A NOTE ON ESTIMATION OF VARIANCE COMPONENTS IN MULTISTAGE SAMPLING WITH VARYING PROBABILITIES

By J. ROY

Indian Statistical Institute, Calcutta

1. Introduction and Summary

In large scale surveys multistage sampling procedure using different probabilities of selection for different units at any particular stage have been used very often. In the Indian National Sample Survey (NSS), for instance, such a scheme is used for selection of units within a stratum. If in the first stage more than one unit is chosen with replacement, the standard error (of the estimate of the mean or of the total) can be easily computed from the variation between the different estimates that can be computed, one based on each of the chosen first stage unit. However, in order to be able to develop a suitable sampling scheme, we require not merely an estimate of the overall sampling error, but also a knowledge of how the error depends on the adjustable parameters at the disposal of the sampler. Cochran (1939) has shown how in the case of multistage simple sampling this leads to a problem of analysis of the total variation into different stage coponents. The corresponding problem when units are chosen with different probabilities (but with replacement) at each stage is dealt with in this paper. The total variation is split up into different meaningful components depending on the type of sampling used, and unbiassed estimators for these components are derived.

2. The sampling schemes

Here we shall consider a three-stage sampling scheme, but the method used is quite general and can be directly extended for any number of stages. We shall further assume that in every stage sampling is with replacement and that in the third stage units are chosen with equal probabilities. This scheme of sampling (with slight modifications) was used for selection of the ultimate unit (household) within a stratum in the first few rounds of the Indian National Sample Survey where within a stratum a tehsil served as a first-stage unit, villages within the tehsil as second stage units and households within a village as the ultimate third stage unit. Tehsils and villages were chosen with different probabilities, generally proportional to population or area and households within a village were selected with equal probabilities. However, sampling was not in general with replacement except in the case of the first stage units.

The simplified sampling scheme that we shall consider is as follows :

stage	number of units in		selection of sample is with replacement and probabilities
	population	sample	
first	N	n	P_i for the i-th first stage unit
second	N_i	n_i	P_{ij} for the j-th second stage unit in the i-th first stage unit.
third	N_{ij}	n_{ij}	equal

Since the use of separate symbols for the variate value in the sample and in the population unnecessarily complicates the notation, we shall use the same symbol X_{ijk} to denote the variate value for the k-th third stage unit in the j-th second stage unit in the i-th first stage unit. The range of i, j, k will show whether we are referring to the sample or to the population. Furthermore, the dropping of a subscript will indicate a summation over the units in that stage, for instance, X_{ij} will stand for the total of the variate values for all third stage units in the j-th second stage unit of the i-th first stage unit, thus $\sum_{k=1}^{N_{ij}} X_{ijk} = X_{ij}$

Similarly $$\sum_{j=1}^{N_i} X_{ij} = X_i \quad \text{and} \quad \sum_{i=1} X_i = X$$

We shall consider the problem of estimating the grand-total X.

From considerations of symmetry the following is taken as the estimate of X

$$t = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{P_i n_i} \sum_{j=1}^{n_i} \frac{N_{ij}}{P_{ij} n_{ij}} \sum_{k=1}^{n_{ij}} X_{ijk}$$

where, of course, X_{ijk} is the variate value for the k-th third stage unit in the j-th second-stage unit in the i-th first-stage unit in the selected sample. Obviously t is unbiassed

$$E(t) = X$$

and a little computation shows that its variance is

$$V(t) = \frac{1}{n} \left\{ \sum_{i=1}^{N} \frac{\theta_i^2 + \sigma_i^2}{P_i n_i} + \sigma^2 \right\}$$

where $$\sigma_{ij}^2 = \frac{1}{N_{ij}} \left\{ \sum_{k=1}^{N_{ij}} X_{ijk}^2 - \frac{X_{ij}^2}{N_{ij}} \right\}$$

$$\theta_i^2 = \sum_{j=1}^{N_i} \frac{N_{ij}^2}{P_{ij}} \frac{\sigma_{ij}^2}{n_{ij}}$$

