ST790 — Fall 2022
Imprecise-Probabilistic Foundations of Statistics

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

Week 01b

Previous lecture

- High-level intro and motivation
- My focus was/is on connections to classical statistics
- I realize that my focus might seem out-dated
- Ideas/methods aren't confined to "old school" problems
  - imprecise prob is quite common in CS/AI/etc
  - ongoing work[1] on, e.g., deep learning w/ imprecise prob
  - even I'm pushing in this direction, e.g., *conformal prediction*[2]
- So, again, there are lots of exciting opportunities

[1] Check out recent issues of SIPTA/BFAS conference proceedings or *IJAR*

[2] Cella & M., links to refs on website

Today's lecture

- Background on (precise) probability
  - interpretations
  - De Finetti's coherence arguments
- Shortcomings, i.e., gaps that imprecision can fill
- Imprecise probabilities — *capacities*
- Some basic properties
- Coherence, revisited

Precise probability

- I assume you're familiar with Kolmogorov's axioms and the associated probability calculus covered in standard texts
- My "review" here will focus on some different aspects
  - interpretation
  - rationale, i.e., de Finetti-style coherence
- Not typically covered in our probability courses
- It's important for us to understand these details in order to transition from precise to imprecise

Precise probability, cont.

- Textbooks usually only briefly mention two interpretations:
  - *frequentist*
  - *subjective*
- "Brief" because textbooks are focused on the probability calculus, which doesn't depend on the interpretation[3]
- But don't let the brevity fool you, questions about the interpretation of probability are important[4]
- These aren't the only interpretations, and finer categorizations are possible, these two are just the most familiar

[3] e.g., even if my P is subjective, I can still simulate realizations from it, do Monte Carlo approximations based on laws of large numbers, etc.

[4] Lots of confusion about "frequentist" vs "Bayesian" statistics stems from misunderstandings about the interpretation of probability

Precise probability, cont.

- *Frequentist* interpretation <u>defines</u> $\mathsf{P}(A)$ as the limiting freq at which event A occurs in an infinite sequence of trials
- This has an air of objectivity, but I don't think it's realistic
- For situations we're interested in, often replications don't make sense, i.e., there's no "sequence of trials"
  - will it rain tomorrow?
  - will my grant proposal get funded?
  - is treatment A better than B?
- A reluctance to accept the frequentist interpretation doesn't make me Bayesian, "anti-frequentist," etc.

Precise probability, cont.

Probability does not exist —Bruno de Finetti

- In STEM, we're taught that *subjective* is a dirty word
- But subjectivity is unavoidable, all probabilities are subjective
- Doesn't mean they're arbitrary or come out of thin air
  - can be based on sound theory, empirical verification, etc.
  - can be a consensus about subjective probabilities
- The point is that I ultimately have to decide on which probabilities describe my degrees of belief
- Accepting that there's nothing inherently objective about precise prob is the first step to appreciating imprecision[5]

[5] In fact, the only way to be "objective" is to be imprecise, to simultaneously consider all of the precise prob's I could choose from

Precise probability, cont.

- Let's pause here for a bit of context...
- Statistical inference:
  - observable X, unknown Θ[6]
  - model for (X, Θ) is a subjective, imprecise prob $(\underline{\mathsf{P}}, \overline{\mathsf{P}})$
  - $\textsc{method}(X)$ answers a particular question about Θ
- Inference based on $X \mapsto \textsc{method}(X)$ shouldn't be wrong with more than a small (subjective) $\overline{\mathsf{P}}$-probability, i.e.,

$$\overline{\mathsf{P}}\{\underbrace{\textsc{method}(X) \text{ gives wrong inference about } \Theta}_{\text{e.g., } (X, \overline{\mathsf{P}}) \mapsto \text{ a set estimator that doesn't contain } \Theta}\} \leq \varepsilon$$

- If my model is sound, then the above warrants inference based on $\textsc{method}(x)$ in individual $X = x$ cases[7]

[6] Upper-case Θ indicates that it's *uncertain*, has an imprecise prior
[7] *Cournot's principle* says, roughly, "small probability events don't happen"

Precise probability, cont.

- A convenient consequence of the "limiting frequency" definition is that the mathematical form of $P(\cdot)$ drops out almost automatically
- e.g., (finite-)additivity holds by definition
- But if P is subjective, then where does the mathematical structure come from?
- De Finetti addressed this problem by introducing ideas of internal rationality, or coherence
- Interprets (subjective) probabilities in a behavioral way, as prices you're willing to pay for well-defined gambles

Precise probability, cont.

- De Finetti's formulation:
  - for each event A, $\Pr(A)$ is the price I believe is *fair* for a gamble that pays \$1 if A happens and \$0 otherwise
  - I agree to buy or sell tickets[8] at my stated prices
  - may be multiple transactions, net winnings calculated
- My pricing scheme is *coherent* if there is no finite collection of transactions that guarantees my winnings are < 0, sure loss
- If I can be made a sure loser, then there's something fundamentally wrong with my pricing scheme

Coherence theorem.

