

Possibilistic inferential models: a review

Ryan Martin*

November 12, 2025

Abstract

An *inferential model* (IM) is a model describing the construction of provably reliable, data-driven uncertainty quantification and inference about relevant unknowns. IMs and Fisher’s fiducial argument have similar objectives, but a fundamental distinction between the two is that the former doesn’t require that uncertainty quantification be probabilistic, offering greater flexibility and allowing for a proof of its reliability. Important recent developments have been made thanks in part to newfound connections with the imprecise probability literature, in particular, possibility theory. The brand of *possibilistic IMs* studied here are straightforward to construct, have very strong frequentist-like reliability properties, and offer fully conditional, Bayesian-like (imprecise) probabilistic reasoning. This paper reviews these key recent developments, describing the new theory, methods, and computational tools. A generalization of the basic possibilistic IM is also presented, making new and unexpected connections with ideas in modern statistics and machine learning, e.g., bootstrap and conformal prediction.

Keywords and phrases: Bayes; confidence distribution; fiducial; frequentist; imprecise probability; possibility theory; validity.

1 Introduction

An *inferential model* (IM) is a model for data-driven uncertainty quantification and inductive inference about relevant unknowns. These unknowns might be parameters in a posited statistical model, or functions thereof, but other situations are possible; see Section 6. More specifically, an IM offers mathematically rigorous, fully conditional, Bayesian-like uncertainty quantification—without requiring prior distributions or Bayes’s theorem—that is provably reliable in the frequentist-like sense that its output is naturally calibrated across repeated sampling. At a high level at least, this is reminiscent of what Fisher’s fiducial argument aimed to do, so it’s important to make clear IM’s novelty at the outset: there are limits to the reliability of probabilistic uncertainty quantification in the absence genuine prior information, and the IM framework sidesteps these limitations not by relaxing “reliability” but instead by relaxing “probabilistic,” by working in a more flexible (but still mathematically rigorous) framework that doesn’t require single or precise probability values to be assigned to each proposition about the unknowns. These

*Department of Statistics, North Carolina State University, rgmarti3@ncsu.edu

details will be explored in-depth in the sections that follow. For now, suffice it to say that this novel use of concepts and tools coming from the theory of *imprecise probability* in the context of statistical inference is surely what inspired Cui and Hannig (2024) to describe IMs as “one of the original statistical innovations of the 2010s.”

It’s been about 10 years since the monograph *Inferential Models: Reasoning with Uncertainty* (Martin and Liu 2015b) was published, and a lot of exciting developments have been made since then on all fronts: foundations, theory, methodology, and computation. An impetus behind these more recent advancements was the recognition that, while IMs’ deviation from ordinary probability theory puts one in unfamiliar territory, that new territory isn’t uninhabited—there’s an entire community of researchers and a body of literature devoted to imprecise probability theory ripe with important insights, understanding, and mathematical/computational tools. These efforts have inspired other developments (e.g., Caprio et al. 2025; Williams 2023; Xie and Wang 2022) and have created opportunities for new-and-improved understanding of Bayes and fiducial inference, bootstrap, conformal prediction, etc. The goal of this paper is to review these recent developments, which are related to but also different from those in the aforementioned monograph, and to present some new insights, methods, and results. I hope that this survey will make these exciting developments more accessible and draw a new generation of researchers’ attention to these foundational advancements and open problems.

Towards this goal, the remainder of the paper is organized as follows. Section 2 sets the scene by reviewing probabilistic uncertainty quantification and highlighting its shortcomings. In particular, Section 2.4 offers a new perspective on the *false confidence theorem* (Balch et al. 2019) establishing that any data-dependent posterior probability distribution—Bayes with any prior, (generalized) fiducial, etc.—will, in vacuous prior information cases like I’m considering here, tend to assign high probability or *confidence* to some false hypotheses. The implications of this are two-fold: that probabilistic uncertainty quantification has an inherent unreliability and that correcting this requires quantification of uncertainty using more flexible tools from the world of imprecise probability. Section 3 dives into the details of the *possibilistic* IM construction, where I’m emphasizing “possibilistic” because this new perspective—which is similar to that presented in Martin and Liu (2015b)—leans heavily on possibility-theoretic interpretations, calculus, computational tools, and mathematical structure. The IM’s key properties are presented next, including a finite-sample validity property which ensures that its imprecise-probabilistic output is properly calibrated and, in particular, that tests and confidence sets derived from the IM’s output have frequentist error rate guarantees. Efficiency is also discussed, including a new possibilistic *Bernstein–von Mises theorem* ensuring that the IM’s output is asymptotically possibilistic-Gaussian and efficient like in the classical sense where the asymptotic variance matches the Cramér–Rao lower bound. But IMs are not solely frequentist—they also offer fully conditional quantification of uncertainty and, in Section 3.4, I survey the relevant Bayesian-like properties that have not received much attention in previous works. IM computation is nontrivial, and Section 3.5 reviews briefly an exciting new development facilitating sampling-based Monte Carlo computations for evaluating the IM’s non-probabilistic output. Section 4 explains how IMs offer much more than a “unification” of frequentism and Bayesianism—the IM framework actually fills the holes in both paradigms! Section 5 deals with the important practical problem of eliminating nuisance parameters, and this motivates some new developments presented in

Section 6, which help IMs reach the next level of uncertainty quantification beyond those cases involving a statistical model. Applications of this idea to inference on risk minimizers and to (conformal) prediction are considered there. The paper ends in Section 7 with a brief summary, mentioning some relevant topics that are not discussed in this review, along with some open questions for future investigation. The appendix/supplement offers supporting technical details and some additional examples.

Some may argue that foundational efforts like these are impractical, but I disagree. For years we’ve heard that statistics as a discipline is at risk of “missing the boat” when it comes to data science; see, e.g., the recent report by He et al. (2025). Statisticians collectively agree that statistics is an important part of data science, so why are we so concerned? A boat’s captain isn’t concerned about missing the boat, neither are the first and second mates; only ancillary crew members and passengers worry about missing the boat. This fear of missing the data science boat betrays our community’s deep-down insecurity—a fear that we’re only ancillary crew members—which surely can’t be a consequence of us not contributing to enough applied projects or not proving enough consistency theorems. It must instead be due to a deeper, more significant shortcoming, say: “How can a discipline, central to science and to critical thinking, have two methodologies, two logics, two approaches that frequently give substantially different answers to the same problems” (Fraser 2011b). Our discipline can’t be confident about its contributions when such fundamental questions remains unanswered. Orienting our discipline by settling these foundational questions would prove that we have unique expertise to contribute, thereby giving us confidence that the boat can’t leave us behind.

2 Background and motivation

2.1 Problem setup and notation

Let Z denote the observable data, taking values in a sample space \mathbb{Z} . It’ll often be the case that Z has components Z_1, \dots, Z_n , where $Z_i = (X_i, Y_i)$ might be an independent–dependent variable pair, etc. When it’s relevant that the data is a sample of size n , e.g., when considering large-sample properties, I’ll write $Z^n = (Z_1, \dots, Z_n)$. Throughout, z (or z^n) will denote a particular realization of the observable data Z (or Z^n).

Next, suppose statistical model $\{\mathbf{P}_\theta : \theta \in \mathbb{T}\}$, consisting of probability distributions supported on (subsets of) \mathbb{Z} , is imposed to quantify the variability or aleatory uncertainty in the observable data Z ; I’ll relax this parametric model restriction in Section 6 and later. The probability distributions \mathbf{P}_θ have a corresponding density/mass function $p_\theta(z)$ and, if $Z = z$ is the observed data, then $\theta \mapsto L_z(\theta) := p_\theta(z)$ is the likelihood function.

Finally, let Θ denote the uncertain true parameter value, i.e., the one such that the statement “ $Z \sim \mathbf{P}_\Theta$ ” is true; this too can and will be relaxed later. Like with my z and Z notation, note that I’m writing θ for generic parameter values and Θ for the uncertain true value to be inferred. Importantly, here I’ll be assuming that prior information about Θ is *vacuous*, i.e., my answer to the question “What is the prior probability that Θ is in A ?” would be “between 0 and 1” for all $A \notin \{\emptyset, \mathbb{T}\}$ —complete ignorance. While ignorance is the state most commonly assumed in the statistics literature, it’s not because complete ignorance is realistic—arguably it’s rare in applications to know *literally nothing* about the quantity of interest. See Section 7 for a brief discussion about recent

developments allowing for non-vacuous, incomplete prior information about Θ .

Since prior information is vacuous, all that’s available is the model/likelihood for Z and the realization z . According to Hacking (1976), “Statisticians want numerical measures of the degree to which data support hypotheses,” which, to me, sounds like *probabilistic uncertainty quantification* (Section 2.2). So, despite having fully vacuous prior information, which puts proper Bayesian inference out of reach, the goal still is to assign data-dependent probabilities (or something similar) to hypotheses about the unknown Θ . Towards this, I’ll follow Fisher—“the world’s master of quantifying uncertainty” (Pearl 2018)—, Jeffreys, Dempster, Berger, Walley, and other leaders.

2.2 Probabilistic uncertainty quantification

For now, probabilistic uncertainty quantification means assigning data-dependent probabilities, say, $Q_z(\cdot)$ to hypotheses about the unknown Θ . The point is that a hypothesis “ $\Theta \in H$ ” determined by a subset $H \subseteq \mathbb{T}$ can be true or not, and a large $Q_z(H)$ assigned to that hypothesis would naturally be interpreted as an indication that data z supports the truthfulness of that hypothesis. (Henceforth, when I refer to a “hypothesis H ” I mean the hypothesis “ $\Theta \in H$.”) So, the magnitudes of these probabilities can be broadly used to assess which hypotheses the data does and doesn’t support. Of course, these probabilities can be used for other purposes (e.g., credible/confidence sets, prediction, and decision-making) but it’s this basic function of assessing broadly where the data lends its support that distinguishes probabilistic uncertainty quantification from other approaches. In Section 2.3 below I give a brief summary of the two most familiar of the existing approaches to probabilistic uncertainty quantification.

It’s important to distinguish probabilistic uncertainty quantification about an unknown and unobservable Θ , like what’s in consideration here, from that about an unknown that’s observable, e.g., a future data point. In the latter case, the probability model in question can be directly tested against observations: if a pre-specified event that the model says has (effectively) zero probability happens, then the model must be wrong. This is *Cournot’s principle*, see Vovk (1993), Shafer (2007), and Shafer and Vovk (2019, Ch. 10). In the former case, however, the true Θ will typically never be revealed, so the probabilistic uncertainty quantification about Θ can’t be *directly* tested against reality. But it can be *indirectly* scrutinized for reliability:

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

In the present context, the “procedure” in question is a mapping $z \mapsto Q_z$ from data z to a probability distribution Q_z on \mathbb{T} , and such procedure gives systematically misleading conclusions if $Q_Z(H)$ tends to be large, as a function of $Z \sim P_\Theta$, when the hypothesis is false, i.e., $H \not\supseteq \Theta$. So, at a high level, a probabilistic quantification of uncertainty is reliable if the “objective” P_Θ -probability that $Q_Z(H)$ is large for false hypotheses H is itself small. In symbols, but still only roughly, if I define the function

$$H \mapsto \sup_{\theta \notin H} P_\theta \{Q_Z(H) \text{ is large}\}, \quad H \subset \mathbb{T},$$

then it’s clearly a desirable property of $z \mapsto Q_z$ that the right-hand side above be small for all $H \subset \mathbb{T}$. Importantly, a more precisely defined version of this function (see Section 2.4) can be evaluated and used to assess the reliability of a given probabilistic uncertainty quantification procedure. That these considerations assess the procedure’s *reliability* is implied by Cournot’s principle, albeit from a different perspective than that above. Indeed, since events with small P_Θ -probability effectively don’t happen, if the right-hand side of the above display is small in the sense above, then Q_Z won’t assign high probability to false hypotheses, hence no “systematically misleading conclusions.”

2.3 Existing approaches

Default-prior Bayes. The idea of using equal probabilities as a default when genuine information is lacking has a long history. It appears in the original work of Bayes (1763) and was adopted by Laplace (1812) and other contemporaries, eventually acquiring the name *principle of insufficient reason* (e.g., Stigler 1986, p. 127–129). Keynes (1921, Ch. 4) later renamed it the *principle of indifference*, which he described as follows:

The Principle of Indifference asserts that... *equal* probabilities must be assigned to each of several arguments, if there is an absence of positive ground for assigning *unequal* ones. (Keynes 1921, p. 45)

On the one hand, the above principle seems generally agreeable at least at first consideration, and it has been used in all sorts of applications and generalized in various ways (e.g., Jaynes 2003). On the other hand, the principle of indifference has been heavily criticized by many authors, including Keynes and Fisher.

Jeffreys responded to Fisher’s criticism with a different perspective. Rather than attempting to give a probabilistic description of ignorance, an impossible task (see Section 2.4), he focused on the construction of otherwise justifiable default priors:

... find a way of saying that the magnitude of a parameter is unknown, when none of the possible values need special attention (Jeffreys 1998, p. 117).

His efforts resulted in the now widely used Jeffreys priors (Jeffreys 1946), later shown to yield posterior distributions with excellent large-sample properties (e.g., Datta and Ghosh 1995; Welch and Peers 1963). Substantial generalizations of Jeffreys formulation, in various directions, are now available (e.g., Berger et al. 2024). But despite this progress, there’s still no general consensus about which—if any—default priors are “right,” so apparently the fundamental problem remains unsolved; see Efron’s quote on page 20.

Fiducial and the like. A novel, non-Bayesian approach to probabilistic uncertainty quantification was taken by Fisher (1930, 1933, 1935a,b)—what Savage (1961) famously described as “a bold attempt to make the Bayesian omelet without breaking the Bayesian egg.” I won’t get into details here, so I refer the reader to Zabell (1992) and Savage (1976) for more on Fisher’s ideas and to Xie and Singh (2013), and Hannig et al. (2016), Schweder and Hjort (2016) for some modern perspectives on fiducial-like inference.

Roughly, Fisher’s fiducial argument takes the model-based, parameter-dependent probabilities assigned to events about the observable data, reinterprets the events as (data-dependent) assertions about the unknown parameter, and then flips the aforementioned

probabilities assigned to these events into subjective probabilities about the unknown parameter, given the observed data. The adjective “fiducial”—which means founded on faith or trust—that Fisher chose to describe his solution makes it clear that he understood his was not a 100% sound mathematical argument. Fisher must have had in mind some principle that justifies the aforementioned faith/trust in his fiducial probabilities but, to my knowledge, he didn’t clearly state any such principle. Dempster (1963, 1964) described it as a *continue to regard* operation; Hannig and others (e.g., Hannig et al. 2016; Murph et al. 2024) coined it a *switching principle* where what’s random and what’s fixed is switched. In any case, the fiducial argument involves a mix of mathematical reasoning and application of principles, hence is not so much different than default-prior Bayes.

Thanks to Fisher’s fame and the mystery surrounding his proposed solution, the fiducial argument received a lot of attention—and heavy scrutiny. Damaging blows to Fisher’s formulation were delivered by Lindley (1958), Dempster (1963, 1964), and Buehler and Fedderson (1963), but these demonstrations, while insightful, largely only confirm that the fiducial argument is not mathematically sound. That the solution offered by Fisher lacks mathematical rigor doesn’t mean that the problem is impractical, immaterial, or impossible, hence fiducial inference remains a sort of Holy Grail for statisticians:

Fisher’s attempt to steer a path between the Scylla of unconditional behaviorist methods which disavow any attempt at “inference” and the Charybdis of subjectivism in science was founded on important concerns, and his personal failure to arrive at a satisfactory solution to the problem means only that the problem remains unsolved, not that it does not exist. (Zabell 1992, p. 382)

2.4 Is probability theory right for the job?

That uncertainty quantification must be formulated using probability theory is almost exclusively taken for granted, at least in the statistics literature. But it’s worth asking if probabilistic uncertainty quantification is capable of achieving the reliability desiderata described in Section 2.2. Spoiler alert—the answer to this question is *No*.

My claim is that probabilistic uncertainty quantification is incapable of achieving the notion of reliability introduced in Section 2.2. For any mapping $z \mapsto \mathbb{Q}_z$ that represents a probabilistic quantification of uncertainty about the unknown Θ , given $Z = z$, there are false hypotheses H , i.e., $H \not\supseteq \Theta$, such that $\mathbb{Q}_Z(H)$ tends to be large as a function of $Z \sim \mathbb{P}_\Theta$. If a large $\Pi_Z(H)$ is interpreted as “confidence” in the truthfulness of H , then the undesirable cases where H is false and $\mathbb{Q}_Z(H)$ is large can be characterized as instances of *false confidence*, and it turns out that every form of probabilistic uncertainty quantification—Bayes, fiducial, etc.—is afflicted by it. That is, however large a probability must be in order to instill confidence, there exists false hypotheses about which the posterior instills confidence with arbitrarily high rate, hence a risk of “systematically misleading conclusions.” The *false confidence theorem* below makes this precise.

Theorem 1 (Balch et al. 2019). *Let $z \mapsto \mathbb{Q}_z$ determine a data-dependent probability distribution absolutely continuous with respect to Lebesgue measure on \mathbb{T} . Define the false confidence rate associated with $z \mapsto \mathbb{Q}_z$ as*

$$\text{FCR}_{\mathbb{Q}}(\alpha, H) = \sup_{\theta \notin H} \mathbb{P}_\theta\{\mathbb{Q}_Z(H) > 1 - \alpha\}, \quad \alpha \in (0, 1), \quad H \subset \mathbb{T}. \quad (1)$$

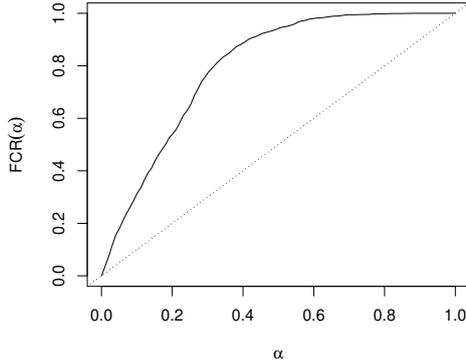


Figure 1: A lower bound on the (empirical) false confidence rate $\alpha \mapsto \text{FCR}_Q(\alpha; H)$ of the Bayes posterior corresponding to the hypothesis H stated in the text.

Then for any $(\alpha, \tau) \in (0, 1)^2$, there exists hypotheses H such that $\text{FCR}_Q(\alpha, H) > \tau$.

In other words, whatever level $1 - \alpha$ is needed to instill confidence in the truthfulness of a hypothesis, there exists hypotheses H for which the false confidence rate $\text{FCR}_Q(\alpha, H)$ is arbitrarily large. Since this is true for any probabilistic uncertainty quantification, the false confidence phenomenon can't be blamed on a poor choice of prior distribution, on issues underlying Fisher's fiducial argument, etc. It's a shortcoming of probability theory with respect to data-driven uncertainty quantification about unknowns.

A criticism that must be dispelled right away is that high false confidence rate is only for "trivial" hypotheses, e.g., H being the complement of a set with zero Lebesgue measure. It's true that such "trivial" H are afflicted by false confidence since $\mathbb{Q}_z(H) \equiv 1$ for all z , but false confidence occurs for lots of seemingly innocuous hypotheses, not just for these trivial ones. One of the many such examples is next.

Consider a simple linear regression model, where $Z = (Z_1, \dots, Z_n)$, with $Z_i = (X_i, Y_i)$ and $(Y_i | x_i, \theta) \stackrel{\text{ind}}{\sim} \mathbb{P}_{\theta, i} := \mathbb{N}(\beta_0 + \beta_1 x_i, \sigma^2)$, for $\theta = (\beta_0, \beta_1, \sigma^2)$. As is customary, I'll treat the x_i 's as given constants that are fixed; in the simulation below, I sampled these values independently from $\text{Unif}(-2, 2)$. A standard Bayesian default prior corresponds to a certain choice of the conjugate normal-inverse gamma prior, so the corresponding posterior inference is rather straightforward. To see how false confidence can creep in, suppose one is interested in the hypothesis

$$H = \{(\beta_0, \beta_1, \sigma^2) : -\beta_0/\beta_1 > -1\},$$

which amounts to a hypothesis that the root of the regression function is larger than -1 . Suppose the true Θ is $(0.3, 0.1, 1)$, so that the above hypothesis is actually *false*. Figure 1 shows a plot of (a lower bound on) the false confidence rate $\alpha \mapsto \text{FCR}(\alpha, H)$ associated with this Bayesian posterior distribution based on 1000 data sets of size $n = 25$. Note that even the lower bound is very high across the entire range of α . This tendency of the Bayesian posterior to assign relatively high probability to a false hypothesis is what creates a risk of systematically misleading conclusions.

The risk associated with formulating statistical uncertainty quantification in terms of probability theory isn't new. For example, Fraser (2013) writes

[Xie and Singh (2013)] are thus recommending that we ignore the restriction to confidence sets or equivalent, and free confidence to allow the production of parameter distributions. Certainly distributions are easier to think about, are largely in accord with Fisher’s original proposal, and are more in the freedom of the Bayes approach, but they do overlook inherent risks...

These risks mainly concern unreliability that results when marginalization is performed (e.g., Balch et al. 2019; Dawid et al. 1973; Fraser 2011a). When Schweder and Hjort (2013) warn that “joint [confidence distributions] should not be sought, we think, since they might easily lead the statistician astray,” they’re concerned that users won’t resist the temptation to do the familiar probabilistic marginalization, thereby creating a risk of unreliability. What causes false confidence or this risk of unreliability more generally remains an open question. The current conjecture is that false confidence tends to occur when the hypothesis concerns values of a non-linear function of the model parameters (Martin 2024b), like the H above is a hypothesis about the ratio; the analyses in Fraser (2011a) and Fraser et al. (2016) offer similar take-away messages.

2.5 If not probability theory, then what?

The issues related to the unreliability of probabilistic uncertainty quantification in the context of statistical inference are general, not specific to any particular brand of probabilistic uncertainty quantification. Therefore, to resolve these issues one must move beyond probabilistic uncertainty quantification to something else. This “something else” should be probability-like, so that uncertainty quantification still makes sense, but it can’t be additive. Capacities, introduced in Choquet (1954), are non-additive set functions, and I’ll show below that a special type of data-dependent capacity can achieve the desired reliability properties that no ordinary probability can.

Loosely speaking, *imprecise probabilities* are normalized capacities having additional properties making them suitable models for quantifying uncertainty. Examples that might be familiar to statisticians include belief functions, first developed by Dempster (1966, 1967, 1968) and later formalized by Shafer (1976), which are infinitely-monotone capacities; the 2-monotone capacities that can be found in robustness investigations (e.g., Berger 1984; Huber 1973, 1981; Wasserman 1990b; Wasserman and Kadane 1990); and the so-called generalized Bayes framework of lower previsions advanced by Walley (1991). No specific details about these particular forms of imprecise probability will be needed in what follows. Here I want to focus on *imprecision* itself and its role.

