INFERENTIAL MODELS

By Ryan Martin and Chuanhai Liu

Indiana University-Purdue University Indianapolis and Purdue University

Probability is a useful tool for describing uncertainty, so it is natural to strive for a system of statistical inference based on probabilities for or against various hypotheses. But existing probabilistic inference methods struggle to provide a meaningful interpretation of the probabilities across experiments in sufficient generality. In this paper we further develop a promising new approach based on what are called inferential models (IMs). The fundamental idea behind IMs is that there is an unobservable auxiliary variable that itself describes the inherent uncertainty about the parameter of interest, and that posterior probabilistic inference can be accomplished by predicting this unobserved quantity. We describe a simple and intuitive three-step construction of a random set of candidate parameter values, each being consistent with the model, the observed data, and a auxiliary variable prediction. Then prior-free posterior summaries of the available statistical evidence for and against a hypothesis of interest are obtained by calculating the probability that this random set falls completely in and completely out of the hypothesis, respectively. We prove that these IM-based measures of evidence are calibrated in a frequentist sense, showing that IMs give easily-interpretable results both within and across experiments.

1. Introduction. Probability is a useful tool for describing uncertainty, and it is natural to strive for a framework of statistical inference in which the statistical evidence for and against an assertion about the unknown parameter can be summarized by a meaningful probability. The well-known Bayesian framework achieves this goal, but the cost is that a prior distribution for the unknown parameter must first be introduced. Early efforts to get probabilistic inference without prior specification include Fisher's fiducial inference (Zabell 1992) and its variants, Fraser's structural inference (Fraser 1966, 1968) and the Dempster-Shafer theory (Dempster 2008; Shafer 1976). These methods generate probabilities for inference, but they may not be easy to interpret, i.e., these probabilities may not be properly calibrated across users or experiments. So, more recently, efforts have focused on incorporating a frequentist element into the Bayesian framework. In particular, objective

AMS 2000 subject classifications: Primary 62A01; secondary 68T37

Keywords and phrases: Bayesian, credibility, evidence function, fiducial, frequentist, random sets, validity

Bayes analysis with default/reference priors (Berger 2006; Berger, Bernardo and Sun 2009; Bernardo 1979) attempts to construct priors for which certain posterior inferences, such as credible intervals, closely match that of a frequentist. The calibrated Bayesian analysis of Rubin (1984), Dawid (1985), and Little (2010) has similar motivations. But difficulties remain in choosing good reference priors for high-dimensional problems so, despite these efforts, a fully satisfactory framework of objective Bayes inference has yet to emerge.

Recently, Martin, Zhang and Liu (2010) propose a promising new framework for statistical inference, based on what are called *inferential models* (IMs). This new approach is based on the very simple and intuitive idea of first identifying the underlying source of uncertainty—a missing but *predictable* quantity—then making probabilistic inference by predicting the predictable quantity in a statistically accurate way. The result is a method that assigns data-dependent probabilities to each relevant assertion about the parameter of interest, without a prior distribution on the parameter space. The goal of this paper is to further develop this new IM approach into a general framework for prior-free probabilistic inference.

Mathematically, an IM determines a data-dependent mapping that assigns values in $[0, 1]$ to subsets of the parameter space; cf. Definition 1. The numerical values assigned to a subset A are meant to summarize the user's uncertainty about the assertion that the unknown parameter lies in A. In this paper we present a simple, intuitive, and easy to implement three-step process for constructing IMs. The following associate-predict-combine steps, described in more detail in Section 3, shall serve this purpose.

A-STEP. Associate the observed data x and the unknown parameter θ with an unobserved auxiliary variable u—the predictable quantity—to obtain a set $\Theta_x(u)$ of candidate parameter values.

P-STEP. Predict u with a credible predictive random set $\mathcal{S}$.

C-STEP. Combine $\Theta_x(u)$ and the predictive random set $\mathcal{S}$ to obtain the random set $\Theta_x(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_x(u)$. Then, for any $A \subseteq \Theta$ of interest, compute the probability that the random set $\Theta_x(\mathcal{S})$ is a subset of A as a measure of the available statistical evidence in favor of A.

The A-step is meant to emphasize the use of predictable quantities in the statistical modeling step. The motivation is clear: to obtain statistical inference based on usual predictive probabilities, *there must be something being predicted*. The P-step is new and unique to the inferential model framework.

Roughly speaking, the credibility condition in the P-step ensures that the numerical values of the probabilities produced in the C-step are consistent with the usual frequency interpretation. This frequency calibration, together with the dependence on the observed data x, makes the output of the C-step meaningful both within and across experiments.

The remainder of this paper is organized as follows. Statistical modeling for probabilistic inference is discussed in Section 2. We argue that a probabilistic inference framework cannot be built from the sampling model alone, and suggest that the sampling model be characterized by an unobserved but predictable auxiliary variable. We refer to the resulting model as an *association model*. Details of the three-step IM construction are described in Section 3. The three steps are illustrated with a simple Poisson running example. The meaningfulness of the IMs is investigated in Section 4. There a validity theorem is proved which formally establishes the frequency calibration of the IM outputs under mild conditions. This result also allows the user additional flexibility in the P-step; see Theorem 2, an extension of Theorem 1 in Martin, Zhang and Liu (2010). Use of IMs for inference is described in Section 5, and two non-trivial examples are given in Section 6. In the first example, a "default" choice of predictive random set $\mathcal{S}$ determines an IM with output similar to the classical frequentist procedure. In the second example, the predictive random set is chosen carefully based on the hypothesis in question, and simulations show that this "problem-specific" IM-based procedure demonstrably outperforms a standard frequentist procedure. Some concluding remarks are made in Section 7.

2. Sampling models and probabilistic inference.

2.1. *Notation and setup.* If X denotes the observable sample data, then the sampling model is a probability distribution P_θ on the sample space $\mathbb{X}$, indexed by a parameter $\theta \in \Theta$. Here X may consist of a collection of n (possibly vector-valued) data points, in which case both P_θ and $\mathbb{X}$ would depend on n. Then the goal of statistical inference is to reach conclusions about θ based on observing $X = x$. In our context, the parameter θ is unknown but fixed. This is essentially without loss of generality because even when θ is a random quantity, the goal is usually to make inference about its realized value in the experiment at hand.

2.2. *The role of sampling models.* For simplicity, suppose that, for each θ, P_θ is dominated by a fixed σ-finite measure λ, such as Lebesgue or counting measure, so that P_θ has a density $p_\theta = d\mathsf{P}_\theta/d\lambda$. In such cases, it is common to summarize the information in the observed data x about θ with

the likelihood function $L_x(\theta) = p_\theta(x)$, now treated as a function of θ for fixed x. Fisher (1922, 1925, 1934) recognized the importance of $L_x(\theta)$ and emphasized that it is not a probability distribution for θ. Rather, its interpretation is postdictive in the sense that it can be used only to compare different explanations of the observed event "$X = x$." The key observation is that while $L_x(\theta)$ is a probability density in x, allowing θ to vary changes the underlying probability space, so, mathematically, the usual laws of probability do not hold and, logically, these probabilities are not predictive in nature. Hence, no system of probabilistic inference—where the output must have a predictive probabilistic interpretation—can be developed based on the likelihood alone. This important point is further emphasized below.

