Some Methods of Model-Based Sampling

Sung-Joon Hong[1], So-Hyung Park[1], Sun-Woong Kim[1], Hong-Yup Ahn[1],
Steven G. Heeringa[2]
[1]Department of Statistics, Dongguk University, 26, 3Ga, Pil-Dong, Jung-Gu, Seoul,
South Korea 100-715
[2]Institute for Social Research, University of Michigan, 426 Thompson Street, Ann Arbor,
MI 48106

Abstract

With respect to commonly used πPS sampling techniques, samplers are often interested in the reduction of the design variance of the Horvitz and Thompson (1952) estimator. We first describe the differences between mechanisms of conventional design-based πPS sampling methods of Mizuno (1952) and Brewer (1963) and two model-based πPS sampling methods developed by Kim, Heeringa, and Solenberger (2006). We also suggest two new model-based πPS sampling methods, and empirically compare the efficiency of the new methods to the previous model-based sampling methods and design-based πPS and non-πPS sampling methods. The case where the sample size is two is of particular interest for empirical comparison. With respect to the design variance, model-based πPS sampling methods are preferable to design-based πPS sampling methods. One of new methods performs best. This new method is comparable to the method of Murthy (1957), which is a design-based non-πPS sampling procedure. Moreover, model-based sampling methods are preferable to design-based sampling methods due to the flexibility in the choice of sampling design for a better stability of the variance estimator.

Key Words: regression superpopulation model, average variance, optimization, design variance, stability of variance estimator, maximum likelihood, restricted maximum likelihood

1. Introduction

Since Hansen and Hurwitz (1943) first suggested the selection of primary sampling units from each stratum with probabilities proportional to size (PPS), a large number of techniques for sampling without replacement with unequal probabilities have been developed.

As discussed by Brewer and Hanif (1983) and Särndal (1996), much research on sample selection has been focused on design-based inclusion probability proportional to size (πPS) sampling procedures in which the second-order inclusion probabilities (or joint probabilities), which indicate the probabilities that any two units in a population are both included in a sample, have a key role in the variance reduction. For example, the methods of Mizuno (1952) and Brewer (1963) that are draw-by-draw procedures and Sampford's (1967) method, a rejective procedure, are well known πPS sampling and commonly employed by samplers. The methods of Brewer (1963) and Sampford (1967) are available in software such as SAS or SPSS. See SAS/STAT (2009) and PASW

Statistics (formerly SPSS Statistics) (2009). Murthy's (1957) method, a non-πPS draw-by-draw procedure, was noted by Rao and Bayless (1969) and Cochran (1977). Murthy's method is available in SAS or SPSS for selecting samples in "two per stratum" designs.

The comparative efficiency of these techniques in actual population sampling applications is an open question. As seen in many empirical studies, this may be due to the fact that the variances of the estimates of interest calculated from a sample selected by any sampling method are sensitive to population characteristics, and hence the user may not be sure that the efficiency of a chosen method would be significantly better than other procedures. This is especially true when a small sample is selected from a population or population stratum. In many national surveys deep stratification with a substantial number of strata is used, and only a small number of cluster units are sampled from each stratum. For example, two per stratum designs are common in national samples. Accordingly, a sampling method whose efficiency is robust in the case of small samples would be preferred.

Although a sample is selected from a finite population, considering the concept of an infinite superpopulation may be useful in the sample selection stage. In fact, an infinite superpopulation model has been often used in the estimation procedures, such as model-assisted estimation and model-dependent estimation. But with regard to sample selection the model has been used by many writers mainly for the theoretical comparisons among sampling procedures, not for the actual selection of a sample.

The model may be used to ensure that the second-order inclusion probabilities involving sampling designs implemented by a πPS sampling procedure would result in reasonable efficiency. Kim, Heeringa, and Solenberger (2006) developed a theory of model-based πPS sampling procedures as a specification of the selection method using the model. Their procedures to yield optimal sampling designs that reduce the variance of the Horvitz and Thompson (1952) estimator were based on fairly practical linear superpopulation models and optimization theory.

In this paper, we first describe the mechanism of design-based πPS sampling of Mizuno (1952) and Brewer (1963) and model-based πPS sampling developed by Kim, Heeringa, and Solenberger (2006). Next, we describe new model-based πPS sampling methods, and empirically compare their efficiency to that for the previous model-based sampling methods and the conventional design-based sampling methods of Mizuno (1952), Brewer (1963) and Murthy (1957). The case where the sample size is two is of particular interest for empirical comparison, both for simplicity and because it is the most important situation in practice. The model in model-based sampling is not be central to the selection problem and is just a means to the end of achieving higher efficiency.

2. Mechanism of Design-Based πPS Sampling

Before presenting model-based πPS sampling procedures in the next section, we first describe design-based πPS sampling method of Mizuno (1952), Brewer (1963), and Sampford (1967).

Consider a finite population of N units, denoted by $U = \{u_1, \cdots, u_i, \cdots, u_N\}$. y_i is the value of the variable of interest, y, for the i th unit u_i. In order to estimate the total $Y = \sum_{i=1}^{N} y_i$, a sample s of size n is selected from the finite population. Let $p_d(s)$ be the sampling design (or sampling plan) indicating the probability of selecting a specified sample in design-based πPS sampling. Let the π_i be the first-order inclusion probabilities, denoted by $\pi_i = \sum_{i\in s} p_d(s)$, and let the π_{ij} be the second-order inclusion probabilities given by $\pi_{ij} = \sum_{i,j\in s} p_d(s)$. When $n = 2$, simply $\pi_{ij} = p(s)$.

