ON ESTIMATING PARAMETRIC FUNCTIONS IN STRATIFIED SAMPLING DESIGNS

By DES RAJ

Indian Statistical Institute, Calcutta

1. The problem considered

In the usual theory of stratified sampling designs, we are generally interested in estimating the mean value or aggregate of a character for a finite population. We may, however, come across situations in which the object is to estimate certain linear functions of the stratum means. As a case in point, we may require to estimate the area under food crops not only for a province as a whole but also for a particular group of strata within the province where for instance famine might have occurred or it is feared that much area is being diverted to crops of one type at the expense of crops of a more beneficial type. As another example, we may wish to estimate not only the average consumption of rice by all the inhabitants in a city but also the average consumptions by the primarily rice-eating classes and primarily wheat-eating classes so that Government may be in a position to find out how much rice should be procured for the city and in what manner it should be rationed to the different classes.

2. Estimation of parametric functions

In general, if the population consists of k strata of sizes $N_j (j = 1, 2, ..., k)$ and mean values

$$\bar{Y}_j (j = 1, 2, ..., k), \qquad \dots \quad (2.1)$$

we are interested in estimating $r < k$ linear functions

$$L_i = \sum_{j=1}^{k} l_{ij} \bar{Y}_j \quad (i = 1, 2, ..., r) \qquad \dots \quad (2.2)$$

of the stratum means, the matrix of co-efficients (l_{ij}) being known.

For a particular distribution $n_j (j = 1, 2, ..., k)$ of the total sample of size n, obviously the best unbiassed estimate of L_i is

$$\hat{L}_i = \sum_{j=1}^{k} l_{ij} \bar{y}_j \qquad (i = 1, 2, ..., r) \qquad \dots \quad (2.3)$$

where $\bar{y}_j$ is the sample mean based on n_j observations in the j-th stratum.

These estimates are best (Basu, 1952) in the sense that for any convex (downwards) loss function they are admissible i.e., no other estimators having a uniformly smaller risk function exist. Also we have

$$V(\hat{L}_i) = \Sigma\, l_{ij}^2\, \sigma_j^2 \left(\frac{1}{n_j} - \frac{1}{N_j}\right)$$

where

$$\sigma_j^2 = \frac{\sum_{k=1}^{N_j} (y_{jk} - \bar{Y}_j)^2}{N_j - 1} \qquad \dots \quad (2.4)$$

is the variance of the j-th stratum. Further the best (in the sense stated above) unbiassed quadratic estimates of $V(\hat{L}_i)$ are given by

$$\hat{V}(\hat{L}_i) = \Sigma\, l_{ij}^2 s_j^2 \left(\frac{1}{n_j}-\frac{1}{N_j}\right)$$

where

$$s_j^2 = \frac{\sum\limits_{k=1}^{n_j} (y_{jk}-\bar{y}_j)^2}{n_j-1}. \qquad \dots \quad (2.5)$$

3. Optimum allocation of sampling units

An important question emerges: How should the total sample be distributed among the different strata ? Obviously, there would not be a single answer to this question. In fact, the answer will depend on what the sample is desired to achieve. In what follows we shall consider some approaches to the problem.

3.1. *Minimisation of cost plus loss :* If the results obtained from the sample are going to form the basis of some practical action, we may be able to calculate in monetary terms the 'loss' that will be incurred in a decision through an error of amount d in the estimate. For example, if this loss be $\mu_i d^2$ (cf. Yates, p. 292) and the estimate be unbiassed, the average loss in a series of samples of the same type and size will be $\mu_i V(\hat{L}_i)$. The purpose in taking the sample may be to diminish the sum of the total expected loss

$$L = \sum_{i=1}^{r} \mu_i\, V(\hat{L}_i) \qquad \dots \quad (3.1.1)$$

and the total cost (cf. Kitagawa, p. 338)

$$C = \Sigma\, c_j n_j^g \ (g > 0). \qquad \dots \quad (3.1.2)$$

Or, for a fixed cost given by (3.1.2), the object may be to diminish the expected loss given by (3.1.1). The cost function used here is more general than the usual cost function

