Valid Inferential Models Offer Performance and Probativeness Assurances

Leonardo Cella(✉) and Ryan Martin

Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA
{lolivei,rgmarti3}@ncsu.edu

Abstract. Bayesians and frequentists are now largely focused on developing methods that *perform* well in a frequentist sense. But the widely-publicized replication crisis suggests that performance guarantees are not enough for good science. In addition to reliably detecting hypotheses that are incompatible with data, users require methods that can *probe* for hypotheses that are actually supported by the data. In this paper, we demonstrate that valid inferential models achieve both performance and probativeness properties. We also draw important connections between inferential models and Deborah Mayo's severe testing.

Keywords: Bayesian · Frequentist · p-value · Possibility measure · Severity

1 Introduction

Important decisions affecting our everyday experiences are increasingly data-driven. But is data helping us make better decisions? The widely-publicized *replication crisis* in science raises serious concerns, e.g., the American Statistical Association's president commissioned a formal *Statement on Statistical Significance and Replicability.*[1] The lack of any clear guidance in that statement reveals that there are important and fundamental questions concerning the foundations of statistical inference that remain unanswered:

> *Should probability enter to capture degrees of belief about claims? ... Or to ensure we won't reach mistaken interpretations of data too often in the long run of experience?* (Mayo 2018, p. xi)

The two distinct roles of probability above correspond to the classical frequentist and Bayesian schools of statistical inference, which have two fundamentally different priorities, referred to here as *performance* and *probativeness*, respectively. Over the last 50+ years, however, the lines between the two perspectives

[1] https://magazine.amstat.org/blog/2021/08/01/task-force-statement-p-value/.

R. Martin—Partially supported by the U.S. National Science Foundation, SES–205122.

S. Le Hégarat-Mascle et al. (Eds.): BELIEF 2022, LNAI 13506, pp. 219–228, 2022.
https://doi.org/10.1007/978-3-031-17801-6_21

and their distinct priorities have been blurred. Indeed, both Bayesians and frequentists now focus almost exclusively on performance. Such considerations are genuinely important for the logic of statistical inference:

> *even if an empirical frequency-based view of probability is not used directly as a basis for inference, it is unacceptable if a procedure... of representing uncertain knowledge would, if used repeatedly, give systematically misleading conclusions* (Reid and Cox 2015, p. 295).

As the replication crisis has taught us, there is more to inference than achieving, say, Type I and II error rate control. Beyond performance, we are also concerned with probativeness, i.e., can methods probe for hypotheses that are genuinely supported by the observed data? Modern statistical methods cannot achieve both performance and probativeness objectives, so a fully satisfactory framework for scientific inferences requires new perspectives.

To set the scene, denote the observable data by Y. The statistical model for Y will be denoted by $\mathsf{P}_{Y|\theta}$, where $\theta \in \Theta$ is an unknown model parameter. Note that the setup here is quite general: Y, θ, or both can be scalars, vectors, or something else. We focus here on the typical case where *no genuine prior information is available/assumed*. So, given only the model $\{\mathsf{P}_{Y|\theta} : \theta \in \Theta\}$ and the observed data $Y = y$, the goal is to quantify uncertainty about the unknown θ for the purpose of making inference. For concreteness, we will interpret "making inference" as making (data-driven) judgments about hypotheses concerning θ. Let $\mathcal{H}$ denote a collection of subsets of Θ, containing the singletons and closed under complementation, and associate $H \in \mathcal{H}$ with a hypothesis about θ.

Section 2.1 briefly describes the Bayesian vs. frequentist *two-theory problem* in our context of hypothesis testing. There we justify our above claim that modern statistical methods fail to meet both the performance and probativeness objectives. This includes the default-prior Bayes solution that aims to strike a balance between the two theories. What holds the default-prior Bayes solution back from meeting the performance and probativeness objectives is its lack of calibration, which is directly related to the constraint that the posterior distribution be a precise probability. Fortunately, the relatively new *inferential model* (IM) framework, reviewed briefly in Sect. 2.2 below, is able to achieve greater flexibility by embracing a certain degree of imprecision in its construction. Our main contribution here, in Sect. 3, is to highlight the IM's ability to simultaneously achieve both *performance* and *probabitiveness*. Two illustrations are presented in Sect. 4 and some concluding remarks are given in Sect. 5.

