Metrika, Volume 31, 1984, page 25–32.

Some Aspects of Random Permutation Models in Finite Population Sampling Theory

By *T.J. Rao*, Calcutta[1])

Summary: We first consider Neyman's optimum allocation of sample size to strata in the light of available auxiliary information for which a suitable random permutation model is assumed. For a special case of this model the allocation of the sample size reduces to the same as when a certain superpopulation regression model is assumed. Motivated by this, more generally, we discuss some optimality results under random permutation models and compare them with the corresponding results when a superpopulation regression model is assumed.

1. Introduction

Consider a finite population of size N divided into k strata of sizes N_i, $i = 1, 2, \ldots, k$. Let Y be the study variate taking values Y_{ij} on the j-th unit of the i-th stratum; values X_{ij} of X, a positive auxiliary variate usually related to the variate Y under study, are available for all units $j = 1, 2, \ldots, N_i$; $i = 1, 2, \ldots, k$. We are interested in estimating parametric functions of Y such as the population mean $\bar{Y} = \Sigma\,\Sigma\, Y_{ij} / N$ or the population total $Y = N\bar{Y}$, based on a stratified sampling design. For Simple Random Sampling With Replacement (SRSWR) in each stratum, we have Neyman's optimum allocation [*Neyman*] given by $n_{i,\text{opt.}} = n\, N_i\, \sigma_i \,/\, \Sigma\, N_i\, \sigma_i$, where n is the total sample size and σ_i^2 is the within stratum variance for the i-th stratum, $i = 1, 2, \ldots, k$. Computation of the $n_{i,\text{opt.}}$'s requires at least the proportionate values of σ_i^2's which are unknown. In practice, values α_i^2's, based on a pilot study or available prior information are substituted for σ_i^2's. These values are usually the within stratum variances of the auxiliary variate related to the study variate.

The justification for the assumption mentioned above that the unknown proportionate values of σ_i^2's are usually not very different from the known α_i^2's has been examined in the light of *a priori* distributions specified by suitable superpopulation models by *Hanurav* [1965] and *Rao* [1968, 1977]. In this paper we shall consider a different

---

[1]) *T.J. Rao*, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, India.
0026-1335/84/010025-32$2.50

superpopulation model applicable when auxiliary information on the variate X is available and the labels of the units are assumed to be uninformative with respect to the values $R_{ij} = Y_{ij} / X_{ij}$ and study the problem of Neyman's optimum allocation with σ_i^2's substituted by α_i^2's where α_i^2's are the within stratum variances for the auxiliary variate X. For a special case of this model, it is found that the allocation turns out to be the same as for the case when a particular superpopulation regression model is assumed. Motivated by this, we look at these two models and draw some parallels between inference from them.

2. Neyman's Optimum Allocation and Random Permutation Models

For SRSWR in each stratum, Neyman's optimum allocation [*Neyman*] is given by $n_{i,\text{opt.}} = nN_i\, \sigma_i / \Sigma N_i\, \sigma_i$, where

$$\sigma_i^2 = (\sum_{j=1}^{N_i} Y_{ij}^2 - Y_i^2 / N_i) / N_i,\ Y_i = \sum_j Y_{ij}. \tag{2.1}$$

When auxiliary information on the variate X is available and further it is assumed that the labels attached to the units are uninformative with respect to the values $R_{ij} = Y_{ij} / X_{ij}$, where Y_{ij} and X_{ij} are respectively the y- and the x-values of the j-th unit of the i-th stratum, we consider a random permutation model [*C.R. Rao; J.N.K. Rao/Bellhouse*] where response errors are not present. For fixed $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, N_i$ we consider $R_{ij} = Y_{ij} / X_{ij}$ and X_{ij} to be unrelated and treat R_{ij}'s as random permutation of an unknown set of N_i numbers with X_{ij}'s fixed. This corresponds to the model [*J.N.K. Rao/Bellhouse*]

$$\left.\begin{aligned} &\epsilon R_{ij} = \bar{R}_i \\ &\epsilon (R_{ij} - \bar{R}_i)^2 = \sigma_{Ri}^2 \\ &\epsilon (R_{ij} - \bar{R}_i)(R_{ij'} - \bar{R}_i) = -\sigma_{Ri}^2 / (N_i - 1) \\ &\qquad\qquad j \neq j' = 1, 2, \ldots, N_i \end{aligned}\right\} \tag{2.2}$$

where $\bar{R}_i = \sum_j R_{ij} / N_i$ and σ_{Ri}^2 for fixed i is the variance of R_{ij}'s and ϵ denotes the Expectation for this random permutation model.

