'Understanding Non-Bayesians'

Francis J. DiTraglia

University of Oxford

Summer of Bayes 2024

Goals of this Presentation[1]

1. Summarize some key points from Sims (2010) - Understanding Non-Bayesians.
2. Relate them to the broader discussion of Bayesian vs. Frequentist inference.

[1] Unless otherwise indicated, all quotes are from Sims (2010).

Fun Facts

- "Understanding Non-Bayesians" was originally written for the Handbook of Bayesian Econometrics in 2010.
- Oxford University Press objected to Sims posting a pre-print on his website.
- Sims favors open access and withdrew from the handbook in protest; the paper remains unpublished.
- In 2011 Sims was awarded the Economics Nobel for "understanding cause and effect in the macroeconomy."
- Take that OUP!

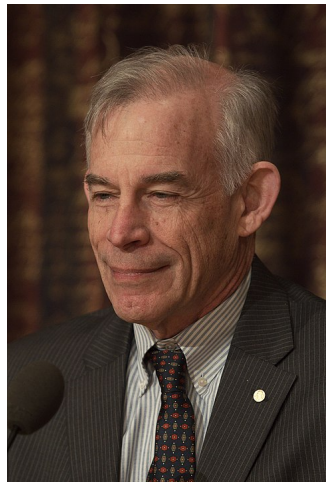


Figure 1: Chris Sims in 2011

Motivation: Why isn't everyone Bayesian?

> *Once one becomes used to thinking about inference from a Bayesian perspective, it becomes difficult to understand why many econometricians are uncomfortable with that way of thinking. But some very good econometricians are either firmly non-Bayesian or (more commonly these days) think of Bayesian approaches as a "tool" which might sometimes be appropriate, sometimes not.*

Some Pedantry

Maybe we should call it "Bayes-Laplace" inference: Laplace developed what we would recognize as "Bayesian" inference.

Figure 2: Laplace in 1775

What is this paper about?

- Purpose: Articulate counterarguments to Bayesian perspective
- Some counterarguments are "easily dismissed"
- Others relate to deep questions about inference in infinite-dimensional spaces

My conclusion is that the Bayesian perspective is indeed universally applicable, but that "non-parametric" inference is hard, in ways about which both Bayesians and non-Bayesians are sometimes careless.

- Some of the insights in the paper are a bit scattered; I've re-arranged a bit.

Differing Interpretations of Probability

- Crucial background to the differences between Bayesians and Frequentists.
- Math is the same either way (Kolmogorov Axioms) but meaning is different.
- Bayesians: **"Belief-type"** interpretation.
- Frequentists: **"Frequency-type"** interpretation.
- Sims doesn't discuss this. The next two slides (including quotes) are based on Chapter 11 of An Introduction to Probability and Inductive Logic by Ian Hacking.

Belief-Type: "It is probable that the dinosaurs were made extinct by a giant asteroid hitting the earth."

Interpersonal / Evidential: Keynes, Jeffreys, Jaynes

- "Relative to the available evidence, the probability that the dinosaurs were made extinct by a giant asteroid hitting the earth is high–about 0.9."
- "She thinks that [the statement] is *interpersonal*–because it is about what is reasonable for any reasonable person to believe. And since the degree of belief should depend on the available *evidence* we call this *interpersonal/evidential*."

Personal Degree of Belief: de Finetti, Savage

- "I personally am very confident that the dinosaurs were made extinct by a giant asteroid hitting earth."
- "If I had to make a bet on it, I would bet 9 to 1 that the dinosaurs were made extinct by a giant asteroid hitting the earth."

Frequency-Type: "The probability of getting heads with this coin is 0.6."

