Model-Based Sampling Designs for Optimum Estimation

Sun-Woong Kim, Dongguk University, Steven G. Heeringa and Peter W. Solenberger, University of Michigan
Sun-Woong Kim, Dongguk University, Jung-Gu Pil-Dong 3 Ga 26, Seoul, South Korea 100-715
Steven Heeringa, Institute for Social Research, 426 Thompson Street, Ann Arbor, Michigan 48106

Key Words: pPS sampling, superpopulation model, model-based sampling, expected variance, optimization problem, linear programming, nonlinear programming

1. Introduction

With respect to survey sampling inferences for finite populations, model-dependent inferences require a superpopulation model. Consequently, we must be concerned over how departures from model assumptions may affect estimation and inference for the target population. But for sampling designs based on models, Hansen, Madow, and Tepping (1983) make the following statement:

"As denoted before, models may also be used to produce model-based designs that are not model-dependent. For example, models of the population may suggest useful procedures for selecting the sample or the estimators. This is often done in probability sampling to great advantage. Thus model-based designs do not need to be model-dependent." (pages 778-779)

"Models are appropriately used to guide and evaluate the design of probability samples, but with large samples the inferences should not depend on the model." (page 791)

These paragraphs deliver an important message on the significant roles and usefulness of population models in model-based sampling designs.

Many sampling schemes have been developed, and one class of popular sampling methods for both model-dependent inferences and design-based inferences may be the traditional pPS (inclusion probability proportional to size) sampling methods. The pPS designs provide efficient estimation in varying probability selections when reasonably good measures of size are available. However, there has been limited research on model-based approaches to traditional pPS sampling designs, and there is a limited statistical literature on this topic.

Raj (1956) suggested a pPS sampling design to minimize the variance of the Horvitz-Thompson (H-T) estimator under a simple model. This method may be called the origin of model-based pps sampling designs. Hanurav (1967) showed that Raj's (1956) method is not valid due to the use of an inappropriate model and exemplified the use of a different model. Rao and Bayless (1969, 1970) empirically compared the expected (or average) variances of several estimators under a superpopulation model.

One of the main differences in the linear models considered in those earlier papers is of whether the line passes through the origin or not. The models that Raj (1956) and Hanurav (1967) used involved the intercept, while Rao and Bayless (1969, 1970) model has a zero intercept.

In fact, the presence of the intercept in the superpopulation model for a variable of interest, y, has been one of the key issues in the debate surrounding model-dependent versus design-based inferences (See Section 2, Hansen, Madow, and Tepping (1983) and Section 3.7, Valliant, Dorfman, and Royall (2000)).

In this paper, we first prove that different population models involving the intercept result in the same optimization problem, and therefore the same model-based pPS sampling design. Second, we revisit the formula of the expected variance of the H-T estimator presented by Rao and Bayless (1969, 1970), which indicates all model-based pPS sampling designs have the same expected variance. Third, we present the expected variance of H-T estimator under a superpopulation model with the intercept and a corresponding optimal pPS sampling design for minimizing this variance. Finally, we propose a model-based pPS sampling design based on the optimization problem for minimizing possible variance estimates subject to constraints on the desirable properties of those estimates.

2. Background

Consider a finite population of N units, denoted by $U = \{u_1, u_2, \cdots, u_N\}$. Let s be a selected sample of size n from U with a sampling plan $P(\cdot)$. Then we define the first-order inclusion probabilities p_i, given by

$$p_i = P(u_i \in s \cup s \in S), \; i = 1, \cdots, N, \qquad (2.1)$$

where S is the collection of all possible samples selected from U.

Also, we define the joint probabilities (or the second-order inclusion probabilities) as follows:

$$p_{ij} = P(u_i \& u_j \in s \cup s \in S), \; i \neq j = 1, \cdots, N \qquad (2.2)$$

Let y_i be the value of the characteristic of the unit u_i in U. We may prefer using the H-T (1952) estimator, which is unbiased for estimating the population total $Y = \sum_{i=1}^{N} y_i$. The H-T estimator is given by:

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{p_i} \qquad (2.3)$$

According to the method of sample selection, the p_i in (2.3) may be replaced by

$$p_i = np_i, \qquad (2.4)$$

where $p_i = x_i / X$, $X = \sum_{i=1}^{N} x_i$, and x_i is a auxiliary variable correlated with the y_i.

