False confidence, non-additive beliefs, and valid statistical inference[1][2]

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

ACMS Department Colloquium
University of Notre Dame
February 8th, 2019

[1] Paper available at https://www.researchers.one/article/2019-02-1
[2] Thanks to friends and collaborators M. Balch, H. Crane, C. Liu...

Introduction

- Statistics has developed a lot in the last 50+ years.
- But important fundamental questions about probability and statistical inference remain unanswered.
- Does it matter? YES!
- Currently we have two dominant schools of thought:
  - *frequentist*
  - *Bayesian*
- Very different views leading to different answers.
- Lots of debate over the years about which one is "right."
- I think it's better to ask:
  - what does science need from statistics?
  - do existing approaches meet this need?
  - if not, then how to fill the void?

Intro, cont.

- Statistical problem:
  - Observable data is Y;
  - Model is $\mathscr{M} = \{\mathsf{P}_{Y|\theta} : \theta \in \Theta\}$ — taken as given.
- Scientific questions correspond to hypotheses about θ.
- Goal is to quantify uncertainty about θ, given $Y = y$.
- Define an *inferential model*

$$(y, \mathscr{M}, \ldots) \mapsto b_y : 2^\Theta \to [0, 1],$$

  where $b_y(A)$ represents the data analyst's degrees of belief about a hypothesis $A \subseteq \Theta$ based on data y, model $\mathscr{M}$, etc.
- The inferential model could be lots of things:
  - a Bayesian posterior distribution;
  - a fiducial or confidence distribution;
  - ...

Intro, cont.

- Pointless if the inferential model isn't "reliable."
- Brad Efron said (roughly) that the construction of reliable, prior-free, inferential models is *the most important unresolved problem in statistical inference*.
- Key insight: *go non-additive!*
    - Familiar things are *additive*, i.e., b_y is a probability.
    - But additivity isn't necessary, might even be a constraint.
    - "There's more to uncertainty than probabilities"[3]
- Take-away messages:
    - additive beliefs are afflicted with false confidence
    - good non-additive beliefs can avoid false confidence
    - there's a way to construct good non-additive beliefs

[3] http://www.isipta2019.ugent.be

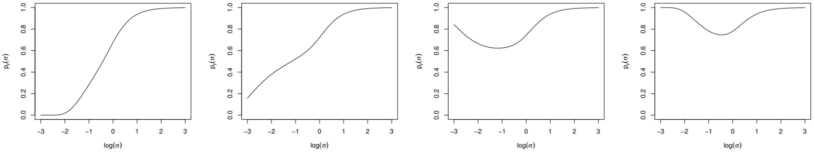
This talk

- The price of additivity:
  - satellite collision example
  - false confidence theorem
- Going non-additive:
  - avoiding false confidence: the validity property
  - non-additive beliefs and random sets
  - construction of valid inferential models
- A few technical remarks:
  - efficiency and dimension reduction
  - a complete-class theorem
- Conclusion

Satellite collision problem

- A satellite orbiting Earth could collide with another object.
- Potential mess, so navigators try to avoid collision.
- Data on position, velocity, etc is used, along with physical and statistical models, to compute a *collision probability*.
- If collision probability is low, then satellite is judged safe; otherwise, some action is taken.
- An unusual phenomenon has been observed: noisier data necessarily makes collision probability small...

- Illustration:
  - $\|y\|$ denotes measured distance between satellite and object.
  - True distance ≤ 1 implies collision.
  - Measurement error variance is σ^2.
  - $p_y(\sigma)$ = probability of non-collision, given y.
- When σ is large, $p_y(\sigma)$ is large, no matter what y is!
- *Potentially misleading conclusions!*



(a) $\|y\|^2 = 0.5$ (b) $\|y\|^2 = 0.9$ (c) $\|y\|^2 = 1.1$ (d) $\|y\|^2 = 1.5$

False confidence

- What's going on here?
- Apparently, data are not sufficiently informative with respect to questions about collision/non-collision.
- *Additivity* forces the probability to go somewhere, happens that it goes to non-collision no matter what data says.
- Even if the satellite is on a direct collision course, probability tells navigators that it's safe.
- *False confidence:*[4] a hypothesis tending to be assigned large probability even though data does not support it.
- Is this a general phenomenon, or just something weird about this particular problem?

[4] Balch, M., and Ferson, `arXiv:1706.08565`

False confidence, cont.

