ON CERTAIN EXTENDED CASES OF DOUBLE SAMPLING

By K. C. SEAL

Statistical Laboratory, Calcutta

1. Introduction

In a previous paper, (Seal, 1951) the author has discussed the accuracies of estimates corresponding to different situations of double sampling. In all types of double sampling, whether with one or many auxiliary variables there are two independent samples in which the variables be chosen according to any desired method of sampling such as random, systematic, fixed etc. The present paper is devoted mainly to the study of an extended case of double sampling in which besides the two independent samples, an additional third sample for the main character is also available. Such a set of samples occurs in practice, as for instance when a sample for the main character has been already taken, during some preliminary investigations or before suitable correlated auxiliary variates could be found out. Estimation of parameters from such a system of samples then becomes similar to estimation from fragmentary data considered by Wilks (1932) or from matched samples discussed by Jessen (1942). Under certain conditions of sampling according to the usual double sampling set-up, it may be possible to combine observations of the auxiliary variables in the two samples to get pooled estimates of their means, so as to obtain a more accurate estimate of the main characters. This problem is investigated in this paper for an auxiliary set of many variables.

2. Normally distributed variables—multivariate auxiliary set

Let the three independent samples be as follows:

Sample (A) consisting of y only and of size $n_1=p_1n$

Sample (B) consisting of $x_1,\ldots,x_k$ and y, of size n

Sample (C) consisting of $x_1,\ldots,x_k$ only, of size $n_2=p_2n$

Assuming y's in sample (A) to follow a univariate normal distribution with mean η and known variance σ_y^2; $x_1,\ldots,x_k$ and y's in sample (B) to follow a multivariate normal distribution with means $\xi_1, \xi_2, \ldots, \xi_k$ and η and known covariance $\sigma_{ij}(i, j = 1, 2,\ldots,k)$ between x_i and x_j respectively; and $x_1, x_2,\ldots,x_k$ in sample (C) to follow another

multivariate normal distribution with the same means and dispersions, the maximum likelihood estimate of η will be given by

$$\hat{\eta} = \begin{vmatrix} \sigma^{(y)}\bar{y}+\frac{p_1}{\sigma_y^2}\bar{y}_1 & \sigma^{(y1)} & \sigma^{(y2)} & \dots & \sigma^{(yk)} \\ \sigma^{(1y)}\bar{y}+p_2\sum_1^k\sigma^{1i}(\bar{x}_{1i}-\bar{x}_i) & \sigma^{(11)}+p_2\sigma^{11} & \sigma^{(12)}+p_2\sigma^{12} & \dots & \sigma^{(1k)}+p_2\sigma^{1k} \\ \sigma^{(2y)}\bar{y}+p_2\sum_1^k\sigma^{2i}(\bar{x}_{1i}-\bar{x}_i) & \sigma^{(21)}+p_2\sigma^{21} & \sigma^{(22)}+p_2\sigma^{22} & \dots & \sigma^{(2k)}+p_2\sigma^{2k} \\ \dots\dots & \dots\dots & \dots\dots & \dots & \dots\dots \\ \sigma^{(ky)}\bar{y}+p_2\sum_1^k\sigma^{ki}(\bar{x}_{1i}-\bar{x}_i) & \sigma^{(k1)}+p_2\sigma^{k1} & \sigma^{(k2)}+p_2\sigma^{k2} & \dots & \sigma^{(kk)}+p_2\sigma^{kk} \end{vmatrix} \Bigg/ \begin{vmatrix} \sigma^{(yy)}+\frac{p_1}{\sigma_y^2} & \sigma^{(y1)} & \sigma^{(y2)} & \dots & \sigma^{(yk)} \\ \sigma^{(1y)} & \sigma^{(11)}+p_2\sigma^{11} & \sigma^{(12)}+p_2\sigma^{12} & \dots & \sigma^{(1k)}+p_2\sigma^{1k} \\ \dots\dots & \dots\dots & \dots\dots & \dots & \dots\dots \\ \sigma^{(1y)} & \sigma^{(k1)}+p_2\sigma^{k1} & \sigma^{(k2)}+p_2\sigma^{k2} & \dots & \sigma^{(kk)}+p_2\sigma^{kk} \end{vmatrix} \quad \dots (2.1)$$