In textbooks, ordinary or precise probability theory is presented in the context of a chance experiment—e.g., rolling an equally weighted six-sided die—where the specifics are all clearly stated but the outcome can’t be predicted with certainty. Then probability quantifies one’s uncertainty about whether the unpredictable outcome of the experiment will satisfy such-and-such property. This type of uncertainty is *aleatory*. But what if the experiment’s specifics weren’t all clearly stated? If there was ambiguity about the to-be-rolled die’s configuration—e.g., *maybe* half of the faces are labeled with “3,” *maybe* it’s asymmetrically weighted favoring “6,” etc.—then clearly there’s no single probability that accurately captures uncertainty about the outcome. This ambiguity is an example of *epistemic* uncertainty, and this can’t be accommodated using ordinary probability theory. In the extreme case of ignorance about the die, applying the principle of indifference and

assuming equal probabilities on each face is unacceptable: how can the assessments not differ between virtually orthogonal situations where one is ignorant about the die and where one is sure that the die is fair? The issue isn't the assumption of fairness, it's the belief that a single probability can capture both aleatory and epistemic uncertainty. Imprecise probability aims to directly address the epistemic uncertainty, the specification-ambiguity. So, imprecision is not an inadequacy due to a poor assessment; it's about making an honest effort to faithfully capture all of what's uncertain.

A bit more specifically, an imprecise probability corresponds to a lower and upper probability pair which, for our purposes here, is denoted by $(\underline{P}, \overline{P})$. That the two elements of the pair are related to each other will become clear shortly (see, also, Section 3.2), but generally neither are probability measures. The simplest interpretation of $(\underline{P}, \overline{P})$ is literally as lower and upper bounds on a collection of precise probabilities. For concreteness, while still taking some mathematical liberties for the sake of conceptual clarity, consider the die example above and let \mathcal{D} denote the set of all the dice that could possibly be used in the experiment; this set captures the epistemic uncertainty. Each die $D \in \mathcal{D}$ has an associated probability P_D quantifying the aleatory uncertainty about the outcome of rolling die D . To also capture the epistemic uncertainty, one can use

$$\underline{P}(\cdot) = \inf_{D \in \mathcal{D}} P_D(\cdot) \quad \text{and} \quad \overline{P}(\cdot) = \sup_{D \in \mathcal{D}} P_D(\cdot).$$

That $(\underline{P}, \overline{P})$ are linked together should now be clear—they're both tied to the collection $\{P_D : D \in \mathcal{D}\}$. In the extreme case where one is ignorant about the to-be-rolled die, the set \mathcal{D} contains literally all possible dice and, consequently, $\underline{P}(A) = 0$ and $\overline{P}(A) = 1$ for all $A \notin \{\emptyset, \emptyset^c\}$. More generally, it's now clear that imprecise probabilities are more nuanced than ordinary or precise probabilities, and the motivation behind this added complexity is to properly handle both aleatory and epistemic uncertainty.

This is relevant to the present goal of uncertainty quantification about Θ because, with *a priori* ignorance, epistemic uncertainty dominates. From this perspective, the idea that data is sufficiently informative to justify a mapping from the vacuous imprecise prior—complete ignorance—to a posterior that's both fully precise and reliable is completely unrealistic. *Imprecision is imperative*. Indeed, the generalized Bayes rule of Walley (1991) applied to a fully vacuous prior returns a vacuous posterior, meaning it's impossible to learn (in a Bayesian way) when one is *a priori* ignorant; see, also, Kyburg (1987), Walley (2002), and, more recently, Gong and Meng (2021). Non-Bayesian learning isn't susceptible to such criticism, but this upper hand apparently comes at the rather steep price of having to abandon probabilism altogether in favor of procedures with no natural fixed-data uncertainty quantification interpretation; see Zabell's quote on page 6.

My claim is, however, is that many of these non-Bayesian learning strategies do correspond to imprecise-probabilistic or, more specifically, *possibilistic* uncertainty quantification, it's just that no one has realized it. Although imprecise probability theory didn't exist when Fisher was active, there are passages in his texts that suggests he may have anticipated an inexact or imprecise probability theory:

- “[A p-value] is more primitive, or elemental than, and does not justify, any exact probability statement about the proposition” (Fisher 1973, p. 46)
- “It is clear, however, that no exact probability statements can be based on [confidence limits]” (ibid., p. 74)

Presumably, non-Bayesians aren't opposed to fixed-data uncertainty quantification interpretations, they just don't know how to justify this without going a Bayesian route and possibly jeopardizing reliability. The developments described below show how to get possibilistic uncertainty quantification that is both reliable and efficient.

3 Possibilistic inferential models

3.1 Perspective

An *inferential model* (IM) is a mapping $z \mapsto (\Pi_z, \bar{\Pi}_z)$ from data z to a lower and upper probability pair, Π_z and $\bar{\Pi}_z$, defined on \mathbb{T} to quantify uncertainty about the unknown Θ , i.e., a model for uncertainty quantification and inference. This includes the probabilistic approaches discussed earlier as a special case. But Martin and Liu (2013, 2015b) were focused on IMs that are provably reliable, which requires special care and consideration, so I'll follow their lead and not give too much attention to this general definition.

Martin and Liu's original construction involved three steps. The first step was an expression of the posited model via a data-generating process, or an *association*, say, $Z = a(\Theta, U)$, between observable data Z , unknown parameter Θ , and an unobservable auxiliary variable U with a known distribution. The same association is used in Fisher's fiducial argument and in the generalizations proposed by Dempster (1967), Fraser (1968), and Hannig et al. (2016). What's different about the IM approach is that attention focuses on the *unobserved value* u of U , since the relation $z = a(\Theta, u)$ between the observed data z and the unknown Θ must hold. The second step, unique to Martin and Liu's IM construction, is where imprecision is introduced: they quantify uncertainty about the unobserved value u using the distribution of a suitable random set. Since the distribution of U is known, it's easy to ensure that this uncertainty quantification about the unobserved u is reliable. Then the third step maps the random set for u to a corresponding random set for Θ , given $Z = z$, via the relation $z = a(\Theta, u)$, and the reliability offered by the former is immediately transferred to the latter. Then uncertainty quantification about Θ , given $Z = z$, is based on the distribution of that latter random set which, again, can be expressed as a belief function.

This general approach is quite powerful, offering provably reliable, prior-free imprecise-probabilistic inference about Θ , which was arguably Fisher's goal with the fiducial argument. But the general IM approach also has some limitations. First, it requires specification of an association " $Z = a(\Theta, U)$," which is not uniquely determined by the model $\{\mathbb{P}_\theta : \theta \in \mathbb{T}\}$. Second, in order for the inference to be efficient, non-trivial efforts to manipulate the association and reduce the dimension of the auxiliary variable are required (Martin and Liu 2015a,c). Third, a justifiably "optimal" choice of the random set for the unobserved u has remained elusive. None of these on their own is a serious obstacle but, in aggregate, they create a barrier preventing the use of IMs, at least by non-experts, in practical applications. A different, *generalized IM* approach was advanced in Martin (2015, 2018) that specifically avoids the aforementioned limitations, and it's the recent developments around this version of the IM construction that I'll focus on here.

One technical point concerning the random set-based construction is needed to provide background for what follows. Theorem 4.3 in Martin and Liu (2015b) says the only admissible random sets for quantifying uncertainty about the unobserved u are nested,

i.e., for any two realizations of the random set, one is a subset of the other. While the distribution of random sets can generally be described by belief functions, the distribution of a nested random set corresponds to a special type of belief function, namely, a consonant belief function; see, e.g., Shafer (1976, 1987). Consonant belief functions correspond to *possibility measures* (e.g., Dubois 2006; Dubois and Prade 1988), and these are similar to the probability distributions that statisticians are familiar with. Since the aforementioned theorem implies that efficient IMs must take the form of possibility measures on \mathbb{T} , I'll focus exclusively on *possibilistic IMs*; see, also, Liu and Martin (2024). Readers unfamiliar with the basics of possibility theory should consult the brief Appendix A for the background relevant to the statistical developments below.

3.2 Construction

The posited model $\{P_\theta : \theta \in \mathbb{T}\}$ and observed data $Z = z$ determine a likelihood $\theta \mapsto p_\theta(z)$ function and a corresponding relative likelihood

$$R(z, \theta) = \frac{p_\theta(z)}{\sup_{\vartheta} p_\vartheta(z)}, \quad \theta \in \mathbb{T}.$$

I will assume throughout that the denominator is finite for almost all z . An imprecise-probabilistic (in fact, possibilistic) interpretation can be given to the relative likelihood directly, and this has been extensively studied (e.g., Denceux 2006, 2014; Shafer 1982; Wasserman 1990a). But the raw, relative likelihood-based possibilistic uncertainty quantification has issues similar to those discussed above for probabilistic uncertainty quantification—one apparently has no control on the false confidence rate. But the relative likelihood serves an important role, namely, ranking different parameter values based on their compatibility with the observed data $Z = z$, which was the role that Fisher envisioned. Arguably the relative likelihood $\theta \mapsto R(z, \theta)$ above is the “best” such ranking function since it’s a minimal sufficient statistic; see, also, Remark 1 in Appendix C. But this is not the only ranking function one might entertain; see Sections 5–6.

The second step of the possibilistic IM construction is “validifying” (Martin 2022a) the relative likelihood (or other ranking function). This amounts to applying a version of the *probability-to-possibility transform* (e.g., Dubois et al. 2004; Hose 2022), and it returns the possibilistic IM’s contour function:

$$\pi_z(\theta) = P_\theta\{R(Z, \theta) \leq R(z, \theta)\}, \quad \theta \in \mathbb{T}. \quad (2)$$

This is indeed a possibility contour since the function is (clearly) non-negative and satisfies $\sup_{\theta} \pi_z(\theta) = 1$; in fact, the supremum is attained at the maximum likelihood estimator, $\hat{\theta}_z = \arg \max_{\theta} R(z, \theta)$, i.e., $\pi_z(\hat{\theta}_z) = 1$. The corresponding possibility measure, or upper probability, is defined via optimization as

$$\Pi_z(H) = \sup_{\theta \in H} \pi_z(\theta), \quad H \subseteq \mathbb{T}. \quad (3)$$

The corresponding necessity measure, or lower probability, is defined via conjugacy: $\Pi_z(H) = 1 - \Pi_z(H^c)$. An example of π_z and its corresponding Π_z is shown in Figure 2. An explanation of how this new possibilistic IM construction is related to the original random set-based IM construction is given in Appendix B.

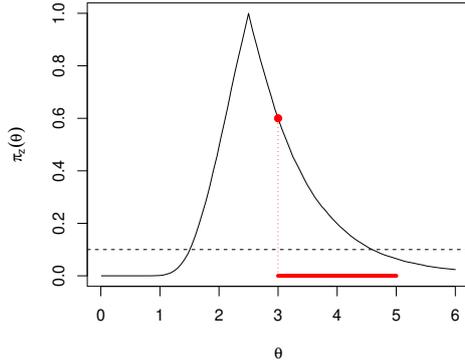


Figure 2: Illustration of a possibility contour $\theta \mapsto \pi_z(\theta)$ and how the possibility measure Π_z is determined from it. Here, the hypothesis H is the interval $[3, 5]$ and the maximum value $\Pi_z(H)$ on contour is highlighted. Horizontal dashed line at $\alpha = 0.1$ determines the 90% confidence interval $C_{0.1}(z)$ in (5).

More details follow, but the key point is that $H \mapsto \{\Pi_z(H), \Pi_z(H)\}$ reliably quantifies uncertainty about Θ , given $Z = z$. For example, H such that $\Pi_z(H)$ is small are implausible because the data provides little/no support in favor of the truthfulness of “ $\Theta \in H$ ” and, similarly, H for which $\Pi_z(H)$ is large are not only plausible but believable since the data provides strong support for the truthfulness of “ $\Theta \in H$.”

3.3 Sampling-based reliability properties

3.3.1 Validity

The essential reliability property of the possibilistic IM is *strong validity*.

Theorem 2. *The possibilistic IM is strongly valid in the sense that*

$$\sup_{\theta \in \mathbb{T}} \mathbb{P}_\theta \{ \pi_Z(\theta) \leq \alpha \} \leq \alpha, \quad \alpha \in [0, 1]. \quad (4)$$

This corresponds to the familiar result for p-values and is a direct consequence of the probability integral transformation taught in the basic math-stat course sequence. While this and some—but not all—of the results below might be familiar in the context of p-values, it’s important to keep in mind that p-values are commonly used for isolated significance tests, not as building blocks for a broad, mathematically rigorous framework for reliable uncertainty quantification. Moreover, the results here align with familiar p-value considerations only because I’m assuming vacuous prior information; the more general case is covered in Martin (2022b) and discussed briefly in Section 7 below.

Strong validity has a number of important consequences. First, (4) immediately implies that the upper α level set of the possibility contour is a $100(1 - \alpha)\%$ confidence region. Note that Bayesian and fiducial credible sets can generally only reach confidence set status asymptotically, as sample size approaches infinity. Figure 2 gives a depiction of the upper level set $C_\alpha(z)$, with $\alpha = 0.1$.

Corollary 1. *The upper α level set of the possibilistic IM contour,*

$$C_\alpha(z) = \{\theta \in \mathbb{T} : \pi_z(\theta) \geq \alpha\}, \quad \alpha \in [0, 1], \quad (5)$$

is a $100(1 - \alpha)\%$ confidence set, i.e., $\sup_{\theta \in \mathbb{T}} \mathbf{P}_\theta\{C_\alpha(Z) \not\ni \theta\} \leq \alpha$.

Second, a more basic property called *validity*, introduced in Martin and Liu (2013), is naturally a consequence of strong validity and the relation (3) between the possibility contour and the possibility measure. The possibilistic IM is said to be *valid* if

$$\sup_{\theta \in H} \mathbf{P}_\theta\{\Pi_Z(H) \leq \alpha\} \leq \alpha, \quad \text{for all } \alpha \in [0, 1], H \subseteq \mathbb{T}. \quad (6)$$

In words, property (6) implies that it's a small probability event that the IM assigns small upper probability to a true hypothesis. Thanks to the relationship between the IM's lower and upper probability, this can equivalently be expressed as

$$\sup_{\theta \notin H} \mathbf{P}_\theta\{\Pi_Z(H) > 1 - \alpha\} \leq \alpha, \quad \text{for all } \alpha \in [0, 1], H \subseteq \mathbb{T}.$$

Then, just like the interpretation of (6): it's a small probability event that the IM assigns large lower probability to a false hypothesis.

A natural question is why should the same quantity, α , appear both inside and outside of the curly brackets in the above two displays. The reason is that the interpretation of numerical probabilities is context-independent. That is, while what probabilities are deemed “small” and “large” may vary from one individual to the next, a statement like *the probability is 0.1* means the same thing to a particular individual whether the topic is tomorrow's weather or the reliability of my data-driven uncertainty quantification about an unknown Θ . So, the scale on which the probabilities about Θ would be interpreted is exactly the same as that on which the model-based probabilities about Z would be interpreted. Hence, the same α —representing whatever value one interprets as “small”—appears inside and outside of the probability statement in (6).

Validity in the sense of (6) has several important consequences concerning the construction of hypothesis testing procedures and the control of false confidence rate. Concerning the latter, while there's no way to eliminate false confidence altogether, except by refusing to learn, i.e., by setting $\Pi_z \equiv 0$ for all z , it's important to control FCR in terms of the confidence threshold α . Bayesian and fiducial uncertainty quantification offers no control over the FCR, but possibilistic IMs do.

Corollary 2. *The possibilistic IM is valid in the sense of (6). Consequently:*

- *The test “reject H if $\Pi_z(H) \leq \alpha$ ” controls the Type I error probability at level α .*
- *If $\text{FCR}_\Pi(\alpha, H)$ is the false confidence rate associated with the possibilistic IM's lower probability $z \mapsto \Pi_z$, analogous to (1), then $\text{FCR}_\Pi(\alpha, H) \leq \alpha$ for all $H \subset \mathbb{T}$.*

Finally, while validity and strong validity were more-or-less treated as equivalent properties in, e.g., Martin and Liu (2013, 2015b), it's important to emphasize that strong validity (4) is indeed stronger than validity (6). This was first established in Cella and Martin (2023), where strong validity in (4) and a uniform-in-hypotheses version of (6) were shown to be equivalent. See Remark 2 in Appendix C for further explanation.

3.3.2 Efficiency

Validity is not the only reliability-relevant concept. In fact, validity is incredibly easy to achieve—just take $\Pi_z(H) \equiv 0$ and $\Pi_z(H) \equiv 1$ for all z and all (non-trivial) H . The problem, of course, with this latter choice is that the output exactly agrees with the vacuous prior; nothing is learned from data. So the goal is to get the most out of the available data but in such a way that reliability is preserved. I have already referred and will continue to refer to this complementary notion of reliability as *efficiency*. The issues here are fundamental, related to basic conceptual notions developed in the early 19th century by Legendre and Gauss concerning least squares and, more generally, the combination of observations (e.g., Stigler 1986) and later, in the 20th century, sufficient statistics, Fisher information, the Cramér–Rao lower bound, etc.

Carefully combining information across distinct sources is critical for efficient inference. In the early IM developments, Martin and Liu (2015a) handled this combination manually, by suitably manipulating the association that links data Z , parameter Θ , and auxiliary variable U . They “rediscovered” classical dimension reduction techniques, such as sufficiency and conditioning on ancillary statistics; they also developed some new insights, beyond the scope of this review. While their manual approach offers greater flexibility and, in turn, the potential for greater efficiency, this is often difficult to carry out. The present construction, with relative likelihood-based ranking, automatically combines observations in a way that’s “optimal,” at least in certain cases, without any manual effort. The results reviewed below, from Martin and Williams (2025), show that the valid possibilistic IM above is asymptotically efficient in a familiar sense. Consequently, the IM’s exact validity (via imprecision) comes at no cost in terms of efficiency.

What’s summarized below is a possibility-theoretic version of the celebrated *Bernstein–von Mises theorem* that appears in the Bayesian and (generalized) fiducial literature, ensuring that the output is asymptotically Gaussian with covariance matrix that agrees with the Cramér–Rao lower bound. The Bernstein–von Mises theorem is important for Bayes and fiducial because it ensures that the credible sets are asymptotic confidence sets. For the possibilistic IM, the contour’s level sets are automatically confidence sets (Corollary 1), so Theorem 3 below strictly concerns the IM’s asymptotic efficiency.

To make sense out of this, I need to introduce the notion of a Gaussian possibility measure. Let g_v denote the d -dimensional Gaussian probability density function, parametrized by a covariance matrix $v \in \mathbb{R}_+^{d \times d}$. Define the corresponding Gaussian possibility measure as the outer possibilistic approximation (e.g., Dubois and Prade 1990, and Appendix A) of $\mathbf{N}_d(0, v)$. Equivalently, the Gaussian possibility measure is just the probability-to-possibility transform of a Gaussian distribution, with contour

$$\gamma_v(y) = \mathbf{P}\{g_v(Y) \leq g_v(y)\} = 1 - F_d(y^\top v^{-1}y), \quad y \in \mathbb{R}^d, \quad Y \sim \mathbf{N}_d(0, v),$$

where F_d is the $\text{ChiSq}(d)$ distribution function. Since we’re concerned with large-sample approximations, replace the generic Z by a vector $Z^n = (Z_1, \dots, Z_n)$ of iid components, and write π_{Z^n} for the IM’s possibility contour. The general theorem in Martin and Williams (2025) says that, under conditions comparable to those in Le Cam (1970), π_{Z^n} and a suitable Gaussian possibility contour—one whose variance agrees with the Cramér–Rao lower bound—merge with probability converging to 1 as $n \rightarrow \infty$. Below is a simpler-to-state version that assumes the stronger conditions typically found in textbooks.

Theorem 3 (Martin and Williams 2025—simpler version). *Under the usual Cramér conditions, the possibilistic IM’s contour π_{Z^n} satisfies*

$$\sup_{u \in \mathcal{U}} |\pi_{Z^n}(\hat{\theta}_{Z^n} + J_{Z^n}^{-1/2}u) - \gamma(u)| \rightarrow 0 \quad \text{in } \mathbb{P}_\Theta\text{-probability as } n \rightarrow \infty,$$

where \mathcal{U} is an arbitrary compact subset of \mathbb{R}^d , $\hat{\theta}_{Z^n}$ is the maximum likelihood estimator, J_{Z^n} is the observed Fisher information matrix, and γ is the standard Gaussian possibility with identity covariance matrix.

For further details, discussion, and illustrations, see Martin and Williams (2025). To summarize, at least when prior information about the model parameter Θ is vacuous and there’s no explicit priority for one assertion about Θ versus another, the relative likelihood-based possibilistic IM is both exactly valid and asymptotically optimal in the classical sense of matching the Cramér–Rao lower bound. But beyond this standard case, improvements may be possible, and that’s the focus of Sections 5–6 below.

3.4 Conditional, fixed-data properties

While it’s natural and important to consider the sampling-based properties of the IM output, an often overlooked advantage of (imprecise) probabilistic uncertainty quantification is that it offers a fully conditional, fixed-data interpretation. This angle isn’t often treated in the literature on default-prior Bayes, (generalized) fiducial, IMs, etc. Unfortunately, I also don’t have the space here for a careful treatment, but see Appendix D.

3.5 Computation

Until recently, only naive and relatively inefficient strategies for computing the IM contour were available. In particular, the go-to strategy was to approximate π_z via

$$\pi_z(\theta) \approx \frac{1}{M} \sum_{m=1}^M 1\{R(Z_{m,\theta}, \theta) \leq R(z, \theta)\}, \quad \theta \in \mathbb{T}, \quad (7)$$

where $Z_{m,\theta}$ are independent copies of data Z , drawn from \mathbb{P}_θ , for $m = 1, \dots, M$. The above computation is feasible at a few different θ values, but all the practically relevant calculations—e.g., identifying the confidence set in (5)—require evaluation over a fine grid covering the relevant portions of \mathbb{T} , and this is expensive. Some improvements to the basic strategy above were presented in Hose et al. (2022). Strategies based on asymptotic approximations are also available, and discussed briefly in Remark 3 in Appendix C.

