Sample Allocation under a Population Model and Stratified Inclusion Probability Proportionate to Size Sampling

Sun Woong Kim[1], Steven Heeringa[2], Peter Solenberger[3]
[1]Statististics, Dongguk University, Seoul, Korea, Republic of
[2]Institute for Social Research, University of Michigan, 426 Thompson, Ann Arbor, Michigan, 48104
[3]Institute for Social Research, University of Michigan

1. Introduction

In stratified sampling, a total sample of n elements is allocated to each of $h = 1, \cdots, H$ design strata and independent samples of n_h elements are selected independently within strata. One of the important roles of the survey sampler is to determine the sample allocation to strata that will result in the greatest precision for sample estimates of population characteristics.

Many studies have focused on sample allocation in stratified random sampling. The following approaches have been popular in survey sampling practice: (i) proportional sample allocation to strata, and (ii) Neyman (1934) sample allocation.

Proportional sample allocation assigns sample sizes to strata in proportion to the stratum population size. Proportional allocation can be used when information on stratum variability is lacking or stratum variances are approximately equal. Since proportional allocation results in a self-weighting sample, population estimates and their sampling variances are easily computed.

Neyman allocation can be used effectively to minimize the variance of an estimator if the survey cost per sampling unit is the same in all strata but element variances, S_h^2 , differ across strata. This allocation method requires knowledge of the values of the standard deviations, S_h , of the variable of interest y for each stratum. This information on stratum-specific variance is often not available in practice.

A sample allocation method with practical advantages over Neyman allocation is termed $x-$ optimal allocation. The $x-$ optimal allocation method uses an auxiliary variable x , highly correlated with the y and replaces the stratum standard deviations of the y with those of the x in the Neyman allocation formula. Of course, this allocation is not strictly optimal if the correlation between x and y is not perfect.

As an alternative, Dayal (1985) showed that a linear model with respect to x and y can be appropriately used in the allocation of a stratified random sample. This technique is called model-assisted allocation.

In fact in many stratified sample designs, especially those employed in business surveys, simple random sampling without replacement can be employed to select elements within strata. But it is well-known that sampling strategies with varying probabilities such as probability proportional to size (PPS) sampling without replacement are superior to simple random sampling with respect to the efficiency of estimator of population totals and related quantities. *PPS* sampling without replacement is often called inclusion probability proportional to size $(IPPS)$ sampling or πPS sampling. A number of πPS sampling schemes have been developed to select samples of size equal to or greater than two, and most of them are not easily applicable in practice. However, some techniques such as Sampford's (1967) method, are not restricted to stratum sample size of $n_h = 2$ and may be an attractive option for reducing sampling variance compared to alternative designs.

Rao (1968) discusses a sample allocation approach that minimizes the expected variance of the Horvitz and Thompson (H-T) (1952) estimator under πPS sampling and a superpopulation regression model without the intercept. Rao's method for sample allocation results in the same expected sampling variance for any πPS sampling design.

Rao's (1968) discussion raises several questions:

(1) It may be desireable to introduce an intercept term into the superpopulation regression model. Considering the intercept term, what is the proper strategy for sample allocation in πPS sampling?
(2) If we use Sampford's (1967) πPS sampling method, what sample allocation strategy would be appropriate?

In this paper, we attempt to answer these questions. We first review Rao's (1968) method. We show that the presence of the intercept in the model produces a more complicated allocation problem, but

one that can be easily solved. In addition, we employ optimization theory to show how to optimally determine stratum sample sizes for Sampford's selection method.

2. Revisiting Rao's method

Consider a finite population consisting of $h = 1, \cdots, H$ strata with N_h units in stratum h. Let s be a sample of size n_h drawn from each stratum by a given sampling design $P(\cdot)$ and let S be the set of all possible samples from each stratum. The total sample size n is :

$$n = \sum_{h=1}^{H} n_h . \tag{2.1}$$

Then the probability that the unit i in the stratum h will be in a sample, denoted π_{hi}, is given by

$$\pi_{hi} = \sum_{i \in s,\ s \in S} P(s),\ h = 1, \cdots, H,\ i = 1, \cdots, N_h , \tag{2.2}$$

which are called the first-order inclusion probabilities.