The probativeness conclusion is a direct consequence of the IM output's imprecision. That the additional flexibility of imprecision creates opportunities for more nuanced judgments is one of the motivations for accounting for imprecision, so this is no big surprise. But our contribution here is valuable for several reasons. First, the statistical community is aware of this need to see beyond basic performance criteria (e.g., Mayo 2018), but no clear, general, and easy-to-follow guidance has been offered. What we are suggesting here, however, is simple: *just follow the general theory of valid IMs and you get both performance and probativeness assurances.* Second, it showcases the importance of the role of belief functions and imprecise probability more generally, by reinforcing the key point that imprecision is *not* due to an inadequate formulation of the problem, but, rather, an essential part of the complete solution.

2 Background

2.1 Two-Theory Problem

In a nutshell, the two dominant schools of thought in statistics are as follows.

Bayesian. Uncertainty is quantified directly through specification of a prior probability distribution for θ, representing the data analyst's *a priori* degrees of belief. Bayes's theorem is then used to update the prior to a data-dependent posterior distribution for θ. The posterior probability of a hypothesis H represents the analyst's degree of belief in the truthfulness of H, given data, and would be essential for inference concerning H. That is, the magnitudes of the posterior probabilities naturally drive the data analyst's judgments about which hypotheses are supported by the data and which are not.

Frequentist. Uncertainty is quantified indirectly through the use of reliable procedures that control error rates. Consider, e.g., a p-value for testing a hypothesis H. What makes such a p-value meaningful is that, by construction, it tends to be not-small when H is true. Therefore, observing a small p-value gives the data analyst reason to doubt the truthfulness of H:

> *The force with which such a conclusion is supported is logically that of the simple disjunction: Either an exceptionally rare chance has occurred, or* [the hypothesis] *is not true* (Fisher 1973, p. 42).

The p-value *does not* represent the "probability of H" in any sense. So, a not-small (resp. small) p-value cannot be interpreted as direct support for H (resp. H^c) or any sub-hypothesis thereof.

The point is that, at least in principle, Bayesians focus on probativeness whereas frequentists focus on performance. But the line between frequentist and modern Bayesian practice is not so clear. Even Bayesians typically assume little or no prior information, as we have assumed here, so default priors are the norm (e.g., Berger 2006; Jeffreys 1946). But with a default prior, the "degree of belief" interpretation or the posterior probabilities is lost,

> [Bayes's theorem] *does not create real probabilities from hypothetical probabilities* (Fraser 2014, p. 249)

and, along with it, the probative nature of inferences based on them,

> *...any serious mathematician would surely ask how you could use* [Bayes's theorem] *with one premise missing by making up an ingredient and thinking that the conclusions of the* [theorem] *were still available* (Fraser 2011, p. 329).

The default-prior Bayes posterior probabilities could still have performance assurances *if* they were suitably calibrated. But the *false confidence theorem* of Balch et al. (2019) shows that this is not the case: there exists false hypotheses to which the posterior distribution tends to assign large probabilities. This implies that inferences based on the magnitudes of default-prior Bayes posterior probabilities can be "systematically misleading" (cf. Reid and Cox). This is perhaps why modern Bayesian analysis focuses less on the posterior probabilities and more on the performance of procedures (tests and credible sets) derived from the posterior. Hence modern Bayesians and frequentists are not so different.

The key take-away message is as follows. Frequentist methods focus on detecting incompatibility between data and hypotheses (performance), so they do not offer any guidance on how to identify hypotheses actually supported by the data (probativeness). Default-prior Bayesian methods are effectively no different, so this critique applies to them too. More specifically, the default-prior Bayes posterior probabilities lack the calibration necessary to reliably check for either incompatibility or support. Therefore, neither of the dominant schools of thought in statistical inference are able to simultaneously achieve both the performance and probativeness objectives.

