Fisher's underworld and the behavioral–statistical reliability balance in scientific inference

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

SIPTA Seminar
October 18th 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

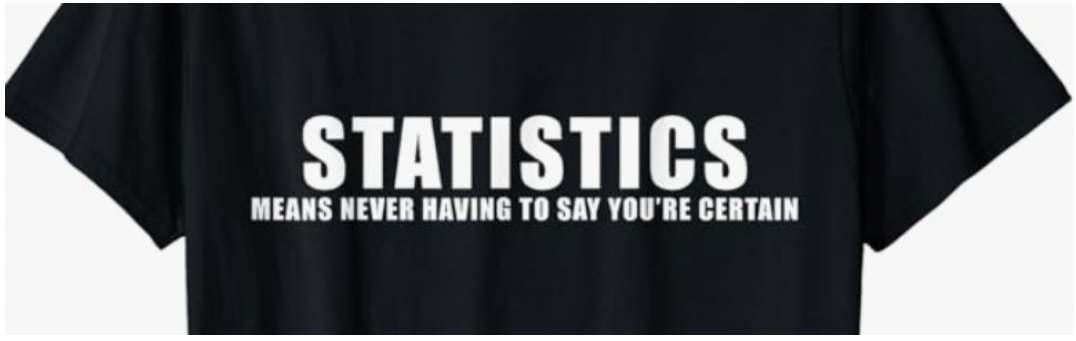
How can a discipline, central to science and to critical thinking, have two methodologies, two logics, two approaches that frequently give substantially different answers to the same problems? —Fraser

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.

- If statistics can't resolve these issues, then who are we?
- With 100+ years of experience, we should be the authorities on uncertainty, but instead we side-step:

  Different measures of uncertainty can complement one another[2]



- Lack of answers $\rightarrow$ confusion and distrust
- Largely no motivation to find answers
- Fraser: *is complacency in the face of contradiction acceptable for a central discipline of science?*

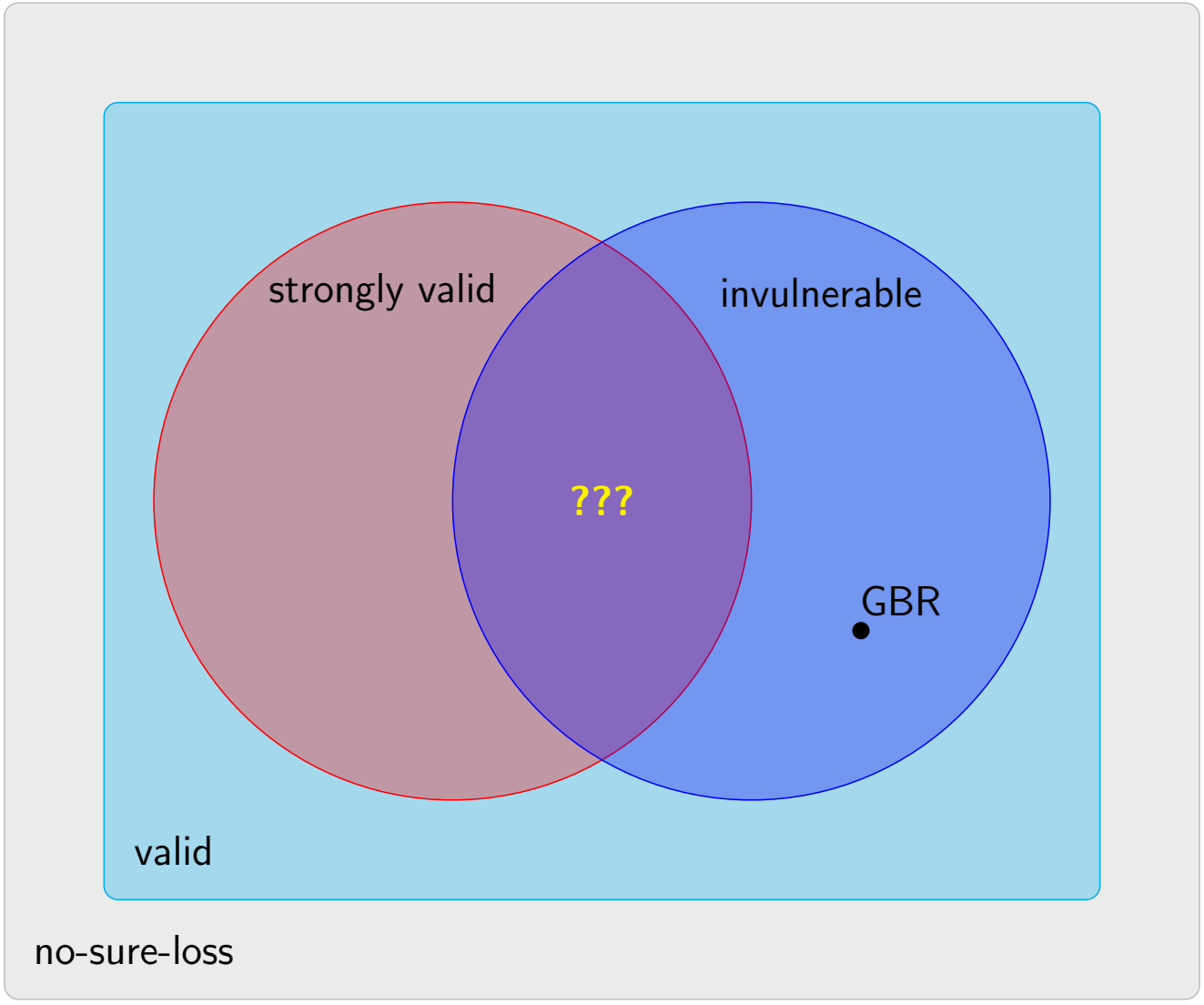
[2] ASA President's Task Force on Statistical Significance statement

