Reliable credence and the foundations of statistics

Jesse Clifton
Department of Statistics, North Carolina State University

Abstract

If the goal of statistical analysis is to form justified credences based on data, then an account of the foundations of statistics should explain what makes credences justified. I present a new account called statistical reliabilism (SR), on which credences resulting from a statistical analysis are justified (relative to alternatives) when they are in a sense closest, on average, to the corresponding objective probabilities. This places (SR) in the same vein as recent work on the reliabilist justification of credences generally [Dunn, 2015, Tang, 2016, Pettigrew, 2018], but it has the advantage of being action-guiding in that knowledge of objective probabilities is not required to identify the best-justified available credences. The price is that justification is relativized to a specific class of candidate objective probabilities, and to a particular choice of reliability measure. On the other hand, I show that (SR) has welcome implications for frequentist-Bayesian reconciliation, including a clarification of the use of priors; complementarity between probabilist and fallibilist [Gelman and Shalizi, 2013, Mayo, 2018] approaches towards statistical foundations; and the justification of credences outside of formal statistical settings. Regarding the latter, I demonstrate how the insights of statistics may be used to amend other reliabilist accounts so as to render them action-guiding. I close by discussing new possible research directions for epistemologists and statisticians (and other applied users of probability) raised by the (SR) framework.

1 Introduction

Despite a century of debate over the foundations of statistics, only the various Bayesian approaches to statistical inference have made sustained contact with epistemology — in particular, with the program of Bayesian epistemology. Philosopher of science Mayo [2018] has offered a frequentist account of statistical testing in the fallibilist tradition (she calls it "severe testing") which accords with much of statistical practice, but being an account of hypothesis testing in particular it falls short of a full-blooded statistical philosophy (as statistics is concerned with more than hypothesis testing). Aside from Mayo's work there is little in the way of philosophical foundations for frequentist statistical practice.

This situation is troubling. Applied statisticians have a broadly frequentist outlook on the justification of statistical inferences (though they may help themselves to Bayesian computational machinery, as do Efron [2005] and Gelman and Shalizi [2013], for instance). What's more, the Bayesianism of classical Bayesian epistemology — which says that degrees of belief need only follow the probability calculus, be updated by Bayes rule, and (for so-called objective Bayesians) respect additional side constraints such as using "uninformative priors" [Yang and Berger, 1996] — faces serious limitations as applied to statistical practice. Among these limitations is the fact that, due to the computational cost of Bayesian methods and the difficulty of eliciting informative priors in complex settings, Bayesian statisticians routinely use "default priors" for reasons of mathematical and computational convenience, rather than out of a sincere effort to encode their beliefs [Mayo, 2013].

Meanwhile, in recent decades there has been significant progress regarding the justification of credences in the formal epistemology literature. Early work in this area sought to place Bayesian epistemology on firmer epistemic ground than the classical pragmatic (Dutch book) arguments by showing that credences which follow the probability calculus and are updated by Bayesian conditionalization are, in some sense, closer to the truth than ones which do not obey these rules [Joyce, 1998, Greaves and Wallace, 2006, Leitgeb and Pettigrew, 2010a,b]. More recently, Dunn [2015] and Tang [2016] have discussed reliabilist accounts of the justification of particular credences (rather than just the Bayesian norms), and Pettigrew [2018] has argued that each of these accounts is extensionally equivalent to (i.e. each deems the same credences justified as) his own favored

epistemic utility account of justified credence. This account, as well as others discussed by Tang [2016], critically depends on probabilities which are *objective* in a frequentist sense; each says (*very* roughly) that a credence in a proposition is justified when it is close to the objective probability of that proposition.

I aim to bridge this gap between the foundations of statistics and recent developments in formal epistemology. I start with the view that the goal of a statistical analysis is to provide justified credences (among a set of alternatives) about observable quantities. Concretely, statisticians study problems like this:

> **Running Example:** Given the working class of models
>
> $$X_i \overset{i.i.d}{\sim} N(\theta_0, 1) : \theta_0 \in \mathbb{R},$$
>
> read "observables X_i are independently and identically distributed (i.i.d) as normal with variance 1 and unknown mean θ", and data $x_1, \ldots, x_n$, what is our best guess at the distribution of X_i?

Note that I will follow the convention of denoting generic random variables with capital letters, and *realizations* of those random variables — i.e. observations from the process encoded by those random variables — with the corresponding lower-case letters.

A typical answer to the question in (Running Example) is $N(\overline{x}_n, 1)$, where $\overline{x}_n$ is the sample mean of $x_1, \ldots, x_n$. In mainstream statistics, arguments for the claim that $\overline{x}_n$ is the best estimator of θ_0 (and therefore, implicitly, $N(\overline{x}_n, 1)$ is the best estimator of the distribution of X_i) are typically in terms of the "frequentist properties" of the estimator $\overline{X}_n$. For instance, assuming the data are truly distributed as i.i.d normal, then $\overline{X}_n$ is the *uniformly minimum variance unbiased estimator* of θ_0. "Unbiased" means that, in repeated samples $X_1, \ldots, X_n$, the corresponding estimates $\overline{X}_n$ are on average equal to θ_0. "Uniformly minimum variance" means that, under the i.i.d normal assumption and across *all* unbiased estimators, $\overline{X}_n$ has the lowest variance regardless of the value of θ_0. Because expected squared error loss decomposes into bias and variance [1] this in turn means that $\overline{X}_n$ has uniformly lowest expected squared error loss in the class of unbiased estimators of θ_0; qualitatively, it is closest to θ_0 on average, among unbiased estimators (again, under the i.i.d normal assumption).