The sampling design satisfying (2.4) is called a pPS design. If the x_i is exactly proportional to the y_i, then the variance of the corresponding H-T estimator is zero. In principle, survey samplers may choose a measure of size, x_i, that is thought to be highly correlated with the y_i and select a sample by a sampling mechanism or sampling plan that assigns each population element a selection probability based on this measure of size.

Most pPS sampling designs including the long-standing methods of Brewer (1963) and Sampford (1967) and comparatively new approaches of Kim, Herringa, and Solenberger (2003, 2004, 2005) may not be efficient for some populations due to the fact that selection probability depends only on the available measure of size, x_i. In section 3, we show that by using population models, more elaborate sampling designs can be established to increase the precision of the H-T estimator.

3. Considering Population Models for Sampling Designs

If we know enough about a population, the population model can guide the choice of the sample design, resulting in increased precision for estimates. Of course, the model-based sample design should be developed by keeping in mind an appropriate estimator such as the H-T estimator. Then the main question becomes, given the population model assumptions, what sampling design is optimal for the H-T estimator?

We define a model-based pPS sampling design that consists of (a) an evaluated quantity (e.g., expected variance of the estimator of Horvitz and Thompson (H-T) (1952)) under an assumed population model, and (b) a sampling plan that each sample is selected with the probability to optimize the quantity.

Assume that from experience with the population to be studied we have enough information on the x_i, correlated with the y_i, to specify a population model.

As the first approach for finding a model-based pPS sampling design, we may choose Raj (1956)' s method for the case of $n = 2$. Raj assumed the following model:

$$y_i = a + b x_i, \quad i = 1, \cdots, N, \qquad (3.1)$$

The linear model includes no error term and no knowledge of the values of a and b is assumed. This model simply indicates that the relation between y_i and x_i is a straight line.

Raj(1956) considered the following form of the variance of the H-T estimator.

$$Var_1\left(\hat{Y}_{HT}\right) = \sum_{i=1}^{N} \frac{y_i^2}{p_i} + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{p_{ij}}{p_i p_j} y_i y_j - Y^2 \qquad (3.2)$$

He showed that since the first and third terms are fixed under the assumed model, the problem of minimizing (3.2) reduces to the optimization problem:

$$Minimize \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{p_{ij}}{p_i p_j} \qquad (3.3)$$

subject to

$$\sum_{j \neq i} p_{ij} = p_i, \quad i = 1, \cdots, N. \qquad (3.4)$$

Note that (3.3) and (3.4) constitute a linear programming (LP) problem and the sums in (3.4) are the linear constraints for achieving pPS sampling designs. Raj gave an illustration to obtain the optimum sampling plan $P_x(\cdot)$ based on the model (3.1) for the three populations A, B and C originally given by Yates and Grundy (1953).

In addition to (3.2), there exist various forms of the variance of the H-T estimator. For example, one alternative form of the variance is:

$$Var_2\left(\hat{Y}_{HT}\right) = \sum_{i=1}^{N} \frac{y_i^2(1 - p_i)}{p_i} + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{y_i y_j (p_{ij} - p_i p_j)}{p_i p_j} \qquad (3.5)$$

Since (3.5) can be expressed as:

$$\sum_{i=1}^{N} \frac{y_i^2(1 - p_i)}{p_i} + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{p_{ij}}{p_i p_j} y_i y_j - 2\sum_{i=1}^{N}\sum_{j>i}^{N} y_i y_j, \qquad (3.6)$$

the minimization of (3.6) reduces to the minimization of the same second term as in (3.2) in Raj's approach. But an alternative LP problem can be obtained by the following theorem.

