A possibility-theoretic solution to Basu's Bayesian–frequentist via media

Ryan Martin[*]

July 23, 2023

Abstract

Basu's *via media* is what he referred to as the middle road between the Bayesian and frequentist poles. He seemed skeptical that a suitable *via media* could be found, but I disagree. My basic claim is that the likelihood alone can't reliably support probabilistic inference, and I justify this by considering a technical trap that Basu stepped in concerning interpretation of the likelihood. While reliable probabilistic inference is out of reach, it turns out that reliable *possibilistic* inference is not. I lay out my proposed possibility-theoretic solution to Basu's *via media* and I investigate how the flexibility afforded by my imprecise-probabilistic solution can be leveraged to achieve the likelihood principle (or something close to it).

Keywords and phrases: conditional inference; fiducial argument; imprecise probability; inferential model; likelihood principle; validity.

1 Introduction

Debabrata Basu (1924–2001) was a giant in the field who made fundamental contributions that have inspired generations of statisticians and helped shape the very core of our subject. It's sincerely an honor and a privilege to contribute this manuscript for possible inclusion in the special volume of *Sankhya* in honor of Basu's birth centenary.

As I was thinking about what to contribute for this volume, I went back and reread some of Basu's classic works. Of course, I read many of these papers when I was younger, even as a PhD student—my introductory stat theory class was taught by Anirban DasGupta who prompted us to *"read Basu"* (Basu 2011, p. xvi)—but I lacked the maturity to fully grasp their depth and quality at the time. Now that I'm more experienced, I can better appreciate Basu's clarity and precision, along with his masterfully constructed examples. Beyond that, I have the context to recognize the courage Basu had to critically challenge the leaders of both Fisher's and Neyman–Pearson–Wald's schools of thought. Drawing inspiration from Basu's courage, here I make some similarly bold claims that I hope will stimulate discussions and help solidify our subject's foundations.

The title of this article references Basu's *via media*, a Latin phrase for "the middle road." This comes from a remark he made in concluding his reply to the discussion of his monumental 1975 essay:

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

> *The Bayesian and Neyman–Pearson–Wald theories of data analysis are the two poles in current statistical thought. Today I find assembled before me a number of eminent statisticians who are looking for a via media between the two poles. I can only wish you success in an endeavor in which the redoubtable R. A. Fisher failed.* (Basu 1975, p. 269)

The endeavor of Fisher's that Basu is referring to is, of course, the *fiducial argument*;[1] see, e.g., Fisher (1933, 1935, 1973), Neyman (1941), Lindley (1958), Fraser (1961, 1965), Seidenfeld (1992), Barnard (1995), Dawid (2020), and many others, including the generalized fiducial developments summarized in, e.g., Hannig (2009) and Hannig et al. (2016). Zabell and others have described fiducial inference as a sort of middle way:

> *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)

In addition to the obvious similarities in how the two authors characterize Fisher's efforts to strike this balance, they both leave open the *possibility* (pun intended) of a resolution, though Basu's remark falls short of Zabell's on the optimism scale. To me, it's imperative to the long-term success of the field of statistics that the *via media* be found and, fortunately, a solution is currently available. The high-level goal of this paper is to motivate and explain this solution, while making connections to Basu's work.

Part of motivating this solution—and even why a solution is needed—is understanding how the priorities of today's statisticians differ from those in Basu's time, when our subject was taking shape. Right or wrong, many view statisticians' role in science as connecting the scientific problem P to a suitable statistical method M to apply. In fact, the courses offered to non-statistics students tend to focus on enumerating the standard (P, M) pairs, e.g., comparing treatments via randomized experiments and the analysis of variance. In the old days, when a scientist encountered a new or unfamiliar problem P', he'd probably consult with a statistician. Nowadays, the scientist can immediately find a method M' to apply in his problem P' by consulting Google, no statistician required. So, while there are exceptions, modern-day statisticians' involvement in the scientific process is more indirect, by having an article that appears near the top of Google's search results. Consequently, those entertaining and expertly-crafted hypothetical dialogues between scientist and statistician in, e.g., Basu (1975, 1980) and Berger and Wolpert (1984), designed to shine light on our foundational questions/answers, no longer ring true.

As I explain in Section 2 below, this shift in the role that statisticians play in the scientific process, from direct to indirect, marks a change in statisticians' priorities. With respect to the two poles that Basu mentioned, now almost everyone is gathered around the frequentist pole—even the Bayesians! The methods-developing statistician simply can't ignore frequentist considerations, the very same frequentist considerations that the aforementioned dialogues crafted by Basu and others aimed to show were irrelevant or

[1] One could argue that Basu is referring to Fisher's ideas on conditional inference. To me, however, Fisher's conditioning (like sufficiency) was largely motivated by a need to reduce a problem's dimension, without loss of information, so that his fiducial argument could be applied.

downright silly. This is an unacceptably wide gap between our practical priorities and what our current foundations say. On the one hand, to jump to Basu's preferred Bayesian pole is tantamount to ignoring the modern priorities. On the other hand, to stay gathered around the frequentist pole and ignore the foundational issues raised by Basu and others is tantamount to concluding that those insights were wrong and/or no longer relevant. In both cases we end up losing our seat at the data science table: in the former, we're ignoring modern priorities and, in the later, we're admitting that our history and experience gives us no upper-hand over our new competitors. Neither of these are desirable outcomes, so a *via media* is imperative for our field's long-term success.

Section 3 presents how I expect the *via media* to look. Like the Bayesian pole, it offers fixed-data "probabilities" which can be used for "probabilistic reasoning" and inference; like the frequentist pole, these "probabilities" satisfy a certain *validity* property which implies that the procedures derived from them have error rate guarantees. This may sound too good to be true, and it is. The catch is that what I referred to above as "probabilities" aren't probabilities in the familiar sense; they're *imprecise probabilities* or, more specifically, they're *possibilities.* The shift from probability theory/calculus to the corresponding possibility theory/calculus is technically simple, but a fundamental change like this can be a conceptually large pill to swallow. My claim is that the likelihood alone can't reliably support probabilistic inference, so sticking with probability isn't an option. To justify this claim, in Section 4, I highlight a technical trap that Basu stepped in, related to the well-known fact that a likelihood function isn't a probability density, i.e., it has no inherent differential element. Other authors (e.g., Shafer 1982; Wasserman 1990) have suggested that the likelihood is more appropriately processed as a differential element-free possibility contour, but their proposals don't go far enough.

Having explained the high-level vision behind my proposed *via media* and justified the transition from probabilistic to possibilistic thinking, I provide a more detailed description of its implementation in Section 5. As I explain, the proposal shares some similarities with what's commonly done in statistical practice, but it's part of a framework that itself is very different. This, to me, is exactly what we'd expect from a suitable *via media*—it must be different from the two poles, but not unrecognizably different. I see this proposal as a modern, likelihood-centric version of the *inferential model* (IM) framework put forward in Martin and Liu (2013, 2015); for many more details, see Martin (2021, 2022a,b). A few Basu-spirited illustrations of my proposal are also presented here.

In Section 6, I consider one of Basu's favorites—the likelihood principle (Birnbaum 1962)—and how the new perspectives afforded by the imprecise-probabilistic formulation of my proposed *via media* can be beneficial. In particular, note that my basic proposal in Section 5 doesn't satisfy the likelihood principle, but it's clear how it can be made to do so without sacrificing on the solution's validity, provided that the data analyst is willing to give up some of their solution's efficiency. It's also possible to be valid and *partially* satisfy the likelihood principle, e.g., to be valid with respect to some user-identified set of plausible stopping rules but not to others, thereby balancing both the efficiency and the stopping-rule invariance that are relevant around the frequentist and Bayesian poles, respectively. The paper concludes with a brief discussion in Section 7.

2 Priorities have changed

Despite the very powerful foundational arguments put forward by Savage (1972), Basu (1975), Berger and Wolpert (1984), and many others in support of a fully-conditional, likelihood-centric approach to statistical inference, it's fair to say that there's effectively no sign of this way of thinking in modern statistics research—even among Bayesians. My claim is that statisticians' priorities have changed.