The truth of this statement seems to have nothing to do with what be believe. We seem to be making a completely factual statement about a material object, namely the coin . . . We may be saying something like:

- *In repeated tossing, the relative frequency of heads settles down to a stable proportion, 6/10.*
- *The coin has a tendency to come down heads far more often than tails.*
- *It has a propensity or disposition to favor heads.*
- *Or we are saying something more basic about the asymmetry of the coin and the tossing device. We may be referring to the geometry and physics of the coin, which cause it to come down more often heads than tails.*

Learn some $*@%&!# physics before you talk to me about coin flips![2]

[2] See Chapter of 10 of Jaynes' Probability Theory: The Logic of Science for more discussion.

Bayesian versus Frequentist Approaches

Frequentist

- "Insists on a sharp distinction between unobserved, but non-random 'parameters' and observable, random data."
- "Works entirely with the probability distributions of data, conditional on unknown parameters–estimators and test statistics, for example–and makes assertions about the distribution of those function of the data, conditional on parameters"

Bayesian

- "Treats everything as random before it is observed, and everything observed as, once observed, no longer random."
- "Aims at assisting in constructing probability statements about anything as yet unobserved (including 'parameters') conditional on the observed data."

Let's unpack this a bit: Frequentists

- Condition on parameters; make probability statements that average over different *datasets* you could potentially observe.
- E.g. $X_1 ..., X_{100}$ is a random sample from a $N(\mu, \sigma^2)$ population with $\sigma = 1$ known.
- "$\bar{X}_n \pm 1.96 \times \sigma/\sqrt{n}$ is a 95% confidence interval for μ."
- Translation: In 95% of the datasets we could possibly observe, the sample mean will land within about ± 0.2 of the true (fixed and unknown) value of μ.
- The observed interval $\bar{x} \pm 0.2$ either contains μ or doesn't: nothing is random after we have observed the data.
- Traditional (Neyman-Pearson) inference is *pre-experimental*: inductive behavior rather than inductive inference.

Let's unpack this a bit: Bayesians

- Condition on *observed data*; make probability statements that average over different *parameter values* that could potentially have generated the data.
- E.g. $X_1 ..., X_{100}$ is a random sample from a $\mathrm{N}(\mu, \sigma^2)$ population with $\sigma = 1$ known.
- Need a prior: just for simplicity choose a "vague" one e.g. $\mu \sim \mathrm{N}(0, 10000)$
- "The posterior distribution of μ is (approximately) $\mathrm{N}(\bar{x}, 1/100)$, so the 95% highest posterior density interval (HPDI) for μ is approximately $\bar{x} \pm 0.2$."
- Translation: Given the observed data, there is around a 95% probability that the population mean μ lies within ± 0.2 of the sample mean $\bar{x}$.
- The observed sample mean $\bar{x}$ is fixed and known; the population mean μ is unknown and treated as random.
- Bayesian inference is *post-experimental* (conditional): inductive inference under an assumed model given observed information.

Implications for Decision-Making

Bayesian inference therefore feeds naturally into discussion of decisions that must be made under uncertainty, while frequentist analysis does not. There are theorems showing that under certtain assumptions is is optimal to make decisions based on probability distributions over unknown quantities, but it is probabily more important that most actual decision makers find the language of probability a natural way to discuss under uncertainty and how the results of data analysis may be relevant to their decisions.

Why does Sims put the word "parameters" in quotes?

Poirier (1996) textbook on econometrics: chapter 5?

- "Bayesians take models seriously not literally; Frequentists take models literally not seriously."
- Quote from Porier about specification testing.

"Easily Dismissed" Objection # 1: "Bayseian Inference is Subjective"

Bayesian inference makes the role of subjective prior beliefs in decision-making explicit, and describes clearly how such beliefs should be modified in the light of observations. But most scientific work with data least to publication, not directly to decision-making. That is, most data analysis is aimed at an audience who face different decision problems and may have diverse prior beliefs. In this situation . . . useful data analysis summarizes the shape of the likelihood. Sometimes it is helpful to apply non-flat, simple, standardized prior in reporting likelihood shape, but ***these are chosen not to reflect the investigator's personal beliefs, but to make the likelihood description more useful to a diverse audience****. A Bayesian perspective makes the entire shape of the likelihood in any sample directly interpretable, whereas a frequentist perspective has to focus on the large-sample behavior of the likelihood near its peak.*