Theorem 3.1. Under the population model (3.1), the minimization of the variance of the H-T estimator given in (3.6) is equivalent to minimizing

$$\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{a+b(x_i+x_j)}{x_i x_j}p_{ij} \tag{3.7}$$

Proof. Since the first and third terms in (3.6) are based on N units in a population, they are fixed under the assumed model. We know that under the pPS sampling design,

$$p_i=\frac{nx_i}{X}, \tag{3.8}$$

as denoted by (2.4) .
On substituting the p_i in the second term in (3.6), we have

$$\frac{2X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{a^2+ab(x_i+x_j)+b^2x_ix_j}{x_ix_j}p_{ij}$$

$$=\frac{2aX^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{a+b(x_i+x_j)}{x_ix_j}p_{ij}+\frac{2b^2X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}p_{ij}$$

$$=\frac{X^2}{n}\left[\frac{2a}{n}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{a+b(x_i+x_j)}{x_ix_j}p_{ij}+b^2(n-1)\right] \tag{3.9}$$

Since all terms in (3.9) except for (3.7) are fixed, the minimization of (3.9) leads to minimization of (3.7). This completes the proof.

Remark 3.1. (3.7) can be regarded as the objective function to minimize the variance of the H-T estimator under the model (3.1).

Remark 3.2. If $a=0$ ($b\neq 0$), then the second term reduces to $b^2X^2(n-1)/n$, which does not depend on the joint probabilities. Thus any pPS sampling design produces the same variance under the model
If $b=0$ ($a\neq 0$), then the objective function reduces to simply:

$$\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{p_{ij}}{x_ix_j} \tag{3.10}$$

Note that (3.10) is equivalent to (3.3) due to (3.8).

Based on Theorem 3.1., the following LP problem to minimize the variance of the H-T estimator can be established under the model (3.1).

$$Minimize\ \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{a+b(x_i+x_j)}{x_ix_j}p_{ij} \tag{3.11}$$

subject to

$$\sum_{j\neq i}p_{ij}=(n-1)p_i\ ,\quad i=1,\cdots,N\ . \tag{3.12}$$

Now consider different models. First, assume the following superpopulation model 1:

$$y_i=a+b\,x_i+e_i\ ,\ i=1,\cdots,N\ , \tag{3.13}$$

where $E_x(y_i)=a\,x_i+b$, $V_x(y_i)=s^2x_i^2$, and $Cov_x(y_i,y_j)=0$. Here E_x denotes the model expectation over all the finite populations that can be drawn from the superpopulation.

The model (3.13) was used by Hanurav (1967) who pointed out several problems with Raj's (1956) approach, and asserted that under the model the expected variance of H-T estimator is minimized for all a and b for given values of x_i's and p_i's, if and only if

$$\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{x_ix_j}{p_ip_j}p_{ij}, \tag{3.14}$$

$$\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{x_i+x_j}{p_ip_j}p_{ij}, \tag{3.15}$$

and

$$\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{p_{ij}}{p_ip_j} \tag{3.16}$$

are simultaneously minimized.

But minimizing the three problems (3.14), (3.15), and (3.16) at the same time is not valid. For example, since (3.14) reduces to a constant under pPS sampling,

$$\frac{X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}p_{ij}=\frac{X^2(n-1)}{2n}\ , \tag{3.17}$$

and (3.14) is not necessary to minimize this component under pPS sampling designs. Also, we show that there exists an optimization problem unlike the above that involves (3.14), (3.15), and (3.16).

Theorem 3.2. Consider the form of variance of H-T estimator given by (3.6) and let superpopulation model (3.13) hold. Then the expected variance under the model is

$$\frac{X^2}{n}\left[\frac{2a}{n}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{a+b(x_i+x_j)}{x_i x_j}p_{ij}+b^2(n-1)\right] +\sum_{i=1}^{N}\left(X/nx_i-1\right)\left(a^2+(b^2+s^2)x_i^2+2ab\,x_i\right) -2\sum_{i}^{N}\sum_{j>i}^{N}\left(a^2+ab(x_i+x_j)+b^2x_ix_j\right) \tag{3.18}$$

Proof. For the first and third terms in (3.6) under the model, we have

$$\begin{aligned} &E_x\left[\sum_{i=1}^{N}\frac{y_i^2(1-p_i)}{p_i}-2\sum_{i=1}^{N}\sum_{j>i}^{N}y_iy_j\right] \\ &=\sum_{i=1}^{N}(X/nx_i-1)E_x(y_i^2)-2\sum_{i=1}^{N}\sum_{j>i}^{N}E_x(y_i)E_x(y_j) \\ &=\sum_{i=1}^{N}(X/nx_i-1)\left(s^2x_i^2+(a+bx_i)^2\right)-2\sum_{i=1}^{N}\sum_{j>i}^{N}E_x(y_i)E_x(y_j) \\ &\qquad(\because V_x(y_i)=E_x(y_i^2)-(E_x(y_i))^2=s^2x_i^2) \\ &=\sum_{i=1}^{N}\left(X/nx_i-1\right)\left(a^2+(b^2+s^2)x_i^2+2ab\,x_i\right) \\ &\qquad-2\sum_{i}^{N}\sum_{j>i}^{N}\left(a^2+ab(x_i+x_j)+b^2x_ix_j\right) \end{aligned} \tag{3.19}$$

