Erkenn (2013) 78:293–318
DOI 10.1007/s10670-011-9317-8

ORIGINAL ARTICLE

Why Frequentists and Bayesians Need Each Other

Jon Williamson

Received: 11 February 2011 / Accepted: 10 August 2011 / Published online: 21 August 2011

Abstract The orthodox view in statistics has it that frequentism and Bayesianism are diametrically opposed—two totally incompatible takes on the problem of statistical inference. This paper argues to the contrary that the two approaches are complementary and need to mesh if probabilistic reasoning is to be carried out correctly.

1 Introduction

1.1 The argument

The argument of this paper proceeds along the following lines.

Most versions of Bayesianism rightly invoke some principle of direct inference—such as the Principal Principle—for ensuring that prior probabilities are calibrated with known physical probabilities. But such a principle presupposes that physical probabilities can be determined independently of Bayesian prior probabilities. Since Bayesian methods for estimating physical probabilities depend on a given prior probability function, and it is precisely the prior that is in question here, this leaves classical (frequentist) estimation methods—in particular confidence interval estimation methods—as the natural candidate for determining physical probabilities. Hence the Bayesian needs the frequentist for calibration.

On the other hand, the frequentist also needs the Bayesian, for the following reason. The physical probabilities invoked by frequentists are generic—i.e., probabilities of repeatedly instantiatable attributes or events. But confidence interval estimation methods are only of interest to the extent that they can be used to generate a single-case interval estimate of a specific quantity, with the confidence

J. Williamson (✉)
Philosophy, SECL, University of Kent, Canterbury CT2 7NF, UK
e-mail: j.williamson@kent.ac.uk

level somehow indicating the extent to which the estimate can be relied upon. Now this confidence in a particular estimate is most naturally explicated using the Bayesian framework, since it is Bayesian probability that expresses the strength that one ought to believe a single-case proposition. Hence the frequentist needs the Bayesian in order to justify the application of frequentist methods to the single case.

1.2 Plan of the paper

The aim of the paper is to flesh out this argument and to show in more detail how frequentism and Bayesianism might fruitfully be combined. Section 2 presents the usual statistical view of the relationship between frequentism and Bayesianism—a view in which the two are incompatible. One needs to move to an epistemological perspective in order to understand how the two should be integrated (Sect. 3). Section 4 presents an extended example which places confidence interval estimation as the locus of integration. Sections 5 and 6 discuss two objections to that analysis. Section 7 draws conclusions and points to ways in which this research might be extended.

2 The Statistical View

2.1 Statistical inference

In this section we will encounter the orthodox view of the relationship between frequentism and Bayesianism—this is the view that they are competing paradigms for statistical inference.

The orthodox view is roughly that statistical inference works like this:

1. Conceptualise the problem and isolate a set $\mathbb{M}$ of models for consideration. Typically models are probability functions or can be thought of as probability functions.[1]
2. Gather evidence E.
3. Apply statistical methods to evaluate models in $\mathbb{M}$ in the light of E. Inferences and decisions will be made on the basis of a set $\mathbb{M}_E \subseteq \mathbb{M}$ of models that are appropriate given E.

Frequentist statistics instantiates this general pattern as follows:

1. Conceptualise the problem. Isolate a set $\mathbb{M}$ of models for consideration. Here $\mathbb{M}$ is a set of candidates for physical probability P^*. Frequentist statistics usually understands physical probability as either limiting relative frequency (von Mises 1928) or generic propensity (Kolmogorov 1933, §2), defined over attributes that can be repeatedly instantiated. Thus probability is relative to a set S which is construed by the limiting-relative-frequency approach as a collective

[1] Often—especially when the models in $\mathbb{M}$ are indexed by a set of parameters—statisticians use the singular word 'model' to refer to the set $\mathbb{M}$ itself. In line with the logicians' use of the term, in this paper the word 'model' will be reserved for a specific member of $\mathbb{M}$.

(an infinite sequence of outcomes) and by the propensity approach as a set of repeatable conditions that would generate a collective.
2. Gather evidence E.
3. Apply statistical methods to evaluate models in $\mathbb{M}$. In this case, inferences and decisions will be made on the basis of a set $\mathbb{M}_E \subseteq \mathbb{M}$ of models that render the evidence sufficiently likely, assuming the evidence is gathered in an appropriate way etc. The formal apparatus of frequentist statistical theory says how likely a model makes the evidence, but it is taken to be a pragmatic question as to whether the evidence is made sufficiently likely for a model to continue to be entertained.

On the other hand, Bayesianism is normally thought of as instantiating the scheme in a very different way:

1. Choose appropriate variables or models, $\mathbb{M}$, and a prior function P defined over $\mathbb{M}$. P is standardly interpreted as a belief function, representing rational degrees of belief, and is defined over single cases rather than repeatably instantiatable outcomes. On a subjective Bayesian account, P is largely a matter of personal choice, while on an objective Bayesian account P is largely constrained by evidence and the domain over which P is defined.
2. Gather evidence, ensuring that E is representable as an element e in the domain of P.
3. Adopt a new belief function $P' = P(\cdot|e)$ over $\mathbb{M}$ (this is *Bayesian conditionalisation*). Typically Bayes' theorem is applied at this stage: $P(m|e) = P(e|m)P(m)/P(e)$. One then can isolate a set $\mathbb{M}_E$ of models with sufficiently high posterior probability P'.

2.2 Incompatible?

From this orthodox perspective, it looks as if the frequentist and Bayesian approaches are simply incompatible ways of doing statistical inference: they implement the general statistical scheme in entirely different ways, employing different concepts and generating different sets $\mathbb{E}$ of models that are deemed appropriate on the basis of the evidence. Hence the standard view is that at most one of these two paradigms can be correct, and much energy has been directed at determining which one is correct.

But there is another way of looking at the relationship between the two approaches. We can ask what question each approach is trying to answer. Apparently, frequentism asks: how does evidence impact on the set of candidate physical probability functions? On the other hand, Bayesianism asks: how does evidence impact on rational degree of belief? Now, rational degree of belief is far from identical to physical probability: rational degree of belief is normally taken to be the basis for rational action (degrees of belief are often interpreted in terms of betting dispositions, for example) while physical probability is a physical quantity, akin to mass, charge or volume. Hence frequentism is intent on describing agent-independent features of the world while Bayesianism is intent on deciding how an

agent should act. Under this perspective, the two approaches do not seem so incompatible after all.

3 The Epistemological View

This last perspective construes Bayesianism as an epistemological theory at root. This view of Bayesianism is called *Bayesian epistemology* and is explored in more detail in this section.

3.1 Bayesian epistemology

Bayesian epistemology is concerned with the following key question: how strongly should an agent with evidence E believe the various propositions expressible in her language $\mathcal{L}$? Here E should be understood as the agent's *total evidence*, everything that is taken for granted in her current operating context: data, assumptions, theoretical knowledge etc.[2] Such evidence need not always be expressible as propositions of $\mathcal{L}$.

There are various Bayesian answers to this basic question, often based around one or more of the following three norms. Arguably, an agent's belief function P_E over $\mathcal{L}$ should satisfy:

Probability. P_E should be a probability function;
Calibration. P_E should satisfy constraints imposed by evidence: in particular, P_E should be calibrated with known physical probabilities where appropriate;
Equivocation. P_E should not award extreme degrees of belief (i.e., near 0 or 1) unless forced to by one of the above two norms: P_E should equivocate sufficiently between the basic possibilities expressible in $\mathcal{L}$.

Those Bayesians who adopt only the Probability norm (typically together with some rule of updating, such as conditionalisation) are known as *strict subjectivists*. (They are subjectivists because the choice of an initial *prior* belief function is largely up to the subject in question.)[3] Those who adopt both Probability and Calibration (again, typically with an updating rule) are sometimes called *empirically-based subjectivists*. Those who adopt all three norms are *objectivists*—the choice of belief function is much more highly constrained and correspondingly there is much less of a role for the subject to determine the belief function. No further updating rule is required in the case of objective Bayesianism, though updates turn

[2] It should be emphasised that such evidence may only be granted defeasibly. If a body of evidence leads to anomalous consequences, its more questionable elements will be withdrawn from the evidence base as they become open to criticism and are no longer taken for granted. See Williamson (2010b, §1.4.1) for further discussion of this notion of evidence.

[3] Advocates of *imprecise probability* reject even the Probability norm, representing a belief function by a set of probability functions rather than a single probability function. While this sort of view is not normally classified as Bayesian, some versions of this view admit analogies with Bayesianism (see, e.g., Walley 1991).

out to accord with conditionalisation in those cases in which conditionalisation is a plausible rule of updating (Williamson 2011b).

Strict subjectivism is a minority view in epistemology. It deems an agent who has evidence $E = \{$the chance of a particle of type S decaying is .01, particle s is of type $S\}$ yet believes that particle s will decay to degree .99, to be rational. However, most would deem such an agent to be irrational on the grounds that the evidence points the other way and warrants much less confidence that the particle will decay.

Calling strict subjectivism a minority view in epistemology is perhaps too generous—indeed, it is not clear that there are any proponents of strict subjectivism as an epistemological position. Bruno de Finetti and Colin Howson both advocate strict subjectivism, arguing that the Probability norm (together with conditionalisation, on de Finetti's account) is the only rational constraint on degrees of belief. However, that is apparently in a logical context—a theory of what it is for an agent's degrees of belief to be consistent—rather than an epistemological context, which requires a theory of what it is for an agent's degrees of belief to be rational (see, e.g., de Finetti 1937; Howson 2001).

