Likelihood-based approaches to weighted data

Christopher A. Sims
Princeton University
sims@princeton.edu

August 23, 2018

Origins of this talk

- I noticed an example in Wasserman's *All of Statistics* that purported to show that any Bayesian approach to estimation in a particular simple, but high-dimensional model resulted in nonsense, while an easy-to-compute, intuitively appealing frequentist estimator was available.

Origins of this talk

- I noticed an example in Wasserman's *All of Statistics* that purported to show that any Bayesian approach to estimation in a particular simple, but high-dimensional model resulted in nonsense, while an easy-to-compute, intuitively appealing frequentist estimator was available.

- I explained in a paper published only on my web site ("Understanding Non-Bayesians") that Wasserman had made a mistake in writing the likelihood function and confused independence with exchangeability in setting the prior.

- Jamie Robins later joined Wasserman in arguing with me about this example on the internet, with each side of the argument claiming victory.

- Somewhat later, I had a student asking me how to write a likelihood function for a model of income distribution when the data were a weighted sample of individual incomes, and I realized that essentially the same issues were at play.

Basu's elephants

- After I initially presented these slides, at a conference honoring Gary Chamberlain, Roger Koenker pointed me to Basu's "elephants" example.

- The example criticizes the Horwitz-Thompson estimator based on ideas similar to those I lay out here. In the course of his career Basu moved toward a firmly Bayesian view of inference on survey sample data, and in the next (8/29/2018) presentation based on these slides there will be discussion both of his elephants example (if the audience does not already know it) and the extent to which his work anticipated what I present here.

Objectives of the talk

- I try to tie together several kinds of weighted-data situations in a Bayesian non-parametric (i.e., infinite-dimensional parameter space) settings.

Objectives of the talk

- I try to tie together several kinds of weighted-data situations in a Bayesian non-parametric (i.e., infinite-dimensional parameter space) settings.

- There is a risk that everything I say is obvious.

Objectives of the talk

- I try to tie together several kinds of weighted-data situations in a Bayesian non-parametric (i.e., infinite-dimensional parameter space) settings.

- There is a risk that everything I say is obvious.

- However, Robins and Wasserman are smart guys, and it's not clear even now that they share my view of the issues, so I thought discussing the issues might be worthwhile.

- The issues are not just "philosophical". They have implications for how one deals with weighting, randomization, and survey data generally.

Easy low-dimensional special case

- Small number of groups $i = 1, \ldots, M$.
- Fairly large number of observations n_i in each group.
- We know the proportion π_i of each group in the population.
- We observe $y_{ij}, j = 1, \ldots, n_i$ in each group, y_{ij}'s independent and, within each group i.i.d., mean μ_i.
- We are interested in the population mean μ of y_{ij}, i.e. $\sum_i \mu_i \pi_i$.

Easy low-dimensional special case

It's obvious what to do here: Take group sample means $\bar{y}_i$, Estimate μ as $1/M \sum_i \pi_i \bar{y}_i$. If the variances of the y_{ij}'s are finite we can assume normality and get a fairly straightforward frequentist distribution theory for this estimator, and this distribution gets simpler and frees itself of the normality assumption as the samples sizes n_i all get larger.

We could be fully Bayesian here, assuming normality, say putting the same highly dispersed conjugate prior, independent across i, on μ_i. But this would be pedantic. The priors would be dominated by the likelihood and produce a posterior mean for μ nearly the same as the straightforward weighted-sum-of-sample-means estimator.

Introducing selection

- Suppose that the sample has been subject to selection. Each draw is in group i with known probability π_i, but then is discarded with probability $1 - p_i$.

- We know p_i for each group.

- The distribution of non-discarded draws across groups gives group i probability

$$\frac{\pi_i p_i}{\sum_j \pi_j p_j}.$$

Another simple estimator

What about estimating μ as

$$\frac{1}{\sum_i n_i} \sum_{i,j} \frac{y_{ij}}{p_i} \cdot \sum_i p_i \pi_i \, ?$$

$$E\left[\frac{y_{ij}}{p_i}\right] = \sum_i \frac{\mu_i}{p_i} \frac{p_i \pi_i}{\sum_j \pi_j p_j} = \frac{\sum_i \mu_i \pi_i}{\sum_j p_j \pi_j} = \frac{\mu}{\sum_i p_i \pi_i}$$

or estimating μ as

$$\frac{1}{N} \sum_{i,j} \frac{y_{ij}}{p_i} \, ,$$

where N is the total number of draws, including those for which y was not observed. $E[n_i/N] = p_i \pi_i$, and n_i is independent of y_i conditional on p_i. So this is also unbiased for μ. Doesn't require we know $\sum p_i \pi_i$.

These simple estimators aren't so good

- They let the randomness in the number of observations per group influence the weighting of group means.
- Since we've assumed we know the correct weights, we're better off using that information.
- When we specialize to the case where y_{ij} is always zero or 1, so $\mu_i = P[y_{ij} = 1]$, the last of these simple estimators, presuming knowledge of N, is the Horwitz-Thompson estimator that Wasserman (and Robins) hold out as the simple frequentist estimator that could never be arrived at by a Bayesian.