There are exceptions, of course, but today's academic statisticians, for the reasons I explained above, are almost exclusively focused on the development of *statistical methods*, i.e., specific tools and software intended to be used off-the-shelf by scientists working on the scientific front lines. The scientist is motivated by results, so their top priority is that a statistical method "works." That is, they're not going to apply a method off-the-shelf unless it's been demonstrated to "work" in some meaningful sense. This begs the question: in what sense could a method "work" that would be meaningful to scientists? It seems necessary that the method has been demonstrated to give a "right answer" in most of the cases in which it's applied. Then a scientist who believes that his problem is similar to those in which it's been demonstrated that the method typically gives a "right answer" has no reason to doubt that his application is one of those typical cases and, consequently, no reason to doubt the result of that method applied to his problem.

The reader surely recognizes that my description of what it means for a method to "work" is very much frequentist. The reader surely is also aware that these frequentist considerations, and my definition of "works," don't align with the fully-conditional likelihood/Bayesian considerations of Basu and others. "Don't align" is an understatement, these two considerations are almost completely incompatible—if a method "works" in the sense above, then it almost always falls short of Basu's foundational bar. So where does this leave us? For sure, a subject that's central to the advancement of knowledge shouldn't abandon its foundations altogether for the *priorité du jour*. But it's similarly embarrassing for the same subject to hold up a foundational standard that's not generally taken seriously by today's methods-developing statisticians. Even modern Bayesian methods fail to meet Basu's high standard. It's of course well-known that the (common) use of default priors is incompatible with the likelihood principle. Moreover, there's a relevant selection bias in the Bayesian literature: the Bayesian methods that appear in publications tend to be those that have been demonstrated to "work" in the sense above, either theoretically or empirically.[2] These demonstrations fix the data-generating process, so adopting a Bayesian solution in an application because the method "works" under certain data-generating processes is a violation of the likelihood principle.

To be fair, the Scylla at the frequentist pole isn't any more pleasant than the Charybdis at Bayesian pole. Aside from not really addressing the question of *inference*, the pure performance focus hasn't proved to ensure reliability, as the replication crisis has revealed. The take-away, again, is that the problem of statistical inference can't be fully resolved at either of the extreme poles; the nuance if a genuine *via media* is necessary.

[2] An anonymous reviewer asked for some evidence to support this claim, so I scanned the 2022 volume of *Bayesian Analysis* and noted which papers offered a demonstration that the proposed method "works" based on a proof of consistency or a simulation study. Of the 45 published papers, I counted 37 papers (about 82%) with a significant focus on the proposed method's frequentist performance.

3 Towards a *via media*

As Zabell wrote in the quote above, the fact that Fisher's attempts to find this *via media* failed doesn't mean that it can't be found. But it will require some outside-the-box ideas, and I'll share these ideas in the subsequent sections. First, I should explain how I think this *via media* ought to look, since what I have in mind is quite different from what is currently done by both Bayesians and frequentists. My main thesis is as follows:

- a fully satisfactory theory of statistical inference ought to produce reliable, data-dependent "probabilities" based on which probabilistic reasoning can be made, i.e., if the data-dependent "probability" assigned to a hypothesis is small, then that should provide good reason to doubt the truthfulness of the hypothesis;
- but the data alone can't reliably support the construction of data-dependent "probabilities" that are genuine probabilities like in Kolmogorov's theory;
- so, to achieve both the probabilistic reasoning that's advantageous for single-case, data-driven inference and the reliability that's necessary from today's perspective with a focus on methods-development, this *via media* can't be contained in the existing/standard theory of probability, i.e., the "probabilities" I'm referring to above can't literally be probabilities in the sense of Kolmogorov.

My justification for the claim in the second bullet point will come in the next section, where things get more technical. In the remainder of this section, I want to focus on the first and third bullet points, which are more conceptual in nature.

For the first bullet: why is probabilistic reasoning so important? A common criticism of the frequentist theory of inference, which isn't based on probabilistic reasoning, is that significance levels, coverage probabilities, etc. are pre-data calculations—they don't speak to what the observed data actually say about the unknown being inferred. P-values aim to bring the observed data into the uncertainty quantification picture, but these too are often (unjustifiably) criticized because they're not probabilities, not measures of the strength of evidence, etc. More recently, some authors, especially Deborah Mayo, have been calling for more than what the classical frequentist solutions offer. She argues in Mayo (2018) that, in addition to determining if a hypothesis is incompatible with the data, via tests and confidence sets, it's important that scientists can "probe" deeper into those hypotheses that are compatible with the data to find sub-hypotheses that might actually be supported by data. This probativeness feature comes fairly naturally when inference is based on data-dependent "probabilities," but not otherwise; Mayo suggests supplementing the frequentist methods with a so-called *severity measure* designed to offer probativeness. Suffice it to say that there are real, practical advantages to probabilistic reasoning that the classical frequentist solutions fail to offer, but these advantages don't come for free just by choosing to write down (artificial) probabilities.

It's in the third bullet where the *via media* starts to reveal itself. Recall that Basu's poles correspond to probability (Bayesian) and not-probability (frequentist). From this perspective, it seems almost obvious that the middle-ground must somehow be both probability and not-probability simultaneously. "Fisher's biggest blunder" (Efron 1998) was just his failure to see that the *via media* can't be achieved entirely within the theory of probability. What I/we have now that Fisher didn't have is (the benefit of hindsight and) more than 60 years of developments—starting with Art Dempster's seminal work in the 1960s (e.g., Dempster 1966, 1967)—in the theory of *imprecise probability*. What I'm proposing in Section 5 falls under the umbrella of imprecise probability, but it's both drastically different from Dempster's approach and surprisingly similar to ideas that can be found in standard statistics textbooks. But before I can get into these details, I need to justify the claim in my second bullet point above, which I'll do next.

4 Likelihood and inference

4.1 What likelihood can't do

To set some notation, let $X \in \mathbb{X}$ denote the observable data and write P_Θ for the posited statistical model, depending on an unknown parameter $\Theta \in \mathbb{T}$. As is customary, here I'll write Θ for the unknown true parameter value, saving θ and ϑ to denote generic parameter values. For fixed $\theta \in \mathbb{T}$, suppose that P_θ admits a probability mass or density function p_θ on $\mathbb{X}$, and define the likelihood function at the observed $X = x$, as $L_x(\theta) = p_\theta(x)$. The name "likelihood" was coined by Fisher and part of the motivation behind this choice of name was to emphasize that, notwithstanding the obvious connection between the likelihood function and the model's probability density/mass function, the likelihood is indeed fundamentally different from probability. In particular:

> *The function of the θ's... is not however a probability and does not obey the laws of probability; it involves no differential element* $d\theta_1\, d\theta_2\, d\theta_3$*...; it does none the less afford a rational basis for preferring some values of θ, or combination of values of the θ's, to others.* (Fisher 1930, p. 552)

Despite the warnings, many have not taken this seriously—including Fisher himself! Indeed, as Basu (1975, p. 33) points out, there are cases in which Fisher's fiducial argument, his proposed *via media*, produces a solution that's equivalent to treating the likelihood as if it were a probability density/mass function for Θ, given $X = x$. If the likelihood doesn't determine a probability distribution for Θ, and if the fiducial argument can produce a solution that's a probability determined by the likelihood, then isn't that an obvious sign something's wrong with the fiducial argument itself?

But Basu stepped into this trap too. In Basu (1975, Sec. 8), he proposes the construction of a data-dependent probability distribution for Θ on $\mathbb{T}$ based on a normalized likelihood function,[3] which, in the present notation, is

$$\bar{L}_x(A) = \frac{\int_A L_x(\theta)\, d\theta}{\int_{\mathbb{T}} L_x(\theta)\, d\theta}, \quad A \subseteq \Theta. \tag{1}$$

Constructing a probability by suitably normalizing the likelihood function as above seems natural and, following a detailed analysis, Basu (1975, p. 33) concludes with:

> *The author can find no logical justification for the often repeated assertion that likelihood is only a point function and not a measure. He does not see what inconsistencies can arise from [treating it as a measure].*

[3] Basu actually assumes $\mathbb{T}$ is finite and defined the above expression with the integrals replaced by sums; see (2). I'm writing integrals here only because it's more common for the parameter space to be a continuum; none of what I have to say here depends on this choice.