2.2 Inferential Models Overview

Inferential models (IMs) were first developed in Martin and Liu (2013, 2015) to balance the Bayesians' desire for belief assignments and the frequentists' desire for error rate control. A key distinction between IMs and the familiar Bayesian and frequentist frameworks is that the output is an *imprecise probability* or, more specifically, a *necessity–possibility measure* pair. This imprecision, however, is not the result of an inability to precisely specify a model, etc., it is a necessary condition for inference to be *valid* in the sense defined in Sect. 3 below.

> *Possibility is an entirely different idea from probability, and it is sometimes, we maintain, a more efficient and powerful uncertainty variable, able to perform semantic tasks which the other cannot* (Shackle 1961, p. 103).

The false confidence theorem establishes that validity cannot be achieved via ordinary probability. More recently it has been shown that the possibility-theoretic formulation is key to achieving the relevant performance-related properties.

The original IM construction put forward in Martin and Liu (2013) relied on suitable random sets, whereas Liu and Martin (2021) recently offered a direct construction using possibility measures. The latter construction starts by associating data Y and unknown parameter θ with an unobservable auxiliary variable U with known distribution P_U via the formula

$$A(Y, \theta, U) = 0, \quad U \sim \mathsf{P}_U.$$

Let π denote a plausibility contour on the U-space such that the corresponding possibility measure is *consistent* with P_U in the sense that $\mathsf{P}_U(B) \leq \sup_{u \in B} \pi(u)$ for all subsets B. Now define its extension to Θ as

$$\pi_y(\vartheta) = \sup_{u:A(y,\vartheta,u)=0} \pi(u), \quad \vartheta \in \Theta.$$

Assuming this is a genuine/normal contour, then it defines an IM for θ having the mathematical form of a possibility measure with upper probability

$$\overline{\Pi}_y(H) = \sup_{\vartheta \in H} \pi_y(\vartheta), \quad H \subseteq \Theta,$$

and lower probability $\underline{\Pi}_y(H) = 1 - \overline{\Pi}_y(H^c)$. The IM's output is meaningful thanks to the properties it satisfies, which we discuss in Sect. 3. The performance-related properties have been the focus in previous work, but it is interesting that the performance properties together with the inherent imprecision in the IM's possibilistic output leads to probativeness properties too.

3 Two P's in the Same Pod

3.1 Performance

As discussed above, the property that gets the most attention in the statistics literature is *performance*, i.e., procedures developed for the purpose of making inference-related decisions (e.g., accept or reject a hypothesis) have error rate control guarantees. This is genuinely important: if statistical methods are not even reliable, then they have no hope of helping to advance science.

Our main result here, which is not new, is that procedures derived from a valid IM achieve the desired performance-related properties. As presented in the reference given in Sect. 2.2 above, we say that an IM without lower and upper probability output $y \mapsto (\underline{\Pi}_y, \overline{\Pi}_y)$ is *valid* if

$$\sup_{\theta \in H} \mathsf{P}_{Y|\theta}\{\overline{\Pi}_Y(H) \le \alpha\} \le \alpha, \quad \text{for all}\, \alpha \in [0,1]\, \text{and all}\, H \in \mathcal{H}. \tag{1}$$

This means that, with respect to the model $\mathsf{P}_{Y|\theta}$, it is a relatively rare event that the IM assigns relatively small upper probability to a true hypothesis about θ. Property (1) closely resembles the defining stochastically-no-smaller-than-uniform property of p-values. As such, Fisher's "logical disjunction" argument also applies to the valid IM's output, giving it objective meaning.