When σ_i^2 of (2.1) which is now rewritten as

$$\sigma_i^2 = (\sum_j R_{ij}^2 X_{ij}^2 - (\sum_j R_{ij} X_{ij})^2 / N_i) / N_i \tag{2.3}$$

is not known, it is usual to compute the known value

$$\alpha_i^2 = (\sum_j X_{ij}^2 - (\sum_j X_{ij})^2 / N_i) / N_i$$

and use that in the derivation of the optimum allocation. We shall now give a justification for doing this in view of the model (2.2) assumed above. Under the model we have the average value of σ_i^2 over permutations of R_{ij}'s, given by

$$\epsilon\, \sigma_i^2 = (\sum_j \epsilon\, R_{ij}^2 X_{ij}^2 - \epsilon\, (\sum_j R_{ij} X_{ij})^2 / N_i) / N_i$$

$$= [\sum_j X_{ij}^2 (\sigma_{Ri}^2 + \bar{R}_i^2) - \{\sum_j X_{ij}^2 (\sigma_{Ri}^2 + \bar{R}_i^2)$$

$$+ \sum_{j \neq j'} \sum X_{ij} X_{ij'} (\bar{R}_i^2 - \sigma_{Ri}^2 / (N_i - 1))\} / N_i] / N_i$$

$$= (\sigma_{Ri}^2 + \bar{R}_i^2)\, \alpha_i^2 + \sigma_{Ri}^2 \sum_{j \neq j'} \sum X_{ij} X_{ij'} / N_i (N_i - 1)$$

$$= (\sigma_{Ri}^2 + \bar{R}_i^2)\, \alpha_i^2 - \sigma_{Ri}^2\, \alpha_i^2 / (N_i - 1) + \sigma_{Ri}^2 \bar{X}_i^2$$

$$\sim (\sigma_{Ri}^2 + \bar{R}_i^2)\, \alpha_i^2 + \sigma_{Ri}^2 \bar{X}_i^2, \text{ when } (N_i - 2) / (N_i - 1) \sim 1. \tag{2.4}$$

Thus the average value of σ_i^2 will be proportional to α_i^2 provided α_i^2 is proportional to $\bar{X}_i^2$ for the special case of the model when σ_{Ri}^2 and $\bar{R}_i$ are equal for all strata. Hence for this special case, in order to obtain the optimum allocation, instead of the unknown σ^2 one can substitute its average value under the model which is in terms of α_i^2 of the known x-values provided the coefficients of variation (c.v.) of the x-variate are equal in all strata. When this condition is satisfied, Neyman's optimum allocation reduces to allocation proportional to $N_i \bar{X}_i = X_i$, the total of x-values in the i-th stratum.

Remark 2.1. The case of Simple Random Sampling Without Replacement in each stratum can be similarly discussed where the optimum allocation is given by $n_{i,\text{opt.}} \propto N_i S_i$, where $S_i^2 = N_i \sigma_i^2 / (N_i - 1)$.

Remark 2.2. Consider a superpopulation regression model where $\underset{\sim}{Y} = (Y_{11}, Y_{12}, \ldots, Y_{kN_k})$ is assumed to be a realization of an N-length random vector with a distribution depending on $\underset{\sim}{X} = (X_{11}, X_{12}, \ldots, X_{kN_k})$ such that

$$\left.\begin{aligned} &E(Y_{ij} \mid X_{ij}) \propto X_{ij} \\ &V(Y_{ij} \mid X_{ij}) \propto X_{ij}^2 \\ &C(Y_{ij}, Y_{i'j'} \mid X_{ij}, X_{i'j'}) = 0, \end{aligned}\right\} \tag{2.5}$$

where the script letters $\mathcal{E}$, $\mathcal{V}$ and $\mathcal{C}$ denote respectively the Expectation, Variance and covariance for this superpopulation.