Intro, cont.

- Problems boil down roughly to competing priorities:
  - behavioral reliability (Bayesian)
  - statistical reliability (frequentist)
- Apparently neither is fully satisfactory — a compromise or balance is needed to resolve the issues
- But what should this balance look like?
- *My basic claim:*
  - imprecision is inherent,[3] plays an essential role
  - great opportunity for imprecise prob!

[3] Fisher wrote about how p-values, etc determined "no exact probability statements" — what else could he mean if not imprecision?

This talk: picture ≈ 1000 words



This talk, cont.

- Fisher's underworld of probability & new perspectives
- Background — inferential models (IMs)
- "Invulnerability" and generalized Bayes
- Towards balance: validity and "minimal vulnerability"
- Strong validity and possibilistic IMs
- Concluding remarks, etc

- Heads-up:
  - probably more questions than answers here
  - suggestions? collaborations? ...

Fisher's underworld of probability

Fisher (1957) considers tossing a six-sided die, one face is called *Ace*

1. Agent 1 is sure that P(Ace) is $\frac{1}{6}$ and is willing to accept bets against an Ace at, say, odds 4:1
2. Agent 2 doesn't know P(Ace) but he believes that at least 10% of the dice in the box from which Agent 1's die will be randomly chosen have probability $\geq \frac{1}{5}$ of turning up Ace; therefore, Agent 2 would accept odds 9:1 that Agent 1 loses money in the long run
3. Agent 3 "does not know and does not think he knows... the probability of throwing an Ace, but believes he knows how the probability is distributed in a consignment of boxes, one of which has been used to supply the die in question, may know with confidence what odds he can profitably accept in betting against the success of [Agent 2]'s wager" (Fisher)

Fisher's underworld, cont.

- In true Fisher form, this is a cryptic paper![4]
- Basic idea (as I see it):
  - probability has different levels
  - e.g., top level is betting with known/assumed odds
  - deeper levels have different assumptions and objectives
  - in particular, agents at lower levels aren't betting on the actual outcome, they're scrutinizing other agents
- Clear hints toward imprecision...
- My thought:
  - statistician is at one level making inferences
  - "society" is deeper scrutinizing the statistician's methods
  - balance by being invulnerable to scrutiny?