False confidence theorem.

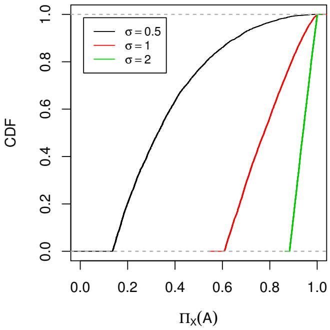
Let Π_Y be any additive belief on Θ, depending on data Y. Then, for any α and any t, there exists $A \subset \Theta$ such that

$$A \not\ni \theta \quad \text{and} \quad \mathsf{P}_{Y|\theta}\{\Pi_Y(A) > t\} \geq \alpha.$$

- In words, there always exists a false hypothesis that tends to be assigned high posterior probability.
- If judgments about the plausibility of a hypothesis are made based on the magnitude of its probability, then the theorem says there's risk of systematic error.
- Reveals a practical price attached to additivity.
- Doesn't say which hypotheses are afflicted, or to what extent, only that they exist.

False confidence, cont.

- Simple version of the satellite example.
  - Let A denote the non-collision hypothesis.
  - Then $\Pi_Y(A)$ as a random variable, with a CDF
  - Plot CDF when data are generated under a collision course.
- *False confidence:* $\Pi_Y(A)$ is almost always large!



Avoiding false confidence

- Theorem says *every additive belief function* is afflicted with false confidence.
- Reid & Cox write:

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

- To avoid false confidence and "systematically misleading conclusions," b_y must be non-additive.
- e.g., $b_y(A) + b_y(A^c) < 1$.
- But not every non-additive belief avoids false confidence.
- So we need some additional restrictions...

Avoiding false confidence, cont.

- First: *Why should we care about false confidence?*
  - Lots of Bayesian success stories, am I'm over-reacting?
  - e.g., have there been satellite collisions?
- The theorem doesn't say that something *will* go wrong, only that there's a risk.
- Statisticians directly involved in the data analysis might be able to manage their risk reasonably well.
- But this level of involvement is rare — we are largely focused on developing methods/software to be used by others.
- To manage risk under this increased exposure, I must either
  - inform users which hypotheses are too dangerous,
  - or avoid false confidence altogether.

Validity property

Definition.

An inferential model $(y, \mathscr{M}, \ldots) \mapsto b_y$ is *valid* if

$$\sup_{\theta \notin A} \mathsf{P}_{Y|\theta}\{b_Y(A) > 1 - \alpha\} \leq \alpha, \quad \forall\, A \subseteq \Theta, \quad \forall\, \alpha \in (0,1).$$

- Validity property implies that assigning high belief to a false hypothesis is a rare event — *no false confidence*.
- α on inside and outside calibrates the belief function values, i.e., so that I know what "large" and "small" means.
- Implies that procedures derived from a valid b_y have frequentist error rate control; details later.

Non-additive beliefs

- Simplest way to construct non-additive beliefs on a space $\mathbb{X}$ is via a *random set* $\mathcal{X} \sim \mathsf{P}_{\mathcal{X}}$ that takes values in $2^{\mathbb{X}}$.
- For a fixed set $A \subseteq \mathbb{X}$, a realization of $\mathcal{X}$ could be
  - fully contained in A,
  - fully contained in A^c,
  - or have non-empty intersection with both.
- Then the containment functional

$$b(A) = \mathsf{P}_{\mathcal{X}}(\mathcal{X} \subseteq A), \quad A \subseteq \mathbb{X},$$

is a continuous, completely monotone Choquet capacity and, in particular, $b(A) + b(A^c) \leq 1$.

Construction

- Express the statistical model, $\mathsf{P}_{Y|\theta}$, as

$$Y = a(\theta, U), \quad U \sim \mathsf{P}_U, \quad \mathsf{P}_U \text{ known.}$$

- Intuition: *If U were observable*, then just solve for θ in terms of Y and U — done!
- Unfortunately, *U is not observable*...
- But, since its distribution is known, we can "guess" its unobserved value with certain degree of reliability.
- This "guess" is based on a random set $\mathcal{S} \sim \mathsf{P}_{\mathcal{S}}$ in the U-space.
- There is theory to guide the choice of $\mathsf{P}_{\mathcal{S}}$.

Construction, cont.