For $n = 2$, the method of Mizuno (1952) gives

$$\pi_i = p_i + (1 - p_i)\frac{1}{N-1} \tag{2.1}$$

and

$$\pi_{ij} = p_d(s) = \frac{1}{N-1}(p_i + p_j) \tag{2.2}$$

which is a simple function of p_i and p_j .

The method of Brewer (1963), which is only for $n = 2$ and every $p_i < 1/2$, gives

$$\pi_i = 2p_i \tag{2.3}$$

and

$$\pi_{ij} = p_d(s) = \frac{2p_i p_j}{Q}\frac{(1 - p_i - p_j)}{(1-2p_i)(1-2p_j)}, \tag{2.4}$$

where $Q = \frac{1}{2}\left(1 + \sum_{i=1}^{N}\frac{p_i}{1-2p_i}\right)$.

The method of Sampford (1967) is an extension of Brewer's (1963) method to samples of any size, and gives $\pi_i = np_i$, which is called the πPS requirement. Like Brewer's method, $p_d(s)$ is obtain according to the selection probability of each unit defined for each draw, and is a function of the relative sizes p_i .

A sampler may prefer a πPS sampling yielding a smaller design variance. But as described above, the $p_d(s)$ in design-based πPS sampling is a certain function of the relative sizes p_i depending on only the values of the auxiliary variable x , and there is no definite indication of the strength and direction of a linear relationship between the variables x and y . Thus, although $p_d(s)$ plays a central role in the reduction of the design variance, it is not clear whether $p_d(s)$ in any design-based πPS sampling procedure would yield a low variance for any population of interest.

3. Mechanism of Model-Based πPS Sampling

A generalized regression (GREG) estimator may be one of the useful estimators for the population total. But it is well-known that it might be appreciably biased for a small sample, although the bias is in modest for large samples. As an alternative, the Horvitz-Thompson (H-T) estimator (1952) in (3.1), which is unbiased for the population total and highly efficient under a good πPS sampling method, can be used.

$$\hat{Y}_{HT} = \sum_{i=1}^{n}\frac{y_i}{\pi_i} \tag{3.1}$$

The H-T estimator is the only unbiased estimator in the subclass of linear estimators denoted by

$$\hat{Y} = \sum_{i=1}^{n} k_i y_i \tag{3.2}$$

where k_i is a constant to be used as a weight for the i th unit whenever it selected for the sample, and hence the best linear estimator of the subclass (Horvitz and Thompson (1952), Godambe (1955)). Also, note that best linear estimate does not exist for the entire class of linear estimators (Godambe (1955)).

The variance of the H-T estimator is

$$Var(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{y_i^2}{\pi_i} + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{\pi_{ij}}{\pi_i \pi_j} y_i y_j - Y^2 \tag{3.3}$$

Model-based πPS sampling method was first suggested by Raj (1956). This method for $n = 2$ is a variance minimization sampling procedure, which first constructs an optimization problem consisting of an objective function and constraints on sampling design $p_m(s)$ for minimizing $Var\left(\hat{Y}_{HT}\right)$ in (3.3) under the model m denoted by $y_i = \alpha + \beta x_i$ reflecting a linear relationship between the variables x and y. The model-based method then attempts to obtain an optimal set of $p_m(s)$. To find a solution, $p_m(s)$, linear programming (LP) is used.

His model-based πPS sampling procedure has the properties:

a) Prior the sample selection, the sampling design $p_m(s) = \pi_{ij}$ for all possible samples is determined by LP. It meets the πPS requirement, that is, $\pi_i = \sum_{i \in s} p_m(s) = np_i$.

b) One selection using $p_m(s)$ samples the whole sample s. This is a whole sample procedure.

His sampling procedure is attractive with respect to the variance reduction achieved by using the model. But his model is unusual because there is no error term. Kim, Heeringa, and Solenberger (2006) developed a theory of model-based πPS sampling procedures using an infinite superpopulation model. They assume that a finite population of N units is drawn from an infinite superpopulation with the regression model ξ, given by

$$y_i = \alpha + \beta x_i + \varepsilon_i\ ,\ i = 1, \cdots, N\ , \tag{3.4}$$

where $E_\xi(\varepsilon_i | x_i) = 0$, $Var_\xi(\varepsilon_i | x_i) = \delta x_i^\gamma$ ($\delta > 0$, $\gamma \ge 0$), and $E_\xi(\varepsilon_i, \varepsilon_j | x_i, x_j) = 0$, $i \ne j$. E_ξ and Var_ξ respectively denote the expected value and variance under the model ξ. It is also assumed that the ε_i are normally distributed.

Note that many writers often prefer the model without the intercept for the purpose of the simplicity of theoretical comparison between sampling procedures, while the model in (3.4) has the intercept for the practical use.

