ON DOUBLE SAMPLING FOR PPS ESTIMATION

By Des Raj

Statistical Service of Greece, Athens

1. Introduction. Usually the double sampling technique is used for purposes of stratification or for ratio or regression estimation. There may, however, be situations in which it is considered important to select the sample units with probability proportionate to some measure of size x, information on which is not readily available but could be collected at moderate cost for a fairly large sample. For example, in sampling a list of agricultural holders it may be considered very desirable to select holdings with probability proportionate to area. In this case one may select a fairly large random sample of holdings from the list, obtain information on areas by interviewing farmers, select a small sample of holdings with probability proportionate to area and collect information on the desired characteristics from this small sample. The units selected may be subsampled if they are large. The purpose of this paper is to develop theory for this technique when the second sample is a subsample of the first (Case 1) or it is selected independently (Case 2). For related work, references at the end may be consulted.

2. Theory for Case 1. Using the notation from Cochran (1963), an unbiased estimate of the population total is given by

$$\hat{Y} = (N/n')(x'/n) \sum_{1}^{n} (y_i/x_i) \tag{1}$$

where x' is the total of the variate x for the initial sample and $\sum$ denotes summation over the second sample. The initial sample of size n' is selected with equal probabilities without replacement and the second sample is selected with pps with replacement. Using theorems on conditional expectations and variances (Raj (1956) p. 271) we have

$$\begin{aligned} V_1E_2(\hat{Y}) &= N^2(n'^{-1} - N^{-1})(N-1)^{-1}\sum_{1}^{N}(y_i - \bar{Y})^2 \\ &= N(N-n')n'^{-1}S_y^2 \end{aligned} \tag{2}$$

$$\begin{aligned} E_1V_2(\hat{Y}) &= \frac{N}{N-1}\,\frac{n'-1}{n'}\,\frac{1}{n}\left(X\sum_{1}^{N}\frac{y_i^2}{x_i} - Y^2\right) \\ &= \frac{N}{N-1}\,\frac{n'-1}{n'}\,V_{\text{pps}} \end{aligned} \tag{3}$$

so that the variance of $\hat{Y}$ is given by

$$V(\hat{Y}) = N(N-1)^{-1}(n'-1)n'^{-1}V_{\text{pps}} + N(N-n')n'^{-1}S_y^2 \tag{4}$$

Received 30 July 1963.

where V_{pps} is the variance based on a pps sample of size n when X is known. This variance may be compared with the large sample approximations to the variances in the case of double sampling for ratio estimation and regression estimation, which are (Cochran (1963) pp. 340):

(5) ratio: $$(1 - n/n')V_{\text{rat}} + (N^2/n')S_y^2$$

(6) regression: $$(1 - n/n')V_{\text{reg}} + (N^2/n')S_y^2$$

where the definitions of V_{rat} and V_{reg} are similar to V_{pps}. Thus, provided the subsampling fractions are small, double sampling for pps estimation will be more precise than double sampling for ratio or regression estimation according as the pps estimate in single-phase sampling is better than the ratio or regression estimates. Some comparisons in single-phase sampling may be found in Raj (1958) and Zarkovich (1960).

Using procedures similar to Raj (1954), an unbiased estimate of the variance of $\hat{Y}$ may be obtained as

$$\begin{aligned} \hat{V}(\hat{Y}) = {} & \frac{N^2}{n'^2}\frac{x'^2}{n(n-1)}\left[\sum_1^n \frac{y_i^2}{x_i^2} - \frac{1}{n}\left(\sum_1^n \frac{y_i}{x_i}\right)^2\right] \\ & + \frac{N(N-n')}{nn'(n'-1)}\left[x'\sum_1^n \frac{y_i^2}{x_i} - \frac{x'^2}{n'(n-1)}\left\{\left(\sum_1^n \frac{y_i}{x_i}\right)^2 - \sum_1^n \frac{y_i^2}{x_i^2}\right\}\right]. \end{aligned} \tag{7}$$

In case the sample units are large and are subsampled, let T_i be an unbiased estimate of the unit total y_i based on subsampling at the second and subsequent stages. The variance of the estimator

$$(N/n')(x'/n)\sum (T_i/x_i) \tag{8}$$

in multistage sampling can then be shown to be the sum of two parts. The first part is given by (4) and the second part is

$$N(N-1)^{-1}(nn')^{-1}[(N-n')\sum V(T_i) + (n'-1)\,X\sum V(T_i)/x_i]. \tag{9}$$

3. Theory for Case 2. In this case the first sample is used solely for estimating X. An independent sample of size n is selected with pps using Lahiri's method (1951) in which it is assumed that an upper bound for x's is known in advance although X is unknown. The estimate of the population total is of the same form as (1) with a variance of

$$[1 + (n'^{-1} - N^{-1})S_x^2/\bar{X}^2]V_{\text{pps}} + (n'^{-1} - N^{-1})N^2R^2S_x^2. \tag{10}$$

The large sample variance of the associated estimator in double sampling for ratio estimation is given by Cochran (1963)

$$V_{\text{rat}} + n'^{-1}N^2R^2S_x^2. \tag{11}$$

Since the square of the coefficient of variation of x, when divided by the size of the initial sample, will generally be very small relative to unity, the issue whether one estimator is better than the other will largely depend on their relative performance in single-phase sampling. An unbiased estimator of the variance in Case 2 may be given as

$$\begin{aligned}(12)\qquad & \frac{N^2}{n(n-1)n'}\left[\sum_1^n \frac{y_i^2}{x_i^2} - \frac{1}{n}\left(\sum_1^n \frac{y_i}{x_i}\right)^2\right]\left[\frac{1}{n'}x'^2 - \left(1 - \frac{n'}{N}\right)s_x^2\right] \\ & \qquad + \frac{N(N-n')}{n'}\left(\frac{1}{n}\sum_1^n \frac{y_i}{x_i}\right)^2 s_x^2\end{aligned}$$

where s_x^2 is the variance of x in the first sample with the divisor $n' - 1$. In case of multistage sampling, the additional contribution to the variance due to subsampling is

$$(11)\qquad [N^2(n'^{-1} - N^{-1})S_x^2 + X^2](1/nX) \sum V(T_i)/x_i .$$

REFERENCES

[1] Cochran, W. G. (1963). *Sampling Techniques.* Wiley, New York.

[2] Lahiri, D. B. (1951). A method for sample selection providing unbiased ratio estimates. *Int. Statist. Inst. Bull.* **33** Part 2 133–140.

[3] Raj, D. (1954). Ratio estimation in sampling with equal and unequal probabilities. *J. Indian Soc. Agric. Statist.* **6** 127–138.

[4] Raj, D. (1956). Some estimators in sampling with varying probabilities without replacement. *J. Amer. Statist. Assoc.* **51** 269–284.

[5] Raj, D. (1958). On the relative accuracy of some sampling techniques. *J. Amer. Statist. Assoc.* **53** 98–101.

[6] Sukhatme, B. V. (1962). Some ratio-type estimators in two-phase sampling. *J. Amer. Statist. Assoc.* **57** 628–632.

[7] Zarkovich, S. S. (1960). On the efficiency of sampling with varying probabilities and the selection of units with replacement. *Metrika* **3** 53–60.