PRINCIPLE. *The sampling model alone is insufficient for probabilistic inference about unknown parameters. Only if unobserved but predictable quantities are associated with the observed data and unknown parameters can predictive probabilistic inference be achieved.*

This principle formally states the well-known fact that the usual frequentist procedures, such as hypothesis tests and confidence intervals, which are built from the sampling model alone, are not probabilistic in nature. But it goes even further, stating what needs to be done to introduce a probabilistic interpretation. Even the basic idea of the familiar Bayesian inference is consistent with the foregoing principle. Indeed, in the fully Bayesian approach, θ plays the role of the predictable quantity—there is no unknown parameter—so the principle has nothing to say. While probabilities are available for inference on θ in the Bayesian setting, these probabilities may not be easy to interpret. In a subjective Bayes setting, for example, where an empirically valid distribution is available for everything, the posterior probabilities are calibrated across samples. But when the postulated prior for θ is not empirically valid, as is the often the case with default priors, the resulting posterior probabilities are not calibrated in general (Ermini Leaf, Hui and Liu 2009), which makes interpretation challenging.

For valid probabilistic inference, we propose to build a full statistical model for observed data and unknown quantities of interest. We call this an *association model* to distinguish it from usual sampling models. Thus, an association model associates three quantities: the observable X, the unknown θ, and a predictable auxiliary variable U. Then probabilistic inference can be achieved by predicting the unobserved auxiliary variable.

2.3. *Association models.* In this paper the sampling model for the observable X is that induced by an auxiliary (a-)variable U, for given θ. Let

$\mathbb{U}$ be a more-or-less arbitrary auxiliary space, on which is defined a probability measure μ. In applications, $\mathbb{U}$ can often be a unit hyper-cube and μ Lebesgue measure. The sampling model P_θ shall be determined by the following algorithm:

$$\text{sample } U \sim \mu \text{ and set } X = a(U, \theta), \tag{2.1}$$

for an appropriate mapping $a : \mathbb{U} \times \Theta \to \mathbb{X}$. The key is the association of the observable X, the unknown θ, and the a-variable U through the relation $X = a(U, \theta)$. This particular formulation of the sampling model is not a restriction. In fact, the two-step construction of the observable X in (2.1) is generally consistent with scientific understanding of the underlying process under investigation; the signal plus noise model is a good example. Model (2.1) is also quite familiar to statisticians in the context of random variable generation or, more generally, in the context of model building. But see Section 2.4 below for discussion of the non-uniqueness issue.

Poisson Example. As a simple running example, consider the problem of inference on the mean θ of a Poisson population based on a single observation X. The association model for X, given θ, may be written as

$$F_\theta(X - 1) \leq U < F_\theta(X), \quad U \sim \mathsf{Unif}(0, 1), \tag{2.2}$$

where F_θ denotes the $\mathsf{Pois}(\theta)$ distribution function.

2.4. *The non-uniqueness issue.* It should not be surprising that there are many association models for a given sampling model. In fact, for a given sampling model P_θ, there are as many association models as there are triplets $(\mathbb{U}, \mu, a_\theta)$, with $a_\theta(\cdot) = a(\cdot, \theta)$, such that P_θ equals the push-forward measure μa_θ^{-1}. The non-uniqueness arises from the fact that it is generally only possible to elicit—from experts, past experience, or exploratory data analysis—a model for the observable X; models for latent variables are more difficult, if not impossible, to pin down. But given the limitations of the sampling model (cf. Section 2.2), this non-uniqueness is apparently the price the statistician must pay for probabilistic inference. This is especially true for complex problems such as multiple testing and variable selection in regression.

An interesting observation is that the choice of association model can be investigated through transformations of the a-variables. To see this, consider two association models $(\mathbb{U}_1, \mu_1, a_{\theta,1})$ and $(\mathbb{U}_2, \mu_2, a_{\theta,2})$ for a given sampling model. Suppose further that the relationship $a_{\theta,1}(u_1) = a_{\theta,2}(u_2)$ defines, for each θ, a one-to-one mapping from $\mathbb{U}_1$ to $\mathbb{U}_2$. Since $a_{1,\theta}(U_1)$ and $a_{\theta,2}(U_2)$ have the same distribution, the measure μ_2 is uniquely defined by μ_1 and

the maps $a_{\theta,1}$ and $a_{\theta,2}$. In other words, the choice of association model is equivalent to the choice of a-variable parametrization, suggesting that any two association models are related via some one-to-one transformations of a-variables, namely $U_2 = \varphi_\theta(U_1)$, possibly depending on the unknown θ. This implies that the non-uniqueness of association models allows for simple alternative ways of doing a-variable transformation for constructing simple and efficient predictive random sets. For example, some association models allow for simple a-variable dimension reduction (Section 7). We will encounter these a-variable transformations again in Sections 4 and 6.

3. Inferential models: a three-step construction. A simple and general three-step representation of an IM is given in Section 1. This section describes each of these three steps in greater detail, with illustrations in the simple Poisson example.

3.1. *Association step.* The association model (2.1) plays two distinct roles. Before the experiment, the association model characterizes the predictive probabilities of the observable X. But once $X = x$ is observed, the role of the association model changes. The key idea is that the observed x and the unknown θ must satisfy

$$x = a(u^\star, \theta) \tag{3.1}$$

for some fixed $u^\star \in \mathbb{U}$. The quantity $u^\star$ in (3.1) is unobserved, but there is information available about the nature of this quantity; in particular, we know that $u^\star$ is the realization of a sample $U \sim \mu$.

Of course, the value of $u^\star$ can never be known, *but if it were*, the inference problem would be simple—just solve the equation $x = a(u^\star, \theta)$ for θ. More generally, one could construct the set of solutions $\Theta_x(u^\star)$, where

$$\Theta_x(u) = \{\theta : x = a(u, \theta)\}, \quad u \in \mathbb{U}. \tag{3.2}$$

For continuous-data problems, $\Theta_x(u)$ is typically a singleton for each u; for other problems, it could be a set. In either case, $\Theta_x(u^\star)$ represents the best possible inference: *the true θ is guaranteed to be in $\Theta_x(u^\star)$*.

Poisson Example (cont). Integration-by-parts reveals that the $\mathsf{Pois}(\theta)$ distribution function satisfies $F_\theta(x) = 1 - G_{x+1}(\theta)$, where G_a is a $\mathsf{Gamma}(a, 1)$ distribution function. Therefore, from (2.2), we get

$$G_{x+1}(\theta) \leq \bar{u} < G_x(\theta), \quad \bar{u} = 1 - u. \tag{3.3}$$

Inverting (3.3) produces the set

$$\Theta_x(u) = (G_x^{-1}(\bar{u}), G_{x+1}^{-1}(\bar{u})]. \tag{3.4}$$

If $u^\star$ was available, then $\Theta_x(u^\star)$ would provide the best possible inference in the sense that the true value of θ is guaranteed to sit in this interval. But there is no additional information available to help pin down further the exact location of θ in $\Theta_x(u^\star)$.

