Bayesian Analysis of Cluster Sampling

Susanna Makela

September 23, 2016

Introduction

One basic goal of survey analysis is making inferences about the finite population from which the sample was drawn, such as means and totals for particular outcomes of interest. We consider the analysis of a survey with a two-stage cluster sampling design, where the first stage involves probability proportional to size (PPS) sampling. Two-stage survey designs are common when generating a sample frame of every unit in the population is infeasible or impractical. For example, in designing a nationally representative household survey, generating a complete listing of every household in the country requires essentially as much effort as a complete census of all households. Instead, the sampling proceeds in stages, first sampling primary sampling units (PSUs) such as counties, cities, or census tracts. The PSUs are sampled with probability proportional to a measure of size, which is commonly the number of secondary units in the PSU but can be a more general measure of size, such as annual revenue or agricultural yield. Secondary sampling units (SSUs) are then sampled within selected PSUs. This design requires a complete listing of PSUs and a complete listing of units only within selected PSUs. Often a fixed number of SSUs are sampled within each selected PSU, yielding what is known as a self-weighting survey design, where each SSU has the same probability of being selected.[1] Because the SSUs are naturally clustered within PSUs, we refer to the PSUs as clusters and the SSUs as simply units.

Unlike other areas of statistics, inference in survey sampling is often design-based, meaning it is based on the randomization distribution of whether a unit is included in the sample instead of directly modeling the survey outcomes themselves (Little, 2004). However, a Bayesian approach to survey inference has many advantages over the design-based approach, including the ability to handle complex design features like clustering, better inference for small-sample problems like small area estimation, incorporation of prior information, and large-sample efficiency (Little, 2004).

Bayesian inference of sample surveys considers both the survey inclusion indicators I and the outcomes Y to be random. Following Gelman et al. (2013), we can decompose the outcome Y into $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} represents the values of Y among the sampled units with $I = 1$ and Y_{mis} represents the values for the nonsampled units with $I = 0$. We denote by X the set of variables determining the sampling mechanism – that

[1] To see this, let M_j represent the measure of size for cluster j and $M = \sum_{j=1}^{J} M_j$ represent the total size in the population. If we sample K clusters and a fixed number of units n in each selected cluster, the probability π_{ij} of unit i in cluster j being included in the sample is $\pi_{ij} = (KM_j/M)(n/M_j) = Kn/M$, which is constant across units i and clusters j.

is, the variables on which the inclusion indicator I depends – and other variables known to be related to Y. In the case of PPS cluster sampling described above, the relevant design variables are the set of cluster sizes M_j for the J clusters in the population.[2] The complete-data likelihood is given by

$$p(Y, I|X, \theta, \phi) = p(Y|X, \theta)p(I|Y, X, \phi),$$

where θ and ϕ are parameters for the distribution of the outcomes and inclusion indicators, respectively.

The posterior distribution of θ in terms of the observed data (Y_{obs}, X, I) is[3]

$$\begin{aligned} p(\theta|Y_{obs}, X, I) &\propto p(\theta|X) \int\int p(\phi|X)p(Y, I|X, \theta, \phi)dY_{mis}d\phi \\ &= p(\theta|X) \int\int p(\phi|X)p(Y|X, \theta)p(I|Y, X, \phi)dY_{mis}d\phi. \end{aligned}$$

If the data collection process is ignorable, then $p(I|Y, X, \phi) = p(I|Y_{obs}, X, \phi)$, meaning that the distribution of the inclusion indicators depends only on observed data (and parameters). In this case, the posterior is

$$\begin{aligned} p(\theta|Y_{obs}, X) &\propto p(\theta|X) \int p(Y|X, \theta)dY_{mis} \int p(\phi|X)p(I|Y_{obs}, X, \phi)d\phi & (1) \\ &\propto p(\theta|X) \int p(Y|X, \theta)dY_{mis} & (2) \\ &= p(\theta|X)p(Y_{obs}|X, \theta) & (3) \end{aligned}$$

since the integral over ϕ is constant with respect to θ. As Gelman et al. (2013) point out, "the posterior distribution of θ and the posterior predictive distribution of Y_{mis} ... are entirely determined by the specification of a data model – that is, $p(Y|X, \theta)p(\theta|X)$ – and the observed values Y_{obs}." While we can never conclusively prove that we have achieved ignorability, its plausibility is strengthened when we include in X all covariates that are