The variance of the H-T estimator, given by Horvitz and Thompson (1952), is

$$Var\left(\hat{Y}_{HT}\right)=\sum_{i=1}^{N}\frac{y_i^2(1-\pi_i)}{\pi_i}+2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{\pi_{ij}}{\pi_i\pi_j}y_i y_j-2\sum_{i=1}^{N}\sum_{j>i}^{N}y_i y_j \tag{3.5}$$

A different expression on the variance of the H-T estimator, suggested by Yates and Grundy (1953), is

$$Var\left(\hat{Y}_{HT}\right)=\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\pi_i\pi_j-\pi_{ij}\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2 \tag{3.6}$$

With respect to inference, the anticipated variance (ANV), introduced by Isaki and Fuller (1982), is used as a measure describing the variability between the total and the estimator of the total under both the sampling design and superpopulation model. If the H-T estimator is used, it simply becomes the average variance (AV), that is, the model expectation of the design variance expressed as

$$E_\xi E_p\left[\left(\hat{Y}_{HT}-Y\right)^2\right]=E_\xi\left[Var\left(\hat{Y}_{HT}\right)\right], \tag{3.7}$$

where E_p denotes the expected value under the sampling design, and both Y and $\hat{Y}_{HT}$ are random variables.

Kim, Heeringa, and Solenberger (2006) showed that in cases of $n=2$, an optimal sampling design $p_\xi(s)$ in a set of possible πPS sampling designs that minimize the AV in (3.7) can be obtained by using one of the following optimization problems:

$$Minimize\ \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{\alpha+\beta(x_i+x_j)}{x_i x_j}p_\xi(s)\,, \tag{3.8}$$

or

$$Minimize\ \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\frac{1}{x_j}-\frac{1}{x_i}\right)\left(\alpha\frac{1}{x_i}+\beta\right)p_\xi(s)\,, \tag{3.9}$$

subject to the linear equality constraints

$$\sum_{i\in s}p_\xi(s)=\pi_i\,,\ i=1,\cdots,N \tag{3.10}$$

Note that the two objective functions in (3.8) and (3.9) are induced from the expressions of $Var\left(\hat{Y}_{HT}\right)$ in (3.5) and (3.6), respectively, and hence a different form of $Var\left(\hat{Y}_{HT}\right)$ may yield a different optimization problem. We call these two optimization problems composed of (3.8) and (3.10), and (3.9) and (3.10), OP1 and OP2, respectively.

4. New Model-Based πPS Sampling

We continue to focus on the design stage for the actual selection of a sample from a finite population, rather than the estimation stage. In other words, although the H-T estimator does not involve the superpopulation model ξ, we assume the model, and seek to find $p_\xi(s)$ to reduce $Var\left(\hat{Y}_{HT}\right)$ for the finite population as well as $E_\xi\left[Var\left(\hat{Y}_{HT}\right)\right]$ for the infinite population.

Here we first derive objective functions different from those in (3.8) or (3.9), and then construct different optimization problems by adding (3.10) and additional constraints, as seen later.

Theorem 4.1. With the variance formula in (3.5), the AV on the H-T estimator under the superpopulation model in (3.4) is

$$\sum_{i=1}^{N}(X/nx_i-1)\left(\delta x_i^{\gamma}+\alpha^2+2\alpha\beta x_i+\beta^2x_i^2\right)-2\sum_{i}^{N}\sum_{j>i}^{N}\left(\alpha^2+\alpha\beta(x_i+x_j)+\beta^2x_ix_j\right)$$
$$+\frac{2\alpha^2X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_ix_j}\pi_{ij}+2\alpha\beta X\frac{n-1}{n}N+\beta^2X^2\frac{n-1}{n} \tag{4.1}$$

Proof. Consider the form of the variance of the H-T estimator in (3.5). Since it is πPS sampling, $\pi_i=np_i$. Then from the first and third terms in (3.5) under the superpopulation model, we have

$$E_\xi\left[\sum_{i=1}^{N}\frac{y_i^2(1-\pi_i)}{\pi_i}-2\sum_{i=1}^{N}\sum_{j>i}^{N}y_iy_j\right]$$
$$=\sum_{i=1}^{N}(X/nx_i-1)E_\xi(y_i^2)-2\sum_{i=1}^{N}\sum_{j>i}^{N}E_\xi(y_iy_j)$$
$$=\sum_{i=1}^{N}(X/nx_i-1)\left(\delta x_i^{\gamma}+\alpha^2+2\alpha\beta x_i+\beta^2x_i^2\right)-2\sum_{i}^{N}\sum_{j>i}^{N}\left(\alpha^2+\alpha\beta(x_i+x_j)+\beta^2x_ix_j\right) \tag{4.2}$$

For the second term in (3.5), we have

$$E_\xi\left[2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{\pi_{ij}}{\pi_i\pi_j}y_iy_j\right]=\frac{2X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{E_\xi(y_iy_j)}{x_ix_j}\pi_{ij}$$
$$=\frac{2X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{\alpha^2+\alpha\beta(x_i+x_j)+\beta^2x_ix_j}{x_ix_j}\pi_{ij}$$
$$=\frac{2\alpha^2X^2}{n^2}\sum_{i}^{N}\sum_{j>i}^{N}\frac{1}{x_ix_j}\pi_{ij}+\frac{2\alpha\beta X^2}{n^2}\sum_{i}^{N}\sum_{j>i}^{N}\frac{x_i+x_j}{x_ix_j}\pi_{ij}+\beta^2X^2\frac{n-1}{n} \tag{4.3}$$

When the second term in (4.3) is expanded, it gives

$$\frac{2\alpha\beta X^2}{n^2}\left[\sum_{i}^{N}\sum_{j>i}^{N}\frac{1}{x_j}\pi_{ij}+\sum_{i}^{N}\sum_{j>i}^{N}\frac{1}{x_i}\pi_{ij}\right]=\frac{\alpha\beta X^2}{n^2}\left[\sum_{j}^{N}\frac{1}{x_j}\sum_{i\neq j}^{N}\pi_{ij}+\sum_{i}^{N}\frac{1}{x_i}\sum_{j\neq i}^{N}\pi_{ij}\right]$$
$$=\frac{\alpha\beta X^2}{n^2}\left[\sum_{j}^{N}\frac{1}{x_j}(n-1)\pi_j+\sum_{i}^{N}\frac{1}{x_i}(n-1)\pi_i\right]$$
$$=2\alpha\beta X\frac{n-1}{n}N \tag{4.4}$$

This completes the proof.