Problems arise because, as Fisher emphasized, the likelihood has no differential element "$d\theta$." While introducing "$d\theta$" and normalization via integration might seem innocuous, this isn't free of consequences. Of course, $A \mapsto \bar{L}_x(A)$ is a measure, so it's additive and monotone: in particular, $A \subseteq B$ implies $\bar{L}_x(A) \leq \bar{L}_x(B)$. That A can't be more compatible with data than B is perfectly logical, but it'll virtually always be that $\bar{L}_x(B)$ is *strictly greater* than $\bar{L}_x(A)$. For example, suppose that $X \sim \mathsf{N}(\Theta, 1)$, and that $x = 7$ is observed. Consider two hypotheses about the unknown Θ: $A = [7.7, 8]$ and $B = [7.7, 20]$. Clearly, A is a proper subinterval of B, and B is much wider than A, which implies that $\bar{L}_x(B) \gg \bar{L}_x(A)$; in this particular case, $\bar{L}_x(B) \approx 3\bar{L}_x(A)$. But is there any sense in which B is *strictly more* compatible with the data than A? No—it's obvious that the inclusion of points that are relatively incompatible with the data doesn't make the hypothesis more compatible with the data. That's the point I think Fisher was trying to make when he emphasized the likelihood involves no differential element.

Similar points were made by economist G. L. S. Shackle in the mid-1900s. Like in Basu (1975, p. 29), Shackle had in mind a finite space $\mathbb{T}$ and was entertaining the option of assigning plausibility[4] to individual elements as

$$\bar{L}_x(\theta) = \frac{L_x(\theta)}{\sum_{\vartheta \in \mathbb{T}} L_x(\vartheta)}, \quad \theta \in \mathbb{T}. \tag{2}$$

Note that the above relationship forces the mass assigned to an individual θ to depend on the cardinality of $\mathbb{T}$. Shackle argues emphatically that the size of (the hypothesis space) $\mathbb{T}$ ought not to influence the plausibility of an individual (hypothesis) θ.

> *To allow the size of the crowd of hypotheses... to influence the value of the [plausibility] assigned to any particular hypothesis, would be like weakening one's praise for the chief actors in a play on the ground that a large number of supers were also allowed to cross the stage.* (Shackle 1961, p. 51)

This begs a fundamental question: does introducing an artificial differential element detail have any practical consequences? Yes! It implies existence of true hypotheses A for which $\bar{L}_X(A)$ tends to be small as a function of X and, similarly, the existence of false hypotheses B for which $\bar{L}_X(B)$ tends to be large as a function of X. Since one would be inclined, e.g., to doubt the truthfulness of a hypothesis for which $\bar{L}_X(A)$ is small, this counter-intuitive behavior raises serious practical concerns about the reliability of inferences based on the normalized likelihood—which is very much relevant to the methods-developing statistician and to the scientist who uses these methods. The root cause of this undesirable behavior is obvious: there are small sets that contain Θ and large sets that don't. That is, the size itself of a hypothesis has no bearing on whether it's true or false and, therefore, no bearing on how compatible it is with the data. This intuition is captured by the likelihood and its lack of a differential element. But the integral-driven normalization forces additivity, which allows the (irrelevant) size of the hypothesis to become relevant, hence the undesirable behavior. This is captured in the main result of Balch et al. (2019), though the connection here to Fisher's warning about the likelihood's lack of a differential element is apparently new.

[4] Shackle didn't specifically mention likelihood in his analysis, but my choice to make this point using likelihood is consistent with Shackle's remarks.

False confidence theorem (Balch et al. 2019). *Let $\bar{L}_x(\cdot)$ be Basu's x-dependent probability distribution* (1) *based on normalizing the likelihood via integration. Then for any pair $(\lambda, \alpha) \in [0,1]^2$ and for any $\Theta \in \mathbb{T}$, there exists a hypothesis $A \subset \mathbb{T}$ such that*

$$A \not\ni \Theta \quad \textit{and} \quad \mathsf{P}_\Theta\{\bar{L}_X(A) > \lambda\} > \alpha.$$

The same conclusion holds true for any data-dependent probability measure, not just that in (1)*, including any Bayes posterior or fiducial distribution.*

In words, the false confidence theorem states that there are false hypotheses to which the artificially additive posterior $\bar{L}_X(\cdot)$ tends to assign high probability; this creates a risk of systematically misleading conclusions and raises doubts about the reliability of inference. This is my justification for the claim made in the second bullet point in Section 3 above: the likelihood (data + model) alone can't reliably support genuine data-dependent *probabilities* and the associated probabilistic inference.

To be clear, the above issues are unrelated to the choice of dominating measure: one can't sidestep the difficulties raised by the false confidence theorem by introducing a default prior density/mass function before normalization. The point, again, is that a hypothesis's size has no direct bearing on its compatibility with the data, and yet it's relevant to any Lebesgue integral. A hypothesis's size is relevant only if the data analyst *believes* that it is, i.e., if he's willing to introduce a genuine prior distribution that connects size to credence. That is, if the prior probabilities are real and part of the posited model, then the differential element is meaningful and there's no false confidence. If the prior probabilities are artificial, then there are no guarantees: *[Bayes's formula] does not create real probabilities from hypothetical probabilities* (Fraser 2014, p. 249).

4.2 What likelihood can do

If the likelihood doesn't have a differential element and, therefore, doesn't reliably support probabilistic inference, but there's still something that can be done! The point is: probability theory is not the only uncertainty quantification game in town. Starting with Dempster's seminal work in the 1960s, there have been major developments in what's called *imprecise probabilities*; see, e.g., the books by Shafer (1976), Dubois and Prade (1988), Walley (1991), Troffaes and de Cooman (2014), Augustin et al. (2014), and Cuzzolin (2021). The simplest among the imprecise probability models is *possibility theory*, with close connections to fuzzy set theory (e.g., Hanss 2005; Zadeh 1965), which dates back to Shackle (1961) and Zadeh (1978); much of modern possibility theory is based on Dubois and Prade (1988) and the extensive subsequent work by the same authors. It is a general-purpose theory of uncertainty quantification applied throughout science and engineering; statistical applications are discussed in Dubois (2006) and the references therein, but the perspective I share here and in the next section is new.

A simple idea, similar to Basu's, starts by defining the *relative likelihood* function

$$\eta_x(\theta) = \frac{L_x(\theta)}{\sup_\vartheta L_x(\vartheta)}, \quad \theta \in \mathbb{T}. \tag{3}$$

Note the difference in normalization—supremum versus Basu's integration—so η_x isn't a probability. But (3) is the driver behind the proposal in Shafer (1976, 1982), which is developed further in Wasserman (1990), Denœux (2014), and elsewhere. This determines a very special imprecise probability structure which has a few different names: here I adopt the possibility theory terminology, so I'll refer to $\theta \mapsto \eta_x(\theta)$ in (3) as a *possibility contour*. Mathematically, what distinguishes η_x as a possibility contour is that, first, it takes values in $[0,1]$ and, second, that it satisfies $\sup_\theta \eta_x(\theta) = 1$. The contour's extension to a *possibility measure* supported on general hypotheses is defined as

$$\eta_x(A) = \sup_{\theta \in A} \eta_x(\theta), \quad A \subseteq \mathbb{T}.$$

This is analogous to Basu's $\bar{L}_x$, just with possibility calculus[5] instead of probability calculus. This way of processing the likelihood function has a number of desirable properties, e.g., it's completely driven by the likelihood-based ranking of the parameter values so it doesn't require introduction of an artificial differential element.

Of course, the set-function $A \mapsto \eta_x(A)$ isn't a measure in the usual sense, but it does have some similar properties. In addition to $\eta_x(\cdot) \geq 0$ and $\eta_x(\mathbb{T}) = 1$, the possibility measure is *maxitive*[6] which implies sub-additivity, in particular,

$$1 \leq \eta_x(A) + \eta_x(A^c), \quad \text{for all } A \subseteq \mathbb{T}. \tag{4}$$

Maxitivity also implies monotonicity, but not the kind of strict monotonicity that often holds for probabilities. Reconsider the simple $X \sim \mathsf{N}(\Theta, 1)$ illustration above, with $A = [7.7, 8]$ and $B = [7.7, 20] = A \cup (8, 20]$. While Basu's $\bar{L}_x$ has $\bar{L}_x(A) \ll \bar{L}_x(B)$, the possibility measure has $\eta_x(A) = \eta_x(B)$, as one would expect: the inclusion of an interval $(8, 20]$ that contains "less-likely" values shouldn't increase the compatibility with x.