3.2. *Prediction step.* The above discussion highlights the importance of the a-variable for inference. Therefore, it is only natural that the inference problem should focus on accurately predicting the unobserved $u^\star$. To predict $u^\star$ with desired certainty, we employ a random set, called a *predictive random set*, or PRS for short. Let $u \mapsto \mathcal{S}_u$ be a mapping from $\mathbb{U}$ to a collection of connected subsets of $\mathbb{U}$, with the constraint that $\mathcal{S}_u \ni u$ for each u. One example of such a mapping $\mathcal{S}$ is given below, and more details about the choice of PRS are given in Section 4. Intuitively, the PRS $\mathcal{S}_U$, for $U \sim \mu$, encodes our certainty about predicting the unobserved $u^\star$.

Poisson Example (cont). In this example we predict the unobserved $\bar{u}^\star = 1 - u^\star$ in (3.3) with a PRS defined by the set-valued mapping

$$\mathcal{S}_u = \{\tilde{u} : |\tilde{u} - 0.5| \leq |u - 0.5|\}, \quad u \in [0, 1]. \tag{3.5}$$

This PRS is can be shown to be credible, as required in the P-step description in Section 1; see Section 4 for the formal definition of credibility and the relevant theorem.

3.3. *Combination step.* To transfer the available information about $u^\star$ to the θ-space, our last step is to combine the information in the association model, the observed x, and the uncertainty about predicting $u^\star$ encoded in the PRS $\mathcal{S}$. Combining $\mathcal{S}_u$ with the association model amounts to expanding the set of solutions $\Theta_x(u)$ in (3.2) for a given u to account for all candidates for $u^\star$ in the set $\mathcal{S}_u$. That is, we now consider

$$\Theta_x(\mathcal{S}_u) = \bigcup_{\tilde{u} \in \mathcal{S}_u} \Theta_x(\tilde{u}), \quad u \in \mathbb{U}, \tag{3.6}$$

which is larger than $\Theta_x(u)$ since $\mathcal{S}_u \ni u$. Again, this expansion of the set $\Theta_x(u)$ of candidate θ values reflects our uncertainty about the guess u for $u^\star$. Intuitively, the set $\Theta_x(\mathcal{S}_u)$ contains those values of θ which are consistent with the observed data and sampling model for at least one candidate $\tilde{u} \in \mathcal{S}_u$.

Now consider an assertion A about the parameter of interest θ. Mathematically, A is just a subset of Θ, but it acts much like a hypothesis in the context of classical statistics. To summarize the statistical evidence in x for the assertion A, we calculate the probabilities that the random set $\Theta_x(\mathcal{S}_U)$, as a function of $U \sim \mu$, is a subset of A. That is

$$\underline{\mathsf{e}}_{x,\mathcal{S}}(A) = \mu\{u : \Theta_x(\mathcal{S}_u) \subseteq A\}. \tag{3.7}$$

We refer to $\underline{\mathsf{e}}_{x,\mathcal{S}}(A)$ as the *lower evidence function* at A. To make decisions about A, it is necessary to know $\underline{\mathsf{e}}_{x,\mathcal{S}}(A^c)$ as well. However, in what follows, it will be more convenient to work with

$$\overline{\mathsf{e}}_{x,\mathcal{S}}(A) = 1 - \underline{\mathsf{e}}_{x,\mathcal{S}}(A^c) = \mu\{u : \Theta_x(\mathcal{S}_u) \not\subseteq A^c\}, \tag{3.8}$$

the *upper evidence function* at A. Therefore, we summarize the statistical evidence about A with the pair $\underline{\mathsf{e}}_{x,\mathcal{S}}(A)$ and $\overline{\mathsf{e}}_{x,\mathcal{S}}(A)$. Intuitively, large values of $\underline{\mathsf{e}}_{x,\mathcal{S}}(A)$ suggests that the data x gives strong support to the assertion, while large values $\overline{\mathsf{e}}_{x,\mathcal{S}}(A)$ suggests that A is highly *plausible* given x.

While the lower and upper evidence functions $\underline{\mathsf{e}}_{x,\mathcal{S}}$ and $\overline{\mathsf{e}}_{x,\mathcal{S}}$ produce bona fide predictive probabilities, there are two important points to keep in mind regarding their interpretation.

- $\underline{\mathsf{e}}_{x,\mathcal{S}}(A)$ and $\overline{\mathsf{e}}_{x,\mathcal{S}}(A)$ are probabilities for the random set $\Theta_x(\mathcal{S}_U)$ conditional on observed $X = x$, similar to Bayesian posterior or fiducial probabilities. But note, for example, that $\underline{\mathsf{e}}_{x,\mathcal{S}}(A)$ does not represent probability that A is true. In fact, no attempt is made to introduce a measure on the parameter space Θ, so no such probability exists.
- $\underline{\mathsf{e}}_{x,\mathcal{S}}$ and $\overline{\mathsf{e}}_{x,\mathcal{S}}$ do not satisfy the usual rules of probability. For example, $\underline{\mathsf{e}}_{x,\mathcal{S}}(A) + \underline{\mathsf{e}}_{x,\mathcal{S}}(A^c) \leq 1$, with equality if and only if $\Theta_x(\mathcal{S}_u)$ is a singleton for μ-almost all u. But from this subadditivity property and the definition of $\overline{\mathsf{e}}_{x,\mathcal{S}}$ in (3.8), we have that

  $$\underline{\mathsf{e}}_{x,\mathcal{S}}(A) \leq \overline{\mathsf{e}}_{x,\mathcal{S}}(A),$$

  which explains the names lower and upper evidence functions. With this in mind, one could view the interval $[\underline{\mathsf{e}}_{x,\mathcal{S}}(A), \overline{\mathsf{e}}_{x,\mathcal{S}}(A)]$ as a sort of range of Bayesian posterior probabilities of A over a class of priors (Wasserman 1990), but we will not take this view.

It should also be mentioned that an adjustment to the evidence functions is required if $\mu\{u : \Theta_x(\mathcal{S}_u) = \varnothing\} > 0$. This will not occur in the examples considered in this paper, but it can happen in general. In such cases, the lower evidence function, for example, must be rewritten as

$$\underline{\mathsf{e}}_{x,\mathcal{S}}(A) = \mu\{u : \Theta_x(\mathcal{S}_u) \subseteq A, \Theta_x(\mathcal{S}_u) \neq \varnothing\} / \mu\{u : \Theta_x(\mathcal{S}_u) \neq \varnothing\},$$

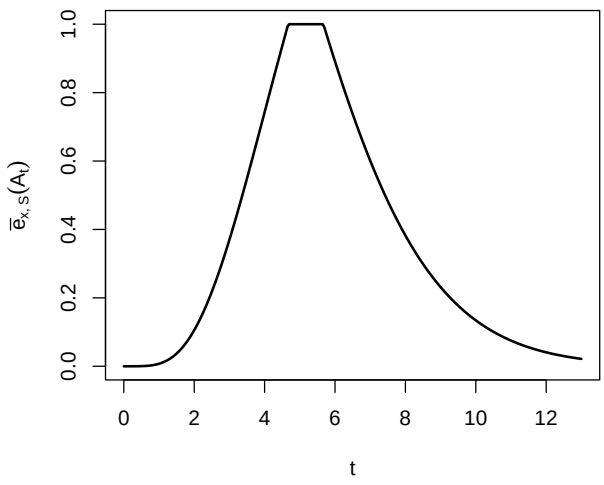

FIG 1. *Plot of the upper evidence function* $\bar{e}_{x,\mathcal{S}}(A_t)$ *in* (3.10)*, with* $A_t = \{t\}$*, as a function of* t *in the Poisson example. We assume that* $X = 5$ *is observed.*

the conditional μ-probability that $\Theta_x(\mathcal{S}_U) \subseteq A$, given $\Theta_x(\mathcal{S}_u)$ is non-empty.

