Old and new developments in consonant belief functions for statistical inference

Ryan Martin[1]
North Carolina State University
Researchers.One
www4.stat.ncsu.edu/~rgmarti3

6th BFAS School
Japan Advanced Institute of Science and Technology
Ishikawa, Japan — online for me :(
October 30th, 2023

[1]Research partially supported by the NSF, SES–2051225

Introduction

> *...as to what probability is and how it is connected with statistics, there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel.* —Savage

> *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

> *Unlike most philosophical arguments, this one has important practical consequences. The two philosophies represent competing visions of how science progresses and how mathematical thinking assists in that progress.* —Efron

Intro, cont.

- With 100+ years of experience, statisticians should be the authorities on uncertainty, but instead we side-step:

  ASA President's Task Force statement: *Different measures of uncertainty can complement one another*

- Lack of answers $\rightarrow$ confusion and distrust[2]
- Not having the answers is OK, but statisticians largely have no motivation to find answers, complacency
- I think this problem is too important to ignore

> *Is complacency in the face of contradiction acceptable for a central discipline of science?* —Fraser

[2] e.g., replication crisis in science

Intro, cont.

- Forward progress on foundations requires new insights
- Fortunately, we have giants' shoulders to stand on
  - Fisher
  - Dempster & Shafer
  - Dubois
  - ...
- My basic claim: *probability theory is lacking, reliable inference & UQ requires different considerations*
- "Different considerations" = imprecise probability
- That is, imprecision is *necessary*, not a choice
- I'm proposing a general imprecise-probabilistic framework:
  - consonant structures, possibility theory
  - balances Bayesian and frequentist reliability
  - can do things Bs & Fs can't

This talk

- precision pushes likelihood too far $\to$ unreliability
- consonant beliefs/possibility $=$ "correct" mode of imprecision
- possibilistic IM via probability-to-possibility transform
- properties & examples
- *partial prior IMs*

Warm-up

- Simple textbook example, hits on the key issues
- Cancer screening
  - data $T = \{\text{test positive}\}$
  - hypothesis $H = \{\text{has cancer}\}$
  - model: $\mathrm{P}(T \mid H) = 0.95$ and $\mathrm{P}(T \mid H^c) = 0.05$
  - prior $\mathrm{P}(H) = 0.01$
- *Question:* given a positive test — cancer or not?
- Polarizing example (re: Fraser)
  - Bayesians infer H^c
  - non-Bayesians infer H

Warm-up, cont.

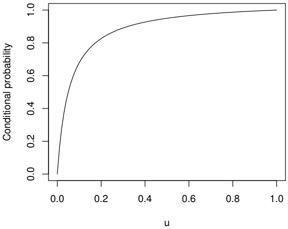
- Non-Bayesian solution:
  - basic *law of likelihood* reasoning

$$\frac{\mathrm{P}(T \mid H)}{\mathrm{P}(T \mid H^c)} = \frac{0.95}{0.05} = 19 \gg 1 \implies \text{infer } H$$

  - or can get a p-value: $\pi_T(H) = 1 \implies$ infer H
- Bayesian solution:
  - prior to posterior update is

$$u \mapsto \frac{0.95\, u}{0.95\, u + 0.05\,(1-u)}$$

  - if $u = 0.01$, then $\mathrm{P}(H \mid T) = 0.16$
  - $0.16 \ll \frac{1}{2} \implies$ infer H^c



Warm-up, cont.

Manski's *Law of Decreasing Credibility*.

Credibility of inference decreases with the strength of assumptions

- If we genuinely know $P(H)$, then
  - that info shouldn't be ignored
  - Bayesian solution *might* make sense
- If we don't know $P(H)$, then
  - absurd to pick $P(H) = 0.5$ (say) and claim "objectivity"
  - better to report the range of $P(H \mid T)$ as prior varies
  - this "Bayes" solution agrees with non-Bayes!
- Most likely case: we know *something* about $P(H)$
  - e.g., *we're pretty sure* $P(H) \in [u_{lo}, u_{hi}]$
  - textbooks are silent about this *most likely case!*

Setup

- Data $Y \in \mathbb{Y}$, observed value y
- Model $\{\mathsf{P}_{Y|\theta} : \theta \in \mathbb{T}\}$
- Uncertain true value is Θ, generic values denoted by θ
- For now, assume no prior info about Θ is available
- **Goal:** UQ/inference about Θ, given $Y = y$

- For example:
  - model $(Y \mid \Theta = \theta) \sim \mathsf{N}(\theta, 1)$
  - observation $Y = y$, e.g, $y = 7$
  - is the hypothesis "$\Theta > 8$" supported?

Setup, cont.

> *It would surely have been astonishing if all the complexities of such a subtle concept as probability in its application to scientific inference could be represented in terms of only three concepts—estimates, confidence intervals, and tests of hypotheses. Yet one would get the impression that this was possible from many textbooks purporting to expound the subject.* —Barnard

- Despite what textbooks say

  inference isn't just estimation, testing, and confidence sets

- A scientist whose never seen a stat textbook isn't going to ask specifically for a confidence interval!
- The UQ I'm after is richer & more informative

Setup, cont.