Mathematics aside, since η_x isn't a probability, we don't have access to the full power of probabilistic reasoning. But a one-sided version is available, what I'll call here *possibilistic reasoning*. That is, possibility theory allows for a direct refutation of a hypothesis "$\Theta \in A$" by showing that $\eta_x(A)$ is small. However, unlike with probabilistic reasoning, if $\eta_x(A)$ is large, then that's not enough to conclude that there's support for "$\Theta \in A$," since (4) doesn't rule out the case that both $\eta_x(A)$ and $\eta_x(A^c)$ are large. In possibilistic reasoning, we need *both* $\eta_x(A)$ large and $\eta_x(A^c)$ small to find support for "$\Theta \in A$."

What does it mean for $\eta_x(A)$ to be "small" or "large"? The methods-developing statistician must suggest to potential users how to make these judgments and, if his method is going to demonstrably "work," then he similarly must take this small/large judgment seriously. One can tailor these small/large possibility thresholds to the problem at hand, or perhaps rely on asymptotic theory to get some unification, but that's different from probabilistic reasoning. Indeed, recall that probability has the same scale across every example to which it's applied: a numerical probability of 0.1 means the same thing whether it's the probability of rain tomorrow or the probability of a patient responding favorably to a new cancer treatment. The basic likelihood-to-possibility setup presented above doesn't share this invariance, i.e., the small/large scale that "works" depends crucially on features of the application at hand. But a different possibility-theoretic framework can do it, as I explain next.

[5] Note that possibility calculus can be described via a suitable Choquet integral instead of the familiar Lebesgue integral; see Choquet (1954) and Troffaes and de Cooman (2014, App. C).

[6] Maxitive means $\eta_x(\bigcup_{k=1}^\infty A_k) = \sup_{k \geq 1} \eta_x(A_k)$ for all $A_1, A_2, \ldots \subseteq \mathbb{T}$.

5 A possibility-theoretic *via media*

The details below are simultaneously both new and familiar, i.e., there are close connections with classical theory but the possibility-theoretic details that make it a full-blown framework are recent developments and are likely unfamiliar to most readers. For the sake of space, I'll only present the immediately-relevant aspects of this theory. In particular, I'll not present the (arguably most interesting) details that showcase how the framework easily incorporates *partial prior information* about Θ. The partial-prior angle is crucial for at least two reasons: first, it's what creates new opportunities for improved methods and, second, it's what justifies this proposal as a bona fide *via media* between the Bayesian and non-Bayesian poles. The interested reader can consult Martin (2022a,b).

It turns out that the relative likelihood function is still very relevant here (see Martin 2022b, Sec. 5.1). But since its role is a bit different, I'm going to use a slightly modified notation: I'll write $\eta(x, \theta)$ instead of $\eta_x(\theta)$. The key idea is that the likelihood offers a data-dependent partial order on $\mathbb{T}$, but even the relative likelihood is lacking a universal scale that "works" for all applications. Following the key developments in Hose (2022), my proposal is to calibrate the relative likelihood in a principled way to construct a new possibility contour—and corresponding possibility measure—that has the same partial order on $\mathbb{T}$ as the likelihood but is universally scaled and provably "works." Define this new likelihood-based possibility contour for Θ as

$$\pi_x(\theta) = \mathsf{P}_\theta\{\eta(X, \theta) \leq \eta(x, \theta)\}, \quad \theta \in \mathbb{T}. \tag{5}$$

The reader may recognize this as a sort of p-value determined by the relative likelihood. This is for the special case where prior information about Θ is vacuous; if (partial) prior information is available, then a different possibility contour emerges. In Martin (2022b), it's shown that (5) corresponds to a familiar operation in the imprecise probability literature, namely, the *probability-to-possibility transform* (e.g., Dubois et al. 2004).

The possibility contour in (5) extends to a full-blown possibility measure for Θ:

$$\overline{\Pi}_x(A) = \sup_{\theta \in A} \pi_x(\theta), \quad A \subseteq \mathbb{T}. \tag{6}$$

There is also a dual *necessity measure* defined via conjugacy, i.e., $\underline{\Pi}_x(A) = 1 - \overline{\Pi}_x(A^c)$, and it's easy to show that $\underline{\Pi}_x(A) \leq \overline{\Pi}_x(A)$ for all $A \subseteq \mathbb{T}$ and all $x \in \mathbb{X}$. Possibilistic reasoning proceeds exactly as described above. The difference here compared to at the end of Section 4 is that now there's a universal possibility scale, so it's easy for the user to decide what it means for $\overline{\Pi}_x(A)$ to be "small"—or, equivalently, what it means for $\underline{\Pi}_x(A)$ to be "large"—and to understand the methodological implications of this decision.

Theorem. *The IM determined by the possibility contour in* (5) *is* (strongly) valid*, i.e.,*

$$\sup_{\Theta \in \mathbb{T}} \mathsf{P}_\Theta\{\pi_X(\Theta) \leq \alpha\} \leq \alpha, \quad \textit{for all } \alpha \in [0, 1], \tag{7}$$

and, consequently, the possibility measure (6) *satisfies*

$$\sup_{\Theta \in \mathbb{T}} \mathsf{P}_\Theta\{\overline{\Pi}_X(A) \leq \alpha \textit{ for some } A \ni \Theta\} \leq \alpha, \quad \textit{for all } \alpha \in [0, 1]. \tag{8}$$

Proof. More general results are covered in Martin (2022b). Claim (7) can be verified directly via the aforementioned connection to the familiar relative likelihood-based p-values. Claim (8)—that calibration holds uniformly over all true hypotheses—follows from (7) and the fact that $\sup_{\theta \in A} \pi_X(\theta) \leq \alpha$ for some $A \ni \Theta$ if and only if $\pi_X(\Theta) \leq \alpha$. □

One important consequence of the above theorem is that the possibilistic IM is not afflicted by false confidence the way probabilistic inference is (Martin 2019, 2021). Specifically, false confidence would arise if the IM tended to assign large $\underline{\Pi}_X$-values to false hypotheses. But it follows immediately from (8) and the definition of $\underline{\Pi}_x$ that

$$\sup_{\Theta \in \mathbb{T}} \mathsf{P}_\Theta\{\underline{\Pi}_X(A) \geq 1 - \alpha \text{ for some } A \not\ni \Theta\} \leq \alpha, \quad \alpha \in [0,1].$$

That is, the event where *any* false hypothesis is assigned a relatively large $\underline{\Pi}_X$-value has relatively small probability—hence no false confidence.

The following corollary establishes that the same IM output that can be used for reliable in-sample possibilistic reasoning can also be used to construct statistical methods or procedures that "work" in the out-of-sample sense described above.

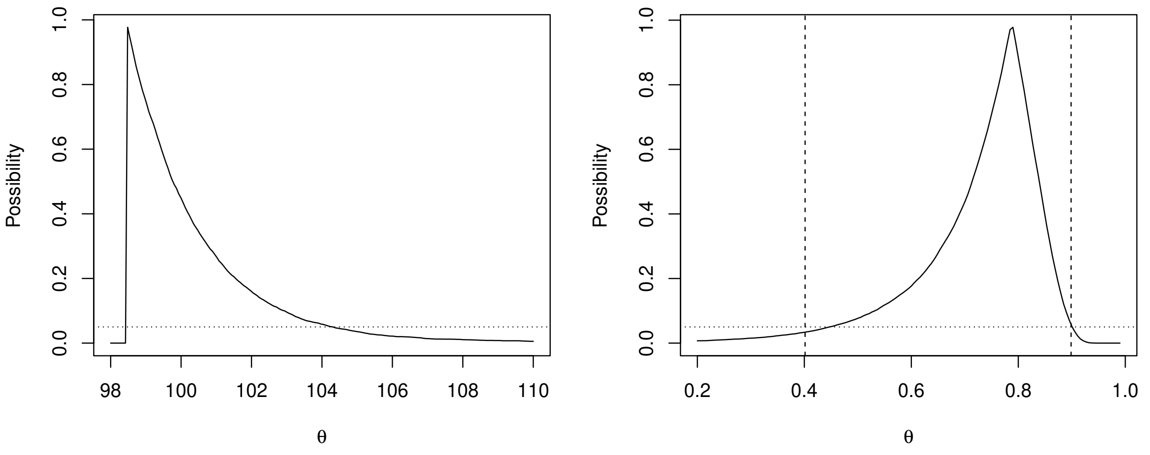
Corollary. *Hypothesis testing and confidence set procedures derived from the IM defined above control frequentist error rates at the nominal levels. That is:*

- *For a hypothesis $H : \Theta \in A$, the test that rejects H if and only if $\overline{\Pi}_X(A) \leq \alpha$ has Type I error probability bounded above by α, and*
- *The set $C_\alpha(X) = \{\theta : \pi_X(\theta) > \alpha\}$ has coverage probability bounded below by $1 - \alpha$.*

These are the usual frequentist properties expected of hypothesis tests and set estimators and they follow immediately from (7) without any conditions on the models involved, the sample size, etc. The key role played by the likelihood suggests that these IM-driven methods would be efficient, and it is often the case in applications that they agree with the optimal or best-available methods. Moreover, the particular result in (8) implies much stronger error rate control than the first part of the above corollary lets on. Indeed, the error rate control is actually *uniform* in the hypotheses A, which has certain probativeness consequences à la Mayo (2018); see Cella and Martin (2023).