Martin (2025b) recently developed a new and efficient IM computational strategy. That proposal replaces (most of) the naive contour evaluations in (7) with Monte Carlo sampling from a sort of “posterior distribution” derived specifically from the IM output—not via Bayes’s theorem. The starting point for these developments is what’s call the *credal set* (e.g., Levi 1980, Ch. 5) associated with the IM output. In general, the credal set is just the collection of precise probabilities that are dominated by a given upper probability; in our present notation, this is defined as

$$\mathcal{C}(\Pi_z) = \{\mathbb{Q}_z \in \text{probs}(\mathbb{T}) : \mathbb{Q}_z(H) \leq \Pi_z(H) \text{ for all measurable } H\}, \quad (8)$$

where $\text{probs}(\mathbb{T})$ is the set of all probability measures supported on (the Borel σ -algebra of) the parameter space \mathbb{T} . Since Π_z is a possibility measure, there are simpler characterizations of the elements of $\mathcal{C}(\Pi_z)$ in (8). Indeed, a now-classical characterization (e.g., Couso et al. 2001) of a possibilistic credal set is

$$\mathbb{Q}_z \in \mathcal{C}(\Pi_z) \iff \mathbb{Q}_z\{C_\alpha(z)\} \geq 1 - \alpha, \quad \text{all } \alpha \in [0, 1], \quad (9)$$

where $C_\alpha(z) = \{\theta : \pi_z(\theta) > \alpha\}$ is defined in (5). Since $\{C_\alpha(z) : \alpha \in [0, 1]\}$ is a collection of confidence sets and since each element \mathbb{Q}_z in $\mathcal{C}(\Pi_z)$ assigns probability at least $1 - \alpha$ to the $100(1 - \alpha)\%$ confidence set, it's fair to describe the elements of $\mathcal{C}(\Pi_z)$ as *confidence distributions*. Among these confidence distributions, clearly the “best” approximation \mathbb{Q}_z^* of Π_z is one that achieves equality in the far-right inequality in (9), i.e.,

$$\mathbb{Q}_z^*\{C_\alpha(z)\} = 1 - \alpha, \quad \alpha \in [0, 1]. \quad (10)$$

From the information provided so far, it's not at all clear what form the elements of $\mathcal{C}(\Pi_z)$ have and how the equality in (10) might be achieved. This leads to a more specific characterization, given in Martin (2025b), which is similar to but different from those in Wasserman (1990b) and in Hose (2022): $\mathbb{Q}_z \in \mathcal{C}(\Pi_z)$ if and only if

$$\mathbb{Q}_z(\cdot) = \int_0^1 \mathbb{K}_z^\beta(\cdot) \mathbb{M}_z(d\beta), \quad (11)$$

for some collection of probability distributions $\{\mathbb{K}_z^\beta : \beta \in [0, 1]\}$ such that \mathbb{K}_z^β fully supported on $C_\beta(z)$ for each $\beta \in [0, 1]$, and some probability measure \mathbb{M}_z on $[0, 1]$ such that a random variable with distribution \mathbb{M}_z is stochastically no smaller than $\text{Unif}(0, 1)$. From this characterization, it's easy to check that a distribution $\mathbb{Q}_z^* \in \mathcal{C}(\Pi_z)$ satisfies (10) if and only if \mathbb{M}_z is $\text{Unif}(0, 1)$ and \mathbb{K}_z^β is fully supported on the boundary, $\partial C_\beta(z)$, of the confidence sets $C_\beta(z)$ for each $\beta \in [0, 1]$. If such a \mathbb{Q}_z^* exists, then it's called the *inner probabilistic approximation* of Π_z ; see, also, Dubois and Prade (1990). Existence is typically not an issue (see Martin 2025b, for these details), but the inner probabilistic approximation is not unique. In the spirit of Keynes's principle of indifference, Martin (2025b) recommends taking \mathbb{K}_z^β to be $\text{Unif}\{\partial C_\beta(z)\}$. To summarize, a draw $(\Theta \mid z) \sim \mathbb{Q}_z^*$, the inner probabilistic approximation, can be obtained by follow these two steps:

1. sample $A \sim \text{Unif}(0, 1)$;
2. sample $(\Theta \mid z, A = \alpha) \sim \text{Unif}\{\partial C_\alpha(z)\}$.

See the gamma example below and those in Appendix E. The second step is simple to state but non-trivial to carry out, in part because the boundary $\partial C_\alpha(z)$ isn't readily available. Martin (2025b) proposed a simple work-around, which is to bound the level set $C_\alpha(z)$ by a suitable ellipsoid, say, $C_\alpha^\sigma(z)$, indexed by a vector σ of scaling parameters, in the sense that $C_\alpha(z) \subseteq C_\alpha^\sigma(z)$ for some σ . In light of the asymptotic normality established in Theorem 3, the level set $C_\alpha(z)$ should be approximately ellipsoidal, so there should be no trouble finding an ellipsoid $C_\alpha^\sigma(z)$ in which $C_\alpha(z)$ fits snugly. The appropriate choice of σ would generally depend on both z and α , so I'll write it as $\sigma(z, \alpha)$. See Figure 3 for an illustration. The key point is that sampling from the boundary of an ellipsoid is easy, so if the ellipsoid is a good, not-anti-conservative approximation, then Step 2 can

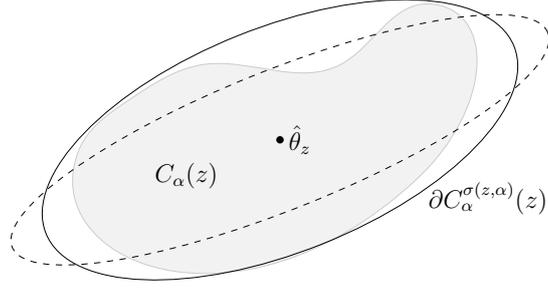


Figure 3: Approximating the IM contour’s α -level set $C_\alpha(z)$ by an ellipse $C_\alpha^\sigma(z)$ with a “good” choice of σ (solid) and with a “bad” choice of σ (dashed).

be reasonably approximated without sacrificing validity. Details about the choice of σ can be found in Cella and Martin (2025) and Martin (2025b). Finally, the samples of Θ drawn according to the above procedure—which are “probabilistic” in the sense that they can be used to approximate probabilities—are transformed back into a possibility measure using an empirical version of the probability-to-possibility transform described in Appendix A; for full details, see Martin (2025b, Sec. 3.4).

3.6 Examples

Normal Example. The particular example considered here—famous for several reasons—is an analysis of Charles Darwin’s data on the difference between cross- and self-fertilized plants. This is the same example that appears in Fisher (1935c, Ch. 3). Context of the experiment aside, the data z consists of $n = 15$ differences in plant heights; the units are eighths of inches. For simplicity, I’ll model these as iid $P_\Theta = \mathbf{N}(\Theta_1, \Theta_2^2)$, i.e., as normal random variables with unknown mean and standard deviation, Θ_1 and Θ_2 , respectively. The maximum likelihood estimators are $\hat{\theta}_{z,1} = 20.93$ and $\hat{\theta}_{z,2} = 36.46$, respectively, and the relative likelihood is

$$R(z, \theta) = \left(\frac{\hat{\theta}_{z,2}^2}{\theta_2^2}\right)^{n/2} \exp\left\{-\frac{n(\hat{\theta}_{z,1} - \theta_1)^2}{2\theta_2^2} - \frac{n}{2}\left(\frac{\hat{\theta}_{z,2}^2}{\theta_2^2} - 1\right)\right\}, \quad \theta_1 \in \mathbb{R}, \quad \theta_2 > 0.$$

Note that $R(Z, \theta)$ is a pivot when Z consists of iid samples from P_θ . Although the distribution doesn’t have a name or a simple form, it is easy to simulate and, therefore, the validation step can easily be done via Monte Carlo. A plot of the possibility contour π_z for Darwin’s data is shown in Figure 4(a). For comparison, samples from the Bayesian posterior distribution based on Jeffreys prior is shown in the background in gray. Both are pointing the same direction, but only the possibilistic IM offers reliable uncertainty quantification guarantees.

Gamma Example. Consider the systems reliability application in Hamada et al. (2004) that involves the breakdown times of $n = 20$ machines. At the 5% significance level, the Kolmogorov–Smirnov test cannot reject the null hypothesis that these data are gamma, so I’ll construct a possibilistic IM for $\Theta = (\Theta_1, \Theta_2)$, the shape and scale parameters of a

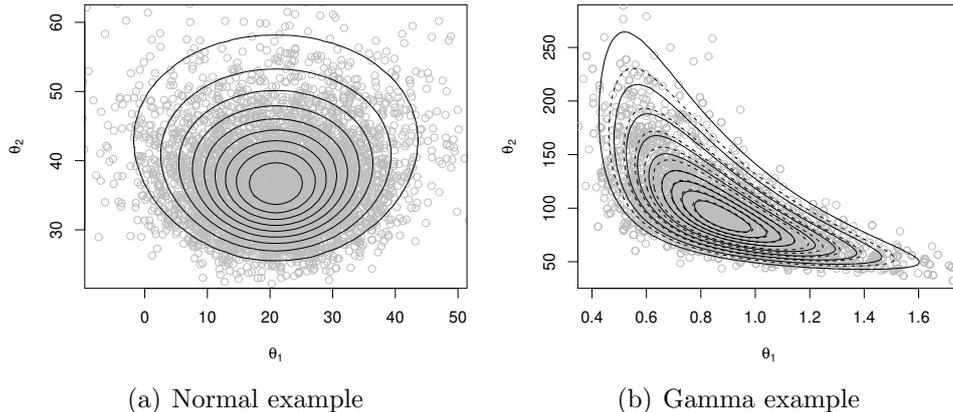


Figure 4: Plot of the joint IM contour for $\Theta = (\Theta_1, \Theta_2)$ in the two examples presented in Section 3.6. In both plots, the solid line is the possibilistic IM contour and the gray points are samples from the Jeffreys prior Bayes posterior. Dashed line in Panel (b) is explained in the text.

gamma model. It is easy to sample from a gamma distribution, and also easy to evaluate the relative likelihood, so the naive approximation (7) of the IM contour is likewise easy; but carrying out these “easy” calculations over a sufficiently fine grid in the (θ_1, θ_2) -space is rather expensive, but the strategy in Martin (2025b) discussed in Section 3.5 is more efficient. A contour plot of this approximation of π_z is shown in Figure 4(b). For comparison, I also show the large-sample Gaussian approximation (dashed line) and samples from a Bayesian posterior distribution using Jeffreys prior (gray points). All three solutions point to roughly the same regions in the (θ_1, θ_2) -space that plausibly contain Θ , and the shapes are similar. Again, only the solid curves are backed by theory that guarantees reliability of uncertainty quantification.

4 Implications for frequentists and Bayesians

4.1 For frequentists

There are good reasons for frequentists to abandon probabilism. The list of reasons includes the reliability-related warnings given in Section 2.4 above, foundational issues as detailed in Mayo (2018), and practically relevant points concerning probabilism’s lack of flexibility, e.g., “The idea that statistical problems do not have to be solved as one coherent whole is anathema to Bayesians but is liberating for frequentists” (Wasserman 2008). But abandoning formal uncertainty quantification altogether because of the shortcomings associated with probabilism is extreme—like throwing the baby out with the bathwater. Indeed, this abandonment is both unnecessary and harmful.

I’ll address the “unnecessary” claim first. Frequentists have their favorite solutions to classical problems, so naturally they’re reluctant to consider new frameworks that propose different solutions to these classical problems. But the likelihood-based possibilistic IM described above often exactly agrees with the classical solution (modulo proper

marginalization, if necessary; see Section 5). Moreover, the likelihood-based formulation can easily be generalized (Section 6), providing more flexibility and broader agreement between the possibilistic IM solutions and go-to frequentist solutions. Even more generally, the result formally stated and proved in Appendix G says roughly the following: *for any test or confidence procedure concerning any feature $\Phi = f(\Theta)$ of the full parameter Θ that has frequentist error rate guarantees, there exists a valid possibilistic IM for Θ —full uncertainty quantification!—that yields a test/confidence procedure for Φ that is at least as good as the given procedure.* This result, which generalizes similar results in Martin and Liu (2014) and Martin (2021a), has an important consequence: there are no genuine frequentist solutions—including those in classical textbooks and those that haven’t been conceived of yet—that are out of the possibilistic IM framework’s reach. Therefore, frequentists are already using possibilistic IMs, so nothing about the proposed brand of uncertainty quantification can be objectionable to them; but they aren’t taking full advantage of all that possibilistic IMs have to offer, which I’ll discuss next.

That frequentists abandoning formal uncertainty quantification is harmful has been widely documented; several recent issues of *The American Statistician* have been devoted to this. The confusion described there stems from textbooks emphasizing that p-values and confidence intervals have no probabilistic interpretation *and then not offering an alternative interpretation.* Without an interpretation, there are at least two ways things can go. Some researchers will craft their own interpretations, but a multitude of different interpretations can only create confusion. Other researchers will simply accept that there is no meaningful interpretation, making statistical analysis a protocol one blindly follows, i.e., a “cult of statistical significance” (Ziliak and McCloskey 2008). This confusion and/or blind trust leads to misuse of statistical tools and, perhaps more importantly, encourages researchers to focus on the relatively narrow scientific questions that they believe can be answered with the simple textbook protocols. Fortunately, this confusion can be overcome, since the frequentist–IM connection offers a simple and mathematically rigorous interpretation of p-values and confidence intervals. Fisher was correct when he said that p-values and confidence intervals warrant no “exact probability statements” about Θ , but that doesn’t mean no statements are warranted. Borrowing Shafer’s description of upper probabilities as measures of plausibility, the aforementioned connection immediately implies that p-values can be interpreted as *the plausibility of H_0 , given data z* and confidence sets can be interpreted as *the set of parameter values that are all individually sufficient plausible, given data z .* This is precisely how p-values and confidence sets are used in practice, and now there’s a mathematically rigorous justification for that interpretation. This is the interpretation of p-values and confidence sets that I teach in my courses, even the introductory-level courses—no need for any technical details about imprecise probability, etc.—and this has been well-received by students.¹

¹For example, a masters student who took my linear models course sent me an email after the course was completed with the following remark after her job interview: “I wanted to let you know that your discussion of plausibility in ST503 made it easier for me to explain what a p-value is when I was interviewing. Thank you for the new perspective on the subject!”

4.2 For Bayesians

In contrast to frequentists, Bayesians are committed to probabilism. Such a commitment is warranted when genuine prior information is available, but otherwise it’s questionable. Since no prior probability distribution that faithfully represents ignorance, no default-prior Bayesian posterior distribution can be “correct” in any sense—“[Bayes’s theorem] does not create real probabilities from hypothetical probabilities” (Fraser 2014). Moreover, even for the pragmatic Bayesian that isn’t concerned about his/her posterior distribution being “correct” still must accept the lack of reliability guarantees from the false confidence theorem. For these and perhaps other reasons, Efron (2013) wrote:

...perhaps the most important unresolved problem in statistical inference is the use of Bayes theorem in the absence of prior information.

Insisting on probabilism is a constraint on the quality and reliability of uncertainty quantification. To emphasize this point, the Society of Imprecise Probabilities Theory and Applications has a motto that reads: *there is more to uncertainty than probability*. The IM formulation openly accepts this limitation of probabilism and acknowledges that, while there is no single “correct” or fully reliable posterior probability distribution in the absence of prior information, there is a set of posterior probabilities that is justifiably reliable, and that set can be characterized by a possibility measure.

Understandably, the reader might be uncomfortable with imprecise probability and, for the sake of simplicity, prefer the familiar probabilistic uncertainty quantification despite its shortcomings. But there are many ways to construct probabilities, and focusing solely on “prior times likelihood” constructions is again a constraint on the quality of uncertainty quantification. The new idea in Martin (2025c) is to *approximate* the IM’s possibilistic output by a probability distribution. Below I give an overview.

Recall the inner probabilistic approximation Q_z^* which, if it exists, satisfies the condition (10). Then such a Q_z^* is exactly probability matching in the sense that its credible sets are confidence sets. Typically, the default-prior Bayes solutions can only achieve this credibility–confidence matching asymptotically. Moreover, under the usual regularity conditions, Q_z^* inherits a probabilistic Bernstein–von Mises theorem from the underlying IM’s possibilistic version (Theorem 3), which implies that the solution Q_z^* is asymptotically efficient and, therefore, the exact probability matching isn’t due to being overly conservative. And while the inner probabilistic approximation generally isn’t a Bayesian posterior under any prior, there are cases in which a direct Bayesian connection can be made. In particular, for the so-called invariant statistical models (see, e.g., Eaton 1989; Schervish 1995, Ch. 6), the Bayesian posterior based on the right Haar prior is an inner probabilistic approximation of the possibilistic IM (e.g., Martin 2023a, 2025c).

To summarize, the limitations of probabilism when prior information is absent, together with the aforementioned benefits of possibilism, warrant abandoning the former in favor of the latter. But even if one insists on probabilism, the Bayesian-style “likelihood times prior” construction itself has limitations: if there was a magical, default prior that solved Efron’s problem, then surely that would’ve been found by now. So, the solution to that problem will likely come from a completely different perspective, one where the posterior isn’t arrived at by application of Bayes’s theorem. Maybe Q_z^* is the no-Bayesian-eggs omelet that solves Efron’s “most important unresolved problem”?

5 Eliminating nuisance parameters

Basu (1977) wrote that “Eliminating nuisance parameters from a model is universally recognized as a major problem of statistics.” Little has changed since Basu’s time—the frequentist impossibility results in, e.g., Gleser and Hwang (1987) and Dufour (1997), and the general unreliability of Bayesian inference as discussed above imply that marginal inference is challenging and requires careful treatment. The possibility-theoretic perspective here offers some new insights, which I discuss below.

A general operation performed in (imprecise) probabilistic inference is *extension*, where uncertainty quantification about one unknown is extended to a related unknown using the uncertainty quantification framework’s calculus. In possibility theory, the relevant calculus is optimization, hence this is the operation used to carry out extension. Following Zadeh (1975, 1978), the building block of the possibilistic extension principle is optimization-based rule for marginalization: using the present notation and terminology, if Θ is unknown, with z -dependent quantification of uncertainty given by the possibilistic IM with contour π_z , and if $\Phi = g(\Theta)$ is a feature of Θ , then the corresponding extension-based marginal IM contour for Φ is defined as

$$\pi_z^{\text{EX}}(\phi) = \sup_{\theta: g(\theta)=\phi} \pi_z(\theta), \quad \phi \in g(\mathbb{T}). \quad (12)$$

While the formality leading up to (12) might be unfamiliar, the operation being carried out is one that statisticians use without second thought: to test a composite hypothesis, one can maximize p-values over all the simple hypotheses that it consists of.

Extension-based marginalization is conceptually simple and reliable in the sense that it preserves the IMs strong validity property like in Theorem 2. That is, for any mapping g , if $\pi_z^{\text{EX},g}$ is the marginal IM contour for $\Phi = g(\Theta)$ as defined in (12), then

$$\sup_{\theta \in \mathbb{T}} \mathbf{P}_\theta \{ \pi_Z^{\text{EX},g}(g(\theta)) \leq \alpha \} \leq \alpha, \quad \alpha \in [0, 1].$$

The simplicity and generality of this strategy are advantageous. But without any tailoring to the specific problem or feature of interest, one should expect the corresponding extension-based marginal IM to be rather conservative.

When interest is in a specific feature $\Phi = g(\Theta)$, it makes sense to modify the ranking step in the IM construction. The rationale is that the goal is to rank values ϕ of Φ in terms of their compatibility with the data—how compatible (z, θ) pairs are is no longer directly relevant. In line with possibility theory’s optimization-centric perspective, a natural strategy is to use a relative profile likelihood:

$$R^{\text{PR}}(z, \phi) = \sup_{\theta: g(\theta)=\phi} R(z, \theta), \quad \phi \in g(\mathbb{T}),$$

where R is the relative likelihood used previously. Then the validation step as usual completes the construction of a profile-based marginal possibilistic IM for Φ , with contour

$$\pi_z^{\text{PR}}(\phi) = \sup_{\theta: g(\theta)=\phi} \mathbf{P}_\theta \{ R^{\text{PR}}(Z, \phi) \leq R^{\text{PR}}(z, \phi) \}, \quad \phi \in g(\mathbb{T}). \quad (13)$$

Note that the outer supremum is needed because, while the relative profile likelihood $R^{\text{PR}}(Z, \phi)$ only directly depends on ϕ , its distribution depends on the model parameter,

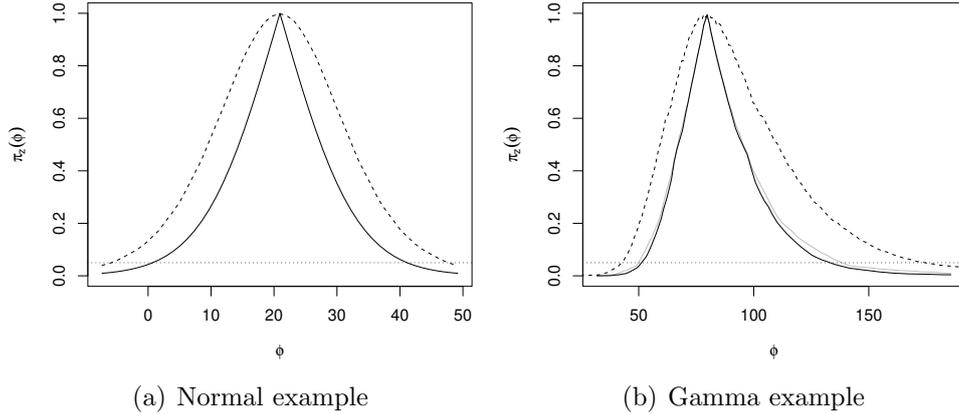


Figure 5: Plots of the marginal IM contours—extension-based (dashed line) and profile-based (solid line)—for the mean Φ in the normal and gamma examples. Gray line in Panel (b) is explained in the text.

which is not fully determined by ϕ . As before, it is easy to show that strong validity is preserved under this profile-based marginal IM construction. The rationale behind why including an optimization in the ranking step is generally superior in terms of efficiency compared to optimizing post-validation step is subtle and I refer the interested reader to Martin (2022b) That profiling tends to be more efficient than extension is easy to see in specific applications; see below. Martin and Williams (2025) show that, while both the extension- and profile-based marginal IM constructions enjoy a large-sample possibilistic Bernstein–von Mises theorem, the latter’s limiting Gaussian generally has a smaller variance, hence its greater efficiency.