Corollary 4.1. With the variance formula in (3.5), the AV on the H-T estimator under the superpopulation model with $\alpha = 0$ in (3.4) does not depend on π_{ij}, and is fixed as

$$\sum_{i=1}^{N}(X/nx_i - 1)\left(\delta x_i^{\gamma} + \beta^2 x_i^2\right) - 2\beta^2 \sum_{i}^{N}\sum_{j>i}^{N} x_i x_j + \beta^2 X^2 \frac{n-1}{n} \tag{4.5}$$

Corollary 4.2. If the superpopulation model in (3.4) is assumed, the minimization of the AV on the H-T estimator given in (4.1) is equivalent to minimizing

$$\sum_{i}^{N}\sum_{j>i}^{N} \frac{1}{x_i x_j} \sum_{i,j \in s} p_{\xi}(s) \tag{4.6}$$

Proof. In (4.1) only the third term depends on π_{ij}, while the other terms do not depend on π_{ij}, and are fixed. Thus, the minimization of the AV amounts to minimizing (4.5).

Remark 4.1. (4.6) does not depend on α, β, δ, and γ, and it is a linear function of $p_{\xi}(s)$.

Corollary 4.3. In cases of $n = 2$, the minimization of the AV on the H-T estimator is equivalent to minimizing

$$\sum_{i}^{N}\sum_{j>i}^{N} \frac{1}{x_i x_j} p_{\xi}(s) \tag{4.7}$$

Remark 4.2. As given in (4.7), in cases of $n = 2$, the minimization of the AV on the H-T estimator is reduced to minimizing a simple linear function of $p_{\xi}(s)$. (4.7) is the same function as Raj (1956) induced to minimize $Var\left(\hat{Y}_{HT}\right)$ under the assumption that $y_i = \alpha + \beta x_i$ without the error term. See page 198, Raj (1956).

Result 4.1. Based on (4.7), in cases of $n = 2$, a simple optimization problem to find model-based πPS sampling design $p_{\xi}(s)$, called OP3, can be given by:

$$Minimize \sum_{i}^{N}\sum_{j>i}^{N} \frac{1}{x_i x_j} p_{\xi}(s) \tag{4.8}$$

subject to the linear equality constraints

$$\sum_{i \in s} p_{\xi}(s) = \pi_i, \; i = 1, \cdots, N \tag{4.9}$$

Now we obtain a different AV by using a variance form different from (3.5).

Theorem 4.2. Using the variance expression in (3.6), the AV on the H-T estimator under the superpopulation model in (3.4) is

$$\frac{\delta X}{n}\sum_{i=1}^{N}x_i^{\gamma-1}-\delta\sum_{i=1}^{N}x_i^{\gamma}-2\alpha\left[\sum_{i=1}^{N}\sum_{j>i}^{N}\left(x_i-x_j\right)\left(\alpha x_i^{-1}+\beta\right)\right]$$

$$+\frac{2\alpha X^2}{n^2}\left[\alpha\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_ix_j}\pi_{ij}-\alpha\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_i^2}\pi_{ij}+\frac{\beta Nn(n-1)}{X}-2\beta\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_i}\pi_{ij}\right] \tag{4.10}$$

Proof. For (3.6), we may write

$$Var\left(\hat{Y}_{HT}\right)=\sum_{i=1}^{N}\sum_{j>i}^{N}\left(p_ip_j-\frac{\pi_{ij}}{n^2}\right)\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2 \tag{4.11}$$

Since

$$E_\xi\left[\frac{y_i}{p_i}-\frac{y_j}{p_j}\right]^2=\frac{2}{p_i^2}(\delta x_i^{\gamma}+\alpha^2+\beta^2x_i^2+2\alpha\beta x_i)-\frac{2}{p_ip_j}(\alpha^2+\alpha\beta(x_i+x_j)+\beta^2x_ix_j)$$

$$=2\delta X^{\gamma}p_i^{\gamma-2}+2\alpha X^2\frac{x_j-x_i}{x_i\,x_j}\left(\alpha x_i^{-1}+\beta\right), \tag{4.12}$$

we have

$$E_\xi\left[\sum_{i=1}^{N}\sum_{j>i}^{N}\left(p_ip_j-\frac{\pi_{ij}}{n^2}\right)\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2\right]$$

$$=2\delta X^{\gamma}\sum_{i=1}^{N}\sum_{j>i}^{N}p_i^{\gamma-2}\left(p_ip_j-\frac{\pi_{ij}}{n^2}\right)+\ 2\alpha X^2\left[\sum_{i=1}^{N}\sum_{j>i}^{N}\left(p_ip_j-\frac{\pi_{ij}}{n^2}\right)\frac{x_j-x_i}{x_i\,x_j}\left(\alpha x_i^{-1}+\beta\right)\right]$$