"Easily Dismissed" Objection # 1: "Bayseian Inference is Subjective"

Though frequentist data analysis makes no explicit use of prior information, ***good applied work does use prior beliefs informally even if it is not explicitly Bayesian****. Models are experimented with, and versions that allow reasonable interpretations of the estimated parameter values are favored. Lag lengths in dynamic models are experimented with, and shorter lag lengths are favored if longer ones add little explanatory power. These are reasonable ways to behave, but they are not "objective".*

"Easily Dismissed" Objection #1: "Bayesian Inference is Subjective"

Researchers who take a Bayesian perspective can take a completely "objective" approach, by aiming at description of the likelihood. Frequentists have no formal interpretation of the global likelihood shape. Frequentist textbook descriptions of methods make no reference to subjective prior beliefs, but everyone recognizes that good applied statistical practice, even for frequentists, entails informal use of prior beliefs when an actual decision is involved. Its supposed "subjectivity" is therefore no reason to forswear the Bayesian approach to inference.

Some Other Views on "Subjectivity"

Gelman et al (2014) - *Bayesian Data Analysis*

> *Bayesian methods are sometimes said to be especially subjective because of their reliance on a prior distribution, but in most problems, scientific judgement is necessary to specify both the "likelihood" and "prior" parts of the model. For example, linear regression models are generally at least as suspect as any prior distribution that might be assumed about the regression parameters.*

McElreath (2020) - *Statistical Rethinking*

> *Within Bayesian data analysis in the natural and social sciences, the prior is considered to be just part of the model. As such it should be chosen, evaluated, and revised just like all the other components of the model . . . priors and Bayesian data analysis are no more inherently subjective thatn likelihoods and the repeated sampling assumptions required for significance testing.*

Fighting Words...

Jaynes (2003) - *Probability Theory: The Logic of Science*

> *Today one wonders how it is possible that orthodox logic continues to be taught in some places year after year and praised as "objective", while Bayesians are charged with "subjectivity". Orthodoxians, preoccupied with fantasies about nonexistent data sets and, in principle, unobservable limiting frequencies – while ignoring relevant prior information – are in no position to charge anybody with "subjectivity". If there is no sufficient statistic, the orthodox accuracy claim based on a single "statistic" simply ignores not only the prior information, but also all the evidence in the data that is relevant to that accuracy: hardly an "objective" procedure. If there are ancillary statistics and the orthodoxian follows Fisher by conditioning on them, he obtains just the estimate that Bayes' theorem based on a noninformative prior would have given him by a shorter calculation. Bayes' theorem would have given also a defensible accuracy claim.*

Easily Dismissed Objection #2: "Bayesian Inference is Harder"

- Frequentist inference is often (wrongly) conflated with "convenient and intuitively appealing estimators" asymptotic approximations.
- Could instead aim for exact finite-sample distribution theory and fully efficient estimators but this is often intractable.
- "Easier to characterize optimal small-sample inference from a Bayesian perspective, and much of the Bayesian literature has insisted that this is a special advantage."
- Simulation-based methods make it easy to explore the shape of complicated likelihood functions in large, non-linear models.

Side Note on Modern Simulation-based Methods

- Probabilistic programming languages (PPLs) like STAN, JAGS, BUGS, PyMC3 make it extremely easy to build and estimate complex Bayesian models.
- We'll have a session on this in September with special guest Alex Andorra.
- StanCon 2024 is in Oxford this year! (September 9-13th, 2024)

From Chapter 1 of *Asymptotic Statistics* by van der Vaart

For a relatively small number of statistical problems there exists and exact, optimal solution . . . If exact optimality theory does not give results, be it because the problem is intractable or because there exist no "optimal" procedures, then asymptotic optimality theory may help . . . ***strictly speaking, most asymptotic results that are currently available are logically useless. This is because most asymptotic results are limit results, rather than approximations consisting of an approximating formula plus an accurate error bound*** *. . . This is why there is good asymptotics and bad asymptotics and why two types of asymptotics sometimes lead to conflicting claims . . . Because it may be theoretically very hard to ascertain that approximation errors are small, one often takes recourse to simulation studies*

But what about GMM, OLS and IV?!

