STATISTICS IN TRANSITION new series, Summer 2012
Vol.13, No. 2, pp. 235-242

FINITE POPULATION SAMPLING: A MODEL-DESIGN SYNTHESIS

Malay Ghosh[1]

ABSTRACT

The paper considers a general class of Bayes estimators for estimating the finite population mean which also achieve design consistency. Some exact results are given where Bayes estimators agree with the Horvitz-Thompson or ratio estimators. For a wider class of priors, asymptotic mathematical equivalence of Bayes estimators with the above estimators is provided.

1. Introduction

There are primarily two basic approaches towards inference from sample survey data. The first, the design-based approach, finds estimators of population quantities based on probability distributions generated by a given selection mechanism. In contrast, a model-based approach assumes the population units to be generated from some superpopulation, and the assumed superpopulation model governs any subsequent inference.

There has been a long-standing debate among survey statisticians regarding which one of the two is the preferred inferential approach. The advocates of design-based methods often criticize model-based inference regarding its failure to guard against any possible model misspecification. On the other hand, those advocating the use of models, question the ability of design-based methods to provide inference with sufficient accuracy in the face of small sample sizes, and often in the neeeded justification of large sample approximations for small or moderate samples. Fortunately, in these days, one notices occasional reconciliation of these two approaches (see e.g. Sarndal, 1984; Prasad and Rao, 1999, among others).

While the basic conceptual disagreement between the two approaches cannot be resolved, from an operational point of view, it is often possible

[1]University of Florida.

to find an agreeement between the two. The present article is a modest attempt to provide some general results showing either exact or large sample agreement. We will illustrate our procedures with several examples.

Section 2 of this paper gives some general results showing model-based interpretation of some of the classical design-based estimators including the celebrated Horvitz-Thompson and ratio estimators. In the process, we revisit the one parameter exponential family model in a slightly non-conventional framework. We show that the said estimators plus others can be deduced as special cases of a general expression for the posterior mean under a certain diffuse prior. In Section 3, we continue with the exponential family model as considered in Section 2, and establish design consistency of Bayes estimators of the finite population mean for a wide class of priors. Here, design consistency is defined in the sense of Lahiri and Mukherjee (2007), and will be made precise in Section 3. Lahiri and Mukherjee concluded that a "subjective Bayes estimator is, in general, not design consistent". They also provided an adjustment to the Bayes estimator to achieve design consistency. While not refuting the word "general" in the statement of these authors, we will show in Section 3 that it is sometimes possible to achieve design consistency exactly in the same sense as of Lahiri and Mukherjee (2007) for a general class of heavy-tailed priors, without seeking any adjustment. We will also point out in this section why the adjustment was needed in their Bayesian framework. Some final remarks are made in Section 4.

2. Some exact results

Consider a finite polpulation with units labelled $1, \ldots, N$. Associated with these units are the characteristics of interest denoted by $y_1, \ldots, y_N$. A sample s of fixed size n is drawn from the population. We will denote by $y(s)$ the set of y_i such that $i \in s$. Similarly, we denote by $\bar{s}$ the set of unsampled population units, and by $y(\bar{s})$ the set of y_j such that $j \in \bar{s}$. The objective is to estimate the finite population mean $m(y) = N^{-1} \sum_{i=1}^{N} y_i$.

Under simple random sampling without replacement, the standard design unbiased estimator of $m(y)$ is the sample mean $\bar{y}_s = \sum_{i \in s} y_i/n$. More generally, with unequal probability sampling, the most well-used estimator of $m(y)$ is

$$N^{-1} \sum_{i \in s} y_i/\pi_i,$$

the Horvitz-Thompson estimator, where π_i denotes the probability of selecting unit i, $i = 1, \ldots, N$. Clearly one must have $\sum_{i=1}^{N} \pi_i = E[\sum_{i=1}^{N} I_{[s \ni i]}] = n$, I denoting the usual indicator function.

