Statistics, imprecise probabilities, and possibilistic inferential models

Ryan Martin[1]
Purdue University
martinrg@purdue.edu

SIPTA School
LMU Munich
30 July 2026

[1]Research partially supported by the NSF, DMS–2412628

Introduction

- *Imprecision is imperative for reliable UQ*
- That is, if the goal is to quantify uncertainty about unknowns based on data in a reliable and probability-like way, then that probability must be imprecise
- Imprecision isn't an option: it's an obligation
- Means that imprecise prob is fundamental (to statistics)
- So, we SIPTAns have opportunities for practical impact

Intro, cont.

- Goal of the lectures: inform & inspire
- I put more weight on *inspire* than on *inform*
- Given inspiration, you can easily learn the details
- Applied topic + *informing* emphasis means my presentation will be more high level than others
- That doesn't mean it's thin on theory!

Outline

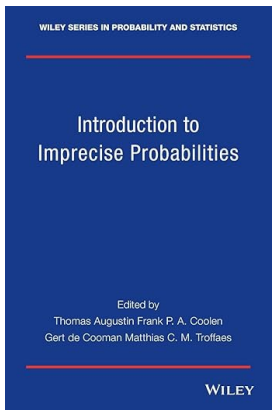
Part I

- what is stat inference?
- ingredients
- sampling distributions
- logic
- likelihood & Bayes
- shortcomings
- *imprecision is imperative*

Part II

- inferential models (IMs)
- possibilism: why & how?
- construction
- properties
- computation
- examples
- extensions...

Dedication — Thomas Augustin



Statistical inference

Thomas Augustin[1], Gero Walter[1], and Frank P. A. Coolen[2]

[1]*Department of Statistics, Ludwig-Maximilians University Munich (LMU), Germany*

[2]*Department of Mathematical Sciences, Durham University, UK*

Dedication, cont.

Research article Get rights and content ↗

Thomas Augustin's Contributions to Imprecise Probability and Statistics

F. Coolen [a], F. Cozman [b], T. Denoeux [c d], C. Jansen [e], H. Küchenhoff [f], R. Martin [g], E. Miranda [h], J. Plass [i], J. Rodema

Cite Add to Mendeley Share | 10.1016/j.ijar.2026.109781 ↗

Thomas was a good person and, alongside all his academic achievements, it is as such that he should be remembered above all else.

Part I

What is statistics?

- *Statistics* can mean lots of different things
- e.g., exploratory data analysis: plots & numerical summaries
- Things I'll cover in this tutorial:
  - statistical inference
  - prediction
  - decision-making
- Other things not covered in this tutorial:
  - classification
  - model assessment/choice
  - ...

What is statistical inference?

- Learning about the world from observations
- A fundamental problem, back to Hume...
- Closely related to induction, just with less committment to the conclusions
- We want be "pretty sure" answers are right
- So, reliability of the methods used is critical

Ingredients of a stat inference problem

- Three main ingredients:
  - scientific objective
  - data
  - statistical model
- Maybe other things I didn't mention...

Ingredients, cont.

- The *scientific objective* provides context
- Data analysis isn't done in a vaccuum
- This is what gives the statistician *skin in the game*
  - makes it more than a math problem
  - gives us something to lose if we're wrong
  - motivates us to question assumptions, push ourselves for the best possible solution, etc.
- Important, but often overlooked in courses and texts, perhaps because it can't be artificially manufactured

Ingredients, cont.

- The *data* is what we actually work with — there's no statistical inference without data
- Mathematically:
  - hypothetical/observable data denoted by X (or X^n)
  - observed data denoted by x (or x^n)
- Critical that data be "relevant" to the objective
- How to collect relevant data is beyond our scope
- Remarks:
  - data is everywhere nowadays
  - driving advances in data science, etc
  - data size can't make up for low quality
  - "garbage in, garbage out"

Ingredients, cont.

- Roughly, the *statistical model* helps us to mathematize the scientific objectives
- Scientific questions concern the "world," which is quantified by the posited statistical model
- Relevant features of the world correspond to features of the model parameter $\Theta \in \mathbb{T}$ — defines our "inferential target"
- Model choice is extremely difficult, beyond our scope

Our statistical problem statement.

The model is $X^n = (X_1, \ldots, X_n) \overset{\text{iid}}{\sim} \mathrm{P}_\Theta$, with Θ *unknown*, and the goal is to "learn" about Θ from observed $X^n = x^n$

Sampling distributions

- Goal: "learn" about unknown θ from observed $X^n = x^n$
- Can't "learn" just from staring at a list of numbers!
- So, start with a summary of the data
- A statistic $T_n = T(X^n)$ is a function of X^n
  - sum or average
  - order statistic or median
  - ...
- The *sampling distribution* of statistic T_n is the distribution derived from the joint distribution of X^n

Sampling distributions, cont.

Replication	Data	Statistic
1	$X_1^{(1)}, \ldots, X_n^{(1)}$	$T_n^{(1)} = T(X_1^{(1)}, \ldots, X_n^{(1)})$
2	$X_1^{(2)}, \ldots, X_n^{(2)}$	$T_n^{(2)} = T(X_1^{(2)}, \ldots, X_n^{(2)})$
$\vdots$		
r	$X_1^{(r)}, \ldots, X_n^{(r)}$	$T_n^{(r)} = T(X_1^{(r)}, \ldots, X_n^{(r)})$
$\vdots$		

- Schematic of a sampling distribution
- A histogram of the right-most column provides a visualization of the sampling distribution of T_n

Sampling distribution, cont.

- In principle, the sampling distribution of T_n can be derived from the joint distribution of X^n
- You learn these strategies in a math-stat course
- Can be difficult to do in general...
- Need ways to approximate the sampling distribution:
  - asymptotic, i.e., limits as $n \to \infty$
  - numerical, i.e., via computer simulations
- Both are useful, involve stochastic convergence:
  - in probability
  - in distribution
  - ...

Sampling distributions, cont.

Definition.

$T_n \to T_\infty$ in probability as $n \to \infty$ if, for all $\varepsilon > 0$,

$$\mathrm{P}\{|T_n - T_\infty| > \varepsilon\} \to 0, \quad \text{as } n \to \infty$$

Definition.

$T_n \to T_\infty$ in distribution as $n \to \infty$ if

$$\mathrm{P}(T_n \leq t) \to \mathrm{P}(T_\infty \leq t), \quad \text{as } n \to \infty$$

for all t at which the right-hand side is continuous

- Former: T_n and T_∞ are close with high probability
- Latter: sampling dist'n of $T_n \approx$ sampling dist'n of T_∞

Sampling distributions, cont.

- A particularly nice and commonly used statistic is the sample mean, $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$
- Averaging operation is stabilizing
- Below are two classical results about $\bar{X}_n$

Law of large numbers.

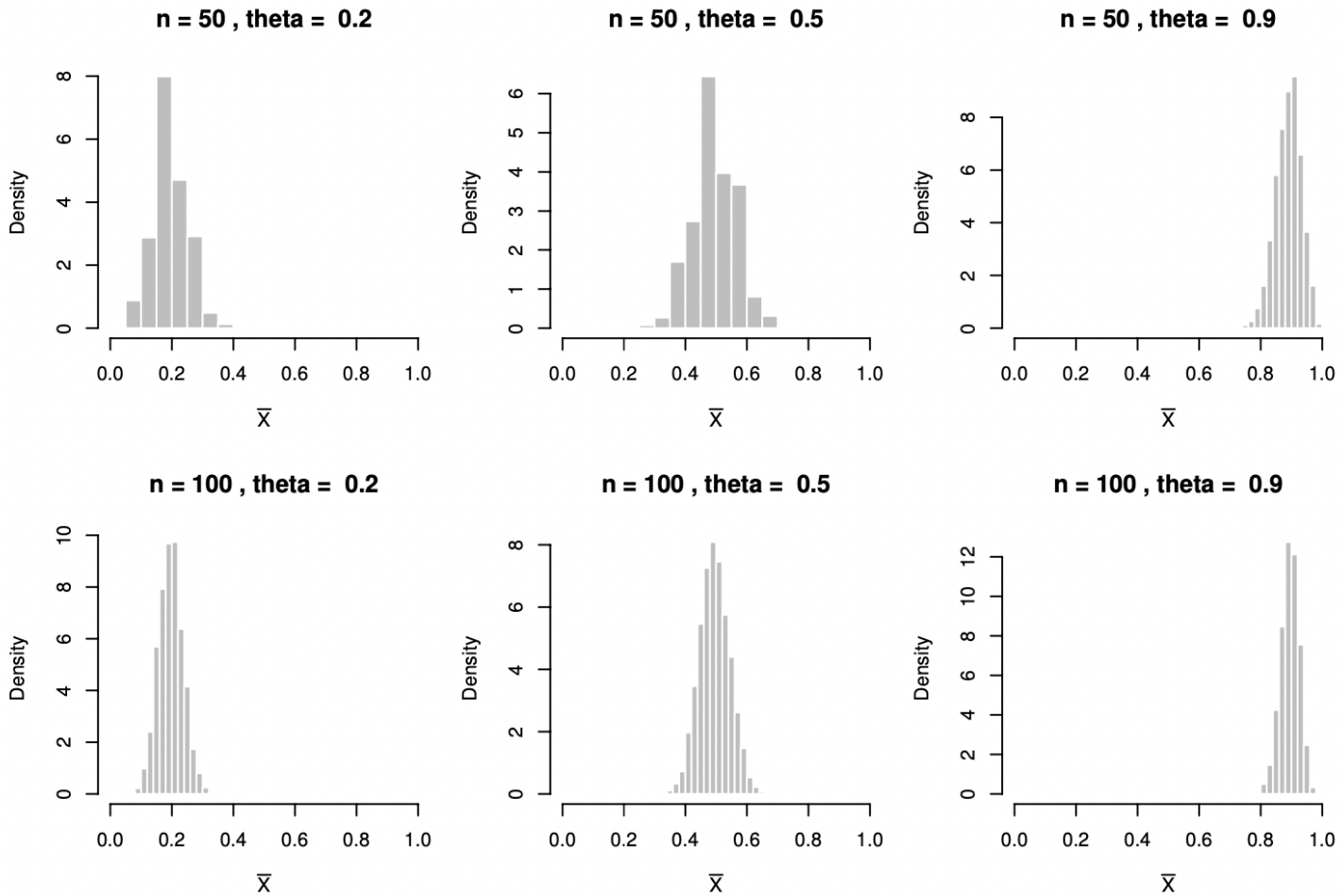
If $X_1, X_2, \ldots$ are iid with mean μ, then $\bar{X}_n \to \mu$ in probability