Normal Example (cont). Here I’ll revisit the normal example based on Darwin’s data as analyzed in Fisher (1935c, Ch. 3). Here, suppose that the interest parameter is $\Phi = g(\Theta) = \Theta_1$, the mean of the normal distribution. Since this study involved paired data, the mean Φ corresponds to the difference between the means of the two marginal populations paired together. Figure 5(a) shows the extension-based (dashed) and profile-based (solid) marginal IM contours for Φ . The latter, profile-based contour $\phi \mapsto \pi_z^{\text{PR}}(\phi)$ exactly corresponds to the p-value for testing $H_0 : \Phi = \phi$ based on the two-sided Student-t statistic or, equivalently, the probability-to-possibility transform of the sampling distribution of that Student-t statistic. Therefore, the value at $\phi = 0$ is exactly the usual t-test’s p-value which, in this case, is about 0.0495. Note that the profile-based contour is more tightly concentrated, e.g., the 95% confidence interval determined by the horizontal line at $\alpha = 0.05$ is narrower, hence the profile-based marginal IM is more efficient.

Gamma Example (cont). Here I revisit the previous analysis of the data in Hamada et al. (2004). Here the focus is inference on the mean $\Phi = \Theta_1\Theta_2$ of the gamma distribution. Two kinds of marginalization can be carried out—extension-based and profile-based—and both of these are shown Figure 5(b). As above, the extension-based contour, which is simple to derive from the joint contour shown in Figure 4(b), turns out to be much wider and lacks the efficiency of the profile-based solution. Figure 5(b) also shows the “exact” profile-based marginal IM contour (gray line) based on the brute-force strategy

in (7). I call this “exact” because that strategy produces pointwise unbiased estimates of the contour. I’m showing these two curves here to highlight the accuracy of the sampling-based Monte Carlo strategy of Martin (2025b) described briefly in Section 3.5: the two solid lines in Figure 5(b) are almost indistinguishable.

Further discussion and illustrations of the profile likelihood-based IM solution are given in Appendix F. Despite its advantages over extension-based marginalization, it’s important to emphasize that the profiling-based marginalization isn’t universal, i.e., there are cases in which profiling is sub-optimal. Issues arise, as expected, when there are many nuisance parameters, like in the famous examples of Neyman and Scott (1948) and of Stein (1959); see Martin (2023b, Sec. 3.6). More specifically, the profile-based marginal IM is always valid, but the efficiency deteriorates as the number of nuisance parameters increases. The reason is that maximum likelihood estimators tend to be inconsistent when the number of nuisance parameters diverges; since the profile-based marginal IM contour’s peak is at the maximum likelihood estimator, if this is off-target, then a wider contour is needed to cover the relevant range than if the peak was on-target. The remedy is to replace the relative profile likelihood-based ranking with something else, such as a marginal or conditional likelihood (e.g., Severini 1993, 1994, 1998), but so far this has only been addressed on a case-by-case basis (Martin 2023b). It deserves to be mentioned a second time that the aforementioned IMs are always valid—unlike Bayes and fiducial, which can be misleading in nuisance parameter problems—so the question is how to properly rank the interest parameter values so that inference is efficient.

Finally, the ideal cases are those where the relative profile likelihood $R^{\text{PR}}(Z, \phi)$ is a pivot under \mathbf{P}_θ with $g(\theta) = \phi$, since the distributional dependence on the parameters drops out and the computationally challenging supremum on the outside can be omitted. The relative profile likelihood is a pivot in some cases, and approximately so in many other cases—thanks to Wilks’s theorem—but not always. When that supremum can’t be ignored, certain adjustments or approximations may be needed. Below are two such “tricks” to side-step the optimization in (13) that will be used in Section 6 below.

Trick 1. One simple and general strategy for eliminating nuisance parameters is conditioning. A familiar example of this is Fisher’s exact test, where the p-value is obtained using the conditional distribution of the test statistic, given the observed value of what is a sufficient statistic under the null. By definition, the conditional distribution of data given a sufficient statistic doesn’t depend on the parameter, hence the nuisance parameter is eliminated. The present goal isn’t getting a p-value to test a hypothesis, but the relevant computations are similar and, therefore, the same strategy can be employed.

Trick 2. Strictly speaking, the supremum on the outside of (13) isn’t necessary. Indeed, the practically-out-of-reach contour defined as

$$\pi_z^{\text{PR},\star}(\phi) = \mathbf{P}_\Theta\{R^{\text{PR}}(Z, g(\Theta)) \leq R^{\text{PR}}(z, \phi)\}, \quad \phi \in g(\mathbb{T}),$$

which has no supremum, leads to a strongly valid marginal IM. I say that this is out of reach because, of course, evaluation requires knowledge of Θ or at least the true nuisance parameter. But this out-of-reach formulation suggests a simple strategy for constructing a marginal IM that is approximately valid. That is, for the above case, and for the one

where θ naturally decomposes as (ϕ, λ) , set

$$\begin{aligned}\pi_z^{\text{PR}}(\phi) &= \mathbf{P}_{\hat{\theta}_z} \{R^{\text{PR}}(Z, g(\hat{\theta}_z)) \leq R^{\text{PR}}(z, \phi)\} \\ \pi_z^{\text{PR}}(\phi) &= \mathbf{P}_{\phi, \hat{\lambda}_z} \{R^{\text{PR}}(Z, \phi) \leq R^{\text{PR}}(z, \phi)\},\end{aligned}$$

where $\hat{\theta}_z$ and $\hat{\lambda}_z$ are (maximum likelihood) estimates of the full parameter Θ and of the nuisance parameter Λ , respectively. Strategies of this type have a bootstrap flavor (Section 6) and have been applied successfully by Cahoon and Martin (2020, 2021) in meta-analysis and survival analysis applications. Improvements to this simple proposal that would be “closer to valid” include, for example, optimizing over nuisance parameters in a neighborhood of $\hat{\lambda}_z$ in the spirit of Berger and Boos (1994) and Xiong (2017). These improvements will be investigated elsewhere.

6 IMs in more general contexts

6.1 Key technical extension

An apparent limitation of the above proposal is that its emphasis on relative likelihood implicitly assumes a statistical model $\{\mathbf{P}_\theta : \theta \in \mathbb{T}\}$ is available. Machine learning applications, for example, tend to avoid such model assumptions. A simple but important observation, one that’s been applied in various contexts—including Section 5 above—is that validation doesn’t require the ranking to be based on relative likelihood. That is, while the posited model’s likelihood function determines the “optimal” choice of ranking relative to that model (Remark 1 in Appendix C), there may be other factors that suggest a different choice. Here are a few key instances:

- If data are coming from multiple sources, e.g., in a meta-analysis or a divide-and-conquer regime (Hector et al. 2025), and in the form of summary statistics, it may not be possible to compute the full likelihood.
- More generally, the posited model might not determine a likelihood function for the parameter of interest, e.g., in quantile regression.
- If partial/incomplete prior information about Θ is available—see Martin (2022b) and Section 7 below—or if the problem’s context suggests some hypotheses are higher priority than others (e.g., Liu and Williams 2025; Yang et al. 2023), then there’s reason to modify the likelihood-based ranking function.

Here I’ll briefly describe this simple-but-important technical extension, and then I’ll apply it to some relevant problems; see, also, Appendix H.

Relax the assumption that Θ is the true parameter in a posited statistical model, by saying only that there is a true distribution \mathbf{P} and that $\Theta \in \mathbb{T}$ is the true value of the functional $\tau : \mathcal{P} \rightarrow \mathbb{T}$ applied to \mathbf{P} , where \mathcal{P} is some specified subset of $\text{probs}(\mathbb{Z})$, not necessarily a parametric model. Consider a ranking function $\rho : \mathbb{Z} \times \mathbb{T} \rightarrow \mathbb{R}$ such that large values of $\rho(z, \theta)$ indicate that data z and feature θ are compatible in some meaningful way. More detailed example will be given below, but a simple case is where

$\rho(z, \theta)$, where $z = (z_1, \dots, z_n)$ and $z_i = (x_i, y_i)$, represents the negative residual sum of squares corresponding to a fit of some regression function with parameters θ , i.e.,

$$\rho(z, \theta) = - \sum_{i=1}^n \{y_i - m_\theta(x_i)\}^2.$$

Note that this setup doesn't assume a parametric model for the data or that the mean structure imposed by m_θ is "correct." Specification of ρ completes the ranking step of the IM construction, and the validation step proceeds just like in Section 5. That is, since this is effectively a marginal inference problem in which all but the feature $\tau(\mathbf{P})$ of $\mathbf{P} \in \mathcal{P}$ is a nuisance parameter to be eliminated using the aforementioned recipe:

$$\pi_z(\theta) = \sup_{\mathbf{P} \in \mathcal{P}: \tau(\mathbf{P}) = \theta} \mathbf{P}\{\rho(Z, \theta) \leq \rho(z, \theta)\}, \quad \theta \in \mathbb{T}. \quad (14)$$

It is easy to verify that this IM is still strongly valid (relative to \mathcal{P}) in the sense that

$$\sup_{\mathbf{P} \in \mathcal{P}} \mathbf{P}\{\pi_Z(\tau(\mathbf{P})) \leq \alpha\} \leq \alpha, \quad \alpha \in [0, 1].$$

Further details about this general IM formulation are given in Appendix H. It's expected that the Bernstein–von Mises theorem in Martin and Williams (2025) established for the likelihood-based possibilistic IM can be extended to cover some more general cases like described here, but the details remain to be worked out.

The challenge, of course, is evaluating the supremum in (14), and this boils down to making strategic choices of ρ and/or applying the marginalization tricks in Section 5. Interesting and practical examples of both cases will be discussed next.

6.2 Inference on risk minimizers

Let $Z^n = (Z_1, \dots, Z_n)$ be iid random variables with distribution \mathbf{P} . Consider a loss function $(z, \theta) \mapsto \ell_\theta(z)$ that measures how (in)compatible a parameter value θ is with a data point z , where large values correspond to a loss due to incompatibility of θ and z . Of interest is the minimizer Θ of the risk, i.e.,

$$\Theta = \arg \min_{\theta} r(\theta), \quad \text{where} \quad r(\theta) = \int \ell_\theta(z) \mathbf{P}(dz).$$

The risk—or expected loss—function depends on the true \mathbf{P} and, of course, so does Θ . But data Z^n carries relevant information about Θ and so the goal is to quantify uncertainty about the risk minimizer Θ , given the observed data $Z^n = z^n$.

Situations like this are common in statistics and machine learning. Aside from the familiar squared error loss, which is used when the goal is inference on the mean (function), other common examples include classification problems with zero–one loss

$$\ell_\theta(z) = 1\{y = c_\theta(x)\}, \quad z = (x, y),$$

for a posited classification function c_θ parametrized by θ , and quantile regression problems with the check loss, depending on a specified quantile level $u \in (0, 1)$,

$$\ell_\theta(z) = \{y - q_\theta(x)\} [u - 1\{y < q_\theta(x)\}], \quad z = (x, y),$$

where q_θ is a posited quantile function.

The standard approach replaces the unknown risk function with its empirical version,

$$r_{z^n}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell_\theta(z_i), \quad \theta \in \mathbb{T}.$$

Then the empirical risk minimizer, $\hat{\theta}_{z^n} = \arg \min_\theta r_{z^n}(\theta)$ is the natural estimator, often called an M-estimator in the statistics literature. Generalizing the strategy based on the relative likelihood above, here I use the empirical risk function to rank the parameter values, thus completing the ranking step:

$$\rho(z^n, \theta) = -\{r_{z^n}(\theta) - r_{z^n}(\hat{\theta}_{z^n})\}, \quad \theta \in \mathbb{T}.$$

Unfortunately, the validation step is highly non-trivial since it involves an optimization over all \mathbb{P} 's having a given value of the risk minimizer. The recommendation in Cella and Martin (2022a), which is equivalent to an application of “Trick 2” in Section 5, is to replace the true \mathbb{P} with $\hat{\mathbb{P}}_{z^n}$, the empirical distribution, in the validation step. That is, they propose the possibilistic IM with contour function

$$\pi_{z^n}(\theta) = \hat{\mathbb{P}}_{z^n} \{\rho(Z^n, \hat{\theta}_{z^n}) \leq \rho(z^n, \theta)\}, \quad \theta \in \mathbb{T}.$$

Note that the data-driven plug-in shows up in two places: one is obvious on the outside of the probability statement; the other is more subtle since, inside the probability statement, the “true” risk minimizer is replaced by the empirical risk minimizer because now, with respect to $\hat{\mathbb{P}}_{z^n}$, the latter is the “true” risk minimizer. To be clear, in the above display, Z^n is an iid sample from $\hat{\mathbb{P}}_{z^n}$, which is equivalent to sampling with replacement from the observed z^n —hence a connection to the bootstrap (e.g., Davison and Hinkley 1997; Efron 1979; Efron and Tibshirani 1993). So, what’s proposed here boils down to the construction of a sort of bootstrap-based marginal possibilistic IM for the risk minimizer. The cost of plugging in the empirical distribution is that exact validity no longer holds, at least not in general. But Cella and Martin (2022a) prove an asymptotic validity result and show empirically that the validity property—which is only guaranteed to hold in large samples—typically holds even in rather small samples.

Admittedly, the solution described above is not fully satisfactory, since validity only holds approximately, as the sample size converges to infinity, instead of exactly in finite samples. But I think there is plenty of room for improvement here, hence my interest in presenting this less-than-fully-satisfactory solution in the present review. Indeed, my conjecture is that the asymptotic validity described above is higher-order accurate, i.e., the convergence of $\pi_{Z^n}(\Theta)$ to uniformity is faster than the usual root- n rate. More generally, I’m certain that variations on the above proposal that are at least “closer to exactly valid” are within reach. I hope this review inspires others to contribute their own ideas to solve this important and challenging open problem.

6.3 Prediction

To showcase the flexibility of the IM formulation, here I’ll consider a different sort of problem, namely, prediction. Assume that $Z^n = (Z_1, \dots, Z_n)$ consists of iid observations

from a common distribution \mathbf{P} , and that the goal is to predict the next observation Z_{n+1} . The setup here can be generalized in at least two directions: first, all that's required is the Z -process is exchangeable; second, if $Z_i = (X_i, Y_i)$, then it's easy to accommodate the case of predicting Y_{n+1} based on a given X_{n+1} and the observed Z^n . It's just for simplicity of presentation that I focus on the simple iid case. The prediction problem can be viewed as an extreme case of marginal inference, where the entirety of \mathbf{P} is a nuisance parameter to be eliminated. A very common method in the statistics and machine learning literature these days is *conformal prediction* (e.g., Shafer and Vovk 2008; Vovk et al. 2005). A close connection between conformal prediction and IMs has already been demonstrated in Cella and Martin (2022b,c), and what I present below offers new perspectives.

Like above, if z^n is the observed data, then the goal is first to rank candidate values z_{n+1} of the next observation in terms of their compatibility or *conformity* with z^n . This will be done via a ranking function $\rho : \mathbb{Z}^n \times \mathbb{Z} \rightarrow \mathbb{R}$ such that large values indicate higher compatibility/conformity. For example, suppose that \hat{z}_{z^n} is a point prediction of Z_{n+1} based on data z^n ; this can be anything, but a reasonable example in this iid case might be $\hat{z}_{z^n} = n^{-1} \sum_{i=1}^n z_i$, the sample mean. Then the ranking function would be

$$\rho(z^n, z_{n+1}) = -d(\hat{z}_{z^n}, z_{n+1}),$$

where $d \geq 0$ is a distance on \mathbb{Z} . The only other constraint on ρ is that it be symmetric in its first argument: that is, $\rho(z^n, z_{n+1})$ doesn't depend on the order of the data z^n . With ρ satisfying the ranking step, the validation step (temporarily) looks like

$$\pi_{z^n}^\dagger(z_{n+1}) = \sup_{\mathbf{P}} \mathbf{P}\{\rho(Z^n, Z_{n+1}) \leq \rho(z^n, z_{n+1})\}, \quad z_{n+1} \in \mathbb{Z},$$

where the supremum is over all marginal distributions for the iid Z -process or, more generally, over all exchangeable joint distributions. Note that the right-hand side above involves a probability calculation with respect to the joint distribution of (Z^n, Z_{n+1}) . It's easy to verify that this predictive possibilistic IM is strongly valid, i.e.,

$$\sup_{\mathbf{P}} \mathbf{P}\{\pi_{Z^n}^\dagger(Z_{n+1}) \leq \alpha\} \leq \alpha, \quad \alpha \in (0, 1), \quad \text{all } n \geq 1. \quad (15)$$

The problem, again, is that the supremum over \mathbf{P} in the definition of $\pi_{z^n}^\dagger$ is computationally infeasible. Fortunately, the structure of the problem allows for a not-so-naive implementation of the IM construction, one that sidesteps the above computational difficulty but without sacrificing on finite-sample strong validity in (15).

As suggested at the end of Section 6.1, a certain conditioning operation can facilitate the elimination of nuisance parameters. The key insight is that, under the ‘‘hypothesis’’ that $Z_{n+1} = z_{n+1}$, the *set of values* $\{z_1, \dots, z_n, z_{n+1}\}$ —or, equivalently, the empirical distribution—is a sufficient statistic. By conditioning on the sufficient statistic, as in ‘‘Trick 1’’ from Section 5, the dependence on the unknown \mathbf{P} is eliminated, thereby simplifying computations. The new predictive possibilistic IM contour is given by

$$\begin{aligned} \pi_{z^n}(z_{n+1}) &= \sup_{\mathbf{P}} \mathbf{P}[\rho(Z^n, Z_{n+1}) \leq \rho(z^n, z_{n+1}) \mid \{z_1, \dots, z_n, z_{n+1}\}] \\ &= \frac{1}{(n+1)!} \sum_{\sigma} 1\{\rho(z^{\sigma(1:n)}, z_{\sigma(n+1)}) \leq \rho(z^n, z_{n+1})\}, \end{aligned}$$

where the sum is over all $(n + 1)!$ many permutations, σ , of the integers $1, \dots, n, n + 1$. Finally, since ρ is symmetric in its first argument, the right-hand side above simplifies:

$$\pi_{z^n}(z_{n+1}) = \frac{1}{n + 1} \sum_{i=1}^{n+1} 1\{\rho(z_{-i}^{n+1}, z_i) \leq \rho(z^n, z_{n+1})\}, \quad z_{n+1} \in \mathbb{Z},$$

where $z_{-i}^{n+1} = \{z^{n+1}\} \setminus \{z_i\}$. The reader will surely recognize the right-hand side above as the so-called “transducer” or “p-value” output produced by the inductive conformal prediction algorithm. In particular, a result establishing what is equivalent to the prediction validity property in (15) can be found in Corollary 2.9 of Vovk et al. (2005). The derivation above, which makes use of conditioning on and sufficiency of the empirical distribution more closely resembles that in Faulkenberry (1973) and, more recently, Hoff (2023). That one can arrive at the very powerful conformal prediction methodology through an IM-driven line of reasoning is quite remarkable.

7 Conclusion

This paper surveys some of the recent developments in possibilistic inferential models (IMs). Most importantly, IMs offer Bayesian-like fully conditional uncertainty quantification which certain frequentist-like calibration properties that imply, among other things, that tests and confidence procedures derived from the IM’s output control frequentist error rates. Neither of the mainstream approaches to statistical inference are able to achieve both the Bayesian-like and frequentist-like goals, so what distinguishes the IM framework is its reliance on imprecise probability, specifically possibility theory. Fisher hinted that significance tests and confidence intervals warrant “no exact probability statements” but he offered no mathematical explanation for this claim. By being precise about the role played by imprecision, I’m now able to rectify what Efron (1998) playfully referred to as “Fisher’s biggest blunder,” namely, fiducial inference. I must also emphasize, again, that embracing imprecision doesn’t lower the quality of inference and uncertainty quantification—possibility theory is mathematically and philosophically sound and the imprecision prevents false confidence, keeps us honest. Moreover, the new possibilistic Bernstein–von Mises theorem ensures that, at least asymptotically, the possibilistic IM solution is efficient. Much of the discussion here and in the references cited focuses on uncertainty quantification about the parameters in a statistical model, but Section 6 describes the first steps towards pushing IMs beyond this relatively narrow case, making key connections to other fundamental ideas in the literature.

Unfortunately, not every recent development could be included in this review. Next is a brief list of topics that weren’t covered in this review. First, uncertainty quantification has many uses, and one important application is decision-making. Following the von Neumann–Morganstern program, the Bayesian framework starts with a loss function, which evaluates the quality of an action for a given parameter value, and then seeks an action that minimizes the expected loss, averaging over parameter values with respect to the posterior distribution. The fiducial framework proceeds similarly (e.g., Taraldsen and Lindqvist 2013). Possibilistic IMs using Choquet integration is a new approach to assessing the quality of an action in terms of an upper expected loss, and the correspond-

ing decision-theoretic framework offers certain reliability guarantees that Bayesian and fiducial theories don't (Martin 2021b, 2025a).

Second, this review focused on quantifying uncertainty about the parameters specific to a given statistical model. But it's often the case that the model itself is uncertain, and this corresponds to an extreme form of marginal inference—all of the model-specific parameters are nuisance. Preliminary work on IMs in this context can be found in Martin and Liu (2015b, Ch. 10) and Martin (2019). What's missing in these first IM attempts is a penalty on model complexity. Bayesians achieve this complexity penalization via prior distributions, and the fiducial efforts control this complexity manually (e.g., Han and Lee 2022; Hannig and Lee 2009; Lai et al. 2015; Shi et al. 2021; Su et al. 2022; Wei and Lee 2023; Williams and Hannig 2019; Wu et al. 2021). My view is that penalizing model complexity is motivated by a (prior) belief that the true model is relatively simple, and, while mathematicizing vague, partial beliefs like these is impossible using probability theory, it is easy to do using imprecise probability theory. So, forthcoming work will show how to treat vague beliefs like “sparsity” as incomplete prior information, formulated as an imprecise probability, and incorporate these into the IM construction resulting in provably reliable uncertainty quantification about the model itself.

Third, I've assumed here that prior information is vacuous. While this is a standard context in the statistical literature, it's arguably rare that investigators literally know nothing about the quantity they aim to infer. The trouble is that it's equally rare for the available information to be sufficiently complete that it justifies a particular prior distribution for inclusion in a Bayesian analysis. The model complexity penalization discussed above is a good example—one might believe a structural assumption like “sparsity,” but maybe don't know anything about the structure-specific parameters. If the only options are to embellish on what's known to create a precise prior distribution, or ignore what's known and assume prior information is vacuous, then the latter is a safe choice. But the relaxed perspective here on uncertainty quantification offers an alternative path, one that encodes *exactly* the available prior information, however vague or incomplete it might be, as an imprecise probability for inclusion in the analysis. This induces a special type of regularization that allows for efficiency gains while retaining validity. These details are being developed in the series of working papers by Martin (2022a,b, 2023b).

There are far too many open problems to list out here, but below are a few that seem particularly interesting, touching on theory, methods, computation, and applications.