> *Statisticians want numerical measures of the degree to which data support hypotheses.* —Hacking

- UQ about Θ, given $Y = y$?
- Hacking says we want "numerical measures of..."
- Gut reaction:
  - "numerical measures of..." $\iff$ probability
  - i.e., return a "posterior" $(\Theta \mid Y = y) \sim \mathsf{Q}_y$
- However: *probability theory generally doesn't provide reliable UQ in the context statistical inference*

Pushing likelihood too far

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

- Data + model $\implies$ likelihood: $\theta \mapsto L_y(\theta)$
- Fisher's warnings "likelihood $\neq$ probability" are familiar
- But we don't take this seriously
- Fisher's point:[3] for a hypothesis "$\Theta \in H$"
  - "no differential element" $\implies$ size of H has no bearing
  - only *relative likelihood* values on H are relevant

[3] M., `arXiv:2303.17425`

Pushing likelihood too far, cont.

- Default prior Bayes (and others):
  - neither prior nor likelihood has meaningful differential element
  - how then can posterior's differential element be meaningful?
- e.g., $\{Y \sim \mathsf{N}(\Theta, 1),\ y = 7\} \rightsquigarrow \mathsf{Q}_y = \mathsf{N}(7, 1)$
  - two hypotheses $H = (6.5, 7]$ and $H' = (-\infty, 7]$
  - clearly, $\mathsf{Q}_y(H') \gg \mathsf{Q}_y(H)$
  - but is H' really more supported by data than H?

> *To allow the size of the crowd of [possibilities]... 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

Why imprecision?

- Pushing likelihood too far has consequences
- In particular, the following would be problematic:
  - for some false hypotheses $H \not\ni \Theta$
  - the rv $Y \mapsto \mathsf{Q}_Y(H)$ tends to be large under $\mathsf{P}_{Y|\Theta}$
- *False confidence*[4] is a particular form of unreliability that <u>all</u> probabilistic UQ suffers from

False confidence theorem.

Let Q_Y be *any* data-dependent probability for Θ. For any thresholds (ρ, τ), there exists hypotheses $H \subset \mathbb{T}$ such that

$$H \not\ni \Theta \quad \text{and} \quad \mathsf{P}_{Y|\Theta}\{\mathsf{Q}_Y(H) > \rho\} > \tau.$$

[4] Balch, M., & Ferson 2019, *Proc Roy Soc A*, `arXiv:1706.08565`

Why imprecision?, cont.

> *Money doesn't grow on trees.* —Dad

- New perspective on the false confidence theorem[5]
  - there exists gambles that are acceptable for each y
  - but are unacceptable, i.e., *loss-prone*, on average
- I'll skip the detailed justification here...
- But note the departure from how (coherent) prob works:
  - if $\mathsf{E}(Z \mid Y = y) \geq 0$ for all y
  - then "$\mathsf{E}(Z)$" shouldn't be negative

[5] Similar results can be found in, e.g., Ch. 7 of Walley's 1991 book, but these don't directly refer to *false confidence*

Why imprecision?, cont.

> *[Fisher] was the world's master of quantifying uncertainty, having developed many of the... procedures for doing so.* —Pearl

- The need for imprecision in UQ wasn't lost on Fisher
  - *[A p-value]... does not justify any exact probability statement*
  - *no exact probability statements can be based on [conf. sets]*
- "Fisher's biggest blunder" (Efron) was just that he didn't find the right mathematical formulation
- However:
  - Fisher + Dempster $\to$ origins of imprecise prob
  - Fisher + Neyman $\to$ confidence sets, inherently imprecise[6]

[6] M., `arXiv:2112.10904`

What imprecision?

- If not probability, then what? *Imprecise probability*
  - imprecise $\neq$ inaccurate
  - Manksi: limiting our precision for credibility's sake
- Roughly, imprecise probability = *sets of probabilities*
- For example:
  - roll a 6-sided die, one face is called "Ace"
  - suppose I only know that $\mathrm{P}(\mathrm{Ace}) \in [0.1, 0.2]$
  - a *set of dice*, don't know which will be rolled
  - should I assume a fair die, and be susceptible to sure monetary loss, just for the sake of precision?
  - or assume only the info that's given, avoid sure loss and maintain credibility, but give up precision?

What imprecision?, cont.

- Yes, whether statisticians acknowledge it or not

  we're dealing with sets of (Y, Θ) *joint distributions!*

- Explanation:
  - e.g., the *confidence level* of a set estimator $C(Y)$ is

$$\sup_{\theta} \mathsf{P}_{Y|\theta}\{C(Y) \not\ni \theta\} = \sup_{\text{all priors Q}} \underbrace{\int \mathsf{P}_{Y|\theta}\{C(Y) \not\ni \theta\}\, \mathsf{Q}(d\theta)}_{\text{linear functional of Q}}$$

  - "$=$" because the supremum of a linear functional over a convex set is on the boundary (point masses)

- Frequentist $=$ "every prior"

What imprecision?, cont.

- If even frequentists have sets of (Y, Θ) joint dist'ns, then
  - imprecision in inference is clear/unavoidable
  - helps to explain the false confidence phenomenon
- Data-dependent imprecise prob for (reliable) UQ...

Definition — *inferential model* (IM).

- given data y, model, etc
- to every $H \subseteq \mathbb{T}$, assign a pair of numbers

$$\underline{\Pi}_y(H) \quad \text{and} \quad \overline{\Pi}_y(H)$$

- y-dependent measures of $\underline{\text{support for}}$ and $\overline{\text{plausibility of}}$ H

What imprecision?, cont.

- There's a wide range of probability alternatives
  - belief functions
  - consonant beliefs/possibility measures
  - credal sets
  - lower/upper previsions
  - ...
- All have advantages and disadvantages
- My focus: *consonant belief functions/possibility measures*
- I'll justify this later; for now I defer to Shafer...