- "There is no reason in principle that very easily implemented estimators like [these] ... have to be interpreted from a Frequentist perspective."
- Kwan (1998) if $\sqrt{n}(\hat{\beta} - \beta) \to_d \text{Normal}(0, \Sigma)$ then, under regularly conditions and for large n, we can approximate $\beta|\widehat{\beta} \approx \text{Normal}(\widehat{\beta}, \Sigma/n)$.
- When the regularity conditions hold, can interpret Frequentist confidence intervals as (approximate) Bayesian posterior credible intervals.
- Only conditions on $\widehat{\beta}$, not the full data: "throws information away."

Stronger Result

> *If we have a trustworthy model for which the likelihood function can be computed, the likelihood function will, under regularity conditions, take on a Gaussian shape in large samples, with the mean at the maximum likelihood estimate (MLE) and the covariance matrix given by the usual frequentist estimator for the covariance matrix of an MLE.*

Differences from Kwan (1998) Result

1. For large n, conditioning on MLE and using a large-sample normal approximation doesn't throw away information: as good as conditioning on all the data.
2. Only a shortcut: from a Bayesian perspective, can always check the quality of the approximation by examining the whole likelihood function.
3. Lacks "robustness" in that: "interpreting the likelihood shape as the posterior generally requires believing that the likelihood is correctly specified."

"Pragmatic Bayesian" Approach

In some circumstances we need to be able to reach conclusions conditional on standard, easily computed statistics, even when they are not sufficient statistics, and that sometimes approximate distribution theory, justified by the fact that in large samples with high probability the approximation is good, is the best that can be done with the available time and reources. Yet at the same time, it is always worthwhile to consider whether such compromises are throwing away a great deal of information or resulting in seriously distorted posterior distributions.

What does it mean to be "conservative"?

These models are conservative in a sense – they promise no more precise estimates, asymptotically, that what can be obtained under the "weakest" of the models defined by the assumptions supporting the frequentist asymptotics. But to say they are conservative may be misleading. Using a model that fails to make efficient use of the data can lead to large, avoidable losses. Concluding that the data do not support a finding that a drug or social policy intervention has important effects is a costly error if the conclusions is incorrect, and reading such a conclusion in a naive attempt to be "robust" by using "weak assumptions" is not in any real sense conservative.

Generating Conservative Models[3]

- Caveats notwithstanding, it might be convenient to set up a "conservative" Bayesian model. How could we do it?
- Suppose we have "weak assumptions" in the form of moment conditions $\mathbb{E}[g(y, \beta)|\beta)] = 0$.
- Goal: find likelihood $f(y|\beta)$ that satisfies these constraints but is otherwise as "uninformative as possible".
- Formally: "minimize Shannon mutual information between y and β"
- Solution: exponential family models!
- E.g. linear regression with normal errors solves the problem subject to

$$\mathbb{E}\left[(y - \mathbf{x}'\beta)\mathbf{x}|\beta\right] = \mathbf{0}$$

[3] Chapter 10 of *Statistical Rethinking* by McElreath explains this better than Sims (2010)