Our interest here is in the finite population mean, given by $Q(Y) = \overline{Y}$, where N is the total number of units in the population. We can write the posterior predictive distribution of $Q(Y)$ given the observed data Y_{obs} as

$$\begin{aligned} p(Q(Y)|Y_{obs}, X) &= \int p(Q(Y), Y_{mis}|Y_{obs}, X)dY_{mis} \\ &= \int p(Q(Y)|Y_{mis}, Y_{obs})p(Y_{mis}|Y_{obs}, X)dY_{mis}, \end{aligned}$$

where $p(Y_{mis}|Y_{obs}, X)$ is the posterior predictive distribution of Y for the nonsampled units, given by

$$p(Y_{mis}|Y_{obs}, X) = \int p(Y_{mis}|\theta, X)p(\theta|Y_{obs}, X)d\theta,$$

[2]Even in the case of a self-weighting sample, these variables are still important because 1) outcomes for units in the same cluster are likely to be correlated and 2) outcomes are often correlated with the measures of size.

[3]Here we make the implicit assumption that the parameters ϕ and θ are distinct, meaning $p(\phi|X, \theta) = p(\phi|X)$ (Gelman et al., 2013).

However, in many (arguably most) practical situations, the set of design variables X is not known for the entire population and is often instead known only for sampled clusters or units. In this case, we need to model the nonsampled values of X before we can make use of the posterior predictive distributions above for inference about $Q(Y)$. For the case of two-stage PPS cluster sampling, we need to model the measures of sizes M_j for the nonsampled clusters.

Existing Bayesian approaches to this problem (Zangeneh et al., 2011; Zangeneh and Little, 2015) separate estimation of the missing cluster sizes and inference for the finite population quantities into two steps and only consider the case of single-stage PPS samples. In contrast, we integrate these steps into one model for a two-stage cluster sample.

Model

We consider a population consisting of $J = 1000$ clusters with measure of size M_j ranging from 100 to 1000 under a data generating model given by

$$\begin{aligned} Y_i &\sim N(\beta_{0j[i]} + \beta_{1j[i]} w_i, \sigma_y^2) \\ \beta_{0j} &\sim N(\gamma_0 + \gamma_1 \log(M_j), \sigma_{\beta_0}^2) \\ \beta_{1j} &\sim N(\alpha_0 + \alpha_1 \log(M_j), \sigma_{\beta_1}^2), \end{aligned}$$

where Y_i is the outcome of interest for unit i, $j[i]$ denotes the cluster j to which unit i belongs, and w_i is a unit-level covariate.[4] A sample of $K < J$ clusters is taken with probability proportional to M_j, and n units are sampled via simple random sampling in each selected cluster. We assume that we know M_j for the K sampled clusters, w_i for every sampled unit and $\overline{w}_j$ for every cluster, and the total measure of size $M = \sum_{j=1}^{J} M_j$. We assume we do not know M_j for nonsampled clusters.

Computational coherence check

To ensure our simulation code is working as expected, we follow the approach of Cook et al. (2006) and conduct a computational coherence check. We repeatedly draw the hyperparameters $\gamma_0, \gamma_1, \alpha_0, \alpha_1, \sigma_y, \sigma_{\beta_0}$, and σ_{β_1} from their priors, simulate population data with the model above, and then fit the model to the full population data (without sampling) using Stan (Stan Development Team, 2015). If our computational procedure is correct, we should be able to recover the true values of the parameters and see approximately 50% of the 50% posterior intervals and 95% of the 95% posterior intervals contain the true parameter values.

In figure 1, we see that the coverage of the 50% and 95% posterior intervals is close to the nominal levels for all the parameters. The point estimates also recover the true values, as seen in figure 2. Here we plot histograms of the posterior means of the hyperparameters across 500 simulations. Because the true values vary across simulations, we plot the differences between the posterior means and the truth; values closer to zero thus indicate more accurate estimates. The parameters in the top row are all drawn from $N(0, 1)$ prior

[4] We can write the population mean $\overline{Y}$ in terms of the cluster-level means $\overline{Y}_j$. When Y_i is normally distributed, we can make inferences for $\overline{Y}_j$ based on $\overline{w}_j$, so in this case it's not necessary to know w_i for each unit in the population. In practice, if w is a demographic covariate, it's often the case that we know demographic characteristics of clusters even if we don't know the overall size.

