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

Sun-Woong Kim

Steven G. Heeringa

Peter W. Solenberger

Dongguk University

&

Survey Research Center, University of Michigan

Overview

- **Allocation of Stratified Random Samples**
- **Alternative Allocations under Stratified Random Sampling**
- **Sampling Strategies with Varying Probabilities**
- **Rao's (1968) Allocation in Stratified $IPPS$ (pPS) Sampling**
- **Raised Questions**
- **Different Models Involving Intercept Term**
- **Sample Allocation for Minimizing Variance Expectation under Model I**
- **Sample Allocation for Minimizing Variance Expectation under Model II**
- **Concluding Remarks**

Allocation of Stratified Random Samples

Many studies have been focused on allocations in stratified random sampling without replacement.

The following have been popular:

- Proportional Allocation:

  Used when stratum-specific information is lacking on data variability

- Neyman (1934) Allocation:

  Attempted to minimize the variance of an estimator if the cost per unit is the same in all strata

The Neyman allocation

- Requires the values of the standard deviations of the study variable of interest y

- Often infeasible in practice because the values are normally unknown.

Alternative Allocations under Simple Random Sampling

Before sample selection the sample designer often knows the variability of an auxiliary variable x thought to be correlated with the characteristic under study. Such an auxiliary variable is often referred to a measure of size.

✓ Dayal (1985)

A linear model with respect to the values of auxiliary characteristic linearly related to the study variable can be used in the allocation of the stratified random sample.

Sampling Strategies with Varying Probabilities

- PPS sampling without replacement is generally more efficient than PPS sampling with replacement or stratified random sampling.

- A number of PPS sampling procedures without replacement have been developed to select samples of size greater than two, and most of these procedures are not easily applicable in practice.

- Due to low variance potential, $IPPS$ (πPS) sampling is an attractive option to survey samplers.

Rao's (1968) Allocation in Stratified $IPPS$ ($p\,PS$) Sampling

- Consider the following population model without the intercept in sample allocation.

$$y_i = b\,x_i + e_i\,,$$

where $E_x(y_i \,|\, x_i) = b\,x_i$, $V_x(y_i \,|\, x_i) = s^{2} x_i^{g}$, and $Cov_x(y_i, y_j \,|\, x_i, x_j) = 0$

Here E_x denotes the model expectation over all the finite populations that can be drawn from the superpopulation.

- Used the following expected variance of the Horvitz-Thompson (1952) estimator under the model:

$$E_x\left(Var(\hat{Y}_{HT})\right) = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\left(\frac{1}{p_{hi}} - 1\right)s^2 x_{hi}^{g},$$

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

Note that the expected variance indicates that any *IPPS* sampling design produces the same expected variance.

- Showed that allocation of the sample size to the strata which minimizes the above expected variance can be given as follows:

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

where $X_h = \sum_{i=1}^{N_h} x_{hi}$

Raised Questions

Q1: It is customary to introduce an intercept term into the model. Considering the intercept term, what is a proper strategy for sample allocation in *IPPS* sampling designs?

Q2: If we use Sampford's (1967) method, which is one of the popular *IPPS* sampling designs, what sample allocation strategy would be appropriate?

Different Models Involving Intercept Term

Model I:

$$y_i = a + b x_i + e_i,$$

where the terms in e_i are numerically negligible, that is, x explains y well.

Model II:

$$y_i = a + b x_i + e_i,$$

where $E_x(y_i | x_i) = a + b x_i$, $V_x(y_i | x_i) = s^2 x_i^g$,
and $Cov_x(y_i, y_j | x_i, x_j) = 0$

Sample Allocation for Minimizing Variance Expectation under Model I

- Using a different form of the variance of H-T estimator

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

Since the first and third terms are fixed under the model, the minimization of the model expectation of $Var(\hat{Y}_{HT})$ in $IPPS$ sampling reduces to minimization of the following :

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

where $A_h = 2X_h^2 \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} \frac{a^2 + ab(x_{hi} + x_{hj})}{x_{hi}x_{hj}} p_{hij}$

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

Note. The A_h and the B_h are known values.

- Sample allocation strategy under Sampford (1967)'s *IPPS* sampling

Asok and Sukhatme (1976) developed the approximate expression for p_{hij} in Sampford's (1967) method.

