Imprecise probability and valid statistical inference

Ryan Martin[1,2]
North Carolina State University
Researchers.One

Van Dantzig Seminar
Netherlands (via Zoom)
April 30th, 2021

[1]Research partially supported by NSF DMS–1811802.

[2]www4.stat.ncsu.edu/~rgmarti3

- Statistics aims to give reliable/valid uncertainty quantification about unknowns based on data, models, etc.
- Two dominant schools of thought:
  - *frequentist*
  - *Bayesian*
- Dichotomy is bad: creates confusion, masks opportunities
- There's clear a spectrum:
  - the more I'm willing to assume, the more precise I can be
  - the less I'm willing to assume, the less precise I can be
- "Unification" isn't enough — spectrum provides a practical opportunity to interpolate between the extremes.

Intro, cont.

- Need a UQ framework that contains the two extremes.
- An advantage of Bayes: (*precise*) probabilities.
- Idea: *imprecise probability*
- Definitely captures Bayes, what about frequentist?
- Meaningful connection can be made if certain (math & stat) constraints are imposed on the imprecise probs.
- Constraints are needed so the imprecise probs respect the frequentists' desire for reliability

This talk

- Problem setup
- UQ via inferential models (IMs), imprecise prob
- Define what validity/reliability means
- Main results:
  1. possibility/consonant plausibility is an imprecise probability structure compatible with validity
  2. position the "frequentist approach" on the imprecise prob spectrum via a sort of converse to Result 1
- How to construct a valid IM?
- Remarks, open questions, etc.

Statistical inference problem

- Observable data: Y in a sample space $\mathbb{Y}$.
- Statistical model: $\mathscr{P} = \{\mathsf{P}_\vartheta : \vartheta \in \Theta\}$.
- θ is the unknown true value; ϑ is a generic value.
- One end of the spectrum: *no prior information*.
- Goal is to learn from data, y, about θ:
  - uncertainty quantification
  - think about a data-dependent "distribution"

Inferential models

Definition.

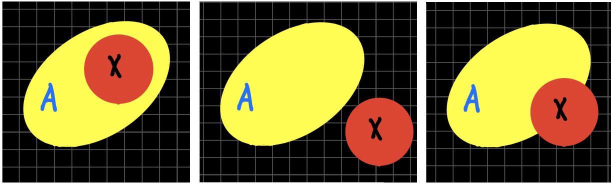
An *inferential model* (IM) is a function whose input is $(y, \mathscr{P}, \ldots)$ and whose output is a capacity[a] $\underline{\Pi}_y : 2^\Theta \to [0,1]$ such that,

$$\underline{\Pi}_y(A) = \text{degree of belief in "}\theta \in A\text{," given } y, \quad A \subseteq \Theta$$

[a] monotone, continuous set function

- The "..." allows for other inputs, e.g., prior information.
- Examples include: Bayes, fiducial, Dempster–Shafer, ...
- Note: the capacity could be *non-additive*.

IMs, cont.



(a) $\mathcal{X} \subseteq A$ (b) $\mathcal{X} \subseteq A^c$ (c) otherwise

- Illustrates shows non-additivity induced by a *random set* $\mathcal{X}$
  - if $\underline{\Pi}(A) := \mathsf{P}(\mathcal{X} \subseteq A)$,
  - then $\underline{\Pi}(A) + \underline{\Pi}(A^c) \leq 1 \quad \leftarrow$ non-additive!
- Under non-additivity, the *dual* to $\underline{\Pi}_y$ is $\overline{\Pi}_y(A) = 1 - \underline{\Pi}_y(A^c)$
  - Fact: $\underline{\Pi}_y(A) \leq \overline{\Pi}_y(A)$
  - $\overline{\Pi}_y(A)$ measures the plausibility of A.

IMs, cont.

- IM output is a pair $(\underline{\Pi}_y, \overline{\Pi}_y)$, lower/upper probs.
- Just like probabilities, these are personal.
- Behavioral interpretation is via gambling:
  - $\underline{\Pi}_y(A)$ is my max buying price for $\$1\{\theta \in A\}$
  - $\overline{\Pi}_y(A)$ is my min selling price for $\$1\{\theta \in A\}$
- Coherence properties:
  - constraints on $A \mapsto \underline{\Pi}_y(A)$ protect me from sure loss
  - my internal sanity check
- More is needed to make my beliefs real-world relevant.
- "More" $=$ statistical constraints.