Many groups

- Now suppose the number of groups, M, is large. So large, in fact, that many groups have very few observations, or even have no observations at all. In the extreme, every observation comes from its own group, with associated p_i and μ_i, and most groups do not appear in the sample.

- Here the straightforward approach we first proposed does not work. That relied on the idea that each group's n_i was large enough that likelihood dominated prior.

- Also, we suggested a prior on the μ_i's that made them i.i.d. across groups. Since now there is a vast number of groups not appearing in the sample, and since the μ_i's for those groups therefore do not appear in the likelihood, the posterior is the prior for all these μ_i's.

Averaging y_i/p_i with many groups

- The argument for unbiasedness of the two estimators based on averaging y_i/p_i did not depend on the number of observations per group, or on all groups being represented in the sample.

- On the other hand, those estimators might still not be very good.

- For example, suppose the observed y_i's all lie in a narrow range, say 10 ± 0.1, but the $1/p_i$ values vary over a wide range.

- Then we might be quite confident that we know the mean value of y for the population with high precision, but treating y_i/p_i's mean and sample variance as characterizing our uncertainty about μ would not reflect this confidence.

A nonparametric Bayesian approach

Getting the likelihood and prior right.

- We must recognize that the p_i values in the sample are observed realizations of a random variable, despite the fact that by assumption we know in advance the value of p at each point in the probability space. That we know this is no more than stating we know the distribution for the draws of p_i,

- Also, it makes no sense to assume a fixed prior distribution, for each μ_i, i.i.d. across draws. We believe we can learn about μ from the sample, which implies the observations contain information about the μ_i's for the many unobserved groups.

- One way to do this is to make the prior exchangeable, rather than independent, across draws, with μ an unknown parameter of the exchangeable distribution.

Likelihood and prior

We assume $\int yq(y \mid \mu)\, dy = \mu$. We also start using the notation $\bar{\mu} = \int \mu h(\mu \mid p)\pi(p)\, d\mu\, dp$, i.e. the population mean of μ. What we observe is either just (y_i, p_i) (in case unselected draws are just discarded), or $(y_i\delta_i, p_i, \delta_i)$, where δ_i is an indicator for whether we see y_i, if we don't discard the observations where y is not seen.

$$(p\, q(y \mid \mu)h(\mu \mid p))^{\delta}(1-p)^{1-\delta}\pi(p) \quad \text{or}$$

$$q(y \mid \mu)h(\mu \mid p)\frac{p\pi(p)}{\int p\pi(p)\, dp}\,.$$

This is a likelihood with an infinite-dimensional unknown "parameter" $h(\cdot \mid \cdot)$

How to proceed along this line in practice?

- For example put a distribution on h that makes it lie in a union of finite-dimensional spaces where $E[\mu \mid p]$ is constant over small intervals of p values, the number of intervals growing and the lengths of the intervals shrinking as we go down the sequence of spaces.

- This is likely to lead to inference that is close to converting the problem to the original simple form, with a finite number of observations per group, though with a systematic rule for moving the focus of attention to higher-dimensional spaces as data accumulates.

Objections to this approach

- Since the parameter space is infinite dimensional, it is likely to be possible to choose forms of $h(\cdot \mid \cdot)$ for which posterior does not converge to the truth, or a sequence of such forms for which convergence is arbitrarily slow.

- But this is almost inevitable in an infinite-dimensional space. Essentially the same problem arises for frequentist asymptotics that provides limiting distributions, rather than just consistency.

Regularity conditions on infinite-dimensional parameter spaces

- Frequentist arguments require conditions on the tail behavior of the distributions that make the parameter space a topologically small subspace of the natural one.

- That is, the regularity conditions assert a priori that many possible $h(\cdot \mid \cdot)$ functions that would be hard to distinguish from the true one based on observed data, are impossible.

- Bayesian priors do the same sort of thing.

Could a Bayesian end up using something like Horwitz-Thompson?

- Yes. There is nothing particularly frequentist or Bayesian about ignoring information.

- If we assume y/p is i.i.d. with mean μ and finite variance, we can do the usual thing, applying the distribution theory as if y/p were normal, which is asymptotically robust under mild assumptions.

- In the original Wasserman example, μ was bounded between 0 and 1, with non-zero probability on the point $y/p = 0$. Something very close to the Horwitz-Thompson estimator then emerges if one looks for a distribution with this support that has the mean of y/p as a sufficient statistic.

Conclusion

- Practical cases can easily lie between these extremes.
- E.g., we might put some non-zero probability on the possibility that the weights are independent of the μ's.
- Or we might have a sample with many heavily-populated groups, but also some with very few observations. The lesson of the likelihood approach is that it can be useful to use a prior/model that allows for drawing information about sparsely populated groups from more heavily populated groups.