Under this model for SRSWR in each stratum, we have the expected value of σ_i^2 proportional to α_i^2 provided the c.v. of x-values are equal in all strata [*Hanurav/Rao*, 1968]. This is the same condition as obtained for the special case of the random permutation model considered above. Thus the conclusions here are the same for special cases of these two models. Motivated by this we now consider the two models (2.2) and (2.5) and look at the expressions for the expected variance of a general homogeneous linear unbiased estimator $\hat{Y} = \sum_{i \in s} \beta_{si} y_i$ of the population total Y and compare them.

3. Superpopulation Regression Model versus Random Permutation Model

Consider a finite population of size N. Let Y_i and X_i be the values taken by the i-th unit on the study variate and the auxiliary variate respectively, $i = 1, 2, \ldots, N$. Let $\hat{Y} = \sum_{i \in s} \beta_{si} y_i$ be an unbiased estimator for the population total $Y = \Sigma Y_i$. For any fixed sample size design p, we have

$$V(\hat{Y}) = \sum_s \left(\sum_{i \in s} \beta_{si} y_i - Y\right)^2 p(s)$$

$$= \sum_i Y_i^2 \left(\sum_{s \ni i} \beta_{si}^2 p(s) - 1\right) + \sum_{i \neq j} \sum Y_i Y_j \left(\sum_{s \ni i,j} \beta_{si} \beta_{sj} p(s) - 1\right),$$

where V denotes the sampling design Variance.

Under the superpopulation regression model of the type (2.5) for which

$$\left.\begin{aligned} &\mathcal{E}(Y_i \mid X_i) = a X_i \\ &\mathcal{V}(Y_i \mid X_i) = \sigma^2 X_i^2 \\ &\mathcal{C}(Y_i, Y_j \mid X_i, X_j) = 0 \end{aligned}\right\} \qquad (3.1)$$

following *Godambe* [1955] we have

$$\mathcal{E}V(\hat{Y}) = \sigma^2 \sum_i X_i^2 \left(\sum_{s \ni i} \beta_{si}^2 p(s) - 1\right) + a^2 V\left(\sum_{i \in s} \beta_{si} x_i\right). \qquad (3.2)$$

Under the random permutation model of the type (2.2) for which the corresponding moments are

$$\left.\begin{aligned} &\epsilon (R_i) = \bar{R} \\ &\epsilon (R_i - \bar{R})^2 = \sigma_R^2 \\ &\epsilon (R_i - \bar{R})(R_j - \bar{R}) = -\sigma_R^2 / N-1 \end{aligned}\right\} \tag{3.3}$$

where $\bar{R} = \Sigma R_i / N$ and σ_R^2 is the variance of $R_i = Y_i / X_i$, we have from *Rao J.N.K.* [1975] that

$$\epsilon V(\hat{Y}) = S_R^2 \sum_i X_i^2 \left(\sum_{i \ni s} \beta_{si}^2 p(s) - 1\right)$$

$$+ V\left(\sum_{i \in s} \beta_{si} x_i\right)(\bar{R}^2 - \sigma_R^2 / (N-1)). \tag{3.4}$$

From (3.2) and (3.4) it follows that

$$EV(\hat{Y}) = \text{constant } \epsilon V(\hat{Y}) \tag{3.5}$$

provided $\sum_{i \in s} \beta_{si} x_i = \text{constant} = X$.

Furthermore, when this condition of model-unbiasedness is satisfied we have that both $E\,V(\hat{Y})$ and $\epsilon\,V(\hat{Y})$ are minimized when $\beta_{si} = 1 / \pi_i$. It is thus easy to see that the optimum strategy in both cases is given by (πPS design, $\hat{Y}_{HT} = \sum_{i \in s} y_i / \pi_i$) provided C_R = c.v. of R_i's $= S_R / \bar{R} \leqslant \sqrt{N}$. This is the 'similarity' mentioned in J.N.K. Rao [1975] between his and *Godambe's* [1955] result.