[4] Hampel: one doesn't understand Fisher, one *feels* Fisher

Background

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

- $(\underline{P}_{Y,\Theta}, \overline{P}_{Y,\Theta})$ is an imprecise joint dist for $(Y, \Theta) \in \mathbb{Y} \times \mathbb{T}$
  - coherent lower & upper probabilities/previsions
  - e.g., precise likelihood for $(Y \mid \Theta = \theta)$ and an imprecise prior for Θ, possibly vacuous
- Goal is to quantify uncertainty about Θ, given $Y = y$
- *Inferential model* (IM): $y \mapsto (\underline{\Pi}_y, \overline{\Pi}_y)$
  - coherent for each y
  - duality $\underline{\Pi}_y(H) = 1 - \overline{\Pi}_y(H^c)$
  - probabilistic reasoning
  - tests, confidence sets, etc can be extracted
- Lots of IMs, choice depends on our priorities

Back to the underworld

- Statistician is like Fisher's Agent 2:
  - doesn't know about the world, about Θ
  - has a model, data, etc
  - constructs an IM, i.e., probability statments about Θ
- Scrutinizer ("society") is like Fisher's Agent 3:
  - looks for vulnerability in statistician's IM
  - aims to expose these and capitalize off of them
  - *not about money, it's about accountability*
- Statisticians should want to be invulnerable
- What does it mean for an IM to be *invulnerable*?

Invulnerability

- Underworld: Agent 3 suspects that Agent 2's probability statements are wrong, vulnerable to loss
- WLOG, assume scrutinizer thinks the statistician's support $\underline{\Pi}_Y(H)$ to hypothesis H tends to be too large
- For suitable $\alpha \in [0, 1]$, the scrutinizer offers the statistician the following (IM-dependent) gamble

$$\begin{aligned} W_{H,\alpha} &= W_{H,\alpha}(Y, \Theta) \\ &= \{1(\Theta \in H) - (1 - \alpha)\}\, 1\{\underline{\Pi}_Y(H) \geq 1 - \alpha\} \end{aligned}$$

- For each y it's *acceptable*:[5] $\underline{\Pi}_y\{W_{H,\alpha}(y, \cdot)\} \geq 0$
- Statistician's IM is vulnerable if at least one of those given-y acceptable gambles is unacceptable averaging over Y

[5] acceptable = almost desirable

Invulnerability, cont.

Definition.

Given $(\underline{\mathsf{P}}_{Y,\Theta}, \overline{\mathsf{P}}_{Y,\Theta})$, the statistician's IM is *invulnerable* if the gambles $\{W_{H,\alpha} : H \subseteq \mathbb{T}, \alpha \in [0,1]\}$ are all acceptable, i.e., if

$$\underline{\mathsf{P}}_{Y,\Theta}(W_{H,\alpha}) \geq 0 \quad \text{for all } (H, \alpha);$$

otherwise, the IM is *vulnerable*.

- Similar to coherence, but I don't think it's the same
- Compare to *Kerckhoff's principle* in cryptography:
  - assume attackers are familiar with your encryption system
  - and design it to be secure anyway

Invulnerability, cont.

- Invulnerability might be a strong condition
- Does any IM satisfy it?
- Obvious first guess — the *generalized Bayes* IM?
- That is, $\underline{\Pi}_y(\cdot) = \underline{\mathsf{P}}_{Y,\Theta}(\Theta \in \cdot \mid Y = y)$

Theorem.

The generalized Bayes IM is invulnerable.

- I don't know (yet) if other IMs are invulnerable :(
- New/interesting connection between Fisher & imprecise prob?

Toward balance

> *One difficulty in using generalized Bayes rule is that the updated previsions it defines may be highly imprecise... This suggests that generalized Bayes rule should be regarded as one of many possible strategies for updating beliefs.* —Walley

- Generalized Bayes is conservative, tends to dilate
- While invulnerability, coherence, etc are nice, GB apparently puts too much emphasis on behavioral reliability
- For example:
  - vacuous prior implies vacuous generalized Bayes IM
  - but one can learn something from data alone
  - statisticians would never accept the vacuous solution
- Other considerations...?

Validity

Definition.