Central limit theorem.

If $X_1, X_2, \ldots$ are iid with mean μ and variance σ^2, then

$$n^{1/2}(\bar{X}_n - \mu) \to \mathrm{N}(0, \sigma^2) \quad \text{in distribution}$$

Sampling distributions, cont.



Logic of statistical inference

> *For everyone who does habitually attempt the difficult task of making sense of figures is, in fact, assaying a logical process of the kind we call induction, in that he is attempting to draw inferences from the particular to the general, from the sample to the population.* —Fisher

Logic, cont.

- Setup:
  - "$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{P}_\Theta$ with unknown $\Theta \in \mathbb{T}$"
  - goal is to "learn" about Θ from observations $X^n = x^n$
- Unlike in mathematics, we can't *deduce*, solely from the given info, statements about Θ that are surely true
- Why? Data is an imperfect representation of the world
  - we might do everything right
  - but we reach a wrong conclusion
- So, the familiar deductive logic doesn't work here

Logic, cont.

- *Inductive logic* is what we need for statistical inference
- We do this all the time without second thought
- Example:
  - I claim a coin is fair
  - to test my claim, you toss the coin 10 times
  - if you observe 10 heads, then you doubt my claim
- What's your reasoning?
  - 10-out-of-10 heads is improbable if my claim were true
  - you don't expect improbable events to occur
  - so your tentative conclusion is that my claim is wrong
- This is intuitive, but there's a lot going on...

Logic, cont.

- There are a number of requirements
  - relevant summary of the data
  - means to evaluate "probability"
  - a rule for classifying events as "improbable"
- Straightforward in the coin flipping example
- More effort needed to flesh this out in general
- Sampling distribution connection should be clear!
- Key points:
  - can't guarantee conclusions are true
  - force of the argument is it's inherent reliability
  - i.e., conclusions are false iff an "improbable" event occurs which, by definition, is improbable

Logic, cont.

- Again: inductive logic doesn't guarantee true conclusions
- e.g., not impossible to get 10-out-of-10 heads from a fair coin
- Some might say this makes statistical inference "soft"
- But (Fisher and) I disagree:
  - the real world is dirty
  - deduction stays clean/pure by working in a fantasy world
  - inductive reasoning's strength is that it can handle the dirt

> *In deductive reasoning all knowledge attainable is already latent in the postulates... In inductive reasoning we are performing part of the process by which new knowledge is created.* —Fisher

Logic, cont.

- General inductive logic for statistical inference?
- To the relevant questions (about Θ), we want to assign a data-driven answer, say $\textsc{ans}(X^n)$
- Answer could be wrong or not-wrong, but since we don't know Θ we can't make this determination
- Ideally,[2] we find procedures $\textsc{ans}(\cdot)$ such that

$$\mathsf{P}_\theta\{\textsc{ans}(X^n) \text{ is "wrong"}\} \quad \text{is small for all } \theta$$

- Definition of "wrong" must be non-data-dependent
- The above is a sampling distribution property

[2] Typically, we can't achieve this ideal, so some relaxation is needed...

Logic to practice

- Suppose the goal is to estimate $\phi(\Theta)$
- An *estimator* is a function $\hat{\phi}_n = \hat{\phi}(X^n)$ of the data X^n
- i.e., $\hat{\phi}_n$ an "answer" to the question "what is $\phi(\Theta)$?"
- Answer is "wrong" if $|\hat{\phi}_n - \phi(\Theta)| > \varepsilon$, for pre-specified $\varepsilon > 0$
- Then we'd want $\hat{\phi}$ to be such that

$$\mathrm{P}_\theta\{\underbrace{|\hat{\phi}(X^n) - \phi(\theta)| > \varepsilon}_{\text{"answer is wrong"}}\} \quad \text{is small for all } \theta$$

- Skin in the game $\Longrightarrow$ find $\hat{\phi}$ to optimize all aspects

Logic to practice, cont.

- Suppose the goal is to *test a hypothesis* $H : \Theta \in A$
- The *test* is a function $\delta : \mathbb{X}^n \to \{0, 1\}$, with

$$\delta(x^n) = \begin{cases} 1 & \text{pick not-}H \\ 0 & \text{pick } H \end{cases}$$

- A *Type I error* corresponds to $\delta(x^n) = 1$ when $\Theta \in A$
- Given $\alpha \in (0, 1)$, goal is that the test $\delta = \delta_\alpha$ have bounded *Type I error probability*, i.e.,

$$\sup_{\theta \in A} \mathsf{P}_\theta\{ \underbrace{\delta_\alpha(X^n) = 1}_{\text{“answer is wrong”}} \} \leq \alpha$$

- Skin in the game $\Longrightarrow$ find such δ_α with “maximal power”

Logic to practice, cont.

- Suppose the goal is to construct a *set estimator*
- Given $\alpha \in (0, 1)$, define $C_\alpha(x^n) \subseteq \phi(\mathbb{T})$
- Goal is that we're "confident" $C_\alpha(x^n)$ contains true $\phi(\theta)$, i.e.,

$$\sup_\theta \mathsf{P}_\theta\{\underbrace{C_\alpha(X^n) \not\ni \phi(\theta)}_{\text{"answer is wrong"}}\} \leq \alpha$$

- C_α that satisfies above property is a *confidence set*
- Skin in the game $\Longrightarrow$ find "smallest" such C_α

Summary so far

- My "induction + skin-in-the-game" logical perspective might not be familiar, so I'm writing a book about it

Fundamentals of Statistical Inference

Intuition, Logic, Theory, and their Use

Ryan Martin
North Carolina State University
www4.stat.ncsu.edu/~rmartin

DRAFT: January 5, 2025

- Next I'll show that the above goals can be accomplished, at least approximately, with methods based on *likelihood*
- But my primary focus is on a broader notion of *reliable uncertainty quantification*
- That's where the connection to imprecise prob is most natural

Likelihood

- Recall: X^n consists of iid samples from P_Θ
- Unknown Θ in $\mathbb{T}$, generic θ
- Fisher was the first to formalize this and establish it's fundamental role in statistical theory
- Roughly, the likelihood function ranks the parameter values in terms of how well they fit the data X^n