With auxiliary information x_i available with the y_i, the well-known ratio estimator of

$$m(y) = [(\sum_{i\in s} y_i)/(\sum_{i\in s} x_i)] \sum_{i=1}^{N} x_i/N.$$

Many other alternative estimators of $m(y)$ have been proposed including an estimator of Hajek (1971), and the celebrated "generalized regression estimator" of Sarndal, Swensson and Wretman (1992), but they will not be considered here.

We provide in this section model-based interpretation of some of the well-known design-based estimators including the Horvitz-Thompson and ratio estimators. It is convenient to begin with a version of the one paramater exponential family model, and obtain the posterior mean of $m(y)$ under a diffuse prior. To this end, we prove the following theorem.

Theorem 1. Suppose $y_i|\theta$ are independently distributed with pdf's $f(y_i|\theta) = \exp[(\theta y_i - a_i\psi(\theta))/\sigma_i^2 + h(y_i)]$, $i = 1, \ldots, N$. Here, θ is an unknown parameter, but the a_i and σ_i^2 are known constants. Consider the prior $\pi(\theta) = c$. Then

$$E[m(y)|y(s)] = N^{-1}[\sum_{i\in s} y_i + (\sum_{i\in s} y_i\sigma_i^{-2}/\sum_{i\in s} a_i\sigma_i^{-2}) \sum_{j\in \bar{s}} a_j]. \quad (1)$$

Proof. First note that solving $E[(d\log f(y_i|\theta)/d\theta)|\theta] = 0$ (the first Bartlett identity), one gets $E(y_i|\theta_i) = a_i\psi'(\theta)$. The posterior

$$\pi(\theta|y(s)) \propto \exp[\theta \sum_{i\in s} y_i\sigma_i^{-2} - \psi(\theta) \sum_{i\in s} a_i\sigma_i^{-2}].$$

Now, by the Bayesian analog of the first Bartlett identity, namely,

$$E[(d\log\pi(\theta|y(s))/d\theta)|y(s)] = 0,$$

one gets

$$E[\psi'(\theta)|y(s)] = \sum_{i\in s} y_i\sigma_i^{-2}/\sum_{i\in s} a_i\sigma_i^{-2}. \quad (2)$$

Next, observe that

$$E[m(y)|y(s)] = N^{-1}[\sum_{i\in s} y_i + \sum_{j\in \bar{s}} E(y_j|y(s))]. \quad (3)$$

Now, noting that for a given $j \in \bar{s}$, $E[y_j|y(s)] = EE[\{y_j|\theta, y(s)\}|y(s)] = a_jE[\psi'(\theta)|y(s)]$, one gets (1) from (2) and (3).

As mentioned earlier, some of the well-known design-based estimators can be derived as special cases of the above result.

Example 1. Let $a_i = \pi_i$ and $\sigma_i^2 = \pi_i/(1-\pi_i)$, where we may recall that π_i is the selection probability of the ith unit and $\sum_{i=1}^N \pi_i = n$. In this case from Theorem 1, $E[m(y)|y(s)]$ simplifies to

$$E[m(y)|y(s)] = N^{-1}[\sum_{i\in s} y_i + \{\sum_{i\in s}((1-\pi_i)/\pi_i)y_i / \sum_{i\in s}(1-\pi_i)\} \sum_{j\in \bar{s}} \pi_j. \quad (4)$$

Since $\sum_{j\in\bar{s}} \pi_j = \sum_{i=1}^N \pi_i - \sum_{i\in s}\pi_i = n - \sum_{i\in s}\pi_i = \sum_{i\in s}(1-\pi_i)$, from (4), one gets $E[m(y)|y(s)] = N^{-1}\sum_{i\in s} y_i/\pi_i$, which is the celebrated Horvitz-Thompson estimator.

Remark 1. Little (2004) gave an asymptotic model-based interpretation of the Horvitz-Thompson estimator. Ghosh and Sinha (1989) provided an exact model-based justification, but restricted only to the normal model. The result is established now under broader generality, and the present result is believed to be new.