Given $(\underline{P}_{Y,\Theta}, \overline{P}_{Y,\Theta})$, the statistician's IM $(\underline{\Pi}_y, \overline{\Pi}_y)$ is *valid* if

$$\overline{P}_{Y,\Theta}\{\underline{\Pi}_Y(H) > 1-\alpha,\ \Theta \notin H\} \leq \alpha \quad \text{for all } (H, \alpha)$$

- Motivated by statistical reliability considerations
  - assigning large belief to a false hypothesis is erroneous
  - validity aims to control probability of errors
- The above definition[6] of validity generalizes that in my earlier work beyond the vacuous prior case
- Behavioral reliability properties...?

[6] M. (2022), `arXiv:2203.06703`

Validity, cont.

Theorem (M., arXiv:2203.06703).

1. generalized Bayes IM is valid
2. validity implies (more than) no-sure-loss

- I didn't expect these behavioral–statistical connections
- Take-away messages:
  - validity isn't a particularly strong condition
  - validity doesn't (and can't) imply coherence
  - it's nestled in between coherence and no-sure-loss
- Potential for balancing distinct priorities?

Validity, cont.

Theorem.

If an IM is invulnerable, then it's valid — i.e., if an IM isn't valid, then there exists (H, α) such that statistician can be duped into accepting $W_{H,\alpha}$ and risk losing money in the "long run"

- Valid = "minimally vulnerable"? i.e.,

$$\text{GBR} \in \{\text{invulnerable}\} \subseteq \{\text{valid}\} \subseteq \{\text{no-sure-loss}\}$$

- New perspective on the *false confidence theorem*[7]
  - original version focused on vacuous prior case
  - described potential statistical unreliability of precise IMs, e.g., default-prior Bayes, fiducial, etc.
  - restatement of the FCT: *wrt vacuous priors, precise IMs are vulnerable because they aren't valid*

[7] Balch, M., and Ferson (2019), *Proc. Roy. Soc. A*, `arXiv:1706.08565`

Towards balance, cont.

- No progress toward balance yet
- i.e., the only valid IM I've mentioned is generalized Bayes
- What other valid IMs are out there?
- This isn't straightforward:
  - my previous work focused exclusively on vacuous priors
  - those ideas applied to this more general context satisfy a stronger *strong validity* property (next)
- Good IMs that are valid but not strongly valid?
  - one idea[8] is to take an IM that's valid wrt a vacuous prior and combine it with the prior using, say, Dempster's rule
  - empirical evidence supports validity, but no proof yet :(

[8] M. (2022), `arXiv:2203.06703`, Sec. 6

Possibilistic IMs

- My instinct: strongest possible statistical reliability
- This led me on an unexpected journey
  - outer consonant approximations
  - imprecise-probability-to-possibility-transforms
  - ...
- Long story short,[9] the new *possibilistic IM* framework
  - incorporates partial prior info
  - is "likelihood-based"
  - is possibility-theoretic $\rightarrow$ relatively simple
  - achieves *strong validity*, etc
  - allows adjustments for improved efficiency
- ISIPTA'23:[10] possibilistic IM's credal sets & fiducial

[9] M. (2022/3), `arXiv:2211.14567`

[10] `https://proceedings.mlr.press/v215/martin23a.html`

Possibilistic IMs, cont.

- For simplicity: precise model with imprecise prior
- $\overline{\Pi}_y$ 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—the workhorse—is

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

  - and R is a "relative likelihood"

$$R(y,\theta) = \frac{p_\theta(y) \cdot \overline{\mathsf{P}}_\Theta(\{\theta\})}{\sup_{\vartheta \in \mathbb{T}} p_\vartheta(y) \cdot \overline{\mathsf{P}}_\Theta(\{\vartheta\})}, \quad (y,\theta) \in \mathbb{Y} \times \mathbb{T}$$

- Vacuous prior: R is the usual likelihood ratio

Possibilistic IMs, cont.

- Contour is (the correct interpretation of) a p-value
- Practicalities:
  - contour is a Choquet integral, not too difficult
  - my naive computational strategies can surely be improved
- Statistical properties...