For the more general superpopulation model (3.1) with $V(Y_i \mid X_i) = \sigma^2 X_i^g$, we have

$$EV(\hat{Y}) = \sigma^2 \sum_i X_i^g \left(\sum_{s \ni i} \beta_{si}^2 p(s) - 1\right)$$

$$+ a^2 V\left(\sum_{i \in s} \beta_{si} x_i\right). \tag{3.6}$$

Further, for the random permutation model with

$$\epsilon R_i' = \bar{R}'$$

$$\epsilon (R_i' - \bar{R}')^2 = \sigma_{R'}^2$$

$$\epsilon (R_i' - \bar{R}')(R_j' - \bar{R}') = -\sigma_{R'}^2 / N-1 \quad \text{for } i \neq j$$

where $R_i' = Y_i / X_i^{g/2}$, $\bar{R}' = \Sigma R_i' / N$ and $\sigma_{R'}^2$ is the variance of R_i''s [see *Rao, J.N.K.*], we have

$$eV(\hat{Y}) = S^2_{R'} \sum_i X_i^g \left(\sum_{s \ni i} \beta^2_{si}\, p(s) - 1 \right)$$

$$+ (\bar{R}'^2 - S^2_{R'} / N)\, V\left(\sum_{i \in s} \beta_{si}\, x_i^{g/2} \right). \quad (3$$

Notice that while the minimization of $EV(\hat{Y})$ is attained for the strategy ($G\pi PS$ design, $\hat{Y}_{HT}$), $\epsilon V(\hat{Y})$ is minimized for the strategy consisting of the πPMS design (π_i's Proportional to Modified Size design) where the Modified Size is $X_i^{g/2}$ and the corresponding Horvitz-Thompson estimator $\hat{Y}_{HT} = \sum_{i \in s} y_i / \pi_i$, where $\pi_i \propto X_i^{g/2}$ when $C_R \leqslant \sqrt{N}$ as before.

Furthermore, it may be observed that the optimum design in the latter case is a straightforward πPS design which is easy to construct where as in the $G\pi PS$ case we have the additional condition that $\sum_{i \in s} x_i^{1-g/2} =$ constant.

Next consider the modified Horvitz-Thompson estimator

$$\hat{Y}^* = \hat{Y}_{HT} + k\,(X - \hat{X}_{HT}) \quad (3.$$

where k is a constant. Under the model (3.1) with $V(Y_i \mid X_i) = \sigma^2 X_i^g$ we have

$$EV(\hat{Y}^*) = \sigma^2 \sum_i \left(\frac{1}{\pi_i} - 1 \right) X_i^g + (a-k)^2\, V(\hat{X}_{HT}).$$

When $g = 2$,

$$\text{Min.}\,(EV(\hat{Y}^*)\big|_{k=a}) = \text{Min.}\ EV(\hat{Y}_{HT}, \pi PS) \quad (3.$$

and for general g,

$$EV(\hat{Y}^*)\big|_{k=a} = \sigma^2 \sum_i \left(\frac{1}{\pi_i} - 1 \right) X_i^g \text{ which is minimized when}$$

$\pi_i \propto X_i^{g/2}$ (subject to the condition $\sum_i \pi_i = n$). Thus

$$\text{Min.}\,(EV(\hat{Y}^*)\big|_{k=a}) = \text{Min.}\ EV(\hat{Y}_{HT}, G\pi PS) \quad (3.10$$

where $G\pi PS$ design is such that $\pi_i \propto X_i^{g/2}$ and $\sum_{i \in s} x_i^{1-(g/2)} =$ constant.