In terms of the second term in (3.6) under the model, we have

$$\begin{aligned} &E_x\left[\frac{2X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{y_iy_j}{x_ix_j}p_{ij}\right]=\frac{2X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{E_x(y_iy_j)}{x_ix_j}p_{ij} \\ &=\frac{2X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{E_x(y_i)E_x(y_j)}{x_ix_j}p_{ij} \\ &\qquad(\because Cov_x(y_i,y_j)=0) \\ &=2\frac{X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{a^2+ab(x_i+x_j)+b^2x_ix_j}{x_ix_j}p_{ij} \\ &=\frac{2aX^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{a+b(x_i+x_j)}{x_ix_j}p_{ij}+\frac{2b^2X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}p_{ij} \\ &=\frac{X^2}{n}\left[\frac{2a}{n}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{a+b(x_i+x_j)}{x_ix_j}p_{ij}+b^2(n-1)\right] \end{aligned} \tag{3.20}$$

Hence (3.18) follows from (3.19) and (3.20).

Remark 3.3. Since the first term in (3.18) is the same as (3.9) and the other terms are fixed, the same LP problem specified by (3.11) and (3.12) can be established to minimize the variance of the H-T estimator. In cases of $n=2$, this problem will be the minimization of the sum of the weighted selection probability for each sample. That is,

$$Minimize \quad \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{a+b(x_i+x_j)}{x_ix_j}p(s) \tag{3.21}$$

subject to

$$\sum_{j\neq i}p(s)=p_i, \quad i=1,\cdots,N \tag{3.22}$$

Also, consider the following superpopulation model 2:

$$y_i=a+b\,x_i+e_i, \quad i=1,\cdots,N, \tag{3.23}$$

where $E_x(e_i)=0$, $V_x(e_i)=s^2x_i^2$, and $E_x(e_i,e_j)=0$.

Although this model looks like a different model than (3.13), they coincide. Note that $E_x(e_i,e_j)=0$ gives $E_x(y_iy_j)=E_x(y_i)E_x(y_j)$, resulting in $Cov_x(y_i,y_j)=0$. Hence the same expected variance as under the model (3.13) and optimization problem is obtained.

Remark 3.4. The three models mentioned above give the same optimization problem to minimize the model expectation of the design-based variance of the H-T estimator. Solving the optimization problem by using some linear programming software would provide an optimal sampling plan $P_x(\cdot)$. A variety of software is available for this problem, but using a powerful software may be recommended because of a number of the unknowns in the optimization problem

Remark 3.5. Under the models discussed above, if $a=0$, the model expectation of the variance is fixed without depending on the joint probabilities p_{ij}. Thus any pPS sampling design is acceptable.

In addition to the forms of the variance of the H-T estimator such as (3.2), (3.5), and (3.6), we may consider one of the popular forms given by

$$Var_3\left(\hat{Y}_{HT}\right)=\sum_{i=1}^{N}\sum_{j>i}^{N}\left(p_ip_j-p_{ij}\right)\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2, \tag{3.24}$$

which is equivalent to Var_1 or Var_2.

Rao and Bayless (1969) (1970) provided results from an empirical study based on a super-population approach for examining the efficiencies of the several estimators including the H-T estimator. They assumed the following superpopulation model with no intercept:

$$y_i = b\,x_i + e_i,\ \ i = 1, \cdots, N, \qquad (3.25)$$

where $E_x(e_i) = 0$, $E_x(e_i^2) = ax_i^g$ ($a > 0$, $g \geq 0$), and $E_x(e_i e_j) = 0$.