> *Specific items of evidence can often be treated as consonant, and there is at least one general type of evidence that seems well adapted to such treatment. This is inferential evidence—the evidence for a cause that is provided by an effect.* —Shafer

What imprecision?, cont.

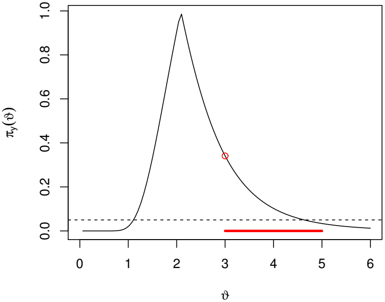
- *Possibility theory* is the simplest imprecise probability
- Dubois and others have written extensively on this
- Similar to probability theory
  - determined by a "density function"
  - different calculus — optimization replaces integration
  - no differential element!
- Specifically:
  - possibility contour $\theta \mapsto \pi_y(\theta)$ with $\sup_\theta \pi_y(\theta) = 1$
  - and then define

$$\overline{\Pi}_y(H) = \sup_{\theta \in H} \pi_y(\theta) \quad \text{and} \quad \underline{\Pi}_y(H) = 1 - \overline{\Pi}_y(H^c)$$

- *What makes the IM output meaningful is its properties*

What imprecision?, cont.

- Possibility contour plot
- Hypothesis $H = [3, 5]$
- $\overline{\Pi}_y(H)$ via optimization
  - "most possible member"
  - no differential element
  - size of H is irrelevant
- Level set defines interval of "sufficiently possible" θs



What imprecision?, cont.

- Possibility theory for statistics isn't my idea!
- A very natural choice of possibility contour/measure in stat applications (e.g., Shafer 1976, Ch. 11)
  - relative likelihood $R(y, \theta) = L_y(\theta) / \sup_\vartheta L_y(\vartheta)$
  - for fixed $Y = y$, set

$$\pi_y(\theta) = R(y, \theta), \quad \theta \in \mathbb{T}$$

  - and then $\overline{\Pi}_y(H) = \sup_{\theta \in H} \pi_y(\theta)$
- Further investigation into this proposal:
  - Wasserman (1990), *Canad. J. Stat.*
  - Denoeux (2014), *IJAR*
  - ...
- Shafer later (*JRSS-B*, 1982) rejected the idea in general
- I think it just doesn't go far enough...

Possibilistic IMs

- Special case of a more general construction[7]
- Workhorse: *imprecise-probability-to-possibility transform*[8]
- In the "no-prior" case, this takes a simple form:
  - same relative likelihood $R(y,\theta) = L_y(\theta) / \sup_\vartheta L_y(\vartheta)$
  - define a possibility contour

$$\pi_y(\theta) = \mathsf{P}_{Y|\theta}\{R(Y,\theta) \leq R(y,\theta)\}, \quad \theta \in \mathbb{T}$$

  - then $\overline{\Pi}_y(H) = \sup_{\theta \in H} \pi_y(\theta)$, etc.
- Not unfamiliar...[9]
- Unexpected connections to fiducial inference (later)

[7] M. arXiv:2211.14567

[8] e.g., Dubois et al (2004), *Rel. Comput.*; Hose's 2022 PhD thesis

[9] M. arXiv:1203.6665 and M. arXiv:1511.06733

Possibilistic IMs, cont.

What makes the IM output meaningful is its properties. —M

It is unacceptable if a procedure... of representing uncertain knowledge would, if used repeatedly, give systematically misleading conclusions. —Reid & Cox

Strong validity theorem.

The new IM construction above is strongly valid, i.e.,

$$\sup_{\Theta} \mathsf{P}_{Y|\Theta}\{\pi_Y(\Theta) \leq \alpha\} \leq \alpha, \quad \alpha \in [0,1]$$

- Remarks:
  - Most basic property, familiar for the no-prior case
  - Implies no false confidence

Possibilistic IMs, cont.

Corollary.

1. the test "reject $\Theta \in H$ iff $\overline{\Pi}_y(H) \leq \alpha$" satisfies

$$\sup_{\Theta \in H} \mathsf{P}_{Y|\Theta}\{\overline{\Pi}_Y(H) \leq \alpha\} \leq \alpha, \quad \alpha \in [0,1], \quad H \subseteq \mathbb{T}$$

2. the region $C_\alpha(y) = \{\theta : \pi_y(\theta) > \alpha\}$ satisfies

$$\sup_{\Theta} \mathsf{P}_{Y|\Theta}\{C_\alpha(Y) \not\ni \Theta\} \leq \alpha, \quad \alpha \in [0,1] \tag{1}$$

- These are the frequentist guarantees we know & love:
- $C'_\alpha(y) = \{\theta : R(y,\theta) > \alpha\}$ doesn't satisfy (1)

Possibilistic IMs, cont.

- There's a sort of converse to the above corollary
- Suggests consonance is inherent in frequentism
- No time to get into the details, but...

Theorem (M., `arXiv:2112.10904`).

Let $\{C_\alpha : \alpha \in [0,1]\}$ be a family of confidence regions for $\Phi = \phi(\Theta)$ that satisfies the following properties:

- Coverage. $\inf_\Theta \mathsf{P}_{Y|\Theta}\{C_\alpha(Y) \ni \phi(\Theta)\} \geq 1 - \alpha$ for all α
- Nested. if $\alpha \leq \beta$, then $C_\beta(y) \subseteq C_\alpha(y)$ for all y
- Compatible. *mild, but too messy...*

There exists a strongly valid possibilistic IM for Θ whose derived marginal plausibility regions $C_\alpha^\star(y)$ for Φ satisfy

$$C_\alpha^\star(y) \subseteq C_\alpha(y) \quad \text{for all } (y, \alpha) \in \mathbb{Y} \times [0,1].$$

Possibilistic IMs, cont.

