ON SAMPLING WITH PROBABILITY PROPORTIONATE TO AGGREGATE SIZE

By Des Raj

National Statistical Service of Greece

1. Introduction

The theory underlying the sampling method in which the sample is selected with probability proportionate to its aggregate measure of size (*ppas*) has been given by Midzuno (1950), Lahiri (1951), Raj (1954) and Sen (1955). This note presents some results relating to the precision of this technique. The discussion is restricted to the situation in which the two units (clusters) selected from a stratum are completely enumerated.

2. Results

Let a stratum contain N units with total measure of size X and let $Y = RX$ be the stratum total to be estimated from a sample of two units. According to the sampling scheme of this paper, the probability that the units u_i and u_j form the sample is proportionate to $x_i + x_j$. As an unbiased estimator of the stratum total we take

$$\hat{Y} = \frac{X(y_i + y_j)}{(x_i + x_j)} \tag{1}$$

and its variance is given by

$$V(\hat{Y}) = \frac{X}{N-1} \sum{}' (y_i + y_j)\left(\frac{y_i + y_j}{x_i + x_j} - R\right) = V_{ppas} \tag{2}$$

where Σ' denotes summation over all different pairs of units in the stratum. It will be noted that the contribution of a number of pairs to V_{ppas} would be negative. This is in marked contrast to *pps* sampling with replacement for which the variance is

$$V_{ppx} = \frac{1}{2} \sum{}' x_i x_j \left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2 \tag{3}$$

and all pairs make a positive contribution to the variance. In order to discuss the relative precision of *ppas* sampling, we shall assume that the N units in the finite population are divided up into a number of arrays with x_i being the (imputed) measure of size for all the N units in the i-th array. Further, let

$$y_{im} = Rx_i + e_{im} \tag{4}$$

where $\sum_m e_{im} = 0$, $\sum_m e_{im}^2 = aN_i x_i^g$. Thus, for a given x, the residuals are assumed to have a mean value of zero and a variance proportionate to $x^g (g > 0)$. Under this model, we have

$$\begin{aligned} V_{ppas} &= \frac{X}{N-1} \sum{}' \frac{(e_{im} + e_{jm'})^2}{(x_i + x_j)} \\ &= \frac{aX}{2(N-1)} \left[2 \sum_i \sum_{j>i} N_i N_j \frac{x_i^g + x_j^g}{x_i + x_j} \right. \\ &\qquad \left. + \sum_i N_i (N_i - 2)\, x_i^{g-1} \right], \end{aligned} \tag{5}$$

$$\begin{aligned} V_{pps} &= \frac{aX}{2} \sum_i N_i\, x_i^{g-1} \\ &= \frac{aX}{2(N-1)} \left[\sum_i \sum_{j>i} N_i N_j \,(x_i^{g-1} + x_j^{g-1}) \right. \\ &\qquad \left. + \sum_i N_i (N_i - 1)\, x_i^{g-1} \right]. \end{aligned} \tag{6}$$

Thus the quantity $\triangle = V_{pps} - V_{ppas}$ is given by

$$\begin{aligned} \triangle = \frac{aX}{2(N-1)} &\left[\sum_i N_i x_i^{g-1} - \sum_i \sum_{j>i} N_i N_j \right. \\ &\left. \times \frac{(x_i - x_j)(x_i^{g-1} - x_j^{g-1})}{x_i + x_j} \right]. \end{aligned} \tag{7}$$

For $g = 0$ and 1, it is obvious that $\triangle > 0$. But, in the case of $g = 1$, the relative increase in variance (with *pps* as the base) is as small as $1/(N-1)$. When $g = 2$, *pps* sampling will be found to be superior whenever the following inequality holds:

$$\sum_i \sum_{j>i} N_i N_j \frac{(x_i - x_j)^2}{x_i + x_j} > X. \tag{8}$$

For $g = 3$, we have

$$\triangle = \frac{aX}{2} \frac{N}{N-1} [\bar{X}^2 - (N-1)\, V(x)]. \tag{9}$$

This *pps* sampling will give a lower variance than *ppas* sampling if $Cv(x) > 1/\sqrt{N-1}$, where $Cv(x)$ stands for the coefficient of variation of x. This shows that for most populations met with in practice (Hansen *et al.*, 1953) *pps* sampling will produce the more precise estimate when $g = 3$. It may be mentioned that it is already known (Raj, 1958) that the range 0 to 1 of g is not favourable to *pps* sampling. In order to make a comparison with equal probability *ep*) sampling, we have for this model

$$V_{ep} = \frac{N(N-2)}{2} \left[R^2 S_x^{\,2} + \frac{a}{N-1} \sum_i N_i x_i^{\,g} \right]. \tag{10}$$

Hence, for $g = 1$, *ppas* sampling is decidedly superior to equal probability sampling and is only slightly better than *pps* sampling.

3. References

1. Hansen, M. H., *et al.* .. *Sample Survey Methods and Theory*, John Wiley & Sons, New York, 1953.

2. Lahiri, D. B. .. "A method of sample selection providing unbiased ratio estimates," *Bull. Int. Stat. Inst.*, 1951, **33**, 133–40.

3. Midzuno, H. .. "An outline of the theory of sampling systems," *Ann. Inst. Stat. Math.*, 1950, **1**, 149–56.

4. Raj, D. .. "Ratio estimation in sampling with equal and unequal probabilities," *Jour. Ind. Soc. Agr. Stat.*, 1954, **6**, 127–38.

5. ———— .. "On the relative accuracy of some sampling techniques," *Jour. Amer. Stat. Assn.*, 1958, **53**, 98–101.

6. Sen, A. R. .. "On the selection of n primary sampling units from a stratum structure ($n \geq 2$)," *Ann. Math. Stats.*, 1955, **26**, 744–51.