Strict subjectivism is more common in statistics. In statistics, the hope is often that one can do away with an explicit Calibration principle by appealing to strict subjectivism and adopting a *pretend-prior* strategy: instead of creating a prior probability function by calibrating directly to available evidence of physical probabilities, create a prior under the pretence that this evidence is not available, and update the pretend prior by conditionalising on the data that gave rise to the evidence of physical probabilities; the resulting posterior probability function can be considered to be the genuine prior function given the evidence of physical probabilities that is actually available. There are several reasons why this strategy is not a live one from an epistemological point of view. First, as mentioned above, forsaking an explicit Calibration norm can lead to intuitively inappropriate degrees of belief, such as degree of belief 0.99 that a particle will decay knowing full well that the chance of it decaying is 0.01, which is but a short step away from Moore's paradox. Second, an explicit Calibration norm admits similar justifications to the Probability norm (see Sect. 3.2) and it is hard to commit to the latter without committing to the former. Third, the pretend-prior strategy leads to well calibrated posteriors only in the asymptotic limit, and only if certain assumptions are satisfied—e.g., the exchangeability assumption, which holds when the agent's prior probabilities do not depend on the order in which outcomes occur. But such assumptions tend to lack independent normative justification and are certainly not a rational requirement of strict subjectivism, and hence there is no guarantee that they will be satisfied. Fourth, one can choose a pretend prior that would yield whatever genuine prior one wishes after updating on the data, so this strategy offers no normative constraint on degrees of belief. Fifth, in many cases—such as cases of testimony—one has evidence of physical probabilities without having access to the data that generated that evidence (e.g., one might have statistics of the whole sample rather than individual sample outcomes), and in these cases the pretend-prior strategy will not normally be implementable. Sixth, this strategy depends on conditionalisation, but there are several important situations in which one's updated degrees of belief should plainly *not* agree with the results of conditionalisation: e.g.,

if the evidence E has (pretend) prior probability 0; if learning the evidence E tells the agent more than simply $P(E) = 1$; if the evidence E is not expressible in the agent's language . This last point is developed in Sect. 7.2.

In sum, under the epistemological perspective, strict subjectivism is on weak ground. It will not be considered a serious option here.

Objective Bayesianism is also a minority view in epistemology (largely because the Equivocation norm is hard to justify), though it is more widely endorsed in the sciences. It is motivated by the consideration that the empirically-based subjectivist will deem an agent who has evidence $E = \{$the chance of a particle of type S decaying is in the interval [.01, .99], particle s is of type $S\}$ yet believes that particle s will decay to degree .99, to be rational, while many would deem such an agent to be irrational on the grounds that the evidence is equivocal and hence fails to endorse such extreme confidence.

Since the Calibration norm is to be the focus of this paper, our arguments will apply equally to empirically-based subjective Bayesianism and to objective Bayesianism and there is no need here to take a stance on which position to adopt.

3.2 An explication

We shall now sketch one way of fleshing out the three norms of Bayesian epistemology. Nothing much will hang on the particular interpretation of the norms presented here, but it will be useful to adopt a concrete explication in which to frame the extended example of Sect. 4. The full details of the approach given here can be found in Williamson (2010b).

The simplest case is perhaps that in which $\mathcal{L}$ is a finite propositional language on propositional variables $A_1, \ldots, A_n$ with sentences $S\mathcal{L}$ formed by applying the usual connectives $\neg, \vee, \wedge, \rightarrow, \leftrightarrow$. In that case the three norms can be explicated as follows:

Probability. P_E should be a probability function:

P1: $P_E(\omega) \geq 0$ for each $\omega \in \Omega = \{\pm A_1 \wedge \cdots \wedge \pm A_n\}$,
P2: $P_E(\tau) = 1$ for some tautology $\tau \in S\mathcal{L}$, and
P3: $P_E(\theta) = \sum_{\omega \models \theta} P_E(\omega)$ for each $\theta \in S\mathcal{L}$.

The justification of this norm usually appeals to the *Dutch book argument*: if an agent bets according to her degrees of belief and the norm is not satisfied, then these betting commitments can be (and in the worst case, will be) exploited to force her to lose money whatever happens (i.e., to force positive expected loss); on the other hand, if the norm is satisfied then it is not possible to exploit the agent and worst-case expected loss is zero. Thus the norm should hold in order to minimise worst-case expected loss.

Calibration. P_E should be compatible with evidence,

C1: $P_E \in \mathbb{E} = \langle \mathbb{P}^*_{\mathcal{L}} \rangle \cap \mathbb{S}$.

Here $\mathbb{P}^*$ is a set of candidate physical probability functions: according to the agent's evidence, the physical probability function P^* lies in $\mathbb{P}^*$. $\mathbb{P}^*_{\mathcal{L}}$ is the specialisation of the information that $P^* \in \mathbb{P}^*$ to $\mathcal{L}$. As to how $\mathbb{P}^*_{\mathcal{L}}$ is to be understood depends on how

physical probability itself is understood. On the one hand, if physical probability is thought of as defined over single cases (P^* is single-case *chance*, applying to events such as the next throw of a particular die) then one can define $\mathbb{P}^*_{\mathcal{L}}$ to be the restriction of the set $\mathbb{P}^*$ of potential chance functions to $\mathcal{L}$: $P \in \mathbb{P}^*_{\mathcal{L}}$ if and only if P is defined on $\mathcal{L}$ and P is not inconsistent with every probability function in $\mathbb{P}^*$, in the sense that it is not the case that there is some sentence θ of $\mathcal{L}$ such that $P(\theta) \neq Q(\theta)$ for every function Q in $\mathbb{P}^*$. (Note that this allows for the possibility that not all the propositions expressible in $\mathcal{L}$ need have determinate physical probabilities, i.e., that $P^*(\theta)$ may not be defined for each sentence θ of $\mathcal{L}$.) On the other hand, if physical probability is thought of as defined over repeatably instantiatable outcomes (P^* is generic propensity or frequency, applying to outcomes such as an arbitrary throw of a particular die), it must first be specialised to the single case in order to isolate a set $\mathbb{P}^*$ of functions defined on $\mathcal{L}$. The infamous *reference class problem* must be tackled at this stage, i.e., one must decide which items of evidence about the generic physical probabilities should be considered when determining single case probabilities $\mathbb{P}^*_{\mathcal{L}}$; this task is unavoidable if single-case probabilities (degrees of belief) are to be calibrated with generic probabilities (propensities or frequencies). As explained in Sect. 7, the theory of *evidential probability* offers one possible protocol for tackling this problem.

$\langle\cdot\rangle$ is the convex hull operator: if two probability functions P and Q are in $\langle\mathbb{P}^*_{\mathcal{L}}\rangle$ then so are any functions on the line segment from P to Q. $\mathbb{S}$ is a set of *structural constraints*—while in many cases evidence constrains an agent's degrees of belief by telling her about physical probabilities, in other cases evidence can constrain degrees of belief in ways not mediated by physical probabilities, and $\mathbb{S}$ is intended to capture those latter constraints. For the purposes of this paper there is no need to discuss $\mathbb{S}$ further, as we will not be considering any structural constraints—see Williamson (2010b, §3.3) for more details.

Betting according to physical probabilities is also justifiable in terms of minimising worst-case expected loss. Suppose evidence says that the physical probability of $\theta \in S\mathcal{L}$ is $x, P^*(\theta) = x$, and that the agent bets according to $P_E(\theta) = q$. Such a bet is interpreted as a payment of qS for a return of S if θ turns out to be true, where stake S is chosen by a stake-maker and can be positive or negative. Expected loss is then $x(q - 1)S + (1 - x)qS = (q - x)S$. If $q > x$ then a stake-maker can (in the worst case, will) choose $S > 0$ to ensure that the expected loss is positive. Similarly if $q < x$ and S is chosen to be negative. Only if $q = x$ is this kind of exploitation of betting commitments not possible. More generally, if E determines some $\mathbb{P}^*_{\mathcal{L}}$ then exploitation is only possible if P_E lies outside the convex hull $\langle\mathbb{P}^*_{\mathcal{L}}\rangle$.

The Calibration norm is a generalisation of what is sometimes called the *Principal Principle*, which says that if one knows that the chance of θ is x, one should set one's degree of belief in θ to be x, as long as there isn't any 'inadmissible' evidence that renders such an assignment inappropriate.[4] The

[4] *Miller's Principle* is a similar principle of direct inference. Lewis (1980) put forward his Principal Principle in order to help elucidate the notion of physical probability for subjectivists, though he advocated an independent 'best-system' analysis of physical probability understood as single-case chance—see Sect. 7.2 on this point.

Principal Principle applies to evidence of individual physical probabilities, while the Calibration norm as formulated above handles evidence of arbitrary constraints on physical probabilities.

Equivocation. P_E should otherwise equivocate sufficiently between the basic possibilities expressible in $\mathcal{L}$.

E1: P_E should be sufficiently close to equivocator function $P_=$ defined by $P_=(\omega) = 1/2^n$ for each state ω of the form $\pm A_1 \wedge \cdots \wedge \pm A_n$.

