

A NOTE ON OPTIMUM INCLUSION PROBABILITIES IN
'WOR-SAMPLING SCHEME' BASED ON SUPER POPULATION
MODEL AND MULTIVARIATE INFORMATION

Pulakesh Maiti and T.P. Tripathi
Indian Statistical Institute, Kolkata, India.
Email: pulakesh@isical.ac.in

ABSTRACT

This paper deals with the problem of obtaining a set of optimum inclusion probabilities $\{\pi_i : i = 1, 2, \dots, N\}$, optimum in the sense of having smallest average (δ -model based) mean square (design-based) of the Horvitz-Thompson estimator \hat{Y}_{H-T} . Since optimum π_i 's are dependent on model parameters, a "near optimum" solution based on the estimates of the model parameters have been proposed.

KEYWORDS AND PHRASES

Inclusion Probabilities; δ -based optimum; Horvitz-Thompson Estimator.

1. INTRODUCTION

Let $U = \{1, 2, \dots, i, \dots, N\}$ of N (Given) well defined, identifiable and observable objects under consideration, denote a finite population, i being a Sampling unit. Let a variate (character) y be a real valued function defined on i and let $y_i = y(i)$, ($i = 1, 2, \dots, N$) be the value of the character y associated with i^{th} unit of the population. In several situations, one may have auxiliary information $\{x_{ij}; j = 1, 2, \dots, q; i = 1, 2, \dots, N\}$ on auxiliary variates $x = (x_1, x_2, \dots, x_q)$ prior to a survey and which may be used in mixed ways at any of the stages namely, (i) at pre-selection stage, (ii) at selection stage and (iii) at estimation stage [Tripathi, (1970)]. The use of auxiliary information in selecting the unit at any draw was first considered by Hansen and Hurwitz (1943) for estimating the population total and since then applications of pps-sampling in which p_i 's are taken proportional to a size measure and being used more and more on various social, economic, demographic and agricultural characters with more and more development of agriculture, trade, commerce and industry [Brewer and Hanif (1983)].

The HT estimator is the only UMVUE for Y in the class of linear unbiased estimators, Horvitz and Thompson (1952) and Koop (1952). The estimator due to Horvitz and Thompson (1952),

$$\hat{Y}_{H-T} = \sum_{i \in s} y_i / \pi_i, \quad (1.1)$$

is the only UMVU estimator for Y in the unbiased subclass of T_2 .

Here in this paper, we have considered the problem of optimum choice of inclusion probabilities $\{\pi_i, i=1, 2, \dots, N\}$, optimality in the sense of having smallest average (δ -model based) mean square (design based) of the estimator \hat{Y}_{H-T} i.e., the criterion of preference adopted here is to minimize,

$$MP_{n,\delta}(\hat{Y}_{H-T}) = \varepsilon_\delta \left[E_{P_n} (\hat{Y}_{H-T} - Y)^2 \right],$$

where, P_n refers to a sampling design of fixed size n and δ is the model representing a super population. This mixed approach was initiated by Cochran (1946), Des Raj (1958).

Similar kind of problem was addressed by Tripathi and Chaubey (1998) in choosing optimum probabilities of Selection in PPS-Sampling.

2. MODEL BASED OPTIMUM π_i 's

It is well known that the sampling strategy

$$\tau_{\pi} = \left\{ PPSWOR, \hat{Y}_{H-T}; \hat{Y}_{H-T} = \sum_{i \in s} y_i / \pi_i, \pi = (\pi_1, \pi_2, \dots, \pi_N) \right\}, \quad (2.1)$$

is unbiased for finite population total $Y = \sum_{i=1}^N y_i$, with the variance

$$V(\tau_{\pi}) = \sum_{i=1}^N y_i^2 / \pi_i + \sum_{i \neq j=1}^N \frac{y_i y_j \pi_i \pi_j}{\pi_{ij}} - Y^2. \quad (2.2)$$

Let the population values $\{y_1, y_2, \dots, y_i, \dots, y_N\}$ be regarded as a random sample from a super population characterized by a model δ and let ε_δ denote model based expected value.