This approach is to be contrasted with (orthodox) Bayesian statistics, wherein the analyst places a prior distribution over the class of candidate models, and updates this prior given data to obtain a posterior over candidate models. This posterior, according to the usual Bayesian story, encodes the analyst's rational degrees of belief about the underlying model. It is also used to obtain the "posterior predictive distribution" over the observable in question, which is supposedly the analyst's rational beliefs about future observables from the same process. Returning to (Running Example), the analyst may place a prior $\theta_0 \sim N(\theta_\pi, \sigma_\pi^2)$ over the possible values of the unknown θ_0. The resulting posterior and posterior predictive distributions turn out to be

$$\begin{aligned}(\theta_0 \mid x_1, \ldots, x_n) &\sim N\left(\frac{1}{\frac{1}{\sigma_\pi^2} + n}\left[\frac{1}{\sigma_\pi^2}\theta_\pi + n\overline{x}_n\right], \frac{1}{\frac{1}{\sigma_\pi^2} + n}\right) \\ (X_{n+1} \mid x_1, \ldots, x_n) &\sim N\left(\frac{1}{\frac{1}{\sigma_\pi^2} + n}\left[\frac{1}{\sigma_\pi^2}\theta_\pi + n\overline{x}_n\right], \frac{1}{\frac{1}{\sigma_\pi^2} + n} + 1\right).\end{aligned} \tag{1}$$

(We will see Equation 1 again in the discussion of prior reliability in Section 2.2.)

On the account I develop, arguments from frequentist properties are used to justify the choice of estimator of the objective probability distribution. On the other hand, the estimator of the objective probability distribution is given an *epistemic* interpretation, just as Bayesians give the posterior predictive distribution an epistemic interpretation. (Indeed, the posterior predictive distribution can be seen as a "Bayes estimator" (defined below) of the objective probability distribution for certain choices of scoring rule.) So my account focuses, like posterior predictive Bayesian analysis, on the quantification of (epistemic) uncertainty over observables, while tying the justification

[1] I.e. For an estimator $\widehat{\theta}$, $\mathbb{E}_{\widehat{\theta}|\theta_0}(\widehat{\theta} - \theta_0)^2 = \mathbb{E}^2_{\widehat{\theta}|\theta_0}(\widehat{\theta} - \theta_0) + \text{Variance}_{\widehat{\theta}|\theta_0}(\widehat{\theta}) = \text{Bias}^2_{\widehat{\theta}|\theta_0}(\widehat{\theta}) + \text{Variance}_{\widehat{\theta}|\theta_0}(\widehat{\theta})$, where subscript $(\widehat{\theta} \mid \theta_0)$ means an expectation taken with respect to the distribution of estimator $\widehat{\theta}$ when θ_0 is the true value of the parameter of interest.

of these epistemic probabilities to their reliability with respect to the objective probabilities, much in the spirit of frequentism.

In Section 2, I develop this reliabilist account of the justification of credences, called statistical reliabilism. While we will see that statistical reliabilism in a way refers to a familiar family of statistical principles, I will argue that taking its implications seriously leads to novel insights about strict versus permissive views on the justification of statistical methodology; the appropriate use of priors; and the separation of statistical activity into justification conditional on a set of models on one hand, and into exploration, model criticism, and model revision on the other.

In Section 3 I explore the connections between my account and other accounts of reliabilist justification of credence discussed in Tang [2016] and Pettigrew [2018]. I will show that, once amended to make these accounts action-guiding (in the sense of making definite statements about which credences are justified) in the absence of knowledge of the relevant objective probabilities, they can be seen as applications of the (SR) framework.

As a final remark on the scope of the article: I do not recapitulate the extensive arguments over reliabilism. However, in Section 3 I do show how my account, similarly to other accounts of reliabilist credal justification, may be defended from several important objections facing reliabilism. Nor do I take up the problem of analyzing objective probability. I will only say that a concept of objective probability is useful and perhaps indispensable in statistical practice, as well as in the explication of reliabilism, which I consider to be the most promising direction for the justification of belief.

2 Statistical reliabilism

Suppose we wish to form justified credences about observable quantity X, with (unknown) objective probability distribution P_0. In this section I give a process-reliabilist account of justification. In statistical language, "estimation" of the true distribution P_0 is an example of a process producing an epistemic probability distribution over X. So the best-justified credence, in the case of statistical estimation, is that produced by the estimator which is closest (in a sense) to the objective probabilities on average.

In statistics, an estimator is just a function from data to an estimate of a quantity of interest. In the introduction, we saw an estimator of the mean of a normal distribution, $T : (X_1, \ldots, X_n) \mapsto \overline{X}_n$. Estimators may instead map data to entire candidate objective distributions, for instance $T' : (X_1, \ldots, X_n) \mapsto N(\overline{X}_n, 1)$ from the same example. For simplicity, I will discuss estimation in terms of the estimation of model parameters. However, keep in mind that the distribution corresponding to the optimal estimator in the space of parameters need not be the distribution which is optimal with respect to a scoring rule that directly measures the distance to the true distribution. (See (Squared error) and (KL divergence) below as an example of this distinction.) Indeed, Lawless and Fredette [2005] show that "plugin" estimators like $N(\overline{X}_n, 1)$ are suboptimal (in a certain sense) with respect to KL-divergence, which is arguably a more appropriate measure of how close the estimated distribution is to the true objective distribution.

Here is a first pass at formulating the account:

> **Statistical reliabilism for estimators, first pass (SRE-α)**: An epistemic probability distribution $\widehat{p}$ over X is justified with respect to a class of estimators $\mathcal{T}$ and scoring rule L[2] iff, given data D and objective probability distribution P_0 for X, $\widehat{p}$ is a realization of $\widehat{P} = T(D)$ and $T \in \arg\min_{T' \in \mathcal{T}} \mathbb{E}_{D|P_0} L(T'(D), P_0)$.

Examples of scoring rules include:

- **Squared error**: Let P_0 belong to a class of models parameterized by a Euclidean vector θ — for instance, $\{N(\theta, 1) : \theta \in \mathbb{R}\}$ in (Running Example) — such that $\theta(P)$ returns the parameter corresponding to distribution P. Then,

$$L_{\text{squared error}} : (P, P_0) \mapsto \|\theta(P) - \theta(P_0)\|^2,$$

  where $\|\cdot\|$ is the Euclidean norm.

[2] In the statistical and machine learning literatures, scoring rules are usually called *loss functions*.

- **Kullback-Leibler (KL) divergence**:

$$L_{\text{KL divergence}} : (P, P_0) \mapsto \int P(x) \log \frac{P(x)}{P_0(x)} \mathrm{d}x.$$

KL divergence directly measures the inaccuracy of distribution P with respect to distribution P_0, rather than measuring the distance between the corresponding parameters.