POISSON EXAMPLE (cont). For the PRS mapping $\mathcal{S}_u$ in (3.5), the set $\Theta_x(\mathcal{S}_u)$ of candidate θ values, for the given u, is

$$\Theta_x(\mathcal{S}_u) = \Big(G_x^{-1}(0.5 - |u - 0.5|), G_{x+1}^{-1}(0.5 + |u - 0.5|)\Big]. \tag{3.9}$$

For a singleton assertion $A_t = \{t\}$, it is easy to see that the lower evidence function is zero. But the upper evidence function is

$$\bar{e}_{x,S}(A_t) = 1 - \max\{1 - 2G_x(t), 0\} - \max\{2G_{x+1}(t) - 1, 0\}. \tag{3.10}$$

A plot of $\bar{e}_{5,\mathcal{S}}(A_t)$ as a function of t is shown in Figure 1. The plateau indicates that no θ values in a small neighborhood of the observed $x = 5$ can be ruled out. Furthermore, the graph shows that θ's in an interval around $x = 5$, say $[3, 9]$, are relatively plausible. More details on the use of the evidence functions for inference are given in Section 5.

In the simple Poisson illustration above, evaluation of the evidence functions is straightforward. The calculations are essentially the same for any single-parameter problem for singleton assertions. More generally, if the a-variable is high-dimensional, like in Section 6.2, then Monte Carlo methods can be used. But once the evidence functions are available, inference can proceed along the lines presented in Section 5.

3.4. *Summary.* The three-step construction of IMs described in Sections 3.1–3.3 is simple and easy to implement. But it is beneficial here to briefly summarize the result of this construction. The A-step specifies the particular association model representation of the sampling model, thus determining the a-variable to be predicted. To predict the unobserved a-variable, we incorporate a PRS $\mathcal{S}$ in the P-step. The intuition here is that predicting this unobserved $u^\star$ with a draw $U \sim \mu$ (or the corresponding singleton set $\{U\}$) is insufficient in the sense that, in many problems, $\Theta_x(\{U\})$ will contain the true θ with μ-probability 0. By expanding the draw U to a set $\mathcal{S}_U$ containing U, give ourselves some opportunity to catch the real θ. The other extreme $\mathcal{S}_u \equiv \mathbb{U}$ is also not useful; the balance between the size of $\mathcal{S}$ and properties of the resulting IM are described in Section 4. The C-step proceeds by combining the information about the relation between (X, θ, U) in the association model, the observed data $X = x$, and the information about the a-variable in $\mathcal{S}_U$ to find a random set $\Theta_x(\mathcal{S}_U)$ of candidate parameter values. For a given assertion A about θ, we summarize the statistical evidence in x for A with $\underline{\mathsf{e}}_{x,\mathcal{S}}(A)$, which is the μ-probability that $\Theta_x(\mathcal{S}_U)$ is a subset of A. We couple this with $\overline{\mathsf{e}}_{x,\mathcal{S}}(A) = 1 - \underline{\mathsf{e}}_{x,\mathcal{S}}(A^c)$, the upper evidence function, and produce the pair $(\underline{\mathsf{e}}_{x,\mathcal{S}}(A), \overline{\mathsf{e}}_{x,\mathcal{S}}(A))$ as a complete summary of our uncertainty about A.

Armed with the necessary concepts and notation, we can now give a formal definition of an inferential model.

DEFINITION 1. For a given sampling model P_θ on $\mathbb{X}$, $\langle(\mathbb{U}, \mu, a), \mathcal{S}\rangle$ defines an *inferential model* (IM) if the triplet $(\mathbb{U}, \mu, a)$ defines an association model as in Section 2.3 and if $\mathcal{S}$ is a PRS as in Section 3.2. Then, given the observed $X = x$, the three-step process in Sections 3.1–3.3 describes how to construct evidence functions $(\underline{\mathsf{e}}, \overline{\mathsf{e}})_{x,\mathcal{S}}$ from the IM for probabilistic inference.

4. Theoretical validity of IMs. In the previous section we described a simple three-step procedure for constructing an IM, and it is clear that, together, the evidence functions $\underline{\mathsf{e}}_{x,\mathcal{S}}(A)$ and $\overline{\mathsf{e}}_{x,\mathcal{S}}(A)$ provide a meaningful summary of evidence favoring A for the given observed data $X = x$. In this section we show that $\underline{\mathsf{e}}_{X,\mathcal{S}}(A)$ and $\overline{\mathsf{e}}_{X,\mathcal{S}}(A)$ are also meaningful as functions of the random variable $X \sim \mathsf{P}_\theta$ for a fixed assertion A. For example, we show that $\underline{\mathsf{e}}_{X,\mathcal{S}}(A)$ is frequency-calibrated in the following sense: if $\theta \notin A$, then $\mathsf{P}_\theta\{\underline{\mathsf{e}}_{X,\mathcal{S}}(A) \geq 1 - \alpha\} \leq \alpha$ for each $\alpha \in [0, 1]$. In other words, the amount of evidence in favor of a false A can be large with only small probability. Mathematically, this property means that the lower evidence function has a predictive probabilistic interpretation, suggesting that the probability in (3.7) is, indeed, appropriately scaled for objective scientific inference. A

similar property holds for $\bar{\mathsf{e}}_{X,\mathcal{S}}(A)$. We refer to this frequency-calibration property *validity* of the IM; cf. Definition 3. See Bernardo (1979), Rubin (1984) and Dawid (1985) for similar considerations leading to investigations of frequency-calibration and of objective priors for Bayesian inference.

Towards establishing this general validity property for IMs, we review some of the concepts from Zhang and Liu (2010) and Martin, Zhang and Liu (2010). First we need a few definitions. Start with the mapping

$$Q_{\mathcal{S}}(u) = \mu\{\tilde{u} : \mathcal{S}_{\tilde{u}} \not\ni u\}, \quad u \in \mathbb{U}, \tag{4.1}$$

which gives the probability that the PRS $\mathcal{S}_U$ misses the specified target u. Ideally, the map $\mathcal{S}$ will be such that the random variable $Q_{\mathcal{S}}(U)$, a function of $U \sim \mu$, will be probabilistically small.