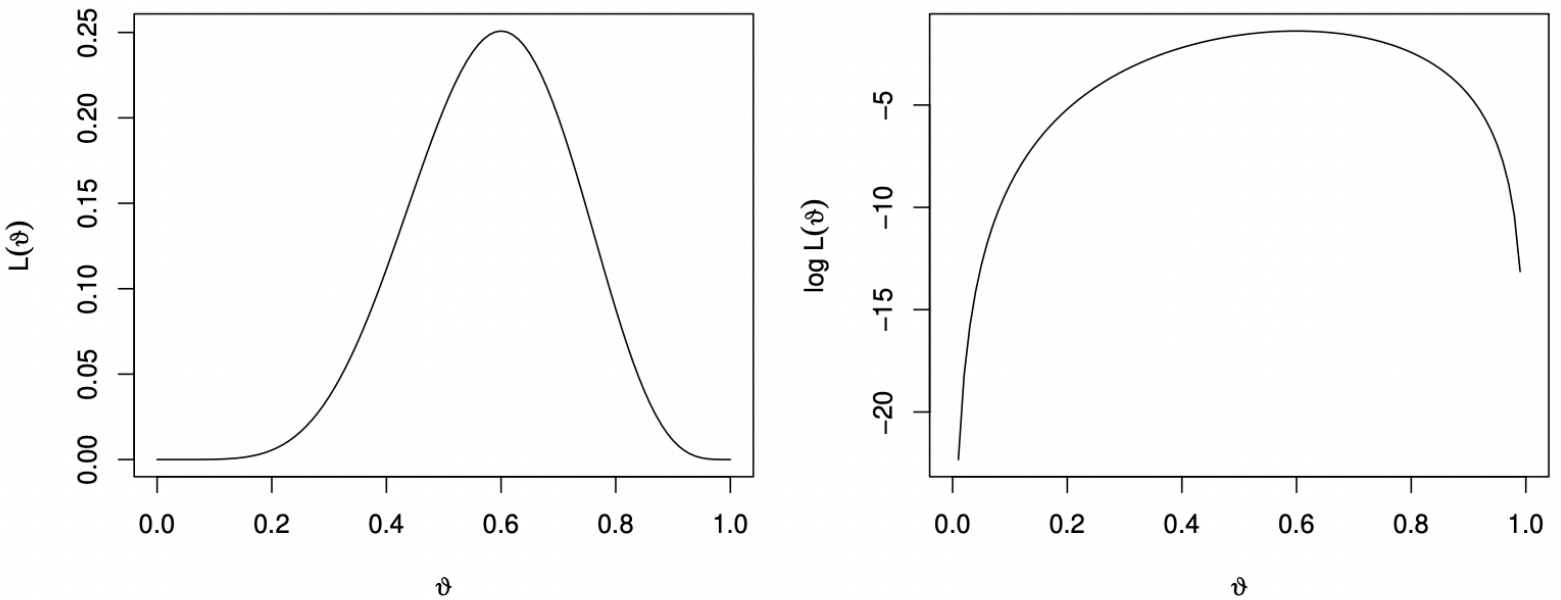
Definition.

The *likelihood function* for Θ is

$$L_{X^n}(\theta) = p_\theta(X^n) = \prod_{i=1}^{n} p_\theta(X_i), \quad \theta \in \mathbb{T}$$

The log-likelihood is $\ell_{X^n}(\theta) = \log L_n(\theta)$

Likelihood, cont.



(a) Likelihood

(b) Log-likelihood

Figure 4.1: Likelihood and log-likelihood functions based on binomial model with observed $X = 6$ and $n = 10$; the corresponding sample proportion is $\hat{\theta} = 0.6$.

Likelihood, cont.

- Recall: likelihood ranks the parameter values
- Natural to seek the "highest ranked" parameter value, which corresponds to the value that best fits the data

Definition.

A *maximum likelihood estimator* (MLE) of Θ, if it exists, is

$$\hat{\theta}_{\mathrm{MLE}} \equiv \hat{\theta}_{X^n} \in \arg\max_{\theta} L_{X^n}(\theta)$$

Can replace likelihood with log-likelihood in the above display

Likelihood, cont.

- Can work out the MLE by hand in simple examples, but generally requires numerical optimization
- Not having a closed-form expression in general means we can only approximate the sampling distribution in general

Theorem.

Under regularity conditions, $\hat{\theta}_{\mathrm{MLE}} \to \Theta$ in P_Θ-probability as $n \to \infty$

Theorem.

Under further regularity conditions, the MLE satisfies

$$n^{1/2}(\hat{\theta}_{\mathrm{MLE}} - \Theta) \to \mathsf{N}(0, I_\Theta^{-1}) \quad \text{in distribution under } \mathsf{P}_\Theta, \text{ as } n \to \infty,$$

where I_Θ is the *Fisher information*

Likelihood, cont.

- Further, define the *relative likelihood*

$$R(X^n, \theta) = \frac{L_{X^n}(\theta)}{L_{X^n}(\hat{\theta}_{\mathrm{MLE}})} \in [0, 1]$$

- Determines the same ranking as $L(\cdot)$, just rescaled
- $R(X^n, \cdot)$ is a *sufficient statistic*
- So, there's a sense in which this is the "best" ranking

Theorem.

Under regularity conditions, $-2 \log R(X^n, \Theta) \to \mathrm{ChiSq}(D)$ in distribution, under P_Θ, as $n \to \infty$, where $D = \dim \Theta$

Likelihood, cont.

- For set estimation, can get approximate confidence sets via

$$C_\alpha(X^n) = \{\theta : -2 \log R(X^n, \theta) \leq \chi^2_{D,1-\alpha}\}$$

- That is, as $n \to \infty$

$$\sup_\theta \mathsf{P}_\theta\{C_\alpha(X^n) \not\ni \theta\} \to \alpha$$

- For $H : \Theta = \theta$, define a *likelihood ratio test*

$$\delta_\alpha(X^n) = \begin{cases} 1 & -2 \log R(X^n, \theta) \text{ is big} \\ 0 & \text{otherwise} \end{cases}$$

- Set "big" to be greater than $\chi^2_{D,1-\alpha}$ and we get

$$\mathsf{P}_\theta\{\delta_\alpha(X^n) = 1\} \to \alpha$$

Uncertainty quantification

- Likelihood-based methods achieve the logical goal, at least approximately, as $n \to \infty$
- But these don't (directly) quantify uncertainty:
  - e.g., consider a hypothesis "$\phi(\theta) \in B$"
  - formal test returns a binary yes/no or accept/reject answer
- Not fully satisfactory, no probability-like assessments
- We should and can do better...

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

Bayesian methods

- Bayesians: "Θ is unknown" $\rightarrow$ quantified with probability
- Amounts to treating unknown-Θ like a random variable, even though it's not actually a random variable
- Prior distribution: $\Theta \sim \mathrm{Q}$, $\textsc{pdf}/\textsc{pmf}$ q
- Bayesians quantify uncertainty about Θ given $X^n = x^n$ using the *posterior distribution* Q_{x^n}
- $\textsc{pdf}/\textsc{pmf}$ of Q_{x^n} given by Bayes's formula:

$$q_{x^n}(\theta) \propto L_{x^n}(\theta)\, q(\theta)$$

- Inference about Θ is (based on) the posterior Q_{x^n}

Bayesian methods, cont.

- Bayesian inference has pros and cons
- These details are mostly beyond the present scope
- An important case in the stat literature[3] is where *prior information about* Θ *is vacuous*
- That is, we apparently don't know anything about Θ and, hence, have no clue about Q to use
- Begs for imprecise-probabilistic considerations... more later

[3] Efron's justification: "scientists want to work on new problems"

Bayesian methods, cont.

- Most common approach is to use a *default prior* Q
- Basically, this amounts to taking flat $q(\theta) \propto 1$[4]
- Justified by *principle of indifference* (Laplace & Keynes)
- Get the posterior Q_{x^n} exactly as before
- Easy to do, and it "works" in the sense described below
- But you're right to be skeptical of this:
  - "Θ unknown" is epistemic uncertainty
  - "$\Theta \sim \mathrm{Q}$" is aleatory uncertainty
  - more generally, "uniformly distributed" $\neq$ ignorant

[4] Jeffreys suggested $q(\theta) \propto (\det I_\theta)^{1/2}$, which is "flat" in a certain sense

Bayesian methods, cont.

- Above suggests a "free lunch"
  - nothing more than likelihood (= model + data)
  - but apparently now we get probabilistic UQ for free
- "No free lunch," so in what sense is this UQ meaningful?
- Preliminary question: how might Q_{X^n} be used?
- For any $H \subseteq \mathbb{T}$,
  - if $\mathrm{Q}_{X^n}(H)$ is large, then infer "$\Theta \in H$"
  - if $\mathrm{Q}_{X^n}(H)$ is small, then infer "$\Theta \notin H$"
- Roughly, a minimal requirement is that, for most/all H,
  - if $H \ni \Theta$, then $\mathrm{Q}_{X^n}(H)$ tends to be large
  - if $H \not\ni \Theta$, then $\mathrm{Q}_{X^n}(H)$ tends to be small
- Related to my "$\textsc{ans}(X^n)$ is wrong" logic

Bayesian methods, cont.

- Bayesian asymptotics provides some positive results
- More complicated, but there are clear parallels with MLEs
- Roughly:
  - false H get vanishing posterior probability
  - Q_{X^n} is approximately $\mathrm{N}(\hat{\theta}_{X^n}, \{nI_\Theta^{-1})$
  - "covariance matching" implies Bayesian credible sets are approximate confidence sets

Theorem — posterior consistency.

Under conditions, $\mathrm{Q}_{X^n}(H) \to 1$ in P_Θ-probability as $n \to \infty$ for all $H \ni \Theta$

Theorem — Bernstein–von Mises.