Next I'll show four quick illustrations of the IM formulation, using some examples that were interesting to Basu. The first pair are problematic "non-regular" examples in which the minimal sufficient statistic has dimension greater than that of the unknown parameter, as considered in (Basu 1964, 1967) and elsewhere.

Example 1. Let $X = (X_1, \ldots, X_n)$ consist of iid $\mathsf{Unif}\{a(\Theta), a(\Theta) + b(\Theta)\}$ random variables, where Θ is an unknown scalar but $a(\cdot)$ and $b(\cdot)$ are known functions. One of Basu's favorites is $a(\theta) = \theta$ and $b(\theta) \equiv 1$, so that θ is a location parameter. That's a special group-invariant case, as studied in, e.g., Basu and Ghosh (1969), and the connection between the proposed IM solution and the fiducial/default-prior Bayes solution is presented in Martin (2023). Here I consider the model $\mathsf{P}_\Theta = \mathsf{Unif}(\Theta, \Theta^2)$, with unknown $\Theta \in \mathbb{T} = (1, \infty)$, with endpoint functions $a(\theta) = \theta$ and $b(\theta) = \theta^2 - \theta$. This problem is "non-regular" in the sense that the minimal sufficient statistic—$(X_{(1)}, X_{(n)})$, the extreme



(a) Example 1, $(x_{(1)}, x_{(n)}) = (281.1, 9689.7)$

(b) Example 2, $\hat\theta_x = 0.788$

Figure 1: Plots of the possibility contour functions for Examples 1–2. In Panel (b), the dashed vertical lines mark the endpoints of an asymptotically efficient 95% confidence interval based on the "$r^\star$" approximation described in Reid (2003) and elsewhere.

order statistics—is two-dimensional while Θ is a scalar. Despite this non-regularity, the maximum likelihood estimator is easy to get, $\hat\theta_x = x_{(n)}^{1/2}$, so the relative likelihood is

$$\eta(x,\theta) = \frac{L_x(\theta)}{L_x(\hat\theta_x)} = \Bigl\{ \frac{x_{(n)} - x_{(n)}^{1/2}}{\theta^2 - \theta} \Bigr\}^n \cdot 1\{x_{(1)} \geq \theta, \, x_{(n)} - x_{(1)} \leq \theta^2 - \theta\}, \quad \theta > 1.$$

The corresponding possibility contour based on (5) is

$$\pi_x(\theta) = \mathsf{P}_\theta\{X_{(n)} - X_{(n)}^{1/2} \leq x_{(n)} - x_{(n)}^{1/2}\}, \quad \theta > x_{(1)}.$$

There is no closed-form expression for this, but it's easy to approximate via Monte Carlo: given $\Theta = \theta$, $X_{(n)}$ is distributed as $\theta + (\theta^2 - \theta)\mathsf{Beta}(n, 1)$, a transformed beta random variable. For an illustration, I follow Hannig et al. (2016, Ex. 4) and consider data of size $n = 25$ with $(x_{(1)}, x_{(n)}) = (281.1, 9689.7)$, so that the maximum likelihood estimator is $\hat\theta_x \approx 98.4$. Figure 1(a) shows the possibility contour and, from this, we can readily evaluate $\overline{\Pi}_x(A)$ for any A and/or extract a $100(1-\alpha)$% confidence interval by thresholding the contour at level α. This IM agrees with that developed in Martin and Lin (2016) and shown there to be both valid and as efficient as existing solutions.

Example 2. Suppose that, $X = (X_1, \ldots, X_n)$ consists of n iid random variable pairs $X_i = (X_{i1}, X_{i2})$ which are bivariate normal with mean 0, variance 1, and unknown correlation $\Theta \in \mathbb{T} = [-1, 1]$. It's easy to check that the minimal sufficient statistic is two-dimensional, compared to one-dimensional Θ, so there's non-regularity like in Example 1. Consequently, inference based on the sampling distribution of, say, the maximum likelihood estimator would be inefficient due to a loss of information. As Basu (1964) showed, there's no guidance on what ancillary statistic one should condition on to recover the lost information—both partial data sets $(X_{i1}, \ldots, X_{n1})$ and $(X_{i2}, \ldots, X_{n2})$ are equally good ancillary statistics—so it's not clear how to proceed. For this reason, care is needed in developing valid and efficient methods for inference on Θ.

There's no closed-form expression for the maximum likelihood estimator and, consequently, there's no closed-form expression for the relative likelihood or the proposed IM's possibility contour $\theta \mapsto \pi_x(\theta)$ in (5). Fortunately, it's not too difficult to perform the required computations numerically (e.g., root-finding and Monte Carlo), and I have done so for a simulated data set of size $n = 10$ with $\Theta = 0.75$. Figure 1(b) shows the possibility contour plot; for reference, the maximum likelihood estimator in this case is $\hat{\theta}_x = 0.788$. One can easily read off a 95% confidence interval by thresholding the contour at level $\alpha = 0.05$. For comparison, the vertical lines mark the endpoints of the 95% confidence interval determined by the efficient "$r^\star$" asymptotic approximation (e.g., Brazzale and Davison 2008; Reid 2003). The difference between these and the IM's confidence limits is negligible, so the IM solution is exactly, provably valid and efficient.

The second pair of illustrations are closely tied to Basu's interests in and deep insights concerning finite-population sampling (e.g., Basu 1969, 1971).

Example 3. Suppose $X = (X_1, \ldots, X_n)$ is an iid sample from $\mathsf{Unif}\{1, 2, \ldots, \Theta\}$, where Θ is an unknown natural number. In this case, the likelihood function is

$$L_x(\theta) = \theta^{-n} \cdot 1(\theta \geq x_{(n)}), \quad \theta = 1, 2, \ldots$$

and the maximum likelihood estimator is $\hat{\theta}_x = x_{(n)}$, so it's easy to show that the IM's possibility contour based on (5) is

$$\pi_x(\theta) = (x_{(n)}/\theta)^n, \quad \theta = x_{(n)}, x_{(n)} + 1, \ldots$$

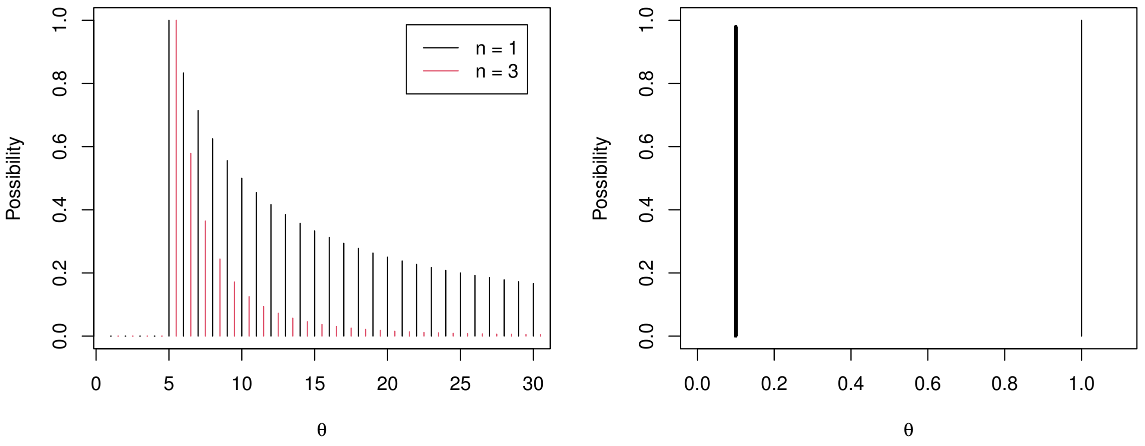
A plot of this contour function is shown in Figure 2(a) based on $x_{(n)} = 5$ with two values of n; the vertical spikes emphasize that it's a function defined only on the integer values. Clearly, these are not probability masses since they don't sum to 1. The maximum possibility value of 1 is attained at $\theta = x_{(n)}$, decreasing thereafter, and the extended possibility measure (6) on general hypotheses can be readily evaluated as needed. Note that the possibility contour vanishes much more rapidly for $n = 3$ compared to $n = 1$, which is sign of the efficiency gain with a larger sample size.