But credence-forming processes need not depend only on data and estimator; we will see this in the discussion of priors in Section 2.2. So we have the more general formulation:

> **Statistical reliabilism, first pass (SR-α)**: An epistemic probability distribution $\widehat{p}$ over X is justified with respect to a class of credence-forming processes $\widehat{\mathcal{P}}$ and scoring rule L iff, given objective probability distribution P_0 for X, $\widehat{p}$ is a realization of $\widehat{P} \in \arg\min_{\widehat{P}' \in \widehat{\mathcal{P}}} \mathbb{E}_{\widehat{P}|P_0} L(\widehat{P}', P_0)$.

Introducing the *risk* function of statistical decision theory

$$R(\cdot, P_0) : T \mapsto \mathbb{E}_{D|P_0} L\big(T(D), P_0\big),$$

we see that (SRE-α) is exactly the requirement that the estimator minimize risk in the class $\mathcal{T}$. The problem, of course, is that R depends on the unknown distribution P_0. Luckily, statistics provides methods of measuring the risk of an estimator across a class of models $\mathcal{P}$ (e.g. $\mathcal{P} = \{N(\theta, 1) : \theta \in \mathbb{R}\}$ from (Running Example)) rather than just at single distribution. I will call such measures *reliability measures*. There are two classical reliability measures in statistical decision theory. The first is worst-case risk over $\mathcal{P}$, and the corresponding best estimators in this sense (when they exist) are called *minimax estimators*. The second is average risk with respect to a weighting function π over $\mathcal{P}$, and the corresponding estimators are called *Bayes estimators* [Berger, 2013]. We can write these as follows:

$$\begin{aligned} \mathcal{R}_{\text{minimax}}(T, \mathcal{P}) &= \sup_{P_0 \in \mathcal{P}} R(T, P_0) \\ \mathcal{R}_{\pi\text{-Bayes}}(T, \mathcal{P}) &= \int_{\mathcal{P}} R(T, P_0) \mathrm{d}\pi(P_0). \end{aligned}$$

Letting T_n be a sequence of estimators defined on a growing dataset D_n, with $T \in \{T_n\}_{n \in \mathbb{N}}$, we can also consider asymptotic measures as n grows. This often simplifies theoretical arguments. For instance, there is the asymptotic maximum risk over small neighborhoods of P_0,

$$\mathcal{R}_{\text{l.a.-minimax}}(T, \mathcal{P}) = \lim_{\delta \downarrow 0} \liminf_{n \to \infty} \sup_{\{P \in \mathcal{P} : \|\theta(P) - \theta(P_0)\| < \delta\}} R(T_n, P_0),$$

whose minimizers are called "local asymptotic minimax" estimators (and can in some cases be obtained without knowing P_0) [Van der Vaart, 2000]. Finally, there are reliability measures which do not depend on a class $\mathcal{P}$. Examples include resampling measures such as the bootstrap and cross-validation [Efron and Gong, 1983] and (penalized) empirical risk [Vapnik, 1992]. For instance, consider datasets $D = \{X_i\}_{i=1}^n$ and scoring rule L defined on pairs $(\widehat{P}, X)$. Then we have:

$$\begin{aligned} \mathcal{R}_{\text{penalized empirical risk}}(T, \mathcal{P}) &= \frac{1}{n} \sum_{i=1} L(T(D), X_i) + \text{Penalty}(T(D)) \\ &= \mathcal{R}_{\text{empirical risk}}(T, \mathcal{P}) + \text{Penalty}(T(D)). \end{aligned}$$

where Penalty is a penalty function which measures the complexity of the estimate $T(D)$. (We write these reliability measures as if they depend on $\mathcal{P}$ in order to keep the notation consistent.) The penalty is introduced to prevent minimizers of $\mathcal{R}_{\text{penalized empirical risk}}$ from over-fitting the data and therefore generalizing poorly to new samples.

Thus we have the amended formulation for estimators:

> **Statistical reliabilism for estimators (SRE)**: An epistemic probability distribution $\widehat{p}$ is justified with respect to a class of estimators $\mathcal{T}$, class of models $\mathcal{P}$, and reliability measure $\mathcal{R}$ iff, given total evidence D, $\widehat{p}$ is a realization of $\widehat{P} = T(D)$ and $T \in \arg\min_{T' \in \mathcal{T}} \mathcal{R}(T', \mathcal{P})$.

We may also define reliability measures on generic credence-forming processes to obtain the general formulation:

> **Statistical reliabilism (SR)**: An epistemic probability distribution $\widehat{p}$ is justified with respect to a class of credence-forming process $\widehat{\mathcal{P}}$, class of models $\mathcal{P}$, and reliability measure $\mathcal{R}$ iff $\widehat{p}$ is a realization of $\widehat{P} \in \arg\min_{\widehat{P}' \in \widehat{\mathcal{P}}} \mathcal{R}(\widehat{P}', \mathcal{P})$.

2.1 Choosing a reliability measure

How do we decide on a reliability measure $\mathcal{R}$? Classical statistical decision theory is divided over minimax and Bayesian choices. Each faces salient objections. Setting aside the fact that minimiax estimators do not always exist, their conservatism seems ill-motivated. On the other hand, the problem of the choice of prior π for the Bayesian choice is vexing if it is to be given an epistemic interpretation, as this requires the further justification of π; in 2.2 I give an alternative account of priors within the (SR) framework which avoids this difficulty. If π is treated simply as a weighting function introduced to allow for the measurement of reliability independent of P_0, and spreads weight evenly (in some sense) over $\mathcal{P}$, then it seems to be an attractive choice, as it avoids the pessimism of minimax while guaranteeing admissibility (i.e. that the estimator is not dominated in risk over $\mathcal{P}$ by another element of $\mathcal{T}$) under certain conditions (by the celebrated complete class theorems [Berger, 2013]). Of course, identifying functions which "spread weight evenly" over $\mathcal{P}$ is fraught, as the history of the principle of indifference attests [Shackel, 2007]. But if the deflationary and heuristical view of the choice of $\mathcal{R}$ mentioned below is appropriate, then these worries too may be deflated.