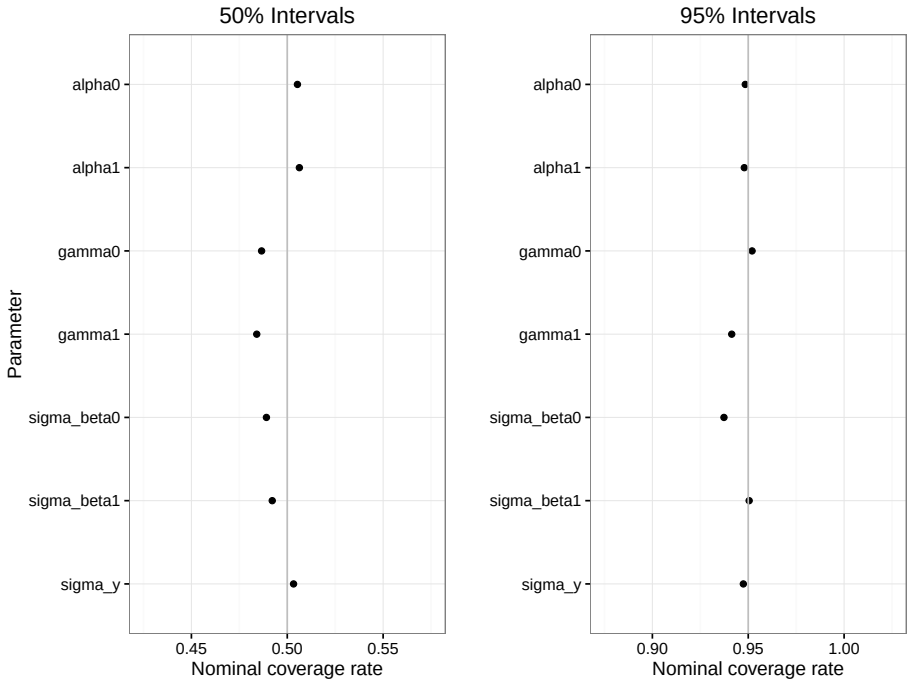

Figure 1: Coverage rates from computational coherence check. Each dot represents the coverage rate of a parameter based on 500 populations generated from repeated draws of the hyperparameters from their prior distributions. The left panel shows coverage rates for 50% posterior intervals and the right panel for 95% posterior intervals.

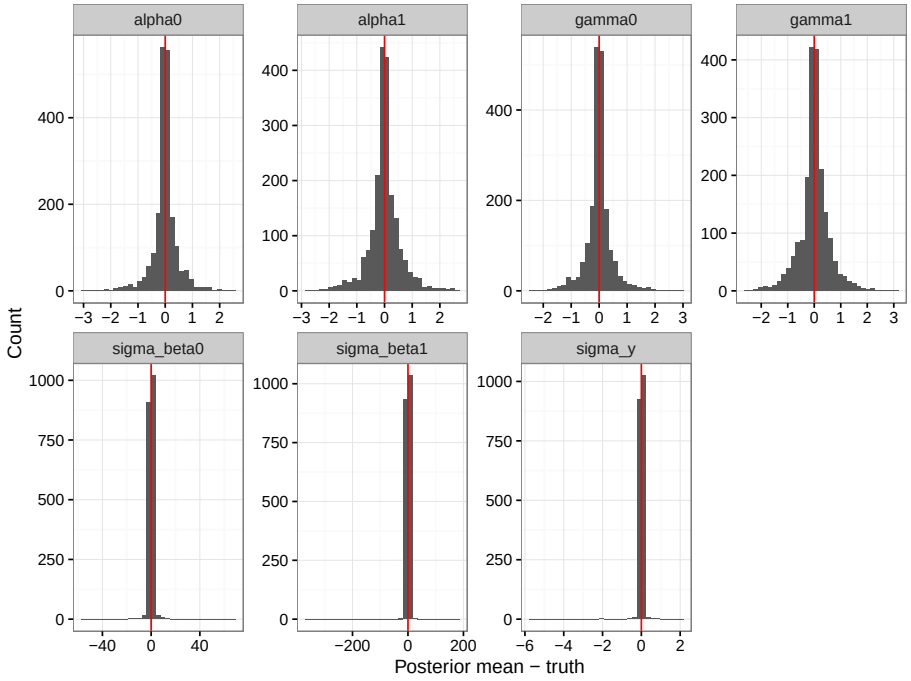

Figure 2: Histograms of posterior means for the hyperparameters across 500 simulations. Because the true hyperparameter values vary across simulations, we plot the difference between the posterior mean and the true value.

distributions, so the shape of the histograms makes sense. The variance parameters in the bottom row are drawn from Half-Cauchy$(0, 2.5)$ prior distributions, so the occasional extreme outliers we observe are to be expected. Recovering the true parameter values with expected confidence coverage rates indicates that our computational procedure is working as expected.