Distance between probability functions is measured by *Kullback-Leibler divergence* (*KL-divergence*), $d(P, Q) = \sum_\omega P(\omega) \log P(\omega)/Q(\omega)$. (In fact, P is closer to the equivocator than Q in terms of KL-divergence just if P has greater *entropy* than Q, $H(P) > H(Q)$ where $H(P) = -\sum_\omega P(\omega) \log P(\omega)$). As to what will count as *sufficiently* close to the equivocator will depend on pragmatic considerations such as the required numerical accuracy of probabilistic predictions.

The Equivocation norm can also be justified by an appeal to minimising worst-case expected loss (Williamson 2010a). Let $L(\omega, Q)$ be the loss the agent incurs by choosing Q as her belief function if ω turns out to be the true state of the world. The expected loss is $L(P, Q) = \sum_\omega P(\omega)L(\omega, Q)$. It turns out that, under natural conditions, the belief function Q that minimises worst-case expected loss (i.e., that minimises the maximum expected loss when P ranges over some set $\mathbb{E}$ of probability functions compatible with evidence) is the belief function in $\mathbb{E}$ that is closest to the equivocator, where the divergence function in question is defined in terms of the loss function (Grünwald and Dawid 2004). Now, in the absence of any specific information about the loss function, one can argue that L should be taken to be logarithmic loss, $L(\omega, Q) = -\log Q(\omega)$, since logarithmic loss is the only loss function that satisfies a list of natural desiderata that one might posit of a default loss function (Williamson 2010a, pp. 133–134). In which case the corresponding divergence function turns out to be KL-divergence, and the Equivocation norm, as formulated above, should hold.

In sum, each of the above three norms can be motivated by the following sort of argument. Degrees of belief are used to determine action. Prudent action demands taking steps to minimise worst-case expected loss. But minimising worst-case expected loss demands satisfying the norm. Hence degrees of belief should satisfy the norm in question.[5]

The key point to note for the purposes of this paper concerns the Calibration norm. The Calibration norm requires that $P_E \in \mathbb{E} = \langle \mathbb{P}^*_{\mathcal{L}} \rangle \cap \mathbb{S}$, where $\mathbb{P}^*$ is the set of physical probability functions that are compatible with the agent's evidence E. But it is prima facie plausible that frequentist statistics, which seeks to determine how

[5] Here one should not necessarily think of loss in financial terms. One might suspect that there are times at which one doesn't care about being financially prudent. For example, betting in a casino might be considered exciting but not financially prudent. In which case one might wonder whether the norms only hold in those cases in which one wishes to be prudent. But prudence is not to be identified with financial prudence: given that one wants excitement it can be prudent to go to a casino—financial losses are outweighed by a lack of excitement. Arguably it is a matter of fact that an ideal action is a prudent action, in the sense of an action that minimises worst-case expected loss, regardless of whether one cares about financial loss.

evidence narrows down the set of candidate physical probability functions, is the natural method to appeal to in order to determine $\mathbb{P}^*$.

In Sect. 2 we concluded with the thought that perhaps frequentism and Bayesianism are not incompatible after all, because they have different goals: the former concerns itself with physical probability while the latter with epistemological probability, i.e., strength of belief. Now it appears that the connection may be somewhat stronger still. Bayesian epistemology, which seeks to use evidence to determine epistemological probability, may need to appeal to frequentist statistics, which seeks to use evidence to determine physical probability. In the next section we shall endeavour to make this connection more precise.

4 Frequentist Statistics for Calibration

4.1 Recap

In Sect. 2 we encountered the three-step view of statistical inference as (1) determining a set $\mathbb{M}$ of models, (2) gathering evidence E; and (3) isolating a subset $\mathbb{M}_E$ of models that are appropriate given the evidence. The standard view of the way in which Bayesianism instantiates this scheme pitches Bayesianism as being incompatible with frequentism. But Bayesianism can be thought of as primarily an epistemological theory, based on the norms of Probability, Calibration and Equivocation, and under that perspective it appears that frequentist methods are required in order to implement the Calibration norm (Sect. 3). In which case Bayesianism (in its empirically-based or objective variants) is best represented as instantiating the three-stage scheme as follows:

1. Let $\mathbb{M}$ be the set of probability functions defined over sentence of the agent's language $\mathcal{L}$.
2. Gather evidence E and isolate a set $\mathbb{E} = \langle \mathbb{P}^*_{\mathcal{L}} \rangle \cap \mathbb{S}$, where $\mathbb{P}^*$ is the set of candidate physical probability functions, given E. This is the stage at which it is natural to apply frequentist methods.
3. The set of belief functions that should be used as a basis for action is $\mathbb{M}_E \subseteq \mathbb{E}$. Empirically-based Bayesians would say that $\mathbb{M}_E = \mathbb{E}$ while objective Bayesians would select only members of $\mathbb{E}$ that are sufficiently equivocal.

This scheme is clearly not incompatible with frequentist methods.

4.2 Extended example

This brief sketch can be fleshed out by means of a simple example.[6]

[6] Depending one's conception of physical probability, one might hesitate as to whether physical probabilities attach to the macroscopic events of this example. The reader should feel free to reinterpret the terms of this example so as to be comfortable that the relevant physical probabilities are all well defined.

4.2.1 Initial evidence

An agent is sampling 100 vehicles at a road T-junction with a view to predicting whether the 101st vehicle will turn left or right. We shall suppose that $\mathcal{L}$ is a language with a predicate L for *turns left* (*turns right* corresponds to $\neg L$), with 101 constant symbols $v_1, \ldots, v_{101}$ for the vehicles in the order in which they are observed, and with the wherewithal to express claims about physical probability P^*. (To keep the example in line with frequentist methods, we shall suppose that physical probability is generic—i.e., defined over repeatably instantiatable outcomes rather than over single cases—and is defined relative to a set of repeatable conditions or a reference class S, which we will occasionally make explicit by adding a subscript S to the relevant variable.) The agent goes ahead and observes $v_1, \ldots, v_{100}$ and finds that 41 of them turn left. The sample does not indicate any dependence of an outcome on the past sequence of outcomes, and the agent is prepared to grant that the outcomes are independent and identically distributed (iid). Since the agent grants the sample and the iid claim, this constitutes the agent's evidence base E.[7]

4.2.2 Step 1. Determine a threshold of acceptance

First assume that τ_0 is the minimum degree to which the agent would need to believe a statement of the form $P^*(L) \in I$ for her to grant it in her current operating context.[8] τ_0 can be thought of as a threshold of acceptance, and, where utilities are available, they can be used to determine the threshold as follows. Consider a utility table of the form:

	θ	$\neg\theta$
Grant θ	S_1	E_2
Don't grant θ	E_1	S_2

Here S_1 is the utility of granting θ when it is true; E_1 is the utility of a type 1 error, i.e., of not granting θ when θ is true; E_2 is the utility of a type 2 error, i.e., of granting θ when it is false; and S_2 is the utility of not granting θ when θ is false. Arguably, one should grant θ iff the expected utility of granting θ outweighs that of not granting θ, i.e., iff

$$P(\theta)S_1 + (1 - P(\theta))E_2 \geq P(\theta)E_1 + (1 - P(\theta))S_2,$$

i.e., iff

[7] Recall that, in the approach to Bayesian epistemology presented in Sect. 3.1, the agent's evidence base includes everything she takes for granted in her current context of inquiry. This includes standard modelling assumptions such as the iid assumption. Such assumptions are retracted from her evidence base if they are no longer granted—e.g., if they are called into question by subsequent evidence.

[8] This assumption will be qualified somewhat in Sect. 6.

$$P(\theta) \geq \frac{S_2 - E_2}{S_1 + S_2 - E_1 - E_2}.$$

Presumably $S_1 \geq E_1$ and $S_2 \geq E_2$, so this threshold lies within the unit interval. For instance, for the utility matrix

	$P^*(L) \in I$	$P^*(L) \notin I$
Grant $P^*(L) \in I$	1	−5
Don't grant $P^*(L) \in I$	−1	1

the threshold of acceptance τ_0 is 6/8 = .75.

4.2.3 Step 2. Determine a confidence interval

Given τ_0, one can then use frequentist confidence-interval methods as above to determine a confidence interval $I(\bar{X}, \tau_0)$ such that $P^*(P^*(L) \in I(\bar{X}, \tau_0)) \approx \tau_0$, as we shall now explain.

Classical frequentist estimation methods routinely yield assertions of the form $P^*(|\bar{X} - P^*(L)| \leq \delta) \approx \tau$. This says that in the limit, in roughly $100\tau\%$ of samples, the proportion $\bar{X}$ of vehicles turning left in the sample will be within δ of the physical probability of vehicles at the junction turning left (L). (Note that $\bar{X} = \bar{X}_S$ is taken to vary over samples within some reference class S and similarly $L = L_V$ varies over the reference class V of vehicles at the junction in question.) Such an assertion might result from taking L to be binomially distributed and $\bar{X} \sim \mathcal{N}(p, p(1-p)/n)$: the Central Limit Theorem implies that the distribution of $\bar{X}$ is approximately normal with mean p and standard deviation $p(1-p)/n$, where $p = P^*(L)$ and n is the sample size (100 in this case). Thus $P^*(\bar{X} \leq r) \approx \Phi((r - p)/\sqrt{p(1-p)/n})$ where Φ is the standard normal distribution function: in our example, if $p = .5$ then $P^*(\bar{X} \leq .41) \approx \Phi(-.09/.05) = 0.0359$.