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

The possibilistic IM is *strongly valid* in the sense that

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

Consequently, if $C_\alpha(y) = \{\theta : \pi_y(\theta) > \alpha\}$, then

$$\overline{\mathsf{P}}_{Y,\Theta}\{C_\alpha(Y) \not\ni \Theta\} \leq \alpha$$

Possibilistic IMs, cont. — tangent 1

- For illustration: *precise model + vacuous prior*
- With GBR, learning is impossible in this case
- Strongly valid IMs can learn!

$$\pi_y(\theta) := \sup_{\vartheta \in \mathbb{T}} \mathsf{P}_{Y|\vartheta}\{R(Y,\vartheta) \leq R(y,\theta)\} \not\equiv 1$$

- $R(Y,\vartheta)$ is often an approximate pivot under $\mathsf{P}_{Y|\vartheta}$, in which case the supremum has little effect
- Also, general principles for an efficiency bump[11]
- In the above case, the efficiency-bumped IM is

$$\pi_y^{\text{bump}}(\theta) := \mathsf{P}_{Y|\theta}\{R(Y,\theta) \leq R(y,\theta)\} \ll \pi_y(\theta)$$

[11] M. (2022/3), `arXiv:2211.14567`, Sec. 5.4

Possibilistic IMs, cont. — tangent 2

- For simplicity: *precise model* + *vacuous prior*
- Strong validity has an interesting consequence:

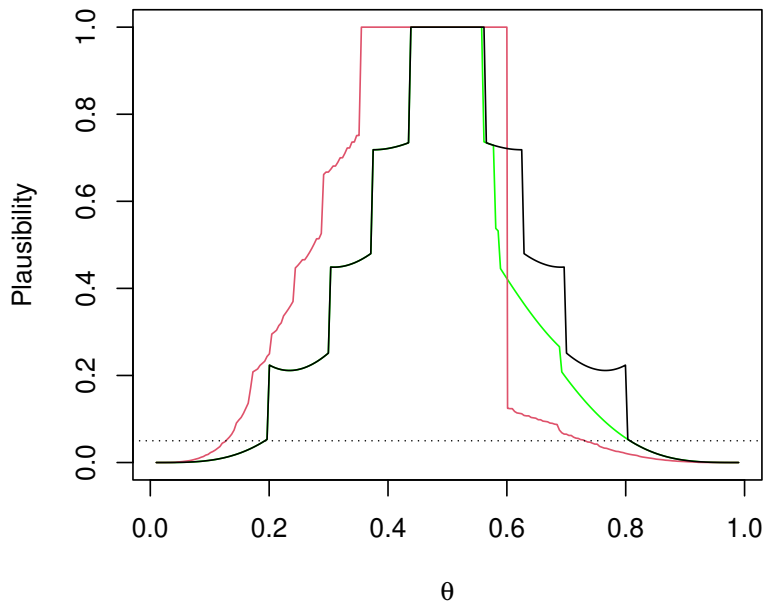
$$\sup_{\theta} \mathsf{P}_{Y|\theta}\{\underline{\Pi}_Y(H) > 1-\alpha \text{ for some } H \not\ni \theta\} \leq \alpha$$

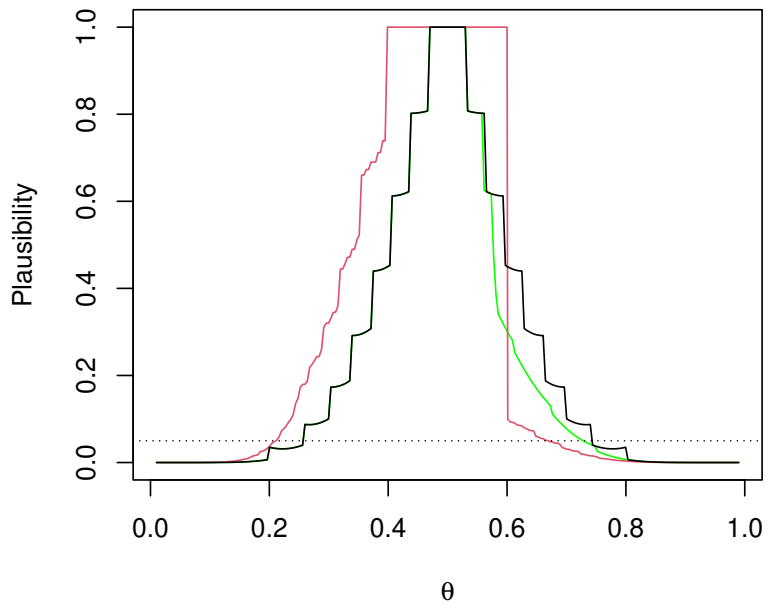
- "Uniformity" in H explains why strong validity $\gg$ validity
- Practical implications:
  - not just controlling error for static null hypotheses
  - uniformity allows for reliable *probing*
- Possibilistic IMs $\gg$ hypothesis testing procedures![12]