If $Z_n = n^{1/2}(\Theta - \hat{\theta}_{X^n})$ is a function of $\Theta \sim \mathrm{Q}_{X^n}$, given X^n, then

$$d_{\mathrm{TV}}\{\mathscr{L}(Z_n \mid X^n), \mathrm{N}(0, I_\Theta^{-1})\} \to 0 \quad \text{in } \mathrm{P}_\Theta\text{-probability as } n \to \infty$$

Bayesian methods, cont.

- Large-n results are nice, but limited:
  - consistency takes H fixed and sends $n \to \infty$
  - BvM approximates $\mathsf{Q}_{X^n}(H)$ as $n \to \infty$
- However, UQ involves varying H for fixed n...
- It turns out[5] that, for any n, there exists (many) H's such that $\mathsf{Q}_{X^n}(H)$ tends to be misleading
- Define the *false confidence rate*

$$\text{FCR}_{\mathsf{Q}}(\alpha, H) = \sup_{\theta \notin H} \mathsf{P}_\theta\{\underbrace{\mathsf{Q}_{X^n}(H) > 1-\alpha}_{\text{confident } H \text{ is true}}\}$$

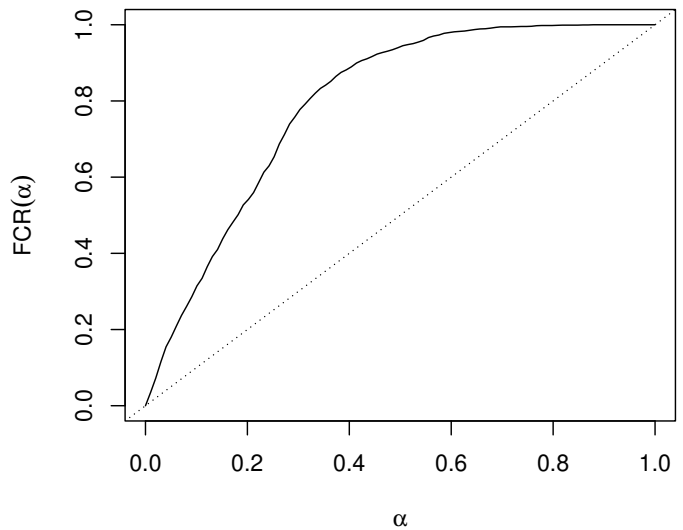
False confidence theorem.

Let Q_{X^n} be *any* data-dependent probability for Θ. For any (α, τ), there exists hypotheses $H \subset \mathbb{T}$ such that $\text{FCR}_{\mathsf{Q}}(\alpha, H) > \tau$.

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

Bayesian methods, cont.

- Simple linear regression: $(X_i \mid u_i) \overset{\text{ind}}{\sim} \mathrm{N}(\beta_0 + \beta_1 u_i, \sigma^2)$
- Covariates are fixed, model parameter $\theta = (\beta_0, \beta_1, \sigma^2)$
- Q_x = Bayes posterior w/ diffuse normal–igamma prior
- Simulation setup:
  - $H = \{(\beta_0, \beta_1, \sigma^2) : -\beta_0/\beta_1 > -1\}$
  - data generated with $\Theta = (0.3, 0.1, 1)$, so H is false



Bayesian methods, cont.

- Two-player perspective on false confidence
  - statistician carrying out (probabilistic) UQ
  - scrutinizer who can bet against the statistician
- Condition "$\mathrm{FCR}_{\mathsf{Q}}(\alpha, H) > \alpha$" implies
  - there exists gambles that are desirable to the statistician, i.e., non-negative Q_x-expected payoff, for each x
  - but are also undesirable to the statistician, i.e., have negative payoff, on average over x for some Θ
- Reveals a type of *incoherence* in probabilistic UQ...
- For details, see `arXiv:2312.14912`

Bayesian methods, cont.

- Examples of false confidence are easy to find
- Applies to *all probabilistic UQ*, not just default-prior Bayes
- This means the UQ's additivity is the culprit
- Therefore, probabilistic UQ is unreliable in this sense
- Aside:
  - Which H are afflicted with false confidence?
  - Current best guess: caused by *non-linearity* in hypotheses
  - For details, see `arXiv:2404.16228`

Part I conclusion

- For specific tasks (estimation, testing, etc) and vacuous prior info, likelihood and Bayes generally work well, for large n
- But reliable UQ requires more and both fall short of this
- In particular, Bayesian/probabilistic UQ is unreliable in the sense of being afflicted with false confidence
- Therefore, to avoid false confidence, and to achieve reliability, *imprecision is imperative*

Part I conclusion, 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" was just that he didn't find the correct mathematical formulation

> *There is still more to be learned from Fisher... than from any other contributor to statistical thinking.* —Dawid

Part II

Outline

- inferential models (IMs)
- possibilism: why & how?
- possibilistic IM construction
- properties
- computation
- examples
- extensions
- ...

Inferential models

- For reliable UQ, *imprecision is imperative*
- So, use data-driven imprecise (lower + upper) probabilities

Definition — *inferential model* (IM).

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

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

- x-dependent measures of support for and plausibility of H

How can UQ be "wrong"?

- For the broad-form UQ that I have in mind, we're interested in lower & upper prob assignments for all H
- Conjugacy: $\text{Ш}_x(H) = 1 - \Pi_x(H^c)$
- So, focus on upper probability
- Upper probability is used to refute hypotheses, i.e.,

$$\text{if } \Pi_x(H) \text{ is small, then doubt that "}\Theta \in H\text{" is true}$$

- Then the upper probability is "wrong" if

$$\Pi_x(H) \text{ is small for some } H \text{ with } H \ni \Theta$$

Preventing UQ from being "wrong"

- Inductive logic: trust the answer if it's a demonstrably low-probability event that it can be wrong
- Leads to a notion of (strong) validity of the IM

Definition — (strong) validity.

The IM $x \mapsto (\sqcup_x, \Pi_x)$ is (strongly) valid if

$$\sup_\theta \mathsf{P}_\theta\{\underbrace{\Pi_X(H) \leq \alpha \text{ for some } H \text{ with } H \ni \theta}_{\text{"answer is wrong"}}\} \leq \alpha$$

Consequences of (strong) validity

Proposition.

For a given $H \subseteq \mathbb{T}$, define the test

$$\delta_\alpha(x) = 1\{\Pi_x(H) \leq \alpha\}$$

Then the test controls Type I error at level α, i.e.,

$$\sup_{\theta \in H} \mathsf{P}_\theta\{\delta_\alpha(X) = 1\} \leq \alpha$$

Proposition.

Define the set estimator

$$C_\alpha(x) = \bigcap \{A \subseteq \mathbb{T} : \amalg_x(A) \geq 1 - \alpha\}$$

Then C_α is a $100(1-\alpha)\%$ confidence set in the sense that

$$\sup_{\theta} \mathsf{P}_\theta\{C_\alpha(X) \not\ni \theta\} \leq \alpha$$

IMs that achieve (stong) validity?

- What mathematical form should the IM take?
- Wide range of imprecise prob options:
  - belief functions
  - consonant belief functions = possibility measures
  - credal sets
  - lower/upper previsions
  - ...
- My focus: *possibility measures* (or *consonant beliefs*)
- I'll justify this later; for now just a quote...

> *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

Achieving (strong) validity, cont.

- Generalized Bayes IM:
  - $(X \mid \Theta = \theta)$ determined by model
  - vacuous prior for Θ
  - combine/condition...
- Fact:[6] generalized Bayes IM is vacuous, i.e.,

$$\text{Ш}_x(H) = 0 \quad \text{and} \quad \Pi_x(H) = 1, \quad H \notin \{\varnothing, \mathbb{T}\}$$

- Consequently, gen Bayes IM is trivially (strongly) valid
- Can we do better? *Yes!*

[6] e.g., Kyburg 1987; Walley 1991; Gong & Meng 2021

Achieving (strong) validity, cont.

- Remember: upper probabilities are for refutation
- Smaller upper probs can refute more than larger upper probs
- So, subject to strong validity, smaller is more informative
- Under strong validity, IMs that are *maxitive*, i.e., possibility measures, are most informative

Proposition (M., `arXiv:2211.14567`.