Example 4. Consider the example in Basu (1975, p. 240) involving an urn that contains 1000 balls: 20 are labeled with Θ and the remaining 980 are labeled with the values $a_1\Theta, \ldots, a_{980}\Theta$, where the a_j's are distinct known values in the interval $[9.9, 10.1]$. Let X denote the value on a single randomly chosen ball from this urn. The likelihood is

$$L_x(\theta) = \begin{cases} 0.02 & \text{if } \theta = x \\ 0.001 & \text{if } \theta \in \{a_1^{-1}x, \ldots, a_{980}^{-1}x\} \\ 0 & \text{otherwise.} \end{cases}$$

Basu designed this example to highlight some unusual behavior of the maximum likelihood estimator, in particular, that $\hat{\theta}_X = X$ is far from Θ with P_Θ-probability 0.98. From here it's not difficult to show that the IM's possibility contour is

$$\pi_x(\theta) = \begin{cases} 1 & \text{if } \theta = x \\ 0.98 & \text{if } \theta \in \{a_1^{-1}x, \ldots, a_{980}^{-1}x\} \\ 0 & \text{otherwise.} \end{cases}$$



(a) Example 3, $x_{(n)} = 5$

(b) Example 4, $x = 1$

Figure 2: Possibility contour plots for Examples 3–4. In Panel (a), the red spikes have been shifted to the right so they don't overlap with the black. In Panel (b), the thick vertical line is made up of 980 spikes of height 0.98 around the value $\theta = 0.1$.

A plot of this is shown in Figure 2(b) with $x = 1$. The aforementioned unusual behavior of the maximum likelihood estimator is not an issue here because there's no compelling reason to single out $\hat\theta_x$ when all the other values—which are very close to Θ when $\hat\theta_x$ isn't—are similarly highly plausible.

To conclude, the framework that I'm proposing here is a viable candidate for Basu's *via media*. It combines the (Bayesian-like) in-sample possibilistic reasoning with the (frequentist-like) calibration that guarantees the derived methods "work" in the out-of-sample sense that's relevant to users of statistical methods. This combination can't be achieved without venturing into imprecise probability territory.

6 Valid IMs and the likelihood principle

The IM framework I put forward in Section 5 doesn't satisfy the likelihood principle. That is, despite being largely relative likelihood-driven, the possibility contour isn't *fully* determined by the relative likelihood—it depends on the model $\{\mathsf{P}_\theta : \theta \in \mathbb{T}\}$, on the sample space $\mathbb{X}$, etc. via the probability calculation in (5)—so it fails to satisfy the likelihood principle. But this doesn't mean that it's impossible to achieve the likelihood principle (or something close enough to it), if desired, through some adjustments. Remember, we're after a *via media*, so certain trade-offs should be expected to meet today's methods-focused needs. These adjustments will also highlight the flexibility that an imprecise-probabilistic framework affords the statistician.

Recall that the posited model $\{\mathsf{P}_\theta : \theta \in \mathbb{T}\}$ and observed data $X = x$ determine a relative likelihood $\eta(x, \cdot)$, but not uniquely. That is, in general there's a class of models that all produce the same $\eta(x, \cdot)$ for (almost) all x. To give the reader some context, what I have in mind are different data-collection procedures, e.g., sampling designs, stopping rules, etc., for investigating the same scientific question. To ensure that these developments make sense mathematically, I'll reinterpret the data X as whatever's needed to determine that relative likelihood. Given $\eta : \mathbb{X} \times \mathbb{T} \to [0,1]$, let

$$\mathscr{P}^\star = \mathscr{P}^\star(\eta, \mathbb{T}) = \{\mathsf{P}_\theta^{(m)} : \theta \in \mathbb{T}, \, m \in \mathbb{M}^\star\},$$

where $m \in \mathbb{M}^\star$ is a generic model index, denote the collection of *all* probability distributions on $\mathbb{X}$, parametrized by $\theta \in \mathbb{T}$, with density/mass function $p_\theta^{(m)}(x)$ that satisfies

$$\frac{p_\theta^{(m)}(x)}{\sup_\vartheta p_\vartheta^{(m)}(x)} = \eta(x, \theta), \quad \text{for all } \theta \in \mathbb{T}, \, m \in \mathbb{M}^\star, \text{ and (almost) all } x \in \mathbb{X}.$$

That is, any one of these candidate models determines η, and then the data analyst collects in $\mathbb{M}^\star$ all the models with equivalent relative likelihood.

For a concrete example, consider a sequence of Bernoulli trials where the data is a pair consisting of the number of trials performed and the number of successes observed; write this as $x = (n, y)$, where n is the number of trials and y is the number of successes. There are, of course, a variety of models for data of this type, depending on how the experiment is performed. If n is fixed in advance, then y would be considered "data," and a binomial model would be appropriate. Alternatively, if y is fixed in advance, then n is the "data," and a negative binomial model would be appropriate. As is well-known, both of these have relative likelihood

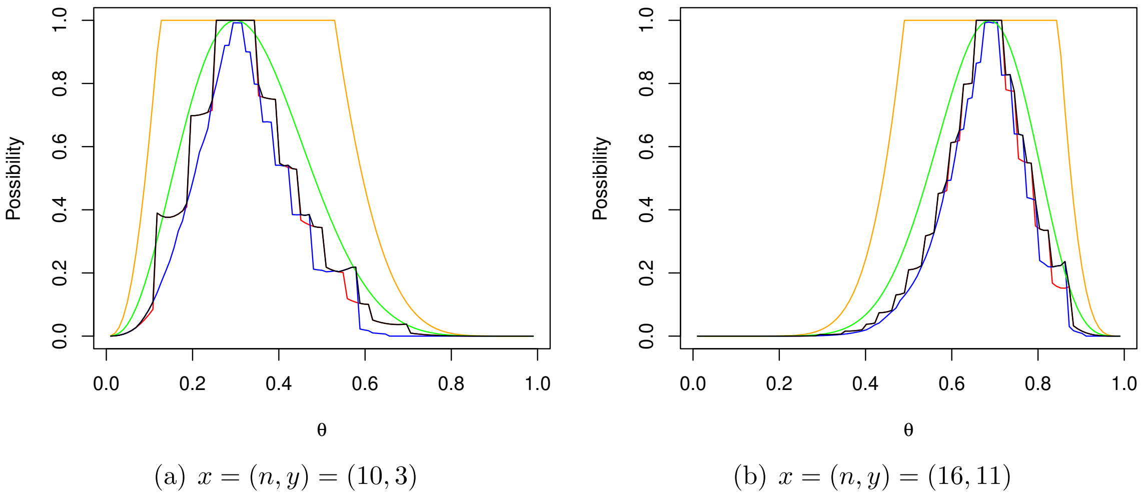
$$\eta(x, \theta) = \Bigl(\frac{n\theta}{y}\Bigr)^y \Bigl(\frac{n - n\theta}{n - y}\Bigr)^{n-y}, \quad \theta \in [0,1], \quad x = (n, y).$$

While the above two designs might be the most common in practice, these aren't the only two models in $\mathbb{M}^\star$ for η as above; there are many more, one for each proper stopping rule. Example 21 of Berger and Wolpert (1984) offers a setup wherein $x = (n, y)$ can take one of three possible values, namely, $(1,1)$, $(2,0)$, or $(2,1)$, i.e., stop the study after the first trial if it's a success, otherwise stop after the second trial.