Also, the probability that both of the units i and j will be included in a sample, denoted π_{hij}, is obtained by

$$\pi_{hij} = \sum_{i,j \in s,\ s \in S} P(s),\ h = 1, \cdots, H,\ i \neq j = 1, \cdots, N_h . \tag{2.3}$$

These are termed the joint selection probabilities or the second-order inclusion probabilities.

Let y_{hi} be the value of y for the unit i in the stratum h. As an estimator of the population total $Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} y_{hi}$, consider the H-T estimator

$$\hat{Y}_{HT} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{y_{hi}}{\pi_{hi}} . \tag{2.4}$$

If $\pi_{hi} > 0$, this estimator is an unbiased estimator of Y, with variance:

$$Var\left(\hat{Y}_{HT}\right) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} \left(\pi_{hi}\pi_{hj} - \pi_{hij}\right) \left(\frac{y_{hi}}{\pi_{hi}} - \frac{y_{hj}}{\pi_{hj}}\right)^2 . \tag{2.5}$$

Rao (1968) considered the following superpopulation regression model without the intercept:

$$y_{hi} = \beta x_{hi} + \varepsilon_{hi} , \tag{2.6}$$

where x_{hi} is the value of x for the unit i in stratum h, $E_\xi\left(y_{hi} \middle| x_{hi}\right) = \beta x_{hi}$, $V_\xi\left(y_{hi} \middle| x_{hi}\right) = \sigma^2 x_{hi}^g$, $1 \le g \le 2$, and $Cov_\xi\left(y_{hi}, y_{hj} \middle| x_{hi}, x_{hj}\right) = 0$. Here E_ξ denotes the model expectation over all the finite populations that can be drawn from the superpopulation.

Then we have the following expected variance under the model (2.6):

$$E_\xi Var\left(\hat{Y}_{HT}\right) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \left(\frac{1}{\pi_{hi}} - 1\right) \sigma^2 x_{hi}^g , \tag{2.7}$$

where, $\pi_{hi} = n_h p_{hi} = n_h\, x_{hi} / X_h$, $X_h = \sum_{i=1}^{N_h} x_{hi}$.

To minimize (2.7) subject to the condition (2.1), using the Lagrange multiplier λ, consider

$$E_\xi Var\left(\hat{Y}_{HT}\right) + \lambda\left(\sum_{h=1}^{H} n_h - n\right) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \left(\frac{1}{n_h p_{hi}} - 1\right) \sigma^2 x_{hi}^g + \lambda\left(\sum_{h=1}^{H} n_h - n\right). \tag{2.8}$$

Equating (2.8) to zero and differentiating with respect to n_h, we have

$$n_h = \frac{1}{\sqrt{\lambda}} \sqrt{\sum_{i=1}^{N_h} \frac{\sigma^2 x_{hi}^g}{p_{hi}}} . \tag{2.9}$$

Substituting n_h in (2.1), we have

$$\frac{1}{\sqrt{\lambda}} = n \Big/ \sum_{h=1}^{H} \sqrt{\sum_{i=1}^{N_h} \frac{\sigma^2 x_{hi}^g}{p_{hi}}} . \tag{2.10}$$

Replacing $1/\sqrt{\lambda}$ in (2.9) with (2.10), we have the following sample allocation in each stratum:

$$n_h = n \frac{\sqrt{X_h \sum_{i=1}^{N_h} x_{hi}^{g-1}}}{\sum_{h=1}^{H} \sqrt{X_h \sum_{i=1}^{N_h} x_{hi}^{g-1}}} . \tag{2.11}$$

Note that if $g = 2$, the allocation under the superpopulation model and πPS sampling reduces to:

$$n_h = n\frac{X_h}{\sum_{h=1}^{H} X_h}, \qquad (2.12)$$

which is a proportional sample allocation to the stratum.

Also, Rao showed that in terms of expected variance, unstratified πPS sampling under the same superpopulation model is inferior to stratified πPS sampling with the allocation (2.11).

Looking at the expected variance in (2.7) and the sample allocation in (2.11), it does not involve the joint probabilities π_{hij} in each stratum. It indicates that under the model without the intercept (2.6) the specific properties of a given πPS sampling scheme (properties that determine the π_{hij}) are not reflected in the sample allocation, resulting in the same sample allocation for any πPS sampling. Hence the following issues, as mentioned in the Introduction, are of interest.