$$=\frac{\delta X^{\gamma}}{n}\sum_{i=1}^{N}(1-np_i)p_i^{\gamma-1}\ +\ 2\alpha X^2\left[\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{x_i\,x_j}{X^2}\frac{x_j-x_i}{x_i\,x_j}\left(\alpha x_i^{-1}+\beta\right)\right]$$

$$-\ \frac{2\alpha X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{x_j-x_i}{x_i\,x_j}\left(\alpha x_i^{-1}+\beta\right)\pi_{ij}$$

$$=\frac{\delta X}{n}\sum_{i=1}^{N}x_i^{\gamma-1}-\delta\sum_{i=1}^{N}x_i^{\gamma}\ -\ 2\alpha\left(\sum_{i=1}^{N}\sum_{j>i}^{N}\left(x_i-x_j\right)\left(\alpha x_i^{-1}+\beta\right)\right)$$

$$+\ \frac{2\alpha X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{x_i-x_j}{x_i\,x_j}\left(\alpha x_i^{-1}+\beta\right)\pi_{ij} \tag{4.13}$$

The last term in (4.13) can be written in the form

$$\frac{2\alpha X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{x_i-x_j}{x_i\,x_j}\left(\alpha x_i^{-1}+\beta\right)\pi_{ij}=\frac{2\alpha X^2}{n^2}\left[\alpha\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{x_i-x_j}{x_i^2x_j}\pi_{ij}+\beta\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{x_i-x_j}{x_ix_j}\pi_{ij}\right]$$

$$=\frac{\alpha X^2}{n^2}\left[2\alpha\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_i\,x_j}\pi_{ij}-2\alpha\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_i^2}\pi_{ij}+\beta\left(2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_j}\pi_{ij}-2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_i}\pi_{ij}\right)\right] \tag{4.14}$$

Also,

$$\begin{aligned}\beta\left(2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_j}\pi_{ij}-2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_i}\pi_{ij}\right)&=\beta\left(2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_j}\pi_{ij}+2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_i}\pi_{ij}-4\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_i}\pi_{ij}\right)\\&=\beta\left(\sum_{i}^{N}\frac{1}{x_i}\sum_{j\neq i}^{N}\pi_{ij}+\sum_{j}^{N}\frac{1}{x_j}\sum_{i\neq j}^{N}\pi_{ij}-4\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_i}\pi_{ij}\right)\\&=2\beta\left(\frac{Nn(n-1)}{X}-2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_i}\pi_{ij}\right)\qquad(4.15)\end{aligned}$$

This completes the proof.

Corollary 4.4. Under the superpopulation model with $\alpha = 0$ in (3.4), (4.10) reduces to

$$\frac{\delta X}{n}\sum_{i=1}^{N}x_i^{\gamma-1}-\delta\sum_{i=1}^{N}x_i^{\gamma}\qquad(4.16)$$

Remark 4.3. (4.16) is different from (4.5), due to the different expressions for the variance of the H-T estimator.

Corollary 4.5. Under the superpopulation model in (3.4), minimizing the AV on the H-T estimator given in (4.10) amounts to minimizing

$$\alpha\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_i x_j}\pi_{ij}-\alpha\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_i^2}\pi_{ij}-2\beta\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_i}\pi_{ij}\ ,\qquad(4.17)$$

where $\pi_{ij}=\sum_{i,j\in s}p_{\xi}(s)$.

Remark 4.4. (4.17) depends on α and β, and it is a linear function of $p_{\xi}(s)$.

Corollary 4.6. In cases of $n = 2$, the minimization of (4.17) reduces to minimizing

$$\sum_{i}^{N}\sum_{j>i}^{N}\left[\alpha\left(\frac{1}{x_i x_j}-\frac{1}{x_i^2}\right)-2\beta\frac{1}{x_i}\right]p_{\xi}(s)\qquad(4.18)$$

Result 4.2. The different optimization problem, called OP4, to obtain a model-based πPS sampling design $p_{\xi}(s)$, for the case of $n = 2$, is given by:

$$Minimize\sum_{i}^{N}\sum_{j>i}^{N}\left[\alpha\left(\frac{1}{x_i x_j}-\frac{1}{x_i^2}\right)-2\beta\frac{1}{x_i}\right]p_{\xi}(s)\qquad(4.19)$$

subject to

$$\sum_{i\in s}p_{\xi}(s)=\pi_i\ ,\ i=1,\cdots,N\ .\qquad(4.20)$$

Remark 4.5. In addition to (4.20), the linear inequality constraints (4.21) can be basically added

$$0 < p_{\xi}(s) \le \pi_i \pi_j \ , \ \ j > i = 1,\cdots,N \ , \tag{4.21}$$

since the well-known variance estimator $\widehat{Var}\left(\hat{Y}_{HT}\right)$ in (4.22), given by Yates and Grundy (1953) and by Sen (1953) from (3.6), is defined if $\pi_{ij} > \mathbf{0}$, and nonnegative if $\pi_i \pi_j \ge \pi_{ij}$.

$$\widehat{Var}\left(\hat{Y}_{HT}\right) = \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^{\mathbf{2}}, \tag{4.22}$$

Also, (4.23) can replace (4.21).

$$c\pi_i \pi_j \le p_{\xi}(s) \le \pi_i \pi_j \ , \ \ j > i = 1,\cdots,N \ , \tag{4.23}$$

where $0 < c < 1$.

Note that the stability of the variance estimator in (4.22) may be improved if c in (4.23) is sufficiently far from 0, as discussed by Hanurav (1967), Nigam, Kumar and Gupta (1984), and Rao and Nigam (1992). Thus, the larger value of c is preferred.