For any strongly valid IM with upper prob Γ_x, there exists a strongly valid *possibilistic* IM Π_x such that

$$\Pi_x(H) \leq \Gamma_x(H) \quad \text{for all } H$$

Possibility theory primer

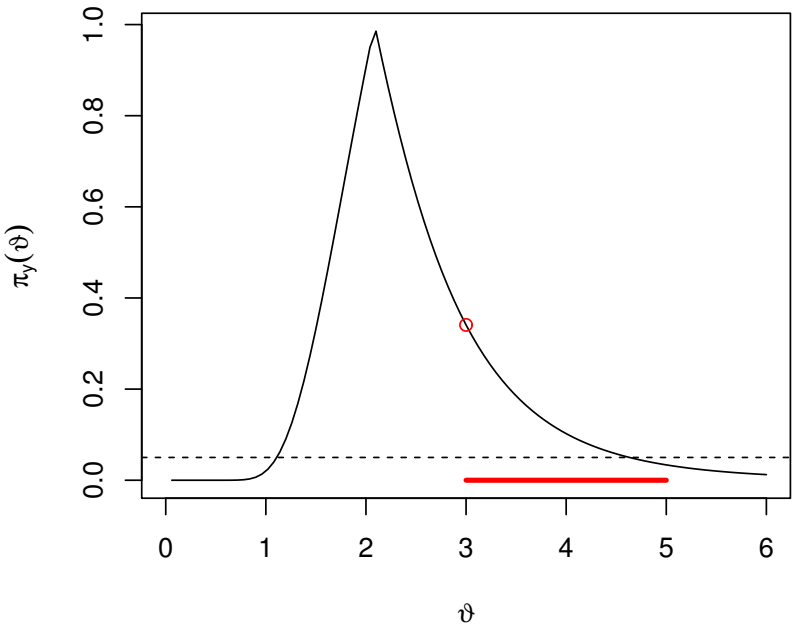
- *Possibility theory*[7] is the simplest imprecise probability
- Similar to probability theory
  - determined by a "density function"
  - different calculus — optimization replaces integration
- Specifically:
  - possibility contour $\theta \mapsto \pi_x(\theta)$ with $\sup_\theta \pi_x(\theta) = 1$
  - and then define
$$\Pi_x(H) = \sup_{\theta \in H} \pi_x(\theta)$$
- Normalization "$\sup_\theta \pi_x(\theta) = 1$" implies coherence[8]
- Simplification: $C_\alpha(x) = \{\theta : \pi_x(\theta) > \alpha\}$
- *What makes the IM output meaningful is its properties*

[7] e.g., Dubois & Prade's *Possibility Theory*

[8] e.g., Troffaes & De Cooman's *Lower Previsions*

Possibility primer, cont.

- Possibility contour plot
- Hypothesis $H = [3, 5]$
- $\Pi_x(H)$ via optimization
- Level set defines interval of "sufficiently possible" θs



Constructing a possibilistic IM

Proposition (Cella & M., *IJAR* 2023).

For an IM $x \mapsto (\sqcup_x, \Pi_x)$, (strong) validity is equivalent to

$$\sup_\theta \mathsf{P}_\theta\{\pi_X(\theta) \leq \alpha\} \leq \alpha, \quad \alpha \in [0,1]$$

where $\pi_x(\theta) = \Pi_x(\{\theta\})$ is the corresponding contour.

- Tells us what we need to achieve our goal:
  - normalization/coherence, i.e., $\sup_\theta \pi_x(\theta) = 1$
  - for each θ, then random variable $\pi_X(\theta)$ should be stochastically no smaller than Unif$(0,1)$ as a function of $X \sim \mathsf{P}_\theta$
- P-values!

Construction, cont.

- Special case of a more general construction[9]
- Workhorse: *imprecise-probability-to-possibility transform*[10]
- In the "no-prior" case, this takes a simple form:
  - same relative likelihood $R(x, \theta) = L_x(\theta) / \sup_\vartheta L_x(\vartheta)$
  - define a possibility contour

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

  - normalization of $\pi_x(\cdot)$ implies by that of $R(x, \cdot)$
- Then $\Pi_x(H) = \sup_{\theta \in H} \pi_x(\theta)$, etc.
- Not unfamiliar... p-value!
- Bayes/fiducial connections...

[9] M. `arXiv:2211.14567`

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

Properties

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

Strong validity theorem.

The possibilistic IM constructed above is strongly valid

- Possibilistic IM aligns with my inductive logic
- Implies no false confidence
- Usual frequentist error rate control, as above

Properties, cont.

- There are converses to the above analysis; one is below
- Suggests possibility theory is fundamental to frequentism

Theorem (arXiv:2507.09007).

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

- $\sup_\theta \mathsf{P}_\theta\{C_\alpha(X) \not\ni \phi(\theta)\} \leq \alpha$ for all α
- if $\alpha \leq \beta$, then $C_\beta(x) \subseteq C_\alpha(x)$ for all x

There exists a valid possibilistic IM for Θ with contour π_x such that

$$\phi(\theta) \in C_\alpha(x) \iff \pi_x(\theta) > \alpha \quad \text{for all } (x, \alpha) \in \mathbb{X} \times [0,1].$$

Computation

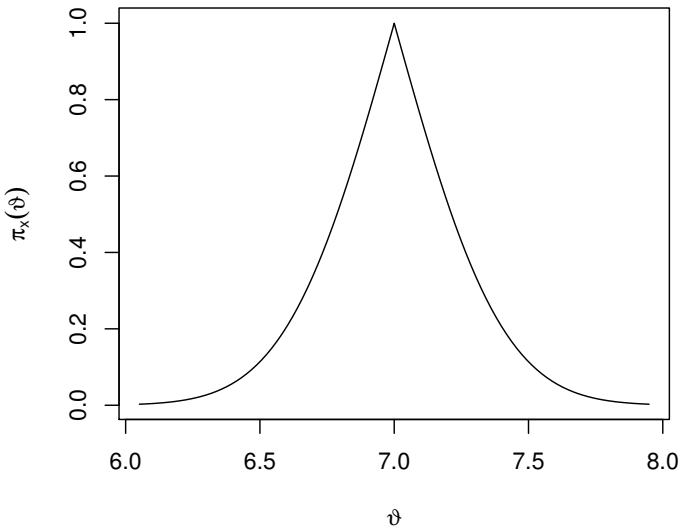
- Need to compute the sampling distribution of $R(X, \theta)$
- Various ways this can be done...
- Almost always requires Monte Carlo
- For each θ on a grid
  - simulate independent $X_\theta^{(m)} \sim \mathrm{P}_\theta$, $m = 1, \ldots, M$
  - approximate the contour by

$$\pi_x(\theta) \approx \frac{1}{M} \sum_{m=1}^{M} 1\{R(X_\theta^{(m)}, \theta) \leq R(x, \theta)\}$$

- Simplifications are possible in specific examples

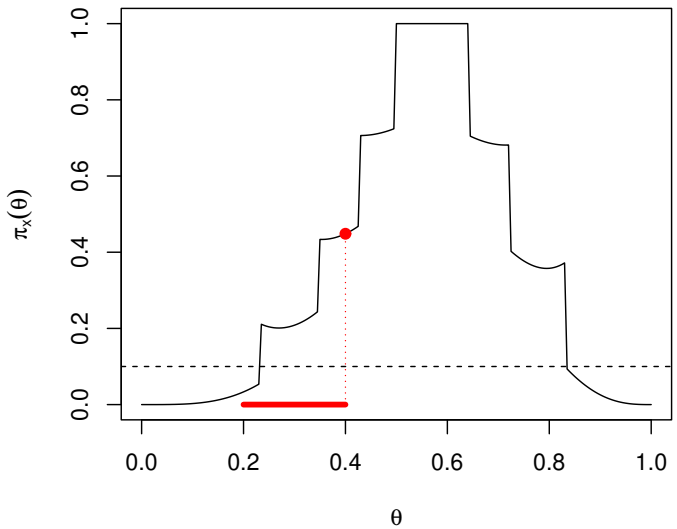
Examples

- $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\Theta, 1)$
- $L(\theta) \propto \exp\{-\frac{n}{2}(\bar{X} - \theta)^2\}$
- $\hat{\theta} = \bar{X}$
- $R(X, \theta) = \exp\{-\frac{n}{2}(\bar{X} - \theta)^2\}$
- Find $\pi_x(\theta)$



Examples, cont.

- $X \sim \text{Bin}(n, \Theta)$
- $L(\theta) \propto \theta^X (1-\theta)^{n-X}$
- $\hat{\theta} = X/n$
- $R(X, \theta) = \left(\frac{n\theta}{X}\right)^X \left(\frac{n-n\theta}{n-X}\right)^{n-X}$
- No simple formula for $\pi_x(\theta)$, but can be evaluated exactly:



$$\pi_x(\theta) = \sum_{y=0}^{n} 1\{R(y,\theta) \le R(x,\theta)\}\, p_\theta(y)$$

Examples, cont.

- $X = (X_1, \ldots, X_n)$ iid $\mathsf{P}_\theta = \mathsf{Exp}(\theta)$
  - likelihood $\theta \mapsto \theta^{-n} \exp(-n\bar{x}/\theta)$
  - relative likelihood $R(x, \theta) = (\theta/\bar{x})^n \exp(-n\bar{x}/\theta)$
- Compute the possibilistic IM contour:

$$\begin{aligned} \pi_x(\theta) &= \mathsf{P}_\theta\{R(X, \theta) \leq R(x, \theta)\} \\ &= \mathsf{P}\{Z^{-n} e^{-Z} \leq (\theta/n\bar{x})^n e^{-n\bar{x}/\theta}\}, \quad \theta > 0 \end{aligned}$$

  where $Z \sim \mathsf{Gamma}(n, 1)$
- Plot from a few slides earlier is based on this

Properties, cont.