(1) The superpopulation regression model which we may wish to employ in many surveys may be :

$$y_{hi} = \alpha + \beta x_{hi} + \varepsilon_{hi}, \qquad (3.1)$$

which is a general form and (2.6) is a special form of (3.1) when $\alpha = 0$.
Considering the intercept term α, we need to reexamine the most appropriate sample allocation strategy for πPS sampling.

(2) Although it will be shown in the following section that using (3.1) gives a sample allocation involving the joint probabilities π_{hij}, and these differ according to the chosen πPS sampling, if we focus on Sampford's (1967) method for πPS sampling, what sample allocation strategy would be appropriate?

Section 3 will address these issues of sample allocation.

3. Alternative Sample Allocations

We assume two different models involving an intercept term:

Model I:

$$y_{hi} = \alpha + \beta x_{hi} + \varepsilon_{hi}, \quad h = 1,\cdots,H,\ i = 1,\cdots,N_h \qquad (3.1)$$

where ε_{hi} is numerically negligible, that is, x explains y well.

Model II:

$$y_{hi} = \alpha + \beta x_{hi} + \varepsilon_{hi}, \quad h = 1,\cdots,H,\ i = 1,\cdots,N_h \qquad (3.2)$$

where $E_\xi\left(y_{hi}\middle|x_{hi}\right) = \alpha + \beta x_{hi}$, $V_\xi\left(y_{hi}\middle|x_{hi}\right) = \sigma^2 x_{hi}^g$, and $Cov_\xi\left(y_{hi}, y_{hj}\middle|x_{hi}, x_{hj}\right) = 0$.

Instead of (2.5) we consider the following form of the variance of the H-T estimator

$$Var\left(\hat{Y}_{HT}\right) = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\frac{y_{hi}^2(1-\pi_{hi})}{\pi_{hi}} + 2\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\frac{\pi_{hij}}{\pi_{hi}\pi_{hj}}y_{hi}y_{hj} - 2\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j>i}^{N_h} y_{hi}y_{hj} \qquad (3.3)$$

Theorem 3.1. Under the *Model I*, the minimization of the expected variance of (2.4) under πPS sampling is equivalent to minimizing

$$\sum_{h=1}^{H}\frac{A_h}{n_h^2} + \sum_{h=1}^{H}\frac{B_h}{n_h}, \qquad (3.4)$$

where,

$$A_h = 2X_h^2\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\frac{\alpha^2 + \alpha\beta(x_{hi}+x_{hj})}{x_{hi}x_{hj}}\pi_{hij} \qquad (3.5)$$

and

$$B_h = X_h\left(\sum_{i=1}^{N_h}\frac{(\alpha+\beta x_{hi})^2}{x_{hi}} - \beta^2 X_h\right). \qquad (3.6)$$

Proof. For the expected variance of (2.4) under *Model I* the third term in (3.3) is a fixed value that does not involve n_h, and the other terms are given by:

$$\left[\sum_{h=1}^{H}\frac{X_h}{n_h}\sum_{i=1}^{N_h}\frac{(\alpha+\beta x_{hi})^2}{x_{hi}} - \sum_{h=1}^{H}\sum_{i=1}^{N_h}(\alpha+\beta x_{hi})^2\right] + \left[2\sum_{h=1}^{H}\frac{X_h^2}{n_h^2}\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\frac{\alpha^2+\alpha\beta(x_{hi}+x_{hj})}{x_{hi}x_{hj}}\pi_{hij} + \beta^2\sum_{h=1}^{H}X_h^2 - \beta^2\sum_{h=1}^{H}\frac{X_h^2}{n_h}\right], \qquad (3.7)$$

by noting $\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\pi_{hij} = n_h(n_h-1)/2$ in the second term in (3.3).

