October 28, 2011

Valid Prior-Free Probabilistic Inference and its Applications in Medical Statistics

Duncan Ermini Leaf, Hyokun Yun, and Chuanhai Liu*

Abstract:

Valid, prior-free, and situation-specific probabilistic inference is desirable for serious uncertain inference, especially in bio-medical statistics. This chapter* introduces such an inferential system, called the Inferential Model (IM) framework, proposed recently. IMs do not require a prior to be specified, yet they produce probabilistic inferential results that have desireable frequency properties. This chapter illustrates the IM framework and demonstrates its potential applications in bio-medical statistics with a collection of benchmark examples, including (i) classification, (ii) inference with subgroup selection, (iii) 2×2 tables, and (v) a many-normal-means problem in meta-analysis.

1. Introduction

1.1. Statistical inference: a brief historical review

For serious statistical applications, such as bio-medical statistics, producing prior-free, frequency-calibrated (or valid), data-dependent (or situation-specific) probabilistic inference has been a longstanding dream of statisticians. Probabilistic inference is attractive as an easy-to-interpret quantitative assessment of uncertainty. It has a long history that can be traced back to Bayes (1763); see also Liang, Liu, and Carroll (2010). Bayes is clearly the most sensible approach when the statistician has a valid prior distribution for unknown quantities. Philosophically, there is nothing wrong and it probably should always be attempted in practice so long as the statistician and his or her clients are willing to accept the consequences of any decision made accordingly. Nevertheless, we shall not discuss philosophical differences here, such as those concerning the different meanings or interpretations of probability. In any case, when a prior is taken for everything, the inference problem is reduced to an exercise of usual probability calculus. Following, *e.g.*, Fraser (2011), for conceptual clarity we simply refer to such models as probability models. This chapter is concerned with statistical inference when there is no known prior for some unknown quantity.

*Department of Statistics, Purdue University, 150 N. University Street, West Lafayette, IN 47907. e-mail: chuanhai@purdue.edu.

*This is prepared as an invited contribution to *Advanced Medical Statistics*, 2nd edition by Ying Lu and Ji-Qian Fang. World Scientific.

For Bayesian approaches to statistical inference, it is a fair comment that not all Bayesian statisticians really put meaning on prior distributions for everything. Most often in practice, part of the prior is used for simplicity to produce inference that matches or beats the best possible frequentist results. This is supported, for example, by the use of Bayes factors and so-called conjugate and invariance priors, among many others. An agreed way of specifying priors, especially in modern, challenging, and large-scale problems, is yet to emerge.

When priors are unknown, it is perhaps due to the fact that Bayesian inferential results are generally not frequency-calibrated that most statisticians have fallen into what is known as the frequentist school. However, frequentist procedures are often created via what can be thought of as a *guess-and-check* process. A proposed procedure is often checked according to the criteria set forth by Fisher (1922), namely, sufficiency, efficiency, and consistency. Due to the limitations of Bayesian and frequentist schools of thought, around 1930 R. A. Fisher started developing an ideal inferential system, which he called fiducial. He spent the rest of his life, about 30 years, working on this. Although, some of the ideas in fiducial inference were reinterpreted by J. Neyman to create confidence intervals that have been central to frequentist statisticians, Fisher's fiducial has been perceived as his *"one great failure"* (Zabell, 1992).

While acknowledging the limited success of his efforts, R. A. Fisher apparently recognized that there was something valuable in fiducial. He wrote (see Savage (1964) p. 926): *"I don't understand yet what fiducial probability does. We shall have to live with it a long time before we know what it's doing for us. But it should not be ignored just because we don't yet have a clear interpretation."* Efforts continuing along this direction can be found in further developments of Fisher's fiducial inference and its variants: Fraser's structural inference (Fraser, 1966, 1968) and Dempster-Shafer theory (Dempster, 2008; Shafer, 1976); see Hannig (2009) for a recent discussion.

What is fiducial? A brief but good answer was given by Savage (1976), who wrote: "*The expressions 'fiducial probability' and 'fiducial argument' are Fisher's. Nobody knows just what they mean, because Fisher repudiated his most explicit, but definitely faulty, definition and ultimately replaced it with a few examples.*" When referring to the fiducial argument, we shall adopt the interpretation of Dempster (1963), namely, *continuing to believe that the pivotal quantity introduced in fiducial has the same distribution after seeing the observed data.* For example, consider inference about θ from an observation x that is assumed to have come from the normal distribution $N(\theta, 1)$ with unknown mean $\theta \in \mathbb{R} = (-\infty, \infty)$ and unit variance. We refer to this problem the one-normal-mean example. To obtain a fiducial probability, one writes

$$x = \theta + Z \qquad (Z \sim N(0,1)), \tag{1.1}$$

where the introduced random variable Z is called a pivotal variable. The connection between (1.1) and the familiar sampling distribution $x \sim N(\theta, 1)$ is made by assuming that $Z \sim N(0,1)$ holds *a priori*, *i.e.*, before seeing x. After seeing x, *i.e.*, conditional on x, with the fiducial argument one continues to believe that $Z \sim N(0,1)$. Since (1.1) implies that $\theta = x - Z$, the distribution on Z defines a distribution on θ

when x is fixed at its observed value. That distribution is $\theta \sim N(x, 1)$ conditional on x, which is the same as the Bayesian posterior of θ obtained by putting the flat prior on θ; see Lindley (1958) and Fraser (2011). When used for inference, this posterior distribution has nice frequency properties for certain assertions or hypotheses on the unknown quantity θ. In general, however, it is questionable to manipulate it as a distribution using the usual probability calculus; see Ermini Leaf et al (2009).

While not as popular as Bayesian inference nowadays, fiducial inference can provide better results than Bayesian methods for certain models; see Section 2. This observation is partially consistent with what Efron (1998) wrote: *". . . but here is a safe prediction for the 21st century: statisticians will be asked to solve bigger and more complicated problems. I believe there is a good chance ... that something like fiducial inference will play an important role in this development. Maybe Fisher's biggest blunder will become a big hit in the 21st century!"* While this "something" remains to be discovered, we and our collaborators recently proposed a new promising inferential framework, called the inferential model (IM) framework, that produces prior-free, frequency-calibrated, and data-dependent probabilities for assessing uncertainty. Although it is both philosophically and technically different from fiducial, the IM framework was motivated by fiducial and its extension, the Dempster-Shafer theory of belief functions. As a matter of fact, the original but fundamental idea of IMs was explored in Martin, Zhang, and Liu (2010), Zhang and Liu (2011) and, Zhang (2010) in the context of Dempster-Shafer theory. In the sequel, we present the IM framework reformulated and extended most recently in Martin and Liu (2011); see also Martin, Hwang, and Liu (2011a,b), and Ermini Leaf and Liu (2011).

1.2. Outline

The key starting point for IMs is to consider reasoning with uncertainty toward predictive or probabilistic inference. Section 2.1 reviews the idea that probabilistic inference requires an unobserved but predictable quantity, called an auxiliary (a)-variable, to be introduced as an integral part of a full statistical model. Given a full statistical model that allows for probabilistic inference, Section 2.2 uses the one-normal-mean to summarize and illustrate a simple but general three-step procedure for constructing IMs. Section 3 reviews conditional and marginal IMs that are necessary for efficient inference.

The second part of this chapter, Sections 4 — 7, presents a sequence of potential applications of IMs in bio-medical statistics, including classification, constrained parameter problems, selective inference, 2×2 tables, and the many-normal-means problem in meta-analysis. The article concludes with a few remarks on further developments of the IM theory and its applications. We hope readers will agree with us that prior-free, valid, and data-dependent probabilistic inference is within reach rather than impossible, as might have been mistakenly perceived from the failure of Fisher's efforts in fiducial inference; see Zabell (1992). Serious statistical applications and challenging large-scale problems demand exact probabilistic inference such as that provided by IMs (see Martin et. al. (2011b)), or similar inferential

methods yet to be developed. We also hope that readers share our vision that this line of thinking has a great future for many, many years to come.

2. Inferential Models

For the sake of simplicity, the one-normal-mean example, *i.e.*, inference about $N(\theta, 1)$ from a single observation x,

$$x \sim N(\theta, 1) \qquad (\theta \in \Theta \equiv \mathbb{R}), \tag{2.1}$$

serves as a running example in this section. Probabilistic inference about the unknown quantity θ from the observed data x is of interest. Any hypothesis about θ corresponds to a subset of Θ called an assertion. In this section, we consider inference about the truth and falsity of the assertion

$$\mathcal{A}_{\theta_0} = \{\theta_0\} \subseteq \Theta \tag{2.2}$$

where θ_0 is some specific value of interest, *e.g.*, $\theta_0 = 0$. For clarity, we note that the subset $\mathcal{A}_{\theta_0}$ of Θ is said to be *true* if the true value of the unknown θ belongs to $\mathcal{A}_{\theta_0}$. Otherwise, $\mathcal{A}_{\theta_0}$ is said to be *false* when the true value of θ is contained in $\mathcal{A}_{\theta_0}^c$, the complement or negation of $\mathcal{A}_{\theta_0}$.

2.1. Association models for probabilistic inference

The postulated model (2.1), known as a sampling model for the observed data x, is a familiar way of starting statistical inference. Perhaps, it is relatively less familiar (see, e.g., Dempster, 1964) that the sampling model (2.1) doesn't allow for desired predictive or probabilistic inference. This is because (i) the sampling model specifies a class of different probability measures indexed by θ, and (ii) the probability measure in the sampling model (2.1) only allows the evaluation of event probabilities to confirm or compare different explanations of observed data in different probability spaces. Since no *missing but predictable quantity* is present, no probabilistic inference is possible. To emphasize this observation, Martin and Liu (2011) introduced the following principle.

Principle 2.1. *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.*

According to this principle, something that is missing but predictable, called an *auxiliary (a)-variable*, must be introduced as an integral part of modeling for statistical inference. Following this principle, IMs require the specification of an *association model* by introducing a-variables. The probability distribution of the a-variables determines the postulated sampling model for the observed data. For

the running example of the one-normal-mean problem, an association model can be specified as in fiducial inference, *i.e.*,

$$x = \theta + z \qquad (z \sim N(0,1)), \tag{2.3}$$

where the a-variable z is missing, but *a priori* assumed to follow the standard normal distribution, $N(0,1)$. Given x and the true value of z, the true value of the unknown quantity θ is then given by $\theta = x - z$. In the general setting, x denotes both the observable and the observed data. An association-model determines the mapping from (x, z) to θ, which in general can be set-valued and thus written as

$$\Theta_x(z) \subseteq \Theta. \tag{2.4}$$

Given an association model, predictive or probabilistic inference about θ amounts to predicting the unobserved realization of the a-variable, z. Motivated by the concepts of Bayesian credible intervals, frequentist confidence intervals that produce inferential results with desirable certainty (i.e., with probabilities close to one or zero), and frequentist rejection regions for hypothesis testing, predicting z can be made simple to create user-friendly inferential outputs. This explains, at least partially, the following use of *predictive random sets* (PRSs). The IM framework, built around this simple idea, is briefly reviewed next.

2.2. The three-step representation of IMs

The IM framework can be represented as a three-step procedure for inference about $\mathcal{A} \subseteq \Theta$, an assertion on θ. This three-step procedure consists of an Association (A) step, a Prediction (P) step, and a Combination (C) step as follows.