IMs, cont.

- What kind of statistical constraints?
- *Basic principle:* if $\underline{\Pi}_Y(A)$ is large, infer A.
- Reid & Cox:

> *it is unacceptable if a procedure... of representing uncertain knowledge would, if used repeatedly, give systematically misleading conclusions.*

- We don't want, e.g., $\underline{\Pi}_Y(A)$ to be large if A is false.
- Idea: require that $y \mapsto \underline{\Pi}_y(\cdot)$ satisfy

$$\{\theta \notin A \text{ and } Y \sim \mathsf{P}_\theta\} \implies \underline{\Pi}_Y(A) \text{ tends to be small.}$$

Valid IMs

Definition.

A no-prior IM $(y, \mathscr{P}) \mapsto \underline{\Pi}_y$ is *valid* if

$$\sup_{\theta \notin A} \mathsf{P}_\theta\{\underline{\Pi}_Y(A) > 1 - \alpha\} \leq \alpha, \quad \text{for all } A \subseteq \Theta,\ \alpha \in [0,1]$$

- Validity controls the frequency at which the IM assigns relatively high beliefs to false assertions.
- There's an equivalent statement in terms of $\overline{\Pi}_y$:

$$\sup_{\theta \in A} \mathsf{P}_\theta\{\overline{\Pi}_Y(A) \leq \alpha\} \leq \alpha, \quad \text{for all } A \subseteq \Theta,\ \alpha \in [0,1].$$

- "For all A" is important, e.g., for marginalization.

Valid IMs, cont.

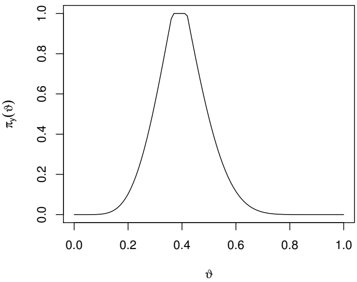
Theorem.

If $(\underline{\Pi}_Y, \overline{\Pi}_Y)$ are valid, then derived procedures control error rates:

- "reject $H_0 : \theta \in A$ if $\overline{\Pi}_y(A) \leq \alpha$" is a size α test,
- the $100(1-\alpha)\%$ plausibility region $\{\vartheta : \overline{\Pi}_y(\{\vartheta\}) > \alpha\}$ has coverage probability $\geq 1-\alpha$.

- IM validity $\Longrightarrow$ usual frequentist validity
- Connection is mutually beneficial:
  - IMs help with interpretation of frequentist output
  - calibration makes IM's $(\underline{\Pi}_y, \overline{\Pi}_y)$ real-world relevant

- $Y \sim \mathsf{P}_\theta = \mathsf{Bin}(n, \theta)$
- Let $(n, y) = (18, 7)$.
- Not difficult to construct a (simple) valid IM
- `[see appendix]`
- Plot of $\vartheta \mapsto \overline{\Pi}_y(\{\vartheta\})$
- It turns out that:
  - $\overline{\Pi}_y(\{\theta_0\})$ is the CP p-value
  - $\{\vartheta : \overline{\Pi}_y(\{\vartheta\}) > \alpha\}$ is the CP confidence interval



Valid IMs, cont.

- *Fact:* Additivity and (no-prior[3]) validity are incompatible.[4][5]
- That is, additive IMs (e.g., Bayes, fiducial) are not valid.
- What's the issue?
  - "The less I'm willing to assume, the less precise I can be"
  - Probabilities are simply too precise.
- Non-additivity creates imprecision:
  - opportunity to be less committal
  - lower our risk of being wrong (cf. Reid & Cox)
  - validity is possible.
- So what's the "right" kind of imprecision?

[3] Generalizations are possible, see appendix
[4] Balch, M., and Ferson (2019), `arXiv:1706.08565`.
[5] M. (2019), `researchers.one/articles/19.02.00001`

Main results: summary

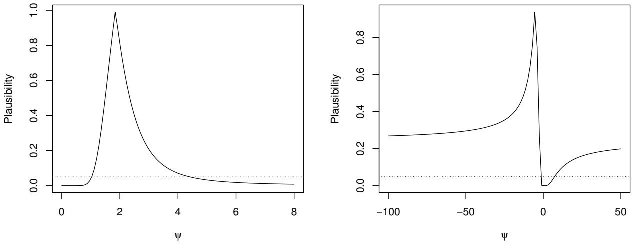
- Recall: I'm sitting at the no-prior end of the IP spectrum
- Goal: position "frequentism" there on the spectrum
- Main results:
  1. *if $\overline{\Pi}_Y$ is a consonant plausibility function, aka a possibility measure, then it can be valid.*
  2. *procedures with frequentist error rate control correspond to valid IMs that are possibility measures.*
- 1 + 2 $\Longrightarrow$ *positioned!!*