At least as concerning, and more neglected, questions of justification arise for asymptotic measures like $\mathcal{R}_{\text{l.a.-minimax}}$. Geyer et al. [2013] make the obvious but overlooked point that

> We know that asymptotics often works well in practical problems be- cause we can check the asymptotics by computer simulation (perhaps what Le Cam meant by "checked on the case at hand"), but conventional theory doesn't tell us why asymptotics works when it does. It only tells us that asymptotics works for sufficiently large n, perhaps astronomically larger than the actual n of the actual data. So that leaves a theoretical puzzle.
>
> - Asymptotics often works.
> - But it doesn't work for the reasons given in proofs.
> - It works for reasons too complicated for theory to handle.

The lack of a logical relationship between asymptotic results and the finite-sample reliability of estimators seems to leave, in the absence of finite-sample results which explain the connection (but also render the asymptotics superfluous), only induction from the observed success of asymptotically sound methods as a justification for their use.

As for "model-free" measures like $\mathcal{R}_{\text{penalized empirical risk}}$, these are typically regarded as *estimators* of an underlying reliability measure (and their use justified by the kind of asymptotic arguments which have just been called into question). However, for the purposes of this presentation I leave open the possibility that the model-free reliability measures may in some cases stand on their own. Indeed, we will see examples in Section 3 where there is no obvious way of specifying a class of models and therefore a model-free measure may be called for.

This leads me to consider a more general perspective on how the choice of reliability measure might be defended. We may treat the selection of $\mathcal{R}$ as a meta-decision problem, based on the modeler's experience with the outcomes of similar choices. Let $\mathcal{C}$ be a set of candidate reliability measures and, for each $\mathcal{R} \in \mathcal{C}$, let $U(\mathcal{R})$ be the observed utility of using $\mathcal{R}$ in similar past situations, or in relevant simulation experiments; $U(\mathcal{R})$ includes things like the predictive accuracy of the estimator derived from $\mathcal{R}$ on newly observed data, and relative computational cost. Then they may choose T by minimizing $\mathcal{R}^* = \arg\max_{\mathcal{R} \in \mathcal{C}} U(\mathcal{R})$. Of course, this meta-decision problem is inevitably informal given the ill-definedness of the terms involved. And such an account raises questions about how and when the empirical performance of statistical methodology in one setting should be expected to hold in another — questions which would likely benefit from philosophical input. But this is at least a plausible account of how experienced modelers do, in fact, choose among

statistical approaches, and lends some justification to the widespread ecumenism and "pragmatism" among statisticians.

Another, perhaps indirect, reason for a deflationary view of disputes over the choice of $\mathcal{R}$ is that justification on (SR) is tentative in the first place. As I discuss at greater length in Section 2.3, the justification of credence relative to a set of models and estimators is only a small part of statistical activity. Much more work goes into exploring the data in order to generate new modeling strategies, and searching for flaws in the assumptions which lead to the credences / estimates. It is difficult, given this point of view, to expend much energy trying to identify a unique solution concept, and is perhaps enough to regard $\mathcal{R}$ as a *heuristic* for the quality of statistically-grounded credences, underdetermined but still justified by qualitative considerations of reliability.

2.2 Priors

Rather than considering only epistemic probability distributions arising from deterministic estimators $T(D)$, we may consider estimators which contain an additional source of randomness: the inclusion of prior information. Priors — at least, informative priors (those not constructed using principle of indifference-like arguments [Yang and Berger, 1996]) — are appropriately regarded as random variables in this framework because they result from a chancy credence-forming process: the elicitation of a prior from an expert. Crucially, Bayes estimators, which are constructed by updating this elicited prior given the data-at-hand, will be more reliable than estimators which use only the data-at-hand only if the prior elicitation process meets a certain threshold of reliability.

To continue with (Running Example): given a certain class of priors, the Bayes estimator of θ_0 in (Running Example) is given by $\widehat{\theta}_{\text{Bayes}} = \alpha_n(\sigma_\pi^2) \cdot \overline{X}_n + (1 - \alpha_n(\sigma_\pi^2)) \cdot \theta_\pi$, where $(\theta_\pi, \sigma_\pi^2)$ are the prior mean and variance and $\alpha_n(\sigma_\pi^2)$ is a weight in $(0, 1)$ that depends on the sample size and prior variance (Equation 1).

On my account, $(\theta_\pi, \sigma_\pi^2)$ is treated as a random variable, and in particular as the output of a chancy process which may be more or less reliable with respect to the true distribution $N(\theta_0, 1)$ on some appropriate reference class. Then the expected squared-error loss with respect to the true mean θ_0 is:

$$\begin{aligned} R(\widehat{\theta}_{\text{Bayes}}, \theta_0) =& \mathbb{E}_{\widehat{\theta}_{\text{Bayes}}|\theta_0}(\widehat{\theta}_{\text{Bayes}} - \theta_0)^2 \\ =& \mathbb{E}_{(X_1,\ldots,X_n,\sigma_\pi^2)|\theta_0}\alpha_n^2(\sigma_\pi^2)(\overline{X}_n - \theta_0)^2 + \mathbb{E}_{(\theta_\pi,\sigma_\pi^2)|\theta_0}(1 - \alpha_n(\sigma_\pi^2))^2(\theta_\pi - \theta_0)^2 + \\ & 2\mathbb{E}_{(X_1,\ldots,X_n,\theta_\pi,\sigma_\pi^2)|\theta_0}\alpha_n(\sigma_\pi^2)(1 - \alpha_n(\sigma_\pi^2))(\theta_\pi - \theta_0)(\overline{X}_n - \theta_0). \end{aligned}$$

So, the risk with respect to the squared error loss decomposes into the risk of the sample mean, the risk of the prior, and the covariance between the error of the sample mean and that of the prior. Thus, if we have reason to believe — either due to evidence from a formal elicitation process, or qualitative domain-specific considerations — that the unreliability of $(\theta_\pi, \sigma_\pi^2)$, $\mathbb{E}_{(\theta_\pi,\sigma_\pi^2)|\theta_0}\alpha_n(\sigma_\pi^2)^2(\theta_\pi - \theta_0)^2$, is high, then we might prefer the no-prior estimator $\overline{X}_n$.