Inferential Model 2.1 (The three-step procedure for IMs)**.** *An IM is built using three steps:*

A-STEP. *Associate the observable x and unknown θ with an a-variable z to obtain a set-valued mapping $\Theta_x(z)$ $(\subseteq \Theta)$, which consists of the candidate values of θ given z and x.*

P-STEP. *Predict z by specifying a credible (PRS) $\mathcal{S}_\theta$, possibly depending on θ; and*

C-STEP. *Combine $\Theta_x(z)$ and $\mathcal{S}_\theta$ to obtain the PRS $\Theta_x(\mathcal{S}_\theta) = \cup_{z \in \mathcal{S}_\theta} \Theta_x(z)$ for θ, and compute the probabilities*

$$\underline{\mathsf{e}}_x(\mathcal{A}) \equiv Pr(\Theta_x(\mathcal{S}_\theta) \subseteq \mathcal{A})$$

and

$$\overline{\mathsf{e}}_x(\mathcal{A}) \equiv 1 - Pr(\Theta_x(\mathcal{S}_\theta) \subseteq \mathcal{A}^c) = Pr(\Theta_x(\mathcal{S}_\theta) \cap \mathcal{A} \neq \emptyset)$$

as the lower evidence and upper evidence for $\mathcal{A}$.

Random variables represent uncertainty in the postulated sampling model. The A-step emphasizes the modeling of data using a-variables for probabilistic inference. For the one-normal-mean example, this step is formally written as

A-STEP. This is given by (2.3), which gives the singleton set-valued mapping

$$\Theta_x(z) = \{\theta : \ \theta = x - z\}. \tag{2.5}$$

A PRS is a random set $\mathcal{S}_\theta$ that takes values in the power set of the space of the a-variables. A realization of the PRS $\mathcal{S}_\theta$ can be interpreted as a set that is believed to include the unobserved value of the a-variable. The strength of this belief is defined by the probability distribution for $\mathcal{S}_\theta$. As mentioned previously, this set of possibilities can depend on the true value of θ. The use of a so-called credible PRS is intended to be user-friendly in the sense that the resulting probabilistic outputs can be interpreted and used for uncertain inference in a practical and straightforward manner; see Liu and Martin (2012) for more discussion on interpreting the evidence functions as belief probability in the context of scientific inference. Roughly speaking, a PRS in the P-step is said to be credible if it produces probabilities that have numerical values consistent with those of frequency or additive probabilities. Formally, the credibility of PRSs is defined as follows.

Definition 2.1 (Credibility)**.** *Let Π_z be the probability distribution of the a-variable z. The PRS S for predicting z is said to be credible if it satisfies*

$$Pr(\mathcal{S} \not\ni z) \overset{Stochastically}{\leq} U \sim Unif(0,1) \qquad (z \sim \Pi_z), \tag{2.6}$$

where $Pr(\mathcal{S} \not\ni z)$ is computed with respect to the distribution of $\mathcal{S}$ for fixed z.

Note that (2.6) can be equivalently expressed as

$$\Pr\left(\Pr\left(\mathcal{S} \not\ni z\right) \geq 1 - \alpha\right) \leq \alpha,$$

for each $\alpha \in (0,1)$, where $\Pr\left(\mathcal{S} \not\ni z\right)$ is a function of the random a-variable, z. This can be interpreted as follows: for any fixed z, we can compute $\Pr\left(\mathcal{S} \not\ni z\right)$ as a measure of belief that z is *not* a realization from Π_z. If we use a threshold $1-\alpha$ to make a decision that z is not the unobserved a-variable realization from Π_z in a particular experiment of interest, then we would make wrong decision at most $100\alpha\%$ of the time for z realizations from Π_z over repeated experiements. For example, the PRS

$$\mathcal{S}_\theta = \{z : |z| \leq |Z|\} = [-|Z|, |Z|] \qquad (Z \sim N(0,1)) \tag{2.7}$$

is credible for predicting an unobserved z taken from $N(0,1)$ due to the fact that

$$\Pr\left(\mathcal{S}_\theta \not\ni z\right) = \Pr\left(|Z| < |z|\right) = F_{|Z|}(|z|) \sim \text{Unif}\,(0,1) \qquad (z \sim N(0,1)).$$

where $F_{|Z|}(.)$ stands for the cdf of $|Z|$. Using this PRS, we have the following P-step for the one-normal-mean problem.

P-STEP. Predict z using the PRS in (2.7).

This credibility requirement for PRSs is essential as it guarantees the validity of inference about θ. This is discussed below along with the explanation of the C-step.

The operations of the C-step should be self-explanatory. Here, we assume that $\Theta_x(\mathcal{S}_\theta) \neq \emptyset$ holds with probability one. For cases where $\Pr(\Theta_x(\mathcal{S}_\theta) = \emptyset) > 0$, Ermini Leaf and Liu (2011) replace the credible PRS with what they call a credible Elastic PRS (EPRS); see Section 4. Most important is that the resulting $\underline{\mathsf{e}}_x(\mathcal{A})$ and $\overline{\mathsf{e}}_x(\mathcal{A})$ probabilities are meaningful in the sense that their numerical values are consistent with a frequency interpretation of probabilities. To be more precise, we need the formal definition of what is called validity.

Definition 2.2 (Validity)**.** *Suppose that $x \sim f(x|\theta)$ and $\mathcal{A} \subseteq \Theta$. An inferential framework is said to be valid for inference about $\mathcal{A} \subseteq \Theta$ if and only if it produces probabilities $\underline{\mathsf{e}}_x(\mathcal{A})$ and $\overline{\mathsf{e}}_x(\mathcal{A})$ for $\mathcal{A}$, as functions of the random variable x, that satisfy*

$$\underline{\mathsf{e}}_x(\mathcal{A}) \overset{stochastically}{\leq} U \sim Unif(0,1)$$

for $\theta \in \mathcal{A}^c$ and

$$\overline{\mathsf{e}}_x(\mathcal{A}) \overset{stochastically}{\geq} V \sim Unif(0,1)$$

for $\theta \in \mathcal{A}$. An inferential framework is said to be valid if and only if it is valid for all assertions $\mathcal{A} \subseteq \Theta$.

The validity theorem (see, *e.g.*, Martin and Liu, 2011) says that *an IM is valid as long as the PRS of the P-step is credible.* The implication can be explained by making analogy with frequentist procedures as follows. The values of $\underline{\mathsf{e}}_x(\mathcal{A})$ and $\overline{\mathsf{e}}_x(\mathcal{A})$ can be interpreted in practice as: *a very large value of the lower evidence $\underline{\mathsf{e}}_x(\mathcal{A})$ practically "confirms" the truth of $\mathcal{A}$, and a very small value of the upper evidence $\overline{\mathsf{e}}_x(\mathcal{A})$ practically "confirms" the falsity of $\mathcal{A}$.* The use of the word "confirm" is in the sense of the Cournot principle (see, e.g., Shafer, 2011): *an event of very small probability will not happen.* This interpretation will not lead to a contradiction because $\underline{\mathsf{e}}_x(\mathcal{A}) \leq \overline{\mathsf{e}}_x(\mathcal{A})$ always holds. For example, given $\alpha = 0.01$ as threshhold for small probabilities, corresponding to the common type-I level of Neyman's test, we have three cases to consider:

1. The case $\underline{e}_x(A) > 1 - .01 = 0.99$: confirms the truth of $\mathcal{A}$;
2. The case $\overline{e}_x(A) < 0.01$: confirms the falsity of $\mathcal{A}$ (or the truth of $\mathcal{A}^c$); and
3. The other case: neither the truth of $\mathcal{A}$ nor the truth of $\mathcal{A}^c$ can be confirmed. In this case, there is not enough evidence to make a decision for the given threshold, or we fail to take a decisive action.

To complete the IM for the one-normal-mean example, we give the following C-step.

C-step. Combine $\Theta_x(z)$ and $\mathcal{S}_\theta$ to obtain $\Theta_x(\mathcal{S}) = [x - |Z|, x + |Z|]$, and compute probabilities

$$\underline{\mathsf{e}}_x(\mathcal{A}_{\theta_0}) = 0 \text{ and } \overline{\mathsf{e}}_x(\mathcal{A}_{\theta_0}) = 2[1 - \Phi(|x - \theta_0|)].$$

for the assertion A_{θ_0}, where $\Phi(.)$ stands for the cdf of the standard normal distribution, $N(0,1)$.

The evidence $\underline{\mathsf{e}}_x(\mathcal{A}_{\theta_0})$ and $\overline{\mathsf{e}}_x(\mathcal{A}_{\theta_0})$ in the C-step can be derived as follows:

$$\begin{aligned}
\underline{\mathsf{e}}_x(\mathcal{A}_{\theta_0}) &= \Pr(\Theta_x(\mathcal{S}) \subseteq \mathcal{A}_{\theta_0}) \\
&= \Pr(\{\theta : |x-z| \le |Z|\} \subseteq \{\theta_0\}) \\
&= \Pr(\{\theta : |x-z| > |Z|\} \supseteq \{\theta_0\}^c) \\
&= \Pr(|Z| < |x-\theta| \text{ for all } \theta \ne \theta_0) \\
&= \Pr\left(|Z| < \min_{\theta \ne \theta_0} |x-\theta|\right) \\
&= 0
\end{aligned}$$

and

$$\begin{aligned}
\overline{\mathsf{e}}_x(\mathcal{A}_{\theta_0}) &= 1 - \Pr\left(\Theta_x(\mathcal{S}) \subseteq \mathcal{A}_{\theta_0}^c\right) \\
&= 1 - \Pr(\{\theta : |x-z| \le |Z|\} \subseteq \{\theta_0\}^c) \\
&= 1 - \Pr(\{\theta : |x-z| > |Z|\} \supseteq \{\theta_0\}) \\
&= 1 - \Pr(|Z| < |x-\theta_0|) \\
&= 2[1 - \Phi(|x-\theta_0|)].
\end{aligned}$$

The logic in these algebraic operations is typical in computing evidence functions. In the sequel, we shall omit detailed calculations of evidence functions to save space.

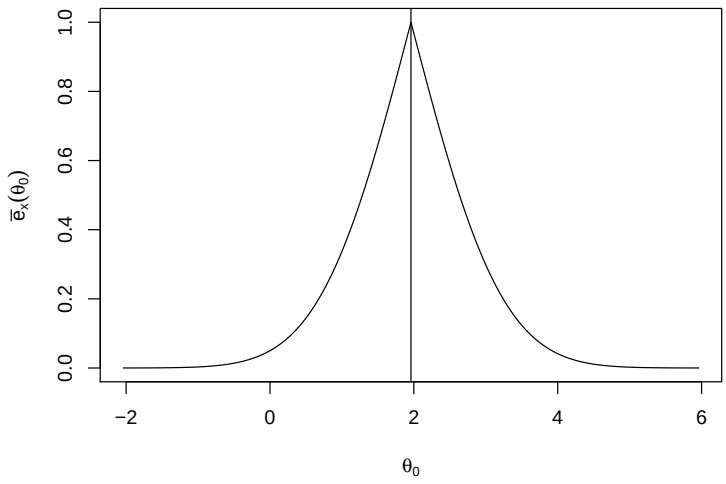

FIG 1. *The most evidence for* $\mathcal{A}_{\theta_0} = \{\theta : \ \theta = \theta_0\}$ *given the observed* $x = 1.96$*, indicated by the vertical line, in the one-normal-mean example in Section 2.*

For a simple numerical example, take $x = 1.96$ and $\mathcal{A}_0 = \{\theta = 0\}$. We have $\underline{\mathsf{e}}_x(\mathcal{A}_0) = 0$ and $\overline{\mathsf{e}}_x(\mathcal{A}_0) = 2[1 - \Phi(1.96)] = 0.05$. Unlike Fisher's p-value, our probability $\overline{\mathsf{e}}_x(\mathcal{A}_0)$ is predictive and therefore provides truly probabilistic inference. This probability has a meaningful probability interpretation for *situation-specific* inference, *i.e.*, it does not need repeated experiments, as in the frequentist school of thought, to help validate its meaningfulness. The results for a sequence of values of θ_0 are displayed in Figure 1. The region with small values of $\overline{\mathsf{e}}_x(\mathcal{A}_0)$ contains the least likely values of the unknown quantity θ. In other words, the evidence *against* $\mathcal{A}_{\theta_0}$ increases as θ_0 deviates away from the observed x. When θ_0 is close to the observed data point x, our uncertainty about the truth or falsity of $\mathcal{A}_{\theta_0}$ is large.