Main result 1

- Let $\pi_y : \Theta \to [0,1]$ be a function with $\sup_\vartheta \pi_y(\vartheta) = 1$ and define

$$\overline{\Pi}_y(A) = \sup_{\vartheta \in A} \pi_y(\vartheta), \quad A \subseteq \Theta.$$

- $\overline{\Pi}_y$ is a *consonant plausibility function*, or *possibility measure*
- and π_y is its *plausibility contour*.
- $\underline{\Pi}_y(A) := 1 - \overline{\Pi}_y(A^c) < \overline{\Pi}_y(A) \quad \leftarrow$ non-additive!



Main result 1, cont.

- Frequentist procedures are driven by "p-values"
  - $(y, \vartheta) \mapsto \pi_y(\vartheta)$, small values imply y and ϑ disagree
  - If $Y \sim \mathsf{P}_\theta$, then $\pi_Y(\theta) \sim \mathsf{Unif}(0,1)$.
- "p-value + consonance" $\to$ valid[6]

Theorem.

If the p-value π_y meets the conditions of a plausibility contour, then the consonant IM with plausibility function

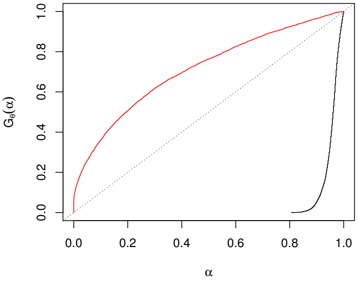
$$\overline{\Pi}_y(A) = \sup_{\vartheta \in A} \pi_y(\vartheta), \quad A \subseteq \Theta,$$

is valid. Moreover, a stronger result with *uniformity in A* holds:

$$\mathsf{P}_\theta\{\overline{\Pi}_Y(A) \leq \alpha \text{ for some } A \ni \theta\} \leq \alpha.$$

[6] M. (2021), `researchers.one/articles/21.01.00002`

- $Y \sim \mathsf{P}_\theta = \mathsf{N}_2(\theta, I)$
- Additive IM: $\Pi_y = \mathsf{N}_2(y, I)$
- Consonant IM:
  - $\pi_y(\vartheta) = 1 - F(\|y - \vartheta\|^2)$
  - $\overline{\Pi}_y$ by supremum
  - $\underline{\Pi}_y$ by duality
- Target is the ratio $\phi = \theta_1/\theta_2$.
- Let $A = \{\vartheta : \phi(\vartheta) \leq 9\}$.
- Suppose $\theta = (1, 0.1)$, so that $\phi = 10$ and A is *false*.
- Then $\Pi_y(A)$ shouldn't be big...



- CDFs of $\Pi_Y(A)$ & $\underline{\Pi}_Y(A)$
  - big → not valid
  - not big → valid

Main result 2

- Combining previous results, we get

$$\{\text{consonant IMs}\} \subseteq \{\text{frequentist}\}.$$

- But to put "frequentist" on the spectrum of valid IMs, I need the opposite inclusion too:

$$\{\text{consonant IMs}\} \supseteq \{\text{frequentist}\}.$$

- Challenge is that the "frequentist approach" has no rules, so the right-hand side contains all sorts of things.
- But I can get basically what I need...

Main result 2, cont.

Theorem.

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

- Coverage. $\inf_\theta \mathsf{P}_\theta\{C_\alpha(Y) \ni \phi(\theta)\} \geq 1 - \alpha$ for all α;
- Nested. if $\alpha \leq \beta$, then $C_\beta \subseteq C_\alpha$;
- Compatible. ....

There exists a *valid & consonant IM* for θ whose derived marginal plausibility region† $C_\alpha^\star(y)$ for ϕ satisfy

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

M. (2021), `researchers.one/articles/21.01.00002`

† $C_\alpha^\star(y) = \{\varphi : \sup_{\vartheta : \phi(\vartheta) = \varphi} \pi_y(\vartheta) > \alpha\}$

Main result 2, cont.

