Statistical Thinking Comment Policy

Comments are welcomed. Be informed, informative, respectful, and criticize ideas, not people. Please read our Comment Policy before commenting.

Ian Smith • a year ago

Nice post!

I was a little bit surprised to not see Wald's sequential probability ratio test (SPRT) mentioned here. The SPRT extends frequentist likelihood-based methods to the fully sequential regime where analyses can be continuously monitored for as long as you like, and type-I error is controlled at arbitrary stopping times (no asymptotics required) [1, 2].

One need not derive any sampling distributions at all (let alone "extremely complicated" ones). The stopping rule is rather simple: reject the null if and when the likelihood ratio crosses 1/alpha. It also has power one (in the limit).

Going a little bit further: in this blog post it was mentioned that "this model is needed just as much by frequentist methods", but it's possible to avoid a likelihood altogether. Sequential estimation of quantiles can be done in a distribution-free manner [3], and means can be estimated in many nonparametric scenarios, e.g. if bounds on the random variable are known [4]. Again, non-asymptotic type-I error control at arbitrary stopping times.

As an aside for those interested in sequential frequentist-Bayesian connections: if we consider the frequentist regime but have a Bayesian working model, then it turns out that the posterior *evaluated at the true parameter* divided by the prior is an anytime-valid p-value (i.e. we can reject the null whenever the p-value drops below alpha). This paper of mine [5] applies it to sampling without replacement but Proposition 2.1 is intentionally stated generally. In other words, if you have a posterior distribution available, you can perform sequential, continuously monitored, nonasymptotic frequentist inference.

[1] Wald's 1945 paper on the SPRT: https://projecteuclid.org/j...

[2] SPRT applied to drug/vaccine safety monitoring https://www.tandfonline.com...

[3] Distribution-free sequential estimation of quantiles and CDFs https://arxiv.org/pdf/1906....

[4] Nonparametric sequential estimation of means https://projecteuclid.org/j...

[5] Sequential frequentist-Bayesian connections and sampling without replacement https://proceedings.neurips...

• Reply • Share ›

Frank Harrell Mod ➚ Ian Smith • a year ago

Thanks for the useful thoughts and references. I studied SPRT in graduate school but have not used it myself. I have a sense that it is conservative compared to some Bayesian approaches but haven't studied it enough to have full faith in that. The fact that frequentists won't use SPRT in randomized studies has put it at a disadvantage. Most frequentist statisticians are fixed sample size persons. The SPRT will not extend easily to complex situations involving multiple outcome variables whereas Bayes is natural for that.

On your point about model assumptions, I don't agree very much. Even methods that are apparently not model-based effectively use models. For example, Wilcoxon tests use the proportional odds model and permutation t-tests assume that the distribution is not heavy tailed. And when you don't have a model it's difficult to extend to censored and longitudinal data. Quantiles are nice for situations where both of these hold: (1) sample size is quite large and (2) distribution is completely continuous. So I don't see a wide role for nonparametric quantile-based methods.

The nonparametric mean method you referenced is a large sample method so is not interesting to me. The confidence sequence method is interesting and different from anything I've seen to date.

• Reply • Share ›

Ian Smith ➚ Frank Harrell • a year ago

Also,
"frequentists won't use SPRT in randomized studies"
I'm not sure how prevalent it is more generally, but it was at least part of the Janssen COVID-19 trial protocol [2]

[2] https://www.jnj.com/coronav...

• Reply • Share ›

Frank Harrell Mod ➚ Ian Smith • a year ago

Great to know this. Thanks for such a timely example.

• Reply • Share ›

Ian Smith ➚ Frank Harrell • a year ago

Thanks for the reply!

“I have a sense that it is conservative compared to some Bayesian approaches”
I’m not sure in what sense the SPRT would be considered conservative. It has optimality properties (e.g. minimizing expected stopping time).

“Even methods that are apparently not model-based effectively use models.”
I agree that this is sometimes true. The referenced methods hold under conditions which ideally are known a priori (e.g. bounds on random variables or their MGFs).

“The nonparametric mean method you referenced is a large sample method so is not interesting to me.”
I’m not sure what you mean by “large sample”. The referenced methods do not rely on asymptotics for validity. If you mean that they are not powerful in small samples, there is recent work trying to improve on it in the case of bounded random variables [1].

[1] https://arxiv.org/pdf/2010....

• Reply • Share ›

Frank Harrell Mod ↗ Ian Smith • a year ago

Interesting stuff - wish I had more time to do justice to it.