where $((\sigma^{(ij)}))$ $(i,j=y,1,\dots,k)$ and $((\sigma^{ij}))(i,j=1,2,\dots,k)$ represent the reciprocals to population dispersion matrices for the samples (B) and (C) respectively; $\bar{y},\bar{x}_i$, $(i=1,2,\dots,k)$ denote the sample means of y and $x_i(i=1,2,\dots,k)$ from the sample (B); $\bar{y}_1$ denote the mean of y's from sample (A) and $\bar{x}_{1i}$ $(i=1,\dots,k)$ denote the mean of x_i's for sample (C).

The large sample variance of $\hat{\eta}$ will be given by

$$V(\hat{\eta}) = \frac{1}{n} \begin{vmatrix} \sigma^{(11)}+p_2\sigma^{11} & \sigma^{(12)}+p_2\sigma^{12} & \dots & \sigma^{(1k)}+p_2\sigma^{1k} \\ \sigma^{(21)}+p_2\sigma^{21} & \sigma^{(22)}+p_2\sigma^{22} & \dots & \sigma^{(2k)}+p_2\sigma^{2k} \\ \dots\dots & \dots\dots & \dots & \dots\dots \\ \sigma^{(k1)}+p_2\sigma^{k1} & \sigma^{(k2)}+p_2\sigma^{k2} & \dots & \sigma^{(kk)}+p_2\sigma^{kk} \end{vmatrix} \Bigg/ \begin{vmatrix} \sigma^{(yy)}+\frac{p_1}{\sigma_y^2} & \sigma^{(y1)} & \sigma^{(y2)} & \dots & \sigma^{(yk)} \\ \sigma^{(1y)} & \sigma^{(11)}+p_2\sigma^{11} & \sigma^{(12)}+p_2\sigma^{12} & \dots & \sigma^{(1k)}+p_2\sigma^{1k} \\ \dots\dots & \dots\dots & \dots\dots & \dots & \dots\dots \\ \sigma^{(ky)} & \sigma^{(k1)}+p_2\sigma^{k1} & \sigma^{(k2)}+p_2\sigma^{k2} & \dots & \sigma^{(kk)}+p_2\sigma^{kk} \end{vmatrix} \quad \dots (2.2)$$

In particular, if there is only one auxiliary variable x, the formulae (2.1) and (2.2) reduce to

$$\hat{\eta} = \frac{(1+p_2)\bar{y}+p_1\bar{y}_1[1+p_2(1-\rho^2)]+p_2\rho\frac{\sigma_y}{\sigma_x}(\bar{x}_1-\bar{x})}{1+p_1+p_2+p_1p_2(1-\rho^2)} \quad \dots (2.11)$$

and $$V(\hat{\eta}) = \frac{\sigma_y^2[1+p_2(1-\rho^2)]}{n[1+p_1+p_2+p_1p_2(1-\rho^2)]} \quad \dots (2.21)$$

where ρ denotes the correlation between x and y in sample (B). It is interesting to note that the equations (2.11) and (2.21) are those given by Wilks (1932) and (2.2)

will agree with the result given by Matthai (1949) when some of the fragmentary samples are assumed to be absent.

3. WEIGHTED LEAST SQUARE ESTIMATES—ONE AUXILIARY VARIABLE

In the previous section, it was assumed that the distributions of y's and x's were known to be normal in each of the following three independent samples:

Sample (A), y only, of size n_1
Sample (B), both x and y, of size n
Sample (C), x only, of size n_2.

Here without assuming any explicit distribution, the estimate of η and its variance, as also the relative loss in efficiency when the distribution is known to be normal are derived.