- Given $\mathcal{S} \sim \mathsf{P}_{\mathcal{S}}$, push it forward to the θ space:

$$\Theta_y(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \{\vartheta : y = a(\vartheta, u)\}.$$

- The *belief function* is just the containment functional

$$b_y(A) = \mathsf{P}_{\mathcal{S}}\{\Theta_y(\mathcal{S}) \subseteq A\}.$$

- Dual is the *plausibility function*

$$p_y(A) := 1 - b_y(A^c) = \mathsf{P}_{\mathcal{S}}\{\Theta_y(\mathcal{S}) \cap A \neq \varnothing\}.$$

- Non-additive, i.e., $b_y(A) \leq p_y(A)$ for all $A \subseteq \Theta$.

Validity theorem

Theorem.

With a "suitable choice of random set $\mathcal{S} \sim \mathsf{P}_{\mathcal{S}}$," the inferential model constructed above satisfies the validity condition.

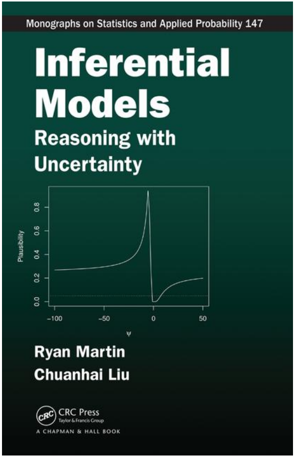
- Re: "suitable choice of $\mathcal{S}$."
  - Let $f(u) = \mathsf{P}_{\mathcal{S}}(\mathcal{S} \ni u)$.
  - Need $f(U) \geq_{\text{st}} \text{Unif}(0,1)$ when $U \sim \mathsf{P}_U$.
- This is actually easy to arrange...
- Consequences of validity:
  - "Reject $H_0 : \theta \in A$ if $p_y(A) \leq \alpha$" is a size α test.
  - The $100(1-\alpha)\%$ plausibility region

$$\{\vartheta : p_y(\{\vartheta\}) \geq \alpha\}$$

    is a $100(1-\alpha)\%$ confidence region.

Commercial

Valentine's Day is coming up — there's still time to get the perfect gift for that special someone



Binomial illustration

- Let $Y \sim \text{Bin}(n, \theta)$, distribution function F_θ.
- Construct an inferential model for θ:

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

  P. "Default" random set

  $$\mathcal{S} = \bigl\{u : |u - 0.5| \leq |\tilde{U} - 0.5|, \ \tilde{U} \sim \text{Unif}(0,1)\bigr\}.$$

  C. Combine to get[5]

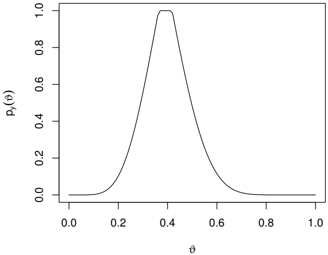
  $$\begin{aligned}\Theta_y(\mathcal{S}) &= \bigcup_{u \in \mathcal{S}} \{\theta : F_\theta(y-1) \leq u < F_\theta(y)\} \\ &= \bigl[1 - G^{-1}_{n-y+1,y}(\tfrac{1}{2} + |\tilde{U} - \tfrac{1}{2}|), 1 - G^{-1}_{n-y,y+1}(\tfrac{1}{2} - |\tilde{U} - \tfrac{1}{2}|)\bigr],\end{aligned}$$

- Note this only depends on $\tilde{U} \sim \text{Unif}(0,1)$...

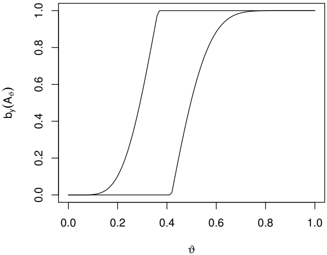
[5] Use fact that $F_\theta(y) = G_{n-y,y+1}(1-\theta)$, beta distribution.

Binomial illustration, cont.

- Data: $n = 18$, $y = 7$.
- Plots of plausibility contour $p_y(\{\vartheta\})$ and of belief and plausibility of $A_\vartheta = [0, \vartheta]$.