[12] Cella and M. (2023), *IJAR*, to appear; `arXiv:2304.05740`

Possibilistic IMs: Example 1

- $(Y \mid \Theta = \theta) \sim \text{Bin}(n, \theta)$
- Partial prior for Θ: "90% sure that $\Theta \leq 0.6$"
- Three different IMs:
  - vacuous prior (with efficiency bump)
  - partial prior
  - partial prior (with efficiency bump)[13]
- Contours: $n = 8$ (left) and $n = 16$ (right), $\hat{\theta} = 0.5$

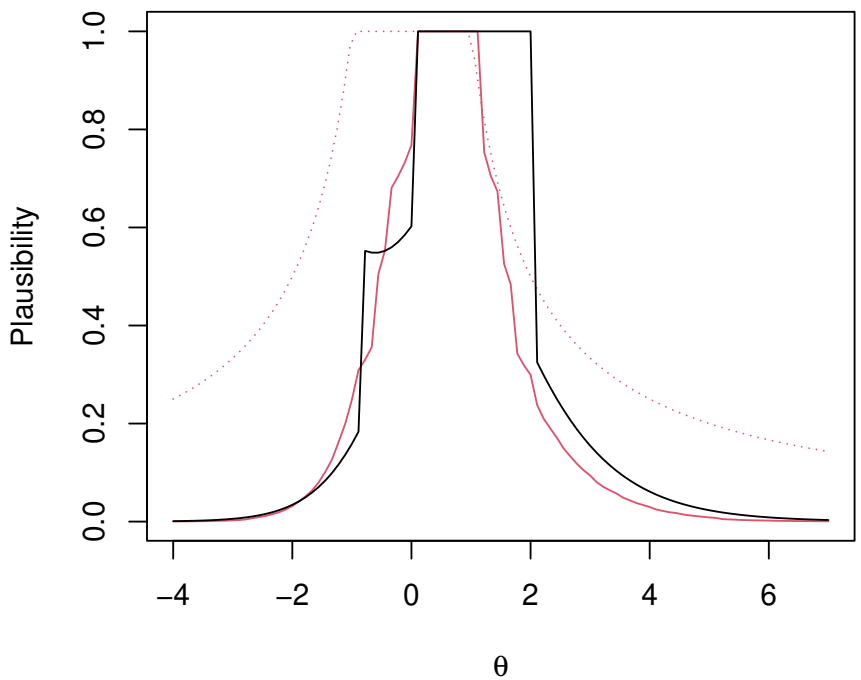




[13] I think partial prior efficiency bump can be improved...

