The Philippine Statistician, 2003
Vol. 52, Nos. 1-4, pp. 17-25

Invited Paper
Inference for Non-ignorable Sampling Designs[1]

Corinne Grace B. Burgos[2]

ABSTRACT

Two methods are presented in obtaining interval estimates for the finite population mean and the mean of the distribution that generates the finite population when selection bias is present. Selection bias, which usually results from the sampling design or procedure, occurs when the sample is not representative of the population. A sample is obtained using Poisson sampling. The two methods considered are non-ignorable methods in the sense that they both use the sampling design information in computing for the interval estimates. The methodology proposed by this paper uses a Bayesian predictive approach. The two methods are compared in terms of accuracy and precision.

Key words: Selection Bias, Bayesian Predictive Inference, Informative Sampling

1. INTRODUCTION

The use of model-based procedures on survey data has been examined closely in recent years because of the observation that a model holding for a sample may be completely different from the model holding for the population. This inappropriate analysis may lead to selection bias, which occurs when the sample is not a good representative of the population, and thus, two different models hold for both sample and population. The selection bias will most likely result to misleading conclusions. Weighting has been suggested to compensate for this difference although model-based analysts consider weighting to be completely unnecessary while those who advocate design-based analysis include these weights in every analysis.

As statisticians begin to realize the advantage in each of these two viewpoints, the concept of an ignorable design is defined. Rubin (1976) and Sugden and Smith (1984) discuss some sampling designs or situations which may be considered ignorable. They also express, in terms of joint densities when a design or scheme is ignorable. An example of an ignorable sampling design cited as Case B in Krieger and Pfeffermann (1992) is one in which a sample is selected with probabilities proportional to z_i with replacement such that at each draw $k = 1,2,...,n$, $\pi_i = P(i \in s) = \frac{z_i}{\sum_{j=1}^{N} z_j}$. The data known to the analyst are $\{y_i, z_i, i \in s\}$ and $\{z_{n+1},...,z_N\}$. With some minimal assumptions about the correlation between Y and Z, the MLE may be obtained.

---

[1] Jan Tinbergen Award for Best Paper from Developing Countries; paper republished from the Proceedings of the 54th Session of the International Statistics Institute held in Berlin, Germany last August 13 to 21, 2003.

[2] Professor at the Mathematics Department, De La Salle University, Taft Avenue, Manila, email: burgosc@dlsu.edu.ph

Pfeffermann (1993) discusses the matter of weighting in modeling survey data. Although his suggestions in this paper are numerous, his general conclusion is that weighting can be used to test and protect against:

1. informative sampling designs and
2. misspecification of the model holding in the population.

Krieger and Pfefferman (1992) consider two designs, D1 (pps with replacement) and D2 (stratified sampling) and the models that result from them. The usual maximum likelihood and weighted maximum likelihood estimators are computed in simulated examples. They note that the weighted maximum likelihood estimators perform well when the relationship between the sample selection probabilities and sample data is expressed correctly.

This issue is addressed well by Chambers, Dorfman and Wang (1998) where they express the sample selection probability as a function of some covariate. In fact, they give a generalization of D1 used by Krieger and Pfeffermann (1992). The model that this paper develops uses this specification of sample selection probabilities. However, as both previously mentioned papers use variations on maximum likelihood estimation, the approach used in this paper is Bayesian.

In this paper, we consider the finite population of size N: $Y_1, Y_2, ..., Y_N$ generated by $N(\mu, \sigma^2)$. A sample of size n from this finite population will be selected via Poisson sampling. In this scheme, we let $I_1, I_2, ..., I_N$ be Bernoulli random variables such that

$$I_i = \begin{cases} 1 & \text{if unit i is included in the sample} \\ 0 & \text{otherwise} \end{cases}$$

The sample selection probability for each Y_i in the population given by $P(I_i = 1) = \pi_i$ are assumed to be known for all populations units. These sample selection probabilities π_i represent the probability that unit i with characteristic Y_i will be included in the sample.

