A noninformative Bayesian approach to domain estimation

Glen Meeden*
School of Statistics
University of Minnesota
Minneapolis, MN 55455
glen@stat.umn.edu

August 2002
Revised July 2003
To appear in
Journal of Statistical Planning and Inference

*Research supported in part by NSF Grant DMS 9971331

SUMMARY

Rather than an estimate of the population mean or total sometimes an estimate of the mean or total of a subpopulation or domain is desired. For such problems the number of units in the sample that fall into the domain is a random variable. This causes some complications for the usual frequentist approach. Here we give a simple and coherent noninformative Bayesian approach to domain estimation.

AMS 1991 subject classifications Primary 62D05; secondary 62C10.

Key Words and phrases: Sample survey, domain estimation, Polya posterior and noninformative Bayes.

1 Introduction

In the Bayesian approach to statistical inference the posterior distribution summarizes the information about a parameter through its posterior distribution. This distribution depends on a probability model and a prior distribution and is conditional on the observed data. In finite population sampling the unknown parameter is just the entire population and a prior distribution must be specified over all possible values of the units in the population. Since units in the sample are assumed to be observed without error the only randomness in the model comes from the probabilities of selection which depend on the sampling design. Given a sample the posterior is just the conditional distribution of the unobserved units given the the values of the observed units computed under the prior distribution for the population. This posterior does not depend on the sampling design which selected the sample. The Bayesian approach to finite population sampling was very elegantly described in the writings of D. Basu. For further discussion see his collection of essays in Ghosh (1988).

Usually, when estimating either the mean or total of a subpopulation or domain it is not known which units belong to the domain. One learns whether or not a unit belongs to the domain only for those units included in the sample. Hence the mean of the units in the sample which fall into the domain is the ratio of two random variables. This estimate is more complicated than the mean of all the units in the sample. To get an estimate of variance for this estimator the usual frequentist method conditions on the number of units in the sample that are in the domain. However when estimating the domain total this conditional argument does not work and an unconditional method is used to get an estimate of variance. More details on these methods can be found in Cochran (1977).

The Polya posterior is a noninformative Bayesian approach to finite population sampling which uses little or no prior information about the population. Even though a prior distribution is not specified given the sample there is a "posterior" distribution which may be used to make inferences. It is appropriate when a classical survey sampler would be willing to use simple random sampling as their sampling design. It has been shown that many standard frequentist methods can be thought of as arising from the Polya posterior. For more details see Ghosh and Meeden (1997). Here we will show that the inferences based on the Polya posterior agree with the two standard methods. This yields a consistent noninformative Bayesian

justification for these procedures.

In section 2 we briefly recall the standard methods for domain estimation. In section 3 we review some facts about the Polya posterior and in section 4 study how it behaves for estimating the domain mean and the domain total. We also present some simulation results that show it gives good frequentist answers for the problem of estimating the difference between two domain medians. In section 5 we give some concluding remarks.

2 Domain estimation

Consider a finite population consisting of N units labeled $1, 2, \ldots, N$. The labels are assumed to be known but contain no information about the characteristic of interest. For each unit i this unknown characteristic is denoted by y_i, a real number. The unknown state of nature, $y = (y_1, \ldots, y_N)$, is assumed to belong to some subset of N-dimensional Euclidean space, $\Re^N$. In addition we let $t_i = 1$ if the ith unit belongs to the domain of interest, say $\mathcal{D}$, and 0 otherwise. The vector $t = (t_1, \ldots, t_N)$ and its sum $\sum_{i=1}^{N} t_i = N_d$, the number of units that belong to $\mathcal{D}$, are both assumed to be unknown.

For simplicity we will assume the sampling design is simple random sampling without replacement of size n. A sample s is a subset of $\{1, 2, \ldots, N\}$ of size n. A sample point consists of the set of observed labels s along with the corresponding values for the characteristic of interest and whether or not they belong to $\mathcal{D}$. If $s = \{i_1, \ldots, i_n\}$ then such a sample point can be denoted by (s, y_s, t_s). Let $n_s(d) = \sum_{i \in s} t_i$ be the number of units in the sample that belong to $\mathcal{D}$.