Definition 2.1:

The sampling strategy τ_{π^*} is said to be δ -better than another strategy τ_{π} if $\varepsilon_\delta V[\tau_{\pi^*}] \leq \varepsilon_\delta V[\tau_{\pi}]$, the equality not holding true identically.

Definition 2.2:

The sampling strategy τ_{π^*} is said to be δ -optimum strategy if $\varepsilon_\delta V[\tau_{\pi^*}] \leq \varepsilon_\delta V[\tau_{\pi}]$, for any other strategy τ_{π} when equality does not hold identically.

Thus, instead of minimizing the design based variance

$$V(\tau_{\pi}) = E_{P_n} [\hat{Y}_{H-T} - Y]^2.$$

We minimize, $\varepsilon_{\delta} V(\tau_{\pi}) = \varepsilon_{\delta} E_{P_n} \left[\hat{Y}_{H-T} - Y \right]^2$.

Theorem 2.1:

Under a super population model δ , the strategy τ_{π} defined in (2.1) would be δ -optimum, if the choice of $\pi_i, (i = 1, 2, \dots, N)$ be made as

$$\pi_i \propto \left[(\varepsilon_{\delta} y_i)^2 + V_{\delta}(y_i) \right]^{-1/2}$$

where V_{δ} stands for the variance under- model.

Proof:

For simplicity of notation, we have used ε in place of ε_{δ} all through; Let

$$\begin{aligned} \varphi(y, \pi, \lambda, \lambda^*, \lambda_i) &= \varepsilon V(\hat{Y}_{H-T}) + \lambda (\sum \pi_i - n) \\ &\quad + \sum \lambda_i \left(\sum_{j \neq i} \pi_{ij} - (n-1)\pi_i \right) + \lambda^* \left(\sum_{i \neq j} \pi_{ij} - n(n-1) \right), \\ &= \text{Constant} + \sum_{i=1}^N \varepsilon (y_i)^2 / \pi_i + \sum_{i < j} \frac{\varepsilon (y_i y_j) \pi_{ij}}{\pi_i \pi_j} \\ &\quad + \sum_{i > j} \frac{\varepsilon (y_i y_j) \pi_{ij}}{\pi_i \pi_j} + \lambda (\sum \pi_i - n) \\ &\quad + \sum \lambda_i \left(\sum_{j \neq i} \pi_{ij} - (n-1)\pi_i \right) + \lambda^* \left(\sum_{i \neq j} \pi_{ij} - n(n-1) \right). \end{aligned}$$

Differentiating φ , w.r.t. π_{ij} ; we have

$$\frac{\partial \varphi}{\partial \pi_{ij}} = \frac{\varepsilon (y_i y_j)}{\pi_i \pi_j} + \lambda^* + \lambda_i \quad (i < j),$$

and

$$\frac{\partial \varphi}{\partial \pi_{ji}} = \frac{\varepsilon (y_i y_j)}{\pi_i \pi_j} + \lambda^* + \lambda_i \quad (i > j).$$

But for a given pair (i, j) , we have $\frac{\partial \varphi}{\partial \pi_{ij}} = \frac{\partial \varphi}{\partial \pi_{ji}}$ and hence,

$$\frac{\varepsilon (y_i y_j)}{\pi_i \pi_j} + \lambda^* + \lambda_i = \frac{\varepsilon (y_i y_j)}{\pi_i \pi_j} + \lambda^* + \lambda_j.$$

Therefore, we have $\lambda_i = \lambda_j$;

Similarly, for another pair (j,k) , we have $\lambda_j = \lambda_k$, thus having,

$$\lambda_i = \lambda_j = \lambda_k .$$

Thus proceeding as before, we have

$$\lambda_1 = \lambda_2 = \dots = \lambda_i = \dots = \lambda_n = c \text{ (Say).}$$