We would like to develop interval estimates for two parameters, the finite population mean $\overline{Y}$ and the superpopulation mean μ. Since the design in use is non-ignorable, we use the model developed by Pfefferman, Skinner, Holmes, Goldstein and Rasbash (1998), and we use some of the specifications used by Chambers, Dorfman and Wang (1998) in a Bayesian predictive model. The results of these two methods will be compared on simulated examples with different degrees of selection bias. This selection bias results from the non-ignorability of the sampling design. We also observe that the selection bias increases with the increase of coefficient of variation given by

$$cv = \frac{\sigma}{\mu} \times 100\%.$$

2. BAYESIAN INFERENCE

The framework that we use for this paper is Bayesian. This means that the parameter of interest θ which may be vector-valued, will have an assumed probability distribution called the *prior distribution*, denoted by $p(\theta)$. We denote the information obtained from the sample by y. The *likelihood* is a function of the parameter for fixed data and is denoted by $f(y|\theta)$ as a function of θ. The probability distribution of θ utilizing the information from the sample is called the *posterior* denoted by $\pi(\theta|y)$. The posterior is expressed below in its unnormalized form:

$$\pi(\theta|y) \propto p(\theta)f(y|\theta).$$

Thus, the conditional distribution of the parameter of interest θ is proportional to the product of the prior and the likelihood. All information for inference thus resides in the posterior.

To summarize the information obtained from the data about the parameter of interest, samples are obtained from the posterior. After obtaining a large enough sample from the posterior, summary statistics and intervals are computed. For example, if we wish to obtain the bounds of a 95% interval, the 2.5th and 97.5th percentiles are identified from the ordered values sampled from the posterior. It is appropriate to state that the probability that θ will be in the interval is equal to 0.95. In cases where the posterior is not a properly defined probability density function, Markov chain Monte Carlo methods enable us to obtain samples from it. The intervals that result from a Bayesian analysis are called *credible intervals*. If available, the shortest credible interval, which may be considered the best, is called the *highest posterior density* (HPD) *interval.*

3. PFEFFERMAN'S METHOD

The first method we use to compute interval estimates is due to Pfefferman, Skinner, Holmes, Goldstein and Rasbash (1998), also from Krieger and Pfeffermann (1992), which is why we refer to it as Pfeffermann's Method in this paper. This method uses sample information and sample inclusion probabilities. For Pfeffermann's method, we assume that $Y_i \sim N(\mu, \pi_i\sigma^2)$. We also assume flat priors for the parameters that generate the observations:

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}.$$

As assumed earlier, the sample inclusion probabilities π_i are known for all population units. Then, the point estimate for both the finite population mean $\overline{Y}$ and the superpopulation mean μ is given as

$$P = \frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i}}{\sum_{i=1}^{n} \pi_i^{-1}}$$

The HPD intervals for both parameters are available in this case. A (1-α)100% HPD interval for the mean μ is

$$P \pm t_{\alpha/2,n-1}\sqrt{\frac{\sum_{i=1}^{n}(y_i - P)^2/\pi_i}{(n-1)\sum_{i=1}^{n}\pi_i^{-1}}}.$$

A (1-α)100% HPD interval for the finite population mean $\overline{Y}$ is

$$P \pm t_{\alpha/2,n-1}\sqrt{\left(1-\frac{n}{N}\right)\left(\frac{\sum_{i=n+1}^{N}\pi_i}{N-n} + \frac{N-n}{\sum_{i=1}^{n}\pi_i^{-1}}\frac{\sum_{i=1}^{n}(y_i - P)^2/\pi_i}{N(n-1)}\right)}.$$

The requirements to implement these interval estimates are the sample values and the complete set of sample selection probabilities.

4. BAYESIAN PREDICTIVE MODEL

In this section, we develop a Bayesian predictive model to compute credible intervals for the two parameters of interest earlier identified. A sample of size n is obtained via Poisson sampling from a finite population of size N generated by a Normal distribution with mean μ and variance σ^2. Sample inclusion probabilities for all population units are known. This is not the limited information situation considered by Chambers, Dorfman and Wang (1998) but we utilize the relationship indicated in their paper between a latent variable υ_i and the sample inclusion probabilities π_i such that

$$\pi_i \propto \nu_i, \quad i = 1, \cdots, N.$$

We specify this relationship as

$$\pi_i = \frac{n\nu_i}{N\bar{\nu}}$$

where

$$\bar{\nu} = \sum_{i=1}^{N}\nu_i/N$$

We then relate these latent variables to the observations through the model

$$\nu_i = \beta_0 + \beta_1 Y_i + e_i, \quad i = 1, \cdots, N$$

such that $e_i|\sigma_e^2 \sim N(0, \sigma_e^2)$. Another constraint incorporated into this model is that the sum of all the sample inclusion probabilities is equal to the sample size n. That is,

$$\sum_{i=1}^{N}\pi_i = n.$$

This constraint is adapted from Chambers, Dorfman and Wang (1998) and Nandram and Sedransk (2001). The former justifies the use of this restriction by noting that any reasonable sampling scheme should have this property.

