(Likelihood-based) Frequentistism and/or Bayesianism

Chris Wymant

I’m not an expert, but I’m still required to have a revealed preference - what I do in practise.

Hopefully uncontroversial:

In cases where
- an answer is urgent and important,
- data are weakly informative,
- a reasonable prior does exist,

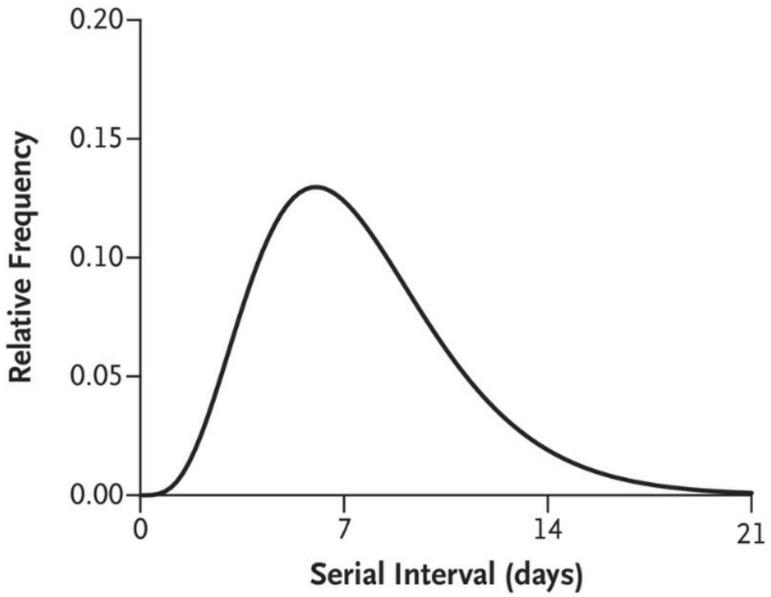
embracing Bayesianism helps.

"an informative prior distribution for the serial interval based on the serial interval of SARS"

+

Data (N=6)

Serial Interval (days)
5
9
7
7
3
7



Li et al, NEJM Jan 2020,
18,000 citations

Definitions

Likelihood $L(\theta) = P(\text{data} \mid \theta)$ considered as a function of θ for fixed data. *Is not* a probability distribution function.

Sampling distribution = $P(\text{data} \mid \theta)$ considered as a function of data (different possible realisations of it) for fixed θ. *Is* a probability distribution function

Frequentist probability:

- Defined only for random variables (e.g. data), not for parameters / hypotheses / models
- Define it, via the sampling distribution, as the fraction of an asymptotically large number of identical stochastic trials with that value of the random variable as an outcome.

Bayesian probability:

- Extend the above definition to *also* include degree of belief / confidence / certainty, quantified consistently (e.g. things sum to 1).
- This extends its scope to parameters / hypotheses / models / …

"Remember that using Bayes' Theorem doesn't make you a Bayesian. Quantifying uncertainty with probability makes you a Bayesian."
Michael Betancourt

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

Is true from the basic laws of probability.

Frequentist:

- Both A and B have to be random variables.
- In this form, difficult to see how this permits *inference* - drawing a conclusion.

Bayesian:

- B can be the target of our inference: P(what we want to know | what we know)
- Cannot get P(B|A) without P(B): the conditional and unconditional probabilities are related.

Choice of priors I

- A Bayesian must specify the prior P(parameters), just as we must usually specify the likelihood P(data | parameters). Both form part of the model for inference.
- You can choose whatever prior you want - your posterior is conditional on it. (When you condition on something you say nothing about how likely it is, c.f. parameters in the likelihood.) e.g. choose a prior for your favourite treatment being as likely to cause harm as benefit. Not your true prior, but the one that's persuasive to sceptics.
- We do not need to meticulously work in all relevant knowledge about a parameter x into P(x) for this to be acceptable inference, just as we do not need a likelihood that models the effect of every potentially entangled phenomenon: models should be as simple as possible but no simpler. With your true prior you would get your true posterior (conditional on the correctness of the likelihood); with a simple weakly informative prior the posterior is then basically just what this dataset is telling us (conditional on the correctness of the model), i.e. not a synthesis of all available evidence.