Although we are not aware of Fisher ever making such a statement, we believe that Fisher's disdain for the Neyman-style behavioral approach to statistical inference at least partially stemmed from the fact that such properties would be immediate consequences of the calibration needed for his "disjunction" argument to apply. So if Fisher's calibration is satisfied, then Neyman's error rate control is a corollary. Indeed, if the IM with output $(\underline{\Pi}_y, \overline{\Pi}_y)$, with corresponding plausibility contour π_y, is valid in the sense of (1), then

- for any fixed $\alpha \in (0,1)$, the test "reject H if and only if $\overline{\Pi}_y(H) \le \alpha$" controls the frequentist Type I error probability at level α, and

- for any fixed $\alpha \in (0,1)$, the set $C_\alpha(y) = \{\vartheta : \pi_y(\vartheta) > \alpha\}$ is a $100(1-\alpha)\%$ confidence set, i.e., its frequentist coverage probability at least $1-\alpha$.

These claims are almost immediate consequences of (1); see, e.g., Martin (2021) for a proof. Therefore, valid IMs offer performance guarantees.

Here it is also worth briefly pointing out that the connection between valid IMs and performance guarantees is even more fundamental. It was recently shown in Martin (2021) that *every* procedure with provable frequentist performance guarantees has, working behind the scenes, a valid IM with the form of a possibility measure. So, not only does the IM framework offer performance guarantees, it is really the only framework that does so. This also highlights the deep connections between frequentist inference and possibility theory.

3.2 Probativeness

The literature on IMs has largely focused on performance, i.e., that (1) implies that the output is suitably calibrated which leads to the results quoted in Sect. 3.1 above. While the IM output does, as discussed above, represent lower and upper probabilities, or degrees of necessity/support and possibility, a clear explanation of their post-data interpretation, and why non-additivity is valuable, has yet to be given. This section aims to fill that gap.

Standard performance metrics, such as Type I and Type II error probabilities, are not data-dependent and, therefore, cannot directly speak to whether the actual observed data offer any direct support to a particular hypothesis. The IM output returns both lower and upper probabilities but, so far, the literature has largely only focused on one of these, typically the upper probability. Perhaps the lower probability will be of some value after all.

Suppose that the data y is such that $\overline{\mathit{\Pi}}_y(H)$ is relatively large, i.e., the data are not incompatible with the hypothesis H. If, instead, $\overline{\mathit{\Pi}}_y(H)$ were small, then we can apply all of what we are about to describe to H^c instead of H. If we determine that the data are not incompatible with H, then a natural follow-up question is to ask if the data actually *support* the hypothesis H or any proper subset, say, $H' \subset H$. For this, we propose to consider the lower probability

$$\underline{\mathit{\Pi}}_y(H') = 1 - \overline{\mathit{\Pi}}_y(H'^c) = 1 - \sup\{\pi_y(\vartheta) : \vartheta \not\in H'\},$$

where the right-most expression is exclusive to the case where the IM output takes the form of a possibility measure, as we consider here. Coincidentally or not, Shafer (1976, Ch. 11) refers to the lower probability function, $H \mapsto \underline{\mathit{\Pi}}_y(H)$, as a *support function*, which is consistent with how we propose to use it here. If $\overline{\mathit{\Pi}}_y(H)$ is not small, then $\underline{\mathit{\Pi}}_y(H') \leq \overline{\mathit{\Pi}}_y(H)$ can be small or (relatively) large, and its magnitude determines the extent to which the data supports the truthfulness of H', beyond just compatibility or plausibility. On the one hand, if $\underline{\mathit{\Pi}}_y(H')$ is small and $\overline{\mathit{\Pi}}_y(H)$ is relatively large, then H' is plausible—or not incompatible—with the data y but there is little direct support in y for its truthfulness. This corresponds to a case with relatively large "don't know" in the sense of Dempster

(2008). On the other hand, if both $\underline{\Pi}_y(H')$ and $\overline{\Pi}_y(H)$ are relatively large, then y is not only compatible with H, it also directly supports H'.