Example 2. Suppose now $a_i = x_i$ and $\sigma_i^2 = 1$. Then, from Theorem 1

$$E[m(y)|y(s)] = N^{-1}[\sum_{i\in s} y_i + (\sum_{i\in s} y_i / \sum_{i\in s} x_i) \sum_{j\in\bar{s}} x_j] =$$

$$= (\sum_{i\in s} y_i / \sum_{i\in s} x_i) \sum_{i=1}^N x_i/N,$$

the well-known ratio estimator.

Example 3. Suppose now $a_i = \sigma_i^2 = x_i$. Then one gets $E[m(y)|y(s)] = N^{-1}\sum_{i\in s} y_i + n^{-1}(\sum_{i\in s} y_i/x_i)\sum_{j\in\bar{s}} x_j$, an example originally considered in Royall (1970). Basu (1971) gave a very intuitive justification of this estimator.

Remark 2. With the available auxiliary information x_i, various choices $a_i = h(x_i)$ and $\sigma_i^2 = v(x_i)$ will produce different model-based estimators of the finite population mean which could potentially be useful also in design-based analysis.

Remark 3. Although phrased in a Bayesian framework, one can think of an alternate model-based interpretation of the result of Theorem 1. To see this, we may note that $E(y_i|\theta) = a_i\psi'(\theta)$ and $V(y_i|\theta) = a_i\sigma_i^{-2}\psi''(\theta)$. The latter follows from the second Bartlett identity $E[\{dlogf(y_i|\theta)/d\theta)\}^2|\theta] = E[(-d^2logf(y_i|\theta)/d\theta^2)|\theta]$.

Then, under the standard quasi-likelihood approach, one obtains the unbiased estimating equation $\sum_{i\in s}\{y_i - E(y_i|\theta_i)\}/V(y_i|\theta_i) = 0$, which is equivalent to $E[\sum_{i\in s}\{y_i - a_i\psi'(\theta)\}^2/(a_i\sigma_i^{-2}) = 0$. This leads to the same estimator of $\psi'(\theta)$ as given in Theorem 1.

Remark 4. The special case $a_i = 1$ for all i and $\sigma_i^2 = \sigma^2$ for all i is of interest. In this case $E[m(y)|y(s)]$ simplifies to $N^{-1}[\sum_{i\in s} y_i + n^{-1}(N - n)\sum_{i\in s} y_i] = n^{-1}\sum_{i\in s} y_i$, the standard design-based estimator of the finite population mean under simple random sampling without replacement. We will revisit this point again in Section 3.

3. Some asymptotic results

The exact results of the previous section require a flat prior for θ. However, design consistency of model-based estimators in an asymptotic sense can often be justified for a wide class of priors. We will show in this section how mathematical limits of certain Bayes estimators result in standard design-based estimators. We use the term "mathematical limit" in the sense of Lahiri and Mukherjee (2007), where the limiting operation is performed in the sense of ordinary calculus, keeping the observations fixed. The latter came to the conclusion that as the sample size goes to infinity, mathematical limits of subjective Bayes estimators of the finite population mean based on the one parameter exponential superpopulation family do not converge in general to the Horvitz-Thompson estimator. They also proposed an adjustment to their subjective Bayes estimators to achieve design consistency. What we show is that a slightly modified version of the one parameter exponential superpopulation family as considered in (1) can indeed lead to design consistent Bayes estimators for a wide class of priors without requiring any adjustment. We will also point out why Lahiri and Mukherjee (2007) needed an adjustment to their Bayes estimators to achieve design consistency.

To this end, we first prove the following theorem. We denote the expression given in the right hand side of (1) as r.