DEFINITION 2 (Credibility). A PRS $\mathcal{S}_U$ is credible at level $\alpha \in (0,1)$ for predicting the unobserved auxiliary variable if

$$\mu\{u : Q_{\mathcal{S}}(u) \geq 1-\alpha\} \leq \alpha. \tag{4.2}$$

In words, credibility implies that the probability that $\mathcal{S}_U$ misses a target u is large for only a small μ-proportion of possible u values. The PRS $\mathcal{S}_u$ in (3.5) is credible. Indeed, it is easy to check that, in this case,

$$Q_{\mathcal{S}}(u) = \max\{1-2u, 0\} + \max\{2u-1, 0\} = |2u-1|.$$

Therefore, if $U \sim \mathsf{Unif}(0,1)$ then $Q_{\mathcal{S}}(U) \sim \mathsf{Unif}(0,1)$ too, and credibility follows immediately. More generally, we have the following recipe for constructing credible PRSs.

THEOREM 1. *Suppose the measure μ on $\mathbb{U}$ is non-atomic, and let f be a continuous real-valued function on $\mathbb{U}$. Then the PRS $\mathcal{S}$ given by*

$$\mathcal{S}_u = \{\tilde{u} : f(\tilde{u}) \leq f(u)\}, \quad u \in \mathbb{U},$$

is credible in the sense of Definition 2.

PROOF. Let F denote the distribution function of $f(U)$ for $U \sim \mu$. Then

$$Q_{\mathcal{S}}(u) = \mu\{\tilde{u} : \mathcal{S}_{\tilde{u}} \not\ni u\} = \mu\{\tilde{u} : f(\tilde{u}) < f(u)\} = F\big(f(u)\big).$$

If $U \sim \mu$, then $Q_{\mathcal{S}}(U) = F\big(f(U)\big) \sim \mathsf{Unif}(0,1)$ and credibility follows. □

It turns out that credibility of the underlying PRS is essentially all that is needed to prove the meaningfulness of the corresponding IM. Here meaningfulness refers to a calibration property of the corresponding evidence function, which is the property we call validity.

DEFINITION 3 (Validity). Suppose $X \sim \mathsf{P}_\theta$ and let A be an assertion of interest. Then the IM $\langle(\mathbb{U}, \mu, a), \mathcal{S}\rangle$ is *valid for* A if, for each $\alpha \in (0,1)$, the corresponding lower evidence function satisfies

$$\sup_{\theta \notin A} \mathsf{P}_\theta\{\underline{\mathsf{e}}_{X,\mathcal{S}}(A) \geq 1-\alpha\} \leq \alpha \tag{4.3}$$

The IM is called *valid* if it is valid for all A.

From the relationship between $\underline{\mathsf{e}}_{x,\mathcal{S}}$ and $\bar{\mathsf{e}}_{x,\mathcal{S}}$ in (3.8), it is easy to check that the validity property can be equivalently stated in terms of the upper evidence function. That is, the IM is valid if, for any assertion A,

$$\sup_{\theta \in A} \mathsf{P}_\theta\{\bar{\mathsf{e}}_{X,\mathcal{S}}(A) \leq \alpha\} \leq \alpha. \tag{4.4}$$

This representation is occasionally more convenient than (4.3).

THEOREM 2. *Consider a one-to-one transformation $v = \varphi_\theta(u)$ such that the push-forward measure $\mu_\varphi := \mu\varphi_\theta^{-1}$ on $\mathbb{V} = \varphi_\theta(\mathbb{U})$ does not depend on θ. Suppose $\mathcal{S}$ is credible for predicting $v^\star = \varphi_\theta(u^\star)$, and $\Theta_x(\mathcal{S}_V) \neq \varnothing$ with μ-probability 1 for all x. Then for any assertion $A \subseteq \Theta$,*

$$\sup_{\theta \notin A} \mathsf{P}_\theta\{\underline{\mathsf{e}}_{X,\mathcal{S},\varphi}(A) \geq 1-\alpha\} \leq \alpha, \tag{4.5}$$

where $\underline{\mathsf{e}}_{x,\mathcal{S},\varphi}(A) = \mu_\varphi\{v \in \mathbb{V} : \Theta_x(\mathcal{S}_v) \subseteq A\}$. In other words, the IM, in terms of transformed a-variables, is valid.

PROOF. Take any $\theta \notin A$. Since $A \subseteq \{\theta\}^c$, it follows from monotonicity of the lower evidence function that

$$\begin{aligned} \underline{\mathsf{e}}_{x,\mathcal{S},\varphi}(A) &\leq \underline{\mathsf{e}}_{x,\mathcal{S},\varphi}(\{\theta\}^c) \\ &= \mu_\varphi\{v : \Theta_x(\mathcal{S}_v) \not\ni \theta\} = \mu_\varphi\{v : \mathcal{S}_v \not\ni v^\star\}. \end{aligned}$$

Credibility of $\mathcal{S}$ implies that the right-hand side, as a function of $V^\star \sim \mu_\varphi$, is stochastically smaller than $\mathsf{Unif}(0,1)$. This, in turn, implies the same of $\underline{\mathsf{e}}_{X,\mathcal{S},\varphi}(A)$ as a function of $X \sim \mathsf{P}_\theta$. Therefore,

$$\mathsf{P}_\theta\{\underline{\mathsf{e}}_{X,\mathcal{S},\varphi}(A) \geq 1-\alpha\} \leq \mathsf{P}\{\mathsf{Unif}(0,1) \geq 1-\alpha\},$$

and the result follows. □

Theorem 2 above, which establishes the validity of the IM framework, extends Theorem 3.1 of Zhang and Liu (2010) in an important direction. In particular, it allows for a change of a-variable which, in addition to being helpful in certain cases, shows that the validity of IMs is independent of the parametrization of the association model. This change of a-variable is also allowed to depend on the unknown parameter, a particularly useful technique in some examples, including those presented in Section 6.

The condition that the set $\Theta_x(\mathcal{S}_V)$ be non-empty with μ_φ-probability 1 for all x is critical and cannot be relaxed. This is easy to arrange for problems where a given x imposes no constraint on the possible θ-values. But there are constrained parameter problems where this condition does not hold; see Ermini Leaf and Liu (2010) for an IM analysis in this more general case.

We conclude this section with a remark related to constructions of PRSs. Validity is a desirable property, but cannot be one's only consideration. This is analogous to the frequentist hypothesis testing problem, with validity playing the role of the significance level. Take, for example, the extreme case where $\mathcal{S}_u \equiv \mathbb{U}$, which implies complete uncertainty about the prediction of $u^\star$. In this case, validity is obvious, just like a test that never rejects is guaranteed to control the significance level at any level α. But we know that such a naive testing rule will perform poorly in terms of power, and the IM with too large of PRS will suffer similarly. For this, a secondary condition called *efficiency*, analogous to power in hypothesis testing, is required. More discussion on efficiency is given in Section 7.

5. Using IMs for inference. Given our interpretation, a natural subjectivist approach would be to examine the relative magnitudes of the evidence functions for several competing assertions. There are similarities here to Jeffreys' Bayes factors and Fisher's p-values, but we believe that the interpretation of evidence functions is easier. For example, $\underline{\mathsf{e}}_{x,\mathcal{S}}(A)$ directly measures the evidence in favor of A, given data $X = x$, while Fisher's p-value measures evidence in favor of A indirectly through the chance of the event $X = x$, given $\theta \in A$.