We may also express latent variables as

$$\nu_i = c_i \bar{\nu} \qquad (1)$$

where $c_i = \frac{N\pi_i}{n}$. As seen from above, since $0 < \pi_i \leq 1$, either all ν_i's are all positive or all negative. We take the case where all of them are positive and this is a second constraint that will be imposed on the model. That is,

$$\nu_i > 0 \text{ for all i} \qquad (2)$$

Also, for identifiability reasons we let β_0=1, as in Nandram and Sedransk (2001). As for the priors assumed for the parameters β_1, μ, σ^2 and σ^2_e, we took a non-informative for (β_1, μ) and proper diffuse priors for σ^{-2} and σ^{-2}_e such that virtually no information is assumed about them. The priors are given as

$$p(\beta_1, \mu) = 1, \quad \sigma^{-2} \sim \Gamma(a/2, a/2), \quad \sigma_e^{-2} \sim \Gamma(a/2, a/2)$$

where a is a very small positive number such as 0.002.

In order to impose the constraints (1) and (2), latent variables ϕ_i, i = 1,...,N are defined such that

$$\phi_i = \nu_i - c_i \bar{\nu}, \quad i = 1, 2, \cdots, N-1$$
$$\phi_N = \bar{\nu}$$

The joint distributions for these variables are then obtained and to impose the constraints, we let $\phi_i = 0$, for i = 1,...,N-1 and $\phi_N > 0$. We note that if we let all $\phi_i = 0$, for i = 1,...,N-1, then $\nu_i = c_i \bar{\nu}$ for the given range of i. Also, if $\phi_N > 0$ then $\bar{\nu} > 0$ which implies that all ν_i's must be positive.

The joint density of all parameters of interest given the sample data with the constraints imposed is given by

$$\pi(\phi_N, \mu, \beta_1, \sigma^2, \sigma_e^2, \underset{\sim}{Y}_{ns} | \underset{\sim}{Y}_{s}) \propto \frac{\frac{1}{\sqrt{2\pi \frac{\sigma_e^2}{N}}} e^{-\frac{N}{2\sigma_e^2}(\phi_N - (\beta_0 + \beta_1 \bar{Y}))^2}}{\Phi\left(\frac{\beta_0 + \beta_1 \bar{Y}}{\sigma_e / \sqrt{N}}\right)} \times \frac{1}{(\sigma_e^2)^{\frac{N-1}{2}}} \times$$

$$e^{-\frac{1}{2\sigma_e^2}(\underset{\sim}{1}(\beta_0+\beta_1\bar{Y}) - \underset{\sim}{c}(\beta_0+\beta_1\bar{Y}) + (\underset{\sim}{1} - \underset{\sim}{c})(\phi_N - (\beta_0+\beta_1\bar{Y})))'(I+J)(\underset{\sim}{1}(\beta_0+\beta_1\bar{Y}) - \underset{\sim}{c}(\beta_0+\beta_1\bar{Y})) + (\underset{\sim}{1} - \underset{\sim}{c})(\phi_N - (\beta_0+\beta_1\bar{Y})))}$$

$$\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y_i - \mu)^2} \times \left(\frac{1}{\sigma^2}\right)^{a/2-1} e^{-a/2\sigma^2} \times \left(\frac{1}{\sigma_e^2}\right)^{a/2-1} e^{-a/2\sigma_e^2}.$$

where $\Phi\left(\frac{\beta_0 + \beta_1 \bar{Y}}{\sigma_e / \sqrt{N}}\right)$ is the usual cumulative standard normal probability and $\underset{\sim}{Y}_{s}$ and $\underset{\sim}{Y}_{ns}$ are the sampled and non-sampled population units respectively. We note that the ϕ_i's disappear

because all except for ϕ_N were equated to zero to impose the constraints. The details of the derivations of this posterior and the conditional posterior densities are given in Burgos (2002).

Samples are obtained from the joint posterior

$$\pi(\underset{\sim}{Y}_{ns}, \phi_N, \sigma_e^2, \sigma^2, \beta_1, \mu | \underset{\sim}{Y}_s)$$

by sampling from the following conditional densities using Metropolis steps:

$$\pi(\underset{\sim}{Y}_{ns} | \phi_N, \sigma_e^2, \sigma^2, \beta_1, \mu, \underset{\sim}{Y}_s)$$

