Statistical Reasoning 101

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

Statistics Power Half-hour
NCSU Library
March 4th, 2021

[1] www4.stat.ncsu.edu/~rmartin

Introduction

- Statistics is hard, for various reasons.
- One is that we never know if we're right.
- So we're guided by *statistical principles*.
- Principles for an effective statistical analysis:[2]
  - data must be relevant
  - model must be sound
  - inference must be valid
- My focus here is mostly on the third principle.

[2] Crane & M. (2018), `https://researchers.one/articles/18.08.00011`

This talk

- Setup
- What is validity? Why is it important?
- Simple comparative example
  - Bayes and probability
  - Fisher and p-value/plausibility
- General comments beyond the example
  - Bayes vs non-Bayes?
  - resolving some confusion about them
- Conclusions

Setup

- Data Y.
- Some unknown θ of interest.
- Connect Y and θ with a model $\mathscr{P}$:
  - classical: $\mathscr{P} = \{Y \sim \mathbb{P}_\theta : \theta \in \Theta\}$
  - Bayesian: $\mathscr{P} = \{(Y, \theta) \sim \mathbb{P} = \text{“}\mathbb{P}_\theta \times \pi\text{”}\}$
- Goal is to use observed $Y = y$ to make inference about θ.
- For concreteness, consider testing[3] a hypothesis H about θ.
- Methods to use:
  - Bayes and (conditional) probability
  - Fisher and p-value/plausibility
  - .....

[3] These considerations are general, even beyond inference to prediction: Cella and M. (2020), https://researchers.one/articles/20.01.00010

Validity

- Statistician can choose any test, $\text{TEST}_H(Y)$.
- Why trust $\text{TEST}_H(Y)$ in a given application?
- We can't evaluate based on it being right/wrong.
- Effectiveness hinges on the method being reliable, or *valid*.
- Very, very roughly: the test is valid, relative to $\mathscr{P}$, if

$$\max_{\mathbb{P} \in \mathscr{P}} \mathbb{P}\{\text{TEST}_H(Y) \text{ is wrong}\} \leq \text{small}.$$

General observation.

Without validity, there's no reason to trust the conclusions drawn based on the method applied to the given data.

Validity, cont.

- Clearly, validity depends on both the method and model.
- That is:
  - one is free to choose any method, but...
  - whether the chosen method is valid depends on the model, i.e., on what one is willing to assume.
- So the goal is to establish validity under a *justifiable* model.
- Obviously, weaker assumptions are easier to justify...

Example

- Standard cancer screening example:

$$\begin{aligned} Y &= 1\{\text{patient tests positive}\} && \text{(observable)} \\ \theta &= 1\{\text{patient has cancer}\} && \text{(unknown)} \end{aligned}$$

- Scanning procedure is such that

$$Y \sim \mathbb{P}_\theta = \begin{cases} \text{Ber}(0.03) & \text{if } \theta = 0 \\ \text{Ber}(0.99) & \text{if } \theta = 1. \end{cases}$$

- Model is $\mathscr{P} = \{\mathbb{P}_\theta : \theta = 0, 1\}$.
- *If a patient tests positive, i.e.,* $y = 1$, *do they have cancer?*
- Goal is to test the hypothesis $H : \theta = 0$.

Example, cont.

- Given $y = 1$, a p-value/plausibility of $H : \theta = 0$,

$$\overline{\pi}_y(H) = \mathbb{P}_0(Y = 1) = 0.03.$$

- The "doesn't have cancer" hypothesis is *implausible*.
- Fact: relative to this model, the following test is valid:

$$\text{TEST}_H(Y) = \text{"reject } H \text{ if } \overline{\pi}_Y(H) \text{ is small"}.$$

- However, the *implausible* conclusion is rather weak:
  - certainly doesn't prove they have cancer
  - doesn't even imply they're "likely" to have cancer
  - only says that $y = 1$ implies having cancer is plausible

Example, cont.

- Bayesian solution is to use conditional probability.
- This requires an extra model assumption:
  - patient is randomly chosen from a population
  - in that population, 99% don't have cancer
- Bayes model: $\mathbb{P} =$ "$\theta \sim \text{Ber}(0.01)$ and $Y \mid \theta \sim \mathbb{P}_\theta$".
- If $y = 1$, then Bayes's theorem says

$$\pi_y(H) = \mathbb{P}(\theta = 0 \mid Y = 1) = 0.75.$$

- Fact: relative to the Bayes model, the following is valid:

$$\text{TEST}_Y(H) = \text{"reject } H \text{ if } \pi_Y(H) \text{ is small"}$$

- Probability claim is stronger than plausibility.

Example, cont.