Theorem.

The possibilistic IM's error rate control is *uniform in H*, i.e.,

$$\sup_{\Theta} \mathsf{P}_{Y|\Theta}\{\overline{\Pi}_Y(H) \leq \alpha \text{ for some } H \ni \Theta\} \leq \alpha$$

- Inference goal isn't just "test one H and done"
- Often the goal is to *probe*
  - e.g., if null is rejected, find other H's that are supported
  - follow-ups depend on data, so H-wise control isn't enough
- Corollary provides error rate control under probing[10]

[10] Cella and M. (202x), *IJAR*, `arXiv:2304.05740`

Possibilistic IMs, cont.

- Computation usually requires Monte Carlo
- For each θ on a grid
  - simulate independent $Y_\theta^{(m)} \sim \mathsf{P}_{Y \mid \theta}$, $m = 1, \ldots, M$
  - approximate the contour by

$$\pi_y(\theta) \approx \frac{1}{M} \sum_{m=1}^{M} 1\{R(Y_\theta^{(m)}, \theta) \leq R(y, \theta)\}$$

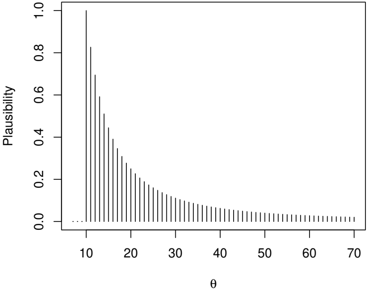
- Simplifications are possible in specific examples
- Importance sampling can also be used to lower expenses

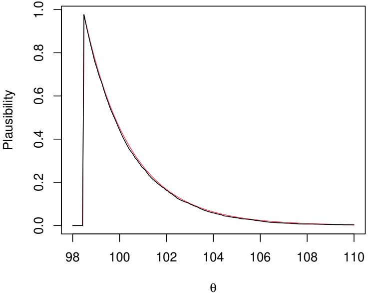
Possibilistic IMs, cont.

- Validity concerns errors like the following:
  - true hypotheses H that I assign small $\overline{\Pi}_y(H)$
  - true Θ is outside my confidence set
- *Efficiency* concerns other kinds of errors:
  - false hypotheses H that I assign not-small $\overline{\Pi}_y(H)$
  - false θ values inside my confidence set
- Efficiency is more difficult to describe mathematically
- One (rare) case where a comparison is clear
  - two IMs with contours π_y and $\tilde{\pi}_y$
  - if $\pi_y \leq \tilde{\pi}_y \ \forall \ y$, then former is more efficient than latter
- Roughly, efficiency to me means that π_Y *tends* to be more tightly concentrated as a function of Y

Examples: a couple uniforms

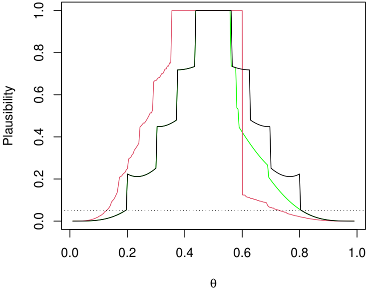
- Plots of $\theta \mapsto \pi_y(\theta)$ for two different uniform models
  - $\mathsf{P}_{Y|\theta} = \mathsf{Unif}\{1, 2, \ldots, \theta\}$, $Y_{(n)} = 10$, $n = 2$
  - $\mathsf{P}_{Y|\theta} = \mathsf{Unif}(\theta, \theta^2)$, $(Y_{(1)} = 281, Y_{(n)} = 9689)$, $n = 25$

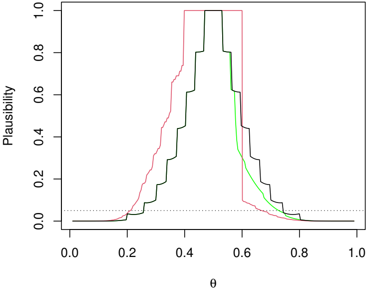




Example: binomial

- $\mathsf{P}_{Y|\theta} = \mathsf{Bin}(n, \theta)$
- Plots of the contour $\theta \mapsto \pi_y(\theta)$ — focus on black curve[11]
- Contours: $n = 8$ (left) and $n = 16$ (right), $\hat{\theta} = 0.5$





[11]Others: partial prior IMs that assume "90% sure that $\Theta \leq 0.6$"

Example: bivariate normal correlation

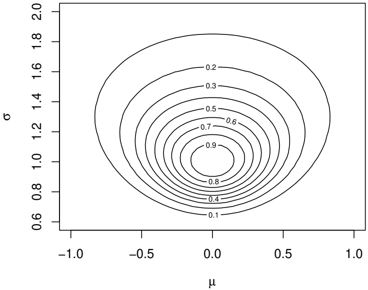
- Bivariate normal model:
  - means are 0, variances are 1
  - unknown correlation Θ
- Simulations to check validity & efficiency

			Θ				
n	Metric	Method	0.05	0.25	0.50	0.75	0.90
10	Coverage	IM	0.957	0.968	0.953	0.947	0.942
		$r^\star$	0.927	0.923	0.933	0.930	0.923
	Length	IM	1.004	0.979	0.888	0.619	0.242
		$r^\star$	0.947	0.919	0.818	0.540	0.238
25	Coverage	IM	0.949	0.960	0.957	0.950	0.941
		$r^\star$	0.938	0.943	0.948	0.944	0.942
	Length	IM	0.729	0.695	0.574	0.311	0.117
		$r^\star$	0.695	0.662	0.545	0.309	0.127

Table 1: Empirical coverage probabilities and expected lengths for the IM- and $r^\star$-based 95% confidence intervals in the bivariate normal correlation case described in Example 5.