Possibilistic IMs: Example 2

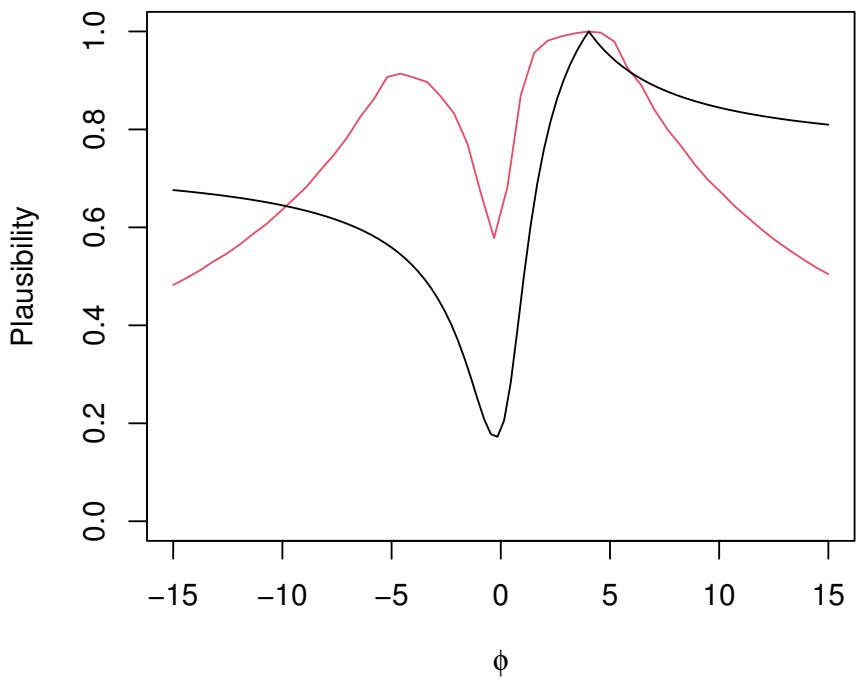
- Two-sample binomial, Θ denotes the log odds ratio
- Data $(n_1, y_1) = (43, 1)$ and $(n_2, y_2) = (39, 2)$
- Nuisance parameters eliminated via conditioning[14]
- Two (marginal) IM contours:
  - vacuous prior (with efficiency bump)
  - partial prior — "$\mathsf{E}|\Theta| \leq 1$"



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

Possibilistic IMs: Example 3

- *Fieller–Creasy:* ratio Φ of two normal means
- *No set estimator has finite length a.s. and coverage prob* > 0[15]
- Two strongly valid marginal IM contours:[16]
  - vacuous prior (with efficiency bump)
  - partial prior — "$\mathrm{E}|\Phi| \leq 5$"



[15] Gleser & Hwang, *Ann. Statist.*, 1987

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

Possibilistic IMs: Example 4

- *Behrens–Fisher:* difference of normal means — "unsolved"
- New marginal IM solution[17] is strongly valid
- Simulation study to check efficiency:
  - difficult unbalanced case, $(n_1, n_2) = (2, 20)$
  - 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**

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

Possibilistic IMs, cont.[18]