The preceding discussion shows how IMs can be used for hypothesis testing after translating a hypothesis into an assertion, $\mathcal{A}$. The upper evidence, $\overline{\mathsf{e}}_x(\mathcal{A})$ can be used in a similar manner to a p-value, although the interpretation is different. When $\overline{\mathsf{e}}_x(\mathcal{A})$ is small, the assertion $\mathcal{A}$ can be rejected. However, large $\overline{\mathsf{e}}_x(\mathcal{A})$ values do not mean that the assertion should be accepted. Only large values of $\underline{\mathsf{e}}_x(\mathcal{A})$ serve to confirm that $\mathcal{A}$ is true. As shown in the preceding example, $\underline{\mathsf{e}}_x(\mathcal{A})$ may be zero for any point assertions of the form $\mathcal{A}_{\theta_0}$. In these cases, the hypothesis that $\theta = \theta_0$ can never be confirmed. However, $\overline{\mathsf{e}}_x(\mathcal{A}_{\theta_0})$ can used for estimation. Simply choose the value of θ_0 that maximizes $\overline{\mathsf{e}}_x(\mathcal{A}_{\theta_0})$. This is the most plausible value of θ given the observed x. In the preceding example, the most plausible value of θ_0 is x (1.96). For an interval estimate, we can find

$$\Gamma_x(\alpha) = \{\theta_0 : \overline{\mathsf{e}}_x(\mathcal{A}_{\theta_0}) > \alpha\}.$$

This is known as a *plausibility interval*. Like a confidence interval, $\Gamma_x(\alpha)$ has $1 - \alpha$ probability of covering a fixed θ value over repeated experiments.

2.3. PRSs and efficiency considerations

Efficiency of IMs is an important issue that is determined by the PRS specified in the P-step. First, it is easy to see that the PRS for predicting a-variables can be made stochastically as small as possible, provided that it is credible (and that the space of the a-variable is metric). For continuous a-variables, it is possible to construct a PRS such that the credibility condition (2.6) is replaced with

$$\Pr(\mathcal{S} \not\ni z) \sim \text{Unif}(0,1) \qquad (z \sim \Pi_z). \tag{2.8}$$

Such a PRS is called an *efficient* PRS. For example, for continuous a-variables, let $b(z)$ be an arbitrary continuous function defined on the a-space, the space of the a-variable. It is easy to show that the random set defined by the boundary function $b(z)$,

$$\mathcal{S} = \{z : b(z) \geq b(Z)\} \qquad (Z \sim \Pi_z), \tag{2.9}$$

is credible and efficient.

It should be noted that an efficient PRS doesn't necessarily mean that the resulting inference for any assertion $\mathcal{A}$ is efficient in the sense of maximizing the

distribution of $\underline{\mathsf{e}}_{x,\mathcal{S}}(\mathcal{A})$ and minimizing the distribution of $\bar{\mathsf{e}}_{x,\mathcal{S}}(\mathcal{A})$ over all efficient and credible PRSs for predicting z. In some sense, the concept of efficient PRSs is meant to emphasize that PRSs do not need to be stochastically larger than necessary for credibility. Among a few topics that can be investigated based on this observation is the construction of assertion-specific PRSs. The shapes of PRSs can be specified according to assertions of interest. This is particularly useful for multiple-parameter inference problems. For single-parameter problems efficiency is less an issue because the inferential results for precise single-parameter assertions, such as those in the one-normal-mean example, based on different credible and efficient PRSs are not very different. Two important classes of methods for making efficient IM inference are discussed next.

3. Conditional and Marginal IMs

Although valid, IMs constructed by predicting a high-dimensional a-variable can be inefficient. Here we review two types of dimensionality reduction for efficient inference. One is to predict only a minimum number of a-variables, typically of the same dimension as the unknown quantity θ. The other is to predict an assertion-specific low dimensional quantity. These two types of dimension reduction of a-variables are considered in Sections 3.1 and 3.2. Inference about the Gaussian model $N(\mu, \sigma^2)$ with unknown μ and known or unknown variance σ^2 from an observed sample $x_1, ..., x_n$, a standard textbook example, is used in this section for both motivation and illustration.

3.1. Combining information via dimension reduction: conditional IMs

Suppose that the observed data $x_1, ..., x_n$ are considered as a sample from the Gaussian model $N(\mu, \sigma^2)$ with unknown mean $\mu \in \mathbb{R}$ and known variance σ^2. Without loss of generality, take $\sigma = 1$. It is natural to extend the discussion of the $n = 1$ case, the one-normal-mean example, to the present $n > 1$ case by writing the baseline association model

$$x_i = \mu + Z_i \qquad (Z_i \overset{iid}{\sim} N(0,1)), \tag{3.1}$$

where the a-variable is the random vector $Z = (Z_1, ..., Z_n)'$. One can predict the n-dimensional a-variable Z for making probabilistic inference about the unknown quantity μ. It is immediately clear that predicting one component of Z, say Z_1, would be sufficient for such a purpose. However, with a closer look at the association model (3.1) we would find that some functions, for example,

$$Z_i - Z_1 = x_i - x_1 \qquad (i = 2, ..., n) \tag{3.2}$$

are fully observed. These observed a-variables can help predict Z_1 more accurately. This motivates the idea of (i) associating x_1 and μ with an unobserved but predictable quantity Z_1

$$x_1 = \mu + Z_1; \tag{3.3}$$

(*ii*) predicting Z_1 using a credible PRS based on its conditional distribution given (3.2); and (*iii*) combining the PRS constructed in (*ii*) and the association model (3.3) to evaluate evidence functions for assertions about μ.

Care must be taken to check that when conditioning on (3.2) there are a-variables available to be predicted for inference about μ, *i.e.*, the fundamental principle of Section 2 is not violated. This is not a problem in this motivating example. For general cases, to preserve the fundamental principle of Section 2, assume that the baseline association model $x = a(U, \theta)$ can be represented by a conditional association model

$$h_1(x) = a_1(V_1, \theta) \tag{3.4}$$

and a conditioning equation

$$h_2(x) = a_2(V_2), \tag{3.5}$$

where $V_1 = \psi_1(U)$ and $V_2 = \psi_2(U)$ form a one-to-one mapping between $V = (V_1, V_2)'$ and U from the the space of a-variable U to that of the new a-variable V. When such a representation exists, a regular conditional IM can be constructed, as suggested by the motivating example above, in the same way as constructing the IM in Section 2. That is

Inferential Model 3.1 (Conditional IM)**.** *The regular conditional IM (CIM) based on (3.4) and (3.5) has the following three steps.*

A-STEP. *Associate $h_1(x)$ and θ through the a-variable V_1, which gives the candidates of θ, $\Theta_{h_1(x)}(V_1) = \{\theta : h_1(x) = a_1(V_1, \theta)\}$, when $h_1(x)$ and V_1 become available, either observed or predicted.*

P-STEP. *Predict V_1 using a credible PRS $\mathcal{S}_\theta$ based on its conditional distribution given (3.5).*

C-STEP. *Combine $\mathcal{S}_\theta$ and $\Theta_{h_1(x)}(V_1)$ to obtain*

$$\Theta_{h_1(x)}(\mathcal{S}_\theta) = \{\theta : \theta \in \Theta_{h_1(x)}(v_1) \text{ for some } v_1 \in S_\theta\}.$$

Then, evaluate evidence functions in exactly the same way as in IM 2.1.

Under mild conditions, mainly that (3.4) is a valid association model that is consistent with the fundamental principle, the CIM is valid; see Martin, Hwang, and Liu (2011a). Here we illustrate the CIM by completing the above motivating example. The A-step gives the candidate set $\{\mu : \mu = x_1 - Z_1\}$. The conditional distribution of Z_1 given $(Z_2 - Z_1, ..., Z_n - Z_1)$ is

$$N\left(-\frac{1}{n}\sum_{i=2}^{n}(x_i - x_1), \frac{1}{n}\right). \tag{3.6}$$

In the P-step, we define a PRS for predicting Z_1 by predicting the residuals from (3.6). The resulting PRS for Z_1 is given by

$$\mathcal{S} = \{z_1 : z_1 = -\frac{1}{n}\sum_{i=2}^{n}(x_i - x_1) + z/\sqrt{n} \text{ for some } z \in (-|V|, |V|)\} \qquad (V \sim N(0,1)).$$

The P-step generates the PRS for μ

$$\{\mu : |\bar{x} - \mu| \leq |V|/\sqrt{n}\} \qquad (V \sim N(0,1)),$$

where $\bar{x}$ denotes the sample mean, i.e., $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$. In other words, the CIM is equivalent to the IM based on the sufficient statistic $\bar{x}$ alone. We note that this CIM can be easily derived by considering the conditional association model

$$\bar{x} = \mu + \frac{1}{\sqrt{n}}V$$

and the conditioning equation

$$x_i - \bar{x} = Z_i - \bar{Z} \qquad (i = 1, ..., n).$$

However, it should be verified that this system of conditioning equations requires only a $(n-1)$ dimensional observed characteristics to be given, leaving a one-dimensional observed quantity for preserving the fundamental principle.