Theorem 2. Consider the one-parameter exponential family as given in Theorem 1. Consider priors $\pi(\theta)$ of θ which are differentiable in θ and satisfy

$$N^{-1}E[\pi'(\theta)/\pi(\theta)|y(s)]\sum_{j\in\bar{s}} a_j / \sum_{i\in s} a_i\sigma_i^{-2} \to 0 \text{ as } n \to \infty. \qquad (5)$$

Then, $E[m(y)|y(s)] - r \to 0$ as $n \to \infty$.

Proof. Once again we use the fact that $E[d\log\pi(\theta|y(s))/d\theta|y(s)] = 0$. In the present set up, this fact leads to the equation

$$\sum_{i\in s} y_i\sigma_i^{-2} - E[\psi'(\theta)|y(s)]\sum_{i\in s} a_i\sigma_i^{-2} + E[\pi'(\theta)/\pi(\theta)|y(s)] = 0,$$

solving which we get

$$E[\psi'(\theta)|y(s)] = \sum_{i\in s} y_i\sigma_i^{-2}/\sum_{i\in s} a_i\sigma_i^{-2} + E[\pi'(\theta)/\pi(\theta)|y(s)]/\sum_{i\in s} a_i\sigma_i^{-2}.$$

The result follows now from Theorem 1 and (5).

While (5) seems somewhat artificial and complicated, it does lead to some simple readily verifiable conditions in some special cases. We begin with the situation where $a_i = \pi_i$ and $\sigma_i^2 = \pi_i/(1-\pi_i)$, $i = 1, \ldots, N$ and $\sum_{i=1}^N \pi_i = n$. Then, r simplifies to the Horvitz-Thompson estimator. In this scenario, $\sum_{j\in\bar{s}} a_j/\sum_{i\in s} a_i\sigma_i^{-2} = 1$ so that (5) simplifies to the condition $N^{-1}E[\pi'(\theta)/\pi(\theta)|y(s)] \to 0$ as $n \to \infty$. For instance, for a prior $\pi(\theta)$ for which $|\pi'(\theta)/\pi(\theta)|$ is bounded uniformly in θ, this condition holds trivially since $n \to \infty$ implies $N \to \infty$.

The boundedness of $|\pi'(\theta)/\pi(\theta)|$ is not all that restrictive either. It holds for many heavy-tailed priors. For example, if $\theta|\sigma^2 \sim N(0, \sigma^2)$ and $\sigma^2 \sim$ inverse gamma$(\beta/2, \alpha/2)$, that is $\pi(\sigma^2) \propto (\sigma^2)^{-\beta/2-1}\exp(-\alpha/2)$, $\alpha > 0$ and $\beta > 0$, then θ has the marginal prior $\pi(\theta) \propto (\theta^2+\alpha)^{-(\beta+1)/2}$. This is immediately recognized as a t-density. In this case $|\pi'(\theta)/\pi(\theta)| = (\beta+1)|\theta|/(\theta^2+\alpha) \le (1/2)(\beta+1)/\alpha^{1/2}$ uniformly in θ. A second example is the logistic prior $\pi(\theta) = \exp(\theta)/[1+\exp(\theta)]^2$ which leads to $\pi'(\theta)/\pi(\theta) = [1-\exp(\theta)]/[1+\exp(\theta)]$. Then, $|\pi'(\theta)/\pi(\theta)| \le 1$ uniformly in θ.

A similar phenomenon occurs for ratio estimators. Here $a_i = x_i$ and $\sigma_i^2 = 1$ for all i. Then, $\sum_{j\in\bar{s}} a_j/\sum_{i\in s} a_i\sigma_i^{-2} = \sum_{j\in\bar{s}} x_j/\sum_{i\in s} x_i$. If now $C_1 \le x_i \le C_2$ for all $i = 1, \ldots, N$, then $\sum_{j\in\bar{s}} x_j/\sum_{i\in s} x_i \le (N-n)C_2/(nC_1)$. Then, (5) reduces to the simple condition $n^{-1}E[\pi'(\theta)/\pi(\theta)|y(s)] \to 0$ as $n \to \infty$. Again, for the t and logistic priors considered in the previous paragraph, (5) trivially holds.