$$C = \Sigma c_j n_j.$$

In the former case, the function to be minimised is

$$G = \Sigma\, c_j n_j^g + \Sigma q_j \sigma_j^2 \left(\frac{1}{n_j}-\frac{1}{N_j}\right) \qquad \dots \quad (3.1.3)$$

where

$$q_j = \sum_i \mu_i l_{ij}^2. \qquad \dots \quad (3.1.4)$$

The stationary value of G is given by

$$n_j^{g+1} = \frac{q_j \sigma_j^2}{g c_j}. \qquad \dots \quad (3.1.5)$$

It is easy to verify that the stationary value is an actual minimum of the function. In the latter case, the solution is given by

$$n_j = \mu\left(\frac{q_j\sigma_j^2}{gc_j}\right)^{\frac{1}{g+1}} \qquad \ldots \quad (3.1.6)$$

where

$$\mu = \frac{C^{1/g}}{\left[\Sigma(q_j\sigma_j^2/g)^{\frac{g}{g+1}}/c_j^g\right]^{1/g}}\,. \qquad \ldots \quad (3.1.7)$$

3.2. *Minimisation of cost:* As an alternative approach to the problem, we may consider the survey to be useful if the parametric functions L_i $(i = 1, 2, \ldots, r)$ are estimated with some desired variances $a_i\,(i = 1, 2, \ldots, r)$. Generally approximate value of L_i's are known on the basis of some previous survey and a_i's are determined by the requirement that the standard errors of the estimates are some specified percentages of the mean values known approximately. In such a case, we have to allocate the total sample to the different strata in such a way that the cost of the survey is made smallest. We have then to minimise

$$f(n_1, n_2, \ldots, n_k) = \Sigma c_j n_j^g$$

subject to the conditions $\phi_i = V(\hat{L}_i) = a_i (i = 1, 2, \ldots, r)$.

The stationary values are given by the equations

$$m_j^{g+1} = gc_j[\sigma_j^2 \sum_i \lambda_i l_{ij}^2]^{-1} \qquad (j = 1, 2, \ldots, k) \qquad \ldots \quad (3.2.1)$$

$$\sum_j l_{ij}^2 m_j \sigma_j^2 = a_i + \sum_j l_{ij}^2 \sigma_j^2/N_j \qquad (i = 1, 2, \ldots, r) \qquad \ldots \quad (3.2.2)$$

where λ's are Lagrange's undetermined multipliers and m_j is the reciprocal of n_j.

These equations are not algebraic in the variables involved. It is, therefore, not possible to provide any explicit mathematical solutions for the general case. One has to solve these equations by iterative processes. For example, if $r = 2$, an approximate solution can easily be obtained by Newton-Raphson method. In this method, we choose λ_1 so that the difference between the values of λ_2 calculated from the two equations obtained from (3.2.2) by substituting for m_j from (3.2.1) is small and positive. We find another value of λ_1 for which this difference is small and negative. By simple interpolation values of λ_1 and λ_2 are obtained which lead to to the optimum allocation.

One may like to verify that the stationary point obtained is an actual minimum of $f(n_1, n_2, \ldots, n_k)$. Putting $f+\lambda_1\phi_1+\lambda_2\phi_2+\ldots+\lambda_r\phi_r = F$ and denoting by λ^{ij} the second partial derivative of h *w.r.t.* n_i and n_j, the condition is that the restricted Hessian (Chaundy, 1935)

$$\begin{vmatrix} F^{11} & F^{12} & \ldots & F^{1k} & \phi_1^1 & \phi_2^1 & \ldots & \phi_r^1 \\ F^{21} & F^{22} & \ldots & F^{2k} & \phi_1^2 & \phi_2^2 & \ldots & \phi_r^2 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ F^{k1} & F^{k2} & \ldots & F^{kk} & \phi_1^k & \phi_2^k & \ldots & \phi_r^k \\ \phi_1^1 & \phi_1^2 & \ldots & \phi_1^k & 0 & 0 & \ldots & 0 \\ \phi_2^1 & \phi_2^2 & \ldots & \phi_2^k & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \phi_r^1 & \phi_r^2 & \ldots & \phi_r^k & 0 & 0 & \ldots & 0 \end{vmatrix} \qquad \ldots \quad (3.2.3)$$

and its principal minors should have the sign $(-1)^r$. It may also be noted that the solution so obtained is formally equivalent to the following:

If $(n_1', n_2, \ldots, n_k')$ is a minimum of f subject to the r conditions

$$\phi_i = V(\hat{L}_i) = a_i \qquad (i = 1, 2, \ldots, r)$$

it is also a minimum or maximum of ϕ_s subject to the r conditions

$$f = f(n_1', n_2', \ldots, n_k'),$$
$$\phi_i = a_i \quad (i = 1, 2, \ldots, s-1,\ s+1, \ldots, r)$$

according as λ_s defined by the k equations

$$\frac{\partial f}{\partial n_p} + \sum_{i=1}^{s-1} \lambda_i \frac{\partial \phi_i}{\partial n_p} + \lambda_s \frac{\partial \phi_s}{\partial n_p} + \sum_{i=s+1}^{k} \lambda_i \frac{\partial \phi_i}{\partial n_p} = 0 \quad (p = 1, 2, \ldots, k)$$

is positive or negative. Moreover, the minimum or maximum value of ϕ_s is a_s. This means that for a certain cost and some specified variances of any $r-1$ estimates, we are minimising the variance of the r-th estimate.

3.3. *Minimisation of variances:* Another type of requirement may be that relative precisions of the different estimates be in some assigned ratios. As a particular case, it may be desired that the coefficients of variation of the different estimates be all equal, the common value of the coefficients of variation necessarily depending on the cost of the survey. In such a situation the variance of one of the variates will be minimised for fixed cost and for stipulated relations between the variances. As an example of such a situation these estimates may be required by different agencies (like State Governments), and if the precision of an estimate is judged by its variance, we may distribute the total sample size so that the relative precisions of the different estimates are all equal.

If the relative precisions of the estimates are governed by

$$V(\hat{L}_1) = v_2 V(\hat{L}_2) = \ldots = v_r V(\hat{L}_r) \qquad \ldots \quad (3.3.1)$$

and the cost is given by

$$\Sigma\, c_j n_j^g = C \qquad \ldots \quad (3.3.2)$$

the optimum allocation is the solution of the equations

$$n_j^{g+1} = \frac{\sigma_j^2}{\lambda_{r+1}\, g c_j}\left[l_{1j}^2 \sum_{i=1}^{r} \lambda_i - \sum_{i=1}^{r} \lambda_i v_i l_{ij}^2 \right], \quad (j = 1, 2, k;\quad \lambda_1 = 1) \qquad \ldots \quad (3.3.3)$$

along with (3.3.2) and the $r-1$ equations given by (3.3.1).

Like the equations considered in the previous section, it is not possible to offer explicit mathematical solutions of these equations. They have to be solved by approximate methods. In particular, if $r = 2$, an approximate solution can be easily obtained using Newton-Raphson method on the lines indicated before.

4. Stratification after selection

Sometimes it happens that the frame for the entire population is available but frames for individual strata are not known. In such a situation, since we cannot sample from individual strata, a simple random sample of size n is taken from the entire population. When the sample data have been collected the units are assigned to the strata by means of the information obtained about them. The best unbiassed estimates are given by (2.3) as before and the expected variances are

$$E[V(\hat{L}_i)] = E\sum_{j=1}^{k} l_{ij}^2 \frac{\sigma_j^2}{n_j} - \sum_{j=1}^{k} l_{ij}^2 \frac{\sigma_j^2}{N_j} \qquad \dots \quad (4.1)$$

Using Stephan's (1945) result

$$E\left(\frac{1}{n_j}\right) = \frac{1}{nw_j} - \frac{1}{n^2w_j} + \frac{1}{n^2w_j^2} \qquad \dots \quad (4.2)$$

approximately where w_j is the relative size of the j-th stratum, we have

$$E[V(\hat{L}_i)] = \sum_j l_{ij}^2\sigma_j^2 \left\{\frac{1}{nw_j} - \frac{1}{n^2w_j} + \frac{1}{n^2w_j^2}\right\} - \sum_j l_{ij}^2 \frac{\sigma_j^2}{N_j}. \qquad \dots \quad (4.3)$$

5. An example

We now give an example to illustrate the various methods stated above. The object is to estimate the average area under wheat per village and the difference in the averages of the two strata, comprising villages with agricultural area below and above 1500 bighas respectively, for Ghaziabad *tahsil* in Uttar Pradesh (India). Data obtained from a previous census (Sukhatme, 1954) are given below in Table 1. We will assume that the stratum sizes N_i are exact as obtained from the previous census. The rest of the material will be used as supplementary information for improving the design of the current survey.