Poisson Example (cont). Suppose $X = 5$ is observed, and the assertion of interest is $A = (0, 2]$, i.e., the Poisson mean θ is no more than 2. With the PRS $\mathcal{S}_u$ in (3.5) and the corresponding $\Theta_x(\mathcal{S}_u)$ in (3.9), it is easy to check that the lower and upper evidence functions are

$$\begin{aligned}\underline{\mathsf{e}}_{5,\mathcal{S}}(A) &= \max\{2G_6(2) - 1, 0\} = 0 \\ \overline{\mathsf{e}}_{5,\mathcal{S}}(A) &= 1 - \max\{1 - 2G_5(2), 0\} = 2G_5(2) = 0.105.\end{aligned}$$

The evidence in favor of A is not overwhelming in this example, so it would not be unreasonable to reject the assertion that θ is no more than 2. For comparison, note that Fisher's p-value for testing $H_0 : \theta \in A$ based on $X = 5$ is 0.0523, exactly half of $\bar{\mathsf{e}}_{x,\mathcal{S}}(A)$ in this case, which has the force of a logical disjunction: either A is false or it is true and a rare event occured. So while both procedures make the same decision, the reasoning is different.

In addition to providing problem-specific measures of certainty about various assertions of interest, the evidence functions can easily be used to design classical inference tools that satisfy the usual frequentist properties. First consider the null hypothesis $H_0 : \theta \in A$. Then an IM-based counterpart to a frequentist testing rule is of the following form:

$$\text{Reject } H_0 \text{ if } \bar{\mathsf{e}}_{x,\mathcal{S}}(A) \leq t, \text{ for a specified } t \in (0,1). \tag{5.1}$$

According to (4.4) and Theorem 2, if the PRS $\mathcal{S}$ is credible, then the probability of a Type I error for such a rejection rule is

$$\sup_{\theta \in A} \mathsf{P}_\theta\{\bar{\mathsf{e}}_{x,\mathcal{S}}(A) \leq t\} \leq t.$$

So in order for the test (5.1) to control the probability of a Type I error at a specified $\alpha \in (0,1)$, one should reject H_0 if the upper evidence is $\leq \alpha$.

Next consider the class of singleton assertions $A_t = \{t\}$, $t \in \Theta$. As a counterpart to a frequentist confidence region, define the *plausibility region*

$$\Pi_x(\alpha) = \{t : \bar{\mathsf{e}}_{x,\mathcal{S}}(A_t) > \alpha\}. \tag{5.2}$$

Now the coverage probability of the plausibility region (5.2) is

$$\begin{aligned} \mathsf{P}_\theta\{\Pi_x(\alpha) \ni \theta\} &= \mathsf{P}_\theta\{\bar{\mathsf{e}}_{X,\mathcal{S}}(A_\theta) > \alpha\} \\ &= 1 - \mathsf{P}_\theta\{\bar{\mathsf{e}}_{X,\mathcal{S}}(A_\theta) \leq \alpha\} \geq 1 - \alpha, \end{aligned}$$

where the last inequality follows from Theorem 2. Therefore, this plausibility region has at least the nominal coverage probability.

Poisson Example (cont). For the PRS determined by $\mathcal{S}$ in (3.5) and the sequence of assertions A_t above, the upper evidence function $\bar{\mathsf{e}}_{x,\mathcal{S}}(A_t)$ is displayed in (3.10). Then, for observed $X = 5$, a nominal 90% plausibility interval for θ is given by

$$\bar{\mathsf{e}}_{5,\mathcal{S}}(A_t) > 0.10 \iff t \in \Pi_5(0.10) = (1.97, 10.51).$$

It is worth pointing out that, if the sample size is increased, then the resulting inference will, as expected, become more definitive. For example, the plausibility interval for the Poisson mean above would be narrower if a sample of size n produced a mean $\overline{X} = 5$. But the required conditioning argument needed to prove this claim is beyond the scope of this paper; see Section 7 for a preview and references.

6. Two non-trivial examples. To illustrate the IM approach, we consider two non-trivial examples. In each example below, there is a sort of *marginalization* required, and this is accomplished by using a cylinder PRS that effectively ignores parts of the a-variable to be predicted. More sophisticated IM-based marginalization techniques are available, but these examples are meant to be as simple as possible.

6.1. *Inference on a standardized mean.* Suppose that $X_1, \ldots, X_n$ are independent observations from a $\mathsf{N}(\xi, \sigma^2)$ population. The goal is to make inference on $\psi = \xi/\sigma$, the standardized mean. Begin with a reduction of the observed data $x = (x_1, \ldots, x_n)$ to the sufficient statistics for $\theta = (\xi, \sigma^2)$, namely $(\overline{x}, s^2)$. Some formal IM-based justification for this reduction is available but we will not discuss this here; see Section 7.

For the A-step, here we take the association model to be

$$\overline{x} = \xi + n^{-1/2}\sigma u_1 \quad \text{and} \quad s = \sigma u_2, \tag{6.1}$$

where the space of a-variables $u = (u_1, u_2)$ is equipped with the measure

$$\mu = \mathsf{N}(0,1) \times \sqrt{\mathsf{ChiSq}_{n-1}/(n-1)}.$$

A bit of algebra reveals that

$$\frac{n^{1/2}\overline{x}}{s} = \frac{n^{1/2}\psi + u_1}{u_2} \quad \text{and} \quad s = \sigma u_2.$$

For $\theta = (\psi, \sigma)$, make a change of a-variable $v = \varphi_\theta(u)$, given by

$$v_1 = F_\psi\Big(\frac{n^{1/2}\psi + u_1}{u_2}\Big) \quad \text{and} \quad v_2 = \frac{\exp\{u_2\}}{1+\exp\{u_2\}},$$

where $F_\psi(\cdot)$ is the distribution function for a non-central t_{n-1} distribution with non-centrality parameter $n^{1/2}\psi$. Note that the full generality of a change-of-variables that depends on the unknown parameter in Theorem 2 is needed here. Then the transformed association model is

$$\frac{n^{1/2}\overline{x}}{s} = F_\psi^{-1}(v_1) \quad \text{and} \quad s = \sigma \log\frac{v_2}{1-v_2},$$

and the measure μ_φ on the space of $v = (v_1, v_2)$ has a $\mathsf{Unif}(0,1)$ marginal on the v_1-space; the distribution on v_1-slices of the v_2 space can be worked out, but it is not needed in what follows. For the P-step, we predict $v^\star = \varphi_\theta(u^\star)$ with a rectangle PRS $\mathcal{S}$ given by

$$\mathcal{S}_v = \{\tilde{v}_1 : |\tilde{v}_1 - 0.5| \leq |v_1 - 0.5|\} \times [0,1], \quad v = (v_1, v_2). \tag{6.2}$$

That this PRS is credible follows Theorem 1. Using a PRS that spans the entire v_2-space for each v has the effect of "integrating out" the nuisance parameter σ. For the PRS $\mathcal{S}$ in (6.2), if $z = n^{1/2}\overline{x}/s$, then the C-step gives the following set of candidate (ψ, σ) pairs:

$$\begin{aligned} \Theta_x(\mathcal{S}_v) &= \Psi_x(\mathcal{S}_v) \times \Sigma_x(\mathcal{S}_v) \\ &= \{\psi : |F_\psi(z) - 0.5| \geq |v_1 - 0.5|\} \times \{\sigma : \sigma > 0\}. \end{aligned} \tag{6.3}$$