We let $\mu(y)$ and $T(y)$ denote the population mean and total while $\mu_d(y)$ and $T_d(y)$ denote the corresponding domain quantities. The natural estimator of $\mu_d(y)$ is

$$\bar{y}_s(d) = \frac{\sum_{i \in s \cap \mathcal{D}} y_i}{n_s(d)} \tag{1}$$

Since this estimator is the ratio of two random variables the usual recommendation is to condition on the value of $n_s(d) > 1$ to get an estimate of its variance. But this depends on N_d which often is unknown. In these cases the ratio $n_s(d)/N_d$ is replaced by n/N to get the estimate of variance

$$(1 - \frac{n}{N}) \frac{v_s(d)}{n_s(d)} \tag{2}$$

where

$$v_s(d) = \sum_{i \in s \cap \mathcal{D}} (y_i - \bar{y}_s(d))^2/(n_s(d) - 1)$$

When N_d is unknown the usually estimator of T_d is

$$\frac{N}{n} \sum_{i \in s \cap \mathcal{D}} y_i = \frac{n_s(d)}{n} N \bar{y}_s(d) \tag{3}$$

However conditioning on $n_s(d)$ no longer gives a sensible estimate of its variance so the following stratagem is employed. For each unit a new variate y'_i is defined, where

$$y'_i = \begin{cases} y_i & \text{if } t_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

Note that the population total for this new variate, say $T(y')$, is equal to the domain total of the variate y. Moreover we see in equation (3) that if we replace y_i with y'_i and sum over the entire sample the value of this estimator remains unchanged. Hence this estimator is just the usual estimator of the population total for the variate y'. So by standard theory an unbiased estimate of its variance is given by

$$(1 - \frac{n}{N}) \frac{N^2 v'_s}{n} \tag{4}$$

where

$$v'_s = \sum_{i \in s} (y'_i - \bar{y'}_s)^2/(n-1)$$

and $\bar{y'}_s$ is the sample mean.

On page 37 of Cochran (1977) it is noted that when computing v'_s

> ... any unit not in the domain is given a zero value. Some students seem to have a psychological objection to doing this, but the method is sound.

Such a student might also wonder when one should condition and what they should condition on. In the next section we will see that the Polya posterior gives a consistent noninformative Bayesian justification for both methods.

3 The Polya posterior

We begin by briefly recalling some facts about the Polya posterior when estimating the population mean $\mu(y)$. Given the sample the Polya posterior is a predictive joint distribution for the unobserved units in the population conditioned on the values in the sample. Given a data point (s, y_s) we now show how to generate a set of possible values for the unobserved units from this distribution. Consider an urn that contains n balls, where ball one is given the value $y_{s_{i_1}}$, ball two the value $y_{s_{i_2}}$ and so on. We begin by choosing a ball at random from the urn and assigning its value to the unobserved unit in the population with the smallest label. This ball and an additional ball with the same value are then returned to the urn. Another ball is chosen at random from the urn and we assign its value to the unobserved unit in the population with the second smallest label. This second ball and another with the same value are then returned to the urn. This process is continued until all $N - n$ unobserved units are assigned a value. Once this is done we have generated one realization of the complete population from the Polya posterior distribution. This simulated, completed copy contains the n observed values along with the $N - n$ simulated values for the unobserved members of the population. Hence by simple Polya sampling we have a predictive distribution for the unobserved given the observed.

One can verify that under this predicted distribution the expected value of the population mean is just the sample mean. This follows because under the Polya posterior, given the sample, the distribution for the unsampled units is exchangeable and the conditional expectation of each unsampled unit is just the observed sample mean. With a bit more work one finds that its posterior variance is

$$(1 - \frac{n}{N})\frac{v_s}{n}\frac{n-1}{n+1} \tag{5}$$

where v_s is just the sample variance. Note that this is approximately the frequentist variance of the sample mean under simple random sampling when $n \geq 25$. Hence inference for the population mean under the Polya posterior agrees with standard methods. Note the design probabilities play no formal role in the inference based on the Polya posterior. But for it to be appropriate in the judgment of the survey sampler the values for the characteristic of interest for the observed and unobserved units need to be roughly exchangeable. This is usually the case when simple random sampling is used to select the sample.

For more details and discussion on the theoretical properties of the Polya posterior see Ghosh and Meeden (1997). In particular it has been shown for a variety of decision problems that procedures based on the Polya posterior are admissible because they are stepwise Bayes. It is the stepwise Bayes nature of the Polya posterior that explains its somewhat paradoxical properties. Given a sample it behaves just like a proper Bayesian posterior but the collection of possible posteriors that arise from all possible samples comes from a family of priors not a single prior. The Polya posterior is related to the Bayesian bootstrap of Rubin (1981). See also Lo (1988).