Since $\pi_i = 2p_i$, (4.21) and (4.23) can be respectively expressed in forms

$$0 < p_{\xi}(s) \le \frac{4}{X^2} x_i x_j \ , \ \ j > i = 1,\cdots,N \tag{4.24}$$

$$\frac{4c}{X^2} x_i x_j \le p_{\xi}(s) \le \frac{4}{X^2} x_i x_j \ , \ \ j > i = 1,\cdots,N \tag{4.25}$$

The constraints in (4.24) or (4.25) can be added to OP1, OP2, and OP3, as in OP4.

5. Empirical Study

The previous model-based *πPS* sampling methods, OP1 and OP2, the suggested model-based *πPS* sampling methods, OP3 and OP4, and the conventional design-based sampling methods of Mizuno (1952), Brewer (1963) and Murthy (1957) were compared for the case of n=2. The comparison used 18 small natural populations described in the paper of Rao and Bayless (1969). There were originally 20 populations in their paper, but 2 populations (numbered 6 and 8 in their paper) were excluded because the model in (3.4) was not successfully applied. For estimation of the parameters of the superpopulation model in model-based approaches, maximum likelihood (ML) estimation and restricted maximum likelihood (REML) estimation were used.

For example, OP3 consists of (5.1), (5.2) and (5.3) or (5.4), and "LP procedure" in SAS/OR (2008) was used to find the solution to the model-based sampling design $p_{\xi}(s)$.

$$Minimize \sum_{i}^{N} \sum_{j>i}^{N} \frac{1}{x_i x_j} p_{\xi}(s) \tag{5.1}$$

subject to

$$\sum_{i \in s} p_{\xi}(s) = \mathbf{2} p_i\,,\ \ i = 1, \cdots, N \tag{5.2}$$

and for $0 < c < 1$,

$$\frac{4c}{X^2} x_i x_j \le p_{\xi}(s) \le \frac{4}{X^2} x_i x_j\,,\ \ j > i = 1, \cdots, N \tag{5.3}$$

or for $c = 0$,

$$0 < p_{\xi}(s) \le \frac{4}{X^2} x_i x_j\,,\ \ j > i = 1, \cdots, N \tag{5.4}$$

OP1, OP2, and OP4 denote that only (5.1) in OP3 is replaced by (3.8), (3.9), and (4.19), respectively.

The design-based sampling design $p_d(s)$ for the methods of Mizuno and Brewer was calculated by (2.2) and (2.4), respectively. The $p_d(s)$ for Murthy's method were computed as:

$$p_d(s) = \frac{p_i p_j (2 - p_i - p_j)}{(1 - p_i)(1 - p_j)} \tag{5.5}$$

As an illustration, when plotting sampling designs from OP3 and the three design-based methods by the values of the auxiliary variable (e.g., x_i and x_j) for population 11 in their paper on a three-dimensional graph, we can know that there is a clear difference between the model-based and design-based sampling methods. The sampling designs from model-based sampling using OP3 with $c = 0$ are scattered, while those from the methods of Mizuno, Brewer, and Murthy tend to concentrate. This causes a smaller variance for model-based sampling and a larger variance for design-based sampling. In contrast, the spread of sampling design from model-based sampling using OP3 with $c = 0.5$ and that from the design-based sampling methods are more similar, yielding more equal variances under the different methods. Also, it seems that there is a trade-off between the reduction of the variance and the stability of the variance estimator. The larger value of c indicates the larger stability of the variance estimator in (4.23). When the value of c is relatively low, sampling designs obtained from model-based sampling method using OP3 tend to be dispersed, resulting in a large reduction in variance, compared to the cases where $c = 0.4$ or $c = 0.5$. Moreover, with respect to any value of c, model-based sampling using OP3 gives a smaller variance than the three design-based sampling methods. In addition, it is flexible in terms of c. If one pursues the larger variance reduction rather than the stability of the variance estimator, using a lower value of c may be appropriate. But if we prefer the stability of the variance estimator, a higher value of c can be used, but for the price is the larger variance. Anyway, it would offer an optimal sampling design under the chosen constraints on the value of c.

Table 1 shows the summary on results of empirical comparison on the relative efficiency (RE) for 18 populations for model-based sampling methods using OP1, OP2, OP3, and OP4 with $c = 0$, 0.1, 0.2, 0.3, 0.4, 0.5 and the three design-based sampling methods. Note that those optimization problems were infeasible for the cases with $c = 0.6$, 0.7, 0.8, and 0.9. The details on Table 1 are illustrated as follows:

For example, "OP1 M" in the table denotes OP1 consisting of the estimates of the model from ML estimation, while "OP1 R" indicates OP1 by the estimates from REML estimation. Here, the RE for model-based πPS sampling is denoted by

$$RE_{\xi,\pi PS} = \left[Var_{PPS}\left(\hat{Y}\right) \Big/ Var\left(\hat{Y}_{HT}\right) \right] \times \mathbf{100} \,, \tag{5.6}$$

where $Var_{PPS}\left(\hat{Y}\right) = \frac{\mathbf{1}}{n}\sum_{i=\mathbf{1}}^{N}\sum_{j>i}^{N} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^{\mathbf{2}}$, which is the variance of the estimate of the population total under the probability proportional to size (PPS) sampling with replacement.