- Validity concerns errors like the following:
  - true hypotheses H that I assign small $\Pi_x(H)$
  - true Θ is outside my confidence set
- *Efficiency* concerns other kinds of errors:
  - false hypotheses H that I assign not-small $\Pi_x(H)$
  - false θ values inside my confidence set
- Efficiency is more difficult to describe mathematically[11]
- Roughly, efficiency to me means that π_X *tends* to be more tightly concentrated as a function of X
- Can look at large-sample concentration properties...

[11] Raner (here), `arXiv:2606.30229`

Properties, cont.

Theorem (possibilistic Bernstein–von Mises).

Simplest-to-state version: under regularity conditions,

$$\sup_{z \in \text{compact}} \left| \pi_{X^n}(\hat{\theta}_{X^n} + J_{X^n}{}^{-1/2} z) - \gamma(z) \right| \to 0 \quad \text{in } \mathsf{P}_\Theta\text{-probability}$$

where $\gamma(z) = \mathsf{P}(\|Z\|^2 > \|z\|^2)$, Gaussian possibility contour

- Key points:[12]
  - IM contour looks like a Gaussian contour for large n
  - concentration determined by the observed Fisher information $J_{X^n} \approx n I_\Theta$, hence efficient
- Case with nuisance parameters covered too

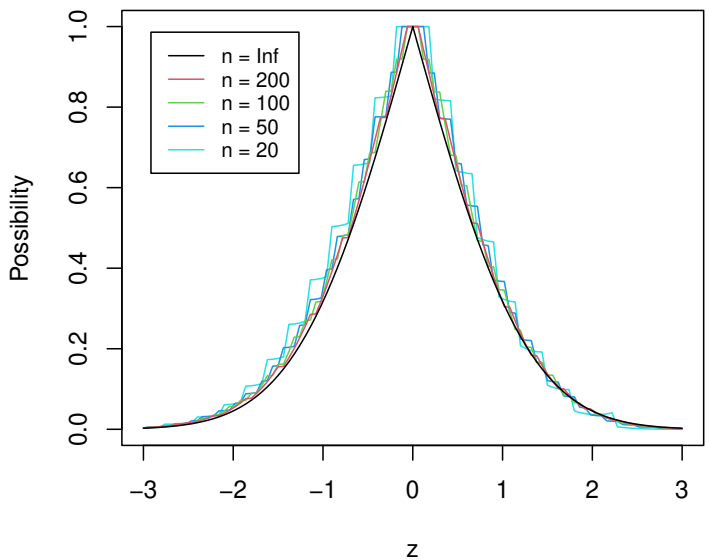
[12] Details in M. and Williams (*IJAR* 2025), `arXiv:2412.15243`

Properties, cont.

- Binomial model from before
- Plot shows contour of the standardized parameter

$$Z = \frac{\sqrt{n}(\Theta - \hat{\theta}_x)}{\sqrt{\hat{\theta}_x(1 - \hat{\theta}_x)}}$$

- Looks Gaussian as n increases



Credal sets

- Π_x corresponds to a set of probabilities
- This (non-empty, closed, convex) set is the *credal set*

$$\mathscr{C}(\Pi_x) = \{\mathsf{Q}_x \in \text{probs}(\mathbb{T}) : \mathsf{Q}_x(\cdot) \leq \Pi_x(\cdot)\}$$

- For possibility measures Π_x, there's a nice characterization:[13]

$$\mathsf{Q}_x \in \mathscr{C}(\Pi_x) \iff \mathsf{Q}_x(\underbrace{\{\theta : \pi_x(\theta) > \alpha\}}_{100(1-\alpha)\% \text{ conf set}}) \geq 1 - \alpha$$

- $\mathscr{C}(\Pi_x)$ consists of (over-)confidence distributions: dist'ns that assign $\geq 1 - \alpha$ probability to $100(1 - \alpha)\%$ confidence sets

[13] e.g., Destercke & Dubois, Ch. 4 of *Intro to IP*

Credal sets, cont.

Theorem (M., `arXiv:2501.10585`).

Each member Q_x of $\mathscr{C}(\Pi_x)$ can be expressed as

$$\mathsf{Q}_x(\cdot) = \int_0^1 \mathsf{K}_x^\beta(\cdot)\, \mathsf{M}(d\beta)$$

- M is a probability stochastically no smaller than $\mathrm{Unif}(0,1)$
- K_x^β is a probability fully supported on $C_\beta(x)$ for each $\beta \in [0,1]$

- *Inner probabilistic approximation* $\mathsf{Q}_x^\star$ of Π_x satisfies

$$\mathsf{Q}_y^\star\{\pi_y(\Theta) > \alpha\} = 1 - \alpha$$

- Often obtain $\mathsf{Q}_x^\star$ via above theorem by taking
  - $\mathsf{M} = \mathrm{Unif}(0,1)$
  - K_x^β fully supported on $\partial C_\beta(x)$ for each $\beta \in [0,1]$

Credal sets, cont.

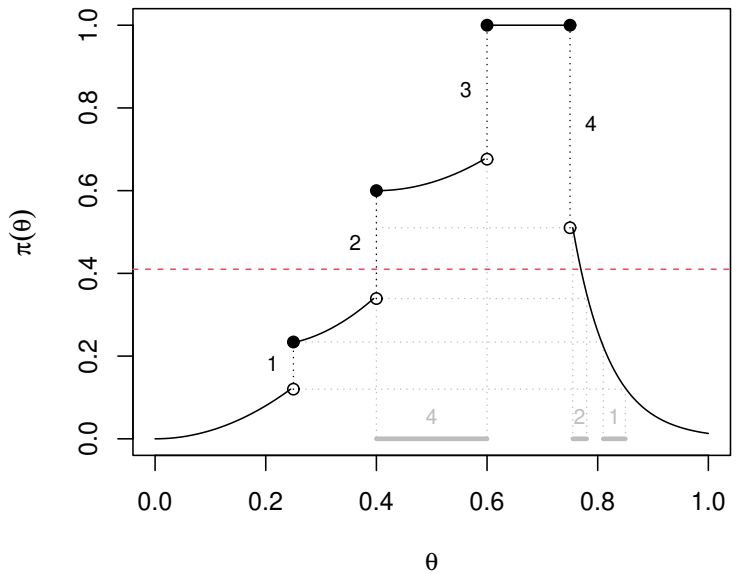
> *Perhaps the most important unresolved problem in statistical inference is the use of Bayes theorem in the absence of prior information.* —Efron

- Inner probabilistic approximation is interesting on its own[14]
- e.g., in problems with group invariant structure, the Bayes posterior with right Haar prior is an inner prob approx
- More generally, appropriate choice of inner prob approx serves as an alternative to Bayes with default priors

[14] M. `arXiv:2503.19748` and forthcoming by Max Raner (here)

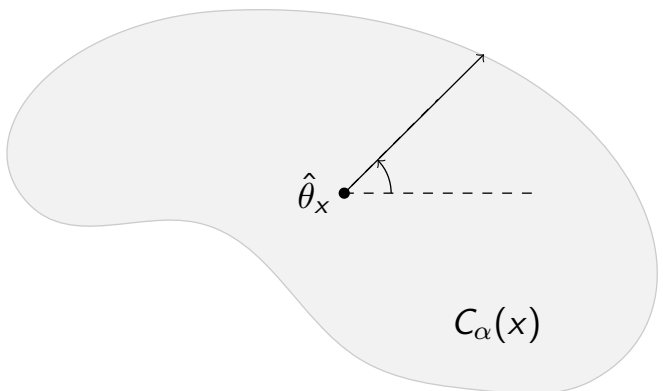
Credal sets, cont.

- Interesting case is where contour has jumps
- e.g., IMs with discrete data
- Inner prob approx will be a mixture of discrete and continuous parts
- Forthcoming work with Sahil Patel (here)



Credal sets, cont.