A pricing scheme is coherent iff Pr is a (finitely-additive) probability.

[8] ticket $=$ promissory note

Precise probability, cont.

- Proof of "only if" (by contraposition[9])
  - Clearly, setting $\Pr(A) > 1$ or $\Pr(A) < 0$ is dumb
  - Suppose, for some $A \cap B = \varnothing$ and some $\varepsilon > 0$, I set

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \varepsilon < \Pr(A) + \Pr(B)$$

  - Your strategy: buy a ticket for $A \cup B$ from me, and sell me a ticket for A and a ticket for B
    - after these transactions, I have $\$(-\varepsilon)$
    - all outcomes (A, B, or $A^c \cap B^c$) give us \$0 net winnings
    - so I'm guaranteed to lose $\$\varepsilon$
- Proof of "if" ....

[9]Prove "If Pr isn't a finitely-additive prob, then it's not coherent"

Precise probability, cont.

- This is a pretty compelling argument for choosing our "subjective probabilities" to be finitely-additive probabilities
- Finitely-additive probabilities aren't very nice though[10]
- Kolmogorov didn't give strong justification for countable additivity,[11] but this extra structure simplifies things a lot
- Advantages:
  - countably-additive $\Longrightarrow$ finitely-additive $\Longleftrightarrow$ coherent
  - do everything with mass/density functions
  - simple numerical approximations, e.g., Monte Carlo
  - it's familiar

[10] Only one "standard" example of a finitely- but not countably-additive probability — see homework

[11] Basically, Kolmogorov said countable additivity is "convenient"

Towards imprecision

- There are some disadvantages to precise probability
- Rarely mentioned in probability texts, for obvious reasons
- Some shortcomings:
  1. precise prob's can't model ignorance
  2. can't distinguish aleatory & epistemic uncertainty
  3. elicitation of precise prob's is impossible
  4. precise prob's are afflicted by *false confidence*
- I find these easiest to explain/discuss in the context of Bayesian statistical inference...

Towards imprecision, cont.

- Too-quick summary of Bayesian statistical inference:
  - Specify a joint distribution for (X, Θ) via

$$(X \mid \Theta = \theta) \sim \mathsf{P}_\theta \quad \text{and} \quad \Theta \sim \Pi$$

  - Use observed $X = x$ to update the prior Π to a posterior distribution Π_x via Bayes's formula, e.g.,[12]

$$\pi_x(\theta) = \frac{p_\theta(x)\,\pi(\theta)}{\int p_\vartheta(x)\,\pi(\vartheta)\,d\vartheta}, \quad \theta \in \mathbb{T}$$

  - Inferences about Θ are drawn using relevant features of Π_x
- Powerful framework, lots of desirable properties
- Most common criticism: *where does the prior Π come from?*

[12] Assumes Π has a density π; similar formula with mass functions

Towards imprecision, cont.

- Often one is *ignorant* about Θ *a priori*
- Efron: "Scientists like to work on new problems"
- A flat prior models *indifference*, not *ignorance*
- More sophisticated attempts (e.g., Jeffreys) to develop *default* priors for Bayesian inference
- These maneuvers ultimately run into trouble because

a precise probability can't model ignorance!

- Proof.....
- First, what does *ignorant* mean?

Towards imprecision, cont.

- Precise prob can't distinguish aleatory/epistemic uncertainty
  - I take a diffuse $N(0, 100)$ prior because I'm *unsure*
  - you take the same $N(0, 100)$ prior because you're *sure*
  - same posterior, but they can't possibly mean the same thing
- Impossible to elicit precise probabilities:
  - if the statistician is ignorant about Θ, makes sense to talk to an expert who isn't ignorant
  - elicitation of a prior boils down to asking experts some questions about what they expect Θ to be
  - this can give at most a finite collection of constraints, not enough to determine a precise prior

Towards imprecision, cont.

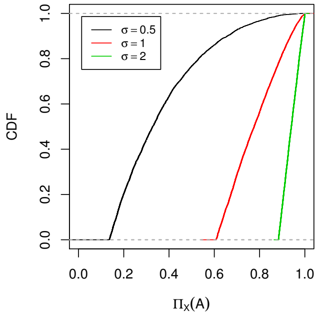
False confidence theorem.

Let Π_X be any data-dependent probability on $\mathbb{T}$. For any (α, β), there exists $A \subset \mathbb{T}$ such that

$$A \not\ni \theta \quad \text{and} \quad \mathsf{P}_\theta\{\Pi_X(A) > \beta\} > \alpha.$$

(Balch, M., & Ferson 2019, `arXiv:1706.08565`)

- Satellite collision example
  - $A = \{\text{non-collision}\}$
  - then $\Pi_X(A)$ as a random variable, with a CDF $\longrightarrow$
  - truth: *on collision course*
  - different noise levels, σ
- *False confidence:* $\Pi_X(A)$ is almost always large!