Then,

$$P^*(|\bar{X} - p| \leq \delta) \approx \Phi\left(\frac{\delta}{\sqrt{p(1-p)/n}}\right) - \Phi\left(\frac{-\delta}{\sqrt{p(1-p)/n}}\right)$$
$$= 2\Phi\left(\frac{\delta}{\sqrt{p(1-p)/n}}\right) - 1 = \tau$$

say. Thus τ can be construed as a function of δ. On the other hand—and more importantly for our analysis—δ can be construed as a function of τ : given τ one can choose $\delta = \Phi^{-1}(1/2 + \tau/2)\sqrt{p(1-p)/n}$ so that $P^*(|\bar{X} - p| \leq \delta) \approx \tau$. Equivalently, $P^*(p \in [\bar{X} - \delta, \bar{X} + \delta]) \approx \tau$. The interval $[\bar{X} - \delta, \bar{X} + \delta]$ is called a $100\tau\%$ *confidence interval* for p; note that $\bar{X}$ is a variable (the sample frequency varies from sample to sample) while p is a constant. The ultimate aim is to instantiate $\bar{X}$ to its

value $\bar{X}_s$ in a particular sample s in class S, and thereby use the confidence interval to provide practical bounds on the unknown p. As yet this is not possible, because δ depends on p and hence is also unknown. But the following procedure is often used to provide an identifiable confidence interval for p. Let $k \stackrel{\text{df}}{=} \Phi^{-1}(1/2 + \tau/2)$. Now, $|\bar{X} - p| \leq \delta$ if and only if

$$|\bar{X} - p| \leq k\sqrt{\frac{p(1-p)}{n}}.$$

Squaring both sides,

$$\bar{X}^2 - 2\bar{X}p + p^2 \leq \frac{k^2 p}{n} - \frac{k^2 p^2}{n},$$

i.e., as a quadratic in p,

$$\left(1 + \frac{k^2}{n}\right)p^2 - 2\left(\bar{X} + \frac{k^2}{2n}\right)p + \bar{X}^2 \leq 0.$$

This inequality holds when p is between the two zeros of this quadratic, i.e., when p is in the interval

$$\left[\frac{\bar{X} + k^2/2n - k\sqrt{\bar{X}(1-\bar{X})/n + k^2/4n^2}}{1 + k^2/n}, \frac{\bar{X} + k^2/2n + k\sqrt{\bar{X}(1-\bar{X})/n + k^2/4n^2}}{1 + k^2/n}\right],$$

an identifiable confidence interval for p. We shall use $I(\bar{X}, \tau)$ to refer to this interval.

In sum, we can apply frequentist methods at this step to infer that $P^*(p \in I(\bar{X}, \tau_0)) \approx \tau_0$, i.e., $P^*(P^*(L) \in I(\bar{X}, \tau_0)) \approx \tau_0$. Note that all probabilities remain generic at this step, since L and $\bar{X}$ are generic (repeatably instantiatable) variables. As yet, there has been no application to the single-case sample of our example.

4.2.4 Step 3. Calibrate

Now, if all that is known about the specific sample s in question is that it is a sample of type S and that $P^*(P^*(L) \in I(\bar{X}, \tau_0)) \approx \tau_0$ for samples of type S, then the Calibration norm arguably implies that $P_E(P^*(L) \in I(\bar{X}_s, \tau_0)) = \tau_0$, i.e., that the agent should believe to degree τ_0 that the physical probability of turning left lies within the confidence interval induced by this specific sample.[9] For example, if $\tau_0 = .75$ then $P^*(P^*(L) \in I(\bar{X}, \tau_0)) \approx \tau_0$ says that in about 75% of samples $t \in S$,

[9] Note that this inference is only appropriate in cases where $I(\bar{X}_s, \tau_0) \subseteq [0, 1]$. Other cases may warrant higher credence in the claim that $P^*(L) \in I(\bar{X}_s, \tau_0)$; see Seidenfeld (1979, Chapter 2) and Mayo (1981, §2) on this point. Expressed in the framework of Sect. 3.2, if $I(\bar{X}_s, \tau_0)) \not\subseteq [0, 1]$ then the single-case consequences $\mathbb{P}^*_{\mathcal{L}}$ of the physical probability information $\mathbb{P}^*$ do not just depend on the explicit information that $P^*(P^*(L) \in I(\bar{X}, \tau_0)) \approx \tau_0$, but also on the further information that $P^*(P^*(L) \in I(\bar{X}, \tau_0)) \in [0, 1]$. In general, any application of the Calibration norm must respect the single-case consequences of the *total* evidence, not just of the information that $P^*(P^*(L) \in I(\bar{X}, \tau_0)) \approx \tau_0$. To put it another way, the *after-trial* evidence differs from the *pre-trial* evidence, and the fact that $\bar{X}_s = .41$ may not only be pertinent with regard to the construction of the interval $I(.41, \tau_0)$, but also in other regards (Hacking 1965, pp. 95–96).

the interval $I(\bar{X}_t, .75)$ bounds the physical probability of turning left; the agent knows just that $s \in S$; hence she should believe to degree .75 that the interval $I(\bar{X}_s, .75)$ bounds the physical probability of turning left. If 41 cars turn left in sample $s, \bar{X}_s = .41$ and the agent should believe to degree .75 that the interval $I(\bar{X}_s, .75) = [.355, .467]$ bounds the physical probability of turning left.

4.2.5 Step 4. Accept

Since τ_0 is the acceptance threshold for statements of the form $P^*(L) \in I$ and the agent believes to degree τ_0 that $P^*(L) \in I(\bar{X}_s, \tau_0)$, the agent should go on to *grant* that $P^*(L) \in I(\bar{X}_s, \tau_0)$. Let E' be her new evidence base after granting this.

4.2.6 Step 5. Recalibrate

Now, if all that is granted about v_{101} is that it is of type V, i.e., a vehicle at the same junction, and that $P^*(L) \in I(\bar{X}_s, \tau_0)$ for vehicles of type V, then the Calibration norm arguably implies that $P_{E'}(Lv_{101}) \in I(\bar{X}_s, \tau_0)$. Hence the agent should believe that the next vehicle will turn left to some degree within the confidence interval $I(\bar{X}_s, \tau_0)$. In our example, the agent grants that $P^*(L) \in [.355, .467]$, i.e., that the proportion of vehicles at the junction that turn left is in the interval [.355, .467], and knows only that vehicle 101 is a vehicle at the junction, so her degree of belief that vehicle 101 turns left should be within the interval [.355, .467].

While the empirically-based subjective Bayesian would stop here, the objective Bayesian would proceed to the following step.

4.2.7 Step 6. Equivocate

Finally, the Equivocation norm says that the agent should believe that the next vehicle will turn left to some degree within the interval that is sufficiently equivocal: $P_{E'}(Lv_{101})$ should lie within the interval and should be sufficiently close to $P_=(Lv_{101}) = 1/2$, where, as before, $P_=$ signifies the equivocator function. In our example, since $1/2 > .467$, the agent should believe that vehicle 101 turns left to degree .467 or thereabouts.

5 Is This Application of Confidence Intervals Legitimate?

5.1 Confidence intervals as functions

We see then that confidence-interval methods form a core part of Step 2 of this analysis. Note though, that the confidence-interval methods applied at Step 2 are uncontentious, because they are a straightforward consequence of the probability axioms: the Central Limit Theorem is a theorem of the probability calculus, and the assertion that $P^*(P^*(L) \in I(\bar{X}, \tau)) \approx \tau$ simply follows from the resulting normal approximation to the binomial distribution. At Step 2, this assertion remains

generic, applying to samples in general—it has not been specialised to the single-case sample in question. This application comes at Step 3, which is cast in a Bayesian rather than frequentist way. It is thus not until Step 3 that a concrete interval is actually isolated and it is asserted that the agent should be confident that $P^*(L)$ lies within this interval.

Howson and Urbach (1989, pp. 240–241) object to an analogous Bayesian casting of confidence-interval methods. While they object to a different formulation of the Calibration norm—namely the Principal Principle—being used to apportion confidence from a confidence interval, their objection does not in fact hinge on the particular way in which the Calibration norm is formulated. In our framework, their objection proceeds as follows: $P^*(P^*(L) \in I(\bar{X}, \tau)) \approx \tau$ does not license the inference to $P_E(P^*(L) \in I(\bar{X}_s, \tau_0)) = \tau$, because $I(\bar{X}, \tau_0)$ is not an interval of numbers, but rather a function of possible experimental outcomes (a function mapping $\bar{X}$ to an interval of numbers).

But this objection cannot be right: by necessity, *any* application of the Calibration norm in which physical probabilities are construed as generic rather than single-case must draw inferences from a function of possible experimental outcomes. Thus a Calibration norm must move from a statement of the form $P^*(\theta(x)) \in Y$, where $\theta(x)$ is repeatedly instantiatable (a function mapping substitutions of x to propositions), to a statement of the form $P_E(\theta(s)) \in Y$, where $\theta(s)$ is single-case (the result of substituting s for x to yield a proposition). For example, an inference from a physical probability of .7 of surviving 5 years after diagnosis with prostate cancer to a degree of belief of .7 that Bob will survive 5 years after diagnosis with prostate cancer is an inference from the probability of a propositional function (x will survive 5 years after diagnosis with prostate cancer) to the probability of a proposition (Bob will survive 5 years after diagnosis with prostate cancer). So the fact that $I(\bar{X}, \tau)$ is a function cannot be problematic in itself..