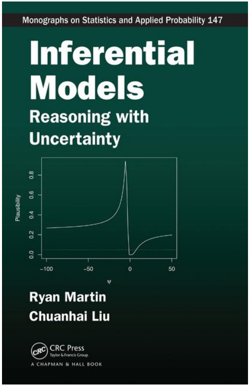
- The compatibility condition is too messy to talk about
  - implies the derived "π_y" has sup $= 1$
  - details in the paper
  - I haven't found a "not-weird" example where it fails
- Analogous result for hypothesis tests.
- Take-away message:
  - anytime you
    - use/teach `[insert favorite method]`
    - or prove a new method controls frequentist error rates,
  - there's a plausibility contour and a valid IM, which you have access to, doing it's thing behind the scenes

Valid IM construction

- Where does a valid IM come from?
- Three strategies for constructing a valid IM:
  - (nested) random sets on an auxiliary space
  - possibility measure on an auxiliary space
  - "Rao–Blackwellization"
- Just a super high-level intro to these...

Valid IM construction, cont.

- $Y = a(\theta, U)$, $U \sim \mathsf{P}_{\text{known}}$.
- If we could observe U, then inference on θ is trivial.
- Idea: use a suitable random set to "guess" the unobserved U.
- Push random set through to a data-dependent possibility measure on θ-space.
- First validity theorems!



Valid IM construction, cont.

- Possibility measures on an auxiliary space.[7]
- Similar to the use of (nested) random sets mentioned above.
- Might be more straightforward in some cases.
- There are some existing "optimality" results in the possibility theory literature.
- Hasn't been carefully explored yet...

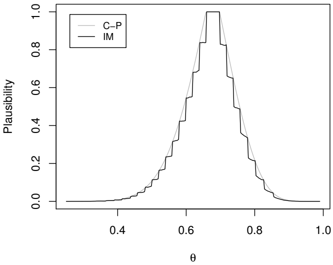
[7] Liu and M. (2020), `researchers.one/articles/20.08.00004`

Valid IM construction, cont.

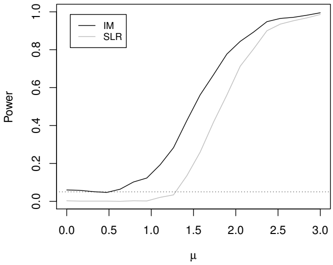
- Proof of Main Result 2 is *constructive*
- That is, the proof gives a sort of "algorithm"
  - input: a decent test/conf region (for ϕ)
  - output: a valid IM (for θ) that's at least as good
- Compare this to the classical Rao–Blackwell result.
- But where does the input come from?
  - we know how to pick inputs in some cases
  - recent development: *universal inference*[8]

[8] Wasserman et al (2020), `arXiv:1912.11436`

- Left: Clopper–Pearson plausibility contour along with that of the "algorithm's" improvement.
- Right: power of the split LR test and along with that of the "algorithm's" improvement.



(d) Binomial plausibility



(e) Power, mixture model

Remarks: interpretation

- Interpretation of inferential output is always non-trivial.
  - Bayes feels easier because probability is familiar.
  - Frequentist felt harder because, e.g., p-values were not well-defined mathematical objects.
- Mathematical facts:
  - p-values measure the *plausibility* of the null hypothesis
  - confidence regions are collections of parameter values that are sufficiently *plausible*
- I teach this interpretation to my students:

  "Your discussion of plausibility in ST503 made it easier for me to explain what a p-value is when I was interviewing. Thank you for the new perspective on the subject!"

Remarks: computation

- Computation with imprecise probs is generally non-trivial, they're more complex than precise probs.
- However, consonance makes them much simpler.
- Computation:
  - evaluate the contour, π_y, may require Monte Carlo
  - do optimization $\overline{\Pi}_y(A) = \sup_{\vartheta \in A} \pi_y(\vartheta)$.
- For standard problems, this is relatively easy.[9][10]
- For more complex, high-dim problems, it's definitely not impossible — interesting intersection of methods?

[9] Joyce Cahoon's PhD thesis and associated papers.

[10] Syring and M. (2021), `arXiv:2103.02659`

Remarks: beyond inference

- Beyond inference, one might be concerned with prediction.
- All of what I said above applies to prediction too.
- Practical challenge:
  - "model-free" prediction is the goal
  - but the above machinery is model-based
- What about a valid *nonparametric* IM?
- Turns out there's a close connection between the imprecise probability stuff here and *conformal prediction*.[11]
- "Conformal + consonance" leads to a powerful and valid nonparametric IM for prediction.