4 Domain estimation revisited

We return to the problem of domain estimation. To simplify notation we will assume that the values of the characteristic of interest are all zero outside the domain. Or in terms of our notation in the previous section y and y' are identical. Now given a sample the number of units belonging to $\mathcal{D}$ and its mean and total are still not completely known. However the Polya posterior induces a posterior distribution for each of this quantities and these posteriors can be used to make inferences about them. In particular under squared error loss the posterior expectation of these quantities are their point estimates and their respective posterior variances yield a measure of precision for these estimators.

First we consider estimating the domain total, $T_d(y)$. From the properties of the Polya posterior given above we see that the posterior expectation of $T_d(y)$ is just the estimate in equation (3) and the corresponding posterior variance is N^2 times the expression in equation (5) which except for the factor $(n-1)/(n+1)$ agrees with equation (4). Note this is just a restatement of the agreement between standard frequentist methods and the Polya posterior when estimating the population total in this special case when all the units outside the domain are zero.

Next we will compute the posterior expectation of the domain mean under the Polya posterior. Remember that the sample values are fixed for this calculation and the design probabilities play no role. What is considered to

be random are the unobserved units of the population.

$$\begin{aligned} E(\mu_d(y)) &= E\{E(\mu_d(y)|N_d)\} \\ &= E\Big\{ \frac{1}{N_d}\Big[\sum_{i \in s \cap \mathcal{D}} y_i + \sum_{j \in s' \cap \mathcal{D}} E(y_j|N_d)\Big]\Big\} \\ &= E\Big\{ \frac{1}{N_d}\Big[n_s(d)\bar{y}_s(d) + (N_d - n_s(d))\bar{y}_s(d)\Big]\Big\} \\ &= E(\bar{y}_s(d)) \\ &= \bar{y}_s(d) \end{aligned}$$

where s' is the complement of s. From this observation it follows easily that this is an admissible estimator of the domain mean.

Next we compute approximately the posterior variance of $\mu_d(y)$

$$\begin{aligned} V(\mu_d(y)) &= E(V(\mu_d(y)|N_d)) + V(E(\mu_d(y)|N_d)) \\ &= E\left[\Big(1 - \frac{n_s(d)}{N_d}\Big)\ \frac{v_s(d)}{n_s(d)}\ \frac{n_s(d)-1}{n_s(d)+1}\right]\ +\ V(\bar{y}_s(d)) \\ &= \Big(1 - \frac{n_s(d)}{N}\Big)\ E\Big[\frac{1}{(N_d/N)}\Big]\ \frac{v_s(d)}{n_s(d)}\ \frac{n_s(d)-1}{n_s(d)+1}\ +\ 0 \\ &\doteq \Big(1 - \frac{n-1}{N}\ \frac{n_s(d)}{n_s(d)-1}\Big)\ \frac{v_s(d)}{n_s(d)}\ \frac{n_s(d)-1}{n_s(d)+1} \end{aligned}$$

The last line follows from the fact that if N is large compared to n then under the Polya posterior the distribution of N_d/N is approximately beta$(n_s(d), n - n_s(d))$. It assumes that $n_s(d) > 1$. Note that if $n_s(d)$ is not too small then the posterior variance of $\mu_d(y)$ given above will be nearly equal to the usual conditional variance given in equation (2). The preceeding gives a noninformative Bayesian justification for the usual point and interval estimates of the domain mean and total.

One advantage of the Bayesian approach is that given a sensible posterior one can estimate a variety of parameters of the population. For example suppose there are two domains of interest and we wish to estimate the difference between their medians. The standard text books often talk about the problem of estimating the difference between two domain means but say little about this problem. But one can use the Polya posterior in a straightforward manner. Given a sample one generates a completed copy of the entire population and then finds the difference between the two domain medians. One

repeats this many times and takes the mean of all the simulated differences as the point estimate. To get a 0.95 credible region for the difference one just computes the lower 0.025 quantile and the upper 0.975 quantile of all the simulated differences. Even though one does not have explicit formulas for these point and interval estimates they can easily be found approximately through simulation.