Since $\sum_{h=1}^{H}\sum_{i=1}^{N_h}(\alpha+\beta x_{hi})^2$ and $\beta^2\sum_{h=1}^{H}X_h^2$ are also fixed, the quantity to be minimized in (3.7) is:

$$\left[\sum_{h=1}^{H}\frac{X_h}{n_h}\sum_{i=1}^{N_h}\frac{(\alpha+\beta x_{hi})^2}{x_{hi}}\right] + \left[2\sum_{h=1}^{H}\frac{X_h^2}{n_h^2}\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\frac{\alpha^2+\alpha\beta(x_{hi}+x_{hj})}{x_{hi}x_{hj}}\pi_{hij}-\beta^2\sum_{h=1}^{H}\frac{X_h^2}{n_h}\right] \tag{3.8}$$

The proof follows from substitution of

$$A_h = 2X_h^2\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\frac{\alpha^2+\alpha\beta(x_{hi}+x_{hj})}{x_{hi}x_{hj}}\pi_{hij}$$

and

$$B_h = X_h\left(\sum_{i=1}^{N_h}\frac{(\alpha+\beta x_{hi})^2}{x_{hi}}-\beta^2 X_h\right)$$

in (3.8).

Remark 3.1. Minimization of (3.4) is a simple problem in terms of n_h because the A_h and the B_h are known values.

Consider Sampford's (1967) πPS sampling method for selecting n_h elements in each stratum. Although we can use (3.4) to decide the stratum sample size, we still don't know the values of the joint probabilities. The following approximate expression for π_{hij} correct to $O(N^{-4})$ may be useful:

$$\pi_{hij} \approx n_h(n_h-1)p_{hi}p_{hj}\left[1+\left\{\left(p_{hi}+p_{hj}\right)-\sum_{k=1}^{N_h}p_{hk}^2\right\} +\left\{2\left(p_{hi}^2+p_{hj}^2\right)-2\sum_{k=1}^{N_h}p_{hk}^3-(n_h-2)p_{hi}p_{hj} +(n_h-3)\left(p_{hi}+p_{hj}\right)\sum_{k=1}^{N_h}p_{hk}^2-(n_h-3)\left(\sum_{k=1}^{N_h}p_{hk}^2\right)^2\right\}\right], \tag{3.9}$$

which was derived by Asok and Sukhatme (1976).

From (3.4) and (3.9) we obtain the following theorem.

Theorem 3.2. Under the *Model I*, the sample allocation problem to minimize the expected variance of (2.4) under Sampford's method when using the joint probabilities, correct to $O(N^{-4})$, given in (3.9) is equivalent to minimizing

$$\sum_{h=1}^{H}C_h n_h+\sum_{h=1}^{H}\frac{D_h}{n_h}, \tag{3.10}$$

where

$$C_h = 2\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left\{\alpha^2+\alpha\beta(x_{hi}+x_{hj})\right\}\pi_{hij1}, \tag{3.11}$$

$$\pi_{hij1} = \left(p_{hi}+p_{hj}\right)\sum_{k=1}^{N_h}p_{hk}^2-p_{hi}p_{hj}-\left(\sum_{k=1}^{N_h}p_{hk}^2\right)^2, \tag{3.12}$$

$$D_h = B_h-2\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left\{\alpha^2+\alpha\beta(x_{hi}+x_{hj})\right\}\pi_{hij2}, \tag{3.13}$$

and

$$\pi_{hij2} = 1+\left\{\left(p_{hi}+p_{hj}\right)-\sum_{k=1}^{N_h}p_{hk}^2\right\} +2\left(p_{hi}^2+p_{hj}^2\right)-2\sum_{k=1}^{N_h}p_{hk}^3 +2p_{hi}p_{hj}-3\left(p_{hi}+p_{hj}\right)\sum_{k=1}^{N_h}p_{hk}^2+3\left(\sum_{k=1}^{N_h}p_{hk}^2\right)^2. \tag{3.14}$$

Proof. Substituting π_{hij} from (3.9) in (3.5) for the first term of (3.4), we get:

$$\sum_{h=1}^{H}\frac{A_h}{n_h^2}=\sum_{h=1}^{H}2\left(1-\frac{1}{n_h}\right)\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left\{\alpha^2+\alpha\beta(x_{hi}+x_{hj})\right\}\pi_{hij0}, \tag{3.15}$$

where:

$$\pi_{hij0}=\left[1+\left\{\left(p_{hi}+p_{hj}\right)-\sum_{k=1}^{N_h}p_{hk}^2\right\} +\left\{2\left(p_{hi}^2+p_{hj}^2\right)-2\sum_{k=1}^{N_h}p_{hk}^3-(n_h-2)p_{hi}p_{hj} +(n_h-3)\left(p_{hi}+p_{hj}\right)\sum_{k=1}^{N_h}p_{hk}^2-(n_h-3)\left(\sum_{k=1}^{N_h}p_{hk}^2\right)^2\right\}\right]. \tag{3.16}$$