Choice of priors II

- Using only very rough reasoning one can usually think of scales for the problem that are plausible and scales that are wildly implausible. Can make your life easier e.g. by modelling the data after standardisation (subtract the sample mean, divide by the sample standard deviation) instead of the raw data, so that variability is roughly O(1).
- Numerical implementations of Frequentist procedures must, I think, consider only finite parameter ranges: a computer cannot handle x^x^x^x^x… where x = number of particles in the universe. A Bayesian procedure can make this explicit.
- For robust inference, different reasonable choices for the prior should lead to roughly the same conclusion. (The same is true for the likelihood.)
- Sometimes we're lazy and don't test different choices for priors; this is excusable if we can see the model fits the data well. (The same is true for the likelihood.) If we see the model fits the data poorly, we must revise our choice for the prior (or likelihood), or think cautiously about having found an unexpected result or a biased dataset.
- A little advice on choosing priors here https://github.com/ChrisHIV/teaching/blob/main/other_topics/2021-09-29_Chris_InferenceOnly.pdf but doubtless much better elsewhere.

Choice of priors III

Criticism: your answer depends on your choice of prior.

However, subjectivity exists already.

- Choice of dataset(s)
- Choice of data cleaning / preparation
- Choice of model linking data to quantity of interest
  - Type/structure of model
  - *Choice of parameterisation of quantity of interest (relative or absolute, ratio or difference, …)*
  - Choice of nuisance parameter values
- Details of implementing the model into an algorithm may have an effect (e.g. parameter translation affects MCMC)
- Which results to report, interpretation, communication, …

“It’s always a leap of faith going from an analysis result to a conclusion about the world” ~ Christophe Fraser

When only uninformative data is available, Frequentist analyses can come to seemingly categorically contradictory results - e.g. masks help or they don't - based on what was considered the null hypothesis (e.g. through choice of parameterisation), the effect of investigator bias on study design, and the binarisation involved in statistical significance.

With highly informative data, different sensible Bayesian estimates should be similar to each other.

(So I disagree with the argument that Frequentist results are likely to have more a desirable level of consistency between studies, by virtue of avoiding differences in subjective priors.)

Hierarchical / multi-level models are widely appropriate: for dataset containing natural groupings/subsets.

Somewhat contrived in a frequentist analysis - the set of parameters **r** that distinguish each group's characteristics from the other groups are treated as random variables rather than parameters, so that we can do exactly what would be natural to a Bayesian: calculate the marginal likelihood without **r** as the integral P(data | **r**) P(**r** | other parameters) d**r**

Useful generally to be able to marginalise over whatever you're not currently interested in, whether it's a random variable or not.

Sir David MacKay
(Information theory heavyweight, founder of the Inference Group, Cambridge University)

“There are two principal paradigms for statistics: sampling theory and Bayesian inference. In sampling theory (also known as ‘frequentist’ or orthodox statistics), one invents estimators of quantities of interest and then chooses between those estimators using some criterion measuring their sampling properties; there is no clear principle for deciding which criterion to use to measure the performance of an estimator; nor, for most criteria, is there any systematic procedure for the construction of optimal estimators. In Bayesian inference, in contrast, once we have made explicit all our assumptions about the model and the data, our inferences are mechanical. Whatever question we wish to pose, the rules of probability theory give a unique answer which consistently takes into account all the given information. Human-designed estimators and confidence intervals have no role in Bayesian inference; human input only enters into the important tasks of designing the hypothesis space (that is, the specification of the model and all its probability distributions), and figuring out how to do the computations that implement inference in that space. The answers to our questions are probability distributions over the quantities of interest.”

(From Information Theory, Inference, and Learning Algorithms)

“The fallacy of placing confidence in confidence intervals” review, Morey et al - long but very helpful!

For clarity, separate a confidence interval procedure from a confidence interval (one output from one such procedure).

An X% CI for a parameter θ is an interval (L, U) generated by a procedure that in repeated sampling has an X% probability of containing the true value of θ, for all possible values of θ (Neyman, 1937)

For a given problem one can define different CI procedures. If you’re careful, choose one that also has high power: excluding false parameters with high frequency. Even then, for some realisations of the data, you can end up with a CI that includes all possible values and/or includes impossible values.