Remark 3.1. *There are cases where regular CIMs may not exist. A many-normal-means problem in meta-analysis in Section 7 provides such an example. In such cases, the generalized CIM of Martin et al (2011a) can be used and illustrated in 7.*

Remark 3.2. *It should be noted that there is difference between conditional IM and Fisher's conditional inference. The latter is frequentist and, thereby, works with the sampling distribution of sufficient statistics conditional an ancillary statistics. That is, it requires other experiments for interpreting inferential results. IM conditional inference is data-dependent or situation-specific and doesn't need other experiments for interpretation.*

3.2. Efficient inference via dimensional reduction: marginal IMs

Suppose that the observed data $x_1, ..., x_n$ are considered as a sample from the Gaussian model $N(\mu, \sigma^2)$ with unknown $\mu \in \mathbb{R}$ and unknown $\sigma \in \mathbb{R}_+ = (0, \infty)$. The natural baseline association model is

$$x_i = \mu + \sigma Z_i \qquad (Z_i \overset{iid}{\sim} N(0,1); i = 1, ..., n).$$

Let s_x^2 be the sample variance of $x_1, ..., x_n$ and let s_Z^2 be the sample variance of the unobserved sample $Z_1, ..., Z_n$ from a known population. Then it is straightforward to show that the CIM obtained by using the conditional association model

$$\bar{x} = \mu + \sigma\bar{Z} \text{ and } s_x^2 = \sigma^2 s_Z^2 \tag{3.7}$$

and the conditioning equation

$$(Z - \mathbf{1}\bar{Z})/s_Z = (x - \mathbf{1}\bar{x})/s_x,$$

i.e., the observed direction of the centered Z, is given by the IM with (3.7) as the baseline association-model. That is, we can make inference about $\theta = (\mu, \sigma^2)$ based on the IM with the association model (3.7). The two a-variables $\bar{Z}$ and s_Z^2 are independent with

$$\bar{Z} \sim N(0, n^{-1}) \text{ and } (n-1)s_Z^2 \sim \chi^2_{n-1}.$$

Now suppose that we are interested in inference about σ alone. In this case, we see that inference can be made by predicting s_Z^2 alone. In other words, we can consider a PRS predicting $(\bar{Z}, s_Z^2)$ but focus more on the accuracy of predicting s_Z^2. This suggests that we construct a credible PRS in such a way that when projected down to space of s_Z^2, it is minimized. Such a PRS is given by a product of a credible PRS for predicting s_Z^2 and a largest possible PRS, i.e., $\mathbb{R}$, for predicting $\bar{Z}$. For example, using a centered (marginal) PRS

$$\{s_Z^2 : |F(s_Z^2) - .5| \leq F(M^2) - .5|\} \qquad ((n-1)M^2 \sim \chi^2_{n-1})$$

for s_Z^2 yields the PRS

$$\mathcal{S} = \mathbb{R} \times \{s_Z^2 : |F(s_Z^2) - .5| \leq F(M^2) - .5|\}$$

for predicting $(\bar{Z}, s_Z^2)$. This approach effectively "integrates" out $\bar{Z}$ or the so-called nuisance parameter μ and results in a marginal IM for inference about σ^2.

Before we discuss inference about μ alone with σ^2 viewed as a "nuisance" parameter, we consider the general case of inferring $\phi = \phi(\theta)$ alone. In what follows, we use a class of nuisance parameters. To be specific, let $(\phi, \xi) = (\phi(\theta), \xi(\theta))$ be a one-to-one mapping from the space Θ to the space of (ϕ, ξ). The approach in the above example of inferring σ^2 is generalized in Martin, Hwang, and Liu (2011b) as follows. Suppose that the baseline association model can be written as a (regular) system of two equations: a marginal association model

$$h(x, \phi) = m(\psi(U), \phi) \tag{3.8}$$

and a nuisance association model

$$c(U, x, \xi) = 0, \tag{3.9}$$

where U is the a-variable. Under the assumption that for any x and U, there exists ξ such that $c(U, x, \xi) = 0$ holds, then the baseline association model is equivalent to the marginal association model (3.8) for inference about ϕ. Thus, under mild conditions a valid IM for inference about ϕ, called a marginal IM (MIM), is obtained.

Inferential Model 3.2 (Marginal IM)**.** *The marginal IM for inference about ϕ has the following three steps.*

A-STEP. *Obtain the candidate set of ϕ from (3.8),*

$$\Phi_x(\psi(U)) = \{\phi : h(x, \phi) = m(\psi(U), \phi)\}.$$

P-STEP. *Construct a credible PRS $\mathcal{S}_\phi$ to predict $\psi(U)$.*

C-STEP. *Compute for any assertion of interest $\mathcal{A}$ on ϕ the evidence functions*

$$\underline{\mathsf{e}}_x(\mathcal{A}) = Pr(\Phi_x(\mathcal{S}) \subseteq \mathcal{A})$$

and

$$\overline{\mathsf{e}}_x(\mathcal{A}) = 1 - Pr(\Phi_x(\mathcal{S}) \subseteq \mathcal{A}^c),$$

where $\Phi_x(\mathcal{S}) = \{\phi : h(x,\phi) = a_1(\psi(U),\phi), \psi(U) \in \mathcal{S}_\phi\}$.

It is easy to see that the above example of inferring σ^2 is an instance of MIM. For another illustrative example, consider marginal inference about μ. Write the baseline association model (3.7) as a regular system of a marginal association model

$$\frac{\sqrt{n}(\bar{x}-\mu)}{s_x} = \frac{\sqrt{n}\bar{Z}}{s_Z} \tag{3.10}$$

and a nuisance association model

$$s_x^2 - \sigma^2 s_Z^2 = 0.$$

Thus, the MIM for μ has the A-step obtained from (3.10), the P-step predicting

$$T = \frac{\sqrt{n}\bar{Z}}{s_Z} \sim t(0,1,n-1),$$

and the standard C-step, where $t(0,1,n-1)$ stands for the standard Student-t distribution with $n-1$ degrees of freedom. The marginal association model

$$\frac{\sqrt{n}(\bar{x}-\mu)}{s_x} = T$$

with the a-variable following $t(0,1,n-1)$ gives IM results that are similar to familiar frequentist results, except that IM results have the desired interpretation for situation-specific uncertainty assessment.

Remark 3.3. *There are cases where regular MIMs may not exist. The Behrens-Fisher problem, a popular statistical problem in bio-medical statistics and elsewhere for inference about the difference of two normal population means with unknown variances, provides such an example. While more research is expected to be done on this interesting topic, Martin, Hwang, and Liu (2011b) proposed a parameter-expansion approach that leads to what they call weak marginal inferential models (WMIMs). With parameter-expansion, they gave an satisfactory IM solution to the Behrens-Fisher problem, a famous benchmark example in statistical inference.*

Remark 3.4. *To help understand Remark 3.2, it is interesting to note that from a frequentist point of view, marginal IM inference can also be considered as Fisherian conditional inference. For example, IM inference about μ gives similar frequentist results obtained by conditioning on s_x^2 (and direction of the residual vector). For IM inference, all observed data are considered as fixed and the efforts are made in predicting the unobserved a-variables in the particular experiment where the observed data were collected.*

4. Constrained Inferential Models

4.1. Constrained Inferential Models

Here we discuss an important issue that occurs when the combination of the PRS $\mathcal{S}$ and the candidate set $\Theta_x(z)$ in the C-step of IMs is empty. We call this problem the constrained inferential model problem. It contains constrained-parameter problems as special cases, but appears to be more general from a constrained inferential model perspective.

As is often the case in IM applications, realizations of $\mathcal{S}$ are here assumed to be *monotone* in the sense that any two realizations S_1 and S_2 of $\mathcal{S}$ satisfy either $S_1 \subseteq S_2$ or $S_1 \supseteq S_2$. A simple way of dealing with the constrained inferential model problem is to discard realizations of $\mathcal{S}$ that do not intersect with $\Theta_x = \cup_z \Theta_x(z)$. Ermini Leaf and Liu (2011) show that the resulting IMs are indeed valid. This idea can be understood as that the discarded realizations are effectively replaced with larger realizations. Computationally, this approach can be implemented using an acceptance-rejection sampling scheme that rejects draws of $\mathcal{S}$ that are too small to intersect with Θ_x. Intuitively, replacing small realizations of $\mathcal{S}$ with large ones maintains the validity, but results in loss of efficiency. To overcome this problem, Ermini Leaf and Liu (2011) proposed what they call the elastic PRS. The basic idea is to replace an overly small realization with one that is large enough rather than throwing it away.

Loosely speaking, the elastic version of a regular PRS $\mathcal{S}$ is a collection of random sets each containing $\mathcal{S}$ as a subset. More precisely, let $\mathbb{S}$ be the sample space of a regular $\mathcal{S}$. Then the elastic PRS can be defined by

$$\mathsf{w}(\mathcal{S}) = \{W : \mathcal{S} \subseteq W \in \mathbb{S}\} \qquad (\mathcal{S} \in \mathbb{S}). \tag{4.1}$$

The constrained IM is obtained by replacing the intersection $\Theta_x(\mathcal{S})$ of the C-step of an IM with

$$\Theta_x(\mathcal{S}) = \cup_{z \in \mathcal{S}^*} \Theta_x(z), \tag{4.2}$$

where $\mathcal{S}^*$ is the "smallest" element in $\mathsf{w}(\mathcal{S})$ that intersects with $\Theta_x = \cup_z \Theta_x(z)$, i.e.,

$$\mathcal{S}^* = \cap_{W \in \mathcal{W}^*} W \quad \text{with} \quad \mathcal{W}^* = \{W : W \in \mathsf{w}(\mathcal{S}), W \cap \Theta_x \neq \emptyset\}.$$

The use of constrained IMs is illustrated below by its application in classification. More applications can be found in Ermini Leaf and Liu (2011); see also Ermini Leaf (2011) and Section 7.

4.2. A classification problem

Here we consider the simple problem of classifying an observation to one of two univariate Gaussian subpopulations. For conceptual clarity, we discuss the case with known populations in Section 4.2.1. The more practical version, with unknown populations, is considered in Section 4.2.2.

4.2.1. Classification with known populations

Suppose that an observation $z \in \mathbb{R}$ is known to have come from either $N(\mu_1, \sigma_1^2)$ or $N(\mu_2, \sigma_2^2)$. It is assumed that there is no prior knowledge about which of the two subpopulations this observation z belongs to, but μ_1, σ_1, μ_2, and σ_2 are all known. Thus we consider the association model

$$z = \mu_i + \sigma_i U \qquad (U \sim N(0,1); i = 1, 2), \tag{4.3}$$

where U is the a-variable and i is the unknown quantity. This association model gives the A-step with

$$\Theta_z(u) = \{i : z = \mu_i + \sigma_i U\}.$$

We use the centered PRS

$$\mathcal{S} = \{u : |u| \leq |U|\} \qquad (U \sim N(0,1))$$

to predict $u^\star$, the unobserved realization of U associated with the observed data z. It is easy to see that the PRS $\mathcal{S}$ and the candidate set Θ_z may have an empty intersection. Equipping the PRS with elasticity, we have the constrained IM that, computationally, acts like a regular IM with the (conditionally enlarged) PRS

$$\mathcal{S} = \left\{ u : |u| \leq \max\left(|U|, \min_{i=1,2} \frac{|z - \mu_i|}{\sigma_i} \right) \right\} \qquad (U \sim N(0,1)).$$

The assertions of interest regarding the classification problem are $\mathcal{A}_1 = \{i = 1\}$ and $\mathcal{A}_2 = \{i = 2\}$. Let $k = \arg\min_{i=1,2} |z - \mu_i|/\sigma_i$. Then routine algebraic operations give for $i = 1, 2$

$$\underline{\mathsf{e}}_z(\mathcal{A}_i) = \begin{cases} 1 - 2\Phi\left(-\frac{|z-\mu_k|}{\sigma_k}\right), & \text{if } i = k; \\ 0, & \text{otherwise,} \end{cases}$$

and

$$\overline{\mathsf{e}}_z(\mathcal{A}_i) = \begin{cases} 1, & \text{if } i = k; \\ 2\Phi\left(-\frac{|z-\mu_i|}{\sigma_i}\right), & \text{otherwise.} \end{cases}$$

4.2.2. Classification with unknown populations

Consider the classification problem of Section 4.2.1 but with unknown parameters μ_1, σ_1^2, μ_2, and σ_2^2. Instead, it is assumed that two samples, $x_1, ..., x_{n_1}$, from $N(\mu_1, \sigma_1^2)$ and $y_1, ..., y_{n_2}$ from $N(\mu_2, \sigma_2^2)$, are available. The problem of inference remains the same, i.e., inferring about the unknown quantities regarding which subpopulation an observation z belongs to.