$$\sigma_i^2 = \sum_{j=1}^{N_i} \frac{X_{ij}^2}{P_{ij}} - X_i^2$$

$$\sigma^2 = \sum_{i=1}^{N} \frac{X_i^2}{P_i} - X^2$$

From the practical point of view, however, the numbers n_i and n_{ij} are not generally defined separately for each of the first and second stages. There are three different ways of fixing up the values of the n_i's and n_{ij}'s. One we may call "equal sampling" at both stages where the same number m of second stage units are selected from each first stage unit and the same number l of third stage units are selected from within each selected second stage unit. The other method may be called "proportionate sampling" at both stages where a fixed proportion of second stage units are selected within each first stage unit and a different fixed proportion of third stage units within each selected second stage unit are sampled. Variants of these two methods, with equal sampling at one stage and proportionate sampling at another stage are also used. A third method is to determine the values of n_i and n_{ij} in such a way that the estimate t comes out as proportional to the total of all variate values in the sample : this is known as "self-weighted" sampling.

Here we shall discuss two of these special types of sampling. The first is equal sampl-sampling at the second stage and proportionate sampling at the third. The other is self-weighted sampling with equal sampling at the second stage.

For equal sampling at the second stage and proportionate sampling at the third we have

$$n_i = m$$

$$n_{ij} = lN_{ij}$$

say. In this case,

$$t = \frac{1}{lmn} \sum_{i=1}^{n} \frac{1}{P_i} \sum_{j=1}^{m} \frac{1}{P_{ij}} \sum_{k=1}^{n_{ij}} X_{ijk}$$

and the variance reduces to

$$V(t) = \frac{A_1}{n} + \frac{B_1}{mn} + \frac{C_1}{lmn}$$

where A_1, B_1, C_1 are independent of l, m, n and given by

$$A_1 = \sigma^2$$

$$B_1 = \sum_{i=1}^{N} \frac{\sigma_i^2}{P_i}$$

$$C_1 = \sum_{i=1}^{N} \frac{1}{P_i} \sum_{j=1}^{K_i} \frac{N_{ij} \sigma_{ij}^2}{P_{ij}}$$

We shall refer to this as the scheme I of sampling.

In the second scheme of sampling (Scheme II), we have

$$n_i = m$$

$$n_{ij} = \frac{l N_{ij}}{P_i P_{ij}}$$

so that the estimate comes out as

$$t = \frac{1}{lmn} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{n_{ij}} X_{ijk}$$

with variance given by

$$V(t) = \frac{A_2}{n} + \frac{B_2}{mn} + \frac{C_2}{lmn}$$

where A_2, B_2, C_2 are constants independent of l, m, n given by

$$A_2 = \sigma^2$$

$$B_2 = \sum_{i=1}^{N} \frac{\sigma_i^2}{P_i}$$

$$C_2 = \sum_{i=1}^{N} \sum_{j=1}^{N_i} N_{ij}\, \sigma_{ij}^2$$

so that $A_1 = A_2$, $B_1 = B_2$ but $C_1 \neq C_2$. We shall write A for A_1 or A_2 and B for B_1 or B_2.

3. The stage components of variance and their estimates

We thus see that in either of the two schemes of sampling under consideration the variance of the estimate depends on only three parameters A, B, C_i independent of the adjustable constants l, m, n. The cost of the survey naturally depends on the values of l, m, n. Therefore if estimates of the parameters A, B, and C_i are available, the information may be of use in planning an optimum survey at a fixed level of cost.

Let us now examine the nature of the three parameters. If it were possible to determine without further sampling the value X_i for a first stage unit selected with probabilities P_i, $\frac{X_i}{P_i}$ would provide an unbiassed estimate of X. The parameter A simply measures the variance of such an estimate, that is A is the variance of the (hypothetical) estimate of X from complete enumeration of a single first stage unit chosen with the probabilities P_i. We may thus look upon A as the "between first stage" variance. Similarly σ_i^2 gives the variance of the (hypothetical) estimate of X_i obtained by completely enumerating a second stage unit drawn with probabilities P_i. Thus σ_{ij}^2 measures variation between second stage units within the i-th first stage unit. Therefore, if the N first stage units were regarded as strata and if from the i-th stratum νP_i second stage units were chosen with probabilities P_{ij} and each completely enumerated, the variance of the estimate of X would be simly B/ν The parameter C_i may be similarly interpreted this way.

We now take up the problem of estimating the parameters A, B, C_i. We shall simply obtain unbiassed quadratic estimators for these parameters. Certain optimum properties of these estimators may be demonstrated from considerations of symmetry but we shall not enter here into a discussion of that type. For problems of estimation we shall consider unrestricted values of n_i's and n_{ij}'s.

