Some Problems in Stratified Sampling

T.J.Rao
Indian Statistical Institute
Calcutta 700 035
India
tjrao@isical.ac.in

1.**Introduction**

Consider a finite population of size N divided into k strata of sizes $N_i, i = 1, 2, 3 \ldots, k$. Suppose that Y_{ij} is the value of the study variate y on the jth unit of the ith stratum, $j = 1, 2, \ldots, N_i; i = 1, 2 \ldots k$. Assume that we have auxiliary information on a variate x closely related to y taking values X_{ij} on all the units. Optimum allocation of sample size to strata subject to total sample size restriction or cost restriction involves the within statum variances of y-values which are unknown. The justification for substituting the within stratum variances of the auxiliary variate under a well known super population model

$\mathcal{E}(Y_{ij} \mid X_{ij}) = aX_{ij}$, $\mathcal{V}\ (Y_{ij} \mid X_{ij}) = \sigma^2 X_{ij}^g$ and $\mathcal{C}\ (Y_{ij}, Y_{i'j'} \mid X_{ij}, X_{i'j'}) = 0$ has been studied earlier for obtaining near optimum allocations.

2.**Near Optimum Allocations**

Under certain conditions usually met with in practice these near-optimum allocations turn out to be X -proportional allocation for stratified SRS as well as stratified $PPSWR$ designs. Under the model , Bankier's allocation also known as 'power' allocation which keeps the stratum c.v.'s also small, reduces to allocation proportional to X_i^q subject to total sample size restriction and proportional to $X_i^q/\sqrt{c_i}$ subject to cost constraint when C.V.'s of x-values in all strata turn out to be equal.

To compare the optimum allocation vector with the actual integer field allocation vector Pukelsheim (1997) has given certain efficiency bounds which can be extended to any design for which optimum allocation becomes proportional to $\sqrt{V_i \mid c_i}$, where V_i is the within stratum variance and c_i is the sampling cost for the ith stratum.

3. **Double sampling for stratification**

Let the first sample be a simple random sample of size n'. Let W_i be the proportion of population units belonging to the stratum i, viz. N_i/N and w_i be the proportion of the n' units selected, falling in stratum i, $i = 1, 2, \ldots, k$. A stratified random sample of size n is drawn and let y_{ij} be the value of the study variate y on the j th unit of the i th stratum. As an estimate of the population mean $\overline{Y} = \sum_{i=1}^{k} W_i \overline{Y}_i$ consider $\hat{\overline{Y}}_{st} = \sum_{i=1}^{k} w_i \overline{y}_i$ where $\overline{y}_i$ is the sample mean based on the n_i units drawn from n'_i of the ith stratum. $\hat{\overline{Y}}_{st}$ is unbiased for Y and

$$V(\hat{\overline{Y}}_{st}) = S^2(\frac{1}{n'} - \frac{1}{N}) + \sum_{i=1}^{k} \frac{W_i S_i^2}{n'}(\frac{1}{f_i} - 1)$$

where $f_i = \frac{n_i}{n'_i}$ and S^2 and S_i^2 have their usual meaning.

Subject to cost restriction $E(C) = C^* = c'n' + n' \sum c_i f_i W_i$, where c' is the cost of classifying a unit and c_i is the sampling cost for ith stratum, minimizing the expected variance under the model considered earlier, we have optimum f_i given by

$f_i^{opt} = \mathcal{E}S_i[\frac{c'}{c_i}(\mathcal{E}S^2 - \sum W_i \mathcal{E}S_i^2)]^{1/2}$ and $n'_{opt} = \frac{C^*}{c' + \sum c_i f_i W_i}$.

If the $c.v.'s$ of x values in all strata are equal, we get $f_i^{opt} \propto \overline{X}_i[\frac{c'}{c_i} \sum W_i(\overline{X}_i - \overline{X})^2]^{1/2}$.

4.**Non-response**

As a special case , considering only two strata of which one is a response stratum in which data on $n_1 = n'_1$ units is available and the other the 'non-response' stratum where data is obtained from a random subsample of size $n_2 = f_2 n'_2$,one can obtain near optimum $k = k_{opt} = \frac{1}{f_2^{opt}}$. If the related auxiliary information is not sensitive it may be easier to get α^2 and α_2^2 or equivalently $\overline{X}$ and $\overline{X}_2$ in this case under certain assumptions.

RESUME

Nous considérons quelques problèms de la répartition de léchantillon dans les strates.

REFERENCES

Pukelsheim, F. (1977).*Efficient rounding of sampling allocations.*Stat.Prob.Lett. **35** 141-143.