Here we call this model the superpopulation model 3. They showed that the average variance of the H-T estimator under this model is as follows:

$$E_x Var_3 = \frac{aX^g}{n} \sum_i^N (1 - np_i) p_i^{g-1} = V_x \qquad (3.26)$$

Note that V_x does not depend on the p_{ij}. Hence all model-based pPS sampling designs using H-T estimator have the same expected variance under the model and any pPS design is good for minimizing the variance.

Since they did not provide the details of the proof for (3.26), we give them as follows:

Consider a different form of Var_3 given by

$$Var_4\left(\hat{Y}_{HT}\right) = \sum_{i=1}^{N} \sum_{j>i}^{N} \left(p_i p_j - \frac{p_{ij}}{n^2}\right) \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2 \qquad (3.27)$$

When taking the model expectation for Var_4, we have

$$E_x\left(Var_4\left(\hat{Y}_{HT}\right)\right) = \sum_{i=1}^{N} \sum_{j>i}^{N} \left(p_i p_j - \frac{p_{ij}}{n^2}\right) E_x\left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2 \qquad (3.28)$$

Noting

$$V_x(y_i) = ax_i^g, \qquad (3.29)$$

$$E_x y_i^2 = ax_i^g + b^2 x_i^2, \qquad (3.30)$$

and

$$E_x(y_i y_j) = b^2 x_i x_j, \qquad (3.31)$$

we have

$$E_x\left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2 = \frac{E_x(y_i^2)}{p_i^2} + \frac{E_x(y_i^2)}{p_i^2} - 2\frac{E_x(y_i y_j)}{p_i p_j}$$

$$= \frac{2}{p_i^2}(ax_i^g + b^2 x_i^2) - \frac{2}{p_i p_j} b^2 x_i x_j$$

$$= 2aX^2 x_i^{g-2} + 2b^2 X^2 - 2b^2 X^2$$

$$= 2aX^2 x_i^{g-2}$$

$$= 2aX^g p_i^{g-2} \qquad (3.32)$$

Then

$$E_x\left(Var_4\left(\hat{Y}_{HT}\right)\right) = 2aX^g \sum_{i=1}^{N} \sum_{j>i}^{N} p_i^{g-2}\left(p_i p_j - \frac{p_{ij}}{n^2}\right)$$

$$= 2aX^g \sum_{i=1}^{N} \sum_{j>i}^{N} \left(p_j - \frac{p_{ij}}{n^2 p_i}\right) p_i^{g-1}$$

$$= 2aX^g \frac{1}{2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \left(p_j - \frac{p_{ij}}{n^2 p_i}\right) p_i^{g-1}$$

$$= aX^g \left(\sum_{i=1}^{N} p_i^{g-1} \sum_{j \neq i}^{N} p_j - \sum_{i=1}^{N} p_i^{g-1} \frac{1}{n^2 p_i} \sum_{j \neq i} p_{ij}\right)$$

$$= aX^g \left(\sum_{i=1}^{N} p_i^{g-1} \sum_{j \neq i}^{N} p_j - \sum_{i=1}^{N} p_i^{g-1} \frac{1}{n^2 p_i} n(n-1) p_i\right)$$

$$= aX^g \left(\sum_{i=1}^{N} p_i^{g-1}(1 - p_i) - \frac{n-1}{n} \sum_{i=1}^{N} p_i^{g-1}\right)$$

$$= aX^g \sum_{i=1}^{N} \left(1 - p_i - \frac{n-1}{n}\right) p_i^{g-1}$$

$$= \frac{aX^g}{n} \sum_{i=1}^{N} (1 - np_i) p_i^{g-1} \qquad (3.33)$$

Hence (3.26) follows.

Note that if $g = 1$ and $a = s^2$, then

$$E_x\left(Var_4\left(\hat{Y}_{HT}\right)\right) = \frac{N-n}{n} X s^2 \qquad (3.34)$$

Now, consider the superpopulation model 4 with intercept term as follows:

$$y_i = a + b\,x_i + e_i,\ \ i = 1, \cdots, N, \qquad (3.35)$$

where $E_x(e_i) = 0$, $E_x(e_i^2) = ax_i^g$ ($a > 0$, $g \geq 0$), and $E_x(e_i e_j) = 0$.

Although the only difference between the model 3 and the model 4 is the intercept, obtaining $E_x\left(Var_4\left(\hat{Y}_{HT}\right)\right)$ under model 4 becomes more complicated. as described below.