Less Easily Dismissed Objection #1: Handy Methods Seem Un-Bayesian

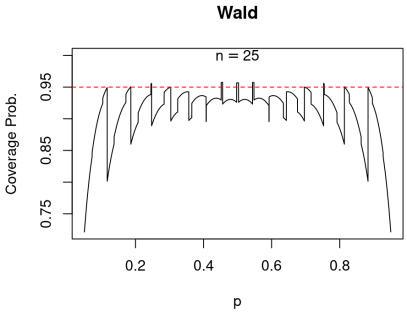
- To me this doesn't really seem like an "objection" – more of a continuation of the "Pragmatic Bayesian" discussion from above.
- Above: IV, GMM etc. can be given "limited information" Bayesian interpretation.
- Frequentist asymptotic distribution can be viewed as an approximate Bayesian posterior under "essentially the same weak assumptions"
- I.e. don't assume known likelihood: only moment/regularity conditions
- "But every application of asymptotic approximate theory relying on 'weak assumptions' involves a Bayesian judgment call."
- In other words: when does the Frequentist asymptotic distribution theory perform well *on its own terms*?
- To rely on the Central Limit Theorem in practice, for example, is to implicitly assume we're in a region of the parameter space where it works well.

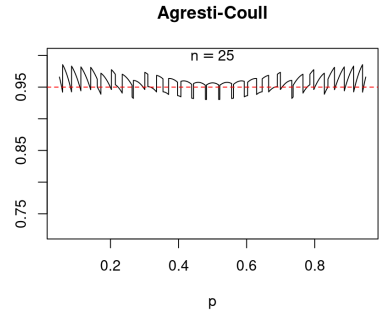
An even simpler example than the one from Sims (2010)...

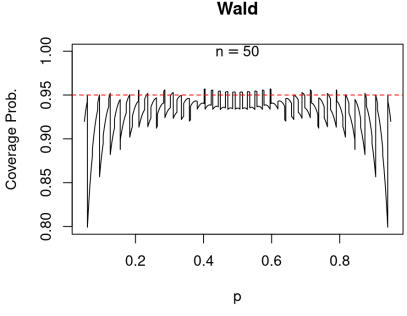
- Suppose that $X_1, ..., X_n \sim$ iid Bernoulli(p) where p is unknown.
- Let $\widehat{p}$ be the sample proportion, i.e. $\frac{1}{n}\sum_{i=1}^{n} X_i$
- The Central Limit Theorem (CLT) gives $\widehat{p} \approx \text{Normal}(p, \widehat{p}(1-\widehat{p})/n)$ for large n.
- Approximate "Limited Information" Posterior: $p \approx \text{Normal}(\widehat{p}, \widehat{p}(1-\widehat{p})/n)$.)
- 95% Confidence / Credible Interval: $\widehat{p} \pm 1.96 \times \sqrt{\widehat{p}(1-\widehat{p})/n}$

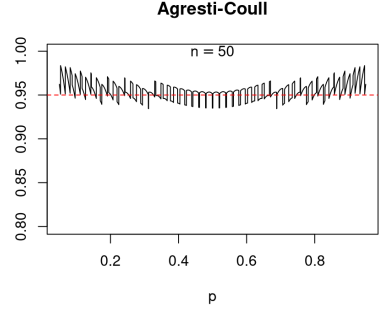
But this approximation is *terrible* when p is far from 0.5...

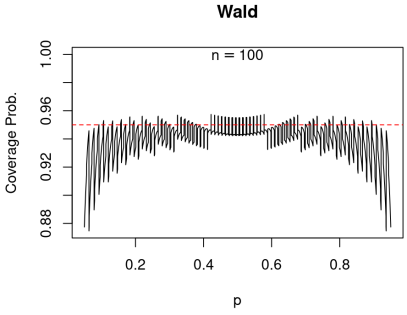
- Bayesian Perspective: bad conditional properties e.g. empty intervals!
- Frequentist Perspective: coverage probabilities are all over the place!
- "Textbook" rule of thumb $np(1-p) > 5$ is a Bayesian judgement call (also false!)

