ON SAMPLING OVER TWO OCCASIONS WITH PROBABILITY PROPORTIONATE TO SIZE

By Des Raj

National Statistical Service of Greece, Athens

1. Introduction. The question of replacement of the sample for making current estimates has been considered among others by Jessen (1942), Patterson (1950) and Cochran (1963) under simple random sampling of the units. In some modern sample surveys, however, it is the sample area selected with probability proportionate to size (pps) that has become the object of rotation over successive occasions. Examples of these areas or clusters are the segment selected with probability proportionate to the estimated number of dwelling units in the U. S. Current Population Survey (Eckler, 1961) and the census block selected with probability proportionate to population (with replacement) in the National Sample Survey of urban India (Lahiri, 1954). In view of this it is of some interest to see how the theory would look if the sample units to be replaced are clusters selected with pps. The purpose of this paper is to present outlines of this theory when sampling is confined to two occasions only and current estimates are of chief interest. Applications of this theory to the double sampling technique are also included.

2. Theory. On the first occasion a sample A_1 of n clusters is selected with replacement with probabilities

$$p_i, \sum_1^N p_i = 1 \tag{1}$$

based on a character z. For estimating the population total X for the character x we have

$$\begin{aligned} \hat{X} &= (1/n)\sum_1^n (x_i/p_i), \\ V(\hat{X}) &= (1/n)\sum_1^N p_i(x_i/p_i - X)^2 = (1/n)V_{\text{pps}}(x). \end{aligned} \tag{2}$$

On the second occasion a simple random sample of $m = n\lambda$ clusters is selected without replacement from A_1 (the matched part) and an independent sample of $u = n\,\mu = n - m$ clusters is selected in the same manner as A_1 (the unmatched part). Based on the matched part an unbiased estimate of the population total Y for the character y will be given by

$$\hat{Y}_m = [m^{-1}\sum_1^m (y_i/p_i) - m^{-1}\sum_1^m (x_i/p_i)] + (1/n)\sum_1^n (x_i/p_i). \tag{3}$$

To find the variance of $\hat{Y}_m$ we keep A_1 fixed in the first instance and then vary A_1. Using theorems on conditional expectations and variances (Raj, 1956) we have

$$V_1E_2(\hat{Y}_m) = (1/n)V_{\text{pps}}(y), \tag{4}$$

Received 24 February 1964; revised 22 July 1964.

$$V_2(\hat{Y}_m) = (1/m - 1/n)(n-1)^{-1} \cdot \textstyle\sum_1^n [(y_i - x_i)/p_i - n^{-1}\sum_1^n (y_i - x_i)/p_i], \tag{5}$$

$$E_1V_2(\hat{Y}_m) = (1/m - 1/n)[V_{\text{pps}}(y) + V_{\text{pps}}(x) - 2\delta\sigma_{\text{pps}}(y)\sigma_{\text{pps}}(x)], \tag{6}$$

where

$$\delta = \textstyle\sum_1^N p_i(y_i/p_i - Y)(x_i/p_i - X)/ [\sum p_i(y_i/p_i - Y)^2 \sum p_i(x_i/p_i - X)^2]^{\frac{1}{2}} \tag{7}$$

is the correlation coefficient between y_i/p_i and x_i/p_i. Hence the variance of $\hat{Y}_m$ is given by

$$V(\hat{Y}_m) = (1/n)V_{\text{pps}}[1/\lambda + (1-\lambda)(1-2\delta)/\lambda] \tag{8}$$

on the assumption that $V_{\text{pps}}(y) = V_{\text{pps}}(x) = V_{\text{pps}}$ when the same character is under study at the two occasions. The unmatched part will give an estimate of Y as

$$\hat{Y}_u = (n\mu)^{-1}\textstyle\sum_1^u (y_i/p_i) \tag{9}$$

and

$$V(\hat{Y}_u) = (n\mu)^{-1}V_{\text{pps}}(y) = (n\mu)^{-1}V_{\text{pps}}\,.$$

Using weights W_m and W_u proportional to inverses of the variances, we have

$$\hat{Y} = W_m\hat{Y}_m + W_u\hat{Y}_u\,, \tag{10}$$

$$V(\hat{Y}) = (1/n)V_{\text{pps}}[1 + (1-2\delta)\mu][1 + (1-2\delta)\mu^2]^{-1}. \tag{11}$$

By differentiating $V(\hat{Y})$ with respect to μ and equating it to zero, the optimum values of λ and μ turn out to be

$$\mu = [1 + 2^{\frac{1}{2}}(1-\delta)^{\frac{1}{2}}]^{-1}, \qquad \lambda = 2^{\frac{1}{2}}(1-\delta)^{\frac{1}{2}}[1 + 2^{\frac{1}{2}}(1-\delta)^{\frac{1}{2}}]^{-1} \tag{12}$$

and the minimum variance achieved this way is

$$V_{\min}(\hat{Y}) = (1/n)V_{\text{pps}}[\tfrac{1}{2} + (1-\delta)^{\frac{1}{2}}/2^{\frac{1}{2}}]. \tag{13}$$

The factor inside the brackets would be less than unity for $\delta > 0.5$. And $n^{-1}V_{\text{pps}}$

TABLE 1

Optimum percent matched and relative gain in precision compared with no matching

δ	Optimum % matched	% gain in precision
0.50	50	0
0.60	47	6
0.70	44	13
0.80	39	23
0.90	31	38
0.95	24	52

represents the variance on the second occasion when an independent sample of n clusters is selected like A_1. Hence, for $\delta > 0.5$, a partially replaced sample will give higher precision than a completely unmatched sample of the same size. Table 1 gives the optimum percentage to match and the relative gain in precision compared with no matching for different values of the correlation coefficient δ.