Theorem 3.3. Assume the population model given in (3.35) and consider the variance of H-T estimator given by (3.27). Then the model expectation of the variance under the model is given by

$$E_x\left(Var_4\left(\hat{Y}_{HT}\right)\right)$$

$$= V_x + 2a\left(\sum_{i=1}^{N}\sum_{j>i}^{N}\left(x_j - x_i\right)\left(a x_i^{-1} + b\right)\right)$$

$$+ \frac{2a X^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\left(x_j^{-1} - x_i^{-1}\right)\left(a x_i^{-1} + b\right)p_{ij}\,, \tag{3.36}$$

where $V_x = \dfrac{aX^g}{n}\sum_{i=1}^{N}(1 - np_i)p_i^{g-1}$.

Proof. Since

$$E_x y_i^2 = a x_i^g + a^2 + b^2 x_i^2 + 2ab\,x_i \tag{3.37}$$

and

$$E_x(y_i y_j) = a^2 + ab(x_i + x_j) + b^2 x_i x_j\,, \tag{3.38}$$

we have

$$E_x\left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2 = \frac{E_x(y_i^2)}{p_i^2} + \frac{E_x(y_i^2)}{p_i^2} - 2\frac{E_x(y_i y_j)}{p_i p_j}$$

$$= \frac{2}{p_i^2}(a x_i^g + a^2 + b^2 x_i^2 + 2ab\,x_i)$$

$$- \frac{2}{p_i p_j}(a^2 + ab(x_i + x_j) + b^2 x_i x_j)$$

$$= 2X^2(a x_i^{g-2} + a^2 x_i^{-2} + 2ab\,x_i^{-1})$$

$$- 2X^2\left(a^2\frac{1}{x_i x_j} + ab\frac{x_i + x_j}{x_i x_j}\right)$$

$$= 2X^2\left(a x_i^{g-2} + a^2\frac{x_j - x_i}{x_i^2 x_j} + ab\frac{x_j - x_i}{x_i x_j}\right)$$

$$= 2X^2 a x_i^{g-2} + 2aX^2\frac{x_j - x_i}{x_i x_j}\left(a x_i^{-1} + b\right)$$

$$= 2aX^g p_i^{g-2} + 2aX^2\frac{x_j - x_i}{x_i x_j}\left(a x_i^{-1} + b\right) \tag{3.39}$$

Then we have

$$E_x\left(Var_4\left(\hat{Y}_{HT}\right)\right) = E_x\sum_{i=1}^{N}\sum_{j>i}^{N}p_i p_j\left(\frac{p_{ij}}{n^2} - \frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2$$

$$= 2aX^g\sum_{i=1}^{N}\sum_{j>i}^{N}p_i^{g-2}p_i p_j\left(\frac{p_{ij}}{n^2}\right)$$

$$+ 2aX^2\left(\sum_{i=1}^{N}\sum_{j>i}^{N}p_i p_j\left(\frac{p_{ij}}{n^2} - \frac{x_j - x_i}{x_i x_j}\left(a x_i^{-1} + b\right)\right)\right)$$

$$= \frac{aX^g}{n}\sum_{i=1}^{N}(1 - np_i)p_i^{g-1}$$

$$+ 2aX^2\left(\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{x_i x_j}{X^2}\frac{x_j - x_i}{x_i x_j}\left(a x_i^{-1} + b\right)\right)$$

$$- \frac{aX^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{x_j - x_i}{x_i x_j}\left(a x_i^{-1} + b\right)p_{ij} \qquad (\because (3.33))$$

$$= V_x + 2a\left(\sum_{i=1}^{N}\sum_{j>i}^{N}\left(x_j - x_i\right)\left(a x_i^{-1} + b\right)\right)$$

$$+ \frac{aX^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\left(x_j^{-1} - x_i^{-1}\right)\left(a x_i^{-1} + b\right)p_{ij}$$

Hence (3.36) follows.

Corollary 3.1. If $a = 0$, then the model expectation of the variance of the H-T estimator under the population model (3.35), V_x^*, reduces to V_x.

Corollary 3.2. Even if the true population model does not involve the intercept a unlike (3.35), the pPS sampling designs obtained from the optimization problem consisting of (3.40), (3.41), and (3.42) have the same model expectation as the variance of the H-T estimator under any other pPS sampling designs.