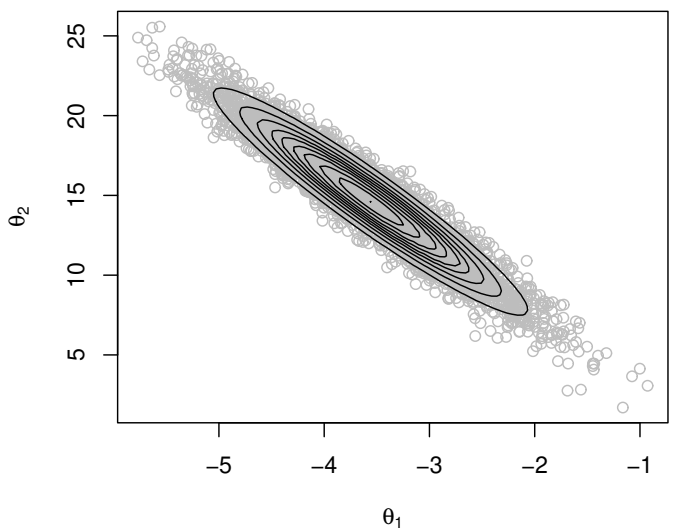
- My initial interest was computational
- Could I use samples from $\mathsf{Q}_x^\star$ to simplify computation of the IM contour π_x in some way?
- Answer is *Yes*, but no time here for details[15]
- Roughly:
  - sample the level α
  - sample a point on $\partial C_\alpha(x)$

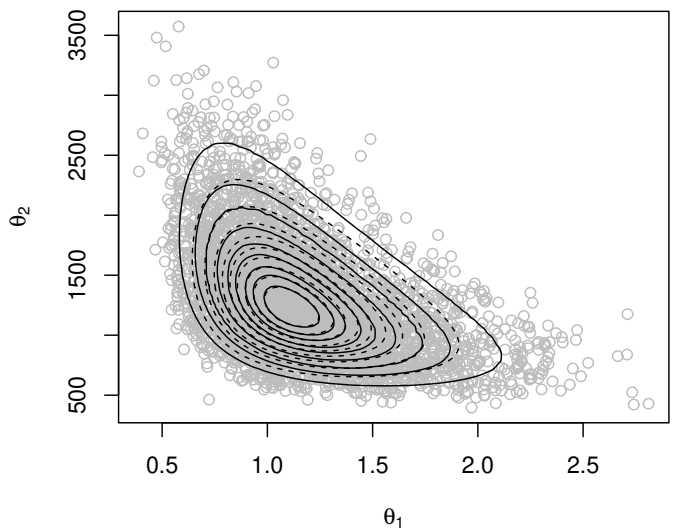


[15] For details, see `arXiv:2501.10585`

Possibilistic IMs, cont.

- Two non-trivial examples
  - logistic regression (left)
  - two-parameter Weibull with right censoring (right)
- Very expensive using the naive computational approach described earlier, but pretty easy with the new approach





Nuisance parameters

- Above, full parameter Θ was the target
- Common for interest to be in some feature $\Phi = f(\Theta)$
- e.g., might only care about the mean and not the variance
- Whatever's in Θ but not Φ is a *nuisance parameter*
- Marginalizing out nuisance parameters reliably is a surprisingly quite difficult problem
- One simple solution is to use possibilistic extension

$$\pi_x^{\mathrm{EX}}(\phi) = \sup_{\theta : f(\theta) = \phi} \pi_x(\theta), \quad \phi \in f(\mathbb{T})$$

- This preserves validity, but is inefficient[16]
- More efficient profiling strategies are available

[16] M. and Williams (*IJAR* 2025) and Raner (`arXiv:2606.30229`)

Extensions

- IM provides reliable UQ for Θ
- Might have other tasks, e.g., decision-making & prediction
- Roughly: *extending* our assessments for certain gambles about Θ to other kinds of gambles
- Workhorse is the *Choquet integral*[17]

$$\Pi_x f \equiv \sup_{Q_x \in \mathscr{C}(\Pi_x)} Q_x f = \int_0^1 \left\{ \sup_{\theta : \pi_x(\theta) > s} f(\theta) \right\} ds, \quad f > 0$$

[17] Troffaes & De Cooman's *Lower Previsions*

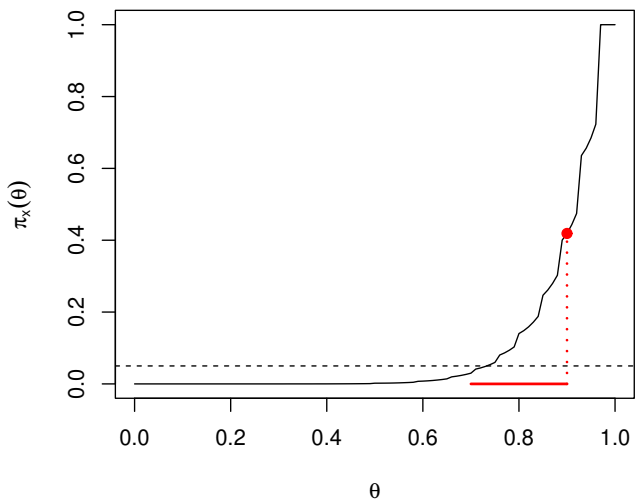
Decision-making

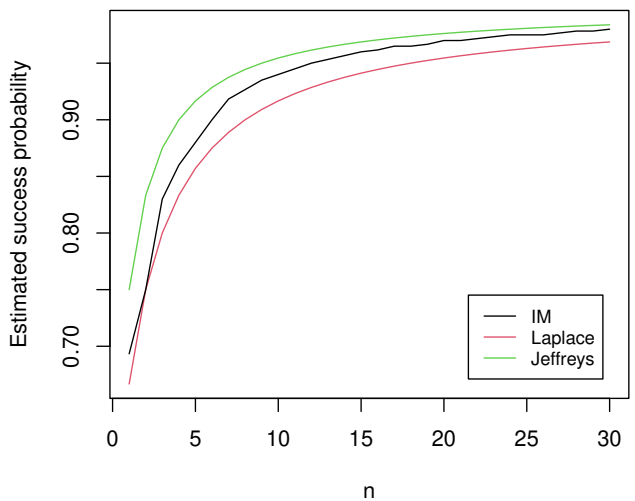
- Loss function $\ell_a : \mathbb{T} \to \mathbb{R}_+$, indexed by actions $\mathbb{A}$
- Can replace loss with utility after obvious adjustments
- Assess the quality of action a wrt Π_x: $a \mapsto \Pi_x \ell_a$
- Very natural choice: $\hat{a}(x) \in \arg\min_a \Pi_x \ell_a$
- Validity-related properties of this assessment...[18]

[18] M., Prim (here), and Williams (*IJAR* 2026), ...

Decision-making, cont.

- Sunrise problem of Hume, Laplace, and others
- Binomial experiment, $X \sim \text{Bin}(n, \Theta)$
- All successes, $X = n$, what can be said about Θ?
- Laplace's famous rule of succession: $\hat{\theta}_{\text{LAP}} = \frac{n}{n+1}$
- IM's risk-minimizing estimator, $\hat{\theta}_{\text{IM}} = \arg\min_a \cdots$





Prediction

- Oracle possibilistic predictor:

$$\pi_x^{\mathrm{OR}}(\tilde{x} \mid \Theta) = \mathsf{P}_\Theta\{p_\Theta(\tilde{X}) \leq p_\Theta(\tilde{x})\}, \quad \tilde{x} \in \mathbb{X}$$

- Average over Θ wrt Π_x via Choquet integration

$$\pi_x^{\mathrm{CH}}(\tilde{x}) = \int_0^1 \left\{ \sup_{\theta : \pi_x(\theta) > s} \pi_x^{\mathrm{OR}}(\tilde{x} \mid \theta) \right\} ds$$

- Promising results so far,[19] no time to discuss details
- This is "parametric" prediction, but "nonparametric" can be handled as well — later

[19] Joint work with James Robertson (here)

Recent references

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2026, VOL. 121, NO. 553, 807-826
https://doi-org.prox.lib.ncsu.edu/10.1080/01621459.2025.2606127

Taylor & Francis
Taylor & Francis Group

Possibilistic Inferential Models: A Review

Ryan Martin

Department of Statistics, North Carolina State University, Raleigh, NC

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
EPUB HEAD-OF-PRINT
https://doi-org.prox.lib.ncsu.edu/10.1080/01621459.2026.2671450

Taylor & Francis
Taylor & Francis Group

An Efficient Monte Carlo Method for Valid Prior-Free Possibilistic Statistical Inference

Ryan Martin

Department of Statistics, North Carolina State University, Raleigh, NC

Beyond the basics

- I've mostly talked about the "basic inference problem"
  - parametric model/likelihood
  - no nuisance parameters
- Most applications don't match this form!
- But the above framework is general, can be tweaked accordingly to address various practical challenges
  - profiling
  - conditioning
- I'll give some details about one important case

Beyond the basics, cont.

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]

More prediction

- Let's be careful about the setup...
  - X^n is observable, interest in X_{n+1} — no covariates
  - assume exchangeable with unknown dist'n P
  - set of probabilities, $\mathscr{P} = \{\text{exchangeable P}\}$
- Let $\rho : \mathbb{X}^n \times \mathbb{X} \to \mathbb{R}$ be a measure of *conformity*
  - e.g., $\rho(x^n, x_{n+1}) = -d(\hat{x}_{x^n}, x_{n+1})$
  - assume ρ is invariant to permutations of its first argument
- Important/basic facts from your math-stat course:
  - conditional distribution of data, given a sufficient statistic, doesn't depend on any unknown parameters
  - in our nonparametric, exchangeable setting, the unordered data values $\{x_1, \ldots, x_n, x_{n+1}\}$ form a sufficient statistic

More prediction, cont.