To derive $\hat{\eta}$ in such a situation, the method of weighted least squares is employed. The merit of this method is that it always gives the best linear unbiased estimate and resolves itself into the maximum likelihood estimate when the distributions of the initial variates are assumed to be normal.

For usual double sampling the estimate of η based on samples (B) and (C) is given by

$$Y_2 = \bar{y} + b(\bar{x}_1 - \bar{x}) \qquad \ldots (3.1)$$

and
$$V(Y_2) \simeq \sigma_y^2\left(\frac{1-\rho^2}{n} + \frac{\rho^2}{n_2}\right) \quad \text{in large samples} \qquad \ldots (3.2)$$

Another estimate of η will be

$$Y_1 = \bar{y}_1 \qquad \ldots (3.3)$$

and
$$V(Y_1) = \frac{\sigma_y^2}{n_1} \qquad \ldots (3.4)$$

Hence the best pooled estimate of η will be given by

$$\hat{\eta}' = \frac{\frac{\bar{y}_1}{V(Y_1)} + \frac{\bar{y} + b(\bar{x}_1 - \bar{x})}{V(Y_2)}}{\frac{1}{V(Y_1)} + \frac{1}{V(Y_2)}}$$

and
$$V(\hat{\eta}') \simeq \frac{\sigma_y^2[n_2(1-\rho^2) + n\rho^2]}{nn_2 + n_1n_2(1-\rho^2) + nn_1\rho^2} \qquad \ldots (3.5)$$

a formula arrived at by Jessen (1942) while studying a very similar problem. Also the variance $V(\hat{\eta})$ of the maximum likelihood estimate $\hat{\eta}$ assuming normal distributions is given by (2.21), i.e.,

$$V(\hat{\eta}) = \frac{\sigma_y^2[n + n_2(1-\rho^2)]}{n^2 + nn_2 + nn_1 + n_1n_2(1-\rho^2)} \qquad \ldots (3.6)$$

Hence
$$\frac{V(\hat{\eta}') - V(\hat{\eta})}{V(\hat{\eta})} = \frac{n^2\rho^2}{[n + n_2(1-\rho^2)][nn_1 + nn_1\rho^2 + n_1n_2(1-\rho^2)]} \qquad \ldots (3.7)$$

$$> 0 \quad \text{when} \quad \rho \neq 0$$
$$= 0 \quad \text{when} \quad \rho = 0$$

Now the function,
$$Z = \frac{[n + n_2 - n_2\rho^2][n_2(n_1+n) - n_1(n_2-n)\rho^2]}{n^2\rho^2}$$
$$= \frac{(a_1 - a_2\rho^2)(b_1 - b_2\rho^2)}{\rho^2} \quad \text{(say)}$$

(where a_1, a_2 and $b_1 \gtrless 0$, and $b_2 \gtreqless 0$ according as $n_2 \gtreqless n$)

$$= \frac{a_1 b_1}{\rho^2} - (a_2 b_1 + a_1 b_2) + a_2 b_2 \rho^2$$

It is found to be a minimum when $\rho^2 = 1$.

Hence from (3.7) the maximum relative gain in efficiency of $\hat{\eta}$ over $\hat{\eta}'$ will be obtained when $\rho = \pm 1$ and is given by

$$\left[\frac{V(\hat{\eta}') - V(\hat{\eta})}{V(\hat{\eta})}\right]_{\max} = \frac{n}{n_1 + n_2} \qquad \ldots \quad (3.8)$$

Thus if $n_1 \neq n_2$ is large compared with n, which is very often the case, the gain in efficiency when the distribution is known to be normal will be negligibly small.

The above weighted least square method can be applied for deriving $\hat{\eta}$ when k auxiliary variates are present in samples (B) and (C). This method can also be applied to all the different types of sampling of x's and y's considered previously by the author (Seal, 1951) when an additional independent sample of y's is also present.