The data analyst might be able to eliminate some of the equivalent models so, in general, consider a sub-collection $\mathbb{M} \subseteq \mathbb{M}^\star$. In the Bernoulli trial illustration above, the data analyst might not know what stopping rule was used, but if he knows that some *weren't* used, then those can be omitted from $\mathbb{M}$. The embellishment I'm suggesting here, natural from an imprecise-probabilistic point of view, is to define a new possibility contour by maximizing the right-hand side of (5) over models:

$$\pi_x(\theta \mid \mathbb{M}) = \sup_{m \in \mathbb{M}} \mathsf{P}_\theta^{(m)}\{\eta(X, \theta) \leq \eta(x, \theta)\}, \quad \theta \in \mathbb{T}. \tag{9}$$

Since there won't be any chance for confusion in what follows, I'll drop the dependence on $\mathbb{M}$ in the notation above, and just write "$\pi_x(\theta)$" for the right-hand side in (9). Note that each $\theta \mapsto \mathsf{P}_\theta^{(m)}\{\eta(X, \theta) \leq \eta(x, \theta)\}$ takes value 1 at the maximum likelihood estimator, so the right-hand side satisfies $\sup_\theta \pi_x(\theta) = 1$, hence is a possibility contour. Therefore, I can define a possibility measure $\overline{\Pi}_x(A) = \sup_{\theta \in A} \pi_x(\theta)$ exactly as before, and the same in-sample possibilistic reasoning can be applied. It's also immediately clear that the IM validity property (7) holds here too, so the derived methods are provably reliable. But the validity conclusions are broader because they hold uniformly over the models in $\mathscr{P}$. The



(a) $x = (n, y) = (10, 3)$

(b) $x = (n, y) = (16, 11)$

Figure 3: Plots of the model-specific possibility contours (binomial is red, negative binomial is blue) and the pointwise maximum (black) in (9). The green curve is the relative likelihood η and the orange curve is the truncated η_Q in (12), with $Q = \mathsf{Unif}(0, 1)$.

broader validity conclusions come at a price though: the supremum over $\mathbb{M}$ implies that the possibility contour in (9) is no more tightly concentrated than that corresponding to any m-specific model, hence a potential loss of efficiency, e.g., larger confidence sets. This loss of efficiency is unavoidable if one wants the likelihood principle and reliability guarantees. In any case, what I'm proposing is very much *via media* in spirit since the practitioner can control how close he is to satisfying the likelihood principle—and how much efficiency he stands to lose—through his choice of $\mathbb{M} \subseteq \mathbb{M}^\star$.

Returning to the Bernoulli trial illustration, suppose that $\mathbb{M}$ contains just the binomial and negative binomial models. Figure 3 shows plots of the two model-specific possibility contours as in (5) and the combined version in (9) for two different data sets $x = (n, y)$. This plot highlights the point that, thanks to sharing the same η, the model-specific possibility contours have overall similar shapes. This means that the pointwise maximum in (9) isn't going to be too much different from the individual curves, which is apparent in the plots. So, for example, the confidence intervals obtained by thresholding the three curves at level α are all about the same. The difference is in which model(s) the coverage probability claims apply to: the interval determined by (9) satisfies the coverage probability claim for *both* the binomial and negative binomial models.

The general case in (9) is computationally intimidating, and I don't presently have any recommendations on how this can be carried out efficiently, but bounds may be available; see below. I imagine, however, that a data analyst who is seriously concerned about both reliability and satisfying the likelihood principle can identify a relatively small finite set of plausible models in $\mathbb{M}^\star$ that deserve consideration. Then the computations wouldn't be much more difficult than those needed to generate the plots in Figure 3. It's in the user's best interest, after all, to be parsimonious in their choice of $\mathbb{M}$, since an overly generous choice will lead to unnecessary loss of efficiency.

I'll conclude this section by discussing the case where $\mathbb{M} = \mathbb{M}^\star$, i.e., where the user is entertaining literally all the models that share a common relative likelihood. Walley

(2002), for instance, develops a framework of (imprecise) probabilistic inference that both satisfies the likelihood principle and achieves a version of the validity result in (7). Let Q denote a generic prior probability distribution for Θ on $\mathbb{T}$ and define

$$\eta_Q(x,\theta) = \frac{L_x(\theta)}{\int L_x(\vartheta)\, Q(d\vartheta)}, \quad \theta \in \mathbb{T}.$$

If θ were some specific parameter value θ_0, then $\eta_Q(x,\theta_0)$ might be referred to as the *Bayes factor* for testing the null hypothesis "$\Theta = \theta_0$" against the alternative hypothesis "$\Theta \sim Q$." In addition to the Bayesian interpretation, η_Q has a few interesting and relevant properties. First, since η_Q only depends on the likelihood function up to proportionality, all the models in $\mathbb{M}^\star$ yield the same η_Q for a given Q, just like with η. Second, the reciprocal of η_Q (but not of η) satisfies

$$\mathsf{E}_\theta^{(m)}\{\eta_Q(X,\theta)^{-1}\} = 1, \quad \text{for all } \theta \in \mathbb{T},\, m \in \mathbb{M}, \tag{10}$$

where $\mathsf{E}_\theta^{(m)}$ denotes expected value with respect to $\mathsf{P}_\theta^{(m)}$. This follows easily because $x \mapsto \int p_\vartheta(x)\, Q(d\vartheta)$ defines a density/mass function. This shows that η_Q^{-1} determines an *e-value* or *e-process* (e.g., Ramdas et al. 2022). An immediate consequence is that η_Q achieves a property similar like in (7): by Markov's inequality and (10)

$$\mathsf{P}_\theta^{(m)}\{\eta_Q(X,\theta) \leq \alpha\} \leq \alpha\, \mathsf{E}_\theta^{(m)}\{\eta_Q(X,\theta)^{-1}\} = \alpha, \quad \alpha \in [0,1],\, \theta \in \mathbb{T},\, m \in \mathbb{M}. \tag{11}$$

Therefore, η_Q can be readily used to construct valid tests and confidence intervals as in the above corollary. Since the "$\leq \alpha$" bound in (11) holds uniformly in $m \in \mathbb{M}$, the conclusions can be strengthened to "anytime valid" (Ramdas et al. 2022), but I'll not explain these details here. Note, however, that $\theta \mapsto \eta_Q(x,\theta)$ is not a possibility contour; since $\eta_Q(X,\theta)^{-1}$ has expected value 1, aside from trivial cases, it'll surely take values greater than 1. But it can be truncated to a possibility contour,

$$\theta \mapsto \eta_Q(x,\theta) \wedge 1, \quad \theta \in \mathbb{T}. \tag{12}$$

One can imagine, however, that the procedures derived from thresholding the Q-specific possibility contour in (12) would be conservative, since the validity guarantees would have to hold for any user-specified Q. This conservatism is apparent in Figure 3.

The relative likelihood η and the corresponding η_Q are related via

$$\eta(x,\theta) = \inf_Q \eta_Q(x,\theta),$$

where the infimum is over all probability measures on $\mathbb{T}$, and it's attained at a measure that assigns probability 1 to set of maximizers of the likelihood for the given data x. Then the following strategy is tempting: first define a Q-specific possibility contour

$$\pi_x(\theta \mid Q) = \pi_x(\theta \mid \mathbb{M}, Q) = \sup_{m \in \mathbb{M}} \mathsf{P}_\theta^{(m)}\{\eta_Q(X,\theta) \leq \eta_Q(x,\theta)\}, \quad \theta \in \mathbb{T},$$

and then try removing the dependence on Q by optimizing again, i.e.,

$$\tilde{\pi}_x(\theta) = \inf_Q \pi_x(\theta \mid Q), \quad \theta \in \mathbb{T}.$$

This satisfies the likelihood principle, since it doesn't depend on any particular model m in $\mathbb{M} = \mathbb{M}^\star$. Moreover, by the bound in (11),

$$\tilde{\pi}_x(\theta) \leq \inf_Q \{\eta_Q(x, \theta) \wedge 1\} = \eta(x, \theta), \quad \theta \in \mathbb{T}.$$

Remember, the relative likelihood on the right-hand side above is a possibility contour but the corresponding possibilistic IM isn't valid—e.g., it doesn't satisfy (10) because $x \mapsto \sup_\vartheta p_\vartheta(x)$ isn't a density/mass function. The problem is that η tends to be too small which, together with the above bound, implies that $\tilde{\pi}_x$ doesn't define a valid IM either. We do, however, get the following insights:

- the relative likelihood-based possibility contour in (9), with $\mathbb{M} = \mathbb{M}^\star$, will tend to be less tightly concentrated than the relative likelihood itself, and
- at least intuitively, the relative likelihood-based possibility contour in (9) ought to be more tightly concentrated then $\eta_Q \wedge 1$ for any particular Q.