What makes the "if $\underline{\Pi}_y(H')$ is relatively large, then infer H'" judgment warranted? Readers familiar with imprecise probability might be surprised by this question—this is precisely what lower probabilities are designed for—but remember that $\underline{\Pi}_y$ is not a subjective assessment of the data analyst's degrees of belief. So the data analyst should require, à la Reid and Cox, that their IM will tend not to lead them to erroneous judgments. Like $\underline{\Pi}_y$ is the dual to $\overline{\Pi}_y$, there is a corresponding dual to the validity property (1):

$$\sup_{\theta \notin H} \mathsf{P}_{Y|\theta}\{\underline{\Pi}_Y(H) > 1-\alpha\} \leq \alpha, \quad \text{for all}\, \alpha \in [0,1]\, \text{and all}\, H \in \mathcal{H}. \tag{2}$$

It is easy to verify that (2) and (1) are equivalent, but it is worth considering both versions because, while the latter refers primarily to assessments of compatibility between data and hypotheses, the former is relevant to judgments about when data actually support a certain hypothesis.

In Sect. 1, we remarked that there have been recent efforts by statisticians to supplement the standard p-values, etc. with measures designed to *probe* for hypotheses supported by the data. In particular, Mayo (2018) proposes a so-called *severity* measure but only gives one concrete example. If we extrapolate her suggestion beyond that one example, then it boils down to what we described above. That is, the map $H' \mapsto \underline{\Pi}_y(H')$ on subcollections of $\mathcal{H}$ can be used to probe for hypotheses that are actually supported by the data.

There is, however, a minor difference between ours and Mayo's perspective. On the one hand, Mayo is thinking in terms of a specific test of a particular hypothesis, so her severity measure is intended to describe how severe the test is, how deep that tests probes for actual support in the data beyond just compatibility or lack thereof. On the other hand, we are thinking in terms of big-picture uncertainty quantification. In light of the fundamental connection between valid IMs and frequentist inference, perhaps it is no surprise that Mayo's proposal, despite coming from a slightly different perspective, ends up directly aligning with what the valid IM does automatically; see Sect. 4.1. It is now clear that probabitiveness is inherent in the valid IM—no supplements needed!

4 Illustrations

4.1 Normal Mean

Mayo (2018, p. 142) describes a hypothetical water plant where the water it discharges is intended to be roughly 150° Fahrenheit. More specifically, under ideal settings, water temperature measurements ought to be normally distributed with mean 150° and standard deviation 10°. To test the water plant's settings, a sample $Y = (Y_1, \ldots, Y_n)$ of $n = 100$ water temperature measurements are taken; then the sample mean, $\bar{Y}$, is $\mathsf{N}(150, 1)$. Since water temperatures higher than 150° might damage the ecosystem, of primary interest are hypotheses $H_\vartheta =$

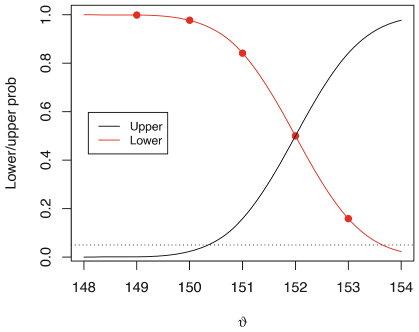


Fig. 1. Results of the valid IM applied to Mayo's normal mean example; the red dots correspond to the values in Table 3.1 of Mayo (2018) (Color figure online)

$(-\infty, \vartheta]$ for ϑ near 150. For hypotheses of this form, the "optimal" IM (Martin and Liu 2013, Sect. 4.3) has upper probability

$$\overline{\mathit{\Pi}}_y(H_\vartheta) = 1 - \Phi(\bar{y} - \vartheta), \quad \vartheta \in \mathbb{R},$$

where Φ denotes the standard normal distribution function.