Question. Which statistical hypotheses are afflicted with false confidence? There is strong theoretical and empirical support for the claim that false confidence is caused by non-linearity, i.e., it's a consequence of probabilistic marginalization through non-linear functions of the full model parameter. But a precise characterization of what these hypotheses look like and how severe the affliction will be remains elusive.

Question. What about IMs constructed based on models learned from training data? The IM literature assumes that the model form is given, which is somewhat unrealistic. It's common in machine learning to use training data to learn about certain aspects of the data-generating process, and then use that partially-trained model for inference and prediction. In the present context, the ranking and/or the validation step can depend on training data. What can be said about the reliability of IMs constructed in this way?

Question. How to scale up to higher dimensions? The validity results presented in Section 3.3.1 hold for all sample sizes and all parameter dimensions; the only result that

assumes the model is “low-dimensional” is Theorem 3 on efficiency. So, concerning scaling up to higher dimensions, the question boils down to computational and statistical efficiency. In high-dimensional problems, statistical efficiency is achieved through proper regularization, as discussed above, and work in this direction is in progress. From a computational perspective, new ideas are needed to leverage cutting edge strategies from both optimization and Monte Carlo integration. I don’t expect that brand new ideas are needed, so a good start would be novel combinations of different ideas. To be fair to IMs, Bayesians and frequentists have been working on high-dimensional problems for a while, and the relevant computational challenges haven’t really been “solved”—we generally know *how* to attempt optimization and sampling in high-dimensional problems, but we generally can’t prove that these attempts actually work.

Question. What about causal inference, differential privacy, etc? Exciting applications nowadays involve causation (e.g., Imbens and Rubin 2015; Pearl 2009), data privacy considerations (e.g., Awan and Wang 2024; Garfinkel 2025), etc. There’s no fundamental barrier preventing IMs from contributing in these directions, especially given the extension in Section 6 above. It’s just a matter of working out the details.

I’ll end with some very high-level thoughts about IMs and the potential role that they have to play in artificial intelligence (AI). Of course, AI is concerned with computational systems having the capability to perform tasks typically associated with human intelligence, such as learning, reasoning, problem-solving, perception, and decision-making. It’s not unreasonable to lump this under the broad umbrella of data-driven uncertainty quantification. Indeed, some psychologists (e.g., Gigerenzer and Murray 1987; Juslin et al. 2007) model cognitive processes as (intuitive) statistical inference: questions are posed, relevant data are collected, and judgments are made based on data, assumed models, etc. An AI–IM connection is surely difficult to see at this point, in large part because the IM constructions discussed here are specifically tailored for statistical applications. But the underlying idea—uncertainty quantification with reliability guarantees—is more general and has broader appeal/applicability. Indeed, the deep learning models employed in modern AI applications are arguably just “fancy nonparametric regression models,” so the IM details discussed in Section 6 above and in the supplementary material are clearly relevant. In any case, just as Shafer’s efforts to develop what was later called “Dempster–Shafer theory” independent of the probabilistic language and statistical focus in Dempster’s early work found real AI applications in the 1980s, I’m optimistic that there’s a sufficiently general IM formulation that can meet the “reliable uncertainty quantification” needs of modern AI.

Acknowledgments

This work is partially supported by the U.S. National Science Foundation, under grants SES–2051225 and DMS–2412628. The author sincerely thanks Chuanhai Liu for his support, encouragement, and insights over the last 15+ years. More specifically, thanks also go out to Leonardo Cella, Emily Hector, Sahil Patel, and James Robertson for helpful suggestions and assistance.

A Possibility theory primer

Possibility measures (e.g., Dubois and Prade 1988) are among the simplest forms of imprecise probability, closely linked to fuzzy set theory (e.g., Zadeh 1978) and Dempster–Shafer theory (e.g., Cuzzolin 2021; Shafer 1976, 1987). Some applications in statistics are described in Dubois (2006). Probability and possibility theory differ philosophically and mathematically, but here I’ll only briefly discuss the latter; for a discussion of the former, the presentation in Shackle (1961) is clear and compelling.

The mathematical differences between probability and possibility theory can be succinctly summarized as follows: optimization is to possibility theory what integration is to probability theory. That is, a possibility measure Π defined on a space \mathbb{T} is determined by a function $\pi : \mathbb{T} \rightarrow [0, 1]$ with the property that $\sup_{t \in \mathbb{T}} \pi(t) = 1$. This function is called the *possibility contour* and the supremum-equals-1 property is a normalization condition akin to the integral-equals-1 property of probability densities. Then the possibility measure is determined by optimizing its contour, i.e., $\Pi(A) = \sup_{t \in A} \pi(t)$, for any $A \subseteq \mathbb{T}$, just like a probability measure is determined by integrating its density.

The different calculus has a number of implications. One is that the aforementioned supremum-equals-1 normalization condition ensures that Π is a coherent upper probability (e.g., De Cooman 1997a; De Cooman and Aeyels 1999; Walley 1997) in the spirit of Walley (1991) and others. This means that Π determines a non-empty (closed and convex) set of ordinary probabilities that it dominates:

$$\mathcal{C}(\Pi) = \{\mathbf{Q} \in \text{probs}(\mathbb{T}) : \mathbf{Q}(H) \leq \Pi(H) \text{ for all measurable } H\}, \quad (16)$$

where $\text{probs}(\mathbb{T})$ is the set of probabilities supported on (the Borel σ -algebra of measurable subsets of) \mathbb{T} . The set $\mathcal{C}(\Pi)$ is called the *credal set* and all (coherent) upper probabilities have an associated credal set. Aside from being relatively simple, an advantage of possibilistic uncertainty quantification is that the associated credal set has a statistically oriented characterization, which is important and beneficial for statistical applications.

Finally, if Π is interpreted as an upper probability, then there is a dual lower probability, $\underline{\Pi}$. This dual is called the necessity measure and is given by $\underline{\Pi}(A) = 1 - \Pi(A^c)$, $A \subseteq \mathbb{T}$. This can be interpreted as a measure of support for or confidence in the truthfulness of an assertion A . But since Π and $\underline{\Pi}$ are linked together through the above relationship, it suffices to work with just one and I prefer to work with Π ; this is why the lower probability $\underline{\Pi}$ hardly appears in this review.

A possibility-theoretic notion and operation relevant to the IM developments in this review is the *probability-to-possibility transform* (e.g., Dubois et al. 2004; Hose 2022; Hose and Hanss 2021). Consider a random variable T with probability density function f supported on a space \mathbb{T} . Suppose one wants to approximate the probability distribution \mathbf{P} that quantifies uncertainty about T by a possibility measure—how would this be done? The probability-to-possibility transform, i.e.,

$$\pi(t) = \mathbf{P}\{f(T) \leq f(t)\}, \quad t \in \mathbb{T},$$

is the best such approximation in a sense to be described. But, first, notice that the probability-to-possibility transform resembles a p-value calculation. I hope that this (and the forthcoming explanation of why this kind of construction is “best”) helps to

explain why the possibilistic IM construction above shares some similarities with p-values. That is, the possibilistic IM construction isn't designed to line up with p-values, that's just what the "best" possibilistic approximations happen to look like. Also notice that what the formula in the above display is doing is assigning possibility values to match certain tail probabilities. This should help explain why, in Section 3.5 above, I propose to approximate the possibility measure by a probability distribution that agrees with the former in the tails—I'm just inverting the probability-to-possibility transform. To understand in what sense the possibility measure Π with contour π as in the above display is the "best" approximation of \mathbf{P} , the credal sets are needed. For Π to be a reasonable approximation of \mathbf{P} , it makes sense that the credal set $\mathcal{C}(\Pi)$ should contain \mathbf{P} . Moreover, the "best" possibilistic approximation should correspond to that having the smallest credal set that contains \mathbf{P} . It turns out that the possibility measure Π determined by the contour in the above display satisfies

$$\mathcal{C}(\Pi) \subseteq \mathcal{C}(\Pi') \quad \text{for all } \Pi' \text{ with } \mathbf{P} \in \mathcal{C}(\Pi').$$

For details, see Dubois et al. (2004, Theorem 3.2) and Hose (2022, Sec. 3.1.4). The intuition behind this notion of "best" is that, while I'm apparently willing to introduce some degree of imprecision when approximating a probability by a possibility, there's no rationale for introducing more imprecision than is absolutely necessary to capture \mathbf{P} in the corresponding credal set.

As an example, Section 3.3.2 in the main text introduced what Martin and Williams (2025) called the *Gaussian possibility measure*. This is nothing but the probability-to-possibility transform applied to a given Gaussian probability distribution. Then the interpretation of the Gaussian possibility is that it's the possibility measure with the smallest credal set that contains the given Gaussian probability distribution.

B Connecting old and new IM constructions

The original IM formulation advanced in Martin and Liu (2013) involves a three-step construction. The first step, called the *association step*, gives an explicit link between the observable data Z , uncertain parameter Θ , and unobservable auxiliary variable U , which I'll write here simply as

$$Z = a(\Theta, U),$$

where a is a known function and U is a random variable with a known distribution, denoted by $\mathbf{P}^{(U)}$, say. The interpretation is that the above relationship, together with the knowledge that $U \sim \mathbf{P}^{(U)}$ describes how data are generated according to the posited model. Of course, for a given model, there are lots of different ways to simulate from that model, so the association is not unique; more on this below. This is exactly the starting point for Fisher's fiducial inference, Dempster's generalization, Fraser's structural inference, Hannig's generalized fiducial inference, among others. Excluding impractically simple problems, all of the aforementioned approaches require some pre-processing before the relationship in the above display can be used for inference, but I won't describe these steps here; suffice it to say that this pre-processing involves making certain transformations, conditioning on features of U that are observed, etc. The point that I want to make is that all of the above approaches, in one way or another, aim to treat U as a

random variable both before and after Z is observed. Of course, if U is interpreted as the generator of the data, then U is random before Z is observed; but once $Z = z$ is observed, U is more-or-less fixed at a value depending on the unknown Θ . Fiducial-like approaches argue that there’s little difference between being *fixed at an unknown point* and being *random*, so, for given $Z = z$, they invert the above relationship to write Θ as a function of the known z and assumed-to-be-random U , yielding a probability distribution for Θ . This is where fiducial-like inference frameworks and the IM frameworks differ.

The IM framework explicitly acknowledges that there’s a fundamental difference in the meaning and form of uncertainty associated with a variable U being *fixed-but-unobserved* and being *random*. The point is that, if the above display is a description of the data-generating process, then knowing both Z and U would effectively allow one to solve for Θ . While $Z = z$ is observable, it is impossible to know the value of U that corresponds to Θ and $Z = z$. It is possible, however, to reliably predict² where the unobserved U is positioned, but not using ordinary probability: for example, drawing new values U' from $\mathbf{P}^{(U)}$ wouldn’t work because actually hitting the target U is a probability-0 event. Martin and Liu’s second step, the *prediction step*, introduces a random set \mathcal{S} , taking values in the power set of the U -space, that is designed to have positive (if not large) probability of hitting the unobservable target U . Again, the point is that a meaningful/reliable inversion of the relationship in the above display requires that the actual unobserved value of U be pinned down. This can be achieved with a suitably constructed random set \mathcal{S} , but not with random draws from a probability distribution. This is an important step and offers some insights as to the need for imprecision in statistical inference. That is, Martin and Liu’s *prediction step* amounts to quantifying uncertainty about the unobservable U , given $Z = z$, using a random set rather than a probability distribution. And since the distribution of random sets can be described by belief functions, possibility measures, and other forms of imprecise probability, the connection becomes clear.

Finally, once the association that determines the unobservable U and the random set for predicting said U are specified, the final step in the IM construction, the *combination step*, puts the different pieces of information together. Specifically, for a realization of the random set \mathcal{S} , define

$$\mathbb{T}_z(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \{\theta : z = a(\theta, u)\}.$$

This is a data-dependent subset of the parameter space and, as a function of the random set \mathcal{S} , $\mathbb{T}_z(\mathcal{S})$ is also a random set. Then the distribution of this random set is the vehicle through which inference is carried out. For example, if H is a relevant hypothesis, then the corresponding upper probability is determined by

$$\Pi_z(H) = \mathbf{P}^{(\mathcal{S})}\{\mathbb{T}_z(\mathcal{S}) \cap H \neq \emptyset\}, \quad H \subseteq \mathbb{T},$$

where $\mathbf{P}^{(\mathcal{S})}$ is the user-specified distribution of the random set \mathcal{S} , designed to meet certain easy-to-satisfy calibration criteria that I won’t discuss here; see Martin and Liu (2013) for details. If these \mathcal{S} -calibration conditions are met, then the IM achieves validity like

²Arguably a better word for this is “postdict” (e.g., Dempster 1971; Gelman et al. 1996) since the guessing is being done after the value of U is chosen and the value $Z = z$ is observed, but that’s an awkward and unfamiliar word so I won’t use it here.

in (6) above, i.e.,

$$\sup_{\theta \in H} \mathbf{P}_\theta \{ \Pi_Z(H) \leq \alpha \} \leq \alpha, \quad \alpha \in [0, 1], \quad H \subseteq \mathbb{T}.$$

I’ve admittedly skipped some of the technical details behind the original IM construction here, but those details can be found in Martin and Liu (2013, 2015b) and the other references cited above. What’s more important here in the context of this review is the new perspective that can be offered: key to achieving the desired goal of reliable (imprecise-)probabilistic inference without prior information is acknowledging that probability theory is inadequate, that there’s a fundamental difference between a unknown fixed value and a random value, and that statistical inference involves the former case and hence a more flexible framework for quantifying uncertainty. All that said, carrying out the above three steps to construct an IM for inference in real applications can be challenging, and that’s what motivated the developments surveyed in this review.

Theorem 4.3 in Martin and Liu (2015b) is a key result implying that the random set \mathcal{S} should be *nested*, i.e., for any two of its possible realizations, one is a subset of the other. By “should be nested” I mean that for any non-nested random set, there exists a nested random set such that the corresponding IM is at least as efficient; in other words, IM constructions using non-nested random sets are *inadmissible*. Since nested random sets are equivalent to possibility measures, another way to view this IM construction is as a choice to quantify uncertainty about the unobservable auxiliary variable U using a possibility measure, and this perspective was explored in Liu and Martin (2024).

Whether one prefers the construction in terms of (nested) random sets or possibility measures, the fact remains that there infinitely many choices of these inputs that result in a valid IM. This begs an important practical question: *how to choose the random set/possibility measure?* One of the issues leading to the multitude of different and apparently acceptable choices is that there’s no unique association between data, parameters, and auxiliary variables. What Martin and Liu recommended was to reduce the dimension of auxiliary variable as much as possible or, in other words, choose an association that could be expressed in terms of an auxiliary variable of lowest possible dimension. This dimension-reduction step can make use of various techniques, as described in Martin and Liu (2015a), related to but more general than classical notions like sufficiency, conditioning on ancillary statistics, etc. Additional opportunities for dimension reduction may be possible when the goal is marginal inference on an interest parameter in the presence of nuisance parameters, as explained in Martin and Liu (2015c). Unfortunately, and not surprisingly, even after putting in the effort to reduce the dimension of the auxiliary variable as much as possible, there’s still no obvious/right/optimal choice of random set. Early efforts focused on constructing optimal random sets when the inferential goal was to assess a particular hypothesis H about the uncertain Θ . In certain cases at least, connections could be drawn between this optimal random set and uniformly most powerful tests, but both the former and the latter are restricted to certain kinds of (low-dimensional) hypotheses. Another downside to these efforts is that the optimal random set for a given H need not be the optimal random set for another H' and, since the goal of the IM framework is to reliably and broadly/coherently quantify uncertainty, having a different IM for each hypothesis kills the breadth/coherence of the overall approach.

Martin (2015, 2018) took a different approach. Rather than specifying an association the fully characterized the data-generating process, he instead formed an association be-

tween data, parameters, and a *scalar* auxiliary variable. This extreme form of dimension reduction is possible only because the direct connection to the data-generating process is lost. But with an appropriate choice of summary—the relative likelihood—all the relevant information in the data about Θ can be preserved, suggesting that this is not a serious sacrifice. Indeed, to my knowledge, there are currently no examples showing that the approach based on a (reasonable) generalized association as described above is grossly inferior to that based on the formulation in Martin and Liu (2015b). To the contrary, there are many examples that are out of reach for the original IM formulation that can be readily handled using the generalized IM formulation (e.g., Cahoon and Martin 2021).

An advantage of the generalized association advanced in Martin (2015, 2018) is that there’s really only one reasonable choice of random set, hence the aforementioned ambiguity is gone. This creates an opportunity to formulate a theory that parallels the (generalized) fiducial and objective Bayes theories that offer a normative approach to statistical inference, one where the theory itself guides the construction of a solution to new problems. To me, this was a major advancement.

The possibilistic IM formulation considered here is technically no different than the generalized IM formulation offered in Martin (2015, 2018). What’s new in recent years, and the motivation behind writing this review paper, is the connection to imprecise probability theory in general, and possibility theory in particular. Without this connection, it’s easy to interpret the proposal as a variation on or extension of significance testing, a derivative of classical frequentist hypothesis testing theory. But the goal is broad, coherent, and reliable uncertainty quantification, and the connection to a formal mathematical framework that fits with these desiderata is critical to the developments. Imprecise probability and possibility theory also provide guidance on how to solve other problems related to inference, including prediction and formal decision-making, although these topics aren’t covered in depth in this review. That same theory also provides some clues on how one can generalize from the “vacuous prior” case in consideration here, and this work is currently under development. Finally, possibility theory also sheds some light on the optimality of the relative likelihood-based construction advocated for here in the review. The interested reader is referred to Section 4 in Martin (2022b).

C IM-relevant technical remarks

Remark 1. Here I aim to justify the claim made in the main text that the relative likelihood-based IM construction is “right” or “optimal,” at least in cases when prior information is vacuous and there’s no justification for prioritizing one hypothesis about Θ over another. For this, I’ll appeal to some more general results and principles, related to some of the details presented in the possibility theory review in Appendix A; I’ll assume that the reader is familiar with the notation and terminology from there. This is based on the arguments in Martin (2022b, Sec. 4.2), which itself closely follows the presentation in Hose (2022, Sec. 2.3.2.1).

What I refer to as the “ranking function” in the main text is called a *plausibility order* in Hose (2022). For the general random variable $T \sim \mathbf{P}$ in \mathbb{T} , write this plausibility order as $\rho : \mathbb{T} \rightarrow \mathbb{R}$, with larger values of $\rho(t)$ indicating higher plausibility. For a given ρ , there is an optimal possibility measure whose contour is given by $t \mapsto \mathbf{P}\{\rho(T) \leq \rho(t)\}$. Prin-

cipally, this is based on the *Principle of Plausibility* which states that “what is plausible must be possible.” Mathematically, this choice of possibility contour is “optimal” in the sense that its corresponding credal set is the smallest among all those possibility measures whose credal sets contain \mathbf{P} and whose contours respect the ordering determined by ρ . That we should desire small credal sets is motivated by the *Principle of Expressiveness* or *Principle of Maximum Specificity* (e.g., Dubois and Prade 1986). So, when it comes to forming a possibilistic representation, the choice comes down to the ranking function or plausibility order ρ . For this, one can appeal to the *Principle of Representation* which says “what is probable must be plausible,” which suggests that plausibility orders be determined by probability density/mass functions. This highlights the probability-to-possibility transform described in the main text and also in Appendix A as the “right” or “optimal” choice for approximating a probability distribution by a possibility measure; see, also, Theorem 3.2 in Dubois et al. (2004).

The statistical context looks a little different only because we interpret likelihood and density functions differently. But the likelihood can be viewed as an “upper density function” when prior information about Θ is vacuous, and then the use of the likelihood function for ranking (Z, Θ) pairs is natural. One last issue is that the likelihood is ranking θ values for given Z , so there’s a sort of “conditioning” step that’s required. Motivated by minimal sufficiency and a desire to avoid the ambiguities associated with *conditioning* in possibility theory (e.g., De Cooman 1997b), the normalized or relative likelihood is used. For a more detailed explanation, the reader is referred to Martin (2022b, Sec. 4.2).

Remark 2. As stated in the main text, strong validity is indeed stronger than validity; the reason is because strong validity is equivalent to a version of the basic validity property that’s *uniform-in-hypotheses*. In words, strong validity implies that it’s a small probability event that the IM assigns small upper probability to *any true hypothesis*, even those chosen by an adversary and/or depending on data. In mathematics,

$$\sup_{\theta \in \mathbb{T}} \mathbf{P}_\theta \{ \Pi_Z(H) \leq \alpha \text{ for some } H \text{ with } H \ni \theta \} \leq \alpha, \quad \alpha \in [0, 1].$$

This reveals how the IM’s uncertainty quantification is more than significance testing. Cella and Martin (2023) argue that this uniformity helps bring to practical use the philosophically appealing notions of *probing* laid out in Mayo (2018).

Remark 3. One very simple strategy for approximating the IM contour function is based on the new possibilistic Bernstein–von Mises theorem presented in Section 3.3.2. This boils down to a simple Wilks-style chi-square approximation:

$$\pi_z(\theta) \approx 1 - F_d(-2 \log R(z, \theta)), \quad \theta \in \mathbb{T} \subseteq \mathbb{R}^d,$$

where F_d is the ChiSq(d) distribution function. The right-hand side is effectively a closed-form expression, and is often quite accurate, even for moderate or small sample sizes. Even more sophisticated approximations using Bartlett corrections, etc. can be employed, as discussed in, e.g., Schweder and Hjort (2016, Ch. 7). But large-sample approximations are not fully satisfactory. Ideally, there’d be a middle-ground between the effective-but-expensive strategy in Equation (7) of the main text and the too-simple large-sample approximations above. That’s what the sampling-based strategy in Martin (2025b) reviewed in the main text aims to achieve.

D Conditional, fixed-data properties of IMs

Famously and provocatively, de Finetti (1990, p. x) starts his treatise on probability by saying “Probability does not exist.” What he means is that probability doesn’t exist in any non-subjective sense, that probability is a mental construct that depends on one’s subjective assessments and judgments. He further argues that, while individuals have inherent beliefs, accurately eliciting these degrees of belief is unrealistic. Therefore, de Finetti focuses on what individuals *do* rather than what they *think*, and suggests relating probabilities to prices an agent is willing to pay, or accept as payment, for a ticket that pays \$1 if a given uncertain event happens (or a given hypothesis is true) and \$0 otherwise. What mathematical properties should an agent’s pricing system satisfy?