To see how the Polya posterior would work for this problem we will present simulation results from four populations. Each population will contain 1,000 units. The two domains of interest will be of size 300 and 200 and the remaining 500 units will be set equal to zero size since their values do not matter. In the first population the first domain is a random sample of 300 from a normal distribution with mean 25 and standard deviation 4 and the second domain is a random sample of 200 from a normal distribution with mean 20 and standard deviation 2. In the second population the first domain is a random sample of 300 from a gamma distribution with shape parameter 25 and scale parameter 1 and the second domain is a random sample of 200 from a gamma distribution with shape parameter 20 and scale parameter 1. In the third population the first domain is a random sample of 200 from a lognormal distribution with mean and standard deviation (of the log) 5.3 and 0.6 and the second domain is a random sample of 300 from a lognormal distribution with mean and standard deviation (of the log) 4.9 and 0.6. The fourth population was generated just like the third expect the two standard deviations (of the log) were 0.7. In each case we were interested in estimating the difference between the median of the first domain and the median of the second domain. The results of the simulation along with the true differences are given in the table. As a check we have included the average of the difference between the sample medians of the two domains and its average absolute error.

First we note that for the point estimation problem the estimator based on the Polya posterior preforms slightly better than the usual estimator. It seems to have slightly less bias and slightly smaller average absolute error. This is not surprising when we remember that for estimating the median of the population the estimator based on the Polya posterior preforms slightly better than the sample median. For more details see section 8.2 of chapter 2 of Ghosh and Meeden (1997). For the interval estimation problem the 0.95 Bayesian credible intervals appear to behave sensible. Their actual frequency of coverage appears to be very close to 95% although in some cases they appear to be a bit conservative.

Put the Table about here

5 Concluding Remarks

Basu argued that in most situations, after the sample has been observed, the design probabilities should play no role when making inferences about the parameter of interest. For a Bayesian this means conditioning on the observed sample and using a posterior distribution to make inferences. Frequentists, however, prefer to argue unconditionally and consider the properties of their procedures under repeated sampling. Sometimes this is difficult to do and they need to invoke other principles to get a sensible estimate of variance. When estimating the domain mean they need to argue conditionally.

It usually is quite difficult to select a prior which reflects one's beliefs about the population. Moreover most such subjective priors will typically yield interval estimators with poor frequentist properties. These two facts have limited the application of the subjectivist Bayesian approach to finite population sampling. Here we have seen that the Polya posterior, an objective noninformative Bayesian method, is a coherent approach to domain estimation which yields procedures with good frequentist properties. In some cases it yields sensible answers where the standard frequentist methods are difficult to apply.

References

[1] W. Cochran. *Sampling techniques*. Wiley, New York, 3rd edition, 1977.

[2] J. K. Ghosh, editor. *Statistical Information and Likelihood: A Collection of Critical Essays by D. Basu*. Springer-Verlag, New York, 1988.

[3] M. Ghosh and G. Meeden. *Bayesian Methods for Finite Population Sampling*. Chapman and Hall, London, 1997.

[4] A. Lo. A Bayesian bootstrap for a finite population. *Annals of Statistics*, 16:1684–1695, 1988.

[5] D. Rubin. The Bayesian bootstrap. *Annals of Statistics*, 9:130–134, 1981.

Table 1: The results of a simulation study of the behavior of the Polya posterior when estimating the difference between the medians of two domains. Given below are the average values and the average absolute errors of the point estimates along with the average lower bounds and average lengths of the 0.95 credible intervals with their relative frequency of coverage. Also included are the average values and the average absolute error of the usual estimator, the difference between the two sample medians. The results are based on 500 random samples of sizes 100 and 200 from four different populations.

Population (True diff)	Sample Size	Ave value	Ave aberr	Ave lowbd	Ave len	Freq. of coverage
normal	100	4.82	0.69	2.78	4.02	0.980
(4.84)		4.82	0.75	the usual estimator		
	200	4.87	0.51	3.50	2.69	0.964
		4.88	0.56	the usual estimator		
gamma	100	5.40	1.26	1.74	7.21	0.972
(5.41)		5.48	1.41	the usual estimator		
	200	5.45	0.86	2.99	4.85	0.966
		5.47	0.93	the usual estimator		
lognorm1	100	77.2	28.3	2.67	158.8	0.966
(71.3)		75.16	31.4	the usual estimator		
	200	77.6	19.1	26.9	106.1	0.964
		76.2	21.6	the usual estimator		
lognorm2	100	58.3	31.4	-25.0	178.8	0.956
(62.1)		56.9	32.7	the usual estimator		
	200	58.8	19.3	5.87	108.0	0.954
		58.8	19.3	the usual estimator		