5.2 Two analogies

Howson and Urbach draw the following analogy:

> For example, the physical probability of getting a number of heads greater than 5 in 20 throws of a fair coin is 0.86 ... That is, $P^*(K > 5) = 0.86$, where K is the number of heads obtained. According to the Principal Principle, $P[(K > 5)_t \mid P^*(K > 5) = 0.86] = 0.86$, so 0.86 is also the confidence that you should place in any particular trial of 20 throws of a fair coin producing a number of heads greater than 5.
>
> Suppose a trial is made and 2 heads are found in a series of 20 throws with a coin that is known to be fair. To infer that we should now be 86 per cent confident that 2 is greater than 5 would be absurd and a misapplication of the Principal Principle. If one could substitute numbers for K in the Principle, it would be hard to see why the substitution should be restricted to the term's first occurrence. But no such substitution is allowed. For the Principal Principle does not assert a general rule for each number K from 0 to 20; the K-term is not in fact a number, it is a function which takes different values

depending on the outcome of the underlying experiment. (Howson and Urbach 1989, p. 240)

There are two problems with their example. First, it is misleading: the problem is chiefly to do with changing information rather than with illegitimate substitution. Before the trial t takes place it is indeed reasonable to believe that $(K > 5)_t$ to degree 0.86—just as at one time it was reasonable to believe that the number of the planets in our solar system is less than 8—because the number of heads K at trial t is unknown. The problem arises because, after the fact, it is known that the number of heads at trial t is 2. This is clearly information, more pertinent than the previous probabilistic information, that thwarts any inference to the claim that one ought to believe that $(K > 5)_t$ to degree 0.86. On Lewis' formulation of the Principal Principle, the chance is now 0 that $(K > 5)_t$, and so one must apply the Principal Principle to this new chance and believe to degree 0 that $2 > 5$. But there are other formulations of the Principal Principle (see, e.g., Hoefer 2007), and one might instead deem the new information to be inadmissible information which prevents any application of the Principal Principle at all.

The point is that the new information blocks the previous application of the Principal Principle because it provides more pertinent information about the number of heads at trial t, not because of any concerns about whether the K-term is a number. Howson and Urbach are right that the K-term (the number of heads at trial t) is not a number, since it is a definite description rather than a number. But the K-term must *pick out* a number, for otherwise the previous application of the Principal Principle would not be legitimate: it makes no sense to ask whether $K > 5$ at trial t if K is not instantiated as a number at trial t. So Howson and Urbach misdiagnose their own example as being one of substitution failure rather than one of gaining more pertinent knowledge.[10]

The second problem with Howson and Urbach's example is that it is not closely analogous to the confidence interval case, since the constant term, 5, is known from the start. The inference we are interested in is from a statement of the form $P^*(p \in I(\bar{X}, \tau)) \approx \tau$ to a statement of the form $P_E(p \in I(\bar{X}_s, \tau)) = \tau$ where p (which does not vary from trial to trial) is unknown. Here, then, is a closer analogy: the move from the claim that, three times out of four, Paul's height (H_p, which is unknown) is greater than that of a randomly selected male of the same species, $P^*(H_p > H) = .75$, to the claim that one ought to believe to degree .75 that Paul's height is greater than that of the next sampled male, who was Steve, and who turned out to be 20 cm high, $P_E(H_p > 20) = .75$ where $H_s = 20$, in the absence of any other pertinent evidence about male heights of that species (the species is not revealed, say). But this is a harmless application of a calibration principle such as the Principal Principle or the Calibration norm C1 of Sect. 3. Suppose that it is known that, three times out of four, sampled males of the species are shorter than

[10] Howson and Urbach are quite right, however, to emphasise that one must guard against substitution failure, as their rebuttal of Miller's paradox does hinge on substitution failure (Howson and Urbach 1989, §15.e).

Paul, $P^*(H_p > H) = .75$, and that it is known that Steve has been randomly sampled, but Steve's height has not yet been obtained. Then surely it is reasonable to believe that Steve is shorter than Paul, $H_p > H_s$, to degree .75. This is a routine application of the Principal Principle. Note that neither H_s nor H_p are known at this stage. Then Steve's height is measured and it is learnt that $H_s = 20$. In the absence of any general knowledge of heights of males of this species, this new knowledge hardly provides any grounds for moving away from degree of belief .75 that $H_p > H_s$. But if one knows that $H_s = 20$ and one believes to degree .75 that $H_p > H_s$ then one ought to believe to degree .75 that $H_p > 20$, i.e., that Paul's height is greater than 20 cm. Analogously, in our example the agent ought to believe to degree .75 that $p \in [.355, .467]$.

In sum, any application of a Calibration norm that appeals to generic probability must draw inferences from propositional functions to propositions. The inference of Step 3 is of just this form and is neither fallacious nor analogous to an inference to the claim that $2 > 5$. In fact it is closely analogous to uncontroversially benign applications of the Calibration norm.

6 Is the Acceptance Assumption Legitimate?

6.1 Narrowest intervals

While the procedure spelt out in Sect. 4 survives Howson and Urbach's objection, it does need to be qualified in order to avoid a more telling objection. As it stands, there is a certain arbitrariness to the above procedure. There are other intervals $I'(\bar{X}, \tau)$ such that $P^*(P^*(L) \in I'(\bar{X}, \tau)) \approx \tau$, and the results of the procedure will depend on the chosen interval.

Consider for example the analysis of Step 2, but reapplied to a one-sided confidence interval. Now,

$$P^*(\bar{X} \geq p - \delta) \approx 1 - \Phi\left(\frac{-\delta}{\sqrt{p(1-p)/n}}\right) = \Phi\left(\frac{\delta}{\sqrt{p(1-p)/n}}\right) = \tau$$

say. Conversely, given τ one can choose $\delta = \Phi^{-1}(\tau)\sqrt{p(1-p)/n}$ so that $P^*(\bar{X} \geq p - \delta) \approx \tau$. Equivalently, $P^*(p \in [0, \bar{X} + \delta]) \approx \tau$. The same procedure as before can be used to yield an identifiable confidence interval: letting $k \stackrel{\text{df}}{=} \Phi^{-1}(\tau), P^*(\bar{X} \geq p - \delta)$ if and only if p is in the interval

$$\left[0, \frac{\bar{X} + k^2/2n + k\sqrt{\bar{X}(1-\bar{X})/n + k^2/4n^2}}{1 + k^2/n}\right],$$

which we shall call $I'(\bar{X}, \tau)$. If $\tau_0 = .75$ then $I'(\bar{X}_s, \tau_0) = [0, .444]$. If this interval had been chosen instead of $I(\bar{X}_s, \tau_0)$ then the empirically-based subjective Bayesian would have required that $P_{E'}(Lv_{101}) \in [0, .444]$ instead of [.353, .467], while the

objective Bayesian would have required that $P_{E'}(Lv_{101}) = .444$ instead of .467. Clearly something is wrong if the same procedure yields inconsistent results.

But what went wrong? One must point the finger at the starting-point of the analysis—the assumption that there is a threshold degree of belief τ_0 above which the agent should grant *any* statement of the form $P^*(L) \in I$. Given τ_0 there may be various sets I for which the agent believes that $P^*(L) \in I$ to this threshold degree of belief. Indeed this family of sets will normally have empty intersection, and so—if the assumption were true—the agent would be forced to believe that $P^*(L)$ is no number at all. Clearly, then, the assumption must be rejected.

Can the procedure be fixed? A standard way round this sort of problem is to restrict the assumption by supposing instead that there is a threshold degree of belief τ_0 above which the agent should grant $P^*(L) \in I$ where I is the *narrowest* interval which meets this threshold (see, e.g., Kyburg Jr and Teng 2001, §11.5). Note that we have already been employing this principle to some extent, since Step 4 accepts $P^*(L) \in I(\bar{X}_s, \tau_0)$ for interval $I(\bar{X}_s, \tau_0)$ *at* the threshold, but ignores all those intervals $I(\bar{X}_s, \tau)$ for $\tau > \tau_0$; these latter intervals are all wider than the former.

6.2 Why the narrowest interval?

Why would the agent be better off with narrowest interval? Simply because the intervals in question are being used for estimation here, and the narrower the interval, the more informative it is about the physical probability being estimated. Recall that the problem the Bayesian faces is that of Calibrating prior probabilities to evidence of physical probabilities. These physical probabilities must be estimated somehow. To the extent that the agent's evidence determines several confidence intervals for some physical probability, some wider than others, she should focus on the narrowest such interval because that interval will convey the most information about the physical probability in question.

Of course, the new assumption will clearly not be appropriate where there is no narrowest interval: suppose a coin is known to be biased but the direction of the bias is not known, and $\tau_0 = 0.5$; then $P^*(H) \in (.5, 1]$ and $P^*(H) \in [0, .5)$ should both be believed to the threshold degree of belief, but granting both would force one to hold that $P^*(H) \in \emptyset$.