Dutch book theorems, advanced in Ramsey (1926), de Finetti (1990), and Savage (1972), state that if, say, my pricing system doesn’t satisfy the properties of a (finitely additive) probability, then I can be made a sure-loser, i.e., another agent can find a finite collection of gambles which are all acceptable to me but together will result in me losing money no matter what happens/is true. Whether the betting context is real or artificial, quantifying uncertainty via a system susceptible to sure-loss is understood to be irrational. This makes a compelling argument in favor of probabilism. Indeed, Bayesians apparently have strong reason to resist my proposed transition from probabilistic to possibilistic uncertainty quantification: taken at superficial face value, since possibility measures aren’t finitely-additive, the Dutch book theorems imply that those employing possibilistic reasoning can be made sure losers.

There is, however, an important condition in these Dutch book theorems that’s easy to overlook: they assume that, for each gamble, the agent is willing to *buy* and *sell* the corresponding \$1 ticket for the stated amount. Such a condition makes sense under the symmetry resulting from perfect information. But if, for example, there is a collection of different dice that could be rolled, but I don’t know which one specifically, then I should base my buying and selling prices for the gamble that pays \$1 if the die lands on “6” on the lower and upper probabilities of landing on “6” over the collection of possible dice, respectively. If one accounts for this asymmetry, then the Dutch book theorems can be generalized: an agent whose system consists of bounds on acceptable buying/selling prices, expressed as an imprecise probability satisfying some very mild conditions (see below), is safe from being made a sure-loser (e.g., Walley 1991, Ch. 2). Intuitively, bets based on imprecise probabilities are more cautious than bets based on ordinary/precise probabilities, so if the latter is safe from sure-loss, then the former must be too. Therefore, imprecise-probabilistic reasoning, despite not being finitely-additive, isn’t subject to the criticism implied by a face-value interpretation of those classical Dutch book theorems.

Returning to the statistical inference context of the present paper, while none of the (imprecise) probabilistic uncertainty quantification perspectives requires subjective prior distributions, all are subjective in de Finetti’s sense: whatever distribution for Θ , given $Z = z$, the data analyst chooses is indeed his/her *choice* and can’t be considered “objective” in any meaningful sense. One might then ask how do these different perspectives stack up in terms of de Finetti’s subjectivist considerations. Default-prior Bayes and (generalized) fiducial return ordinary probabilities, so these are safe from sure-loss for any fixed z . In the case of IMs, for fixed z , I can construct a pricing system out of my

IM output as follows:

$$\begin{aligned}\Pi_z(H) &= \text{my supremum buying price for } \$1(\Theta \in H), \text{ given } z \\ \Pi_z(H) &= \text{my infimum selling price for } \$1(\Theta \in H), \text{ given } z.\end{aligned}$$

Here and elsewhere, $1(E)$ denotes the indicator function of the event E . The idea is as follows: if another agent is selling $\$1(\Theta \in H)$ for less than $\Pi_z(H)$, then I'd buy it from him/her; if another agent is willing to buy $\$1(\Theta \in H)$ for more than $\Pi_z(H)$, then I'd sell it to him/her; otherwise, the gamble is “too risky” so I'd neither buy nor sell. Since the possibilistic IM above satisfies the aforementioned “very mild conditions,” it follows that, for each fixed z , this IM-based pricing system is also safe from sure-loss.

I emphasized “for any fixed z ” because there are more extensive notions of coherence that consider potential losses incurred when gambles are made first based on prices determined *a priori*, then prices are updated in light of observation $Z = z$, and later debts are settled. This involves properties of the “(prior, data) to posterior” mapping, and such considerations are more interesting when prior information is non-vacuous; these cases are beyond the present review's scope, but see Martin (2022a,b).

What are those “very mild conditions” mentioned above? Analogous to the collection of dice in the motivating example in the main text, define the collection of precise probability distributions compatible with the imprecise probability (Π_z, Π_z) , i.e.,

$$\mathcal{C}(\Pi_z) = \{\mathbb{Q}_z \in \text{probs}(\mathbb{T}) : \mathbb{Q}_z(H) \leq \Pi_z(H) \text{ for all measurable } H\},$$

where $\text{probs}(\mathbb{T})$ is the set of all probability measures supported on (the Borel σ -algebra of) the parameter space \mathbb{T} . This is called the *credal set* (e.g., Levi 1980, Ch. 5) associated with the imprecise probability pair (Π_z, Π_z) , defined in (8) and discussed in Section 3.5 in the main text. The individual elements \mathbb{Q}_z generally depend on z because the upper bound Π_z does. The same credal set can be defined in terms of the lower probability Π_z , just with the inequality “ \leq ” reversed. Then the aforementioned mild condition is that the credal set be non-empty, that there is at least one precise probability compatible with the given imprecise probability. The idea is that if there exists a probability \mathbb{Q}_z such that the (Π_z, Π_z) pricing system is more cautious than \mathbb{Q}_z , then the former's safety from sure-loss implies the same of the latter. For possibility measures in general, which includes the possibilistic IMs in particular, the credal set being non-empty corresponds to a very simple condition concerning the contour, namely, $\sup_{\theta} \pi_z(\theta) = 1$ (e.g., Troffaes and de Cooman 2014, Prop. 7.14). As was mentioned in the construction, this holds for the possibilistic IM, hence no sure-loss.

The IM output can also be evaluated with respect to the classical statistical principles, such as the sufficiency, conditionality, and likelihood principles. The short explanation is that the relative likelihood-based possibilistic IM satisfies the sufficiency principle—what Mayo (2014) and others refer to as the *weak likelihood principle*—but it does not automatically satisfy the conditionality and likelihood principles. As Martin (2024a) explains, this is not a shortcoming, but I point the reader to that paper for further explanation.

E More examples

E.1 Binomial regression

I'll focus here on a common example involving binomial regression, namely, the o-ring failure experiment associated with the 1986 crash of the Challenger space shuttle. In this study, 6 tests were conducted at a range of different temperatures (x) and the number of o-ring failures (y) out of 6 tests were recorded. The data `orings` are available in, e.g., the `faraway` package (Faraway 2022). Figure 6 shows the data and the estimated o-ring failure probability curve obtained by fitting a binomial regression model with logistic link. Note that the fitted curve is decreasing very fast, suggesting that o-rings are very likely to fail at low temperatures. This is relevant because the Challenger space shuttle launch took place on a day when the temperature was 31 degrees Fahrenheit. Here I present some details of an IM analysis of these data.

Data Z consists of pairs $Z_i = (X_i, Y_i)$, for $i = 1, \dots, n$, and a conditionally binomial model for Y_i , given X_i , i.e.,

$$(Y_i \mid X_i = x_i) \sim \text{Bin}\{6, F(\theta_1 + \theta_2 x)\}, \quad i = 1, \dots, n,$$

where where $F(u) = (1 + e^{-u})^{-1}$ is the logistic distribution function. The corresponding likelihood cannot be maximized in closed-form, but this is easy to do numerically, and the maximum likelihood estimator and the corresponding observed information matrix lead to the asymptotically valid inference reported by standard statistical software. Exact inference, however, implies a heavier computational burden, since evaluating the exact IM contour over a sufficiently fine grid of θ values is prohibitively expensive. Alternatively, the Monte Carlo sampling method presented in Section 3.5 is easy to implement and fast to run. Figure 7(a) shows the 5000 Monte Carlo samples of (Θ_1, Θ_2) from \mathbf{Q}_z^* along with the corresponding approximate possibility contour.

A practically relevant question concerns the (log) odds of an o-ring failure at temperature 31 degrees, the temperature on the day of the Challenger launch. Under this binomial model, the log odds of an o-ring failure at 31 degrees is

$$\phi = \log \left\{ \frac{F(\theta_1 + \theta_2 \cdot 31)}{1 - F(\theta_1 + \theta_2 \cdot 31)} \right\} = \theta_1 + \theta_2 \cdot 31.$$

For inference on $\Phi = \Theta_1 + \Theta_2 \cdot 31$, the true log odds of failure at 31 degrees, I use the previously-obtained samples (Θ_1, Θ_2) from \mathbf{Q}_z^* to get a corresponding set of samples of Φ . Figure 7(b) plots the corresponding estimate of the possibility contour for Φ . If the log odds of failure at 31 degrees are roughly between 2 and 8, then that corresponds to a very high probability of o-ring failure, very dangerous.

E.2 Categorical data

A fundamental problem is inference on the probability vector in a multinomial model. I say “fundamental” because this is the canonical nonparametric inference problem—one that, e.g., inspired Dempster’s original developments (Dempster 1966) of what’s now called Dempster–Shafer theory and still a relevant topic of research, including the discussion paper of (Jacob et al. 2021) published recently in *JASA*.

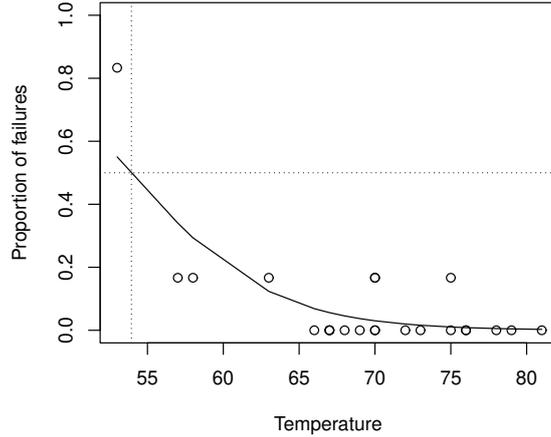


Figure 6: Plot of the o-ring data along with the fitted failure probability curve. The vertical line at 53.94 marks the estimated temperature at which is the o-ring failure probability is 0.5.

To set the scene, a random sample of size n is taken from a population consisting of K -many categories, labeled $1, 2, \dots, K$. The model specifies a probability vector $\theta = (\theta_1, \dots, \theta_K)$ with

$$P_\theta(\text{observation is of category } k) = \theta_k, \quad k = 1, \dots, K.$$

The parameter space \mathbb{T} is the K -dimensional probability simplex, i.e.,

$$\mathbb{T} = \{u \in \mathbb{R}^K : u_k \geq 0, \sum_{k=1}^K u_k = 1\}.$$

The observable data $Z = (Z_1, \dots, Z_K)$ consists of a frequency table listing the number of observations in each of the K categories; note that $\sum_{k=1}^K Z_k = n$. The true probability vector Θ is unknown and to be inferred based on observations $Z = z$. More precisely, let $(Z \mid \Theta = \theta) \sim \text{Mult}_K(n, \theta)$. This determines a likelihood function $L_z(\theta) \propto \prod_{k=1}^K \theta_k^{z_k}$, for $\theta \in \mathbb{T}$. Then the relative likelihood is

$$R(z, \theta) = \prod_{k=1}^K \left(\frac{n\theta_k}{z_k} \right)^{z_k}, \quad \theta \in \mathbb{T}.$$

From here it's straightforward to evaluate the possibilistic IM contour function, but expensive to do so over a dense grid of points in the simplex when K is 2 or more. Fortunately, the computational strategy presented in Martin (2025b) and reviewed in Section 3.5 can be readily applied to relieve this computational burden. In my illustration below, I'll use the basic ellipsoidal approximations as mentioned in Section 3.5 of the main text, but note that this may not be sufficiently accurate when the sample size is small or if there are sparse cells in the observed z .

For illustration, I consider a down-scaled version of an example presented in Agresti (2002, Sec. 3.2.2). The original example involves a 3×3 cross-tabulation of religious beliefs and education levels. I'm treating this as a multinomial inference problem with

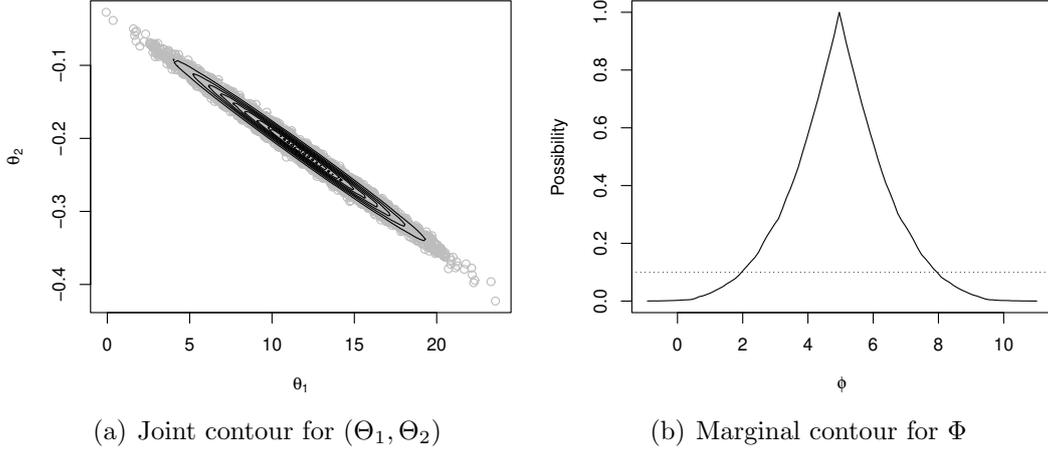


Figure 7: Panel (a) shows the samples of (Θ_1, Θ_2) from \mathbf{Q}_z^* and the corresponding approximate possibility contour. Panel (b) shows the approximate possibility contour for Φ , the log odds of an o-ring failure at 31 degrees.

$K = 9$ categories. In this form, the original data is

$$z^{\text{orig}} = (178, 138, 108, 570, 648, 442, 138, 252, 252),$$

so that $n^{\text{orig}} = 2726$. This is a large sample, making it difficult to visualize the results since the answers are very precise. So, instead, I down-scale this so the sample proportions are roughly the same but the overall sample size is small; in particular, I use

$$z = (10, 8, 6, 34, 38, 26, 8, 15, 15),$$

with corresponding sample size $n = 160$. The full possibilistic IM solution is obtained as mentioned above, resulting in samples of probability vectors from the probabilistic approximation \mathbf{Q}_z^* , which can then be transformed back into a possibility contour. This relatively high-dimensional solution is difficult to visualize so, for that reason, I'll consider a lower-dimensional feature:

$$\Phi = (\Theta_1 + \Theta_2 + \Theta_3, \Theta_4 + \Theta_5 + \Theta_6, \Theta_7 + \Theta_8 + \Theta_9).$$

In the original example, this corresponds to the marginal distribution for education level. The maximum likelihood estimator of Φ is

$$\hat{\phi}_z = (0.1500, 0.6125, 0.2375),$$

so that category 2 dominates on this margin. The results of applying the simple probabilistic marginalization that maps the samples of Θ from \mathbf{Q}_z^* to samples of Φ , and then transforming those samples of Φ into a possibility contour are shown in Figure 8.

E.3 Variance components

Random effect models are common across business, engineering, and science applications. The simplest such model is often expressed via the following data-generating equation,

$$Z_{ij} = \lambda + T_i + E_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, n,$$

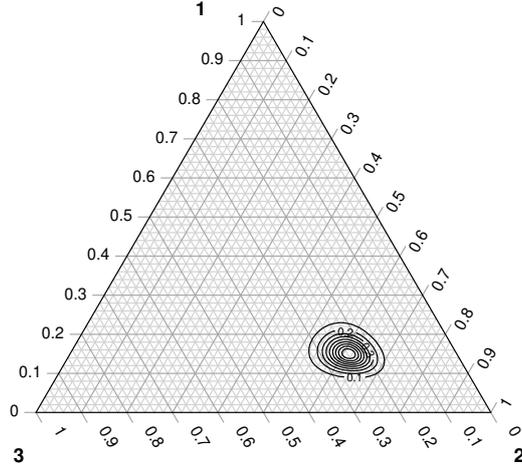


Figure 8: Level sets of the marginal possibilistic IM contour for Φ in the multinomial illustration described in Appendix E.2.

where $Z = (Z_{ij})$ is the observable data and T_i and E_{ij} are mutually independent random effects, with $E_{ij} \stackrel{\text{iid}}{\sim} \mathbf{N}(0, \phi_1)$ and $T_i \stackrel{\text{iid}}{\sim} \mathbf{N}(0, \phi_2)$. Then the three-dimensional parameter that indexes this model is $\theta = (\lambda, \phi_1, \phi_2)$, where λ is the overall mean and (ϕ_1, ϕ_2) is the pair of variance components associated with error/replication and treatment, respectively. The overall mean is not of primary interest, and it can be easily marginalized out, as shown below. The focus will be on the two variance components.

The likelihood function for this model (e.g., Tiao and Tan 1965, Eq. 1.4) can be expressed as

$$L(\lambda, \phi_1, \phi_2) = \phi_1^{-m(n-1)/2} (\phi_1 + n\phi_2)^{-m/2} \exp\left[-\frac{1}{2} \left\{ \frac{S_1}{\phi_1} + \frac{S_2 + mn(\bar{Z} - \lambda)^2}{\phi_1 + n\phi_2} \right\}\right],$$

where $\bar{Z} = (mn)^{-1} \sum_{j=1}^n \sum_{i=1}^m Z_{ij}$ is the overall sample mean and

$$S_1 = \sum_{j=1}^n \sum_{i=1}^m (Z_{ij} \bar{Z}_i)^2 \quad \text{and} \quad S_2 = n \sum_{i=1}^m (\bar{Z}_i - \bar{Z})^2,$$

with $\bar{Z}_i = n^{-1} \sum_{j=1}^n Z_{ij}$ the group-specific sample mean. The dependence on λ is very simple, so profile likelihood for (ϕ_1, ϕ_2) can be readily derived:

$$L^{\text{PR}}(\phi_1, \phi_2) = \phi_1^{-m(n-1)/2} (\phi_1 + n\phi_2)^{-m/2} \exp\left[-\frac{1}{2} \left\{ \frac{S_1}{\phi_1} + \frac{S_2}{\phi_1 + n\phi_2} \right\}\right].$$

Then the relative profile likelihood $R^{\text{PR}}(Z, \phi_1, \phi_2)$ is easy to obtain via (numerical) optimization. It's also clear that the distribution of the relative profile likelihood doesn't depend on the overall mean λ . Consequently, the computation

$$\pi_z^{\text{PR}}(\phi_1, \phi_2) = \mathbf{P}_{\lambda, \phi_1, \phi_2} \{R^{\text{PR}}(Z, \phi_1, \phi_2) \leq R^{\text{PR}}(z, \phi_1, \phi_2)\},$$

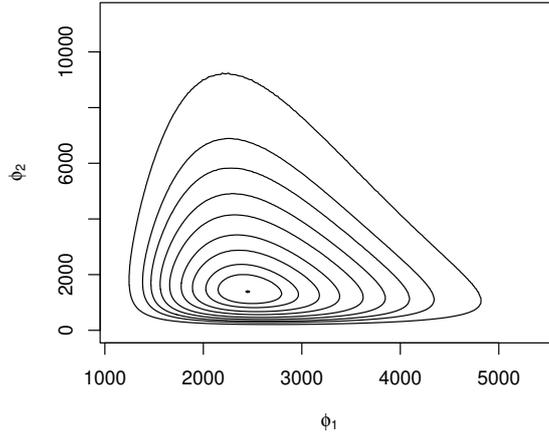


Figure 9: Plots of the marginal possibilistic IM contour for the variance components (ϕ_1, ϕ_2) as described in Appendix E.3.

can be carried out by fixing $\lambda = 0$, say. In what follows, I will employ the Bayesian-like sampling strategy using the basic elliptical approximations (applied on the log-variance scale) as described in Section 3.5 of the main text to evaluate the marginal possibilistic IM contour $\pi_z(\phi_1, \phi_2)$ for the variance components.

For illustration, I’ll reconsider the real-data example summarized in Tiao and Tan (1965). The experiment in question was designed to determine if variations in quality of an intermediate product affect the yield of dyestuff prepared from it. In this experiment, there are $m = 6$ samples of the intermediate product used and $n = 5$ preparations of the dyestuff made from each sample. The observations were measured in grams of standard colors. Figure 9 shows a plot of the marginal IM contour for (ϕ_1, ϕ_2) based on the data from this experiment. The point that represents the peak in the IM contour is the maximum likelihood estimator, $(\hat{\phi}_1, \hat{\phi}_2) = (2450.7, 1388.6)$.

F Eliminating nuisance parameters, cont.

The goal here is to showcase two different ways in which the (marginal) IM solution is superior to existing solutions when nuisance parameters are present.

Example (Behrens–Fisher). The classical *Behrens–Fisher problem* (Fisher 1935a, 1939) is simple to state: independent samples of size n_1 and n_2 , respectively, are available from two distinct normal populations, $\mathbf{N}(\Theta_{11}, \Theta_{12}^2)$ and $\mathbf{N}(\Theta_{21}, \Theta_{22}^2)$, with $\Theta = (\Theta_{11}, \Theta_{12}, \Theta_{21}, \Theta_{22})$ unknown, and the goal is marginal inference on the difference $\Phi = \Theta_{21} - \Theta_{11}$ of the two means. If the two variances are known or if their ratio is known, then the problem is straightforward; but there is no consensus on which solution is “right” or “best” when the variances are fully unknown. See Kim and Cohen (1998) for a survey.

The most common solution modifies the basic Student-t pivot with the degrees of freedom approximations due to Welch (1938, 1947); this is implemented in R’s `t.test` function. Other standard approaches include the simple-but-conservative solution proposed by Hsu (1938) and Scheffé (1970), and the Bayesian solution proposed by Jef-

Method	Coverage probability	Mean length
Bayes with Jeffreys prior	0.861	2.31
Bayes with right Haar prior	0.929	3.26
<code>t.test</code> with Welch df	0.881	2.79
Profile-based marginal IM	0.904	2.71

Table 1: Coverage probabilities and mean lengths for the 90% confidence intervals returned by the various methods. Standard errors for all of the coverage probabilities and mean lengths are roughly 0.003 and 0.01, respectively.

freys (1940), based on the right Haar prior for Θ , which is mathematically equivalent to Fisher’s fiducial solution. Ironically, the solution proposed by Jeffreys is different from the Bayesian solution based on the Jeffreys prior, both in its construction and its performance. A conservative IM solution was advanced in Martin and Liu (2015c), but the approach here, based on the strategy outlined in the main text, is different.

The approach here is based on validifying the profile relative likelihood. This was first proposed in Martin (2023b) and an illustration of its use in the Behrens–Fisher problem is considered in Example 5 there. This is a conceptually simple application of what was summarized in Section 5 in the main paper. The only wrinkle is that the relative profile likelihood doesn’t have a closed-form expression in the Behrens–Fisher problem, and its distribution depends on a nuisance parameter; this is precisely what makes the Behrens–Fisher problem challenging. This increases the computational cost, but I think the validity and efficiency achieved is worth it.