Expressing (3.16) in terms of n_h, we have:

$$\pi_{hij0} = n_h\pi_{hij1}+\pi_{hij2}. \tag{3.17}$$

Substituting (3.17) in (3.15), we obtain

$$\sum_{h=1}^{H}\frac{A_h}{n_h^2}=2\sum_{h=1}^{H}n_h\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left\{\alpha^2+\alpha\beta(x_{hi}+x_{hj})\right\}\pi_{hij1} +2\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left\{\alpha^2+\alpha\beta(x_{hi}+x_{hj})\right\}\pi_{hij2} -2\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left\{\alpha^2+\alpha\beta(x_{hi}+x_{hj})\right\}\pi_{hij1} -2\sum_{h=1}^{H}\frac{1}{n_h}\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left\{\alpha^2+\alpha\beta(x_{hi}+x_{hj})\right\}\pi_{hij2}.$$

(3.18)

Since the second and third terms in (3.18) are the known values, the minimization of (3.18) reduces to minimizing:

$$2\sum_{h=1}^{H} n_h \sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left\{\alpha^2+\alpha\beta(x_{hi}+x_{hj})\right\}\pi_{hij1} - 2\sum_{h=1}^{H}\frac{1}{n_h}\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left\{\alpha^2+\alpha\beta(x_{hi}+x_{hj})\right\}\pi_{hij2}\,. \tag{3.19}$$

Adding $\sum_{h=1}^{H}\frac{B_h}{n_h}$ in (3.19), we have the following equivalent minimization problem to the minimization of (3.4):

$$2\sum_{h=1}^{H} n_h \sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left\{\alpha^2+\alpha\beta(x_{hi}+x_{hj})\right\}\pi_{hij1} + \sum_{h=1}^{H}\frac{1}{n_h}\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left[B_h - 2\left\{\alpha^2+\alpha\beta(x_{hi}+x_{hj})\right\}\pi_{hij2}\right]. \tag{3.20}$$

This completes the proof.

Remark 3.2. (3.10) is a simple allocation problem in terms of n_h because the C_h and the D_h are the known values.

Remark 3.3. We can define the following optimization problem with respect to n_h:

$$\text{Minimize } \sum_{h=1}^{H} C_h n_h + \sum_{h=1}^{H}\frac{D_h}{n_h} \tag{3.21}$$

subject to,

$$n_h \le N_h,\ h=1,\cdots,H, \tag{3.22}$$

$$n_h \ge 2,\ h=1,\cdots,H, \tag{3.23}$$

and

$$\sum_{h=1}^{H} n_h = n. \tag{3.24}$$

This problem may be easily handled by convex mathematical programming algorithms and the solution provides an efficient sample allocation strategy when using Sampford's method under the model assumption of (3.1).

We obtain the following theorem regarding the minimization of the variance of the H-T estimator (2.4) in πPS sampling under the assumption of the model (3.2).

Theorem 3.3. Under *Model II*, minimizing the expected variance of (2.4) under πPS sampling amounts to minimizing:

$$\sum_{h=1}^{H}\frac{A_h^*}{n_h^2}+\sum_{h=1}^{H}\frac{B_h^*}{n_h}\,, \tag{3.25}$$

where,

$$A_h^* = 2\alpha X_h^2\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left(x_{hj}^{-1}-x_{hi}^{-1}\right)\left(\alpha x_{hi}^{-1}+\beta\right)\pi_{hij} \tag{3.26}$$

and

$$B_h^* = \sigma^2 X_h\sum_{i=1}^{N_h} x_{hi}^{g-1}\,. \tag{3.27}$$

Proof. Consider a different form of (2.5) using $\pi_{hi}=n_h p_{hi}$:

$$Var\left(\hat{Y}_{HT}\right)=\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left(p_{hi}p_{hj}-\frac{\pi_{hij}}{n_h^2}\right)\left(\frac{y_{hi}}{p_{hi}}-\frac{y_{hj}}{p_{hj}}\right)^2. \tag{3.28}$$