This simultaneously vindicates the Bayesian insistence that prior information be considered — *insofar as* the inclusion of prior information improves reliability — and makes sense of efforts to elicit reliable priors, which are difficult to justify in the classical Bayesian framework which places minimal constraints on credences. (The entire subfield of prior elicitation is dedicated to developing methods for improving the reliability of prior elicitation methods, for instance by developing exercises which improve the calibration of domain experts whose priors are to be elicited [O'Hagan et al., 2006].) The reliabilist understanding of priors can also explain the hesitancy of statisticians to use informative priors in cases where it is unclear how to "correctly" represent prior knowledge, even if a formal prior elicitation process is not feasible for the analysis in question. On the other hand, (SR) leaves open the possibility of incorporating prior information in ways other than via Bayesian conditionalization. For instance, Boos and Monahan [1986] use priors in combination with the bootstrap to obtain posterior-like uncertainty quantification, and Schweder and Hjort [1999] describe methods for combining likeilhoods with "prior confidence distributions" as a frequentist analog to the Bayesian procedure for incorporating prior beliefs.

2.3 Contingency, exploration, and criticism

(SR) is an account of justification *with respect to* a class of models $\mathcal{P}$. While a limitation, I also take it to be a virtue of (SR) that it respects the open-endedness of statistical practice. Indeed,

forming credences based on a fixed set of models is only a small part of the working statistician's job. Exploring the data and critiquing the choice of $\mathcal{P}$ is often the bulk of the work. For instance, a statistician analyzing data from (Running Example) might examine diagnostic plots of $x_1, \ldots, x_n$, or conduct formal statistical tests which measure how unlikely the data are under the model assumptions, to determine whether $\{X_i \overset{i.i.d}{\sim} N(\theta, 1) : \theta \in \mathbb{R}\}$ is an appropriate class of models. Evidence of (say) heavy tails might lead them consider a wider class of candidate distributions, while evidence of (say) correlation between consecutive observations might lead them to re-evaluate the assumption of independence.

So (SR) is complementary to approaches which emphasize open-ended model-building, criticism, and revision such as Gelman and Shalizi [2013] and Mayo [2018] (though a rigorous account of how $\mathcal{P}$ should itself be chosen and revised is beyond the scope of the current work). And, while the dependence of data-based beliefs on model assumptions and the importance of data exploration and iterative model-checking may be truisms among mainstream statisticians, similar insights are perhaps neglected in the formal epistemology literature. In Section 3 I will place (SR) in the context of existing reliabilist theories of justified credence, and show how the this and other lessons of the statistical point of view may shed light on issues of the justification of belief beyond statistics.

3 Statistical reliabilism beyond statistics

Returning to the formal epistemology literature, I explore connections between (SR) and two other reliabilist accounts of justified credence. In Section 3.1, I develop amendments to accounts presented in Tang [2016] and Pettigrew [2018] which render them action-guiding by allowing them to make statements about justified credences without knowledge of objective probabilities. I then show that the amended accounts coincide with applications of (SR). In Section 3.2, I show how (SR) may be defended from salient objections to reliabilism in a similar fashion as previous accounts, but in a way that follows directly from the (SR) framework without needing to posit additional epistemic principles.

3.1 (SR) as an action-guiding version of other reliabilist accounts

On the first account considered here — (Brier) as discussed by Tang [2016] — credences are justified if they are produced by a process which yields credences that are close to the truth in expected squared error loss. Here is a comparative version, which I denote (Brier-α) to parallel SR-α; and where Tang only writes that credences should have "a low average Brier score", I write out the expectation to make its relation to the present account explicit:

> **Brier-α**: For any set $\widehat{\mathcal{P}}$ of credence- forming processes, a credence of $\widehat{p}$ in q is justified relative to $\widehat{\mathcal{P}}$ iff $\widehat{p}$ is a realization of $\widehat{P}$ and $\widehat{P} \in \arg\min_{\widehat{P}' \in \widehat{\mathcal{P}}} \mathbb{E}(\widehat{P} - X)^2$, where $X = \mathbb{1}(Q)$ are indicators of propositions Q[3] on an appropriate reference class for q, and the expectation is taken with respect to the joint distribution of $(\widehat{P}, X)$ on some relevant reference class of propositions.

(Brier-α) is inadequate because it is not action-guiding without knowledge of the relevant (objective) expectation for each $\widehat{P}$; without knowing the true Brier scores (expected squared errors) we do not know which credence to accept. However, (Brier-α) is actually a special case of (SR-α) and may therefore be amended to render an (action-guiding) version of (SR), which we can call (SR-Brier). Let $\{q_i\}_{i=1}^n$ be propositions in a relevant reference class and define $x_i = \mathbb{1}(q_i)$. For each $\widehat{P} \in \widehat{\mathcal{P}}$ let $\{\widehat{p}_i\}_{i=1}^n$ be n realizations of $\widehat{P}$. Then, one reliability measure is

$$\mathcal{R}_{\text{empirical risk}}(\widehat{P}, \mathcal{P}) = \frac{\sum_{i=1}^n (\widehat{p}_i - x_i)^2}{n},$$

which is exactly the (empirical) Brier score used to evaluate real-world forecasters and forecasting systems [Brier and Allen, 1951, Gneiting and Raftery, 2007].

Tang rejects (Brier-α). He objects that on (Brier-α), processes which form intermediate credences will never be *perfectly* justified, as their Brier score will never be 0. He raises the further

[3] That is, $\mathbb{1}(q') = \begin{cases} 0 & \text{when } q' \text{ is false;} \\ 1 & \text{when } q' \text{ is true.} \end{cases}$

worry that if "perfect justification" is relativized to a reasonable class of credence-forming processes, then it may be *too easy* to have perfectly justified credences; for instance, if an agent has access to only one, highly unreliable credence-forming process, (Brier-α) allows their credence to be perfectly justified. Tang says this is the wrong conclusion. I reply that it is beside the point whether a credence is "perfectly justified" if it is the only game in town. *However*, the unreliability of all available processes may prompt a rational agent to search for *better credence-forming processes*. This point is a generalization of the points made in Section 2.3 about the complementarity between justification contingent on a set of models, and processes which generate new and improved sets of models (or estimators). If we take the view that forming credences is part of an iterative process of fixing and expanding the set of available credence-forming processes, Tang's worry about the "easiness" of credal justification under (Brier-α) is defused.