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

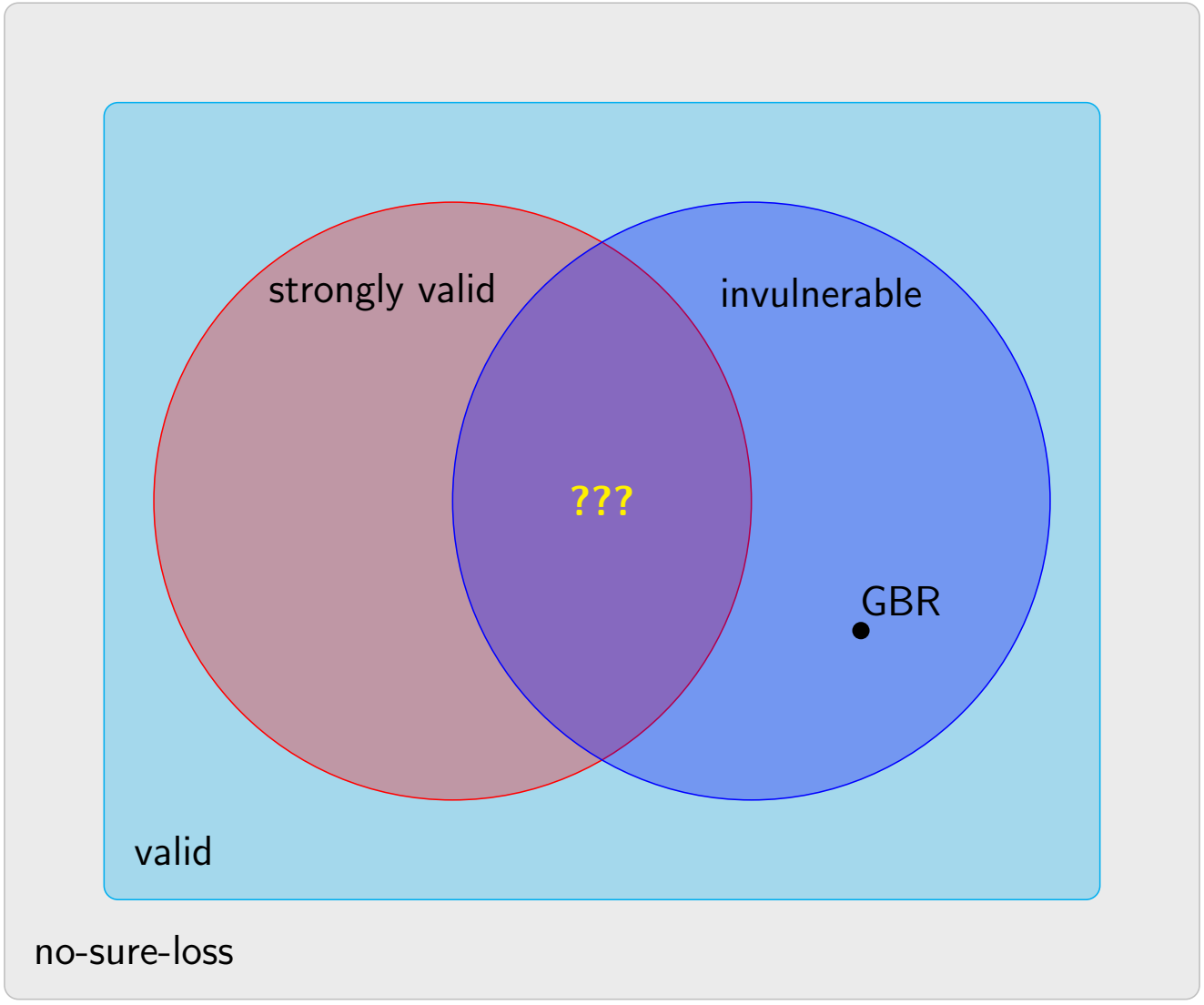
Leonardo Cella [a], Ryan Martin [b]

[18] "Possibilistic" connection drawn in M. (2023), arXiv:2309.13454

Towards balance, cont.

- Back to Fisher and the underworld...
- Where does strong validity fit in?
  - strong validity $\implies$ validity $\implies$ no-sure-loss
  - but GBR isn't strongly valid
  - so invulnerable $\not\implies$ strongly valid
  - and, therefore, connection to invulnerability isn't clear
- *Question:* Are any strongly valid IMs invulnerable?
- I don't know yet... :(

1000+ words later



Conclusion

- Recall the goal:
  - two-theory (B vs F) problem threatens statistics' identity
  - need to balance behavioral & statistical reliability
- Started with Fisher's underworld of probability
- Used this to motivate *invulnerability*
- Showed that
  - generalized Bayes IM is invulnerable
  - invulnerability implies validity (which implies no-sure-loss)
- New connections between Fisher and imprecision
- Still no clear behavioral–statistical reliability balance :(

Conclusion, cont.

- Possibilistic IMs are my go-to solution
  - strong validity
  - all the desired statistical properties
  - opportunities for efficiency gains
  - computation is doable
- Possibilistic IMs have extra structure, so their connection in invulnerability isn't clear yet
- In particular, $\{\text{strongly valid}\} \cap \{\text{invulnerable}\} \neq \varnothing$?
  - if non-empty, then balance is within reach
  - if empty, then I/we have to keep looking

Conclusion, cont.

- Open questions I already mentioned:
  - is generalized Bayes the only invulnerable IM?
  - construct valid but not strongly-valid IMs?
  - improvements to the partial prior efficiency bump?
  - strongly valid IMs that are also invulnerable?
- Other open questions:
  - efficient Choquet integral computation?
  - high-dim cases: computation & elicitation?
  - valid (possibilistic) model assessment/selection?

Conclusion, cont. — tangent

- Statisticians generally don't understand this spectrum
- Classical problems: in some sense it doesn't matter much
- High-dim problems: *regularization matters!*
- Real problems fall in the middle of the spectrum
- Frequentists & Bayesians fumble focusing on the endpoints
- Excellent opportunity for (IP and) partial-prior IMs!

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!