To showcase the marginal IM’s performance, I reproduce a simulation study presented in Martin (2025c). I focus on a rather severely imbalanced case—with $n_1 = 3$ and $n_2 = 20$ —to ensure that the differences in performance are apparent; this setting was one of the more challenging cases considered in Fraser et al. (2009). Otherwise, the simulation settings are standard: $\Theta_{11} = 2$, $\Theta_{21} = 0$, $\Theta_{12}^2 = 1$, and $\Theta_{22}^2 = 2$. I generated 10000 samples under this setup, and Table 1 presents the coverage probability and expected length of the various 90% confidence intervals for Φ . Notably, only the profile-based marginal IM solution is able to hit the target coverage probability, and, as desired, it’s more efficient than the valid-but-conservative Bayes/fiducial solution based on the right Haar prior.

Example (Nonparametric quantile). This example differs from the others in that it involves a nonparametric problem. To keep the notation consistent, let θ be an infinite-dimensional parameter such that \mathbf{P}_θ , for $\theta \in \mathbb{T}$, spans the entire class $\text{probs}^{\text{AC}}(\mathbb{Z})$ of absolutely continuous probability distributions supported on (subsets of) $\mathbb{Z} = \mathbb{R}$; therefore, the true Θ is in one-to-one correspondence with the underlying true distribution, which has a density with respect to Lebesgue measure on \mathbb{R} . The data Z consists of n iid observations from distribution \mathbf{P}_Θ and the goal is marginal inference on $\Phi = g(\Theta)$, the q^{th} quantile of \mathbf{P}_Θ for some fixed $q \in (0, 1)$. While there is no likelihood function for Φ in the usual sense, Wasserman (1990a, Theorem 5) showed that a relative profile likelihood—or empirical likelihood ratio (e.g., Owen 2001)—is well-defined and given by

$$R^{\text{PR}}(z, \phi) = \left(\frac{q}{v}\right)^v \left(\frac{1-q}{n-v}\right)^{n-v},$$

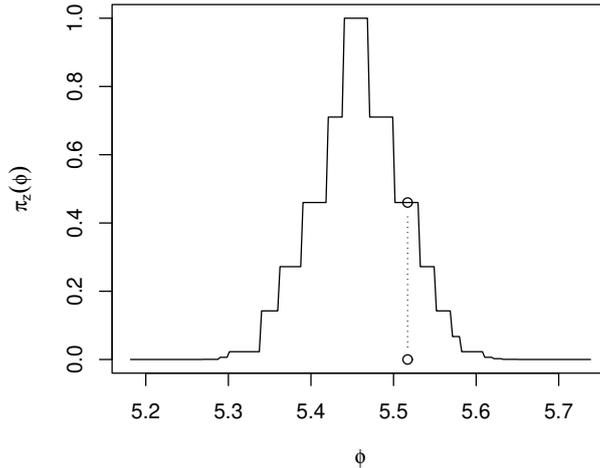


Figure 10: Plot of the marginal IM possibility contour for the median Φ , based on Cavendish’s data and the nonparametric relative profile likelihood from Wasserman (1990a). The vertical line highlights the possibility contour evaluated at the modern “true” value, 5.517, of the Earth’s density.

where

$$v = v_q(z, \phi) = \begin{cases} \sum_{i=1}^n 1(Z_i \leq \phi), & \phi \leq \hat{\phi}_z \\ \sum_{i=1}^n 1(Z_i < \phi), & \phi > \hat{\phi}_z, \end{cases}$$

and $\hat{\phi}_z$ is the q^{th} sample quantile. Following Martin (2015, Example 6), I propose to construct a marginal possibilistic IM contour $\pi_z^{\text{PR}}(\phi)$ exactly as in (13). For what it’s worth, a very similar nonparametric IM construction can be made based on the so-called *marginal likelihood* derived in Kalbfleisch (1978), but I didn’t find any significant differences between these two solutions.

For illustration, consider the $n = 29$ measurements of the density of the Earth relative to water, taken by Cavendish back in 1798 (Stigler 1977, Table 8). These experimental results were used to indirectly calculate the gravitational constant in Newton’s formula for the force of gravitational attraction between two bodies. Here I’ll treat the true density of the Earth as the median Φ of the population of such density measurements. As above, I construct a marginal IM contour for Φ based on Cavendish’s data, and the result is plotted in Figure 10. The modern “true” value of the Earth’s density is 5.517 and it’s clear from the plot that this true value is easily contained in the marginal IM’s 95% confidence interval for Φ .

Related to the second example above, some general comments can be made concerning how the IM’s uncertainty quantification framework differs from that of Bayes and fiducial. In particular, the quantile example reveals that the IM framework can, at least in certain cases, make direct inference about the target quantity of interest. From the remark quoted in Section 4.1 of the main text, Wasserman (2008) continues:

To estimate a quantile, an honest Bayesian needs to put a prior on the space of all distributions and then find the marginal posterior. The frequentist need

not care about the rest of the distribution and can focus on much simpler tasks.

Of course, Bayesians (e.g., Ferguson 1973; Ghosal and van der Vaart 2017) and (generalized) fiducialists (e.g., Cui and Hannig 2019) can solve nonparametric problems, but they do so by first making inference on the full model parameter—in this case, the unknown distribution function—and then marginalizing to the quantity of interest. As Wasserman mentions, frequentists can address the marginal inference problem directly, and apparently so too can IM users, at least in this quantile example. The point is that one doesn’t have to give up on data-dependent, fully conditional uncertainty quantification in exchange for a more appealing direct attack on the relevant unknowns.

G IM characterization of frequentism

The possibilistic IM formulation presented in the main text is intended to be normative; that is, I’m offering a rather general construction that is provably valid and at least approximately efficient in a wide range of applications. I started with a basic likelihood-based construction in Section 3 and that was generalized a bit in Section 6. It’s easy, however, to confuse normativity for rigidity, to think that “possibilistic IMs” are only those obtained through those constructions. From that perspective, one would surely believe that frequentist inference is more flexible than IMs, since there are effectively no rules dictating how the latter is carried out. But remember that a possibilistic IM is simply a mapping from data z (perhaps depending on a posited model, available prior information, or other things) to an imprecise-probabilistic output whose upper probability has the mathematical form of a possibility measure. There are, of course, lots of different ways this can be achieved, and fewer—but still many—that would satisfy the validity requirement. Some of these will be specifically good at answering very specific questions, but lousy at answering other questions. The point of focusing on the particular likelihood-based construction here and elsewhere in the literature is to offer a solution that’s generally good across all relevant questions. What might be surprising to the reader is that there are exactly as many valid, possibilistic IMs as there are frequentist solutions, as I show formally below. This justifies the claim made in the main text, namely, that there are no genuine frequentist solutions—including classical those in textbooks as well as those that haven’t even been developed yet—that are out of reach by the possibilistic IM framework. What’s notable about the result below is that, while the given frequentist procedure might be a test of a specific null hypothesis or a confidence set for a feature parameter Φ , the resulting possibilistic IM offers full uncertainty quantification about Θ . The possibilistic IM constructed in the proof below to agree with the given frequentist procedure may not answer every question about Θ efficiently, but that’s not relevant; after all, the frequentist procedure it starts with also can’t answer every question efficiently. The point is simply that whatever frequentist priorities one might have, there’s no loss of generality or efficiency focusing on possibilistic IMs.

Versions of the characterization result presented below have been given in Martin and Liu (2014) and Martin (2021a), with increasing generality. The result presented below is embarrassingly simple, and requiring only a basic “nestedness” condition on the frequentist procedure in consideration. This amounts to no practical restriction on

the characterization result because non-nested confidence sets or tests with non-nested rejection regions are difficult to interpret and rarely if ever used.

Proposition. Consider a model $\{\mathbf{P}_\theta : \theta \in \mathbb{T}\}$ for observable data Z with uncertain true parameter value $\Theta \in \mathbb{T}$.

(a) Let $H_0 \subset \mathbb{T}$ be a relevant null hypothesis and consider a test with a nested family of rejection regions $\{\mathcal{R}_\alpha : \alpha \in [0, 1]\}$, i.e., $\mathcal{R}_\alpha \subset \mathcal{R}_{\alpha'}$ for all $\alpha < \alpha'$, such that

$$\sup_{\theta \in H_0} \mathbf{P}_\theta(Z \in \mathcal{R}_\alpha) \leq \alpha, \quad \alpha \in [0, 1].$$

Then there exists a strongly valid possibilistic IM for Θ , with contour π_z and corresponding upper probability Π_z , such that

$$z \in \mathcal{R}_\alpha \iff \Pi_z(H_0) < \alpha.$$

(b) Let $\{C_\alpha^f(z) : \alpha \in [0, 1]\}$ be a family of confidence sets for $\Phi = f(\Theta)$, nested in the sense that $C_\alpha^f(z) \supset C_{\alpha'}^f(z)$ for all $\alpha < \alpha'$ and non-empty in the sense that $\bigcap_\alpha C_\alpha^f(z) \neq \emptyset$, with the property that

$$\sup_\theta \mathbf{P}_\theta\{C_\alpha^f(Z) \not\ni f(\theta)\} \leq \alpha, \quad \alpha \in [0, 1].$$

Then there exists a strongly valid possibilistic IM for Θ , with contour π_z , such that

$$f(\theta) \in C_\alpha^f(z) \iff \pi_z(\theta) > \alpha, \quad \alpha \in [0, 1].$$

Proof. Starting with Part (a), define the possibility contour π_z as

$$\pi_z(\theta) = \begin{cases} \sup\{\beta : z \notin \mathcal{R}_\beta\} & \theta \in H_0 \\ 1 & \theta \notin H_0. \end{cases}$$

Clearly π_z satisfies the normalization condition required to call it a possibility contour. Then define the corresponding possibility measure Π_z via optimization as usual:

$$\Pi_z(H) = \begin{cases} \sup\{\beta : z \notin \mathcal{R}_\beta\} & H \subseteq H_0 \\ 1 & H \not\subseteq H_0. \end{cases}$$

Then it's easy to see that

$$\Pi_z(H_0) < \alpha \iff \sup\{\beta : z \notin \mathcal{R}_\beta\} < \alpha \iff z \in \mathcal{R}_\alpha.$$

It's similarly easy to see that

$$\mathbf{P}_\theta\{\pi_z(\theta) < \alpha\} = \begin{cases} \mathbf{P}_\theta(Z \in \mathcal{R}_\alpha) & \theta \in H_0 \\ 0 & \theta \notin H_0. \end{cases}$$

From this and the assumed Type I error rate control property of the given frequentist test, the strong validity claim follows immediately.

For Part (b), start with defining a possibility contour π_z^f for Φ as follows:

$$\pi_z^f(\phi) = \sup\{\beta : C_\beta^f(z) \ni \phi\}, \quad \phi \in f(\mathbb{T}).$$

The non-emptiness of the given confidence sets implies that there exists a point at which π_z^f equals 1, hence the normalization condition is satisfied and it's fair to call this a "possibility contour." Now extend this to a possibility contour for Θ in the natural way:

$$\pi_z(\theta) := \pi_z^f(f(\theta)), \quad \theta \in \mathbb{T}.$$

Then it's clear that $f(\theta) \in C_\alpha^f(z)$ if and only if $\pi_z(\theta) > \alpha$ and, moreover, strong validity holds because

$$\mathbb{P}_\theta\{\pi_Z(\theta) \leq \alpha\} = \mathbb{P}_\theta\{C_\alpha^f(Z) \not\ni f(\theta)\} \leq \alpha,$$

where the right-most inequality holds by the assumed coverage properties of the C_α^f 's. \square

H General IM construction and properties

The main text focused primarily on IM constructions for a finite-dimensional parameter indexing the posited statistical model. The reason for this focus is purely practical: most real-world applications are of this type. But these are not the only cases that the IM framework can handle, and the later sections of the paper briefly mentioned various extensions. Since this review paper is intended both to showcase the developments that have been made so far and to highlight various opportunities for the future of IMs, I think it's worth sketching some of these details, even if it's a bit beyond the paper's scope.

The general setup in Section 6 of the main paper suggested a possibilistic IM contour $\pi_z(\theta)$ for the relevant feature $\Theta = \tau(\mathbf{P})$ of the underlying distribution $\mathbf{P} \in \mathcal{P}$ as

$$\pi_z(\theta) = \sup_{\mathbf{P} \in \mathcal{P} : \tau(\mathbf{P}) = \theta} \mathbb{P}\{\rho(Z, \theta) \leq \rho(z, \theta)\}, \quad \theta \in \mathbb{T}.$$

Of course, the corresponding possibility measure or upper probability is defined via optimization just like before: $\Pi_z(H) = \sup_{\theta \in H} \pi_z(\theta)$, for any $H \subseteq \mathbb{T}$. Accurate and efficient computation of the IM contour in this very general case is an important open question, so I don't have anything to say about this point here. My focus in this present discussion is on what kind of properties the ranking function ρ must satisfy in order for the IM constructed above to be valid and efficient.

One basic requirement is that π_z as constructed above should be such that $\sup_\theta \pi_z(\theta) = 1$ for each z . A sufficient condition is that the maximum $\max_\theta \rho(z, \theta)$ be independent of z . This can be achieved, for example, if ρ is such that $\theta\rho(z, \theta)$ is a possibility contour for each z —that is, if $\sup_\theta \rho(z, \theta) = 1$ for each z . The relative likelihood and its variations satisfy this condition, but so do many other functions. To simplify the discussion that follows, I'll assume that ρ takes values between 0 and 1, again, like the relative likelihood and its variants.

Validity, too, only requires minimal assumptions on ρ . In particular, the distribution of the (scalar) random variable $\rho(Z, \theta)$ should not have an atom at 0 under any distribution \mathbf{P} with $\tau(\mathbf{P}) = \theta$. The reason is that this will imply that $\pi_z(\theta)$ is bounded away from 0 and, therefore, the random variable $\pi_Z(\theta)$ can't be stochastically no smaller

than $\text{Unif}(0, 1)$. Any practically reasonable choice of ranking function would achieve this since, otherwise, there'd be positive probability of observing data that's fully incompatible (relative to ρ) with the true parameter, which clearly makes ρ a poor measure of compatibility. In any case, mathematically, no atom at 0 in the sense above implies that the distribution function $G_{\mathbf{P}}$ of $\rho(Z, \tau(\mathbf{P}))$ under $Z \sim \mathbf{P}$ satisfies $\sup_{\mathbf{P} \in \mathcal{P}} G_{\mathbf{P}}(0) = 0$.

Concerning efficiency, let me turn to large-sample considerations. For this, write Z^n for an iid sample of size n from unknown distribution \mathbf{P} . A first basic question is what does it take for the possibilistic IM to be *consistent*, i.e.,

$$\Pi_{Z^n}(H) \rightarrow 0 \quad \text{in } \mathbf{P}\text{-probability as } n \rightarrow \infty \quad \text{for each } H \not\supseteq \tau(\mathbf{P}).$$

This is analogous to Bayesian (and fiducial) consistency as explored in, e.g., Barron et al. (1999), Ghosal et al. (2000), and Walker (2004). To achieve this, the ranking function must be able to distinguish “good” and “bad” values of the interest parameter θ as efficiently as possible. In light of the above constraints, and the exponential growth rate that can be achieved in many problems (e.g., Cover and Thomas 2006, Ch. 11), there's effectively no choice but to take ρ of the form

$$\rho(z^n, \theta) = \frac{\prod_{i=1}^n f(z_i, \theta)}{\sup_{\vartheta} \prod_{i=1}^n f(z_i, \vartheta)} = \exp \left[\sum_{i=1}^n \{ \log f(z_i, \theta) - \log f(z_i, \hat{\theta}_{z^n}) \} \right],$$

where $\hat{\theta}_{z^n} = \arg \max_{\vartheta} \prod_{i=1}^n f(z_i, \vartheta)$, for suitable function f . So, either directly or indirectly, explicitly or implicitly, $\Theta = \tau(\mathbf{P})$ corresponds to the maximizer of the function $\theta \mapsto \mathbf{E} \log f(Z_1, \theta)$. This is the case with the relative likelihood as discussed in the main text, since the true parameter is the maximizer of the expected log density function.

Let $\Theta = \tau(\mathbf{P})$ be the true value of the interest parameter and take $\theta \neq \Theta$; then θ is a “wrong” value of the parameter and the IM is expected to rule this value out as the data grows more and more informative, i.e., $\pi_{Z^n}(\theta) \rightarrow 0$ in \mathbf{P} -probability as $n \rightarrow \infty$; a result of this type was established in Martin (2015, Theorem 3). Towards this, since $\hat{\theta}_{Z^n}$ is, by definition, a maximizer, it follows that

$$\rho(Z^n, \theta) \leq \exp \left[\sum_{i=1}^n \{ \log f(Z_i, \theta) - \log f(Z_i, \Theta) \} \right].$$

Under mild integrability conditions, the law of large numbers implies

$$\frac{1}{n} \sum_{i=1}^n \log \frac{f(Z_i, \theta)}{f(Z_i, \Theta)} \rightarrow \mathbf{E} \log \frac{f(Z_1, \theta)}{f(Z_1, \Theta)}, \quad \text{in } \mathbf{P}\text{-probability.}$$

Note that the right-hand side above is negative by definition. Then, by monotonicity of the distribution function,

$$\pi_{Z^n}(\theta) = \sup_{\mathbf{P} \in \mathcal{P}: \tau(\mathbf{P}) = \theta} G_{\mathbf{P}} \{ \rho(Z^n, \theta) \} \leq \sup_{\mathbf{P} \in \mathcal{P}: \tau(\mathbf{P}) = \theta} G_{\mathbf{P}} \{ e^{\sum_{i=1}^n \{ \log f(Z_i, \theta) - \log f(Z_i, \Theta) \}} \}.$$

Clearly, the argument to the distribution function behaves like $e^{-n\gamma}$ for some $\gamma > 0$, hence is converging to 0 very rapidly. Since there's no atom at 0, the entire right-hand side is vanishing in \mathbf{P} -probability. Therefore, $\pi_{Z^n}(\theta) \rightarrow 0$ in \mathbf{P} -probability for any $\theta \neq \Theta$.

Extension to the proper consistency property as stated above involves establishing that the aforementioned property holds appropriately uniformly in θ . I won't get into these details here but surely this is doable and would make an interesting contribution.

Beyond the basic consistency result discussed above, it would be desirable to have concentration rate and generalizations of the Bernstein–von Mises property due to Martin and Williams (2025) reviewed in the main text. These details are beyond the scope of high-level discussion here, but these too are deserving of further exploration.

Finally, as I mentioned briefly in Section 7 of the main text, there are even more general IM constructions that can incorporate any available, perhaps incomplete prior information about the relevant unknowns. These details are too new and, therefore, beyond the scope of this review paper, but they can be found in the working papers Martin (2022a,b, 2023b). This incorporation of prior information is done in a novel, non-Bayesian way that allows for efficiency gains akin to those obtained by more familiar forms of regularization but while retaining a generalized notion of validity that's meaningful relative to the (incomplete) prior knowledge available. A conjecture is that this non-Bayesian style of prior-to-posterior updating is more efficient in the sense that it can avoid the undesirable dilation phenomenon that plagues generalized Bayes solutions (e.g., Gong and Meng 2021; Kyburg 1987; Walley 1991).

References

- Agresti, A. (2002). *Categorical Data Analysis*. Wiley-Interscience [John Wiley & Sons], New York, second edition.
- Awan, J. and Wang, Z. (2024). Simulation-based, finite-sample inference for privatized data. *J. Amer. Statist. Assoc.*, to appear; [arXiv:2303.05328](https://arxiv.org/abs/2303.05328).
- Balch, M. S., Martin, R., and Ferson, S. (2019). Satellite conjunction analysis and the false confidence theorem. *Proc. Royal Soc. A*, 475(2227):2018.0565.
- Barron, A., Schervish, M. J., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27(2):536–561.
- Basu, D. (1977). On the elimination of nuisance parameters. *J. Amer. Statist. Assoc.*, 72(358):355–366.
- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philos. Trans. Roy. Soc.*, pages 330–418.
- Berger, J. (1984). The robust Bayesian viewpoint. In *Robustness of Bayesian Analyses*, volume 4 of *Stud. Bayesian Econometrics*, pages 63–144. North-Holland, Amsterdam. With comments and with a reply by the author.
- Berger, J., Bernardo, J., and Sun, D. (2024). *Objective Bayesian Inference*. World Scientific Publishing Co.
- Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *J. Amer. Statist. Assoc.*, 89(427):1012–1016.

- Buehler, R. J. and Fedderson, A. P. (1963). Note on a conditional property of Student's t . *Ann. Math. Statist.*, 34(3):1098–1100.
- Cahoon, J. and Martin, R. (2020). A generalized inferential model for meta-analyses based on few studies. *Statistics and Applications*, 18(2):299–316.
- Cahoon, J. and Martin, R. (2021). Generalized inferential models for censored data. *Internat. J. Approx. Reason.*, 137:51–66.
- Caprio, M., Sale, Y., and Hüllermeier, E. (2025). Conformal prediction regions are imprecise highest density regions. [arXiv:2502.06331](https://arxiv.org/abs/2502.06331).
- Cella, L. and Martin, R. (2022a). Direct and approximately valid probabilistic inference on a class of statistical functionals. *Internat. J. Approx. Reason.*, 151:205–224.
- Cella, L. and Martin, R. (2022b). Valid inferential models for prediction in supervised learning problems. *Internat. J. Approx. Reason.*, 150:1–18.
- Cella, L. and Martin, R. (2022c). Validity, consonant plausibility measures, and conformal prediction. *Internat. J. Approx. Reason.*, 141:110–130.
- Cella, L. and Martin, R. (2023). Possibility-theoretic statistical inference offers performance and probativeness assurances. *Internat. J. Approx. Reason.*, 163:109060.
- Cella, L. and Martin, R. (2025). Computationally efficient variational-like approximations of possibilistic inferential models. *Internat. J. Approx. Reason.*, 186:Paper No. 109506.
- Choquet, G. (1953–1954). Theory of capacities. *Ann. Inst. Fourier, Grenoble*, 5:131–295 (1955).
- Couso, I., Montes, S., and Gil, P. (2001). The necessity of the strong α -cuts of a fuzzy set. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems*, 9(2):249–262.
- Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition.
- Cui, Y. and Hannig, J. (2019). Nonparametric generalized fiducial inference for survival functions under censoring. *Biometrika*, 106(3):501–518.
- Cui, Y. and Hannig, J. (2024). Demystifying inferential models: A fiducial perspective. *Statist. Sci.*, to appear; [arXiv:2205.05612](https://arxiv.org/abs/2205.05612).
- Cuzzolin, F. (2021). *The Geometry of Uncertainty*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, Cham.
- Datta, G. S. and Ghosh, J. K. (1995). On priors providing frequentist validity for Bayesian inference. *Biometrika*, 82(1):37–45.
- Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*, volume 1. Cambridge University Press, Cambridge.