A standard normal prior for θ leads to $\pi'(\theta)/\pi(\theta) = \theta$ and in this case (5) may not hold true because of the unboundedness of θ. Accordingly, the subjective Bayes estimator of Ericson (1969) may not achieve design robustness except under very special circumstances, for example, under simple random sampling with replacement. But even under the normal superpopulation

model, which is a special case of (1), a heavy-tailed prior such as the t or logistic can lead to design consistency of the Bayes estimator of the finite population mean.

It is important to point out why Lahiri and Mukherjee (2007) needed an adjustment to their subjective Bayes estimator. They introduced a new random variable T_n to this end. If one examines carefully the pdf of T_n given in their (2.4), then it is clear that it is equivalent to the posterior pdf given in our Section 2 with the flat prior $\pi(\theta) = c$, $a_i = 1$ and $\sigma_i^2 = \sigma^2$ (ϕ in their notation) for all i. As pointed out in our Remark 4, one then has $E[\psi'(T_n)] = \bar{y}_s$. This is an exact relation. Obviously, $\bar{y}_s$ is not necessarily equal to a weighted estimator, say, $\bar{y}_w$ unless the design is self weighted, which was noted also by Lahiri and Mukherjee. On the other hand, as evidenced in our Section 2, what one needs is a suitable choice of the a_i and σ_i^2 to agree with the Horvitz-Thompson or the ratio estimator.

4. Summary and conclusion

The paper derives Bayes estimators under the one-parameter exponential family superpopulation model, and demonstrates design consistency of these estimators under certain conditions. It is needless to say that not all model-based estimators can achieve design consistency. Many authors have shown design consistency of certain model-based estimators in a probabilistic sense. What we have shown here is that often it is possible to achieve design consistency of Bayes estimators in a pure mathematical way, holding the observations as fixed numbers. A possible extension of our work is to consider the multiparameter situation, and show design consistency of model-based estimators, for example, in the regression model, by considering mathematical rather than probabilistic limits.

Acknowledgement

This research was partially supported by an NSF Grant SES-1026165.

REFERENCES

(1) BASU, D. (1971). An essay on the logical foundations of survey sampling, part 1. In *Foundations of Statistical Inference*. Eds. V.P. Godambe and D.A. Sprott. Holt, Rinehart and Winston, Toronto, Canada, 203-242.

(2) SARNDAL, C-E, SWENSSON, B. and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

(3) ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations (with discussion). *J. Roy. Statist. Soc. B*, 31, 195-233.

(4) GHOSH, M. and SINHA, B.K. (1989). On the consistency between model and design based estimators in survey sampling. *Comm. Statist.*, 20, 689-702.

(5) HAJEK, J. (1971). Discussion of 'an essay on the logical foundations of survey sampling, part one' by D. Basu. In *Foundations of Statistical Inference*. Eds. V.P. Godambe and D.A. Sprott. Holt, Rinehart and Winston, Toronto, Canada, p 236.

(6) LAHIRI, P. and MUKHERJEE, K. (2007). On the design consistency property of hierarchical Bayes estimators in finite population sampling. *Ann. Statist.*, 35, 724-737.

(7) LITTLE, R.J.A. (2004). To model or not to model? Comparing modes of inference for finite population sampling. *J. Amer. statist. Assoc.*, 99, 546-556.

(8) PRASAD, N.G.N. and Rao, J.N.K. (1999). On robust small area estimation using a single random effects model. *Survey Methodology*, 25, 67-72.

(9) ROYALL, R.M. (1970). On finite population sampling theory under certain regression models. *Biometrika*, 57, 377-387.

(10) SARNDAL, C-E. (1984). Design-consistent versus model-dependent estimation in small domains. *J. Amer. statist. Assoc.*, 79, 624-631.