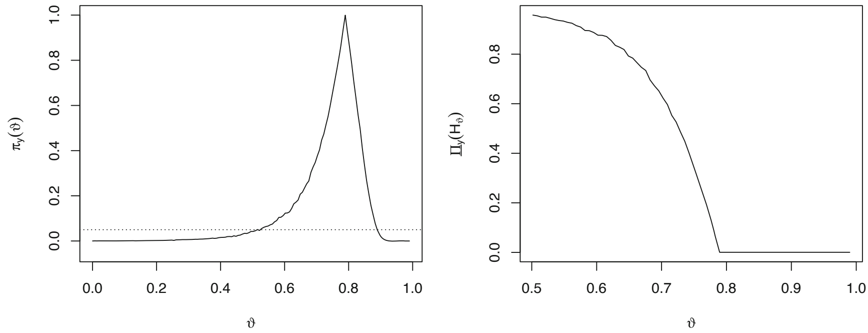
Suppose we observe $\bar{y} = 152$, which is potentially incompatible with the hypothesis H_{150}. Indeed, a plot of the upper probability is shown in Fig. 1(a) and we see that, at $\vartheta = 150$, the upper probability is smaller than 0.05, so we would be inclined to reject the hypothesis $\theta \leq 150$. To probe for support of subsets of the alternative hypothesis, we also plot the lower probability

$$\underline{\mathit{\Pi}}_y(H_\vartheta^c) = \Phi(\bar{y} - \vartheta), \quad \vartheta \in \mathbb{R},$$

and we see that there is, in fact, non-negligible support in the data for, say, $H_{151}^c = (151, \infty)$. These results agree exactly with the analysis presented in Mayo (2018) based on her supplement of the ordinary p-value with a severity measure. Mayo elaborates on this example in a couple different ways but, for the sake of space, suffice it to say that our analysis perfectly agrees with hers.

4.2 Bivariate Normal Correlation

Suppose that Y consists of n independent and identically distributed pairs $Y_i = (Y_{1,i}, Y_{2,i})$ having a bivariate normal distribution with zero means, unit variances, and correlation $\theta \in [-1, 1]$. Let $\mathsf{P}_{Y|\theta}$ denote the corresponding joint distribution. An asymptotic pivot based on the maximum likelihood estimator, $\hat{\theta}$, can be constructed and the corresponding Wald test would look very similar to that in Sect. 4.1. This bivariate normal correlation problem, however, corresponds to one of those "curved exponential families" where $\hat{\theta}$ is not a sufficient statistic so some efficiency is lost in the Wald test for finite n. So we take a different approach here, which extends us beyond the cases Mayo considers.



(a) Plausibility contour, π_y

(b) $\vartheta \mapsto \underline{\mathit{\Pi}}_y(H_\vartheta)$

Fig. 2. Results of the valid IM applied to Efron's law school admissions data.

Let $\vartheta \mapsto L_y(\vartheta)$ denote the likelihood function for θ based on data y. Following Martin (2015, 2018), a valid IM can be constructed based on the relative likelihood, $r_y(\vartheta) = L_y(\vartheta)/L_y(\hat{\theta})$, with plausibility contour function

$$\pi_y(\vartheta) = \mathsf{P}_{Y|\vartheta}\{r_Y(\vartheta) \leq r_y(\vartheta)\}, \quad \vartheta \in [-1, 1].$$

This resembles the p-value function for a suitable likelihood ratio test. The IM's output, $(\underline{\mathit{\Pi}}_y, \overline{\mathit{\Pi}}_y)$, is determined by optimizing the contour function.

As an illustration of the ideas presented above, consider the law school admissions data analyzed in Efron (1982), which consists of $n = 15$ data pairs with $Y_1 =$ LSAT scores and $Y_2 =$ undergrad GPA. For our analysis, we standardize these so that the mean zero–unit variance is appropriate. Of course, this standardization has no effect on the correlation, which is our object of interest. In this case, the sample correlation is 0.776; the maximum likelihood estimator, which has no closed-form expression, is $\hat{\theta} = 0.789$. A plot of the plausibility contour π_y for this data is shown in Fig. 2(a). The horizontal line at $\alpha = 0.05$ determines the 95% plausibility interval, which is an exact 95% confidence interval. It is clear that the data shows virtually no support for $\theta = 0$, but there is some marginal support for the hypothesis $H = (0.5, 1]$. To probe this further, consider the class of sub-hypotheses $H_\vartheta = (\vartheta, 1]$, $\vartheta > 0.5$. A plot of the function $\vartheta \mapsto \underline{\mathit{\Pi}}_y(H_\vartheta)$ is shown in Fig. 2(b). As expected from Panel (a), the latter function is decreasing in ϑ and we clearly see no support for H_ϑ as soon as $\vartheta \geq \hat{\theta}$. But there is non-negligible support for H_ϑ with ϑ less than, say, 0.65–0.70.