$$\pi(\phi_N | \sigma_e^2, \sigma^2, \beta_1, \mu, \underset{\sim}{Y}_s)$$

$$\pi(\beta_1 | \phi_N, \sigma_e^2, \sigma^2, \mu, \underset{\sim}{Y}_s)$$

$$\pi(\mu | \sigma_e^2, \sigma^2, \beta_1, \underset{\sim}{Y}_s)$$

$$\pi(\sigma^2 | \sigma_e^2, \mu, \beta_1, \underset{\sim}{Y}_s)$$

$$\pi(\sigma_e^2 | \phi_N, \mu, \sigma^2, \beta_1, Y_s).$$

Sampling from the conditional densities of $\underset{\sim}{Y}_{ns}$, β_1 and σ_e^2 utilizes the Metropolis-Hastings algorithm (Chib and Greenberg, 1995). We sample from these conditional densities and the Metropolis steps ensure that the sampled values indeed come from the given distribution via a rejection step. Samples from the conditional density of ϕ_N uses a step prescribed by Devroye (1986). This ensures that the value of ϕ_N is greater than zero, which is one of the two constraints imposed on the model. Sampling from the conditional densities of μ and σ^2 is straightforward.

5. SIMULATED EXAMPLES

We wish to compare the performance of the two aforementioned methods whenever selection bias is present. Enumerated below are the simulation steps performed for each example.

1. Generate a population of size N from a $N(20, \sigma^2)$ population.
2. From the population values, compute values of υ_i such that

$$\nu_i = \beta_0 + \beta_1 Y_i + e_i, \quad i = 1, \cdots, N$$

where $\beta_0 = \beta_1 = 1$ and $e_i | \sigma_e^2 \sim N(0, \sigma_e^2)$ where $\sigma_e^2 = 10$.

3. Compute the sample inclusion probabilities from using the υ_i's.
4. Draw the sample via Poisson sampling.
5. Compute intervals for the two parameters for Pfeffermann's method.
6. Sampling from the joint posterior density uses 11,000 iterates with a 'burn-in' of 1,000. That is, the first 1,000 iterates are deleted to wash out the autocorrrelations

and every tenth iterate remains out of the remaining 10,000. The remaining 1,000 consist of the sample from the joint posterior.

7. The 2.5^{th} and 97.5^{th} percentiles are used to define the upper and lower limits of he interval estimates for the parameters of interest.

The examples given in Table 1 use two sets of population size N and sample size n. The sampling fraction, however, is constant at 5%. The value of σ^2 starts at 1 then is increased to 100 then to 400 to have values of cv equal to 0.05, 0.5 and 1. With the increase in cv, selection bias also becomes more severe. It would be interesting to note which method performs better under extreme selection bias. In Table 1, FPM is the finite population mean for the example, HT is the Horvitz-Thompson estimator, $\bar{Y}_s$ is the sample mean, BP is the symbol for the Bayesian predictive model while PF stands for Pfeffermann's method. The two parameters of interest are FPM and μ, which is the mean of the Normal distribution that generates the population. In our examples, this value is 20. SE is the standard error of the estimate while the interval estimate is given in the last column. An accurate interval estimate contains the value of the parameter it estimates and a precise one does it with the least interval width.

Table 1. Comparison of Estimators at 5% Sampling Fraction

σ^2	N	n	FPM	HT	$\bar{Y}_s$	Mdl	Par	Mean	SE	Interval
1	200	10	19.92	19.19	20.21	BP	FPM	19.12	0.69	(17.74, 20.21)
							μ	19.12	0.69	(17.74, 20.21)
						PF	FPM	20.16	0.35	(19.47, 20.84)
							μ	20.16	0.35	(19.45, 20.86)
1	400	20	19.92	20.21	20.11	BP	FPM	19.45	0.42	(18.30, 20.03)
							μ	19.45	0.42	(18.31, 20.02)
						PF	FPM	20.12	0.17	(19.77, 20.46)
							μ	20.12	0.18	(19.76, 20.47)
100	200	10	19.89	20.37	28.02	BP	FPM	20.30	1.03	(18.21, 22.38)
							μ	20.29	1.09	(18.19, 22.50)
						PF	FPM	25.08	3.24	(18.70, 31.47)
							μ	25.08	3.35	(18.50, 31.67)
100	400	20	20.05	19.43	24.82	BP	FPM	19.38	0.78	(17.86, 20.99)
							μ	19.38	0.80	(17.85, 21.01)
						PF	FPM	20.40	2.39	(15.68, 25.11)
							μ	20.40	2.45	(15.55, 25.24)
400	200	10	20.89	20.74	29.87	BP	FPM	19.45	0.90	(17.68, 21.23)
							μ	19.46	0.98	(17.58, 21.47)
						PF	FPM	21.34	5.27	(10.97, 31.71)
							μ	21.34	5.42	(10.69, 31.99)
400	400	20	18.25	16.95	30.16	BP	FPM	19.02	0.65	(17.76, 20.35)
							μ	19.01	0.72	(17.65, 20.35)
						PF	FPM	20.39	3.60	(13.29, 27.48)
							μ	20.39	3.71	(13.07, 27.71)