For the Random Permutation Model we have

$$\epsilon V(\hat{Y}^*) = \sum_i \left(\frac{1}{\pi_i} - 1 \right) S_R^2\, X_i^2 + (k - \bar{R})^2\, V(\hat{X}_{HT})$$

$$- (\sigma^2 / N - 1)\, V(\hat{X}_{HT}).$$

Hence

$$\epsilon V(\hat{Y}^*)\Big|_{k=\bar{R}} = \sum_i \left(\frac{1}{\pi_i} - 1\right) S_R^2 X_i^2 - (\sigma^2 / N - 1)\, V(\hat{X}_{HT})$$

and

$$\text{Min. } \epsilon V(\hat{Y}^*)\Big|_{k=\bar{R}} = \text{Min. } \epsilon V(\hat{Y}_{HT}, \pi PS) \tag{3.11}$$

a result similar to (3.9). On the other hand, when we consider the model with X_i replaced by $X_i^{g/2}$ we get

$$\text{Min. } \epsilon V(\hat{Y}^*)\Big|_{k=\bar{R}} = \text{Min. } \epsilon V(\hat{Y}_{HT}, \pi PX^{g/2}) \tag{3.12}$$

where on the r.h.s. of (3.12) the design is a simple πPS design, sizes being $X_i^{g/2}$, $i = 1, 2, \ldots, N$ while the design on the r.h.s. of (3.10) is a Generalized πPS design.

Ramakrishnan [1970] considered the class of ϵ-unbiased estimators $\hat{Y} = \sum_{i \in s} \beta_{si} y_i$ of Y and demonstrated that the optimum value of β_{si} which minimizes the average m.s.e. of $\hat{Y}$ is given by $\beta_{si} = 1 + (X - \Sigma x_i) / nx_i$ which gives the optimum estimator

$$\hat{Y}_2 = \sum_{i \in s} y_i + (X - \sum_{i \in s} x_i) \sum_{i \in s} (y_i / x_i) / n. \tag{3.13}$$

From this it follows that, for the class of estimators defined by

$$\hat{Y}' = \sum_{i \in s} y_i + (X - \sum_{i \in s} x_i) (\sum_{i \in s} \gamma_{si} y_i / \sum_{i \in s} \gamma_{si} x_i) \tag{3.14}$$

where γ_{si} are arbitrary weights,

ϵ M.S.E. $(\hat{Y}')$ is minimum for the choice of weights γ_{si} given by

$$\gamma_{si} x_i / \Sigma \gamma_{si} x_i = 1/n$$

which leads to $\gamma_{si} \propto x_i^{-1}$ leading to the estimator (3.13). Under a different context *Brewer* [1979] considers the class of estimators of the type (3.14) under a super-population regression model.

References

Brewer, K.R.W.: A class of robust sampling designs for large-scale surveys. J. Amer. Statist.Assoc. 74, 1979, 911–915.

Godambe, V.P.: A unified theory of sampling from finite populations. J. Roy. Statist. Soc. B, 17, 1955, 269–278.

Hanurav, T.V.: Optimum sampling strategies and some related problems. Ph. D. Thesis submitted to the Indian Statistical Institute, 1965.
Neyman, J.: On two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. J. Roy. Statist. Soc. 97, 1934, 558–625.
Ramakrishnan, M.K.: Optimum estimators and strategies in survey sampling. Ph. D. Thesis submitted to the Indian Statistical Institute, 1970.
Rao, C.R.: Some aspects of statistical inference in problems of sampling from finite populations. In: Foundations of statistical inference. Ed. by V.P. Godambe, and D.A. Sprott. Toronto 1971, 177–202.
Rao, J.N.K.: On the foundations of survey sampling. In: A survey of statistical design and linear models. Ed. by J.N. Srivastava. The Hague 1975, 489–505.
Rao, J.N.K., and *D.R. Bellhouse*: Optimal estimation of a finite population mean under generalized random permutation models. J. Stat. Plan. and Inference 2, 1978, 125–141.
Rao, T.J.: On the allocation of sample size in stratified sampling. Ann. Inst. Stat. Math. 20, 1968, 159–166.
–: Optimum allocation of sample size and prior distributions: a review. Inst. Stat.Rev. 45, 1977, 173–179.

Received November 10, 1980
(revised version March 23, 1981)