[11] Cella and M. (2021), `researchers.one/articles/20.01.00010`

What's next?

- Model-free inference is an interesting problem.[12]
- Valid IMs for general nonparametric problems?
- What does the "spectrum" I mentioned above look like?
- Imprecise prob is flexible, allows for *partial prior info*.
- High-dim problems are my motivation:
  - structural assumptions, e.g., sparsity, parsimony, etc.
  - this is just partial prior information
- How to use imprecise prob to incorporate this in a way that's both valid and efficient along the spectrum?

[12] Leo Cella's PhD thesis and associated papers.

Conclusion

- We all know statistical reasoning is imprecise.
- Now we can see this in the math:
  - validity is crucial to the logic, and
  - imprecision is necessary for validity.
- *Plausibility/possibility* is especially important:
  - for the theory
  - for the interpretation
  - for the computation
- Hopefully can capture the spectrum from no prior info to complete prior info, with a notion of validity.
- New and exciting territory.

The end!

Thanks for your attention!

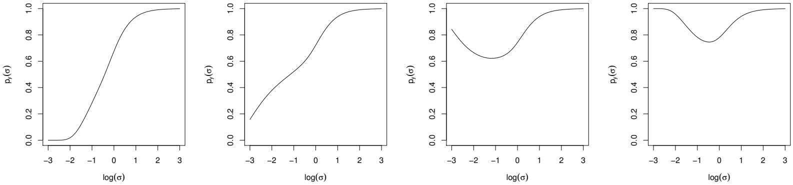
`www4.stat.ncsu.edu/~rgmarti3`
`https://researchers.one`

Appendix

More details, if there's time.

1. False confidence
2. A general definition of validity
3. Details for the binomial example

- Satellite conjunction analysis:
  - Orbiting satellite could collide with another object.
  - To avoid this, analysts compute a *non-collision probability*.[13]
  - Satellite judged to be safe if non-collision probability is high.
  - *Noisier data makes non-collision probability large*[14]



(f) Close (g) Kinda close (h) Kinda far (i) Far

[13] Details in Balch, M., and Ferson (2019), `arXiv:1706.08565`.

[14] M. (2019), `researchers.one/articles/19.02.00001`

1. False confidence, cont.

- *False confidence:* Hypotheses tending to be assigned high probability even if data don't support them.
- Probabilities suffer from false confidence, *not valid*.

False confidence theorem (Balch, M., and Ferson 2019).

Let Π_Y be a probability on Θ, depending on data Y. Then, for any $\alpha \in (0,1)$ and any $\theta \in \Theta$, there exists $A \subset \Theta$ such that

$$A \not\ni \theta \quad \text{and} \quad \mathsf{P}_\theta\{\Pi_Y(A) > 1 - \alpha\} > \alpha.$$

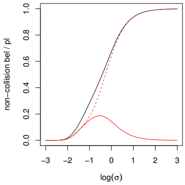
R.ONE

1. False confidence, cont.

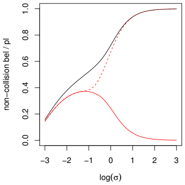
- This result is not too surprising:
  - probabilities are too precise
  - data may not be informative enough to reliably support that level of precision
  - e.g., mean vector vs. a weird function of it
- This is not the consequence of using a "bad prior," etc., it's entirely due to additivity.
- Validity and precision are both strong requirements, too strong to be compatible.
- Imprecision is needed for (this kind of) validity.
- Practical effects of imprecision below.

1. False confidence, cont.

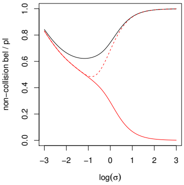
- Satellite collision illustration: $A = \{\text{non-collision}\}$.
  - $\Pi_y(A)$ (black)
  - $\underline{\Pi}_y(A)$ (red) and $\overline{\Pi}_y(A)$ (red, dashed).
- Gap between $\underline{\Pi}_y(A)$ and $\overline{\Pi}_y(A)$ is increasing in σ.