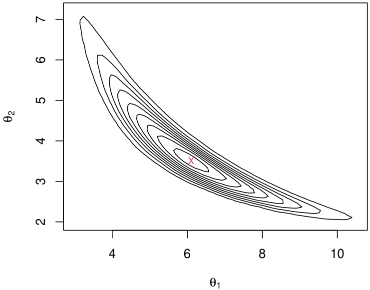
Example: two-parameter normal

- $\mathsf{P}_{Y|\theta} = \mathsf{N}(\mu, \sigma^2)$
- Data: $n = 10$, $(\hat{\mu}, \hat{\sigma}) = (0, 1)$
- Map of the IM possibility contour



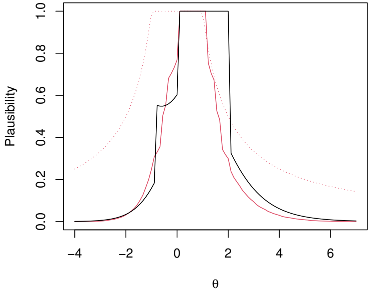
Example: two-parameter gamma

- $\mathsf{P}_{Y|\theta} = \text{Gamma}(\theta_1, \theta_2)$
- Simulated data, $n = 25$, $\theta_1 = 7$ and $\theta_2 = 3$
- Map of the IM possibility contour
- Computation via importance sampling



Example: two binomials

- Clinical trial:
  - association between treatment & control?
  - let Θ denote the log odds ratio[12]
- Data $(n_1, y_1) = (43, 1)$ and $(n_2, y_2) = (39, 2)$
- Focus on the black curve only[13]



[12] Nuisance parameters eliminated via conditioning (M., `arXiv:2309.13454`)
[13] Red is a partial prior IM that assumes "$\mathsf{E}|\Theta| \leq 1$"

Example: Behrens–Fisher

- Difference of normal means, general variances — *unsolved!*
- New marginal IM solution[14]
- Strong validity is guaranteed, sim to check efficiency:
  - difficult unbalanced case, $(n_1, n_2) = (2, 20)$
  - $\mu_1 = 2$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$
  - compare coverage prob of 90% confidence intervals

Method	Coverage Prob
Hsu–Scheffe	0.9738
Jeffreys	0.9296
Ghosh & Kim	0.7873
Welch	0.8362
1st order	0.7399
Fraser et al	0.8617
IM	**0.9082**

[14] M. (2023), `arXiv:2309.13454`

Possibilistic IMs in ML[15]

International Journal of Approximate Reasoning
Volume 150, November 2022, Pages 1-18

Valid inferential models for prediction in supervised learning problems ☆

Leonardo Cella, Ryan Martin

International Journal of Approximate Reasoning
Volume 151, December 2022, Pages 205-224

Direct and approximately valid probabilistic inference on a class of statistical functionals

Leonardo Cella [a], Ryan Martin [b]

---

[15] "Possibilistic" connection drawn in M., `arXiv:2309.13454`

IMs and fiducial

- Imprecise prob corresponds to sets of probabilities
- This (non-empty) set is called the *credal set*

$$\mathscr{C}(\overline{\Pi}_y) = \{\mathsf{Q}_y \in \mathsf{probs}(\mathbb{T}) : \mathsf{Q}_y(\cdot) \leq \overline{\Pi}_y(\cdot)\}$$

- For possibility measures $\overline{\Pi}_y$, there's a characterization:[16]

$$\mathsf{Q}_y \in \mathscr{C}(\overline{\Pi}_y) \iff \mathsf{Q}_y(\underbrace{\{\theta : \pi_y(\theta) > \alpha\}}_{100(1-\alpha)\%\text{ pl region}}) \geq 1 - \alpha$$

- i.e., elements of $\mathscr{C}(\overline{\Pi}_y)$ are confidence distributions
- More can be said, with additional structure

[16] Destercke & Dubois, Ch. 4 of *Intro to IP*

IMs and fiducial, cont.

- Suppose the model has a group transformation structure
- e.g., location–scale models
- With such structure, all the following methods give the same solution for probabilistic UQ:
  - default-prior Bayes
  - (generalized) fiducial
  - ...

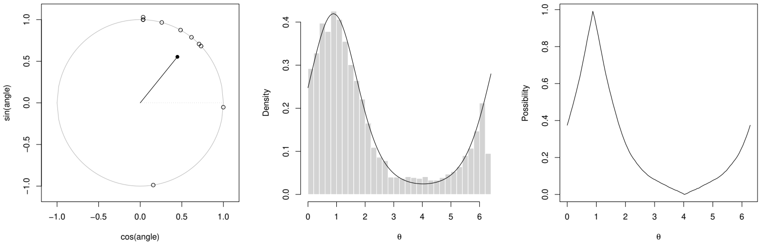
Theorem (M., ISIPTA'23, `arXiv:2303.08630`).