TABLE 1. RESULTS OF A PRELIMINARY CENSUS

strata	agr. area in bighas	no. of villages N_i	average area under wheat	standard deviation of area under wheat
(1)	(2)	(3)	(4)	(5)
1	0– 500	63	112	56
2	501–1500	199	277	116
3	1501–2500	53	558	186
4	2501 and above	25	960	361

Let the cost function be simply the number of villages and the expected loss be the variance of the estimate. The linear functions to be estimated are

$$L_1 = \frac{1}{N}\Sigma N_j \bar{Y}_j,$$

$$L_2 = \frac{N_1}{N_1+N_2}\bar{Y}_1 + \frac{N_2}{N_1+N_2}\bar{Y}_2 - \frac{N_3}{N_3+N_4}\bar{Y}_3 - \frac{N_4}{N_3+N_4}\bar{Y}_4.$$

(a) For a fixed cost of 34 villages, the optimum allocation minimising the total loss is given in column (2) of Table 2.

TABLE 2. OPTIMUM ALLOCATIONS IN DIFFERENT SITUATIONS

strata	case (a)	Neyman's allocation	case (b)	case (c)
(1)	(2)	(3)	(4)	(5)
1	2	3	3	2
2	7	17	19	10
3	8	7	20	11
4	17	7	18	11

The total loss in this case is 3903. The coefficients of variation of the estimates $\hat{L}_1$ and $\hat{L}_2$ are 8.27 percent and 12.46 percent respectively. If we had used Neyman's allocation based on the estimation of the mean only, the total loss would be 4292 and the individual coefficients of variation would be 6.29 percent and 13.77 percent respectively.

(b) If it be desired that the c.v. for $\hat{L}_1$ be 5 percent and that for $\hat{L}_2$ be 7.5 percent, we have to minimise the cost subject to the conditions

$$V(\hat{L}_1) = 280.77,$$

$$V(\hat{L}_2) = 1136.60.$$

The optimum allocation is given in column (4) of Table 2.

(c) In case it is desired that the c.v. of $\hat{L}_1$ be approximately 2/3 c.v. $(\hat{L}_2)$ and that the cost of the survey be 34 villages we minimise $V(\hat{L}_1)$ subject to the conditions

$$\Sigma n_j = 34,$$

$$V(\hat{L}_1) = .25 V(\hat{L}_2).$$

The allocation is given in column (5) of Table 2. It is found that

$$\text{c.v. } (\hat{L}_1) = 7.11 \text{ percent},$$

$$\text{c.v. } (\hat{L}_2) = 10.77 \text{ percent}.$$

REFERENCES

BASU, D. (1952): On symmetric estimators in point estimation with convex weight functions, *Sankhyā* 12, 45-52.

CHAUNDY, T. (1935): *The Differential Calculus*, Oxford University Press, Oxford.

COCHRAN, W. G. (1953): *Sampling Techniques*, John Wiley and Sons, New York.

KITAGAWA, T. (1956): Some contributions to the design of sample surveys. *Sankhyā*, 14, 317-362.

STEPHAN, F. F. (1945): The expected value and variance of the reciprocal and other negative powers of a positive Bernoullian variate, *Ann. Math. Stat.* 16, 50–61.

SUKHATME, P. V. (1954): Sampling theory of surveys with applications. *Ind. Soc. Agr. Stat.*, New Delhi, 121.

YATES, F. (1949): *Sampling Methods for Censuses and Surveys*, Charles Griffin and Company Ltd.,London.

Paper received: May, 1954.