Because the first and second terms are fixed in (3.36) under the model (3.35), we may consider the following optimization problem to minimize the model expectation of the variance of the H-T estimator.

$$Minimize \sum_{i=1}^{N}\sum_{j>i}^{N}(x_j^{-1} - x_i^{-1})\left(a x_i^{-1} + b\right)p_{ij} \tag{3.40}$$

under the linear inequality constraints

$$c p_i p_j \le p_{ij} \le p_i p_j\,,\ \ j > i = 1, \cdots, N\,, \tag{3.41}$$

where c is a real number between 0 and 1,

and

$$\sum_{j \ne i}^{N} p_{ij} = (n-1)p_i\,,\ \ i = 1, \cdots, N\,. \tag{3.42}$$

Remark 3.6. The optimization problem consisting of (3.40), (3.41), and (3.42), based on V_x^*, does not depend on $a(>0)$ or $g(\ge 0)$ under the model (3.35).

Next, consider model-based pPS sampling designs for minimizing expected variance estimates under a superpopulation model. This approach may be encouraged to reduce the length of the confidence

intervals based on the model of a population to be sampled.

The variance estimator of the H-T estimator given by Sen (1953) and Yates and Grundy (1953), termed the Sen-Yates-Grundy (S-Y-G) estimator is denoted by

$$var_{SYG}(\hat{Y}_{HT}) = \sum_{i=1}^{n}\sum_{j>i}^{n} \frac{p_i p_j - p_{ij}}{p_{ij}}\left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2, \quad (3.43)$$

which is expressed as a different form

$$var_{SYG}(\hat{Y}_{HT}) = \sum_{i=1}^{n}\sum_{j>i}^{n}\left(\frac{p_i p_j}{p_{ij}} - \frac{1}{n^2}\right)\left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2 \quad (3.44)$$

Theorem 3.4. Consider the variance estimator of H-T estimator given by (3.44) and let a population model be (3.35). Then the model expectation of the variance estimator under the model (3.35) is

$$E_x\left(var_{SYG}(\hat{Y}_{HT})\right) = 2\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{x_{ij}}{p_{ij}} - \frac{2aX^g}{n^2}\sum_{i=1}^{n}\sum_{j>i}^{n} p_i^{g-2} - \frac{2a\,X^2}{n^2}\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{x_j - x_i}{x_i x_j}\left(a\,x_i^{-1} + b\right) \quad (3.45)$$

where $x_{ij} = a x_i^{g-1} x_j + a(x_j - x_i)\left(a\,x_i^{-1} + b\right)$. (3.46)

Proof. Taking the model expectation for the S-Y-G variance estimator, we have

$$\begin{aligned} E_x\left(var_{SYG}(\hat{Y}_{HT})\right) &= E_x\sum_{i=1}^{n}\sum_{j>i}^{n}\left(\frac{p_i p_j}{p_{ij}} - \frac{1}{n^2}\right)\left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2 \\ &= 2aX^g\sum_{i=1}^{n}\sum_{j>i}^{n} p_i^{g-2}\left(\frac{p_i p_j}{p_{ij}} - \frac{1}{n^2}\right) \\ &\quad + 2a\,X^2\left(\sum_{i=1}^{n}\sum_{j>i}^{n}\left(\frac{p_i p_j}{p_{ij}} - \frac{1}{n^2}\right)\frac{x_j - x_i}{x_i x_j}\left(a\,x_i^{-1} + b\right)\right) \\ &= 2aX^g\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{p_i^{g-1}p_j}{p_{ij}} - \frac{2aX^g}{n^2}\sum_{i=1}^{n}\sum_{j>i}^{n} p_i^{g-2} \\ &\quad + 2a\,X^2\sum_{i=1}^{n}\sum_{j>i}^{n}\left(\frac{(x_i x_j)/X^2}{p_{ij}} - \frac{x_j - x_i}{x_i x_j}\left(a\,x_i^{-1} + b\right)\right) \\ &\quad - \frac{2}{n^2}a\,X^2\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{x_j - x_i}{x_i x_j}\left(a\,x_i^{-1} + b\right) \end{aligned} \quad (3.47)$$

The first and third term in (3.47) are re-expressed as

$$2\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{x_{ij}}{p_{ij}}, \quad (3.48)$$

where $x_{ij} = a x_i^{g-1} x_j + a(x_j - x_i)\left(a\,x_i^{-1} + b\right)$.