These observations are apparent in Figure 3. An interesting open question is whether the vague notion of "tight" that I'm using above could be related to the familiar, well-defined notion of *specificity* in the possibility theory literature (e.g., Dubois and Prade 1986). In any case, I'd feel comfortable upper bounding the possibility contour (9) in the challenging case with $\mathbb{M} = \mathbb{M}^\star$ by $\eta_Q \wedge 1$ for some not-too-tightly-concentrated Q.

7 Conclusion

In this paper, I revisited the *possibility* of achieving a middle-ground between Basu's Bayesian and frequentist poles. Resolving this open question would go a long way towards pinpointing statisticians' contribution and securing our seat at the data science table. While Fisher's efforts fell short, my claim is that there's still hope. The key new observation, as I described in Section 4, is that likelihood (model + data) is insufficient to reliably support probabilistic inference. This helps to justify consideration of other non-traditional modes of uncertainty quantification. Furthermore, I've argued (here and elsewhere) that likelihod can reliably support possibilistic inference, and I've offered a framework in which this can be carried out. There's still some work to be done, but I think almost all of the relevant details have been worked out in Martin (2022b). If I'm wrong and this isn't the *via media* that Fisher and others have been looking for, then I urge the reader to reach out and let me know what I'm missing.

For further developments, I'm very excited about the potential for incorporating partial prior information into the possibilistic IM, like I mentioned briefly in Section 5. A practically important and challenging problem—another favorite of Basu's (e.g., Basu 1977, 1978)—that tends to get overlooked is marginal inference in the presence of nuisance parameters. The possibility-theoretic framework offers a straightforward marginalization procedure that preserves validity; this is via the *extension principle* of Zadeh (1975). The downside is that this straightforward marginalization tends to be inefficient. To avoid this loss of efficiency, some form of dimension reduction is needed. The general profiling strategy I proposed in Martin (2022b, Sec. 7) seems promising, but I've since realized that, in certain cases, more efficient marginal inference can be achieved using other strategies besides profiling. So there are still more insights to be gleaned from Basu on this important question, and I'll report on these details elsewhere.

Acknowledgments

The author thanks an anonymous reviewer for valuable feedback on a previous version of the manuscript. This work is partially supported by the U.S. National Science Foundation, grant SES–2051225.

References

Augustin, T., Coolen, F. P. A., de Cooman, G., and Troffaes, M. C. M., editors (2014). *Introduction to Imprecise Probabilities*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

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.

Barnard, G. A. (1995). Pivotal models and the fiducial argument. *Int. Statist. Rev.*, 63(3):309–323.

Basu, D. (1964). Recovery of ancillary information. *Sankhyā Ser. A*, 26:3–16.

Basu, D. (1967). Problems relating to the existence of maximal and minimal elements in some families of statistics (subfields). In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*, pages 41–50. Univ. California Press, Berkeley, Calif.

Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā Ser. A*, 31:441–454.

Basu, D. (1971). An essay on the logical foundations of survey sampling. I. In *Foundations of statistical inference (Proc. Sympos., Univ. Waterloo, Waterloo, Ont., 1970)*, pages 203–242. Holt, Rinehart and Winston of Canada, Toronto, Ont. With comments by G. A. Barnard, V. P. Godambe, J. Hájek, J. C. Koop and R. Royall and a reply by the author.

Basu, D. (1975). Statistical information and likelihood. *Sankhyā Ser. A*, 37(1):1–71. Discussion and correspondance between Barnard and Basu.

Basu, D. (1977). On the elimination of nuisance parameters. *J. Amer. Statist. Assoc.*, 72(358):355–366.

Basu, D. (1978). On partial sufficiency: a review. *J. Statist. Plann. Inference*, 2(1):1–13.

Basu, D. (1980). Randomization analysis of experimental data: the Fisher randomization test. *J. Amer. Statist. Assoc.*, 75(371):575–595.

Basu, D. (2011). *Selected works of Debabrata Basu*. Selected Works in Probability and Statistics. Springer, New York. Edited by Anirban DasGupta.

Basu, D. and Ghosh, J. K. (1969). Invariant sets for translation-parameter families of measures. *Ann. Math. Statist.*, 40:162–174.

Berger, J. O. and Wolpert, R. L. (1984). *The Likelihood Principle.* Institute of Mathematical Statistics Lecture Notes—Monograph Series, 6. Institute of Mathematical Statistics, Hayward, CA.

Birnbaum, A. (1962). On the foundations of statistical inference. *J. Amer. Statist. Assoc.*, 57:269–326.

Brazzale, A. R. and Davison, A. C. (2008). Accurate parametric inference for small samples. *Statist. Sci.*, 23(4):465–484.

Cella, L. and Martin, R. (2023). Possibility-theoretic statistical inference offers performance and probativeness assurances. `arXiv:2304.05740`.

Choquet, G. (1953–1954). Theory of capacities. *Ann. Inst. Fourier, Grenoble*, 5:131–295 (1955).

Cuzzolin, F. (2021). *The Geometry of Uncertainty.* Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, Cham.

Dawid, A. P. (2020). Fiducial inference then and now. `arXiv:2012.10689`.

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.

Denœux, 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). The principle of minimum specificity as a basis for evidential reasoning. In *Proceedings of International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems*, pages 75–84. Springer.

Dubois, D. and Prade, H. (1988). *Possibility Theory.* Plenum Press, New York.

Efron, B. (1998). R. A. Fisher in the 21st century. *Statist. Sci.*, 13(2):95–122.

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. (1935). The fiducial argument in statistical inference. *Ann. Eugenics*, 6:391–398.

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

Fraser, D. A. S. (1961). On fiducial inference. *Ann. Math. Statist.*, 32:661–676.

Fraser, D. A. S. (1965). Fiducial consistency and group structure. *Biometrika*, 52:55–65.

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.

Hannig, J. (2009). On generalized fiducial inference. *Statist. Sinica*, 19(2):491–544.

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.

Hanss, M. (2005). *Applied Fuzzy Arithmetic.* Springer Berlin, Heidelberg.

Hose, D. (2022). *Possibilistic Reasoning with Imprecise Probabilities: Statistical Inference and Dynamic Filtering.* PhD thesis, University of Stuttgart.

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

Martin, R. (2019). False confidence, non-additive beliefs, and valid statistical inference. *Internat. J. Approx. Reason.*, 113:39–73.

Martin, R. (2021). An imprecise-probabilistic characterization of frequentist statistical inference. `arXiv:2112.10904`.

Martin, R. (2022a). Valid and efficient imprecise-probabilistic inference under partial priors, I. First results. `arXiv:2203.06703`.

Martin, R. (2022b). Valid and efficient imprecise-probabilistic inference under partial priors, II. General framework. `arXiv:2211.14567`.

Martin, R. (2023). Fiducial inference viewed through a possibility-theoretic inferential model lens. `arXiv:2303.08630`.

Martin, R. and Lin, Y. (2016). Exact prior-free probabilistic inference in a class of non-regular models. *Stat*, 5:312–321.

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. (2015). *Inferential Models*, volume 147 of *Monographs on Statistics and Applied Probability.* CRC Press, Boca Raton, FL.

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

Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. *Biometrika*, 32:128–150.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2022). Game-theoretic statistics and safe anytime-valid inference. `arXiv:2210.01948`.

Reid, N. (2003). Asymptotics and the theory of inference. *Ann. Statist.*, 31(6):1695–1731.

Savage, L. J. (1972). *The Foundations of Statistics.* Dover Publications, Inc., New York, revised edition.

Seidenfeld, T. (1992). R. A. Fisher's fiducial argument and Bayes' theorem. *Statist. Sci.*, 7(3):358–368.

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.

Troffaes, M. C. M. and de Cooman, G. (2014). *Lower Previsions.* Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*, volume 42 of *Monographs on Statistics and Applied Probability.* Chapman & Hall Ltd., London.

Walley, P. (2002). Reconciling frequentist properties with the likelihood principle. *J. Statist. Plann. Inference*, 105(1):35–65.

Wasserman, L. A. (1990). Belief functions and statistical inference. *Canad. J. Statist.*, 18(3):183–196.

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

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8:338–353.

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.