For assertions $A_t = \{\psi = t, \sigma > 0\}$, the lower evidence function is zero, but the upper evidence function is given by

$$\begin{aligned} \overline{\mathsf{e}}_{x,\mathcal{S}}(A_t) &= \mu_\varphi\{v : \Theta_x(\mathcal{S}_v) \cap A_t \neq \varnothing\} \\ &= \mu_\varphi\{v : \Psi_x(\mathcal{S}_v) \cap \{t\} \neq \varnothing\} \\ &= \mu_\varphi\{v : \Psi_x(\mathcal{S}_v) \ni t\} \\ &= 1 - |1 - 2F_t(z)|. \end{aligned}$$

In this case, the $100(1-\alpha)\%$ plausibility interval $\Pi_x(\alpha)$ for ψ is obtained by inverting the inequality $1 - |2F_t(z) - 1| > \alpha$, i.e.,

$$\Pi_x(\alpha) = \{t : \alpha/2 < F_t(z) < 1 - \alpha/2\}.$$

The reader should note that this is exactly the same as the standard frequentist confidence interval based on the sampling distribution of the standardized sample mean. The frequentist approach, however, relies on an informal marginalization. On the other hand, the IM approach above formally shows exactly how σ is ignored and, in fact, there are available more sophisticated IM analyses that handle unknown σ more gracefully; these details are beyond our present scope, but see Section 7.

6.2. *Inference on a Poisson process.* In this section we take a second look at an example presented in Martin, Zhang and Liu (2010) in which we investigate whether an observed Poisson process is homogeneous. This example shares a number of similarities with the now very common high-dimensional multiple testing problems; see, for example, Efron (2010) and the references therein. Our focus here is to simplify the presentation in

Martin, Zhang and Liu (2010) and to emphasize the three-step construction of an IM for probabilistic inference.

A process is to be monitored over a pre-specified interval of time, say $[0, \tau]$. Suppose, during that period of time, we observe n events at times $0 < T_1 < \cdots < T_n \leq \tau$; the $(n+1)$st event, which takes place at time $T_{n+1} > \tau$, is unobserved. We model the time between events $X_i = T_i - T_{i-1}$, $i = 1, \ldots, n$, with $T_0 \equiv 0$, as independent exponential random variables, $X_i \sim \mathsf{Exp}(\theta_i)$, with unknown rates $\theta_1, \ldots, \theta_n$. The goal is to produce a plausibility measure of the assertion

$$A = \{\text{process is homogeneous}\} = \{\theta_1 = \cdots = \theta_n\}. \tag{6.4}$$

Start, in the A-step, with the following simple association model:

$$x_i = u_i/\theta_i, \quad i = 1, \ldots, n.$$

In this case, $\mathbb{U} = (0, \infty)^n$ and μ is the n-fold $\mathsf{Exp}(1)$ product measure. Make a change of a-variables $v = \varphi(u)$ as follows:

$$v_0 = \sum_{i=1}^n u_i \quad \text{and} \quad v_i = u_i/v_0, \quad i = 1, \ldots, n.$$

The new vector $v = (v_0, v_1, \ldots, v_n)$ lives in $\mathbb{V} = (0, \infty) \times \mathbb{S}_{n-1}$, where $\mathbb{S}_{n-1}$ is the $(n-1)$-dimensional probability simplex in $\mathbb{R}^n$, and the corresponding measure μ_φ is the product $\mathsf{Gamma}(n, 1) \times \mathsf{Dir}_n(p_n)$ with $p_n = n^{-1} 1_n$. Then the modified association model is

$$x_i = v_0 v_i/\theta_i, \quad i = 1, \ldots, n. \tag{6.5}$$

For the P-step, we shall consider the following PRS:

$$\mathcal{S}_v = \{\tilde{v} : f(\tilde{v}) \leq f(v)\}, \tag{6.6}$$

where

$$f(v) = -\sum_{i=1}^{n-1} \{a_i \log s_i(v) + b_i \log[1 - s_i(v)]\},$$

with $s_i(v) = \sum_{j=1}^i v_i$, $a_i = 1/(n-i-0.3)$, and $b_i = 1/(i-0.3)$. The would-be last term in $f(v)$, with $i = n$, is omitted since $s_n(v)$ is identically 1 for all v. A few remarks are in order regarding this choice of PRS.

- The random vector $(s_1(V), \ldots, s_{n-1}(V))$, for $V \sim \mu_\varphi$, has the distribution of a vector of sorted $\mathsf{Unif}(0, 1)$ random variables. The problem

of predicting sorted uniforms is important in a number of IM applications. The PRS in (6.6) with $f(v)$ as above is shown by Zhang (2010, Sec. 3.4.2) to provide an easy-to-compute alternative to the hierarchical PRS used in Martin, Zhang and Liu (2010).

- The fact that f is continuous implies, via Theorem 1, that this choice of PRS is credible in the sense of Definition 2.
- The first component v_0 of the v-vector is essentially ignored in the construction of the PRS. This is partly for convenience, and partly because v_0 is related to the overall scale of the problem which is irrelevant to the homogeneity assertion.

For the C-step, combining the observed data, the association model (6.5), and the PRS (6.6), we get the following random set for θ:

$$\Theta_x(\mathcal{S}_V) = \{\theta : f(v(x,\theta)) \leq f(V)\},$$

where $v(x,\theta) = (\theta_1 x_1, \ldots, \theta_n x_n)/\sum_{j=1}^n \theta_j x_j$, and $V = (V_0, V_1, \ldots, V_n) \sim \mu_\varphi$. The lower evidence function is zero for the homogeneity assertion. It is important to note that if θ is a constant vector, then $v(x,\theta)$ is independent of the constant, i.e., $v(x,\theta) = v(x, 1_n)$, which greatly simplifies computation of the upper evidence function at A. Indeed,

$$\bar{\mathsf{e}}_{x,\mathcal{S},\varphi}(A) = \mu_\varphi\{v : f(v(x,1)) \leq f(v)\},$$

which can easily be evaluated using Monte Carlo. As described in Section 5, the level α IM-based tests rejects if $\bar{\mathsf{e}}_{x,\mathcal{S},\varphi}(A) \leq \alpha$.