The discussion of Section 4.2.1 motivates us to consider inference about $(z - \mu_i)/\sigma_i$ for classification. The baseline association model is given by

$$z = \mu_i + \sigma_i U \qquad (U \sim N(0,1)),$$

$$x_i = \mu_1 + \sigma_1 V_i^{(x)} \qquad (V_i^{(x)} \overset{iid}{\sim} N(0,1); i = 1, ..., n_1),$$

and

$$y_i = \mu_2 + \sigma_2 V_i^{(y)} \qquad (V_i^{(y)} \overset{iid}{\sim} N(0,1); i = 1, ..., n_2),$$

where $U, (V_1^{(x)}, ..., V_{n_1}^{(x)})$ and $(V_1^{(y)}, ..., V_{n_2}^{(y)})$ are independent of each other. Applying the standard conditioning approach yields the reduced association model

$$\bar{x} = \mu_1 + \sigma_1 \bar{V}_x \quad \text{and } s_x = \sigma_1 M_1 \tag{4.4}$$

and

$$\bar{y} = \mu_2 + \sigma_2 \bar{V}_y \quad \text{and } s_y = \sigma_2 M_2, \tag{4.5}$$

where $\bar{x}$ and s_x are the sample mean and standard deviation of $x_1, ..., x_{n_1}$, $\bar{V}_x$ and M_1 are the "sample" mean and standard deviation of $U_1, ..., U_{n_1}$. The quantities $\bar{y}$, s_y, $\bar{V}_y$, and M_2 are introduced similarly but for $y_1, ..., y_{n_2}$ and the associated a-variables $V_1^{(y)}, ..., V_{n_2}^{(y)}$. The four a-variables $\bar{V}_x$, $\bar{V}_y$, M_1, and M_2 are independent with

$$\bar{V}_x \sim N(0, 1/n_1),\ \bar{V}_y \sim N(0, 1/n_2),\ (n_1{-}1)M_1^2 \sim \chi^2_{n_1-1}, \text{ and } (n_2{-}1)M_2^2 \sim \chi^2_{n_2-1}.$$

For inference about $\delta_i \equiv (z - \mu_i)/\sigma_i$ for $i = 1, 2$, write (4.4) as

$$\frac{z - \bar{x}}{s_x} = \frac{\delta_1 + \bar{V}_x}{M_1} \quad \text{and } s_x = \sigma_1 M_1 \tag{4.6}$$

and (4.5) as

$$\frac{z - \bar{y}}{s_y} = \frac{\delta_2 + \bar{V}_y}{M_2} \quad \text{and } s_y = \sigma_2 M_2. \tag{4.7}$$

For efficient inference, marginalizing out σ_1 and σ_2 and noticing that $\delta_i = U$ leave the dimension-reduced association model for classification

$$\frac{z - \bar{x}}{s_x\sqrt{1 + \frac{1}{n_1}}} = \frac{U + \bar{V}_x}{M_1\sqrt{1 + \frac{1}{n_1}}} \quad \text{or} \quad \frac{z - \bar{y}}{s_y\sqrt{1 + \frac{1}{n_2}}} = \frac{U + \bar{V}}{M_2\sqrt{1 + \frac{1}{n_2}}}.$$

Let $F_1(.)$ denote the cdf of $(U + \bar{V}_x)/(M_1\sqrt{1 + 1/n_1})$, the standard Student-t with $n_1 - 1$ degrees of freedom, and, likewise, let $F_2(.)$ denote the cdf of $(U + \bar{V}_y)/(M_2\sqrt{1 + 1/n_2})$, the Student-t with $n_2 - 1$ degrees of freedom. Thus, via a change-of-variable we can write the reduced association model as

$$\frac{z - \bar{x}}{s_x\sqrt{1 + \frac{1}{n_1}}} = F_1^{-1}(U) \quad \text{or} \quad \frac{z - \bar{y}}{s_y\sqrt{1 + \frac{1}{n_2}}} = F_2^{-1}(U) \qquad (U \sim \text{Unif}\,(0,1)) \tag{4.8}$$

with U playing the role of the a-variable.

Technically, classification based on the reduced association model (4.8) with the uniform a-variable U is the same as that with known subpopulations discussed in Section 4.2.1. Let $k = \arg\min_{i=1,2} |F_i - .5|$, where $F_1 = F_1\left((z - \bar{x})/(s_x\sqrt{1 + \frac{1}{n_1}})\right)$

and $F_2 = F_2\left((z-\bar{y})/(s_y\sqrt{1+\frac{1}{n_2}})\right)$. Take the centered PRS for predicting U and equip it with elasticity. The evidence functions for the assertions $\mathcal{A}_i = \{i\}$ in classification are given by

$$\underline{\mathsf{e}}_{z,x,y}(\mathcal{A}_i) = \begin{cases} 1 - 2F_k\left(-\frac{|z-\mu_k|}{\sigma_k\sqrt{1+\frac{1}{n_k}}}\right), & \text{if } i = k; \\ 0, & \text{otherwise,} \end{cases}$$

and

$$\overline{\mathsf{e}}_{z,x,y}(\mathcal{A}_i) = \begin{cases} 1, & \text{if } i = k; \\ 2F_i\left(-\frac{|z-\mu_i|}{\sigma_i\sqrt{1+\frac{1}{n_i}}}\right), & \text{otherwise} \end{cases}$$

for $i = 1, 2$.

5. Selective Inference

In this section, we consider a situation where the hypothesis of interest is not determined until after observing the data; see Ermini Leaf (2011) for more details. Suppose that a population has two subgroups: female and male. A treatment is administered to a sample of size n from each group. Without considering subgroups, the entire sample might not have significant evidence for a positive treatment effect in the entire population. However, when looking separately at each subgroup mean, there might be evidence for a positive treatment effect within one subgroup. Let $\overline{y}_{\mathrm{F}}$ and $\overline{y}_{\mathrm{M}}$ be the mean observed treatment effects for each group. Suppose further that $\overline{y}_{\mathrm{F}}$ comes from a $\mathrm{N}(\mu_{\mathrm{F}}, \sigma^2/n)$ distribution and, similarily, $\overline{y}_{\mathrm{M}}$ comes from a $\mathrm{N}(\mu_{\mathrm{M}}, \sigma^2/n)$. The standard deviation, $\sigma > 0$ is assumed to be known. Suppose it is observed that

$$\overline{y}_{\mathrm{F}} > \overline{y}_{\mathrm{M}}. \tag{5.1}$$

This suggests there might be evidence for the hypothesis that $\mu_{\mathrm{F}} > 0$. An IM for this hypothesis is preseneted in this section.

Bias can occur when a hypothesis is tested after observing data that suggests it is true. From a frequentist perspective, if the hypothesis is actually false, *i.e.*, $\mu_{\mathrm{F}} \leq 0$, and we only test it when $\overline{y}_{\mathrm{F}} > \overline{y}_{\mathrm{M}}$ is observed, then we would decide that the hypothesis is true too often over repeated experiments. The inference would not be valid. The way to overcome this bias is to condition the inference procedure on the event that led to the choice of hypothesis. In this case, the event (5.1) will be incorporated into the IM.

The association model can be built using the conditional cdf of each sample mean. Let

$$u_{\mathrm{M}} = \frac{\int_{-\infty}^{\overline{y}_{\mathrm{M}}} \frac{\sqrt{n}}{\sigma}\phi(\frac{t-\mu_{\mathrm{M}}}{\sigma/\sqrt{n}})\left(1 - \Phi(\frac{t-\mu_{\mathrm{F}}}{\sigma/\sqrt{n}})\right)\,dt}{\Phi\left(\frac{\mu_{\mathrm{F}}-\mu_{\mathrm{M}}}{\sqrt{2\sigma^2/n}}\right)} \tag{5.2}$$

and

$$u_{\mathrm{F}} = \frac{\Phi(\frac{\overline{y}_{\mathrm{F}}-\mu_{\mathrm{F}}}{\sigma/\sqrt{n}}) - \Phi(\frac{\overline{y}_{\mathrm{M}}-\mu_{\mathrm{F}}}{\sigma/\sqrt{n}})}{1 - \Phi(\frac{\overline{y}_{\mathrm{M}}-\mu_{\mathrm{F}}}{\sigma/\sqrt{n}})} \tag{5.3}$$

for $\overline{y}_{\mathrm{F}} > \overline{y}_{\mathrm{M}}$. The a-variables, u_{M} and u_{F}, each have a Unif$(0,1)$ distribution. However, the association model does not have a closed form expression.

The mean treatment effect in the male population subgroup, μ_{M}, is a nuisance parameter. It plays no role in the assertion of interest: $\mathcal{A} = \{\mu_{\mathrm{F}} : \mu_{\mathrm{F}} > 0\}$. A marginal IM (see Section 3.2) can be created by predicting u_{M} with the largest possible PRS, $[0,1]$. Since $\mathcal{A}$ is one-sided, the PRS for u_{F} can also be one-sided. The PRS for $(u_{\mathrm{F}}, u_{\mathrm{M}})$ is:

$$\mathcal{S} = \{u : u \geq U\} \times [0,1], \quad (U \sim \mathrm{Unif}(0,1))$$

The C-step requires the mapping (5.3) to be inverted with respect to μ_{F} for fixed u_{F}. This can be done numerically using a bracketing method (see, *e.g.*, chapter 2 of Givens and Hoeting, 2005), but technical details are omitted here.

Fig. 2 shows the one-sided 95% plausibility interval for μ_{F} with different values of $\overline{y}_{\mathrm{F}}$ when $\overline{y}_{\mathrm{M}} = -1.5$. The upper bound of the usual one-sided confidence interval,

$$\overline{y}_{\mathrm{F}} + 1.645\sigma/\sqrt{n},$$

is also shown. When $\overline{y}_{\mathrm{F}} \gg \overline{y}_{\mathrm{M}}$, the plausibility interval bound approaches the confidence interval bound. However, as $\overline{y}_{\mathrm{F}}$ approaches $\overline{y}_{\mathrm{M}}$ the plausibility interval bound drops to $-\infty$. The plausibility interval has exact 0.95 coverage probability for fixed θ over repeated experiments with $\overline{y}_{\mathrm{F}} > \overline{y}_{\mathrm{M}}$. The confidence interval is too wide.

6. Inference for binomial data

This section considers inference about the parameters of a binomial distribution, $Bin(n,\theta)$, based on an observed binomial count, x. First, we simply consider inference about the success probability, θ. This idea is then extended to inference about the odds ratio in a 2×2 contingency table.