Bayesian bootstrap

One method for predicting unobserved cluster sizes is the modified Bayesian bootstrap of Zangeneh and Little (2015), based on earlier work by Little and Zheng (2007). Let $w_1, \ldots, w_B$ represent the B unique values of the measures of size M_j observed in a sample of K clusters out of a population of J clusters. Let $n_1, \ldots, n_B$ represent the corresponding counts of these unique sizes, so that $\sum_b n_b = K$. We can then model the counts $n = (n_1, \ldots, n_B)$ as multinomially distributed with parameters $\psi = (\psi_1, \ldots, \psi_B)$. The ψ's are given a noninformative Haldane prior, so $p(\psi_1, \ldots, \psi_B) = Dir(0, \ldots, 0)$. The posterior distribution is then

$$p(\psi_1, \ldots, \psi_B | n_1, \ldots, n_B) = Dir(n_1, \ldots, n_B),$$

and the counts of the unique sizes for the $J - K$ nonsampled clusters, $n^* = (n_1^*, \ldots, n_B^*)$, can be drawn from the posterior predictive distribution

$$\begin{aligned} p(n^*|n) &= \int p(n^*|\psi)p(\psi|n)d\psi \\ &= \int \text{Multin}(n^*|J - K, \psi)\text{Dir}(\psi|n)d\psi. \end{aligned}$$

Directly drawing from this distribution, however, would not account for the PPS sampling design. To adjust for this, Little and Zheng (2007) draw values of ψ from their posterior $\text{Dir}(\psi|n)$ and then replace ψ_b with

$$\psi_b^* = \frac{c\psi_b(1 - \pi_b)}{\pi_b},$$

where

$$\pi_b = \frac{KM_b}{M}$$

is the selection probability for units with measure of size M_b and c is a constant ensuring that $\sum_{b=1}^{B} \psi_b = 1$. This approach essentially adjusts the probability of resampling an observed size M_b by the odds of a cluster of that size not being sampled, so smaller sizes are upweighted relative to larger ones. However, this approach restricts the draws for the nonsampled cluster sizes to come from the set of observed cluster sizes.

Figure 3 shows density plots of the estimated finite population mean $\overline{Y}$ across 2,000 simulations. Because true value of $\overline{Y}$ varies across simulations, we again plot the difference between the estimated value and the truth, and the panels are for scenarios where we sample 20, 50, and 100 units per sampled cluster. As we would hope, the densities are sharply peaked around zero. The coverage rates for the 50% and 95% posterior intervals shown in Table are close to nominal rates, with 50% intervals slightly more conservative than the 95% ones.