The second account is Pettigrew [2018]'s epstemic value reliabilism:

> **Epistemic value reliabilism for justified credence (ERC):** credence of $\widehat{p}$ in proposition q by an agent S is justified iff

(ERC1) S has ground g;

(ERC2) The credence $\widehat{p}$ in q by S is based on ground g;

(ERC3) If S has ground $g' \subseteq g$, then the objective probability of q given that the agent has ground g' approximates or equals $\widehat{p}$ — that is $P_0(q \mid S \text{ has } g') \approx \widehat{p}$.

Pettigrew's account also fails to be action-guiding when the objective conditional probability in (ERC3) is unknown. However, we may reconcile (ERC) with (SR) by the following procedure. Suppose we have data $\{(x_i, \widehat{p}_i^j, g_i\}_{i=1}^n$ for each credence-forming process $\widehat{P}^j$. Then, given grounds g we can adopt the credence from credence-forming process corresponding to $j^* = \arg\min_j \frac{1}{\#\{g_i = g\}} \sum_{g_i = g} (\widehat{p}_i^j - x_i)^2$, i.e. the optimal empirical Brier score for cases where the credences are based on grounds g. This is just an application of (SR-Brier) to the subset of relevant (credence, proposition) pairs where the credence was formed based on g. Notice that this conditional Brier score is a measure of the average of squared errors $(P_0(q \mid S \text{ has } g) - \widehat{P}^j)^2$; thus we preserve the spirit of (ERC3) while having a procedure which allows us (at least in an idealized setting) to actually identify a (relatively) justified credence.

This leads to a generalization of (SR) in which epistemic probability distributions depend on the available grounds for belief:

> **Statistical reliabilism with grounds (SRG):** An epistemic probability distribution $\widehat{p}$, given the most inclusive available grounds for belief g, is justified with respect to a class of credence-forming processes $\widehat{\mathcal{P}}$ (which are random maps from grounds to probability distributions), class of models $\mathcal{P}$, and reliability measure $\mathcal{R}$ if $\widehat{p}$ is a realization of $\widehat{P}(g)$ and
> $\widehat{P}(\cdot) \in \arg\min_{\widehat{P}'(\cdot) \in \widehat{\mathcal{P}}} \mathcal{R}(\widehat{P}(\cdot)', \mathcal{P})$,

where reliability measures here act (in general) on *collections* of random variables indexed by grounds, i.e. random maps $g \mapsto \widehat{P}(g)$.

(SRG) is a generalization of (SR) because, in the simplest case, we may regard the credence-forming process $\widehat{P}$ as a black box (i.e. we do not have access to the grounds on which the credences are based), and therefore the grounds may be modeled as the constant $g = \{\}$. In this case, we can drop the dependence of $\widehat{P}(\cdot)$ on g and recover (SR).

To make (SRG) more concrete, take an adapted version of Pettigrew's story about the reliability of flower identification. Suppose a friend and I spend morning and evening looking at flowers in a field. For the i^{th} flower we encounter, we each write down our beliefs $\widehat{p}_i^1, \widehat{p}_i^2$ in the proposition $q_i =$ [Flower i is a corncockle], look up the answer $x_i = \mathbb{1}(q_i)$ in our guidebook, and mark it down. Upon seeing flower $n+1$, we each form credences $\widehat{p}_{n+1}^1, \widehat{p}_{n+1}^2$. Now I wish to adopt a *justified* credence from the two available credences. I write down the (overly simplistic) class of models $\mathcal{P}_1$:

$$(X_i, \widehat{P}_i^1) \overset{i.i.d}{\sim} P_0^1$$
$$(X_i, \widehat{P}_i^2) \overset{i.i.d}{\sim} P_0^2$$

where P_0^1 and P_0^2 are unknown, objective joint probability distributions.

$\mathcal{P}_1$ treats our respective credence-forming processes as black boxes. But suppose at each flower i we are additionally tracking the lighting level ℓ_i. We each know that our confidence is modulated by ℓ_i (we are less confident in low light), and therefore the grounds on which we each base our credences about the i^{th} flower are $g_i = \{\ell_i\}$. I may then build a more complex (though still overly simplistic) class of models $\mathcal{P}_2$:

$$((X_i, \widehat{P}_i^1) \mid \ell) \overset{i.i.d}{\sim} P_0^1(\ell)$$
$$((X_i, \widehat{P}_i^2) \mid \ell) \overset{i.i.d}{\sim} P_0^2(\ell),$$

(where P^1 and P^2 are now maps from grounds to conditional probability distributions), and invoke (SRG) to obtain (relatively) justified credences about X_{n+1} given grounds $\{\ell_{n+1}\}$. For concreteness, suppose $\ell_i \in \{\text{Dim}, \text{Bright}\}$, and that an equal number of flowers have been seen under Dim and Bright conditions. Then, suppose $\ell_{n+1} = \text{Dim}$. Taking $\mathcal{R}_{\text{empirical risk}}$ as a reliability measure, I have two ways to choose my (relatively) justified credence from $\{\widehat{p}_{n+1}^1, \widehat{p}_{n+1}^2\}$ given grounds $\{\ell_{n+1} = \text{Dim}\}$:

$$\widehat{p}_{n+1}^{\mathcal{P}_1}(\{\ell_{n+1}\}) = \widehat{p}_{n+1}^{j^*}, \text{ where } j^* = \underset{j \in \{1,2\}}{\arg\min} \quad \frac{1}{n} \sum_{i=1}^{n} (\widehat{p}_i^j - x_i)^2$$
$$\widehat{p}_{n+1}^{\mathcal{P}_2}(\{\ell_{n+1}\}) = \widehat{p}_{n+1}^{j^*}, \text{ where } j^* = \underset{j \in \{1,2\}}{\arg\min} \frac{1}{2n} \sum_{\ell_i = \text{Dim}} (\widehat{p}_i^j - x_i)^2.$$

At first blush, $\widehat{p}_{n+1}^{\mathcal{P}_2}$ seems clearly preferable to $\widehat{p}_{n+1}^{\mathcal{P}_1}$, because it is based on more inclusive grounds. However, it is based on *half as much data*. This means that the random variable $\widehat{P}_{n+1}^{\mathcal{P}_2}$ will have greater *variance* than $\widehat{P}_{n+1}^{\mathcal{P}_1}$, and variance reduces reliability (see Footnote 1, for example). So it is not automatically the case that the credence based on more inclusive grounds is preferable from the perspective of (SRG).