^ | v • Reply • Share ›

Ahmed • a year ago

Dear Prof Harrell,
I would like to ask you about the work here, http://hbiostat.org/proj/co...
I have now real data, do you think I can apply the same approach on it? I do not know how start really.
Thanks in advance.

^ | v • Reply • Share ›

Frank Harrell Mod ↗ Ahmed • a year ago

Ahmed that's such a wide-open question that I don't know where to start. My suggestion is to carefully study that document plus the first link under https://hbiostat.org/proj/c... . I'm sorry I'm not able to give specific advice for specific studies here.

^ | v • Reply • Share ›

Frank Harrell • 5 years ago

I can't disagree with that, though I'm not convinced of the relevance of a frequentist strategy that no one uses in practice and that hurts type II error.

^ | v • Reply • Share ›

Unknown • 5 years ago

I would disagree with your characterization of the frequentist approach to optional stopping. As I showed in "Almost sure hypothesis testing and a resolution of the Jeffreys-Lindley paradox", one can make the probability of incorrectly stopping an experiment arbitrarily small by selecting a sequence of significance levels that decrease to 0 at an appropriate rate. The frequentist only needs to bound the probability of incorrectly terminating the experiment.

^ | v • Reply • Share ›

Frank Harrell • 5 years ago

Hi Donald. Sorry to be slow in responding due to attending the ASA Statistical Inference conference. It was good to see you there also.

The experimental design and assumptions between studies can dictate in all statistical paradigms that the parameter is the same. For example, you can adjust for the subject's sex in the model as a covariate but assume that the sex regression coefficient is zero, meaning that you assume that mu is the same for females and males. You can also pool studies if you are fairly certain they are generated from the same mu. And with Bayes, it is not necessarily the case that there is fuzz in mu. Bayes is consistent with their being one true single-valued mu. It's just that we don't know what it is, and the less we know the more uncertain will be the final Bayes result (wider posterior distribution). I think of Bayes this way: We don't know what an experiments is throwing at the Bayesian analysis. The Bayesian model you specify should be able to handle a variety of possible values. If your theory dictates that some values are impossible (e.g., with blood pressure it can't be below about 30mmHg or the patient would not have lived long enough to enroll in the study), those values will be excluded in the formulation of the prior distribution to improve the final estimates. After such considerations, we want the Bayesian model to properly deal with whatever the experiment is throwing at it. We demonstrate that by simulating mu from a prior consistent with our prior beliefs/knowledge. Uncertainty is coming from absence of knowledge of the single true mu, not from any inherent fuzziness in the true mu that is generating the data at hand for one experimental design.

Another way to think about this is that in the frequentist world we waste an incredible amount of funding by designing studies to detect a single mu, where that detectable value of mu has been set with too little knowledge. In Bayesian sample size estimation, uncertainty around mu, through a prior distribution, goes into the Bayesian power calculation, resulting in much more honest sample sizes. With Bayes you can compute the expected sample size, the 0.95 quantile of the sample size, etc., because of admission of uncertainty about the true mu.

^ | v • Reply • Share ›

Frank Harrell • 5 years ago

For my situation the pertinent thing to emphasize would be the width of say a 0.95 credible interval. And you'll find that the posterior mean tracks the true mean (efficacy) no matter what the N. Feel free to modify the code to do any of this.

^ | v • Reply • Share ›

Donald Williams • 5 years ago

Hi:
The value differs because of sampling variability, but in an unbiased situation the average frequency should be the assumed point estimate. So, for any given iteration, the value (in this case the mean) will never be exactly the same;

will never be exactly the same.

```
mu_f <- replicate(1000, mean(rnorm(100, 0, 1)))
mean(mu_f)
# here values span from -.3 to .3.
```

In contrast, the Bayes assumption has this variability and incorporates the fact that we do not know the exact value for a treatment. However, we can still think one value exists, which is the assumption of Bayesian methods, but we also allow for uncertainty surrounding the "true" effect. This is different than suggesting many effect exist. Of course, I think the question is not which assumption is "true", but understanding the context under which assuming a set of assumptions allows for richer inference.

```
mu_b <- replicate(1000, mean(rnorm(100, rnorm(1, 0, .2), 1)))
mean(mu_b)
# here, the values span from < - 0.5 and > .5
```

In both cases, the returned mean is basically zero but the Bayes has captured uncertainty in the effect. This does not necessarily mean that we think the treatment effect is actually different, as indicated by the unbiasedness of the estimator.

• Reply • Share ›

Unknown • 5 years ago

Frank, in your simulations please calculate the efficacy, effect sized, and variances across the various Ns and plot them. What does a plot of efficacy against N look like with a dashed line of the actual efficacy in the population?