From (3.47) and (3.48) we have (3.45). Hence the proof is completed.

We may consider minimizing the possible variance estimates of the H-T estimator under the model (3.35). It is clear that under the model (3.35) the second and third terms in (3.45) are fixed. Thus for minimizing the possible variance estimates, the following can be used

$$\sum_{s\in S} E_x\left(var_{SYG}(\hat{Y}_{HT})\right) = \frac{2(N-2)!}{(n-2)!(N-n)!}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{x_{ij}}{p_{ij}} \quad (3.49)$$

Thus minimizing (3.49) reduces to

$$Minimize \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{x_{ij}}{p_{ij}}, \quad (3.50)$$

where $x_{ij} = a x_i^{g-1} x_j + a(x_j - x_i)\left(a\,x_i^{-1} + b\right)$.

With respect to the objective function (3.50) to reduce the possible variance estimates, the constraints (3.51) and (3.52) can be added for the optimization problem and the following constraint can be added to maintain some desirable properties for variance estimation.

$$c p_i p_j < p_{ij} \le p_i p_j, \; j > i = 1, \cdots, N, \quad (3.51)$$

where c is a real number between 0 and 1.

$$\sum_{j\neq i}^{N} p_{ij} = (n-1)p_i, \; i = 1, \cdots, N. \quad (3.52)$$

If $a = 0$, the problem (3.50) will be minimizing

$$\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{x_i^{g-1}x_j}{p_{ij}}, \quad (3.53)$$

which depends on g in the model.

Note that (3.50) and (3.53) are the nonlinear programming (NLP) problem and they will be solved by using an optimization software program.

4. Discussion

Raj's (1956) approach only uses a simple model, but it is useful in developing other pPS sampling designs based on the superpopulation models.

We have suggested several approaches for model-based pPS sampling designs. Introducing the intercept term into the model may be desirable with respect to flexibility in sampling design. It does not result in any loss of model expectation of variance by a given sampling design, even though the model may pass through the origin. The structure of the optimization problem for the minimization of model expectation of variance estimates depends on the expressions of the variance as well as model assumptions. The construction of the sampling plan, which is a solution of the optimization problem, may be possible by using LP or NLP software.

Empirical comparison studies between model-based pPS sampling design and traditional pPS sampling design under a diversity of superpopulation assumptions may be useful.

We did not deal with pPS sampling designs involving more complicated models such as a multiple regression model or a nonlinear model. Examining those models for sampling designs may be recommended.

References

Bayless, D. L. and Rao, J. N. K. (1969). "An empirical study of stabilities of estimators and variance estimators in unequal probability sampling ($n = 3$ and $n = 4$)," *Journal of the American Statistical Association*, 65, 1645-1667.

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

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983). "An evaluation of model-dependent and probability-sampling inferences in sample surveys," *Journal of the American Statistical Association*, 78, 776-793.

Hanurav, T. V. (1967). "Optimum utilization of auxiliary information: pps sampling of two units from a stratum," *Journal of the Royal Statistical Society, Series B*, 29, 374-391.

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.

Kim, S. W., Heeringa, S. G., and Solenberger, P. W. (2003). "A probability sampling approach for variance minimization," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, 2168-2173.

Kim, S. W., Heeringa, S. G., and Solenberger, P. W. (2004). "Inclusion Probability Proportional to Size Sampling: A Nonlinear Programming Approach to Ensure a Nonnegative and Stable Variance Estimator," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, 3821-3828.

Kim, S. W., Heeringa, S. G., and Solenberger, P. W. (2005). "An empirical comparison of efficiency between optimization and non-optimization probability sampling of two units from a stratum," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, 3211-3219.

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.

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

Sen, A. R. (1953). "On the estimate of the variance in sampling with varying probabilities," Journal of the Indian Society of Agricultural Statistics, 5. 119-127.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000). Finite population sampling and inference: a prediction approach, New York: John Wiley and Sons, Inc.

Yates, F. and Grundy, P. M. (1953). "Selection without replacement from within strata with probability proportional to size," *Journal of the Royal Statistical Society, Series B*, 15, 253-261.