The REs for the design-based πPS sampling methods of Mizuno or Brewer are also computed by (5.6), and with a distinction, the REs are denoted by $RE_{d,\pi PS}$ instead of $RE_{\xi,\pi PS}$. The RE for Murthy's method, which is a non-πPS sampling method, is calculated by:

$$RE_M = \left[Var_{PPS}\left(\hat{Y}\right) \Big/ Var_M\left(\hat{Y}\right) \right] \times \mathbf{100} \,, \tag{5.7}$$

where

$$Var_M\left(\hat{Y}\right) = \sum_{i=\mathbf{1}}^{N}\sum_{j>i}^{N} \frac{p_i p_j (\mathbf{1} - p_i - p_j)}{\mathbf{2} - p_i - p_j} \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^{\mathbf{2}} \tag{5.8}$$

According to the empirical study of Rao and Bayless (1969), for 18 populations, PPS sampling with replacement always had a larger variance than Brewer's method. Also, it is theoretically clear that $Var_M\left(\hat{Y}\right) < Var_{PPS}\left(\hat{Y}\right)$.

The frequencies in column "f" in the table denote the number of populations where

$$RE_{\xi,\pi PS} > RE_{d,\pi PS} \tag{5.9}$$

or

$$RE_{\xi,\pi PS} > RE_M \tag{5.10}$$

For example, the first "16" in terms of "OP1 M" and "Mizuno" in the column of "f" in the table indicates that of 18 populations, 16 populations satisfy (6.9). More specifically, for 16 populations, the REs for model-based sampling using "OP1 M" are larger than in design-based sampling of Mizuno, whereas for 2 populations they are smaller.

The frequencies in "f1," "f2," and "f3" in the table respectively denote the number of populations that are

$$0 < RE_{\xi,\pi PS} - RE_{d,\pi PS} \leq 10 \tag{5.11}$$

or

$$0 < RE_{\xi,\pi PS} - RE_M \leq 10 \,, \tag{5.12}$$

$$11 \leq RE_{\xi,\pi PS} - RE_{d,\pi PS} \leq 20 \tag{5.13}$$

or

$$11 \leq RE_{\xi,\pi PS} - RE_M \leq 20 \,, \tag{5.14}$$

and

$$RE_{\xi,\pi PS} - RE_{d,\pi PS} \geq 21 \tag{5.15}$$

Table 1: Comparison of frequency of populations that model-based sampling shows a better efficiency than design-based sampling

Design-based	Model-based	c=0				c=0.1				c=0.2				c=0.3				c=0.4				c=0.5			
		f	f1	f2	f3	f	f1	f2	f3	f	f1	f2	f3	f	f1	f2	f3	f	f1	f2	f3	f	f1	f2	f3
Mizuno	OP1 M	16**	2	3	11	16**	2	1	13	16**	5	0	11	14**	2	1	11	14**	3	1	10	13**	3	1	9
	OP1 R	16**	0	4	12	16**	3	1	12	16**	4	1	11	14**	2	1	11	15**	4	1	10	13**	2	2	9
	OP2 M	14**	3	1	10	13**	2	1	10	14**	2	2	10	13**	3	1	9	14**	2	2	10	13**	2	2	9
	OP2 R	14**	3	1	10	13**	2	1	10	14**	2	2	10	13**	3	1	9	14**	2	2	10	13**	2	2	9
	OP3	15**	0	4	11	16**	3	1	12	16**	3	2	11	14**	2	1	11	14**	3	1	10	13**	3	1	9
	OP4 M	14**	3	1	10	13**	2	1	10	14**	3	1	10	13**	2	2	9	14**	3	1	10	13**	2	2	9
	OP4 R	15**	4	1	10	14**	3	1	10	14**	2	2	10	13**	2	2	9	14**	3	1	10	12**	1	2	9
Brewer	OP1 M	12**	5	6	1	12**	5	3	4	8	4	4	0	12**	9	0	3	6	5	1	0	5	5	0	0
	OP1 R	12**	5	5	2	10**	3	3	4	8	3	5	0	8	4	1	3	7	6	1	0	6	6	0	0
	OP2 M	8	6	2	0	8	5	3	0	9*	7	2	0	7	7	0	0	7	7	0	0	8	8	0	0
	OP2 R	8	6	1	1	6	4	2	0	10**	9	1	0	8	8	0	0	9*	9	0	0	7	7	0	0
	OP3	10**	6	2	2	11**	7	1	3	13**	11	2	0	9*	7	1	1	9*	7	2	0	9*	9	0	0
	OP4 M	10**	8	2	0	11**	9	2	0	8	6	2	0	8	8	0	0	6	5	1	0	7	7	0	0
	OP4 R	9*	7	2	0	10**	8	2	0	9*	7	2	0	8	8	0	0	7	7	0	0	9*	9	0	0
Murthy	OP1 M	8	3	4	1	8	2	3	3	6	4	2	0	7	4	0	3	5	3	2	0	3	3	0	0
	OP1 R	9*	3	4	2	8	2	3	3	7	3	4	0	7	3	1	3	5	3	2	0	4	4	0	0
	OP2 M	6	3	3	0	6	4	2	0	8	7	1	0	3	3	0	0	5	5	0	0	6	6	0	0
	OP2 R	6	3	3	0	5	4	1	0	8	7	1	0	3	3	0	0	5	5	0	0	6	6	0	0
	OP3	9*	5	2	2	10**	7	1	2	8	6	2	0	7	6	0	1	9*	7	2	0	6	6	0	0
	OP4 M	7	5	2	0	8	7	1	0	5	3	2	0	4	4	0	0	4	3	1	0	7	7	0	0
	OP4 R	7	5	2	0	8	7	1	0	6	4	2	0	3	3	0	0	4	4	0	0	7	7	0	0

*exactly half of 18 populations, **over half of 18 populations or

$$RE_{\xi,\pi PS} - RE_M \geq 21\,. \tag{5.16}$$

Here, (5.11) or (5.12), (5.13) or (5.14), and (5.15) or (5.16) denote that the REs on model-based sampling are respectively "slightly better," "much better," and "very much better," than those on design-based sampling. Note that f = f1 + f2 + f3. For example, the first "2" in the column of "f1," the first "3" in "f2," and the first "11" in "f3" in the table indicates that of 16 populations in "f," 2 populations satisfy (5.11), 3 populations do (5.13), and 11 populations do (5.15).