Equating $\frac{\partial \phi}{\partial \pi_{ij}} = \frac{\partial \phi}{\partial \pi_{ji}} = 0$, we have

$$\frac{\varepsilon(y_i y_j)}{\pi_i \pi_j} = -(\lambda^* + c) \text{ for each } (i, j). \quad (2.3)$$

$$\begin{aligned} \text{Again } \frac{\partial \phi}{\partial \pi_i} &= -\frac{\varepsilon(y_i^2)}{\pi_i^2} - 2 \sum_{j \neq i} \frac{\varepsilon(y_i y_j)}{\pi_i^2 \pi_j} + \lambda - (n-1)c \\ &= -\frac{\varepsilon(y_i^2)}{\pi_i^2} + \frac{2(\lambda^* + c)}{\pi_i} + \left(\sum_{j \neq i} \pi_{ij} \right) + \lambda - (n-1)c \quad [\text{by (2.3)}] \\ &= -\frac{\varepsilon(y_i^2)}{\pi_i^2} + \frac{2(\lambda^* + c)(n-1)\pi_i}{\pi_i} + \lambda - (n-1)c \\ &= -\frac{\varepsilon(y_i^2)}{\pi_i^2} + 2\lambda^*(n-1) + 2c(n-1) + \lambda - (n-1)c \\ &= -\frac{\varepsilon(y_i^2)}{\pi_i^2} + 2\lambda^*(n-1) + \lambda + c(n-1) \\ &= -\frac{\varepsilon(y_i^2)}{\pi_i^2} + \lambda', \end{aligned}$$

where, $\lambda' = 2\lambda^*(n-1) + \lambda + c(n-1)$.

Now equating $\frac{\partial \phi}{\partial \pi_i} = 0$, we have,

$$\pi_i^2 \alpha \varepsilon(y_i^2) \text{ i.e., } \pi_i \alpha \sqrt{\varepsilon(y_i^2)} \text{ for } i = 1, 2, \dots, N .$$

Let the optimum choice be

$$\pi_{0i} = K' \sqrt{\varepsilon(y_i^2)}. \quad (2.4)$$

Thus, $\pi_0 = (\pi_{01}, \pi_{02}, \dots, \pi_{0i}, \dots, \pi_{0N})$ is an extreme point; For π_0 to be a minimum point, we proceed as follows:

Let $\pi_* = (\pi_{*1}, \pi_{*2}, \dots, \pi_{*i}, \dots, \pi_{*n})$ be any other inclusion probability vector. We have

$$\begin{aligned} V(\tau_{\pi_*}) - V(\tau_{\pi_0}) &= \sum_{i=1}^N \left(\frac{1}{\pi_{*i}} - \frac{1}{\pi_{0i}} \right) \varepsilon(y_i^2) = K \sum_{i=1}^N \left(\frac{1}{\pi_{*i}} - \frac{1}{\pi_{0i}} \right) \pi_{0i}^2 \\ &= K \sum \pi_{*i} \left(\frac{\pi_{0i}^2}{\pi_{*i}^2} - \frac{\pi_{0i}}{\pi_{*i}} \right) = K \sum \pi_{*i} \left(\frac{\pi_{0i}}{\pi_{*i}} - 1 \right)^2 > 0. \end{aligned} \tag{2.5}$$

Now, combining (2.4) and (2.5), we have the desired result.

3. ‘NEAR-OPTIMUM’ PROBABILITIES UNDER A MODEL

As may be seen optimum inclusion probabilities $\pi_{0i} (i = 1, 2, \dots, N)$ will depend on the exact value of the model parameters and can not be used in practice unless they are known exactly. If one considers the super-population model specified as

$$\begin{aligned} \varepsilon_{\delta}(y_i | x_i) &= \beta_0 + \beta' x_i \\ V_{\delta}(y_i | x_i) &= \sigma^2 \end{aligned} \tag{3.1}$$

and

$$Cov(y_i y_j | x_i, x_j) = 0,$$

the π_{0i} 's ($i = 1, 2, \dots, N$) will depend on model parameters $\beta_0, \beta = (\beta_1, \beta_2, \dots, \beta_q)$ and σ^2 . What we propose here is to find near-optimum inclusion probabilities which are obtained when estimates of model parameters are used in defining π_0 .