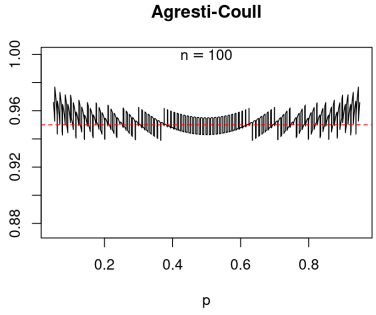

Wald
n = 25
Coverage Prob.
0.75
0.85
0.95
0.2
0.4
0.6
0.8
p

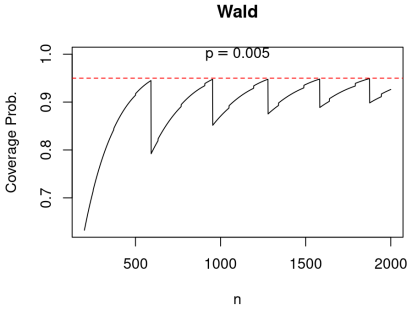

Agresti-Coull
n = 25
0.75
0.85
0.95
0.2
0.4
0.6
0.8
p

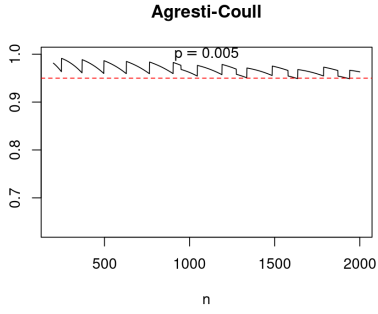

Wald
n = 50
Coverage Prob.
0.80
0.85
0.90
0.95
1.00
0.2
0.4
0.6
0.8
p


Agresti-Coull
n = 50
0.80
0.85
0.90
0.95
1.00
0.2
0.4
0.6
0.8
p


Wald
n = 100
Coverage Prob.
1.00
0.96
0.92
0.88
0.2
0.4
0.6
0.8
p


Agresti-Coull
n = 100
1.00
0.96
0.92
0.88
0.2
0.4
0.6
0.8
p


Wald
p = 0.005
Coverage Prob.
1.0
0.9
0.8
0.7
500
1000
1500
2000
n


Agresti-Coull
p = 0.005
1.0
0.9
0.8
0.7
500
1000
1500
2000
n

Challenges in Non-parametrics

- Infinite-dimensional parameter spaces
- Consistency issues in Bayesian inference
- Pitfalls in high-dimensional spaces:
  - Priors can be unintentionally dogmatic
  - Importance of careful prior specification

Example: Angrist and Krueger (1991) Quarter of Birth

- The Wasserman problem is about non-parametrics and you can read about it here.
- But we don't need anything too exotic to see the issues Sims is talking about.
- If you're not an economist and don't know what instrumental variables is, here's a very quick introduction.
- Give the introduction.
- Then make the point of Chamberlain & Imbens (1996)
- Point out that the Frequentist solution is also terrible in this case since it corresponds to an insane prior!
- Useful dialogue between Bayesians and Frequentists: what prior does the frequentist solution correspond to? Frequency properties of Bayesian estimators?

Example 1: The Wasserman Problem

- Setup: Observing (ξ, R, Y) with unobserved θ
- Goal: Estimate $\psi = \mathbb{E}[\theta]$
- Bayesian approaches:
  1. Independence case
  2. Dependence case (sieve method)
  3. Limited information approach

Critique of Wasserman's Conclusions

I probably still want to mention these points, but I don't really want to get into the Wasserman example since it won't be as familiar to the audience.

- Bayesian methods are not necessarily insensitive to data
- Importance of appropriate prior specification
- Pitfalls of high-dimensional parameter spaces

Example 2: Robust Variance Estimates in Regression

Not sure how much I should say about this one, but if I do mention it then it might be worth mentioning the Leamer "White-washing" stuff along with the paper where he talks about the "sandwich" estimator versus GLS and something about when the point estimates will change.

- OLS with sandwich covariance matrix
- Efficiency bounds (Chamberlain, 1987)
- When is OLS with sandwich appropriate?
  - Large samples
  - Likely nonlinear regression function
  - Interest in best linear predictor

Conclusion

- Bayesian perspective is universally applicable
- Importance of careful modeling in high-dimensional spaces
- Pragmatic Bayesian approach:
  - Recognize limitations of asymptotic approximations
  - Consider model improvements when appropriate
  - Use OLS with sandwich judiciously

Questions?