- Two analyses come up with different answers:
  - non-Bayesian analysis implies H is *implausible*
  - Bayesian analysis implies H is *probable*
- More specifically:
  - former commits to the data: $Y = 1 \rightarrow$ cancer plausible
  - latter ignores the data: $Y = 0, 1 \rightarrow$ cancer improbable
- Which one is better? It depends...
  - If the stronger assumptions of the Bayes model are justified, then the Bayes method is valid—and stronger;
  - otherwise, the Bayes method is not valid so the non-Bayes method is preferred despite the weaker conclusions.
- Not a philosophical choice, it's about being justifiably valid.

Beyond this example

- The considerations in the above example are typical.
- That is, one is faced with the question: Bayes or not?
- The cancer screening example is a special one:
  - real/reliable prior information is available
  - so Bayes model is justifiable, and validity follows
- But perhaps most scientific applications have no genuine prior information available.
- Brad Efron: *Scientists want to work on new problems.*

Comments on probability

- A common criticism of p-values is that they don't measure **the** probability of the hypothesis.
- This complaint assumes "probability of H" is real.
- Bruno de Finetti: *Probability does not exist.*
- Therefore: "probability of H" is only in the statistician's mind
- Is what's in the statistician's mind practically relevant?
  - only if its connection to the real world is established
  - that's what validity/reliability aims to do.

Comments on probability, cont.

- Can't fake it with default/non-informative priors...
- That is, one can't both have and eat cake:
  - stronger probabilistic conclusions aren't free
  - price is making stronger assumptions, harder to justify
- Don Fraser:
  - *Bayes's theorem does not create real probabilities from hypothetical probabilities.*
  - *Any serious mathematician would ask how you could use a lemma with one premise missing by making up an ingredient and thinking that the lemma's conclusions were still available.*

Comments on probability, cont.

- Fact: Under the Bayes model for Y, based on prior π,

$$\mathbb{E}\{\pi_Y(H)\} = \pi(H)$$

- Therefore, under the Bayes model, validity follows because the posterior is near the real prior.
- Alternatively, under the non-Bayes model, $\mathbb{P}_{\theta^\star}$,
  - the posterior can learn (e.g., consistency)
  - but generally isn't valid, i.e., a risk of *false confidence*[4][5]

[4] Balch, M., and Ferson (2019), `arXiv:1706.08565`

[5] M. (2019). `https://researchers.one/articles/19.02.00001`

Comments on plausibility

- Fisher/p-values/significance are misunderstood.
- A small p-value...
  - *does not imply a scientific discovery*
  - suggests a notable finding, worthy of further investigation.
- The "plausibility" interpretation makes this clear:
  - small p-value means H is implausible
  - not-H being plausible is far from scientific proof of not-H.
- Fisher was clear that p-values are not probabilities, but there was no alternative to probability at that time...

Comments on plausibility, cont.

- Plausibility theory isn't something I made up:
  - mathematically rigorous framework for UQ
  - part of *imprecise probability*
- Plausibility provides a convenient way to interpret p-values and confidence intervals.
- I teach this to students in my classes:

  > *"...your discussion of plausibility in ST503 made it easier for me to explain what a p-value is at my job interview. Thank you for the new perspective on the subject!"*

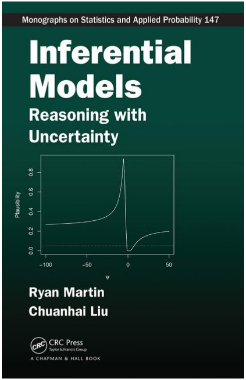
- Strong validity results, no prior necessary.
- Conclusions based on plausibility are weaker
  - but p-values do exactly what they advertise to do
  - do data suggest that not-H is worth investigating further?

Conclusion

- Take-away messages:
  - validity is crucial to statistical reasoning
  - any method can be valid
  - valid or not depends on assumptions one is willing to make
- Bayes and probability
  - strong probabilistic conclusions
  - no validity guarantees without strong assumptions
- Fisher and p-values/plausibility
  - validity under weaker assumptions
  - weaker plausibility-based conclusions
- Remember: statistics is hard.
- So we have to calibrate our expectations:
  - nobody can say that one approach is always right
  - validity under justifiable assumptions can't be wrong

Commercial

- Some recent work about these ideas:
  - `https://researchers.one/articles/19.02.00001`
  - `https://researchers.one/articles/21.01.00002`
- How I got started thinking about these things:



The end

- Questions? Email me at rgmarti3@ncsu.edu.
- Papers, etc. at www4.stat.ncsu.edu/~rmartin/
- Check out *Researchers.One*:
    - https://researchers.one
    - www.twitter.com/ResearchersOne

Thanks for your attention!