The likelihood principle. *6*.

Bernardo, J. M. (2011). Integrated objective bayesian estimation and hypothesis testing. *Bayesian statistics*, *9*, 1–68.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1–3.

Colling, L. J., & Szűcs, D. (2021). Statistical inference and the replication crisis. *Review of Philosophy and Psychology*, *12*(1), 121–147.

Crupi, V., Nelson, J. D., Meder, B., Cevolani, G., & Tentori, K. (2018). Generalized information theory meets human cognition: Introducing a unified framework to model uncertainty and information search. *Cognitive science*, *42*(5), 1410–1456.

De Heide, R., & Grünwald, P. D. (2021). Why optional stopping can be a problem for bayesians. *Psychonomic Bulletin & Review*, *28*(3), 795–812.

Doob, J. L. (1971). What is a martingale? *The American Mathematical Monthly*, *78*(5), 451–463.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological review*, *70*(3), 193.

Eva, B., & Hartmann, S. (2018). Bayesian argumentation and the value of logical validity. *Psychological Review*, *125*(5), 806–821.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh: Oliver; Boyd.

Fletcher, S. C. (2024). The stopping rule principle and confirmational reliability. *Journal for General Philosophy of Science*, *55*(1), 1–28.

Gandenberger, G. (2017). Differences among noninformative stopping rules are often relevant to bayesian decisions. *arXiv preprint arXiv:1707.00214*.

García, M. J. B., & Berger, J. (2005). The interplay of bayesian and frequentist analysis. *Quality control and applied statistics*, *50*(3), 321–322.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, *102*(477), 359–378.

Good, I. (1991). A comment concerning optional stopping. *Journal of Statistical Computation and Simulation*, *39*(3), 191–192.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *14*(1), 107–114.

Good, I. J. (1966). On the principle of total evidence. *The British Journal for the Philosophy of Science*, *17*(4), 319–321.

Grünwald, P., de Heide, R., & Koolen, W. (2024). Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *86*(5), 1091–1128.

Hahn, U. (2014). The bayesian boom: Good thing or bad? *Frontiers in psychology*, *5*, 765.

Jaynes, E. T. (2007). Prior probabilities. *IEEE Transactions on systems science and cybernetics*, *4*(3), 227–241.

Landes, J., & Williamson, J. (2013). Objective bayesianism and the maximum entropy principle. *Entropy*, *15*(9), 3528–3591.

Levinstein, B. A. (2017). A pragmatist's guide to epistemic utility. *Philosophy of Science*, *84*(4), 613–638.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*(1/2), 187–192.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.

Mayo, D. G., & Kruse, M. (2001). Principles of inference and their consequences. In *Foundations of bayesianism* (pp. 381–403). Springer.

McCutcheon, R. G. (2019). In favor of logarithmic scoring. *Philosophy of Science*, *86*(2), 286–303.

Meder, B., Crupi, V., & Nelson, J. D. (2022). What makes a good query?: Prospects for a comprehensive theory of human information acquisition. In I. Cogliati Dezza, E. Schulz, & C. M. Wu (Eds.), *The drive for knowledge: The science of human information seeking* (pp. 101–123). Cambridge University Press.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*, 289–337.

Pettigrew, R. (2016). *Accuracy and the laws of credence*. Oxford University Press.

Rainforth, T., Foster, A., Ivanova, D. R., & Bickford Smith, F. (2024). Modern bayesian experimental design. *Statistical Science*, *39*(1), 100–114.

Ramdas, A., Grünwald, P., Vovk, V., & Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, *38*(4), 576–601.

Rosenkrantz, R. D. (1992). The justification of induction. *Philosophy of Science*, *59*(4), 527–539.

Rouder, J. N. (2014). Optional stopping: No problem for bayesians. *Psychonomic bulletin & review*, *21*(2), 301–308.

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on bayesian hypothesis tests. *Psychonomic bulletin & review*, *21*(2), 283–300.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with bayes factors: Efficiently testing mean differences. *Psychological methods*, *22*(2), 322–339.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Shultz, T. R. (2007). The bayesian revolution approaches psychological development. *Developmental science*, *10*(3), 357–364.

Storey, J. D. (2025). False discovery rate. In *International encyclopedia of statistical science* (pp. 944–949). Springer.

Tamhane, A. C. (1996). 18 multiple comparisons. *Handbook of statistics*, *13*, 587–630.

Turner, R. J. (2023). *Safe anytime-valid inference: From theory to implementation in psychiatry research* [Doctoral dissertation, Leiden University].

Vovk, V., & Wang, R. (2024). Nonparametric e-tests of symmetry. *The New England Journal of Statistics in Data Science*, *2*(2), 261–270.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, *14*(5), 779–804.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on psychological science*, *7*(6), 632–638.

Williamson, J. (2010). *In defence of objective bayesianism*. Oxford University Press.