The findings from Table 1 are summarized as follows:

(1) Model-based sampling methods (using OP1 M, OP1 R, OP2 M, OP2 R, OP3, OP4 M, and OP4 R) are consistently more efficient than Mizuno's method, regardless of the value of c. For at least half of 18 populations, they show "very much better" efficiency.

(2) When the value of c is low, model-based sampling methods are overall more efficient relative to Brewer's method. For some populations, when the value of c is low, the methods using OP1 or OP3 show "very much better" efficiency. Model-based sampling using OP3 consistently shows a better efficiency than the other model-based methods. .

(3) Model-based sampling method using OP3 compares favorably with the method of Murthy, when the value of c is low, and for some populations, it has "very much better" efficiency as well as "much better" efficiency. Other model-based sampling methods are less efficient than the one of Murthy.

(4) ML estimation and REML estimation give different estimates of the model in (4.3), and it seems that model-based methods using optimization problems involving these different estimates of the model may yield different efficiencies.

(5) For model-based sampling methods, there is a trade-off between the reduction of variance and the stability of the variance estimator because the REs tend to be reduced as the value of c is increased.

7. Conclusion Remarks

We have suggested two model-based πPS sampling strategies using the optimization problems of OP3 and OP4. The method using OP3 is empirically preferable to the method using OP4, as well as the previous methods using OP1 and OP2. Compared to others, OP3 is the simpler optimization problem, and it does not depend on the parameters in the superpopulation model.

Those four model-based πPS sampling methods are flexible in terms of the choice of sampling design because one may choose the value of c, which is related to the stability of variance estimator. But one should be careful in choosing the value, since there is a trade-off between the variance reduction and the stability of the variance estimator. With regard to the efficiency, regardless of the value of c, the model-based methods are shown empirically to be superior to design-based πPS sampling of Mizuno, and when the value of c is low, they are preferable to the one of Brewer. Also, in such a case, the method using OP3 is comparable to the method of Murthy.

References

Brewer, K. R. W. (1963). "A model of systematic sampling with unequal probabilities," *Australian Journal of Statistics*, 5, 5-13.

Brewer, K. R. W. and Hanif, M. (1983). *Sampling with Unequal Probabilities*, New York: Springer-Verlag.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed. New York: Wiley.

Godambe, V. P. (1955). "A unified theory of sampling from finite populations," *Journal of the Royal Statistical Society*, B17, 269-278.

Hansen, M. H., and Hurwitz, W. N. (1943). "On the theory of sampling from finite populations," *Annals of Mathematical Statistics*, 14, 333-362.

Horvitz, D. G. and Thompson, D. J. (1952). "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47, 663-685.

Isaki, C. T. and Fuller, W. A. (1982). "Survey design under the regression superpopulation model." *Journal of the American Statistical Association*, 77, 89-96.

Kim, S. W., Heeringa, S. G., and Solenberger, P. W. (2006). "Model-based sampling designs for optimum estimation," In *JSM Proceedings*, Survey Research Methods Section, American Statistical Association, 3245-3252.

Mizuno, H. (1952). "On the sampling system with probability proportional to sum of sizes," *Annals of the Institute of Statistical Mathematics*, 3, 99-107.

Murthy, M. N. (1957). "Ordered and unordered estimators in sampling without replacement," *Sankhya*, 18, 379-390.

Nigam, A. K., Kumar, P., and Gupta, V. K. (1984). "Some methods of inclusion probability proportional to size sampling," *Journal of the Royal Statistical Society*, B46, 564-571.

PASW Statistics (2009). *PASW Complex Samples*, Version 18, Chicago, IL: SPSS Inc. http://www.spss.com/media/collateral/statistics/complex-samples.pdf

Raj, D. (1956). "A note on the determination of optimum probabilities in sampling without replacement," *Sankhya*, 17, 197-200.

Rao, J. N. K. and Bayless, D. L. (1969). "An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum," *Journal of the American Statistical Association*, 64, 540-559.

Rao, J. N. K. and Nigam, A. K. (1992). "Optimal controlled sampling: a unified approach," *International Statistical Review*, 60, 89-98.

Sampford, M. R. (1967). "On sampling without replacement with unequal probabilities of selection," *Biometrika*, 54, 499-513.

Särndal, C. E. (1996). "Efficient estimators with simple variance in unequal probability sampling," *Journal of the American Statistical Association*, 91, 1289-1300.

SAS/OR (2008). *User's Guide: Mathematical Programming*, Version 9.2, Cary, NC: SAS Institute Inc.

SAS/STAT (2009). *User's Guide: Survey Data Analysis*, Version 9.2, Cary, NC: SAS Institute Inc.