Moreover, the narrowest confidence interval will typically be the interval $[\bar{X} - \delta, \bar{X} + \delta]$ that is symmetric about the sample proportion $\bar{X}$; as discussed in Sect. 4, this interval is not identifiable because δ is defined in terms of p, the unknown quantity that is being estimated. The agent can hardly grant the narrowest interval estimate if the narrowest interval is unknowable. The best the agent can do is grant the narrowest interval estimate *from all those interval estimates that she can formulate*.

In sum, we can modify the acceptance assumption by supposing that there is a threshold degree of belief τ_0 above which the agent should grant $P^*(L) \in I$, in those cases in which the agent knows of no other interval I' that is at least as narrow as I and for which her degree of belief in $P^*(L) \in I'$ also meets the threshold τ_0.

7 Discussion

7.1 Summary

This paper has argued that one should not view frequentism and Bayesianism as chalk and cheese; rather, their relationship should be one of harmonious symbiosis. One particular way in which the two positions can be integrated proceeds along the following lines:

Step 1. Let τ_0 be the minimum degree to which the agent would need to believe a statement of the form $P^*(L) \in I$ for her to grant it. Here I is an interval, and it is understood that, if more than one such statement reaches the threshold degree of belief, the agent will only grant the statement involving the narrowest interval (if there is precisely one narrowest interval).

Step 2. Given τ_0, find a confidence interval $I(\bar{X}, \tau_0)$ such that $P^*(P^*(L) \in I(\bar{X}, \tau_0)) \approx \tau_0$.

Step 3. The Calibration norm implies that $P_E(P^*(L) \in I(\bar{X}_s, \tau_0)) = \tau_0$.

Step 4. The agent should *grant* that $P^*(L) \in I(\bar{X}_s, \tau_0)$.

Step 5. The Calibration norm implies that $P_{E'}(Lv_{101}) \in I(\bar{X}_s, \tau_0)$.

Step 6. The Equivocation norm implies that $P_{E'}(Lv_{101})$ should lie within the interval and should be sufficiently close to $P_=(L\ v_{101}) = 1/2$.

This last step applies to an objective Bayesian account but not to an empirically-based subjective Bayesian account.

Regardless of whether the details of this specific integration are accepted, the following two general points can be made.

7.2 The Bayesian needs the Frequentist

Standard forms of Bayesianism invoke a calibration principle along the lines of the Principal Principle or the Calibration norm C1 of Sect. 3. The idea here is that, in cases where the agent has evidence of physical probabilities, she would be irrational if she did not bet according to those physical probabilities (at least in situations where the reference class problem is resolvable). But such a principle presupposes that (1) physical probabilities can be known, at least approximately, and (2) that they are obtained before the prior belief function is set, i.e., they can be estimated independently of the agent's prior belief function. This second presupposition rules out subjectivist Bayesian methods for estimating physical probabilities by updating the prior belief function in the light of individual items of sample data, since the prior has not yet been determined.[11] The only plausible remaining estimation methods are frequentist estimation methods. Hence the Bayesian needs to employ frequentist estimation methods in order to calibrate with physical probabilities.

[11] Jaynes (1976, §IIIa) maintains that Bayesian interval estimates with respect to a uniform prior are close to, but slightly narrower than, frequentist confidence intervals.

There are various possible ways in which one might try to undermine this argument. First, one might be sceptical of the existence of physical probabilities. De Finetti himself famously claimed that 'probability does not exist' (de Finetti 1970, p. x). This is not the place for a general defence of physical probabilities—suffice it to say that even the sceptic can accept that one ought to calibrate one's degrees of belief with sample frequencies, especially where the sample contains plentiful, good quality data.[12]

Second, one might be sceptical as to whether one really should calibrate degrees of belief with physical probabilities. Again, a sustained defence of the Calibration norm is beyond the scope of this paper—see Williamson (2010b, §3.3) for a fuller discussion. But, as pointed out in Sect. 3, a defence can be given in terms of minimising worst-case expected loss. If one were to reject this defence, one would also have to rescind the standard Dutch book argument for the Probability norm, which is itself essentially a justification in terms of minimising worst-case expected loss. (And it is argued in Williamson 2010b, §§3.1–3.2 that other ways of justifying the Probability norm are far less compelling.)

Third, one might be sceptical of whether frequentist estimation methods are appropriate methods for estimating physical probabilities. But, as emphasised above, the frequentist confidence-interval methods of Step 2 are totally uncontentious. The contentious step is Step 3, which turns out to be a routine application of the Principal Principle or the Calibration norm. So the Bayesian seeking to calibrate degrees of belief with physical probabilities cannot deny that confidence-interval methods are an appropriate way of estimating physical probabilities.

Fourth, one might have a particular view of physical probabilities which fits more naturally with Bayesian methods. For example, under the Ramsey-Lewis *best-system* view (Ramsey 1928; Lewis 1994), facts about physical probabilities are determined by the best systematisation of reality—the deductive system, from all those that yield true conclusions about fundamental matters of fact, that offers the best balance between simplicity, strength and fit. Now a calibration principle such as the Principal Principle would require roughly that, for all possible systemisations S, one ought to set one's degrees of belief, conditional S being the best systematisation, to what S says the physical probabilities are. In order to determine a prior function over the agent's language $\mathcal{L}$, one then needs, for each S, to award some prior probability to the claim that S is the best systematisation. This is clearly a Bayesian resolution to the task at hand: there is no use of frequentist methods here.

However, while this kind or move might be natural for someone who is interested in the Principal Principle insofar as it elucidates the metaphysics of probability, it faces difficulties in the present, epistemological context. This move translates the problem of setting a prior over $\mathcal{L}$ into the problem of setting a prior over claims of the form 'S offers the best balance between simplicity, strength and fit', for all possible systematisations S of all fundamental matters of fact. From the point of view of Bayesian epistemology this is unsatisfactory because it replaces an intuitive

[12] In this paper, the term 'physical probability', rather than the more common term 'objective probability', is used to refer to non-epistemic probability, in order to avoid confusion in the case of objective Bayesianism, which is objective in the sense that it admits little room for subjective choice, but which cannot be classified as objective in the non-epistemic sense.

problem, about which there is substantial agreement even if some aspects are contentious, with a much larger and less tractable problem, about which there are few firm intuitions. One might suggest that one ought to adopt a uniform prior over the systematisation partition, i.e., over the partition of claims of the form 'S offers the best balance between simplicity, strength and fit'. But such a suggestion is not applicable to the case at hand: while it might be appropriate in the total absence of evidence, here we concerned with the question of how one should best calibrate *given evidence E which offers substantial information about physical probabilities*. While there might be some dispute as to whether frequentist interval estimates should explicate this substantial information, it is much more straightforward to determine the consequences of this evidence over $\mathcal{L}$ than over the systematisation partition.

In response, the proponent of this sort of move might say something like this: a prior in the presence of evidence E should match the posterior that would be formed by updating the 'pretend' prior that one would have adopted in the total absence of evidence, by Bayesian conditionalisation on E. Thus one might update a uniform prior ever the systematisation partition. Some general problems with this pretend-prior strategy were discussed in Sect. 3.1, one of which is particularly pertinent here. This strategy can fail from a practical point of view, since E need not be expressible in the agent's language $\mathcal{L}$, in which case the relevant conditional probabilities are undefined and conditionalisation is simply not possible. One might suggest then, that one should expand the domain of the belief function to include not only $\mathcal{L}$ and the systematisation partition but also all possible sets of evidence E, so that the relevant conditional probabilities can be defined. But the prior over this new domain should not be uniform if some sets of evidence are to favour some systematisations more than others, since the uniform equivocator function $P_=$ renders logically independent propositions probabilistically independent (Williamson 2011a). Since all possible evidence sets are under consideration, the problem of determining an appropriate prior over the new domain in the total absence of evidence is harder if anything than the problem of determining a prior over the systematisation partition.[13] Hence the epistemological difficulties remain for the advocate of the best-system view of physical probabilities.

7.3 The frequentist needs the Bayesian

The second general point that can be made is this. Frequentist confidence-interval estimation methods are used to estimate unknown constants—e.g., the true value of a parameter of a parameterised set of probability functions. Thus the standard line in statistics text books is this:

> With the realisation that a particular value of an estimator T, called a point estimate, is almost surely wrong, it is natural to want to indicate the degree of

[13] Efforts have been directed at resolving this sort of problem in the area of machine learning—e.g., stemming from the ideas of Solomonoff (1964). However, these efforts have been primarily directed at the more restricted problem of balancing simplicity and fit, and even there, nothing approaching consensus has been reached.

> fuzziness or anticipated error by giving an *interval* of parameter values that, with some assurance or confidence, will contain the true value of the parameter. (Lindgren et al., 1957, p. 160)

The point is that one wants to say that, with some appropriate fixed degree of confidence, a particular interval of numbers (which has been identified) contains a constant (which is unknown). There are two things to note about such a conclusion: (i) a claim is made about a single case (that the value of a constant is within a certain interval), and (ii) it apportions a fixed confidence level to this claim. Now the frequentist needs to justify this conclusion, for otherwise confidence-interval methods will not be compelling. But the frequentist cannot even formulate the conclusion in the frequentist framework. As is well known, while physical probability statements are important steps on the way to drawing such an inference, the conclusion itself cannot be formulated as a physical probability statement: physical probabilities are not normally single-case and they are not degrees of confidence. It is Bayesian probabilities that are single-case and are degrees of confidence. Indeed, this conclusion can be articulated using Bayesian probability as a statement of the form $P_E(p \in I) = \tau$, say. Moreover, as we have seen, the Bayesian can justify this conclusion, by appealing to the frequentist methods and then calibrating degrees of belief to the resulting physical probabilities. So the frequentist needs the Bayesian to justify the application of confidence-interval estimation methods.