Let us write

$$s_{ij}^2 = \frac{1}{n_{ij}-1}\left(\sum_{k=1}^{n_{ij}} X_{ijk}^2 - \frac{x_{ij}^2}{n_{ij}} \right) \text{ where } x_{ij} = \sum_{k=1}^{n_{ij}} X_{ijk}$$

Then for fixed first and second stage units

$$E\, s_{ij}^2 = \sigma_{ij}^2$$

Hence, if we write

$$c_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{P_i^2 n_i} \sum_{j=1}^{n_i} \frac{N_{ij} s_{ij}^2}{P_{ij}^2}$$

$$c_2 = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{P_i n_i} \sum_{j=1}^{n_i} \frac{N_{ij} s_{ij}^2}{P_{ij}}$$

then c_1 and c_2 provide unbiassed estimators for C_1 and C_2 respectively.

Now construct

$$y_{ij} = \frac{1}{P_{ij}} \; \frac{N_{ij}}{n_{ij}} \; x_{ij}$$

Let

$$s_i^2 = \frac{1}{n_i - 1}\left(\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_i^2}{n_i} \right) \text{ where } y_i = \sum_{j=1}^{n_i} y_{ij}$$

For fixed first stags units $E(s_i^2) = V(y_{ij})$.

If first and second stage units are fixed

$$E\, y_{ij} = \frac{X_{ij}}{P_{ij}} \qquad V(y_{ij}) = \frac{1}{P_{ij}^2} \; \frac{N_{ij}^2}{n_{ij}} \; \sigma_{ij}^2$$

and therefore when second stage units are allowed to vary, that is when only first stage units are fixed

$$V(y_{ij}) = V\left(\frac{X_{ij}}{P_{ij}} \right) + E\left(\frac{1}{P_{ij}^2} \; \frac{N_{ij}^2}{n_{ij}} \; \sigma_{ij}^2 \right)$$

$$= \sigma_i^2 + E\left(\frac{1}{P_{ij}^2} \; \frac{N_{ij}^2}{n_{ij}} \; \sigma_{ij}^2 \right)$$

Therefore if we write

$$w_i^2 = s_i^2 - \frac{1}{n_i}\sum_{j=1}^{n_i} \frac{1}{P_{ij}^2} \; \frac{N_{ij}^2}{n_{ij}} \; s_{ij}^2$$

for fixed first state unit $E\, w_i^2 = \sigma_i^2$ and consequently

$$b = \frac{1}{n}\sum_{i=1}^{n} \frac{w_i^2}{P_i^2}$$

provides an unbiassed estimate for B.

Finally to estimate A construct

$$z_i = \frac{1}{P_i n_i}\, y_i$$

Then, for fixed first stage unit,

$$E(z_i) = \frac{X_i}{P_i} \qquad V(z_i) = \frac{1}{P_i^2 n_i} V(y_{ij})$$

and therefore for fixed first stage unit

$$E\left(\frac{s_i^2}{P_i^2 n_i}\right) = V(z_i)$$

Consequently for unrestricted variations

$$V(z_i) = V\left(\frac{X_i}{P_i}\right) + E\{V(z_i)\}$$

$$= \sigma^2 + E\{V(z_i)\}.$$

Therefore, if we write

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} z_i^2 - \frac{z^2}{n}\right) \text{ where } z = \sum_{i=1}^{n} z_i$$

we have

$$a = s^2 - \frac{1}{n}\sum_{i=1}^{n} \frac{s_i^2}{P_i^2 n_i}$$

for an unbiassed estimate of A.

We may note in this connection, the well known result that since

$$t = \frac{1}{n}\sum_{i=1}^{n} z_i$$

an unbiassed estimate of its variance is given by $\frac{s^2}{n}$ but if we are interested in the separate components of the variance, we have to compute a, b, c_i separately. One disadvantage of the estimates a, b, c_i is that sometimes these may turn out to be negative.

REFERENCES

COCHRAN, W. G. (1939): The use of the analysis of variance in enumeration by sampling. *Amer. Stat. Ass.*, 34, 492-510.

LAHIRI, D. B. (1954): Technical paper on some aspects of the development of the sample design. *The National Sample Survey* No. 5, Department of Economic Affairs, Ministry of Finance, Government of India.

Paper received : April, 1956.