In order to make an unbiased estimate of the variance of $\hat{Y}$, we make separate estimates of $V(\hat{Y}_u)$ and $V(\hat{Y}_m) = E_1V_2(\hat{Y}_m) + V_1E_2(\hat{Y}_m)$. This can be done by observing that

$$E_1V_2(\hat{Y}_m) \cong (m^{-1} - n^{-1})(m - 1)^{-1}\textstyle\sum_1^m [(y_i - x_i)/p_i - m^{-1}\sum_1^m (y_i - x_i)/p_i]^2,$$

$$V_1E_2(\hat{Y}_m) \cong n^{-1}(m - 1)^{-1}\textstyle\sum_1^m (y_i/p_i - m^{-1}\sum_1^m y_i/p_i)^2,$$

$$V(\hat{Y}_u) \cong u^{-1}(u - 1)^{-1}\textstyle\sum_1^u (y_i/p_i - u^{-1}\sum_1^u y_i/p_i)^2,$$

where $\cong$ stands for "is estimated unbiasedly by".

It may be noted that the estimator $\hat{Y}$ (formula 10) depends on the correlation coefficient δ which, in general, will not be known. It will therefore be necessary to replace δ by some past value. This will retain the unbiased character of the estimator $\hat{Y}$ although the variance will increase.

3. Application to double sampling. Now the problem is to estimate the population total Y by taking an initial sample of size n' in which information is collected inexpensively on a correlated character x and a subsequent small sample of size n in which both y and x are measured. The second sample would ordinarily be a subsample of the first. But it could be independent too when, for instance, it is found at the time of analysis that information on x is available with a different agency on a much larger sample and it is proposed to use that for achieving greater precision. The n' units in the first sample are selected like A_1 of Section 2. Let the second sample be a subsample of A_1 selected with equal probabilities without replacement. Let k stand for a reasonably good guess, made on the basis of previous knowledge, of the ratio of y to x in the population. Then

$$\hat{Y} = (1/n)\textstyle\sum_1^n y_i/p_i - (k/n)\sum_1^n x_i/p_i + (k/n')\sum_1^{n'} x_i/p_i, \tag{14}$$

$$V(\hat{Y}) = (1/n)V_{\text{pps}}(y) + (n^{-1} - n'^{-1})[k^2V_{\text{pps}}(x) - 2k\delta\sigma_{\text{pps}}(x)\sigma_{\text{pps}}(y)], \tag{15}$$

$$\begin{aligned}\hat{V}(\hat{Y}) = {} & [n'(n - 1)]^{-1}\textstyle\sum_1^n (y_i/p_i - n^{-1}\sum y_i/p_i)^2 \\ & + (n^{-1} - n'^{-1})(n - 1)^{-1}\textstyle\sum_1^n (d_i/p_i - n^{-1}\sum d_i/p_i)^2,\end{aligned} \tag{16}$$

where

$$d_i = y_i - kx_i.$$

It is of interest to make a comparison with the technique of double sampling for ratio estimation when the units are selected with equal probabilities. In this case the variance is given by (Cochran, 1963)

$$V_{\text{rat}}(\hat{Y}) = n^{-1}V_{\text{ran}}(y) + [n^{-1} - (n')^{-1}][R^2V_{\text{ran}}(x) - 2R\rho\sigma_{\text{ran}}(y)\sigma_{\text{ran}}(x)], \tag{17}$$

where 'ran' refers to random sampling (with equal probabilities). Provided k is near to R, the superiority of the present technique to the method of double sampling for ratio estimation would seem to lie in the selection of clusters with unequal probabilities.

In case the two samples are independent, the form of the population total estimate is the same but the variance becomes

$$V(\hat{Y}) = (1/n)[V_{\text{pps}}(y) + k^2V_{\text{pps}}(x) - 2k\delta\sigma_{\text{pps}}(y)\sigma_{\text{pps}}(x)] + (k^2/n')V_{\text{pps}}(x), \tag{18}$$

the estimator of variance being

$$\hat{V}(\hat{Y}) = k^2[n'(n' - 1)]^{-1}\textstyle\sum_1^{n'} (x_i/p_i - n'^{-1}\sum x_i/p_i)^2 + [n(n - 1)]^{-1}\sum_1^n (d_i/p_i - n^{-1}\sum d_i/p_i)^2. \tag{19}$$

The large sample variance of the comparable ratio estimator is

$$V_{\text{rat}}(\hat{Y}) = n^{-1}[V_{\text{ran}}(y) + R^2V_{\text{ran}}(x) - 2R\rho\sigma_{\text{ran}}(y)\sigma_{\text{ran}}(x)] + (R^2/n')V_{\text{ran}}(x). \tag{20}$$

A different application of the double sampling technique involving sampling with probabilities proportionate to size is given by Raj (1964).

REFERENCES

Cochran, W. G. (1963). *Sampling Techniques* (2nd edition), Wiley, New York.

Eckler, A. R. (1961). The continuous population and labour force survey in the United States. *Family Living Studies*, International Labour Office, Geneva.

Jessen, R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agr. Sta. Res. Bull.* 304.

Lahiri, D. B. (1954). Technical paper on some aspects of the development of the sample design. *The National Sample Survey No. 5*, Ministry of Finance, India.

Patterson, H. D. (1950). Sampling on successive occasions with partial replacement of units. *J. Roy. Statist. Soc. Ser. B* **12** 241–255.

Raj, D. (1964). On double sampling for pps estimation. *Ann. Math. Statist.* **35** 900–902.

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