The only permissible interpretation of any given X% CI is that it is one realisation of procedure with that property. Interpreting it as “there is an X% chance that the true value lies within this particular interval” quickly implies things that are logically inconsistent with each other.

“Once one has collected data and computed a confidence interval, how does one then interpret the interval? The answer is quite straightforward: one does not – at least not within confidence interval theory… Frequentist theory is a “pre-data” theory. It looks forward, devising procedures that will have particular average properties in repeated sampling”

“The fallacy of placing confidence in confidence intervals” review, Morey et al - long but very helpful!

Guidelines:
- Report credible intervals instead of confidence intervals
- Do not use confidence procedures whose Bayesian properties are not known
- Warn readers if the confidence procedure does not correspond to a Bayesian procedure
- Authors choosing to report CIs have a responsibility to keep their readers from invalid inferences, because it is almost certain that without a warning readers will misinterpret them
- *Never* report a confidence interval without noting the procedure and the corresponding statistics (there are many different ways to construct confidence intervals, and they will have different properties)
- Consider reporting likelihoods or posteriors instead (an interval provides fairly impoverished information)

Misconception: Bayesian and frequentist inference lead to the same inferences, and hence all confidence intervals can simply be interpreted in a Bayesian way. Actually they can differ markedly. Even when true, this “is actually no defense at all. One must first choose which confidence procedure, of many, to use; if one is committed to the procedure that allows a Bayesian interpretation, then one’s time is much better spent simply applying Bayesian theory.”

Gelman:
“Bayesian statistics is about *making* probability statements, frequentist statistics is about *evaluating* probability statements…
I’m not quite sure what a “frequentist method” is, but I will assume that the term refers to any statistical method for which a frequency evaluation has been performed... However, I don’t quite understand Wasserman’s statement, “If the Bayes estimator has good frequency behavior then we might as well use the frequentist method.” As far as I know, there is no “frequentist method” for coming up with an estimator… the frequentist method, as I understand it, is an approach for evaluating inferences—in which case, I have no problem with Wasserman or anyone else taking a Bayesian inference and labeling it as “frequentist.”

DOI:10.1214/08-BA318REJ

Gelman on significance testing:

“there are various ways to evaluate a model, and one of these is model checking, comparing fitted model to data. This is the same as significance testing or hypothesis testing, but with two differences:

(1) I prefer graphical checks rather than numerical summaries and p-values;

(2) I do model checking to test the model that I am fitting, usually not to test a straw-man null hypothesis. I already know my model is false, so I don’t pat myself on the back for finding problems with the fit (thus “rejecting” the model); rather, when I find problems with fit, this motivates improvement to the model.”

https://statmodeling.stat.columbia.edu/2018/06/17/bayesians-are-frequentists/

Summary of my take

	Pro-Frequentist Arguments	**Pro-Bayesian Arguments**
Valid	• For problems with standard likelihoods, informative data, and ML parameters not near boundaries of parameter space etc. then usually (but when? and which ones?) frequentist CIs are like Bayesian CIs but easier to calculate using standard tools a la glm(output ~ predictor1 + predictor2, link = …) • Some people want to distort good inference for their own ends; if they can get away with not (clearly) reporting their biased priors and not showing model fit, allowing them the Bayesian option is worse. By doing good Bayesian analysis perhaps we facilitate bad Bayesian analysis?	• The combined effect of the choice of parameterisation and choice of prior is made explicit, and easier to critique & revise • The posterior means what we want it to mean • posterior is a pdf not a 1D interval • Posterior links parameters - can calculate derived quantities and reparameterise after inference
Intermediate	For other problems it's quicker? Unclear once one is proficient with a method like Stan.	Principled model selection and model averaging via P(model i | data). True in theory but apparently highly impractical in practise - this quantity depends very sensitively on the prior even when the posterior does not.
Not valid	"Removes subjectivity." See previous slides.	"We should always be including relevant information from previous studies in our prior." Not valid because your answer is conditional on the prior - given these assumptions, this is what follows. Assumptions are given, not calculated; use whatever is most useful, as per your choice of likelihood.