Along the same line as in Tripathi and Choubey (1998), we have the following theorem.

Theorem 3.1:

Let the model parameters in (3.1) be estimated by ordinary least squares method, then ‘near-optimum’ inclusion probabilities for the strategy τ_{π_0} are given by

$$\begin{aligned} \pi_{0i} &\propto \left[1 + N \sum_{\alpha=1}^q \sum_{\alpha=1}^q \rho_{0\alpha} C_{\alpha} \bar{x}_{\alpha} \bar{x}_{\alpha} \cdot \sigma^{\alpha\alpha} \left(P_{\alpha i} - \frac{1}{N} \right) \right] \\ &\times \left[1 + \frac{N}{N-q-1} \frac{(1-\rho_{0(1-q)}) c_o^2}{1 + N \sum_{\alpha} \sum_{\alpha} \rho_{0\alpha} \cdot C_{\alpha} \bar{x}_{\alpha} \bar{x}_{\alpha} \cdot \sigma^{\alpha\alpha} \left(P_{\alpha i} - \frac{1}{N} \right)} \right]^{1/2} \end{aligned}$$

where, $\bar{x}_\alpha = \sum_{i \in J} X_{\alpha i} / N$; $P_{\alpha i} = X_{\alpha i} / N \bar{x}_\alpha$; $\rho_{0,\alpha}$ denote the correlation coefficient of y with X_α ; C_0 and C_α denote the coefficient of variation of y and x_α respectively, $\sigma^{\alpha\alpha}$ is the (α, α) element of the inverse of the variance covariance matrix of (x_1, x_2, \dots, x_q) and $\rho_{0(1,\dots,q)}^2$ stands for the multiple correlation coefficient between y and $\underline{x} = (x_1, x_2, \dots, x_q)$.

Proof: See Tripathi and Chaubey (1998).

Since, complete information on the auxiliary variables are available, the knowledge of $C_\alpha, \bar{x}_\alpha, \sigma^{\alpha\alpha}$ can be easily obtained, but $\rho_{0,\alpha}, \rho_{0(1,\dots,q)}^2$ and C_0 has to be obtained through previous census, surveys or a pilot survey.

REFERENCES

1. Agrawal, S.K. and Singh, M. (1980). Use of multivariate auxiliary information in selection of units in probability proportional to size with replacement. *Jour. Ind. Soc. Agri. Stat.*, XXXII, No. 3, 71-81.
2. Brewer, K.R.W. and Hanif, M. (1983). *Sampling with unequal probabilities*. Springer-Verlag, New Work.
3. Hansen, M.H. and Hurwitz, W.N. (1983). On the theory of sampling from finite populations. *Ann. Math. Stat.*, 1014, 333-362.
4. Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, 17, 663-685.
5. Koop, J.C. (1963). On axioms of sample formation and their bearing on construction of linear estimators in sampling theory of finite universes, *Metrika*, 7(1), 81-114.
6. Maiti, P. and Tripathi, T.P. (1976)/ The Use of multivariate auxiliary information in selecting the sampling units. *Symposium on Survey methodology*, March 22-27, 1976, ISI, Calcutta.
7. Maiti, P. and Tripathi, T.P. (2002). Optimum Inclusion probabilities in 'WOR-Sampling Scheme' based on Super Population Model and Multivariate information. *Proceedings of the V International Symposium on Optimization and Statistics*, 28-30.
8. Tripathi, T.P. and Chaubey, Y.P. (1998). *Optimum probabilities of selection in pps-sampling based on super population model and multivariate information; Probability and Statistics*, Edited by S.P. Mukherjee, S.K. Basu and B.K. Sinha. 365-369.
9. Tripathi, T.P. and Maiti, P. (1980). Use of information on two auxiliary variates in selecting the sampling units. *Stat-Math Tech. Report No. 18/80*, ISI, Calcutta.
10. Tripathi, T.P. (1970). *Contributions to the theory using multivariate information*. Ph.D. thesis submitted to the Punjabi University, Patiala.