• Reply • Share ›

Frank Harrell • 5 years ago

Donald I'm not clear on 'the value also differs when assuming one value.' Perhaps this whole issue can become more clear by stating that whether or not we assume the same value of the parameter in different studies comes from the experimental design. If you had two independent replications of a study you could have pooled all the raw data together and estimated a single parameter whether Bayesian or frequentist. The bigger point is that describing uncertainty in known quantities is very conveniently done using probability distributions, and this also results in good solutions to everyday problems as well as to highly complex ones.

• Reply • Share ›

Donald Williams • 5 years ago

Interesting. I have done similar simulations, but not for optional stopping. Rather, looking at estimation in longitudinal models across hypothesized distributions. I describe it slightly differently, however. Sure, the value does differ when drawing from prior but the value also differs when assuming one value. The bayesian approach has two sources of uncertainty: 1) in the effect; 2) sampling variability. Frequentist only has sampling variability. Lastly, both approaches assume that there is a true value, but differ in how this is described. Bayes allows the "truth"-eg, effect of interest-to have a probabilty distribution that allows for incorporating uncertainty into the simulations.

• Reply • Share ›

Frank Harrell • 5 years ago

Hi Stephen - I think the simulation is more consistent with your world view than you think. The value of the parameter generating your data is fixed. The value doesn't need to change the next time you generate data. The use of a probability distribution behind the scene is a useful device to deal with "would that we only knew the value now." Parameters don't need to vary across time, cultures, etc. But for any dataset we have at present THE value is unknown and we find probabilities helpful for dealing with uncertainties.

If you wanted to use the same value of mu each time, you have to effectively force the prior to be degenerate, and the Bayesian posterior will behave exactly as you wish. There is no way to use a prior that is smooth but having the data be forced to be generated from only one value. I think these are very subtle points, and worthy of further discussion.

I don't see how you can be Bayesian and need to keep track of sampling intentions.

• Reply • Share ›

Stephen Martin • 5 years ago

Excellent post Frank. Although it still highlights a difference in philosophy about the DGP that I have.

This assumes that the generative parameter indeed varies according to the prior, as you state. Such that, mu ~ prior; y ~ mu. My issue with that still remains that it's assuming that the true parameter, in the universe, does vary according to some distribution. Whereas parameters may change over time or across subpopulations, I'm not sold on the idea that the generative parameter is indeed drawn from some probability distribution for any given collection effort. Likewise, if you go into the discrete world of things, we see a similar philosophy. E.g., BFs and p(H_k|y) are shown to be calibrated such that

of things, we see a similar philosophy. E.g., BFs and p(H_k|y) are shown to be calibrated such that when the BF is 6, then across repeated samples, there is indeed a 6:1 odds that the data were generated under H1 than under H0. But my issue with that is it assumes, in the universe, that H1 and H0 are simultaneously true, but vary in probability. E.g., 80% of the time H1 is true, and 20% of the time H0 is true. I think that's a weird statement to make about the universe --- That two hypotheses about how the universe works can simultaneously be true any given proportion of the time. It's drawing a truth statement from an urn.

The same notion is present in this simulation --- The state of truth randomly varies; ASSUMING this is true, then sure it is calibrated. But I think from a philosophy of science standpoint, it's a little weird. "Theta = .26 is true for the moment you are collecting data; but Theta = -1.2 is true for another moment you are collecting data".

Again, not to say parameters can't actually vary a bit over time, across cultures, across subpopulations (and for this, longitudinal, mixture, or random effects models can capture that tendency). But to say at any given moment the generative parameter is a totally random process is strange to me.

When you don't assume the above is true, when you instead assume a generative parameter is stable but unknown (but data are fixed, because we observed them), the probability of making decisions does change, as Kruschke, I, and many others have shown. E.g., if mu = 1.2 [which you know, because you are creating the universe at hand] and y ~ N(mu, sigma); but as an unknowing scientist you estimate a model: mu ~ prior; sigma ~ prior; then decision rates can change with sequential sampling. So it seems to me that the only way Bayesian models are calibrated, is if the true generative parameter can indeed change in the universe each time you sample from it, and that's difficult for me to digest. (Otoh, other Bayesians indeed don't assume that, and use a model that adjusts for sampling intentions in a fully Bayesian way through likelihood or prior modifications).

Related

- Bayesian vs. Frequentist Statements About Treatment Efficacy
- Statistical Errors in the Medical Literature
- My Journey From Frequentist to Bayesian Statistics
- Commentary on Improving Precision and Power in Randomized Trials for COVID-19 Treatments Using Covariate Adjustment, for Binary, Ordinal, and Time-to-Event Outcomes