To illustrate the performance of this IM-based approach to testing for homogeneity, we will reproduce the simulation study presented in Martin, Zhang and Liu (2010). Here we compare our results with the basic likelihood ratio test, which is based on the test statistic

$$\lambda(x) = \{(\textstyle\prod_{i=1}^n x_i)^{1/n}/\overline{x}\}^n,$$

a power of the ratio of geometric to arithmetic means. The sampling distribution of $\lambda(X)$ is invariant to scale transformations, so its distribution under the homogeneity assumption is independent of the common θ value. We compare the power of the IM and likelihood ratio tests in several different cases. In each setup, exactly $n_1 + n_2 = 100$ events are observed, but the first n_1 exponential rates equal 1 while the last n_2 equal θ. That is, the data is sampled from a Poisson process with a single change point. Figure 2 summarizes the power of the two tests in this problem over a range of θ values for two configurations of (n_1, n_2). Here we see that, in both cases,

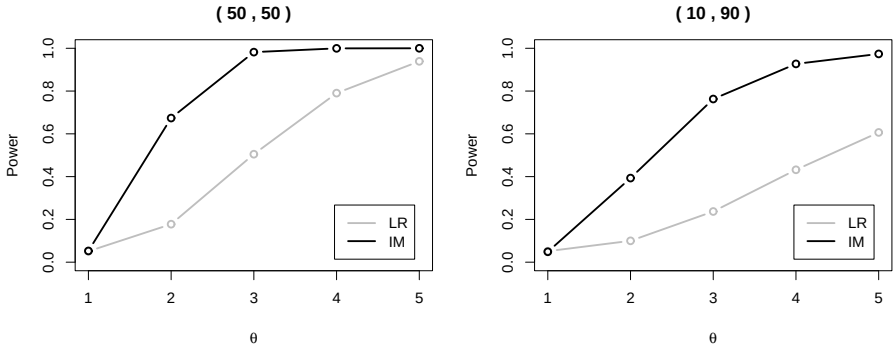

FIG 2. *Powers of the IM and likelihood ratio tests for the simulation described in Section 6.2. The left shows* $(n_1, n_2) = (50, 50)$ *and the right shows* $(n_1, n_2) = (10, 90)$*. In both cases,* θ *is the ratio of the rate of the last* n_2 *observations to that of the first* n_1*.*

the IM-based test has much larger power than the likelihood ratio test. We should mention that the choice of PRS used for the IM procedure is closely related to the particular choice of assertion/hypothesis, while the likelihood ratio test is more of a black-box procedure, so the better performance of the former is not so surprising. However, there does not appear to be a straightforward classical test which is particularly good against the alternative of at least one change-point.

7. Discussion. In this paper we have described a simple three-step procedure to construct IMs for prior-free probabilistic inference. The basis for this new approach is that probabilistic inference requires a relationship between data, parameters, and a predictable quantity. The proposed association model does just that, and fits in nicely in the three-step construction. Other methods are available for constructing probabilistic inference but, as described above, it is challenging to guarantee that the numerical values of such probabilities are meaningful across users or experiments. We proved in Section 4 that an IM yields frequency-calibrated probabilities under very general conditions. The big point is that the values of the corresponding plausibility function are meaningful both within and across experiments, accomplishing both the frequentist and Bayesian goals simultaneously.

Admittedly, the final IM depends on the user's choice of association model and PRS, but we do not believe that this is particularly damning. Efforts to define and construct optimal association models and PRSs are ongoing, but a case can be made to prefer the "arbitrariness" of the choice of PRS over that of a frequentist's choice of statistic or Bayesian's choice of prior.

The point is that neither a frequentist sampling distribution nor a Bayesian prior distribution adequately describes the source of uncertainty about θ. As we argued above, this uncertainty is fully characterized by the fact that, whatever the association model, the value of $u^\star$ is missing. Therefore, it seems only natural to prefer the IM framework that features a direct attack on the source of uncertainty over another that attacks the problem indirectly.

To elaborate a bit more on the non-uniqueness issue, we note that differences between IM outputs from different PRSs are slight for assertions involving one-dimensional quantities. However, for high-dimensional a-variables, the non-uniqueness issue deserves special attention. The basic idea of IMs is to construct PRSs by accurately predicting functions of a-variables that are most relevant to specific assertions of interest. This often corresponds to dimension reduction for a-variables or their functions to be predicted. It is interesting that this approach to a-variable dimension reduction has some close connections to Fisher's theory of sufficient statistics; see Martin, Hwang and Liu (2011a) for details. For nuisance parameter problems, like those in Section 6, there is a different form of dimension reduction required, namely marginalization. Martin, Hwang and Liu (2011b) discuss this technique and apply their methods to the famous Behrens-Fisher problem.

Of course, compared to Bayesian and frequentist methods, which are well-developed in the last century, IMs have many open problems. Both theoretical work, including conditioning and marginalization techniques for dimension reduction, and applications, including large-scale multiple testing and variable selection, have shown that the IM framework is promising. Given the attractive properties of IMs developed here and in the references above, we expect to see more exciting advancements in IMs or new inferential frameworks that are probabilistic and have desirable frequency properties.

References.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Anal.* **1** 385–402 (electronic). MR2221271

Berger, J. O., Bernardo, J. M. and Sun, D. (2009). The formal definition of reference priors. *Ann. Statist.* **37** 905–938. MR2502655

Bernardo, J.-M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B* **41** 113–147.

Dawid, A. P. (1985). Calibration-based empirical probability. *Ann. Statist.* **13** 1251–1285. With discussion. MR811493

Dempster, A. P. (2008). Dempster-Shafer calculus for statisticians. *Internat. J. of Approx. Reason.* **48** 265–277.

Efron, B. (2010). *Large-scale inference. Institute of Mathematical Statistics Monographs* **1**. Cambridge University Press, Cambridge. MR2724758

Ermini Leaf, D., Hui, J. and Liu, C. (2009). Statistical inference with a single observation of $N(\theta, 1)$. *Pak. J. Statist.* **25** 571–586.

Ermini Leaf, D. and Liu, C. (2010). A weak belief approach to inference on constrained parameters: elastic beliefs. *Working paper*.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* **222** 309–368.

Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **22** 200–225.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc. A* **144** 285–307.

Fraser, D. A. S. (1966). Structural probability and a generalization. *Biometrika* **53** 1–9. MR0196840

Fraser, D. A. S. (1968). *The structure of inference*. John Wiley & Sons Inc., New York. MR0235643

Little, R. (2010). Calibrated Bayes, for statistics in general, and missing data in particular. *Statist. Sci.* To appear.

Martin, R., Hwang, J.-S. and Liu, C. (2011a). Conditional inferential models. *Working manuscript*. Old version: www.math.iupui.edu/~rgmartin.

Martin, R., Hwang, J.-S. and Liu, C. (2011b). Marginal inferential models. *Working manuscript*. Old version: www.math.iupui.edu/~rgmartin.

Martin, R., Zhang, J. and Liu, C. (2010). Dempster-Shafer theory and statistical inference with weak beliefs. *Statist. Sci.* **25** 72–87.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR760681

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J. MR0464340

Wasserman, L. (1990). Prior envelopes based on belief functions. *Ann. Statist.* **18** 454–464.

Zabell, S. L. (1992). R. A. Fisher and the fiducial argument. *Statist. Sci.* **7** 369–387. MR1181418

Zhang, J. (2010). Statistical inference with weak beliefs PhD thesis, Purdue University.

Zhang, J. and Liu, C. (2010). Dempster-Shafer inference with weak beliefs. *Statist. Sinica*, to appear. Preprint www.stat.purdue.edu/~chuanhai.

Department of Mathematical Sciences
Indiana University-Purdue University Indianapolis
402 North Blackford Street, LD270
Indianapolis, IN 46202, USA
E-mail: rgmartin@math.iupui.edu

Department of Statistics
Purdue University
250 North University Street
West Lafayette, IN 47907, USA
E-mail: chuanhai@stat.purdue.edu

date: February 14, 2011