韓國數學敎育學會誌<數學敎育>
1984. 3. Vol. XXII, No. 2.

Journal of the Korea Society of Mathematical Education
Mar. 1984. Vol. XXII, No. 2.

A Study on Stratified Sampling Variance of Double Sampling

by Jee Euen Sook
Kwangwoon University, Seoul, Korea

A number of sampling techniques depend on the possession of advance information about an auxiliary variate x_i. Ratio and regression estimates require a knowledge of the population mean $\bar{X}$. If it is desired to stratify the population according to the values of the x_i, their frequency distribution must be known.

When such information is lacking, it is sometimes relatively cheap to take a large preliminary sample in which x_i alone is measured. The purpose of this sample is to furnish a good estimate of $\bar{X}$ or of the frequency distribution of x_i. In a survey whose function is to make estimates for some other variate y_i, it may pay to devote part of the resources to this preliminary sample, although this means that the size of the sample in the main survey on y_i must be decreased. This technique is known as double sampling or two-phase sampling. As the discussion implies, the technique is profitable only if the gain in precision from ratio or regression estimates or stratification more than offsets the loss in precision due to the reduction in the size of the main sample.

The population is to be stratified into L classes (strata). The first sample is a simple random sample of size n'.

Let

$W_h = N_h/N$=proportion of population falling in stratum h

$w_h = n_h'/n'$=proportion of first sample falling in stratum h

Then w_h is an unbiased estimate of W_h.

The second sample is a stratified random sample of size n in which the y_{hi} are measured: n_h units are drawn from stratum h. Usually the second sample in stratum h is a random subsample from the n_h' in the stratum. The objective of the first sample is to estimate the strata weights; that of the second sample is to estimate the strata means $\bar{Y}_h$.

The population mean $\bar{Y} = W_h\bar{Y}_h$. As an estimate we use

$$\bar{y}_{st} = \sum_{h=1}^{L} w_h \bar{y}_h$$

The problem is to choose n' and the n_h to minimize $V(\bar{y}_{st})$ for given cost.

We must then verify whether the minimum variance is smaller than can be attained by a single simple random sample in which y_i alone is measured. In presenting the theory, we assume that the n_h are a random subsample of the n_h'. Thus, $n_h = v_h n_h'$, where $0 < v_h \leq 1$ and the v_h are chosen in advance. Repeated sampling implies a fresh drawing of both the first and the second samples, so that the w_h, n_h and $\bar{y}_h$ are all random variables. The problem is therefore one of stratification in which the strata sizes are not known exactly.

Two approximations will be made for simplicity. The first sample size n' is assumed large enough so that every $w_h > 0$. Second, when we come to discuss optimum strategy, every optimum v_h as found by the formula is assumed $\leqslant 1$.

Theorem. *If the first sample is random and of size n', the second sample is a random subsample of the first, of size $n_h = v_h n_h'$ where $0 \leqslant v_h \leqslant 1$ and the v_h are fixed,*

$$V(\bar{y}_{st}) = s^2\left(\frac{1}{n'} - \frac{1}{N}\right) + \sum_h^L \frac{W_h S_h^2}{n'}\left(\frac{1}{V_h} - 1\right) \tag{1}$$

where S^2 is the population variance.

Proof. The proof is easily obtained by the following device. Suppose that the y_{hi} were measured on all n_h' first-sample units in stratum h, not just on the random subsample of n_h. Then, since $w_h = n_h'/n'$,

$$\sum_h^L w_h \bar{y}_h' = \bar{y}'$$

is the mean of a simple random of size n' from the population. Hence, averaging over repeated selections of sample of size n',

$$V\left(\sum_h^L w_h \bar{y}_h'\right) = S^2\left(\frac{1}{n'} - \frac{1}{N}\right) \tag{2}$$

But

$$\bar{y}_{st} = \sum_h^L w_h \bar{y}_h = \sum_h^L w_h \bar{y}_h' + \sum_h^L w_h (\bar{y}_h - \bar{y}_h') \tag{3}$$

Let the subscript 2 refer to an average over all random subsamples of n_h units that can be drawn from a given n_h' units. Clearly, $E_2(\bar{y}_h) = \bar{y}_h'$. Results that follow immediately are:

$$COV\ [\bar{y}_h',\ (\bar{y}_h - \bar{y}_h')] = 0:$$
$$COV\ (\bar{y}_h', \bar{y}_h) = V(\bar{y}_h'):\ V(\bar{y}_h - \bar{y}_h') = V(\bar{y}_h) - V(\bar{y}') \tag{4}$$

Hence, for fixed w_h,

$$V_2[\sum w_h (\bar{y}_h - \bar{y}_h')] = \sum w_h^2 s_h^2\left(\frac{1}{n_h} - \frac{1}{n_h'}\right) = \sum \frac{w_h S_h^2}{n'}\left(\frac{1}{v_h} - 1\right) \tag{5}$$

since $n_h = v_h n_h' = v_h w_h n'$.

Averaging over the distribution of the w_h obtained by repeated selections of the first sample, we have, from (2), (3) and (4),

$$V(\bar{y}_{st}) = S^2\left(\frac{1}{n'} - \frac{1}{N}\right) + \sum_h^L \frac{w_h s_h^2}{n'}\left(\frac{1}{v_h} - 1\right) \tag{6}$$

Papers by Robson (1952) and Robson and King (1953) extend the stratification theory to two-stage sampling, applying it to the estimation of magazine readership.

References

1. Cochran W.G. (1977). *Sampling Techniques.* John Wiley & Sons, New York, third edition.
2. Neyman J. (1938). Contribution to the theory of sampling human populations. *Jour. Amer. Stat. Assoc.*, *33*, 101-116.

3. Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, ***60***, 125-133.
4. Robson, D.S. (1952). Multiple sampling of attributes. *Jour. Amer. Stat. Assoc.*, ***47***, 203-215.
5. Robson, D.S., and King, A.J. (1953). Double sampling and the Curtis impact survey. Cornell Univ. *Agr. Exp. Sta. Mem.*, ***231***.