By using

$$E_\xi y_{hi}^2=\sigma^2 x_{hi}^g+\alpha^2+\beta^2 x_{hi}^2+2\alpha\beta x_{hi} \tag{3.29}$$

and

$$E_\xi(y_{hi}y_{hj})=\alpha^2+\alpha\beta(x_{hi}+x_{hj})+\beta^2 x_{hi}x_{hj}\,, \tag{3.30}$$

we obtain

$$E_\xi\left(\frac{y_{hi}}{p_{hi}}-\frac{y_{hj}}{p_{hj}}\right)^2 = 2\sigma^2 X_h^g p_{hi}^{g-2} + 2\alpha X_h^2\frac{x_{hj}-x_{hi}}{x_{hi}x_{hj}}\left(\alpha x_{hi}^{-1}+\beta\right). \tag{3.31}$$

Then we get:

$$E_\xi Var\left(\hat{Y}_{HT}\right) = 2\sigma^2\sum_{h=1}^{H}X_h^g\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}p_{hi}^{g-2}\left(p_{hi}p_{hj}-\frac{\pi_{hij}}{n_h^2}\right) + 2\alpha\sum_{h=1}^{H}X_h^2\left(\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left(p_{hi}p_{hj}-\frac{\pi_{hij}}{n_h^2}\right)\frac{x_{hj}-x_{hi}}{x_{hi}x_{hj}}\left(\alpha x_{hi}^{-1}+\beta\right)\right)$$

$$= EV + 2\alpha\sum_{h=1}^{H}\left(\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left(x_{hj}-x_{hi}\right)\left(\alpha x_{hi}^{-1}+\beta\right)\right) + 2\alpha\sum_{h=1}^{H}\frac{X_h^2}{n_h^2}\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left(x_{hj}^{-1}-x_{hi}^{-1}\right)\left(\alpha x_{hi}^{-1}+\beta\right)\pi_{hij} \tag{3.32}$$

with $EV=\sigma^2\sum_{h=1}^{H}\sum_{i=1}^{N_h}\frac{X_h^g}{n_h}(1-n_h p_{hi})p_{hi}^{g-1}$

$$= \sigma^2 \sum_{h=1}^{H} \sum_{i=1}^{N_h} \frac{X_h^g}{n_h} (1 - n_h p_{hi}) \left(\frac{x_{hi}}{X_h} \right)^g \left(\frac{1}{p_{hi}} \right)$$
$$= \sum_{h=1}^{H} \sum_{i=1}^{N_h} \left(\frac{1}{n_h p_{hi}} - 1 \right) \sigma^2 x_{hi}^g$$
$$= \sigma^2 \sum_{h=1}^{H} \sum_{i=1}^{N_h} X_h \frac{x_{hi}^{g-1}}{n_h} - \sigma^2 \sum_{h=1}^{H} \sum_{i=1}^{N_h} x_{hi}^g \,. \qquad (3.33)$$

Since the second term in (3.32) and the second term in (3.33) are fixed in terms of n_h, the minimization of the model expectation of (3.28) reduces to minimizing:

$$2\alpha \sum_{h=1}^{H} \frac{X_h^2}{n_h^2} \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} \left(x_{hj}^{-1} - x_{hi}^{-1} \right) \left(\alpha x_{hi}^{-1} + \beta \right) \pi_{hij} + \sigma^2 \sum_{h=1}^{H} \frac{X_h}{n_h} \sum_{i=1}^{N_h} x_{hi}^{g-1} \,. \qquad (3.34)$$

Since (3.34) equals (3.25), the proof is completed.

Remark 3.4. Minimizing (3.25) is a simple problem in terms of n_h because the A_h^* and the B_h^* are the known values.

Remark 3.4. (3.33) is a different form of (2.7). The model expectation of (3.28) involves (2.7) plus the other terms due to *Model II* with the intercept term, as shown in (3.32).