This point goes beyond the comparison of $\mathcal{P}_1$ and $\mathcal{P}_2$. We may consider still more realistic models; it is sensible, for example, to model $\widehat{P}_i^j(\{\ell_i\})$ as a dependent process whose accuracy is improving over time. Refer to some formalization of this set of candidate processes as $\mathcal{P}_3$. $\mathcal{P}_3$ is presumably a more faithful representation of the credence-forming processes, it is again higher-variance than $\mathcal{P}_1$ and $\mathcal{P}_2$ due to its greater complexity. In a formal setting, we might consider the (hierarchical) model class $\mathcal{P}_4 = \{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3\}$ and use model selection techniques (which are constructed to select the model class which optimally trades off complexity and fit to the data, in a reliabilist sense) [4] to arrive at a justified credence via (SRG).

Now, I do not envision people actually assessing their credences in this way. But the account exhibits several properties which should figure into reliabilist accounts of belief generally. First, judgements about the reliability of credences over observables depend on assumptions about the structure of the underlying process. Second, those assumptions may be critiqued and revised, both by reasoning about the plausible mechanisms underlying the reliability of the credence-forming process (as we did when we formulated model class $\mathcal{P}_3$), and by comparing (if only intuitively) how well the candidate models trade off simplicity and fit to the data (as would be formalized in the process of selection among the models in $\mathcal{P}_4$).

3.2 The Generality and Graining problems

The requirement in (ERC) that credences be based on the most inclusive available grounds is motivated (following Comesaña [2006] and Tang [2016]) by two problems for reliabilism: the Generality Problem [Conee and Feldman, 1998], and what Pettigrew calls the Graining Problem. The Generality Problem is that there seems to be no non-arbitrary way of saying what count as instances of a "process", and that some ways of defining the process by which a credence is produced will render that process reliable, while others will render it unreliable. The most-inclusive-grounds requirement aims to solve this problem by introducing a non-arbitrary constraint on the definition of the process-type: the process should be one of forming credences based on the most inclusive

[4] See Forster and Sober [1994] for a philosophical discussion of a popular model selection technique.

available grounds. This simultaneously addresses the Graining Problem, which is the problem that credences may be based on different subsets of the available evidence.

If (ERC) is indeed safe from Generality and Graining, then (SRG) is, too, by also requiring that credences be based on the most inclusive grounds g (and, in the case of statistical estimation, the full data set D). But the role grounds and data play in the (SR) framework is somewhat subtle, and helps to explain the intuition motivating the Comesaña/Tang/Pettigrew most-inclusive-grounds requirement. This is because (SRG) does not require that the entire dataset or grounds is "active" in the same way that (ERC) requires credences to depend on the most inclusive grounds; on (SR), we are allowed to consider estimators which discard any amount of the information at hand. To take an extreme case, we may even include in $\mathfrak{T}$ estimators of the form $T : D \mapsto c$ for some constant c. For rich-enough $\mathfrak{T}$, however, such estimators will be *automatically* discarded as a consequence of our formulation, because estimators which throw away information in this way are known to be unreliable. As an example, consider the squared error loss of estimators $T_1 : D \mapsto c$ and $T_2 : D \mapsto \overline{X}_n$ of the true mean of $X_i \sim N(\theta_0, 1)$. For any θ_0, the risks of each with respect to the squared error loss turn out to be (given n "draws" from $N(\theta_0, 1)$):

$$R(T_1(D), \theta_0) = \mathbb{E}_{D|\theta_0}(T_1(D) - \theta_0)^2 = (c - \theta_0)^2$$
$$R(T_2(D), \theta_0) = \mathbb{E}_{D|\theta_0}(T_2(D) - \theta_0)^2 = \frac{1}{n}.$$

Thus any of the reliability measures we have seen (with the exception of Bayes estimators which use a weighting function concentrated on c) would prefer $N(\overline{X}_n, 1)$ to $N(c, 1)$ as an epistemic probability distribution for X_{n+1}.

Or, consider Pettigrew [2018]'s example of a person forming a credence about the color of a ball drawn from an urn. Half of the balls are black and half are white and each is marked with a unique number. The person has drawn a ball which they see as White and marked with the number 73. What credence should they form on the basis of their visual experience $g = \{\text{Color=White}, \text{Number=73}\}$? Pettigrew points out that it is irrational to base our beliefs only on the grounds $\{\text{Number=73}\}$ (and therefore form a credence of $\frac{1}{2}$ that the ball is White), which is the sort of observation that motivates the "most-inclusive grounds" stipulation. However, from the point of view of (SR) there is no need to invoke an additional principle. Consider two credence forming-processes, one based only on $\{\text{Number}\}$ and the other based on the full data $\{\text{Color}, \text{Number}\}$:

$$\widehat{P}_1 : \{\text{Color}, \text{Number}\} \mapsto P(\text{Color} = \text{White} \mid \text{Number}) = \frac{1}{2}$$
$$\widehat{P}_2 : \{\text{Color}, \text{Number}\} \mapsto \mathbb{1}(\text{Color} = \text{White}),$$

Then (assuming the person's vision is intact) $\widehat{P}_2$ is perfectly reliable, but $\widehat{P}_1$ is not; its expected squared error loss, for instance, is 0.25.