(e) Plausibility contour



(f) $b_y(A_\vartheta)$ and $p_y(A_\vartheta)$

Normal CV example

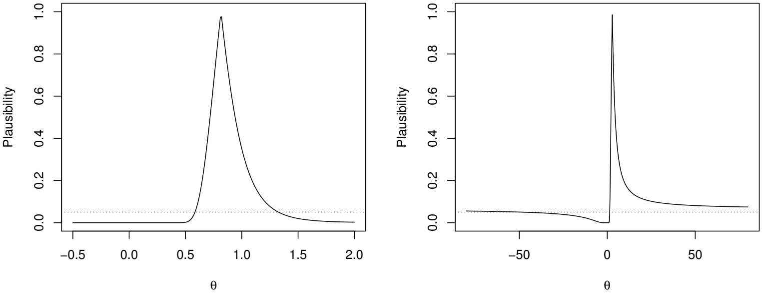
- Let $Y = (Y_1, \ldots, Y_n)$ be iid $\mathsf{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma)$.
- Goal is inference on $\phi = \sigma/\mu$, coefficient of variation.
- μ in the denominator makes this a difficult problem.
- Details are too lengthy to present here, so just the highlights:
  - build a (marginal) association:

$$\frac{n^{1/2}\bar{Y}}{S} = F_{n,1/\phi}^{-1}(U), \quad U \sim \mathsf{Unif}(0,1).$$

  - random set $\mathcal{S}$ for U yields a (marginal) inferential model for ϕ
  - guaranteed valid!

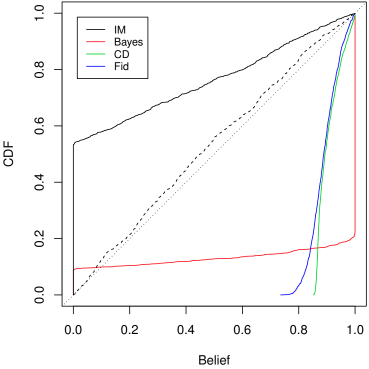
Normal CV example, cont.

- Simulate data of size $n = 30$ from $\mathsf{N}(\mu, \sigma^2)$, with $\sigma = 1$.
- Plots of plausibility contour for ϕ; based on "default" $\mathcal{S}$.
- Left: $\mu = 1$; right: $\mu = 0$.



Normal CV example, cont.

- Simulation: $n = 10$ from $\mathsf{N}(0.1, 1)$, so that $\phi = 10$.
- Consider $A = (-\infty, 9]$, which *does not* contain $\phi = 10$.
- Compute CDFs of various (marginal) beliefs b_y.
- Colored lines have false confidence, black lines don't!



Dimension reduction

- Inferential model construction and validity result is general, but more care is needed to get efficient solutions.
- I hid these steps in the normal CV example above.
- What's often needed is a way to reduce the dimension of U before introducing the random set to guess its value.
- For example, if we have iid data with

$$Y_i = a(\theta, U_i), \quad i = 1, \ldots, n,$$

  then the common θ and observed $y_1, \ldots, y_n$ imposes a constraint on the unobserved values $U_1, \ldots, U_n$.
- In other words, *there is a feature of U that is observed*.

Dimension reduction, cont.

- There exists $y \mapsto (T(y), H(y))$ and $u \mapsto (\tau(u), \eta(u))$ such that the above association can be re-expressed as

$$T(Y) = b(\theta, \tau(U)) \quad \text{and} \quad H(Y) = \eta(U).$$

- Key point is that the 2nd expression has no θ in it, hence the value of $\eta(U)$ is observed, even though U isn't.
  - Don't need to guess $\eta(U)$, hence dimension reduction;
  - Observed $\eta(U)$ helps improve guess of unobserved $\tau(U)$.
- How to find (T, H) and (τ, η)?
  - sufficient statistics
  - group invariance
  - *solving a partial differential equation*
  - ...

Dimension reduction, cont.

- Let $u_{y,\theta}$ be a solution to the equation $y = a(\theta, u)$.
- $\eta(U)$ would be observed if $\eta(u_{y,\theta})$ is constant in θ.
- That is, we seek η that satisfies

$$\frac{\partial \eta(u_{y,\theta})}{\partial \theta} = 0.$$

- I don't know about existence in general, but sometimes I can solve it, e.g., method of characteristics.
- Sometimes the solution depends on the parameter, which is bad, but there's a neat *localization* trick to get around this.
- Sorry for the lack of details...

Dimension reduction, cont.