Theorem 3.4. Under the *Model II*, the sample allocation problem under Sampford's sampling scheme to minimize the expected variance of (2.4), when using the joint probabilities correct to $O(N^{-4})$ given in (3.9), is equivalent to minimizing:

$$\sum_{h=1}^{H} C_h^* n_h + \sum_{h=1}^{H} \frac{D_h^*}{n_h}, \qquad (3.35)$$

where

$$C_h^* = 2\alpha \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} \left\{ \left(x_{hi} - x_{hj} \right) \left(\alpha x_{hi}^{-1} + \beta \right) \pi_{hij1} \right\} \qquad (3.36)$$

and

$$D_h^* = B_h^* - 2\alpha \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} \left\{ \left(x_{hi} - x_{hj} \right) \left(\alpha x_{hi}^{-1} + \beta \right) \pi_{hij2} \right\} \qquad (3.37)$$

with

$$\pi_{hij1} = \left(p_{hi} + p_{hj} \right) \sum_{k=1}^{N_h} p_{hk}^2 - p_{hi} p_{hj} - \left(\sum_{k=1}^{N_h} p_{hk}^2 \right)^2,$$

$$\pi_{hij2} = 1 + \left\{ \left(p_{hi} + p_{hj} \right) - \sum_{k=1}^{N_h} p_{hk}^2 \right\} + 2\left(p_{hi}^2 + p_{hj}^2 \right) - 2 \sum_{k=1}^{N_h} p_{hk}^3 + 2 p_{hi} p_{hj} - 3\left(p_{hi} + p_{hj} \right) \sum_{k=1}^{N_h} p_{hk}^2 + 3 \left(\sum_{k=1}^{N_h} p_{hk}^2 \right)^2,$$

and

$$B_h^* = \sigma^2 X_h \sum_{i=1}^{N_h} x_{hi}^{g-1} \,.$$

Proof. Substituting (3.9) in the first term of (3.25) and using (3.17) with (3.12) and (3.14), we obtain

$$\sum_{h=1}^{H} \frac{A_h^*}{n_h^2} = 2\alpha \sum_{h=1}^{H} \frac{X_h^2}{n_h^2} n_h (n_h - 1) \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} \left\{ \left(x_{hj}^{-1} - x_{hi}^{-1} \right) \left(\alpha x_{hi}^{-1} + \beta \right) p_{hi} p_{hj} \pi_{hij0} \right\}$$
$$= 2\alpha \sum_{h=1}^{H} \left(1 - \frac{1}{n_h} \right) \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} \left\{ \left(x_{hi} - x_{hj} \right) \left(\alpha x_{hi}^{-1} + \beta \right) \left(n_h \pi_{hij1} + \pi_{hij2} \right) \right\}$$
$$= 2\alpha \sum_{h=1}^{H} n_h \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} \left\{ \left(x_{hi} - x_{hj} \right) \left(\alpha x_{hi}^{-1} + \beta \right) \pi_{hij1} \right\}$$
$$+ 2\alpha \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} \left\{ \left(x_{hi} - x_{hj} \right) \left(\alpha x_{hi}^{-1} + \beta \right) \pi_{hij2} \right\}$$
$$- 2\alpha \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} \left\{ \left(x_{hi} - x_{hj} \right) \left(\alpha x_{hi}^{-1} + \beta \right) \pi_{hij1} \right\}$$
$$- 2\alpha \sum_{h=1}^{H} \frac{1}{n_h} \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} \left\{ \left(x_{hi} - x_{hj} \right) \left(\alpha x_{hi}^{-1} + \beta \right) \pi_{hij2} \right\}. \qquad (3.38)$$

Since the second and third terms in (3.38) are equal, the minimization of (3.38) reduces to minimizing the other terms, that is,

$$2\alpha \sum_{h=1}^{H} n_h \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} \left\{ \left(x_{hi} - x_{hj} \right) \left(\alpha x_{hi}^{-1} + \beta \right) \pi_{hij1} \right\} - 2\alpha \sum_{h=1}^{H} \frac{1}{n_h} \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} \left\{ \left(x_{hi} - x_{hj} \right) \left(\alpha x_{hi}^{-1} + \beta \right) \pi_{hij2} \right\}. \qquad (3.39)$$