Under the structure above, let $\mathsf{Q}_y^\star$ be the fiducial solution and $\overline{\Pi}_y$ the possibilistic IM. Then $\mathsf{Q}_y^\star \in \mathscr{C}(\overline{\Pi}_y)$ and, moreover,

$$\mathsf{Q}_y^\star(\{\theta : \pi_y(\theta) > \alpha\}) = 1 - \alpha, \quad \text{all } \alpha \in [0,1]$$

IMs and fiducial, cont.

- Directional data, i.e., angles or points on a circle
  - simple von Mises model with unknown mean direction Θ
  - group structure: Θ plays the role of a rotation
- Plots show data,[17] fiducial density, and IM contour[18]



[17] From Example 1 in Mardia's *Directional Statistics*

[18] Details in M., ISIPTA'23, `arXiv:2303.08630`

Example: causal inference[19]

- Does the treatment cause a change in the outcome?
- Focus here on randomized experiments in finite populations
  - potential outcomes $\{Y_i(0), Y_i(1)\}$ are fixed, $i = 1, \ldots, n$
  - randomization determines which of $Y_i(0)$, $Y_i(1)$ are observed
  - covariates x_i might also be available
- Interest is in the differences "$Y_i(1) - Y_i(0)$"
- Neyman defines average treatment effect

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\}$$

and then does the usual "estimate $\pm$ 2 $\times$ std err"

[19] This is new/ongoing work

Example: causal inference, cont.

> *Causal inference requires causal hypotheses.* —Pearl

- My opinion: Neyman falls short of "causal hypotheses"
  - knowing τ doesn't *explain* the effect of treatment
  - i.e., the unit level differences $Y_i(1) - Y_i(0)$ could vary wildly, positive and negative, and give the same τ
- My view:
  - no free lunch, we need to work for causation
  - specifically, we need to posit a sort of "model" that directly explains the unit-level differences
- But (a new perspective on) Fisher's approach[20] does offer explanations and, hence, a route to valid causal inference

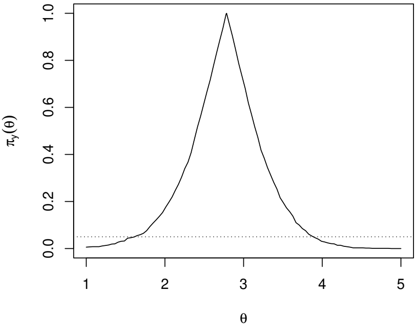
[20] Often called "Fisher's randomization test"

Example: causal inference, cont.

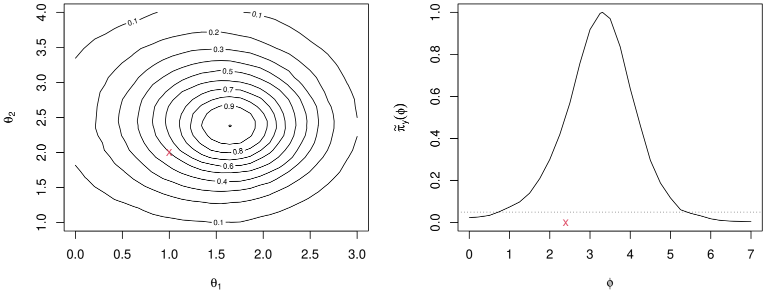
- My (new?) perspective on Fisher's proposal:
  - user posits $Y_i(1) - Y_i(0) = h(x_i, \Theta)$ for all i
  - where h is a given function
  - and Θ is the *causal effect* to be inferred
- If I know $\Theta = \theta$ and observed y,
  - then I know the missing potential outcomes
  - hence an *explanation* of treatment effect
- Too-quick explanation of the strategy:
  - posit a value θ, determines $\{Y_i(0), Y_i(1)\}$
  - if I know the potential outcomes, then I can get evaluate the contour (p-value) at θ via randomization/Monte Carlo
  - repeat for all θ on a grid

Example: causal inference, cont.

- Simple case, no x's
- Monte Carlo approx'n of IM contour via random treatment allocations
- Simulated data:
  - $n = 20$
  - true effect is 3



- Slightly more complicated example:
  - $n = 30$
  - one covariate
  - $h(x, \theta) = \theta_1 + \theta_2 x$
- Contour for Θ (left), marginal contour for $h(x^\star, \Theta)$ (right)



FAQs

- There are a couple questions I'm often asked
  - why consonant structures?
  - you don't employ the full power of the DS calculus — why?
- Below I offer answers to these questions

FAQs: Why consonance?

- Consonant beliefs are very special, extra structure
- In particular, if $(\underline{\Pi}_y, \overline{\Pi}_y)$ is consonant, then

$$\underline{\Pi}_y(H) > 0 \implies \overline{\Pi}_y(H) = 1$$
$$\overline{\Pi}_y(H) < 1 \implies \underline{\Pi}_y(H) = 0$$

- Some might say a consonant IM is "too imprecise"
- So justification is needed for imposing such structure
- I have *two* reasons for this, see below

FAQs: Why consonance?, I

- In general, strong validity is needed to ensure that the IM induces confidence sets with nominal coverage probability
- If we take strong validity + efficiency as desiderata, then consonance is necessary

Proposition (M., `arXiv:2211.14567`).