6.1. Inference about a Binomial proportion

Following Martin et al. (2011a), suppose that the observed data x follows binomial distribution $Bin(n,\theta)$ with known size n and unknown parameter θ. We are interested in inference about θ. In the A-step, we associate x and the unknown θ with an a-variable u as follows:

$$x = \min\{k : 1 - u < F_{n,\theta}(k)\} \ \ (u \sim Unif(0,1)), \tag{6.1}$$

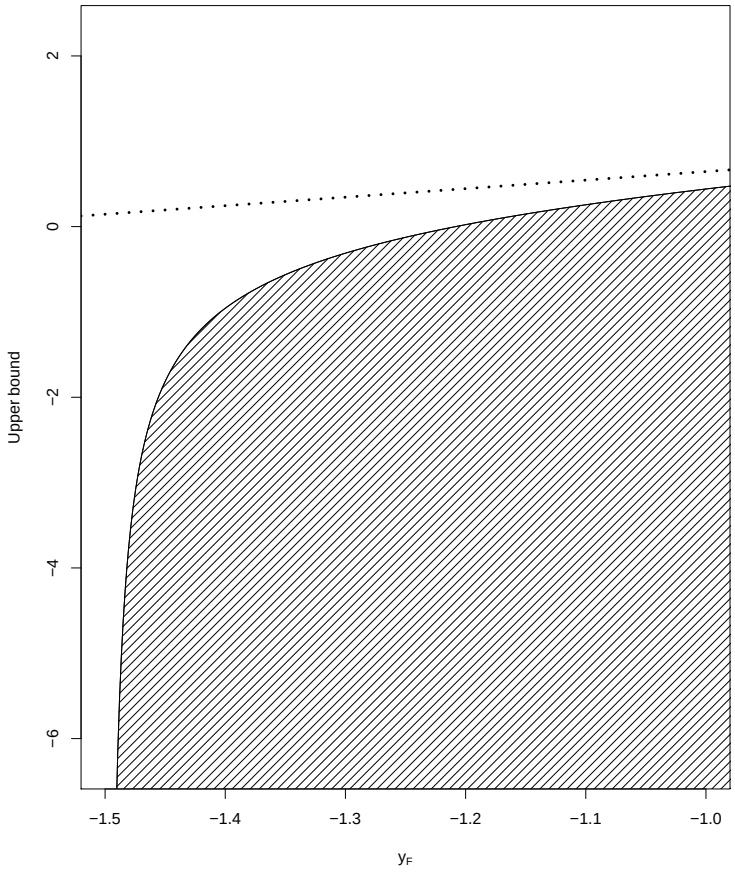

FIG 2. *Lower level-0.95 plausibility interval (shaded) for* μ_F *when* $y_M = -1.5$ *and* $\sigma/\sqrt{n} = 1$. *The dotted line is the upper bound of the usual one-sided confidence interval.*

where $F_{n,\theta}(\cdot)$ is the distribution function for the binomial distribution $Bin(n, \theta)$.

To complete the A-step, we derive $\Theta_x(u)$. From (6.1), we have

$$F_{n,\theta}(x-1) \leq 1-u < F_{n,\theta}(x). \tag{6.2}$$

Let $G_{a,b}(\cdot)$ be the distribution function of beta distribution with parameters a and b. The distribution function of the binomial distribution is related to that of beta distribution by the following equation

$$F_{n,\theta}(x) = G_{n-x,x+1}(1-\theta) = 1 - G_{x+1,n-x}(\theta), \tag{6.3}$$

which can be shown using integration by parts. Therefore, (6.2) can be rewritten as

$$G^{-1}_{x,n-x+1}(u) \leq \theta < G^{-1}_{x+1,n-x}(u). \tag{6.4}$$

Therefore,

$$\Theta_x(u) = \{\theta : G^{-1}_{x,n-x+1}(u) \leq \theta < G^{-1}_{x+1,n-x}(u)\}. \tag{6.5}$$

Note that unlike previous examples, $\Theta_x(u)$ is *not* single-valued due to the discrete nature of the sampling distribution.

For the P-step, u can be predicted by the PRS

$$\mathcal{S}_\theta = \{u : |u - 0.5| \leq |U - 0.5|\} \quad (U \sim Unif(0,1)). \tag{6.6}$$

Finally, in the C-Step $\Theta_x(u)$ and $\mathcal{S}_\theta$ are combined to obtain

$$\Theta_x(\mathcal{S}_\theta) = [G^{-1}_{x,n-x+1}(0.5 - |U - 0.5|), G^{-1}_{x+1,n-x}(0.5 + |U - 0.5|)), \tag{6.7}$$

with $U \sim Unif(0,1)$.

6.2. 2 by 2 Contigency Table

Now let us consider the more complicated case of observing counts from two independent binomial distributions,

$$x_1 \sim Bin(n_1, \phi_1) \quad \text{and} \quad x_2 \sim Bin(n_2, \phi_2), \tag{6.8}$$

when we are interested in inference about the odds ratio: $\theta = \frac{\phi_1}{1-\phi_1} / \frac{\phi_2}{1-\phi_2}$. Assume the sizes, n_1 and n_2, are known but parameters ϕ_1 and ϕ_2 are not known. When we are particularly interested in whether $\theta = 1$, it becomes the classical problem of testing association between two kinds of classification.

For the A-step, let $F_{n_1,\phi_1,n_2,\phi_2}(\cdot)$ be the distribution function of the sum of x_1 and x_2 so that

$$F_{n_1,\phi_1,n_2,\phi_2}(k) = \sum_{x_1+x_2 \leq k} \binom{n_1}{x_1} \phi_1^{x_1}(1-\phi_1)^{n_1-x_1} \binom{n_2}{x_2} \phi_2^{x_2}(1-\phi_2)^{n_2-x_2}. \tag{6.9}$$

A CIM can be built using an a-equation for $z = x_1 + x_2$:

$$z = \min\{k : v_2 < F_{n_1,\phi_1,n_2,\phi_2}(k)\} \quad (v_2 \sim Unif(0,1)). \tag{6.10}$$

Given z from the a-equation (6.10), x_1 follows Fisher's noncentral hypergeometric distribution. Thus, letting $H_{n_1,n_2,z,\theta}(\cdot)$ be the distribution function of x_1,

$$x_{\min} = \max(0, z - n_2), \tag{6.11}$$

$$x_{\max} = \min(z, n_1), \tag{6.12}$$

$$H_{n_1,n_2,z,\theta}(k) = \frac{\sum_{x_{\min} \leq x_1 \leq k} \binom{n_1}{x_1}\binom{n_2}{z-x_1}\theta^{x_1}}{\sum_{x_{\min} \leq x_1 \leq x_{\max}} \binom{n_1}{x_1}\binom{n_2}{z-x_1}\theta^{x_1}}, \tag{6.13}$$

the conditional association model can be derived as:

$$x_1 = \min\{k : v_1 < H_{n_1,n_2,z,\theta}(k)\} \quad (v_1 \sim Unif(0,1)). \tag{6.14}$$

Note that unlike (3.5), the conditioning equation is not free of parameters ϕ_1 and ϕ_2. However, we can still conduct a conditional inference based on (6.14) using generalized conditional IM approach (Martin et al., 2011a). Thus, the conditional association model can be simplified as:

$$\Theta_{x_1}(v_1) = \{\theta : H_{n_1,n_2,z,\theta}(x_1 - 1) < v_1 \leq H_{n_1,n_2,z,\theta}(x_1)\}. \tag{6.15}$$

For the P-step, we can predict v_1 by the marginal PRS

$$\mathcal{S}_\theta = \{v_1 : |v_1 - 0.5| \leq |V_1 - 0.5|\} \ \ (V_1 \sim Unif(0,1)). \tag{6.16}$$

In the C-step, although $\Theta_{x_1}(v_1)$ cannot be described as simply as in (6.7), for the assertion $\mathcal{A}_{\theta_0} = \{\theta_0\}$, it is possible to derive a simple expression for $\underline{e}(\mathcal{A}_{\theta_0})$ and $\overline{e}(\mathcal{A}_{\theta_0})$. Trivially, $\underline{e}(\mathcal{A}_{\theta_0}) = 0$ for any point assertion $\mathcal{A}_{\theta_0}$, since (6.15) is multiple-valued almost surely, thus $\Theta_{x_1}(v_1) \not\subseteq \mathcal{A}_{\theta_0}$ almost surely. Also, from (6.15) we have

$$\Theta_{x_1}^{-1}(\theta_0) = [a_{x_1}(\theta_0), b_{x_1}(\theta_0)). \tag{6.17}$$

where

$$a_{x_1}(\theta_0) = \frac{\binom{n_1}{x_1-1}\binom{n_2}{z-(x_1-1)}\theta^{x_1-1}}{\sum_{x_{\min}\leq t\leq x_{\max}}\binom{n_1}{t}\binom{n_2}{z-t}\theta^t},$$

and

$$b_{x_1}(\theta_0) = \frac{\binom{n_1}{x_1}\binom{n_2}{z-x_1}\theta^{x_1}}{\sum_{x_{\min}\leq t\leq x_{\max}}\binom{n_1}{t}\binom{n_2}{z-t}\theta^t}.$$

Therefore, when PRS in (6.16) is used,

$$\overline{\mathsf{e}}_x(\mathcal{A}_{\theta_0}) = \begin{cases} 1 & \text{if } a_{x_1}(\theta_0) \leq 0.5 \leq b_{x_1}(\theta_0); \\ 2(0.5 - b_{x_1}(\theta_0)) & \text{if } b_{x_1}(\theta_0) < 0.5; \\ 2(a_{x_1}(\theta_0) - 0.5) & \text{if } a_{x_1}(\theta_0) > 0.5. \end{cases}$$

For a concrete example, let us consider the case of Kidney stone treatment from Julious and Mullee (1994). Table 1 describes the result of applying treatment A to patients with small stones (n_1) and large stones (n_2), where x_1 and x_2 are the number of successful treatments for each group respectively. Fig. 3 plots the log of $\overline{\mathsf{e}}_x(\mathcal{A}_{\theta_0})$ versus the log of θ_0. One can find a level $1-\alpha$ plausibility interval, by locating α on the vertical axis and choosing all the θ_0 values for which $\overline{\mathsf{e}}_x(\mathcal{A}_{\theta_0}) > \alpha$.

Note that conducting conditional inference based on (6.15) is similar to Fisher's exact test. In this respect, the IM approach provides another way of viewing and validating Fisher's exact test and conditional likelihood methods.

$x_1 = 81$	$n_1 - x_1 = 6$	$n_1 = 87$
$x_2 = 192$	$n_2 - x_2 = 71$	$n_2 = 263$
$x_1 + x_2 = 273$	$n_1 + n_2 - x_1 - x_2 = 77$	$n_1 + n_2 = 350$

TABLE 1
Kidney stone experiment data

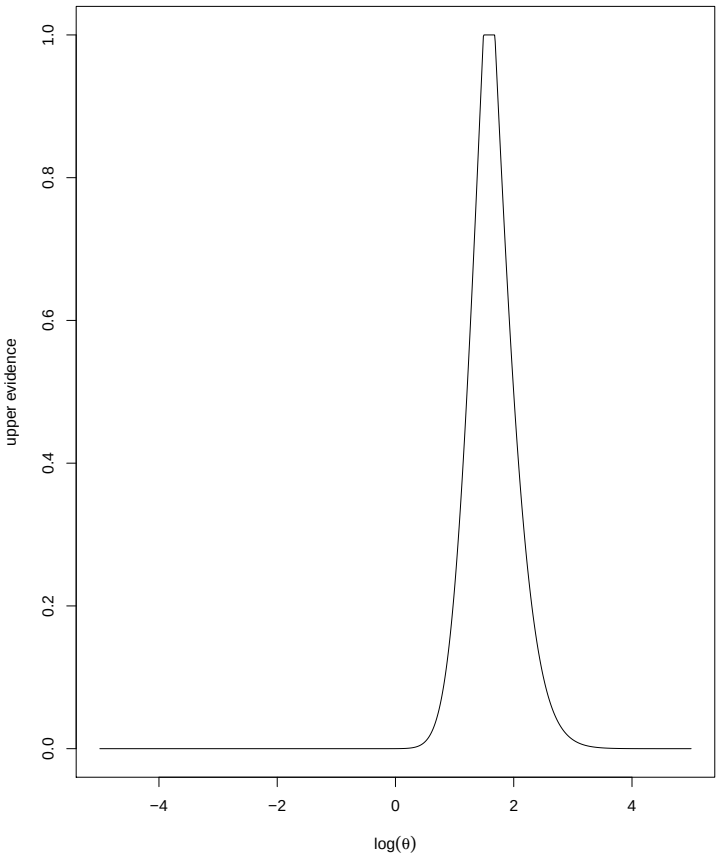

FIG 3. *Plot of upper evidence function for each point assertion* $\mathcal{A}_{\theta_0}$.