Substituting the expression for p_{hij} in $\sum_{h=1}^{H}\frac{A_h}{n_h^2}+\sum_{h=1}^{H}\frac{B_h}{n_h}$, we get

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

where $C_h = 2\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left[\left(a^2+ab\,(x_{hi}+x_{hj})\right)\left((p_{hi}+p_{hj})\sum_{i=1}^{N_h}p_{hi}^2 - p_{hi}p_{hj} - (\sum_{i=1}^{N_h}p_{hi}^2)^2\right)\right]$

$$p_{hi} = \frac{x_{hi}}{X_h}$$

$$D_h = X_h\left(\sum_{i=1}^{N_h}\frac{(a+b\,x_{hi})^2}{x_{hi}} - b^2 X_h\right)$$
$$-2\sum_{i=1}^{N_h}\sum_{j>i}^{N_h}\left[\left\{a^2+ab\,(x_{hi}+x_{hj})\right\}\left\{(2\,p_{hi}p_{hj}-3(p_{hi}+p_{hj})\sum_{i=1}^{N_h}p_{hi}^2+3(\sum_{i=1}^{N_h}p_{hi}^2)^2\right.\right.$$
$$\left.\left.+1+(p_{hi}+p_{hj})-\sum_{i=1}^{N_h}p_{hi}^2+2(p_{hi}^2+p_{hj}^2)-2\sum_{i=1}^{N_h}p_{hi}^3\right\}\right]$$

Note. The C_h and the D_h are known values.

If the following constraints are added, the sample allocation problem to minimize $\sum_{h=1}^{H} C_h n_h + \sum_{h=1}^{H} \frac{D_h}{n_h}$ can be solved by mathematical programming.

$$\sum_{h=1}^{H} n_h = n,$$
$$n_h \le N_h, \quad h = 1, 2, \cdots, H$$
$$n_h \ge 2, \quad h = 1, 2, \cdots, H$$

Sample Allocation for Minimizing Variance Expectation under Model II

- Using the variance of H-T estimator

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

The minimization of the model expectation of $Var(\hat{Y}_{HT})$ in *IPPS* sampling is equivalent to minimization of the following:

.

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

where $A_h^* = a X_h^2 \sum_{i}^{N_h}\sum_{j>i}^{N_h}\left(x_{hj}^{-1} - x_{hi}^{-1}\right)\left(a x_{hi}^{-1} + b\right)p_{hij}$

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

Note. The A_h^* and the B_h^* are known values.

- Sample allocation under Sampford's *IPPS* sampling

Using Asok and Sukhatme's (1976) formula for p_{hij}, we have

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

where

$$C_h^* = a \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} \left[\left((x_{hi} - x_{hj})(a\, x_{hi}^{-1} + b) \right) \left((p_{hi} + p_{hj}) \sum_{i=1}^{N_h} p_{hi}^2 - p_{hi} p_{hj} - (\sum_{i=1}^{N_h} p_{hi}^2)^2 \right) \right]$$

$$\begin{aligned} D_h^* = B_h^* - a \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} \Bigg[\left((x_{hi} - x_{hj})(a\, x_{hi}^{-1} + b) \right) \Bigg\{ & (2 p_{hi} p_{hj} - 3(p_{hi} + p_{hj}) \sum_{i=1}^{N_h} p_{hi}^2 \\ & + 3(\sum_{i=1}^{N_h} p_{hi}^2)^2 + 1 + (p_{hi} + p_{hj}) - \sum_{i=1}^{N_h} p_{hi}^2 + 2(p_{hi}^2 + p_{hj}^2) - 2 \sum_{i=1}^{N_h} p_{hi}^3 \Bigg\} \Bigg] \end{aligned}$$

Note. The C_h^* and the D_h^* are known values.

The sample allocation problem under Sampford's *IPPS* sampling below can be easily solved by nonlinear programming algorithms.

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

$$\text{subject to} \quad \sum_{h=1}^{H} n_h = n,$$

$$n_h \le N_h, \quad h = 1, 2, \cdots, H$$

$$n_h \ge 2, \quad h = 1, 2, \cdots, H$$

Concluding Remarks

- We have used more general population models relative to the model Rao (1968) used.
- We have proposed a quite straightforward approach for sample allocation in stratified *IPPS* sampling.
- Although it seems that the minimization problems are complicated, they can be easily solved by using software involving nonlinear programming.
- In addition to Sampford's *IPPS* sampling, the approach described here can be applied to a variety of sampling without replacement designs.

- The structure of minimization problem regarding the model expectation of the variance depends on the expression of the variance.

- Allocation under more complicated models and allocation under the situations where each stratum has a different model should be studied.