Now the main way in which one might try to avoid the force of this argument is to deny that the confidence-interval conclusion in question is a statement about confidence. The conclusion might, for example, be formulated as a statement of reliability (see, e.g., Mayo 1996, p. 272): $100\tau\%$ of the time, interval estimates formulated in this way will be successful. But such a statement of reliability can at best be used to justify using confidence-interval estimates in the long term—it cannot, as it stands, be used to justify the use of the procedure in a particular case. (One might normally use such methods, but, in one particular case, wonder whether to or not; since a single case will make negligible difference to asymptotic reliability, there is little that a reliability claim can do to motivate the use of such methods in that case.) To move from the general to the single case, one needs a further claim that one should conform in each single case in accordance with the general rule. But this further claim is precisely what the frequentist lacks and what the Bayesian can provide in virtue of having a Calibration norm.

Another way in which one might attempt to deny that the confidence-interval conclusion in question is a statement about confidence is to treat it behaviourally: one should *act* as if the interval estimate is correct, but one need not be confident that it is (see, e.g., Neyman 1955, §13). But this response is not really at odds with the Bayesian position. Bayesian epistemology is itself a behavioural theory according to which the claim that one should behave as if the interval estimate is correct is equivalent to the claim that one should bet that it is correct, which is in turn equivalent to the claim that one should be confident that is correct (Ramsey 1926). So the claim that one should merely act as if the interval estimate is correct

implies that one should be confident—in the Bayesian sense—that it is correct, and the question of a Bayesian justification of confidence intervals remains pertinent.

If one accepts that confidence-interval methods concern confidence, one might try to avoid the force of the Bayesian justification by claiming that Bayesian epistemology fails as a theory of confidence. Again—as Fermat might say—the margins of this paper are too small for a general defence of Bayesian epistemology. We may simply note that this tactic will not help the frequentist, who lacks a viable alternative explanation as to why confidence should be apportioned using confidence intervals in the single case.[14]

The general conclusion then is that—like it or not—frequentists and Bayesians need each other.

7.4 Questions for further research

Of course there remain questions for further research, three of which are particularly pressing.

First, while this paper argues that confidence interval estimation is one locus for an integration of frequentism and Bayesianism, it leaves open the question of whether there are other loci for integration. There is a well-known duality between confidence interval estimation and hypothesis testing, for example, and the question arises as to whether there is also a need for frequentists and Bayesians to join forces to test hypotheses.

Two other questions for further research involve the reference-class problem and the acceptance phase of the integration of Sect. 4. We shall now present these questions in a little more detail.

7.4.1 The reference class problem

The reference class problem has been set aside in this paper. This is appropriate as our running example was simple enough, with a single physical probability statement $P^*(P^*(L) \in I(\bar{X}, \tau_0)) \approx \tau_0$, for clashes of reference class not to arise. But the problem still has to be tackled in general, and it is an open question as to how it might best be tackled.

One approach to tackling the problem is to apply the theory of evidential probability as follows. Suppose the agent's evidence yields two statements at the end of Step 2 of Sect. 4, $P^*(P^*(L) \in I(\bar{X}_S, \tau_0)) \approx \tau_0$ and $P^*(P^*(L) \in I(\bar{X}_T, \tau_0)) \approx \tau_0$, where S and T are different reference classes—trials of type S and T select individuals from different classes. If the agent performs sample s of reference class S *and* sample t of class T, then at the end of Step 4 we have that the agent grants

[14] The only other justification of single-case applications of confidence-interval methods seems to be Fisher's *fiducial argument*; however, this seems to require a calibration principle (Hacking 1965, p. 137), so it is apparently a Bayesian justification. Since the fiducial argument is highly controversial, only applicable in specific situations and hard to apply even there, the more straightforward justification of Sect. 4 is preferred here; the exact relationship between the two justifications remains a question for further research. See Seidenfeld (1979, Chapters 4 and 5) and Haenni et al. (2011, Chapter 5) for further discussion of the fiducial argument.

both $P^*(L) \in I(\bar{X}_s, \tau_0)$ and $P^*(L) \in I(\bar{X}_t, \tau_0)$. Suppose $\bar{X}_s \neq \bar{X}_t$ so that these two claims would yield conflicting conclusions if the agent calibrated with respect to one or the other, i.e., $P_E(Lv_{101}) \in I(\bar{X}_s, \tau_0)$ and $P_E(Lv_{101}) \in I(\bar{X}_t, \tau_0)$ where $I(\bar{X}_s, \tau_0) \neq I(\bar{X}_t, \tau_0)$. According to the precepts of evidential probability (Kyburg Jr and Teng 2001), the former statement trumps the latter if:

Richness. $P^*(P^*(L) \in I(\bar{X}_S, \tau_0)) \approx \tau_0$ was obtained from a richer domain (i.e., attributes $\{L, A_1, \ldots, A_n\}$ are measured in samples of class S, while attributes $\{L, A_1, \ldots, A_m\}$ are measured in samples of class T, and $m < n$).
Specificity. Trials of class S are trials of class T but not vice versa.
Precision. $I(\bar{X}_s, \tau_0) \subset I(\bar{X}_t, \tau_0)$.

In general, these principles are applied in the above order to yield a smaller set of untrumped relevant statistical statements. Then the convex hull is taken of the remaining intervals (this is an application of the *Principle of Strength*). So if neither $P^*(L) \in I(\bar{X}_S, \tau_0)$ nor $P^*(L) \in I(\bar{X}_T, \tau_0)$ trumps the other, one sets $P_E(Lv_{101}) \in \langle I(\bar{X}_s, \tau_0), I(\bar{X}_t, \tau_0)\rangle$, the narrowest interval containing the two intervals.

The principle of Richness is intended to explicate the intuition that joint distributions are more informative than marginal distributions, and hence more pertinent. Specificity is the principle of the narrowest reference class: if trials of type S involve sampling vehicles at the junction while trials of type T involve sampling any moving object at the junction—including vehicles, pedestrians, birds etc.—and if v_{101} is a vehicle, then the data from an S-trial is taken to be more pertinent than that from a T-trial. The principle of Precision embodies the agent's need for more precise estimates of physical probabilities, and is redundant given our assumption that the agent will only grant $P^*(L) \in I(\bar{X}_S, \tau_0)$ for the narrowest such interval that reaches the threshold (Sect. 6). Finally, the application of the principle of Strength is also redundant in our framework, in which convex hulls are built into the formulation of the Calibration norm.

The theory of evidential probability, then, can be viewed as a precise theory of how to determine the convex hull of the single-case consequences of potentially conflicting statistical statements—i.e., it can be used to determine the set $\langle \mathbb{P}^*_{\mathcal{L}} \rangle$ that is required by the Calibration norm C1 of Sect. 3. The question of how well it succeeds in doing this would be an interesting question for further research.

7.4.2 Postponing acceptance

Step 4 of the procedure outlined in Sect. 4 involves granting a belief that reaches a threshold, with a view to drawing further conclusions which would not be drawn were the threshold belief not granted. Now any inference must be an inference from what is previously granted, so the act of granting is not a problem in itself. But our success as reasoners depends on our taking the right things for granted. Any act of granting involves a loss of information—it involves neglecting the possibility that what is granted may in fact be false. In our case this possibility has degree of belief $1 - \tau_0$, and this is deemed sufficiently small as to be negligible. But the question remains as to whether more accurate and hence more successful inferences could be

drawn if this possibility weren't neglected, i.e., if the act of granting or acceptance were postponed to a later stage in the reasoning process.

While this is a question for future research, two possible approaches stand out.

One approach is to factor in the neglected uncertainty of the above procedure. Let us focus on Steps 3–5:

Step 3. The Calibration norm implies that $P_E(P^*(L) \in I(\bar{X}_s, \tau_0)) = \tau_0$.
Step 4. The agent should *grant* that $P^*(L) \in I(\bar{X}_s, \tau_0)$.
Step 5. The Calibration norm implies that $P_{E'}(Lv_{101}) \in I(\bar{X}_s, \tau_0)$.