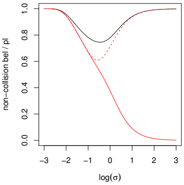
(j) Close



(k) Kinda close



(l) Kinda far



(m) Far

2. General validity

- Work thus far has focused on the no-prior case, very recently I started thinking about the more general case.[15]
- What happens if there's partial/imprecise prior info?
- Notation:
  - Partial prior info encoded in a set $\mathscr{Q}$ of priors Q
  - Model $P_\cdot$ + prior $Q \to$ joint dist $(Y, \theta) \sim M_Q$
  - Upper envelope: $\overline{M}_{\mathscr{Q}}(\cdot) = \sup_{Q \in \mathscr{Q}} M_Q(\cdot)$
- Questions:
  - what does validity mean in this more general case?
  - what kind of IM $(\underline{\Pi}_Y, \overline{\Pi}_Y)$ achieves validity?
  - .....

[15] M. (2021), almost done...

2. General validity, cont.

Definition.

An IM $(y, \mathscr{P}, \mathscr{Q}) \mapsto \underline{\Pi}_y$ is *valid* if

$$\overline{M}_{\mathscr{Q}}\{\overline{\Pi}_Y(A) \leq \alpha, \, \theta \in A\} \leq \alpha, \quad \text{all } (\alpha, A).$$

- Special cases:
  - If $\mathscr{Q} = \{\text{all priors}\}$, then back to previous definition.
  - If $\mathscr{Q} = \{Q\}$, then it's the classical Bayes setup and the posterior is valid (wrt to M_Q).
- Can prove:
  - validity implies "no sure loss"
  - validity implies IM-based procedures "control error rates"
  - generalized Bayes posterior is valid

3. Binomial example

- Following the approach in the IM book.
- Let $Y \sim \mathsf{P}_\theta = \mathsf{Bin}(n, \theta)$, distribution function F_θ.
- Construct a valid IM for θ:

  A. $F_\theta(Y-1) \leq U < F_\theta(Y)$, $U \sim \mathsf{Unif}(0,1)$.

  P. Simple (but not best) "default" random set

$$\mathcal{S} = \{u : |u - 0.5| \leq |U - 0.5|\}, \quad U \sim \mathsf{Unif}(0,1).$$

  C. Combine to get[16]

$$\Theta_y(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \{\theta : F_\theta(y-1) \leq u < F_\theta(y)\}$$
$$= \left[1 - G^{-1}_{n-y+1,y}\left(\tfrac{1}{2} + |U - \tfrac{1}{2}|\right), 1 - G^{-1}_{n-y,y+1}\left(\tfrac{1}{2} - |U - \tfrac{1}{2}|\right)\right],$$

[16] Use fact that $F_\theta(y) = G_{n-y,y+1}(1-\theta)$, beta distribution.

3. Binomial example

- Output: new random set $\Theta_y(\mathcal{S})$.
- Distribution of $\Theta_y(\mathcal{S})$ only depends on $U \sim \text{Unif}(0, 1)$.
- Now use this random set to calculate the IM:

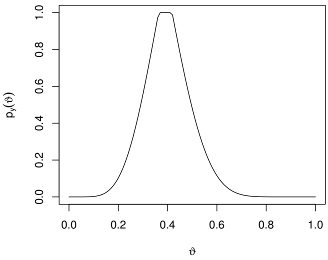
$$\underline{\Pi}_y(A) = \mathsf{P}_{\mathcal{S}}\{\Theta_y(\mathcal{S}) \subseteq A\}$$
$$\overline{\Pi}_y(A) = \mathsf{P}_{\mathcal{S}}\{\Theta_y(\mathcal{S}) \cap A \neq \varnothing\}$$

- There are (messy) closed-form expressions.
- IM is guaranteed valid by construction, e.g.,

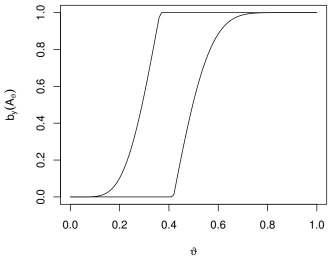
$$\sup_{\theta \in A} \mathsf{P}_\theta\{\overline{\Pi}_Y(A) \leq \alpha\} \leq \quad \forall\, (A, \alpha).$$

3. Binomial example, cont.

- Data: $n = 18$, $y = 7$.
- Plots of plausibility contour $\pi_y(\vartheta)$ and of the lower and upper probabilities of $A_\vartheta = [0, \vartheta]$.



(n) Plausibility contour



(o) lower/upper probabilities