For any strongly valid IM with upper prob $\overline{\Gamma}_y$, there exists a strongly valid consonant IM with upper prob $\overline{\Pi}_y$ such that

$$\overline{\Pi}_y(H) \leq \overline{\Gamma}_y(H) \quad \text{for all } H$$

FAQs: Why consonance?, II

- IM's output: *not bounds on a "true" precise prob for* Θ
- So, consonance $\not\Longrightarrow$ loose bounds on this "true" prob
- In fact, tighter bounds wouldn't add value to the IM
- Recall *Cournot's Principle*[21]
  - roughly, only the small & large probability values are meaningful in the real world
  - i.e., suggests what won't & will happen, respectively
- For us, Cournot says *inference* can be made only when

$$\underline{\Pi}_y(H) \text{ is large} \quad \text{or} \quad \overline{\Pi}_y(H) \text{ is small}$$

- That the complementary term is trivial doesn't matter

[21] Shafer & Vovk (2019), e.g., Ch. 10.2

FAQs: Why no DS rules?

- Note: I'm not following any formal rules (e.g., Dempster's)
- that is, $\underline{\Pi}_{(y_1,y_2)}(\cdot) \neq \underline{\Pi}_{y_1}(\cdot) \star \underline{\Pi}_{y_2}(\cdot)$ for any rule $\star$
- This is problematic to some...
- My perspective is different:
  - my top priority is validy + efficiency
  - enforcing a $\star$-identity like above is a constraint
  - can't be more efficient with constraint than without

FAQs: Why no DS rules?, cont.

- Of course, if necessary, I could use such a rule
  - either because I insist on it
  - or if the problem doesn't give me access to the full data (y_1, y_2)
  - e.g., if I only have access to π_{y_1} & π_{y_2}, so couldn't directing construct $\pi_{(y_1,y_2)}$ as described above
- Natural idea: use a consonance-preserving rule,[22] e.g.,

$$\pi_{y_1}(\theta) \star \pi_{y_2}(\theta) = \frac{\pi_{y_1}(\theta) \circ \pi_{y_2}(\theta)}{\sup_\vartheta \pi_{y_1}(\vartheta) \circ \pi_{y_2}(\vartheta)}, \quad \circ = \text{t-norm}$$

- *Question:* Is this (strongly) valid?

[22] Dubois & Prade (*Comp Intell* 1988); M. & Syring, ISIPTA'19

Partial prior IMs

frequentist		Bayesian
every prior	← some priors →	one prior

- This goes back to the warm-up example...
- There's a spectrum of problems indexed by prior info
- Key observations:
  - spectrum can't be described with ordinary probability
  - applying Bayes to a set of priors is inefficient (dilation)

Partial prior IMs, cont.

frequentist		Bayesian
every prior	← some priors →	one prior

- Especially important in high-dim problems
- Main-stream methods for regularization are awkward:
  - Fs' penalties aren't part of the model/assumptions
  - Bs' priors impose more structure than is justified
- New idea:
  - incorporates into the model exactly the structural assumptions that can be justified, no more & no less
  - different from generalized/robust Bayes

Partial prior IMs, cont.

- For simplicity:
  - precise model with likelihood $L_y(\theta)$
  - imprecise prior via possibility contour $q(\theta)$
- $\overline{\Pi}_y$ is a possibility measure given by

$$\overline{\Pi}_y(H) = \sup_{\theta \in H} \pi_y(\theta), \quad H \subseteq \mathbb{T}$$

  - where the contour function[23] is

$$\pi_y(\theta) = \int_0^1 \Big[\sup_{\vartheta : q(\vartheta) > s} \mathsf{P}_{Y|\vartheta}\{R(Y, \vartheta) \leq R(y, \theta)\} \Big] \, ds$$

  - and R is a "relative likelihood"

$$R(y, \theta) = \frac{L_y(\theta)\, q(\theta)}{\sup_{\vartheta \in \mathbb{T}} L_y(\vartheta)\, q(\vartheta)}$$

[23] This is a special kind of Choquet integral...

Partial prior IMs, cont.

Strong validity theorem.

The new IM construction above is strongly valid (relative to the partial prior information), i.e.,

$$\overline{\mathsf{P}}_{Y,\Theta}\{\pi_Y(\Theta) \leq \alpha\} \leq \alpha, \quad \alpha \in [0,1]$$

- $\overline{\mathsf{P}}_{Y,\Theta}$ is the upper joint dist'n:

$$\overline{\mathsf{P}}_{Y,\Theta}(Y \in A, \Theta \in H) = \sup_{\mathsf{Q} \in \mathscr{C}(q)} \int_H \mathsf{P}_{Y|\theta}(Y \in A)\, \mathsf{Q}(d\theta)$$

- Above isn't as strict as validity wrt vacuous prior
  - creates an opportunity for improved efficiency
  - without sacrificing error control guarantees

Partial prior IMs, cont.

- Statistical procedures (tests & conf sets) derived from the partial prior IM have properties similar to those above
- There are also some corresponding behavioral properties
- In particular,[24]
  - treated as a data-dependent update of a partial prior to a "posterior" $(\underline{\Pi}_y, \overline{\Pi}_y)$, it avoids sure loss
  - it satisfies a little more than no-sure-loss, but not coherent
- Observation deserving further investigation:
  - generalized Bayes is coherent but often inefficient
  - partial prior IM is less conservative/more efficient
  - IM's contraction–dilation balance?

[24] M., `arXiv:2211.14567`

Partial prior IMs, cont.