- Condition on the unordered set of values[20] $\{x_1, \ldots, x_n, x_{n+1}\}$

$$\pi_{x^n}(x_{n+1}) = \sup_{P \in \mathscr{P}} P\big[\rho(X^n, X_{n+1}) \leq \rho(x^n, x_{n+1}) \mid \{x^n, z_{n+1}\}\big]$$

$$\text{sufficiency} \to = \frac{1}{(n+1)!} \sum_{\sigma} 1\{\rho(x^{\sigma(1:n)}, x_{\sigma(n+1)}) \leq \rho(x^n, x_{n+1})\}$$

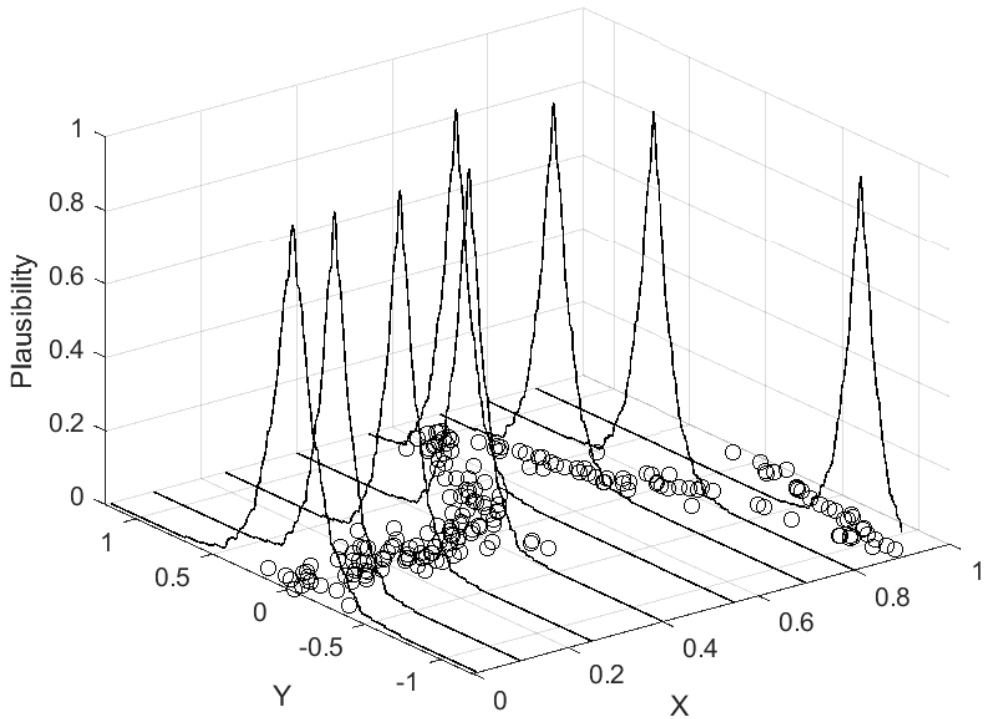
$$\text{invariance} \to = \underbrace{\frac{1}{n+1} \sum_{i=1}^{n+1} 1\{\rho(x_{-i}^{n+1}, x_i) \leq \rho(x^n, x_{n+1})\}}_{\text{conformal prediction ``transducer''}}$$

- Take-aways:
  - IM can be computed
  - strongly valid, fully conditional & coherent possibilistic UQ
  - agrees with conformal prediction in certain ways

[20] Faulkenberry (*JASA* 1973), Hoff (*Bernoulli* 2023)

More prediction, cont.

- For feature–label pairs $Z = (X, Y)$, IM produces possibility contour functions for Y at each $X = x$



Partial prior possibilistic IMs

- There's a spectrum of problems indexed by prior info
- Key observations:
  - to reliably solve real problems, we can't ignore the spectrum
  - spectrum can't be described with ordinary probability
  - applying Bayes to a set of priors is inefficient (dilation)
- These considerations are essential for high-dim problems and cases involving model uncertainty
- Framework above generalizes, but no time to discuss details

Partial prior IMs, cont.

- 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 $L_x(\theta)$
  - imprecise prior via possibility contour $q(\theta)$
- Π_x is a possibility measure given by

$$\Pi_x(H) = \sup_{\theta \in H} \pi_x(\theta), \quad H \subseteq \mathbb{T}$$

  - where the contour function is a Choquet integral:

$$\pi_x(\theta) = \int_0^1 \Big[\sup_{\vartheta : q(\vartheta) > s} \mathsf{P}_\vartheta\{R(X,\vartheta) \leq R(x,\theta)\} \Big] \, ds$$

  - and R is a "relative likelihood"

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

Partial prior IMs, cont.

Strong validity theorem.

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

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

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

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

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

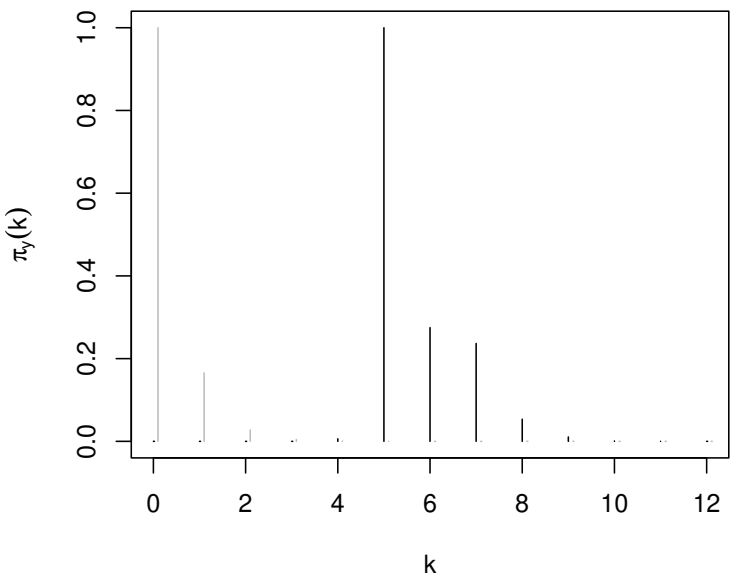
Partial prior IMs, cont.

- 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,[21]
  - treated as a data-dependent update of a partial prior to a "posterior" $(\text{Ш}_x, \Pi_x)$, 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?

[21] M., `arXiv:2211.14567`

Partial prior IMs, cont.

- Interesting application is model assessment in linear regression
- Prior info: most features aren't important, sparsity!
- Can encode this as partial prior info on the model complexity
- Details in M., Singer, and Williams (`arXiv:2511.09305`)



Conclusion

- *Imprecision is imperative for reliable UQ*
- IM aims to address this challenge
- Concrete construction of an imprecise-probabilistic UQ that provably achieves reliability
- Possibility theory seems like the "right" representation
- Computation was/is a challenge, but progress has been made
- Contributions welcome!

Conclusion, cont.

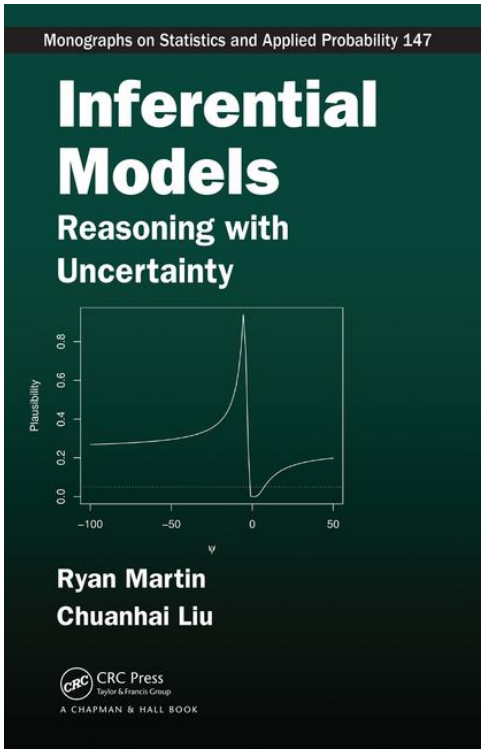
- Semi-open methodological questions:
  - causal inference
  - discrete data problems
  - model assessment
- Open methodological questions:
  - efficient computation?
  - partial prior elicitation?
- Open theoretical questions:
  - IM vs generalized Bayes contraction–dilation
  - models constructed with training data?
  - 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

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



The end — *Thanks!*

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

(New version of the course to be offered in Spring 2027)

`martinrg@purdue.edu`