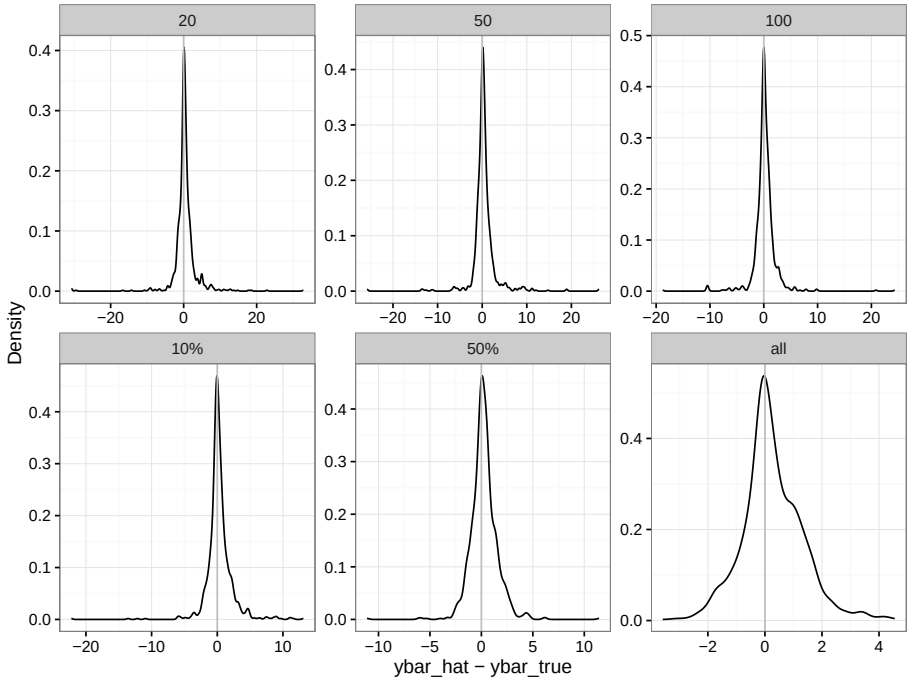

Figure 3: Density plots of the estimated finite population mean across 2,000 simulations. Because true value of $\overline{Y}$ varies across simulations, we again plot the difference between the estimated value and the truth. The panels are for scenarios where we sample 20, 50, and 100 units per sampled cluster.

Units per cluster	50% interval	95% interval
20	0.61	0.93
50	0.65	0.97
100	0.66	0.98

Table 1: Coverage rates of 50% and 95% posterior intervals for the finite population mean, calculated across 2,000 simulations.

A future approach: the negative binomial

An alternative to the Bayesian bootstrap is to take a superpopulation approach to estimating the nonsampled cluster sizes. Because the clusters are sampled PPS, we know that the sizes we observe in the sample are biased toward the larger sizes in the population. G. P. Patil (1978) consider a random variable X with pdf $f(x)$ and a probability $w(x)$ of recording the observation $X = x$ and give the density of X^w, the recorded observation, as

$$f^w(x) = \frac{w(x)f(x)}{\omega},$$

where ω is the normalizing constant. They define the case where $w(x) = x$ as the size-biased distribution. We can see how this distribution is derived in the case of PPS sampling, where the probability of sampling a cluster of size M_j is cM_j, where c is a normalizing constant. The probability of a cluster of size M_j occurring in the superpopulation is $f(M_j)$, so the overall probability of observing a cluster of size M_j in the PPS

sample is proportional to $M_j f(M_j)$ and equal to

$$\frac{M_j f(M_j)}{\sum_{M'} M' f(M')},$$

where in this case the denominator is just the expected value of M'.

One reasonable superpopulation distribution for cluster sizes is the negative binomial distribution, a count model that allows for overdispersion. If the superpopulation distribution is $NB(k, p)$ so that

$$Pr(X = x|k, p) = \binom{k + x - 1}{x}(1 - p)^x p^k,$$

then the size-biased distribution of X^w works out to $p(X^w) \sim 1+NB(k+1, p)$ (G. P. Patil, 1978).

The advantage of this approach is that it allows for cluster sizes not observed in the sample to be estimated for the nonsampled units.

References

Samantha R Cook, Andrew Gelman, and Donald B Rubin. Validation of software for bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692, 2006. doi: 10.1198/106186006X136976. URL `http://dx.doi.org/10.1198/106186006X136976`.

C. R. Rao G. P. Patil. Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, 34(2):179–189, 1978. ISSN 0006341X, 15410420. URL `http://www.jstor.org/stable/2530008`.

Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian Data Analysis, Third Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, London, third edition, November 2013. ISBN 1439840954.

Roderick J. Little. To Model or Not To Model? Competing Modes of Inference for Finite Population Sampling. *Journal of the American Statistical Association*, 99:546–556, January 2004. URL `https://ideas.repec.org/a/bes/jnlasa/v99y2004p546-556.html`.

Roderick J.A. Little and H Zheng. The Bayesian approach to the analysis of finite population surveys. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 8*, pages 283–302 (with discussion and rejoinder). Oxford University Press, Oxford, 2007.

Stan Development Team. Stan: A c++ library for probability and sampling, version 2.8.0, 2015. URL `http://mc-stan.org/`.

Sahar Z. Zangeneh and Roderick J. A. Little. Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample. *Journal of Survey Statistics and Methodology*, 3(2):162–192, 2015. doi: 10.1093/jssam/smv002. URL `http://jssam.oxfordjournals.org/content/3/2/162.abstract`.

Sahar Z. Zangeneh, Robert W. Keener, and Roderick J. A. Little. Bayesian nonparametric estimation of finite population quantities in absence of design information on nonsampled units. In *JSM Proceedings. Section on Survey Research Methods. Miami Beach, FL, USA. American Statistical Association-IMS*, pages 3429–3440, 2011.