On the other hand, as discussed in Section 2.2, "throwing away" data may sometimes *improve* reliability, if that data enters via an unreliable process — for instance, via systematically bad prior elicitation in the case of Bayesian estimation. Similarly, because more complicated models have higher variance (and therefore, all else equal, lower reliability), we may sometimes prefer credences which do *not* "activate" the most inclusive grounds. The upshot is that the statistical-reliabilist point of view explains why we should *generally* base our beliefs on all of the available evidence: evidence provides information, and more information leads to more reliable beliefs (in a way that can be quantified in formal statistical settings). At the same time (SR) explains why we may sometimes rightfully reluctant to include certain sources of information when forming beliefs.

4 Conclusion

There are several issues I have not had space to discuss here. One is how the choice of scoring rule L figures into the justification of the corresponding reliability measure. There has been substantial discussion of the epistemic significance of scoring rules in both statistics and formal epistemology [Gneiting and Raftery, 2007, Leitgeb and Pettigrew, 2010a, Levinstein, 2017], and there is surely more to be said about how various putative epistemic requirements for scoring rules (such as "propriety") interact with other considerations (such as computational efficiency) in the (SR) framework. Another area I have neglected is the *confidence distribution* approach to statistical

foundations. While little-discussed in the philosophy literature, the confidence distribution approach purports to unify frequentist and Bayesian approaches to statistics by giving an account of epistemic probability ("confidence") which is grounded in frequentist properties [Schweder, 2018]. In particular, confidence distributions are distributions on the space of parameters for a class of models which make calibrated statements of confidence, in the sense that (say) 95% confidence intervals derived from such a distribution will (under the model assumptions) contain the true parameter value in 95% of samples [Xie and Singh, 2013]. This makes the confidence approach a rival to (SR) as a reliabilist foundation for statistics. However, confidence distributions need not be optimal in the (SR) sense (though they turn out to be in some cases; see Lawless and Fredette [2005]), and therefore do not exhibit what I regard as the appropriate sense of reliability. Indeed, this is an instance of the often-noted inadequacy of calibration relative to accuracy as a notion of reliability for probability statements; see the discussion of (Calibration) versus (Brier) in Tang [2016], for example.

The (SR) account also points to promising opportunities for collaborations between formal epistemologists and statisticians. First, as touched on briefly in 2.1, the nature of the justification of a statistical procedure via asymptotic arguments and simulation experiments would likely benefit from philosophical clarification. Second, the adoption of reliabilism for credences raises questions about what a reliabilist *decision theory* might look like; for instance, should agents calculate expected utilities by averaging utilities against an (SR)-justified epistemic probability distribution? Or should expected utility be estimated directly, just as epistemic probabilities are obtained by estimating objective probabilities in (SR)? This question parallels questions in the literature on optimal sequential decision-making over "model-based" versus "model-free" approaches [Sutton and Barto, 2018], as well as "direct" versus "indirect" methods of optimizing expected utility [Zhao and Laber, 2014].

References

James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

Dennis D Boos and John F Monahan. Bootstrap methods using prior information. *Biometrika*, 73 (1):77–83, 1986.

Glenn W Brier and Roger A Allen. Verification of weather forecasts. In *Compendium of meteorology*, pages 841–848. Springer, 1951.

Juan Comesaña. A well-founded solution to the generality problem. *Philosophical Studies*, 129(1): 27–47, 2006.

Earl Conee and Richard Feldman. The generality problem for reliabilism. *Philosophical Studies*, 89(1):1–29, 1998.

Jeff Dunn. Reliability for degrees of belief. *Philosophical Studies*, 172(7):1929–1952, 2015.

Bradley Efron. *Modern science and the Bayesian-frequentist controversy*. Division of Biostatistics, Stanford University, 2005.

Bradley Efron and Gail Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983.

Malcolm Forster and Elliott Sober. How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, 45 (1):1–35, 1994.

Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013.

Charles J Geyer et al. Asymptotics of maximum likelihood without the lln or clt or sample size going to infinity. In *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton*, pages 1–24. Institute of Mathematical Statistics, 2013.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Hilary Greaves and David Wallace. Justifying conditionalization: Conditionalization maximizes expected epistemic utility. *Mind*, 115(459):607–632, 2006.

James M Joyce. A nonpragmatic vindication of probabilism. *Philosophy of science*, 65(4):575–603, 1998.

JF Lawless and Marc Fredette. Frequentist prediction intervals and predictive distributions. *Biometrika*, 92(3):529–542, 2005.

Hannes Leitgeb and Richard Pettigrew. An objective justification of bayesianism i: Measuring inaccuracy. *Philosophy of Science*, 77(2):201–235, 2010a.

Hannes Leitgeb and Richard Pettigrew. An objective justification of bayesianism ii: The consequences of minimizing inaccuracy. *Philosophy of Science*, 77(2):236–272, 2010b.

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

Deborah G Mayo. Discussion: Bayesian methods: Applied? yes. philosophical defense? in flux. *The American Statistician*, 67(1):11–15, 2013.

Deborah G Mayo. *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press, 2018.

Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons, 2006.

Richard Pettigrew. What is justified credence? *Episteme*, 2018. doi:0.1017/epi.2018.50.

Tore Schweder. Confidence is epistemic probability for empirical science. *Journal of Statistical Planning and Inference*, 195:116–125, 2018.

Tore Schweder and Nils Lid Hjort. Frequentist analogues of priors and posteriors. *Preprint series. Statistical Research Report http://urn. nb. no/URN: NBN: no-23420*, 1999.

Nicholas Shackel. Bertrand's paradox and the principle of indifference. *Philosophy of Science*, 74 (2):150–175, 2007.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Weng Hong Tang. Reliability theories of justified credence. *Mind*, 125(497):63–94, 2016.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992.

Min-ge Xie and Kesar Singh. Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, 81(1):3–39, 2013.

Ruoyong Yang and James O Berger. *A catalog of noninformative priors*. Institute of Statistics and Decision Sciences, Duke University, 1996.

Ying-Qi Zhao and Eric B Laber. Estimation of optimal dynamic treatment regimes. *Clinical Trials*, 11(4):400–407, 2014.