- Computation is more challenging now...
- Naive approach:[25]
  - for each ϑ on a grid in $\mathbb{T}$, take iid samples $Y_\vartheta^{(m)} \sim \mathsf{P}_{Y|\vartheta}$, $m = 1, \ldots, M$
  - then, for each $\theta \in \mathbb{T}$, use the approximation

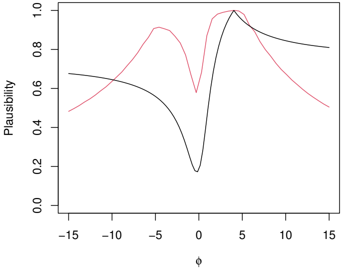
$$\pi_y(\theta) \approx \int_0^1 \max_{\vartheta : q(\vartheta) > s} \Big[\frac{1}{M} \sum_{m=1}^M 1\{R(Y_\vartheta^{(m)}, \vartheta) \leq R(y, \theta)\} \Big] \, ds$$

- This is doable in applications — I've done it!
- Surely more efficient strategies are available

[25] Hose, Hanss, M., *BELIEF'21*

Example: Fieller–Creasy

- Ratio of two normal means[26] (e.g., instrumental variables)
- Impossibility theorem:[27] *No set estimator has finite length with probability 1 and coverage probability* > 0
- Two marginal IM contours
  - vacuous prior
  - "$\mathsf{E}|\Theta| \leq 5$"
- Partial-prior IM
  - $\overline{\mathsf{P}}_{Y,\Theta}\{\pi_Y(\Theta) \leq \alpha\} \leq \alpha$
  - relaxed coverage prob
  - impossible $\to$ possible?

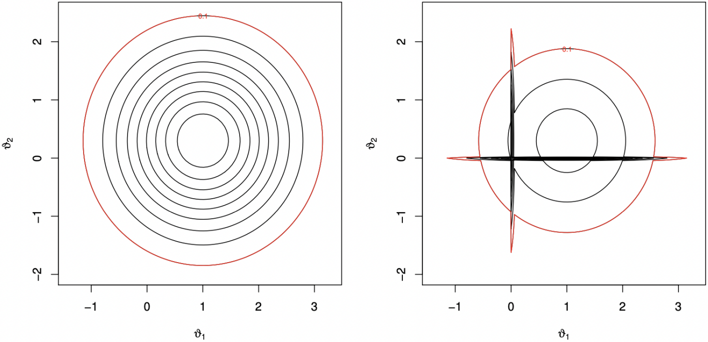


[26] M., arXiv:2309.13454

[27] Gleser & Hwang (1987), *Ann. Statist.*; Dufour (1994), *Econometrica*

Example: "high-dim" case

- $(Y \mid \Theta = \theta) \sim \mathsf{N}_2(\theta, I)$, observed $y = (1, 0.3)$
- Sparsity encouraging partial prior
- Compare vacuous-prior and partial-prior IMs



(a) vacuous-prior IM (b) sparsity prior consonant IM

Warm-up $\searrow$ cool-down

- Recall: $H = \{\text{cancer}\}$
- Practically realistic case: "$\mathsf{P}(H) \leq 0.05$" (say)
- Can encode this with a partial prior contour:

$$q(H) = 0.05 \quad \text{and} \quad q(H^c) = 1$$

Vacuous Prior	*Partial Prior*
$\underline{\Pi}_{\text{neg}}(H) = 0.00$	$\underline{\Pi}_{\text{neg}}(H) = 0.00$
$\overline{\Pi}_{\text{neg}}(H) = 0.05 \implies \text{infer } H^c$	$\overline{\Pi}_{\text{neg}}(H) < 0.01 \implies \text{infer } H^c$
$\underline{\Pi}_{\text{pos}}(H) = 0.95 \implies \text{infer } H$	$\underline{\Pi}_{\text{pos}}(H) = 0.00$
$\overline{\Pi}_{\text{pos}}(H) = 1.00$	$\overline{\Pi}_{\text{pos}}(H) = 0.05 \implies \text{infer } H^c$

Conclusion

> *Perhaps the most important unresolved problem in statistical inference is the use of Bayes theorem in the absence of prior information.* —Efron

- Statistical principles/foundations matter!
- Fisher had the right ideas but...
- Imprecision is imperative for reliable UQ
- New possibilistic IM framework:
  - likelihood-based
  - relatively simple (consonance)
  - achieves strong validity
  - incorporates partial prior info

Conclusion, cont.

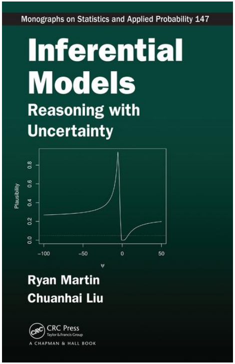
- Open methodological questions:
  - efficient Choquet integral computation?
  - partial prior elicitation?
  - high-dim problems?
  - valid (possibilistic) model assessment/selection?
- Open theoretical questions:
  - IM vs generalized Bayes contraction–dilation
  - decision-theoretic considerations?
  - imprecision & asymptotics?
  - revisiting "impossibility theorems" & paradoxes?
- Open philosophical questions:
  - "what makes IM output meaningful is its properties"?
  - frequentist–Bayesian spectrum?
- Open applications: *all of them!*

References

- Learned a lot since my first book
- *n*-part series of working papers:

  Valid and efficient imprecise-probabilistic inference with partial priors

- $n = 3$ parts so far
  1. arXiv:2203.06703
  2. arXiv:2211.14567
  3. arXiv:2309.13454
- Working towards a new book...



The end

`https://wordpress-courses2223.wolfware.ncsu.edu/st-790-001-fall-2022/`

- Papers, talks, etc? `www4.stat.ncsu.edu/~rmartin/`
- Question, etc? `rgmarti3@ncsu.edu`

Thanks for your attention!