Towards imprecision, cont.

- Identified some issues with the use of precise prob's[13]
- Take-away: precise probability doesn't do all the things we might want it to do
- The *false confidence* issue is new and of a different nature
    - risk of systematic errors when using precise prob's for UQ
    - practical vs philosophical: false conf demonstrates a sense in which precise prob "doesn't work"
- So, it's worth exploring what imprecise prob's can offer[14]

[13] Not the only issues, e.g., a group of individuals generally won't have a consensus on their degrees of belief

[14] For example, Walley's framework for statistical inference is, roughly, Bayesian inference with imprecise probabilities

- *What is an imprecise probability?*
- Mathematically, a probability is just a function with certain properties, so let's just define a more general function
- A *capacity*[15] on $\mathbb{X}$ is a map $\gamma : 2^{\mathbb{X}} \to [0, 1]$ that satisfies
  - $\gamma(\varnothing) = 0$
  - $\gamma(\mathbb{X}) = 1$
  - $A \subseteq B$ implies $\gamma(A) \leq \gamma(B)$, i.e., monotonicity
- Clearly, probabilities are capacities, but not conversely
- Given γ, define its *dual* or *conjugate* as

$$\tilde{\gamma}(A) = 1 - \gamma(A^c), \quad A \subseteq \mathbb{X}$$

- Probabilities are self-conjugate but, in general, $\tilde{\gamma} \neq \gamma$

[15] First studied by Choquet, 1950s

Imprecise probability, cont.

- A capacity is called *super-additive* if

$$\gamma(A \cup B) \geq \gamma(A) + \gamma(B), \quad \text{all } A \cap B = \varnothing$$

- *Sub-additive* if the inequality is reversed
- A capacity is *2-monotone* if

$$\gamma(A \cup B) + \gamma(A \cap B) \geq \gamma(A) + \gamma(B), \quad \text{all } A,\ B$$

- *2-alternating* if the inequality is reversed
- Clearly, 2-monotone $\implies$ super-additive
- Simple properties:
  - if γ is super-additive, then $\gamma(A) \leq \tilde{\gamma}(A)$ for all A
  - if γ is 2-monotone, then $\tilde{\gamma}$ is 2-alternating

Imprecise probability, cont.

- 2-monotone capacities appear in various contexts:
  - game theory (Shapley)
  - decision theory (Gilboa & Schmidler[16])
  - robust statistics (Huber & Strassen; Kadane & Wasserman)
  - ...
- This is the most basic kind of imprecise probability, for reasons described below
- All the imprecise prob models we consider are 2-monotone
- In fact, they have much more regularity,[17] 2-monotone capacities are too complex

[16] Generalizations to the von Neumann & Morganstern theory

[17] Higher-order monotonicity, etc.

Imprecise probability, cont.

- There's an obvious issue we need to settle right away
- De Finetti: *only probabilities are coherent*
- If we switch to something more general, then we're at risk of some internal irrationality, right?
- But De Finetti makes a strong assumption, easy to overlook
  - $\rightarrow$ *For every gamble, I can precisely specify my fair price and I commit to buy/sell at that price*
- A weaker, more realistic assumption:
  - specify a **max** price at which I'm willing to **buy**
  - specify a **min** price at which I'm willing to **sell**
- "Lower/upper prices" $\rightarrow$ 2-monotone capacity and its dual

Imprecise probability, cont.

- Weaker requirement on the gambler creates more flexibility, an opportunity for other things to be coherent
- Now a pricing scheme sets lower and upper prices

$$\underline{\mathrm{Pr}} = \text{max price to buy} \qquad \overline{\mathrm{Pr}} = \text{min price to sell}$$

- A pricing scheme *avoids sure loss*[18] if there is no finite collection of transactions that ensures winnings < 0

"No-sure-loss theorem."

A pricing scheme avoids sure loss if $\underline{\mathrm{Pr}}$ is a 2-monotone capacity and $\overline{\mathrm{Pr}}$ is its dual

[18] For precise probabilities, coherence $\equiv$ avoids sure loss; but for imprecise probabilities, coherence $\gg$ avoids sure loss

Imprecise probability, cont.

- For a capacity γ, define the credal set

$$\mathscr{C}(\gamma) : \{\mathrm{P} : \mathrm{P}(A) \geq \gamma(A) \text{ for all } A\},$$

  the set of probabilities that *dominate* γ
- Theorem is a consequence of the following two facts:
  - if γ is 2-monotone, then $\mathscr{C}(\gamma) \neq \varnothing$
  - if $\mathscr{C}(\underline{\mathrm{Pr}}) \neq \varnothing$, then pricing scheme avoids sure loss
- Direct proof of $\mathscr{C}(\gamma) \neq \varnothing$:[19]
  - constructing P with $\mathrm{P} \geq \gamma$
  - homework

[19] e.g., Chateauneuf & Jaffray, 1989

Next lecture

- Random sets
- Properties of the induced capacities
- Examples
- ...