5 Conclusion

Here we showed that there is more to the IM framework than what has been presented in the existing literature. Specifically, the validity property, together with its inherent imprecision implies both performance and probativeness assurances. This is of special interest to the belief function/possibility theory community as it showcases the fundamental importance of its brand of imprecision.

We also identified a connection between IMs and Mayo's severe testing framework. This is beneficial to severe testers, as the IM construction exposed in Sect. 2.2 provides a general recipe for assessing severity in a wide range of modern applications. We also find it attractive that the IM framework has this notion of probativeness built in, as opposed to being an add-on to classical testing. Illustrations in cases beyond the simple, low dimensional problems above will be reported elsewhere, as well as the extension of the notion of probativeness/severity to statistical learning problems.

References

Balch, M.S., Martin, R., Ferson, S.: Satellite conjunction analysis and the false confidence theorem. Proc. Roy. Soc. A **475**(2227), 2018.0565 (2019)

Berger, J.: The case for objective Bayesian analysis. Bayesian Anal. **1**(3), 385–402 (2006)

Dempster, A.P.: The Dempster-Shafer calculus for statisticians. Internat. J. Approx. Reason. **48**(2), 365–377 (2008)

Efron, B.: The Jackknife, the Bootstrap and other Resampling Plans. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 38. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1982)

Fisher, R.A.: Statistical Methods and Scientific Inference, 3rd edn. Hafner Press, New York (1973)

Fraser, D.A.S.: Rejoinder: "Is Bayes posterior just quick and dirty confidence?". Statist. Sci. **26**(3), 329–331 (2011)

Fraser, D.A.S.: Why does statistics have two theories? In: Lin, X., Genest, C., Banks, D.L., Molenberghs, G., Scott, D.W., Wang, J.-L. (eds.) Past, Present, and Future of Statistical Science, chap. 22. Chapman & Hall/CRC Press (2014)

Jeffreys, H.: An invariant form for the prior probability in estimation problems. Proc. Roy. Soc. London Ser. A **186**, 453–461 (1946)

Liu, C., Martin, R.: Inferential models and possibility measures. In: Handbook of Bayesian, Fiducial, and Frequentist Inference arXiv:2008.06874 (2021, to appear)

Martin, R.: Plausibility functions and exact frequentist inference. J. Amer. Statist. Assoc. **110**(512), 1552–1561 (2015)

Martin, R.: On an inferential model construction using generalized associations. J. Statist. Plann. Inference **195**, 105–115 (2018)

Martin, R.: An imprecise-probabilistic characterization of frequentist statistical inference (2021). https://researchers.one/articles/21.01.00002

Martin, R., Liu, C.: Inferential models: a framework for prior-free posterior probabilistic inference. J. Amer. Statist. Assoc. **108**(501), 301–313 (2013)

Martin, R., Liu, C.: Inferential Models: Reasoning with Uncertainty. Monographs on Statistics and Applied Probability, vol. 147. CRC Press, Boca Raton (2015)

Mayo, D.G.: Statistical Inference as Severe Testing. Cambridge University Press, Cambridge (2018)

Reid, N., Cox, D.R.: On some principles of statistical inference. Int. Stat. Rev. **83**(2), 293–308 (2015)

Shackle, G.L.S.: Decision Order and Time in Human Affairs. Cambridge University Press, Cambridge (1961)

Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)