- Location model illustration: $Y = \theta 1_n + U$.
- $u_{y,\theta} = y - \theta 1_n$.
- Goal is to solve

$$0 = \frac{\partial \eta(u_{y,\theta})}{\partial \theta} = \frac{\partial \eta(u)}{\partial u}\Big|_{u=u_{y,\theta}} \cdot \frac{\partial u_{y,\theta}}{\partial \theta}.$$

- Of course, right-most factor is 1_n in this case.
- Need η such that $\partial \eta(u)/\partial u$ sends constant vectors to 0.
- Take $\eta(u) = Mu$ where, e.g., $M = I_n - n^{-1} 1_n 1_n^\top$.
- Then $U_i - \bar{U} = Y_i - \bar{Y}$ are *observed*.

Dimension reduction, cont.

- A variation on the above location model:

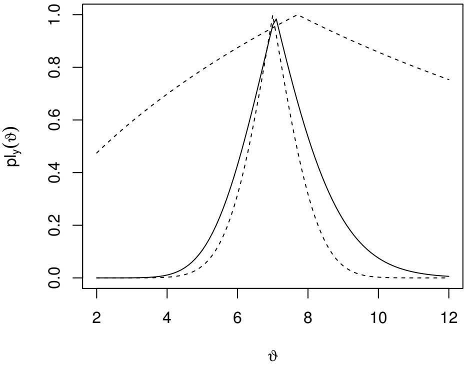
$$Y_1 = \theta + U_i, \quad Y_2 = \theta^{1/3} + U_2, \quad U_1, U_2 \overset{\text{iid}}{\sim} \mathsf{N}(0,1).$$

- "Non-regular," no reduction via sufficiency is possible.
- Same PDE as before can be set up, but solving it requires the localization trick I mentioned.
- Skipping details, the (localized) plausibility contour is

$$p_y(\{\vartheta\}) = 1 - \Bigl| 2\Phi\Bigl(\frac{y_1 - \vartheta + \frac{1}{3\vartheta^{2/3}}(y_2 - \vartheta^{1/3})}{(1 + \frac{1}{9\vartheta^{4/3}})^{1/2}} \Bigr) - 1 \Bigr|.$$

Dimension reduction, cont.

- Cube-root model, true value is $\theta = 7$.
- Plot plausibility contours based on individual observations and the one based on solving the PDE.

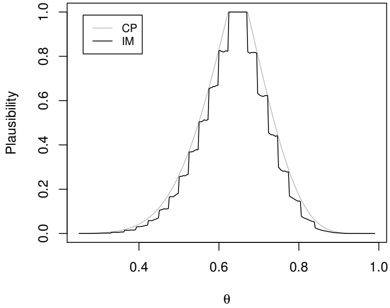


Confidence and beliefs

- Valid inferential models yield nominal confidence regions.
- Interesting reversal:
  - Given a nested family of confidence regions, create a function by stacking up its contours
  - This is a valid plausibility function.
- So, valid belief/plausibility is fundamental to statistics.
  - We always tell students that confidence *isn't* probability, but we never tell them what it *is*.
  - Confidence is plausibility: *a confidence set contains all parameter values that are sufficiently and justifiably plausible based on the observed data.*
  - Similar statements can be made about p-values.

Confidence and beliefs, cont.

- Complete-class theorem:[6]
  - (basically) for every confidence region there exists a random set such that the corresponding inferential model's is valid and its plausibility region is the same or smaller,
  - and there's an algorithm for constructing the random set.
- Binomial example...



[6] M., `arXiv:1707.00486`.

Concluding remarks

- Non-additive beliefs are fundamental to statistics.
- Additive beliefs put user at risk of false confidence and "systematically misleading conclusions."
- Avoid this risk with suitable non-additive beliefs.
- Main drawback of the proposed approach is that it's not always easy to do — we still don't understand it that well.
- Lots of interesting questions that remain to be answered:
  - marginalization mysteries
  - "optimal" random sets
  - incorporating prior info without losing validity
  - model uncertainty
  - computation!
  - ...

One more commercial

- The peer review system is broken in various ways.
- A particular concern of mine:
  - Feedback and judgment is through the same process
  - "Positive feedback" is good for our careers, but not for science
- Successful reform requires new ideas.
- Harry Crane and I developed a new open-access publication platform, featuring an *author-driven* peer review process.

RESEARCHERS.ONE

- For details, check us out at

www.researchers.one

www.twitter.com/@ResearchersOne

Thank you!

rgmarti3@ncsu.edu
www4.stat.ncsu.edu/~rmartin

RES3ARCHERS.ONE