Here the inference $P_{E'}(Lv_{101}) \in I(\bar{X}_s, \tau_0)$ can be viewed as correct if the agent is right to grant that $P^*(L) \in I(\bar{X}_s, \tau_0)$. Thus it is correct with (Bayesian) probability τ_0. On the other hand, it is incorrect with probability $1 - \tau_0$, in which case another inference is appropriate: if $P^*(L) \notin I(\bar{X}_s, \tau_0)$ then $P^*(L) \in [0, 1] \backslash I(\bar{X}_s, \tau_0)$; taking the convex hull, as long at neither endpoint of $I(\bar{X}_s, \tau_0)$ is 0 or 1, we have that $\langle [0, 1] \backslash I(\bar{X}_s, \tau_0) \rangle = [0, 1]$ and the trivial inference $P_{E'}(Lv_{101}) \in [0, 1]$ becomes the appropriate inference with probability $1 - \tau_0$.[15] If $P_{E'}(Lv_{101}) \in [0, 1]$ then one might assume that the probability that $P_{E'}(Lv_{101}) \in I(\bar{X}_s, \tau_0)$ is just the width of that interval, $|I(\bar{X}_s, \tau_0)|/|[0, 1]| = |I(\bar{X}_s, \tau_0)|$. Hence we have that $P_{E'}(Lv_{101}) \in I(\bar{X}_s, \tau_0)$ with probability $\tau_0 + (1 - \tau_0)|I(\bar{X}_s, \tau_0)|$. In our example this probability is $.75 + (0.25 \times (.467 - .355)) = .778$. Hence we can conclude that the agent ought to believe to degree .778 that she ought to believe Lv_{101} to some degree in the interval [.355, .467]. This may or may not be sufficiently high for the agent to grant that she ought to believe Lv_{101} to some degree in the interval [.355, .467]. The point is that this procedure allows one to postpone the acceptance phase until the end of the chain of reasoning, and thereby to take extra uncertainties into account. More generally, one can extend the theory of evidential probability to take account of the uncertainties attaching to its inferences: this yields the theory of *second-order evidential probability*, which is described in Wheeler and Williamson (2011) and Haenni et al. (2011).

Note that this procedure is itself based on a certain assumption: we granted that the probability that $P_{E'}(Lv_{101}) \in I(\bar{X}_s, \tau_0)$ is just the width of that interval. The question arises as to whether there is any approach which avoids this sort of assumption, which, being based on a principle of indifference, may be more palatable to the objective Bayesian than to the empirically-based subjective Bayesian. A second possible approach which does avoid this assumption is based on de Finetti's representation theorem and can be sketched as follows. First, use one-sided confidence intervals as in Sect. 6 to yield statements of the form $P^*(P^*(L) \leq \bar{X} + x) \approx \tau_x$ for each $x \in (0, 1]$. Then calibrate to yield $P_E(P^*(L) \leq \bar{X}_s + x) = \tau_x$ for each $x \in (0, 1 - \bar{X}_s)$. Similarly $P^*(P^*(L) \geq \bar{X} - x) \approx \tau_x$ for each $x \in (0, 1]$ and $P_E(P^*(L) \leq \bar{X}_s - x) = 1 - \tau_x$ for each $x \in (0, 1 - \bar{X}_s)$. These two kinds of claim fully specify $P_E(P^*(L) \leq y)$ for $y = \bar{X} \pm x \in [0, 1]$—i.e., the Bayesian probability distribution of the physical probability distribution is fully determined. Now

[15] Recall that it is assumed that $I(\bar{X}_s, \tau_0) \subseteq [0, 1]$ in order for the original inference to be legitimate. See footnote 9.

de Finetti's representation theorem says that $Lv_{101}, \ldots, Lv_{100+n}$ are exchangeable with respect to P_E if and only if there is a distribution function F such that for all n,

$$P_E(\pm Lv_{101} \wedge \cdots \wedge \pm Lv_{100+n}) = \int_0^1 y^{r_n}(1-y)^{n-r_n} dF(y),$$

where r_n is the number of positive instances in $\pm Lv_{101} \wedge \cdots \wedge \pm Lv_{100+n}$ (de Finetti, 1937). It turns out that F is the distribution function of the limiting relative frequency of L : $F(y) = P_E(\bar{X}_\infty \leq y)$ where $\bar{X}_\infty$ is the limiting proportion of L as the number of vehicles tends to infinity. Since exchangeability is rather natural when the agent already grants that the relevant variables are iid with respect to physical probability (Gillies 2000, pp. 77–83), and since physical probabilities almost always coincide with limiting relative frequencies, the representation theorem can be interpreted as implying that the agent's degrees of belief can be thought of as formed by adopting a Bayesian belief distribution over physical probabilities. Hence one can take $F(y) = P_E(P^*(L) \leq y)$, which as we saw above, is determined from the first hundred sampled vehicles, and, in particular, $P_E(Lv_{101}) = \int_0^1 y dF(y)$.

This sketch glides over a lot of details—for example, confidence interval estimation gets rather subtle when $P^*(L)$ is close to 0 or 1 (Brown et al. 2001), and further argument is needed before one can be convinced that $F(y)$ will turn out to be a continuous distribution function—so this second approach to postponing acceptance must be regarded as much more speculative than the first. But it would be interesting to see whether this general idea can be fleshed out and whether its steps can be adequately justified.[16]

Acknowledgments I am very grateful to the British Academy for supporting this research and to David Corfield, Jan-Willem Romeijn, Jan Sprenger and Gregory Wheeler for helpful comments.

References

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science, 16*(2), 101–117.

de Finetti, B. (1937). Foresight. its logical laws, its subjective sources. In H. E. Kyburg & H. E. Smokler (Eds.). *Studies in subjective probability* (pp. 53–118). Huntington, New York: Robert E. Krieger Publishing Company. Second (1980) edition.

de Finetti, B. (1970). *Theory of probability*. New York: Wiley.

Gillies, D. (2000). *Philosophical theories of probability*. London and New York: Routledge.

Grünwald, P., & Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics, 32*(4), 1367–1433.

Hacking, I. (1965). *Logic of statistical inference*. Cambridge: Cambridge University Press.

Haenni, R., Romeijn, J.-W., Wheeler, G., & Williamson, J. (2011). *Probabilistic logics and probabilistic networks. Synthese library*. New York: Springer.

Hoefer, C. (2007). The third way on objective probability: a sceptic's guide to objective chance. *Mind, 116*, 549–696.

[16] In statistics, research into *predictive probability matching priors* is also beginning to show interesting connections between Bayesian priors and frequentist confidence intervals in the non-parametric setting (Sweeting 2008).

Howson, C. (2001). The logic of Bayesian probability. In D. Corfield, & J. Williamson (Eds.). *Foundations of Bayesianism* (pp. 137–159). Dordrecht: Kluwer.

Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach*. Chicago, IL: Open Court. Second (1993) edition.

Jaynes, E. T. (1976). Confidence intervals vs Bayesian intervals. In W. L. Harper & C. A. Hooker (Eds.). *Foundations of probability theory, statistical inference, and statistical theories of science* (Vol. 2, pp. 175–257). Dordrecht: D. Reidel.

Kolmogorov, A. N. (1933). *The foundations of the theory of probability*. New York: Chelsea Publishing Company (1950).

Kyburg, H. E. Jr., & Teng, C. M. (2001). *Uncertain inference*. Cambridge: Cambridge University Press.

Lewis, D. K. (1980). A subjectivist's guide to objective chance. In *Philosophical papers* (Vol. 2, pp. 83–132). Oxford: Oxford University Press (1986).

Lewis, D. K. (1994). Humean supervenience debugged. *Mind, 412*, 471–490.

Lindgren, B. W., McElrath, G. W., & Berry, D. A. (1957). *Introduction to probability and statistics*. New York: Macmillan. 1978 Edition.

Mayo, D. G. (1981). In defense of the Neyman-Pearson theory of confidence intervals. *Philosophy of Science, 48*, 269–280.

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

Miller, D. (1966). A paradox of information. *British Journal for the Philosophy of Science, 17*, 59–61.

Neyman, J. (1955). The problem of inductive inference. *Communications on Pure and Applied Mathematics, 8*, 13–46.

Ramsey, F. P. (1926). Truth and probability. In H. E. Kyburg & H. E. Smokler (Eds.). *Studies in subjective probability* (pp. 23–52). Huntington, New York: Robert E. Krieger Publishing Company. Second (1980) edition.

Ramsey, F. P. (1928). Reasonable degree of belief. In D. H. Mellor (Ed.). *Philosophical papers*. Cambridge: Cambridge University Press. 1990 Edition.

Seidenfeld, T. (1979). *Philosophical problems of statistical inference: learning from R. A. Fisher*. Dordrecht: Reidel.

Solomonoff, R. (1964). A formal theory of inductive inference. *Information and Control, 7*(1, 2), 1–22, 224–254.

Sweeting, T. J. (2008). On predictive probability matching priors. In B. Clarke & S. Ghosal (Eds.). *Pushing the limits of contemporary statistics: Contributions in honor of Jayanta K. Ghosh* (pp. 46–59). Beachwood: Institute of Mathematical Statistics.

von Mises, R. (1928). *Probability, statistics and truth*. London: Allen and Unwin. Second (1957) edition.

Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. London: Chapman and Hall.

Wheeler, G., & Williamson, J. (2011). Evidential probability and objective Bayesian epistemology. In P. S. Bandyopadhyay & M. Forster (Eds.). *Philosophy of statistics, handbook of the philosophy of science* (pp 307–331). Amsterdam: Elsevier.

Williamson, J. (2010a). Bruno de Finetti: Philosophical lectures on probability. *Philosophia Mathematica, 18*(1):130–135.

Williamson, J. (2010b). *In defence of objective Bayesianism*. Oxford: Oxford University Press.

Williamson, J. (2011a). An objective Bayesian account of confirmation. In D. Dieks, W. J. Gonzalez, S. Hartmann, T. Uebel, & M. Weber (Eds.). *Explanation, prediction, and confirmation. New trends and old ones reconsidered* (pp 53–81). Dordrecht: Springer.

Williamson, J. (2011b). Objective Bayesianism, Bayesian conditionalisation and voluntarism. *Synthese, 178*, 67–85.