- Dawid, A. P., Stone, M., and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc. Ser. B*, 35:189–233. With discussion and reply by the authors.
- De Cooman, G. (1997a). Possibility theory. I. The measure- and integral-theoretic groundwork. *Internat. J. Gen. Systems*, 25(4):291–323.
- De Cooman, G. (1997b). Possibility theory. II. Conditional possibility. *Internat. J. Gen. Systems*, 25(4):325–351.
- De Cooman, G. and Aeyels, D. (1999). Supremum preserving upper probabilities. *Inform. Sci.*, 118(1-4):173–212.
- de Finetti, B. (1990). *Theory of Probability. Vol. 1*. Wiley Classics Library. John Wiley & Sons, Ltd., Chichester.
- Dempster, A. P. (1963). Further examples of inconsistencies in the fiducial argument. *Ann. Math. Statist.*, 34:884–891.
- Dempster, A. P. (1964). On the difficulties inherent in Fisher’s fiducial argument. *J. Amer. Statist. Assoc.*, 59:56–66.
- Dempster, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Ann. Math. Statist.*, 37:355–374.
- Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.*, 38:325–339.
- Dempster, A. P. (1968). A generalization of Bayesian inference. (With discussion). *J. Roy. Statist. Soc. Ser. B*, 30:205–247.
- Dempster, A. P. (1971). Model searching and estimation in the logic of inference. In Godambe, V. P. and Sprott, D. A., editors, *Foundations of Statistical Inference*, pages 56–78. Holt, Rinehart and Winston, Toronto.
- Dencœux, T. (2006). Constructing belief functions from sample data using multinomial confidence regions. *Internat. J. of Approx. Reason.*, 42(3):228–252.
- Dencœux, T. (2014). Likelihood-based belief function: justification and some extensions to low-quality data. *Internat. J. Approx. Reason.*, 55(7):1535–1547.
- Dubois, D. (2006). Possibility theory and statistical reasoning. *Comput. Statist. Data Anal.*, 51(1):47–69.
- Dubois, D., Foulloy, L., Mauris, G., and Prade, H. (2004). Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliab. Comput.*, 10(4):273–297.
- Dubois, D. and Prade, H. (1986). A set-theoretic view of belief functions. Logical operations and approximations by fuzzy sets. *Internat. J. Gen. Systems*, 12(3):193–226.

- Dubois, D. and Prade, H. (1988). *Possibility Theory*. Plenum Press, New York.
- Dubois, D. and Prade, H. (1990). Consonant approximations of belief functions. *Internat. J. Approx. Reason.*, 4(5-6):419–449.
- Dufour, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica*, 65(6):1365–1387.
- Eaton, M. L. (1989). *Group Invariance Applications in Statistics*. Institute of Mathematical Statistics, Hayward, CA.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26.
- Efron, B. (1998). R. A. Fisher in the 21st century. *Statist. Sci.*, 13(2):95–122.
- Efron, B. (2013). Discussion: “Confidence distribution, the frequentist distribution estimator of a parameter: a review” [mr3047496]. *Int. Stat. Rev.*, 81(1):41–42.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.
- Faraway, J. (2022). *faraway: Functions and Datasets for Books by Julian Faraway*. R package version 1.0.8.
- Faulkenberry, G. D. (1973). A method of obtaining prediction intervals. *J. Amer. Statist. Assoc.*, 68:433–435.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230.
- Fisher, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 26:528–535.
- Fisher, R. A. (1933). The concepts of inverse probability and fiducial probability referring to unknown parameters. *Proc. R. Soc. Lond. A.*, 139:343–348.
- Fisher, R. A. (1935a). The fiducial argument in statistical inference. *Ann. Eugenics*, 6:391–398.
- Fisher, R. A. (1935b). The logic of inductive inference. *J. Roy. Statist. Soc.*, 98:39–82.
- Fisher, R. A. (1935c). *The Design of Experiments*. Oliver and Boyd, Edinburgh.
- Fisher, R. A. (1939). The comparison of samples with possibly unequal variances. *Ann. Eugenics*, 9:174–180.
- Fisher, R. A. (1973). *Statistical Methods and Scientific Inference*. Hafner Press, New York, 3rd edition.
- Fraser, D. A. S. (1968). *The Structure of Inference*. John Wiley & Sons Inc., New York.

- Fraser, D. A. S. (2011a). Is Bayes posterior just quick and dirty confidence? *Statist. Sci.*, 26(3):299–316.
- Fraser, D. A. S. (2011b). Rejoinder: “Is Bayes posterior just quick and dirty confidence?”. *Statist. Sci.*, 26(3):329–331.
- Fraser, D. A. S. (2013). Discussion: “Confidence distribution, the frequentist distribution estimator of a parameter: a review” [mr3047496]. *Int. Stat. Rev.*, 81(1):42–48.
- Fraser, D. A. S. (2014). Why does statistics have two theories? In Lin, X., Genest, C., Banks, D. L., Molenberghs, G., Scott, D. W., and Wang, J.-L., editors, *Past, Present, and Future of Statistical Science*, chapter 22. Chapman & Hall/CRC Press.
- Fraser, D. A. S., Bédard, M., Wong, A., Lin, W., and Fraser, A. M. (2016). Bayes, reproducibility and the quest for truth. *Statist. Sci.*, 31(4):578–590.
- Fraser, D. A. S., Wong, A., and Sun, Y. (2009). Three enigmatic examples and inference from likelihood. *Canad. J. Statist.*, 37(2):161–181.
- Garfinkel, S. L. (2025). *Differential Privacy*. The MIT Press, Cambridge, Massachusetts.
- Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica*, 6(4):733–807. With comments and a rejoinder by the authors.
- Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.
- Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*, volume 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.
- Gigerenzer, G. and Murray, D. J. (1987). *Cognition as Intuitive Statistics*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.
- Gleser, L. J. and Hwang, J. T. (1987). The nonexistence of $100(1 - \alpha)\%$ confidence sets of finite expected diameter in errors-in-variables and related models. *Ann. Statist.*, 15(4):1351–1362.
- Gong, R. and Meng, X.-L. (2021). Judicious judgment meets unsettling updating: dilation, sure loss and Simpson’s paradox. *Statist. Sci.*, 36(2):169–190.
- Hacking, I. (1976). *Logic of Statistical Inference*. Cambridge University Press, Cambridge-New York-Melbourne.
- Hamada, M., Johnson, V., Moore, L. M., and Wendelberger, J. (2004). Bayesian prediction intervals and their relationship to tolerance intervals. *Technometrics*, 46(4):452–459.
- Han, Y. and Lee, T. C. M. (2022). Uncertainty quantification for sparse estimation of spectral lines. *IEEE Trans. Signal Process.*, 70:6243–6256.

- Hannig, J., Iyer, H., Lai, R. C. S., and Lee, T. C. M. (2016). Generalized fiducial inference: a review and new results. *J. Amer. Statist. Assoc.*, 111(515):1346–1361.
- Hannig, J. and Lee, T. C. M. (2009). Generalized fiducial inference for wavelet regression. *Biometrika*, 96(4):847–860.
- He, X., Madigan, D., Yu, B., and Wellner, J. (2025). Statistics at a crossroads; who is for the challenge? [arXiv:2503.22945](https://arxiv.org/abs/2503.22945).
- Hector, E. C., Cella, L., and Martin, R. (2025). Divide-and-conquer with finite sample sizes: valid and efficient possibilistic inference. [arXiv:2503.22812](https://arxiv.org/abs/2503.22812).
- Hoff, P. (2023). Bayes-optimal prediction with frequentist coverage control. *Bernoulli*, 29(2):901–928.
- Hose, D. (2022). *Possibilistic Reasoning with Imprecise Probabilities: Statistical Inference and Dynamic Filtering*. PhD thesis, University of Stuttgart. https://dominikhose.github.io/dissertation/diss_dhose.pdf.
- Hose, D. and Hanss, M. (2021). A universal approach to imprecise probabilities in possibility theory. *Internat. J. Approx. Reason.*, 133:133–158.
- Hose, D., Hanss, M., and Martin, R. (2022). A practical strategy for valid partial prior-dependent possibilistic inference. In Le Hégarat-Masclé, S., Bloch, I., and Aldea, E., editors, *Belief Functions: Theory and Applications (BELIEF 2022)*, volume 13506 of *Lecture Notes in Artificial Intelligence*, pages 197–206. Springer.
- Hsu, P. L. (1938). Contributions to the theory of “Student’s” t -test as applied to the problem of two samples. In *Statistical Research Memoirs*, pages 1–24. University College, London.
- Huber, P. J. (1973). The use of Choquet capacities in statistics. *Bull. Inst. Internat. Statist.*, 45(book 4):181–191. With discussion by P. Gänszler, W. S. Hall, L. Schmetterer, Charles Stein, and G. N. Wilkinson and with reply by the author.
- Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference—for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, New York.
- Jacob, P. E., Gong, R., Edlefsen, P. T., and Dempster, A. P. (2021). A Gibbs sampler for a class of random convex polytopes. *J. Amer. Statist. Assoc.*, 116(535):1181–1192.
- Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge.
- Jeffreys, H. (1940). Note on the Behrens-Fisher formula. *Ann. Eugenics*, 10:48–51.
- Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. London Ser. A*, 186:453–461.

- Jeffreys, H. (1998). *Theory of Probability*. Oxford Classic Texts in the Physical Sciences. The Clarendon Press, Oxford University Press, New York. Reprint of the 1983 edition.
- Juslin, P., Winman, A., and Hansson, P. (2007). The naïve intuitive statistician. *Psychol. Rev.*, 114(3):678–703.
- Kalbfleisch, J. D. (1978). Likelihood methods and nonparametric tests. *J. Amer. Statist. Assoc.*, 73(361):167–170.
- Keynes, J. M. (1921). *A Treatise on Probability*. Macmillan and Co.
- Kim, S.-H. and Cohen, A. S. (1998). On the Behrens-Fisher problem: A review. *Journal of Educational and Behavioral Statistics*, 23(4):356–377.
- Kyburg, Jr., H. E. (1987). Bayesian and non-Bayesian evidential updating. *Artificial Intelligence*, 31(3):271–293.
- Lai, R. C. S., Hannig, J., and Lee, T. C. M. (2015). Generalized fiducial inference for ultrahigh dimensional regression. *J. Amer. Statist. Assoc.*, 110:760–772.
- Laplace, P. S. (1812). *Théorie Analytique des Probabilités*. Courcier, Paris.
- Le Cam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.*, 41:802–828.
- Levi, I. (1980). *The Enterprise of Knowledge*. The MIT Press, Boston.
- Lindley, D. V. (1958). Fiducial distributions and Bayes’ theorem. *J. Roy. Statist. Soc. Ser. B*, 20:102–107.
- Liu, C. and Martin, R. (2024). Inferential models and possibility measures. In Berger, J., Meng, X.-L., Reid, N., and Xie, M.-g., editors, *Handbook of Bayesian, Fiducial, and Frequentist Inference*, pages 344–363. Chapman & Hall/CRC Press.
- Liu, Y. and Williams, J. P. (2025). An improved solution to the two normal means problem via regularization. [arXiv:2508.13012](https://arxiv.org/abs/2508.13012).
- Martin, R. (2015). Plausibility functions and exact frequentist inference. *J. Amer. Statist. Assoc.*, 110(512):1552–1561.
- Martin, R. (2018). On an inferential model construction using generalized associations. *J. Statist. Plann. Inference*, 195:105–115.
- Martin, R. (2019). On valid uncertainty quantification about a model. In De Bock, J., de Campos, C. P., de Cooman, G., Quaeghebeur, E., and Wheeler, G., editors, *Proceedings of the Eleventh International Symposium on Imprecise Probabilities: Theories and Applications*, volume 103 of *Proceedings of Machine Learning Research*, pages 295–303, Thagaste, Ghent, Belgium. PMLR.
- Martin, R. (2021a). An imprecise-probabilistic characterization of frequentist statistical inference. [arXiv:2112.10904](https://arxiv.org/abs/2112.10904).

- Martin, R. (2021b). Inferential models and the decision-theoretic implications of the validity property. [arXiv:2112.13247](#).
- Martin, R. (2022a). Valid and efficient imprecise-probabilistic inference with partial priors, I. First results. [arXiv:2203.06703](#).
- Martin, R. (2022b). Valid and efficient imprecise-probabilistic inference with partial priors, II. General framework. [arXiv:2211.14567](#).
- Martin, R. (2023a). Fiducial inference viewed through a possibility-theoretic inferential model lens. In Miranda, E., Montes, I., Quaeghebeur, E., and Vantaggi, B., editors, *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications*, volume 215 of *Proceedings of Machine Learning Research*, pages 299–310. PMLR.
- Martin, R. (2023b). Valid and efficient imprecise-probabilistic inference with partial priors, III. Marginalization. [arXiv:2309.13454](#).
- Martin, R. (2024a). A possibility-theoretic solution to Basu’s Bayesian–frequentist via media. *Sankhya A*, 86:43–70.
- Martin, R. (2024b). Which statistical hypotheses are afflicted by false confidence? In Bi, Y., Jousselme, A.-L., and Denoeux, T., editors, *BELIEF 2024*, volume 14909 of *Lecture Notes in Artificial Intelligence*, pages 140–149, Switzerland. Springer Nature.
- Martin, R. (2025a). Decision-theoretic properties of possibilistic inferential models. In Destercke, S., Erreygers, A., Nendel, M., Riedel, F., and Troffaes, M. C. M., editors, *Proceedings of the Fourteenth International Symposium on Imprecise Probabilities: Theories and Applications*, volume 290 of *Proceedings of Machine Learning Research*, pages 183–195. PMLR.
- Martin, R. (2025b). An efficient Monte Carlo method for valid prior-free possibilistic statistical inference. [arXiv:2501.10585](#).
- Martin, R. (2025c). No-prior Bayesian inference reIMagined: probabilistic approximations of possibilistic inferential models. [arXiv:2503.19748](#).
- Martin, R. and Liu, C. (2013). Inferential models: a framework for prior-free posterior probabilistic inference. *J. Amer. Statist. Assoc.*, 108(501):301–313.
- Martin, R. and Liu, C. (2014). A note on p-values interpreted as plausibilities. *Statist. Sinica*, 24(4):1703–1716.
- Martin, R. and Liu, C. (2015a). Conditional inferential models: combining information for prior-free probabilistic inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 77(1):195–217.
- Martin, R. and Liu, C. (2015b). *Inferential Models*, volume 147 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL.

- Martin, R. and Liu, C. (2015c). Marginal inferential models: prior-free probabilistic inference on interest parameters. *J. Amer. Statist. Assoc.*, 110(512):1621–1631.
- Martin, R. and Williams, J. P. (2025). Asymptotic efficiency of inferential models and a possibilistic Bernstein–von Mises theorem. *Internat. J. Approx. Reason.*, 180:Paper No. 109389.
- Mayo, D. (2014). On the Birnbaum argument for the strong likelihood principle. *Statist. Sci.*, 29(2):227–239.
- Mayo, D. G. (2018). *Statistical Inference as Severe Testing*. Cambridge University Press, Cambridge.
- Murph, A., Hannig, J., and Williams, J. P. (2024). Introduction to generalized fiducial inference. In Berger, J., Meng, X.-L., Reid, N., and Xie, M.-g., editors, *Handbook of Bayesian, Fiducial, and Frequentist Inference*, pages 276–299. Chapman & Hall/CRC Press.
- Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16:1–32.
- Owen, A. B. (2001). *Empirical Likelihood*, volume 92 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC Press, Boca Raton, FL.
- Pearl, J. (2009). *Causality*. Cambridge University Press, Cambridge, second edition. Models, Reasoning, and Inference.
- Pearl, J. (2018). *The Book of Why*. Basic Books, New York. The New Science of Cause and Effect.
- Ramsey, F. P. (1926). Truth and probability. In Braithwaite, R. B., editor, *The Foundations of Mathematics and other Logical Essays*, pages 156–198. Routledge, London.
- Reid, N. and Cox, D. R. (2015). On some principles of statistical inference. *Int. Stat. Rev.*, 83(2):293–308.
- Savage, L. J. (1961). The foundations of statistics reconsidered. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 575–586. Univ. California Press, Berkeley, Calif.
- Savage, L. J. (1972). *The Foundations of Statistics*. Dover Publications, Inc., New York, revised edition.
- Savage, L. J. (1976). On rereading R. A. Fisher. *Ann. Statist.*, 4(3):441–500. Edited posthumously by John W. Pratt, With a discussion by B. Efron, Churchill Eisenhart, Bruno de Finetti, D. A. S. Fraser, V. P. Godambe, I. J. Good, O. Kempthorne and Stephen M. Stigler.
- Scheffé, H. (1970). Practical solutions of the Behrens–Fisher problem. *J. Amer. Statist. Assoc.*, 65:1501–1508.

- Schervish, M. J. (1995). *Theory of Statistics*. Springer-Verlag, New York.
- Schweder, T. and Hjort, N. L. (2013). Discussion: “Confidence distribution, the frequentist distribution estimator of a parameter: a review”. *Int. Stat. Rev.*, 81(1):56–68.
- Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability*, volume 41 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, New York.
- Severini, T. A. (1993). Local ancillarity in the presence of a nuisance parameter. *Biometrika*, 80(2):305–320.
- Severini, T. A. (1994). On the approximate elimination of nuisance parameters by conditioning. *Biometrika*, 81(4):649–661.
- Severini, T. A. (1998). Likelihood functions for inference in the presence of a nuisance parameter. *Biometrika*, 85(3):507–522.
- Shackle, G. L. S. (1961). *Decision Order and Time in Human Affairs*. Cambridge University Press, Cambridge.
- Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J.
- Shafer, G. (1982). Belief functions and parametric models. *J. Roy. Statist. Soc. Ser. B*, 44(3):322–352. With discussion.
- Shafer, G. (1987). Belief functions and possibility measures. In Bezdek, J. C., editor, *The Analysis of Fuzzy Information, Vol. 1: Mathematics and Logic*, pages 51–84. CRC.
- Shafer, G. (2007). From Cournot’s principle to market efficiency. In Touffut, J.-P., editor, *Augustin Cournot: Modelling Economics*, pages 55–95. Edward Elgar.
- Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:371–421.
- Shafer, G. and Vovk, V. (2019). *Game-Theoretic Foundations for Probability and Finance*. Wiley, Hoboken, NJ.
- Shi, W. J., Hannig, J., Lai, R. C. S., and Lee, T. C. M. (2021). Covariance estimation via fiducial inference. *Stat. Theory Relat. Fields*, 5(4):316–331.
- Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.*, 30:877–880.
- Stigler, S. M. (1977). Do robust estimators work with real data? *Ann. Statist.*, 5(6):1055–1098. With discussion and a reply by the author.
- Stigler, S. M. (1986). *The History of Statistics*. The Belknap Press of Harvard University Press, Cambridge, MA. The Measurement of Uncertainty before 1900.

- Su, Y., Hannig, J., and Lee, T. C. M. (2022). Uncertainty quantification in graphon estimation using generalized fiducial inference. *IEEE Trans. Signal Inform. Process. Netw.*, 8:597–609.
- Taraldsen, G. and Lindqvist, B. H. (2013). Fiducial theory and optimal inference. *Ann. Statist.*, 41(1):323–341.
- Tiao, G. C. and Tan, W.-Y. (1965). Bayesian analysis of random-effect models in the analysis of variance. I. Posterior distribution of variance-components. *Biometrika*, 52:37–53.
- Troffaes, M. C. M. and de Cooman, G. (2014). *Lower Previsions*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.
- Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer, New York.
- Vovk, V. G. (1993). A logic of probability, with application to the foundations of statistics. *J. Roy. Statist. Soc. Ser. B*, 55(2):317–351. With discussion and a reply by the author.
- Walker, S. (2004). New approaches to Bayesian consistency. *Ann. Statist.*, 32(5):2028–2043.
- Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*, volume 42 of *Monographs on Statistics and Applied Probability*. Chapman & Hall Ltd., London.
- Walley, P. (1997). Statistical inferences based on a second-order possibility distribution. *Internat. J. Gen. Systems*, 26(4):337–383.
- Walley, P. (2002). Reconciling frequentist properties with the likelihood principle. *J. Statist. Plann. Inference*, 105(1):35–65.
- Wasserman, L. (2008). Why isn’t everyone a Bayesian? In Morris, C. and Tibshirani, R., editors, *The Science of Bradley Efron*, pages 260–261. Springer, New York.
- Wasserman, L. A. (1990a). Belief functions and statistical inference. *Canad. J. Statist.*, 18(3):183–196.
- Wasserman, L. A. (1990b). Prior envelopes based on belief functions. *Ann. Statist.*, 18(1):454–464.
- Wasserman, L. A. and Kadane, J. B. (1990). Bayes’ theorem for Choquet capacities. *Ann. Statist.*, 18(3):1328–1339.
- Wei, Z. and Lee, T. C. M. (2023). High-dimensional multi-task learning using multivariate regression and generalized fiducial inference. *J. Comput. Graph. Statist.*, 32(1):226–240.
- Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29:350–362.
- Welch, B. L. (1947). The generalization of ‘Student’s’ problem when several different population variances are involved. *Biometrika*, 34:28–35.

- Welch, B. L. and Peers, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. Ser. B*, 25:318–329.
- Williams, J. P. (2023). Model-free generalized fiducial inference. [arXiv:2307.12472](https://arxiv.org/abs/2307.12472).
- Williams, J. P. and Hannig, J. (2019). Nonpenalized variable selection in high-dimensional linear model settings via generalized fiducial inference. *Ann. Statist.*, 47(3):1723–1753.
- Wu, S., Hannig, J., and Lee, T. C. M. (2021). Uncertainty quantification for principal component regression. *Electron. J. Stat.*, 15(1):2157–2178.
- Xie, M. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review. *Int. Stat. Rev.*, 81(1):3–39.
- Xie, M. and Wang, P. (2022). Repro samples method for finite- and large-sample inferences. [arXiv:2206.06421](https://arxiv.org/abs/2206.06421).
- Xiong, S. (2017). Local optimization-based statistical inference. *Electron. J. Stat.*, 11(1):2295–2320.
- Yang, J., Wang, X., and Liu, C. (2023). Partial conditioning for inference of many-normal-means with Hölder constraints. *Internat. J. Approx. Reason.*, 159:Paper No. 108946, 20.
- Zabell, S. L. (1992). R. A. Fisher and the fiducial argument. *Statist. Sci.*, 7(3):369–387.
- Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning. I. *Information Sci.*, 8:199–249.
- Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1(1):3–28.
- Ziliak, S. T. and McCloskey, D. N. (2008). *The Cult of Statistical Significance*. Economics, Cognition, and Society. University of Michigan Press, Ann Arbor, MI.