Thus, the minimization of (3.25) with (3.26) and (3.27) amounts to the one of

$$2\alpha \sum_{h=1}^{H} n_h \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} \left\{ \left(x_{hi} - x_{hj} \right) \left(\alpha x_{hi}^{-1} + \beta \right) \pi_{hij1} \right\}$$

$$-2\alpha\sum_{h=1}^{H}\frac{1}{n_h}\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left\{\left(x_{hi}-x_{hj}\right)\left(\alpha x_{hi}^{-1}+\beta\right)\pi_{hij2}\right\} +\sum_{h=1}^{H}\frac{B_h^*}{n_h}. \tag{3.39}$$

Accordingly, the following reduced form from (3.39) can be obtained.

$$2\alpha\sum_{h=1}^{H}n_h\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left\{\left(x_{hi}-x_{hj}\right)\left(\alpha x_{hi}^{-1}+\beta\right)\pi_{hij1}\right\} +\sum_{h=1}^{H}\frac{1}{n_h}\left[B_h^*-2\alpha\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left\{\left(x_{hi}-x_{hj}\right)\left(\alpha x_{hi}^{-1}+\beta\right)\pi_{hij2}\right\}\right] \tag{3.40}$$

Hence, we have proved the theorem.

Remark 3.5. (3.35) is a simple allocation problem in terms of n_h since the C_h^* and the D_h^* are the known values.

Remark3.6. In order to find a solution for n_h, we may define the following optimization problem:

$$\text{Minimize } \sum_{h=1}^{H}C_h^*n_h+\sum_{h=1}^{H}\frac{D_h^*}{n_h} \tag{3.41}$$

subject to

$$n_h \le N_h,\ h=1,\cdots,H \tag{3.42}$$

and

$$n_h \ge 2,\ h=1,\cdots,H\ . \tag{3.43}$$

It is noted that the condition (2.1) may not be used as the constraint, different from *Remark 3.3*.

Corollary3.1. Under M*odel II,* without the intercept the minimization of the expected variance of (2.4) under πPS sampling is equivalent to minimizing:

$$\sum_{h=1}^{H}\sum_{i=1}^{N_h}\frac{1}{n_h}X^* \tag{3.44}$$

where

$$X^* = X_h x_{hi}^{g-1} \tag{3.45}$$

Proof. When $\alpha=0$, (3.32) in *Theorem 3.3* reduces to simply EV, which is expressed as (3.33). σ^2 and the second term in (3.33) are fixed values with respect to n_h, and the minimization of (3.33) reduces to the one of (3.44). Hence, we have the corollary.

Remark 3.7. (3.44) is quite a simple allocation problem in terms of n_h not depending on the joint probabilities π_{hij}.

4. Discussion

We have addressed the topic of efficient sample allocation in stratified samples using more general superpopulation regression models than those investigated by Rao (1968). Under more general models that include an intercept term, we have developed several theorems to be useful for deciding sample allocation in πPS sampling designs. Also, through the theorems we have showed how to apply this sample allocation theory for Sampford's (1967) sampling method, one of the more common πPS sampling designs used in survey practice.

We determined that the sample allocation approaches to mimizing the model expectation of the variance of the H-T estimator may depend on the expressions of the variance.

Based on the theorems developed in this paper, the optimization problem with respect to the stratum sample sizes can be solved by using software involving convex mathematical programming algorithms. This is a straightforward approach for sample allocation when using more efficient πPS sampling methods.

In addition to Sampford' sampling, the approach can be applied to a variety of πPS sampling without replacement designs.

In future work it will be important to extend the theory and methods described here to allocation problems under more complicated superpopulation models and situations where the superpopulatin model can vary across strata

References

Dayal, S. (1985). "Allocation of sample using values of auxiliary characteristic," *Journal of the Statistical Planning and Inference*, 11, 321-328.

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.

Neyman, J. (1934). "On two different aspects of the representative method: the method of stratified sampling and the method of purposive selection," *Journal of the Royal Statistical Society*, 97, 558-606.

Rao, T. J. (1968). "On the allocation of sample size in stratified sampling," *Annals of the Institute of Statistical Mathematics*, 20, 159-166.

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