TABLE 2
The data of Rubin (1981) on SAT coaching experiments

School (i)	Estimated Treatment-Effects (y_i)	Associated Standard-Error (s_i)
1	28.39	14.9
2	7.94	10.2
3	-2.75	16.3
4	6.82	11.0
5	-0.64	9.4
6	0.63	11.4
7	18.01	10.4
8	12.16	17.6

7. Meta-Analysis: many-normal-means problems

We refer to Zhou et al. (2003) for a comprehensive review of meta-analysis. Here we focus on a popular statistical model for meta-analysis, namely, the many-normal-means model

$$y_i \overset{ind}{\sim} N(\theta_i, s_i^2) \qquad (i = 1, ..., n), \tag{7.1}$$

where y_i is the estimated or observed treatment effect in the i-th study. The mean, θ_i, is the unknown treatment effect to be inferred. The associated standard deviation s_i^2 is taken to be known, assuming that the sample sizes in all the studies are not too small.

For a numerical example of our exposition, we take the real-data example of Rubin (1981), which has the same data structure and inference problems. Rubin (1981) considers assessing effects of SAT coaching programs based on parallel randomized experiments in $n = 8$ schools. The estimated individual effects and the associated standard deviations are tabulated in Table 2. An argument for the use of the sampling model (7.1) is that each y_i was obtained from a data set that is large enough for the acceptance of the normality assumption and the assumption that s_i^2's are known.

Section 7.1 discusses a formal IM approach to uncertainty assessment of the assertion or hypothesis

$$\mathcal{A}_0 = \{\theta : \theta_1 = ... = \theta_n\}. \tag{7.2}$$

Section 7.2 considers the hierarchical model specified by imposing a second stage structure

$$\theta_i \overset{iid}{\sim} N(\mu, \sigma^2) \qquad (i = 1, ..., n) \tag{7.3}$$

with unknown $\mu \in \mathbb{R}$ and unknown variance σ^2. The resulting model is also known as a random-effects model. Inference with unknown σ^2 has been challenging. For

example, the method of maximum likelihood is problematic. Bayesian inference requires sensible operational priors to be carefully chosen for (μ, σ^2). This serves as a good example for the argument that practical Bayesian methods are approximating frequentist methods with resulting inference not necessarily valid in the sense of Section 2. After a brief discussion on an IM method for checking the postulated hierarchical model specified by (7.1) and (7.3), an IM approach to inference with the hierarchical model without requiring priors for (μ, σ^2) is presented in Section 7.2.

7.1. Inference about equal means or homogeneity

The sampling model (7.1) can be expanded for probabilistic inference as

$$y_i = \theta_i + s_i Z_i \qquad (Z_i \overset{iid}{\sim} N(0,1);\ i = 1, ..., n), \tag{7.4}$$

where $Z_1, ..., Z_n$ are the a-variables. The problem here is to assess the truth and falsity of the assertion $\mathcal{A}_0$ given in (7.2).

Note that the assertion $\mathcal{A}_0$ is a point assertion in an $(n-1)$-dimensional space. That is, the common value of θ_i under the truth of $\mathcal{A}_0$ can be located anywhere on the real line $\mathcal{R}$. This means that no precise inference is needed for this unknown common mean. Thus, we consider an MIM that effectively integrates out a nuisance parameter representing this common value. We simply take the MIM consisting of the nuisance association model

$$y_1 = \theta_1 + s_1 Z_1$$

and the marginal association model

$$y_i - y_1 = \phi_i + U_i \qquad (i = 2, ..., n), \tag{7.5}$$

where $\phi_i = \theta_i - \theta_1$ and $U_i = s_i Z_i - s_1 Z_1$ for $i = 2, ..., n$. Since for any y and Z, there is a θ_1 such that the nuisance association model holds, we proceed with the marginal association model (7.5) alone.

Write $D = (y_2 - y_1, ..., y_n - y_1)'$, $\phi = (\phi_2, ..., \phi_n)'$, and $U = (U_2, ..., U_n)'$. Then (7.5) can be written in a vector form as

$$D = \phi + U \qquad (U \sim N_{n-1}(0, W)), \tag{7.6}$$

where the covariance matrix of U is

$$W = \text{diag}(s_2^2, ..., s_n^2) + s_1^2 \mathbf{1}\mathbf{1}'$$

with $\mathbf{1} = (1, ..., 1)'$, denoting the vector of $(n-1)$ ones. In terms of ϕ, the assertion of interest $\mathcal{A}_0$ is given by

$$\mathcal{A}_0 = \{\phi : \phi_2 = ... = \phi_n = 0\}.$$

For this precise or point assertion, we consider a credible PRS in the $n-1$ dimensional space of U. It is seen from (7.6) that the propagation of uncertainty from U to ϕ is subject to only a translation operation (i.e., additive transformation). To create an efficient PRS for predicting U, we construct smallest possible subsets with fixed coverage probabilities, that is, the highest-density regions

$$\mathcal{S} = \{u : u'W^{-1}u \leq U'W^{-1}U\} \qquad (U \sim N_{n-1}(0, W))$$

or, equivalently,

$$\mathcal{S} = \{u : u'W^{-1}u \leq R^2\} \qquad (R^2 \sim \chi^2_{n-1}). \tag{7.7}$$

It should be noted that the projected one-dimensional predictive random intervals in any one-dimensional space have the same coverage probabilities. This observation, together with the use of high-density sets and the translation operation from U to ϕ, explains intuitively the efficiency of the PRS (7.7).

Routine operations, using, for example, familiar matrix identities that can be easily established with the sweep operator, give the following evidence functions

$$\underline{\mathsf{e}}_y(\mathcal{A}_0) = 0 \tag{7.8}$$

and

$$\overline{\mathsf{e}}_y(\mathcal{A}_0) = 1 - F_{n-1}\left(\sum_{i=1}^{n} w_i(y_i - \bar{y})^2\right), \tag{7.9}$$

where $w_i = 1/s_i^2$ for $i = 1, ..., n$, $\bar{y} = \sum_{i=1}^n w_i y_i / \sum_{i=1}^n w_i$, and $F_{n-1}(.)$ denotes the cdf of the chi-square distribution with $n-1$ degrees of freedom. The evidence for $\mathcal{A}_0$ from the data in Table 2 are

$$\underline{\mathsf{e}}_y(\mathcal{A}_0) = 0 \text{ and } \overline{\mathsf{e}}_y(\mathcal{A}_0) = 0.7131,$$

indicating no strong evidence is available for inference about the homogeneity of θ_i's.

We note that the results do not depend on the choice of $y_1 = \theta_1 + s_1 Z_1$ for the marginal inference. These results agree with the familiar classic results, but have the desirable interpretation for situation-specific inference.

7.2. A hierarchical model

Although there is no strong evidence *against* the assertion that the effects of all the coaching programs are the same, the evidence *for* this assertion is not strong either. Most importantly, it can be argued that these effects cannot be precisely the same. One purpose of the study is to evaluate how much difference there is among coaching programs, that is, the individual θ_i effects are viewed as random. A simple model for characterizing such an estimand is the hierarchical model given by (7.1) and (7.3):

$$y_i|\theta \overset{ind}{\sim} N(\theta_i, s_i^2) \text{ and } \theta_i \overset{iid}{\sim} N(\mu, \sigma^2).$$

Integrating out the random θ_i effects yields the sampling model for the observed data given by the association model

$$y_i = \mu + \sqrt{s_i^2 + \sigma^2} Z_i \qquad (Z_i \overset{iid}{\sim} N(0,1); i = 1, ..., n). \tag{7.10}$$

7.2.1. Model checking

In practice, it is necessary to check if this postulated model fits the observed data. A simple IM approach can be carried out by producing evidence for the assertion that *there exists some value of* (μ, σ^2) *such that*

$$\frac{y_i - \mu}{\sqrt{s_i^2 + \sigma^2}} \qquad (i = 1, ..., n)$$

is a sample from the standard normal distribution. The IM proposed for such an assertion in Liu and Xie (2010) can be applied here with a simple modification obtained by including the corresponding optimization over (μ, σ^2). We refer to Liu and Xie (2010) for the technical details regarding the essential part of this IM method. For the present example, we applied this IM method to the data in Table 2. The pair of lower and upper evidence for this model-checking assertion is $(0, 0.95)$, indicating that the model fits the data well. In what follows, we consider inference about σ^2 based on the association model (7.10) with the observed data $y_1, ..., y_n$.

7.2.2. A simple IM for inference about σ^2

Inference about σ^2 appears to be a challenging problem. It is easy to see that the $n = 2$ case and the $s_1 = ... = s_n$ case are simple constrained parameter problems. In the $n = 2$ case, marginalizing out μ gives the association model

$$y_2 - y_2 = \sqrt{s_1^2 + s_2^2 + 2\sigma^2} Z \qquad (Z \sim N(0,1))$$

with the unobserved a-variable Z to be predicted. This is effectively a constrained parameter problem because the unknown quantity $s_1^2 + s_2^2 + 2\sigma^2$ is subject to the constraint that $s_1^2 + s_2^2 + 2\sigma^2 \geq s_1^2 + s_2^2$. Inference about σ^2 in this case can be done using the method in Section 4. IM inference can also be carried out easily for the case where $s \equiv s_1 = ... = s_n$ by making inference about $\phi^2 = s^2 + \sigma^2$ subject to the constraint $\phi^2 \geq s^2$. For the general case, we present a simple IM approach below and discuss its efficiency; see Remark 7.1.

A simple IM approach is obtained by noticing the fact that

$$\sum_{i=1}^{n} \frac{(y_i - \bar{y}(\sigma^2))^2}{s_i^2 + \sigma^2} = S^2 \qquad (S^2 \sim \chi^2_{n-1}) \tag{7.11}$$

where S^2 is the a-variable and $\bar{y}(\sigma^2) = \sum_{i=1}^n w_i(\sigma^2) y_i / \sum_{i=1}^n w_i(\sigma^2)$ with $w_i(\sigma^2) = 1/(s_i^2 + \sigma^2)$ for $i = 1, ..., n$. Using a PRS to predict S^2 alone results in valid inference about σ^2; see Remark 7.1 for more discussion.