The selection bias is practically non-existent in the case where cv is small, that is, σ^2=1. The sample mean is close to the values of μ and FPM. This means that the units

included in the sample represent those in the population. Both models perform well because they both contain the values of the parameters. However, as the value of **cv** increases, there is a marked increase in the sample mean compared to the two parameters. This means that larger values are entering the sample producing an unusually large sample mean. The two methods produce intervals that are accurate yet, in examples where selection bias is severe (σ^2=400), those produced by Pfeffermann's method are much wider compared to those produced by the Bayesian predictive method. This means that the Bayesian predictive method produce smaller standard errors compared to Pfeffermann's method only under extreme selection bias.

Since the results for the simulated examples are interesting, we simulate 200 more examples for the cases where selection bias is severe, that is, **cv** = 0.5 and **cv** = 1.

Table 2. Summary of 200 Simulated Examples for Each Case

	PF		BP	
	Coverage	**Width**	**Coverage**	**Width**
a. N = 200, n = 10, σ^2=100				
μ	0.315	4.081608	0.95	4.036
FPM	0.295	3.8184	0.965	3.8393
b. N = 400, n = 20, σ^2=100				
μ	0.23	2.5381	0.95	2.9297
FPM	0.31	3.2438	0.95	2.8524
c. N = 200, n = 10, σ^2=400				
μ	0.795	20.272	0.95	3.9909
FPM	0.805	20.014	0.95	3.8112
d. N = 400, n = 20, σ^2=400				
μ	0.285	5.623	0.95	2.9427
FPM	0.04	5.1356	0.955	2.7435

Note: Width is the average width of the 200 95% credible intervals for the two parameters; **Coverage** is the proportion of intervals containing the true value of the parameter out of the 200 simulated examples.

In the 200 simulated examples for all cases, the Bayesian predictive (BP) method attains the required accuracy while Pfeffermann's (PF) method does not reach this level. In the cases where selection bias is severe (**cv** = 1), BP produces significantly shorter intervals than those produced by PF.

6. CONCLUSIONS

In this paper, two non-ignorable methods for obtaining interval estimates for the finite population mean and the mean of the generating function for the population are examined. Although both utilized sampling design information in obtaining these estimates, the method proposed in this paper uses Bayesian predictive inference. It also incorporates two constraints, given as (1) and (2) in Section 4, which contributes to the precision of the BP

method under severe selection bias. Also, the BP method attains the required level of accuracy which the PF method was unable to do.

This paper shows the great potential of the Bayesian method when used with samples selected using non-ignorable sampling designs. Although, it is still not easy to identify situations where selection bias is present, the Bayesian predictive method introduced in this paper attempts to correct this problem while producing precise interval estimates.

References

BURGOS, C. (2002) Bayesian Predictive Inference for the Finite Population Mean Under Selection Bias, Ph.D. Dissertation; University of the Philippines Los Baños .

CHAMBERS, R.L., DORFMAN A.H. and WANG S. (1998) Limited information like-lihood analysis of survey data. *J.R. Stastist. Soc. B,* **60**. Part 2. 397-411.

CHIB, S., GREENBERG, E. (1995) Understanding the Metropolis-Hastings Algorithm, *The American Statistician*, **49** , 327-335.

DEVROYE L. (1986) Non-uniform Random Variate Generation. Springer-Verlag, New York.

KRIEGER, A.B. and PFEFFERMAN, D. (1992) Maximum likelihood estimation from complex surveys. *Surv. Methodol.*, **18**, 225-239.

NANDRAM, B. and SEDRANSK, J. (2001) A Bayesian predictive inference for the finite population mean under informative sampling. Technical report, Worcester Polytechnic Institute, Worcester, MA.

PFEFFERMANN, D. (1993) The role of sampling weights when modeling survey data. *Int. Statist. Rev.*, **61**, 2 317-337.

PFEFFERMANN, D., SKINNER, C.J., HOLMES, D.J., GOLDSTEIN, H. and RASBASH, J. (1998) Weighting for unequal selection probabilities in multilevel models. *J.R. Stastist. Soc. B,* **60**. Part 1. 23-40.

RUBIN, D.B. (1976) Inference and missing data. *Biometrika*, **53**, 581-592.

SUGDEN, R.A. and SMITH, T.M.F. (1984) Ignorable and informative designs in survey sampling inference. *Biometrika*, **71**, 495-506.