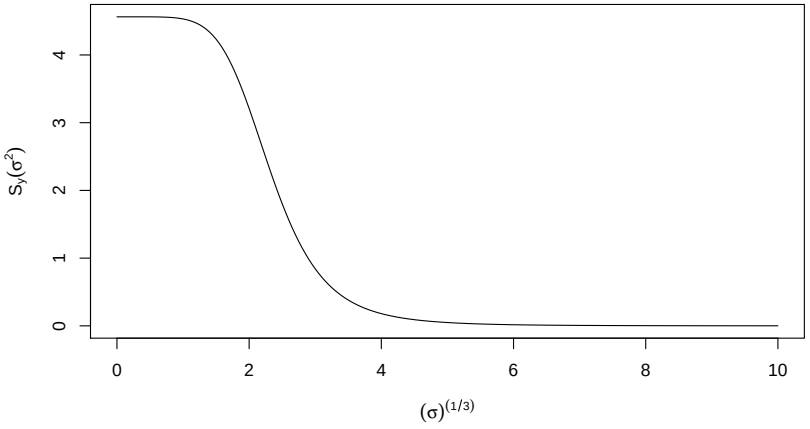

FIG 4. *The function $S_y(\sigma^2)$ for the SAT coaching example.*

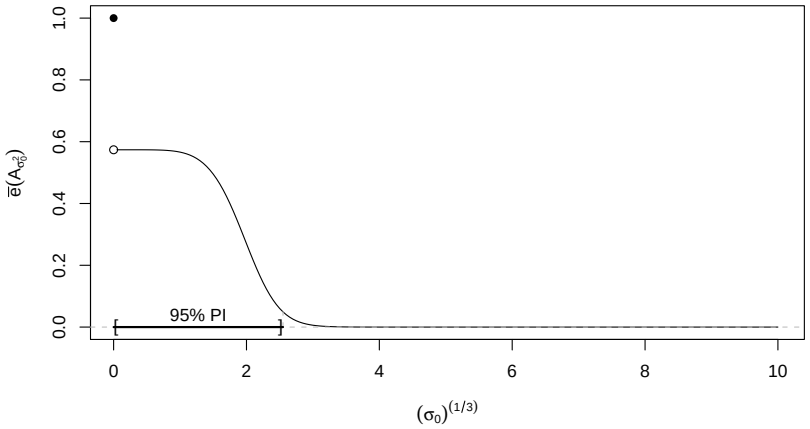

FIG 5. *The upper evidence function for the point assertion $\mathcal{A}_{\sigma_0^2} = \{\sigma_0^2\}$ and the 95% plausibility interval for σ,* $[0, 16.50]$.

Let

$$S_y(\sigma^2) = \sum_{i=1}^{n} \frac{(y_i - \bar{y}(\sigma^2))^2}{s_i^2 + \sigma^2}.$$

The function $S_y(\sigma^2)$ for the SAT data is displayed in Figure 4, which shows that $S_y(\sigma^2)$ is bounded from above. Thus the EPRS $\mathcal{S}$ for predicting S^2 is used to ensure that $\mathcal{S}$ has non-empty intersection with the interval $[0, \max_{\sigma^2} S_y(\sigma^2)]$. We use, for example, the equal-tails/two-sided PRS to build an EPRS $\mathcal{S}$. In this example, the median of χ^2_7 is 6.3458 and $\max_{\sigma^2} S_y(\sigma^2) = 4.5632$. Thus, EPRS $\mathcal{S}$ has a point mass of 0.4262 at $\mathcal{S} = [4.5632, 8.5460]$. The resulting upper evidence function for the sequence of point assertions

$$\mathcal{A}_{\sigma_0^2} = \{\sigma_0^2\}$$

is shown in Figure 5, where the 95% plausibility interval for σ is $[0, 16.50]$.

Remark 7.1. *Write*

$$V(y, \sigma^2) \equiv \left(\frac{y_1 - \bar{y}(\sigma^2)}{\sqrt{s_1^2 + \sigma^2}}, \ldots, \frac{y_n - \bar{y}(\sigma^2)}{\sqrt{s_n^2 + \sigma^2}} \right)'$$

and

$$U(Z, \sigma^2) \equiv \left(Z_1 - \frac{\tilde{Z}(\sigma^2)}{\sqrt{s_1^2 + \sigma^2}}, \ldots, Z_n - \frac{\tilde{Z}(\sigma^2)}{\sqrt{s_n^2 + \sigma^2}} \right)',$$

where

$$\tilde{Z}(\sigma^2) = \frac{\sum_{i=1}^n \frac{Z_i}{\sqrt{s_i^2+\sigma^2}}}{\sum_{i=1}^n \frac{1}{s_i^2+\sigma^2}}.$$

A formal argument for the validity of the simple approach in the last section can be made by writing the association model (7.10), after integrating out μ, as a system of equations consisting of (7.11), i.e.,

$$\|V(y, \sigma^2)\| = \|U(Z, \sigma^2)\|, \tag{7.12}$$

representing the length of the vector, and

$$\|V(y, \sigma^2)\|^{-1} V(y, \sigma^2) = \|U(Z, \sigma^2)\|^{-1} U(Z, \sigma^2), \tag{7.13}$$

for the direction of $V(y, \sigma^2)$. For the case of $s = s_1 = \ldots = s_n$, the unknown parameter σ^2 disappears from (7.13). This results in a simple conditional IM, from which we infer σ^2 based on the association (7.12) with $\|U(Z, \sigma^2)\|$ conditioned on the observed direction of $U(Z, \sigma^2)$ in (7.13). Since the length and direction of $U(Z, \sigma^2)$ are independent of each other, the simple approach is in fact based on a CIM and, thereby, is efficient. When the length and direction are not independent, one can ignore (7.13) or, equivalently, predict the direction of $U(Z, \sigma^2)$ with a (projected) vacuous PRS. This implies that although valid, efficiency may be gained by making use of (7.13). Research on developing general IM approaches for such problems, under the umbrella of the so-called generalized CIM (GCIM) approach, is on-going. Nevertheless, a Bayesian analysis is given below for indirectly verifying the efficiency of the above IM approach.

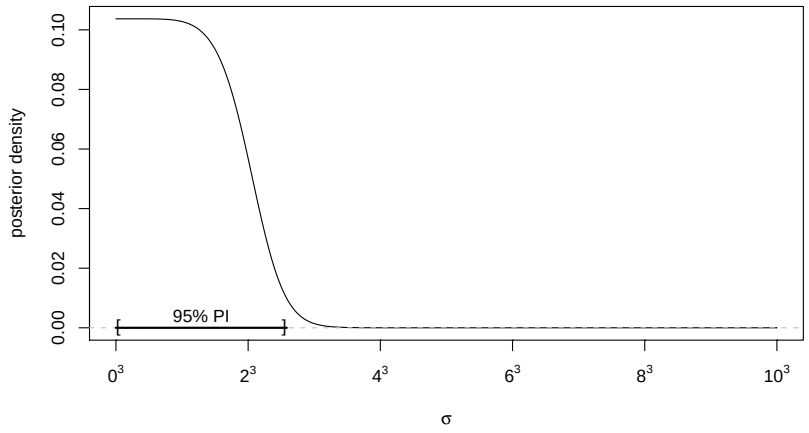

FIG 6. *The (marginal) Bayesian posterior density and 95% credible interval (CI),* $[0, 17.27]$*, for* σ*, obtained with the prior* $\pi(\mu, \sigma) \propto 1$.

For a comparison, consider the Bayesian inference using the (flat) prior distribution (see, e.g., Rubin (1981) and Gelman et al. (2005)):

$$\pi(\mu, \sigma) \propto \text{const.}$$

The posterior density of σ and the 95% Bayesian credible inference (CI) are shown in Figure 6. Note that the Bayesian CI, $[0, 17.27]$, is slightly longer than IM plausibility interval (PI), $[0, 16.50]$. It should also be noted that the posterior can be very sensitive to the specification of the prior for σ. For example, different results will be produced if the alternative prior $\pi(\mu, \sigma^2) \propto 1/(\sigma^2 + \min_{1 \le i \le n} s_i^2)$ is used.

References

1. Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. *Phil. Trans.*, **53**, 370–418.
2. Dempster, A. P. (1963). Further examples of inconsistencies in the fiducial argument. *Ann. Math. Statist.*, **34**, 884-891.
3. Dempster, A. P. (1964). On the difficulties inherent in Fisher's fiducial argument. *American Statistical Ass. J.*, **59**, 56-66.
4. Dempster, A. P. (2008). Dempster-Shafer calculus for statisticians. *Internat. J. of Approx. Reason.*, **48**, 265-277.

5. Efron, Bradley. (1998). R. A. Fisher in the 21st Century. *Statistical Science* **13**, 95-122.

6. Ermini Leaf, D. (2011). *Inferential Models and Restricted Spaces*. PhD thesis. Purdue University.

7. Ermini Leaf, D., Hui, J., and Liu, C. (2009). Statistical inference with a single observation of $N(\theta, 1)$. *Pakistan J. Statist.* **25**(4), 571–586.

8. Ermini Leaf, D., and Liu, C. (2011). Inference about constrained parameters using the elastic belief method. Submitted.

9. Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society*, A, **222** 309-368.

10. Fraser, D. A. S. (1966). On the definition of fiducial probability. *Bull. Int. Statist. Inst.*, **40**, 842-856.

11. Fraser, D. A. S. (1968). *The Structure of Inference*. Krieger, Huntington, New York.

12. Fraser, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence? (with discussion). to appear in *Statistical Science*.

13. Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.

14. Givens, G. H. and Hoeting, J. A. (2005). *Computational statistics.* Wiley Series in Probability and Statistics. Wiley-Interscience. Hoboken, NJ.

15. Hannig, J. (2009). On generalized fiducial inference. *Statist. Sinica* **19**, 491-544.

16. Liang, F., Liu, C. and Carroll, R.J. (2010). *Advanced Markov chain Monte Carlo: Learning from Past Samples*, Wiley.

17. Lindley, D. V. (1958). Fiducial distributions and Bayes theorem. *J. Roy. Statist. Soc. Ser.* B **20** 102-107.

18. Julious, S. A. and Mullee, M. A. (1994). Confounding and Simpson's paradox. *BMJ*, **309**, 1480-1481.

19. Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem. *J. Roy. Statist. Soc. Ser. B.* **20**, 102-107

20. Liu, C. and Martin. R. (2012). *Inferential Models: reasoning with uncertainty*, monograph in preparation.

21. Liu, C. and Xie, J. (2010). Probabilistic Inference for Multiple Testing. Technical Report, Purdue University.

22. Martin, R., Hwang, J.-S. and Liu, C. (2011a). Conditional inferential models. Working manuscript. Old version:
www.stat.purdue.edu/∼chuanhai

23. Martin, R., Hwang, J.-S. and Liu, C. (2011b). Marginal inferential models. Working manuscript. Old version: www.stat.purdue.edu/∼chuanhai

24. Martin, R. and Liu, C. (2011). Inferential models. Submitted, www.stat.purdue.edu/∼chuanhai

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

26. Rubin, D. B. (1981). Estimation in Parallel Randomized Experiments. *Journal of Educational Statistics* **6**, 377-401.

27. Savage, L. J. (1964). Discussion. *Bull. Inst. Internat. Statist.* **40**, 925-927.

28. Savage, L. J. (1976). On Rereading R. A. Fisher, *The Annals of Statistics*, **4**, 441-500.

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

30. Shafer, G. (2011). A betting interpretation for probabilities and Dempster–Shafer degrees of belief. *Journal International Journal of Approximate Reasoning*, **52**, 127-136.

31. Zabell, S. L. (1992) R. A. Fisher and the fiducial argument. *Statist. Sci.*, **7**(3), 369-387.

32. Zhang, J. and Liu, C. (2011). Dempster-Shafer inference with weak beliefs. *Statist. Sinica* **21**, 475-494.

33. Zhang, Z., Xu, H., Martin, R., and Liu, C. (2011). Inferential models for linear regression. To appear in *Pakistan Journal of Statistics and Operation Research*.

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

35. Zhou X, Fang J, Yu C, Xu Z, Lu Y. (2003). Meta analysis method, in *Advanced Medical Statistics*, 1st edition, eds. Lu Y, Fang J., River Edge, NJ; World Scientific.