4. Allocation of sample sizes

The variance of the maximum likelihood estimate of η in the double sampling procedure with one auxiliary variate when another random sample of y's (size n_1) had already been taken is given by

$$V(\hat{\eta}) = \frac{\sigma_y^2[n + n_2(1-\rho^2)]}{n(n+n_2) + n_1[n + n_2(1-\rho^2)]}$$

$$\therefore \qquad V^{-1}\sigma_y^2 = I_1 = \frac{n(n+n_2)}{n + n_2(1-\rho^2)} + n_1 \qquad \ldots \quad (4.1)$$

If the problem is to choose n and n_2 in such a manner as to maximise (4.1) for a given cost T, where

$$T = C_0 + (C_1 + C_2)n + C_2 n_2 \qquad \ldots \quad (4.2)$$

C_0 being the overhead cost, C_1 and C_2 denoting the cost of taking one unit of y and of x respectively, the following solutions are obtained:

$$\frac{n_2}{n} = \sqrt{\frac{C_1}{C_2}} \frac{|\rho|}{\sqrt{1-\rho^2}} - 1 \qquad \ldots \quad (4.3)$$

and

$$n = \frac{T - C_0}{\sqrt{C_1}} \frac{\sqrt{1-\rho^2}}{\sqrt{C_1}\sqrt{1-\rho^2} + \sqrt{C_2}|\rho|} \qquad \ldots \quad (4.4)$$

The expression (4.3) will ordinarily be positive, but in the exceptional cases when it is negative the best choice seems to be to take the sample of y's only, without considering x's at all.

It was shown in sec. 3 that another alternative estimate of η without any explicit assumption of the normal distribution will be usually quite efficient.

Here

$$V(\hat{\eta})' = \frac{\sigma_y^2[n\rho^2 + n_2(1-\rho^2)]}{nn_2 + n_1[n\rho^2 + n_2(1-\rho^2)]}$$

$$\therefore \qquad V'^{-1}\sigma_y^2 = I'_1 = \frac{nn_2}{n\rho^2 + n_2(1-\rho^2)} + n_1 \qquad \ldots \quad (4.11)$$

and

$$T = C_0 + (C_1 + C_2)n + C_2 n_2$$

The optimum solutions of n_2 and n will then be given by

$$\frac{n_2}{n} = \sqrt{\frac{C_1+C_2}{C_2}}\,\frac{|\rho|}{\sqrt{1-\rho^2}} \qquad \ldots (4.31)$$

and the optimum n for a given T would be

$$n = \frac{T-C_0}{\sqrt{C_1}+\sqrt{C_2}}\,\frac{\sqrt{1-\rho^2}}{\sqrt{C_1+C_2}\sqrt{1-\rho^2}+\sqrt{C_2}|\rho|} \qquad \ldots (4.41)$$

We may make here certain observations in this connection:

(1) Comparing the equations (4.3) and (4.4) with the equations (4.31) and (4.41) it is evident that the latter alternative method will tend to give somewhat lower values for n. But since in usual double sampling procedure n will be very large compared to n_2, the discrepancy between (4.3) and (4.31) will not be much.

(2) The optimum allocations of n and n_2 as given in (4.3), (4.4), (4.31) and (4.41) would be unaltered for any sample size $n_1(\geq 0)$ of y's. Thus for ordinary double sampling method where the variance of the estimate $\hat{\eta}$ is given approximately by (4.11) with $n_1=0$, the solutions (4.31) and (4.41) are identical with those given by Ghosh (1949).

(3) The estimate of η in the double sampling procedure when $n_1=0$ is not the maximum likelihood estimate when the distributions are considered to be normal.

(4) The above comparisons between (4.3), (4.4), (4.31) and (4.41) would also hold good between the allocations for corresponding ordinary double sampling and maximum likelihood estimate (2.11) with $p_1=0$.

5. Possibility of improved estimate from double sampling set-up

In the usual double sampling procedure with one auxiliary variate the estimate Y of the main character η where

$$Y = \bar{y}_n + b_n(\bar{x}_N - \bar{x}_n) \qquad \ldots (5.1)$$

is not the same as the maximum likelihood estimate Y'. The estimate Y' can be written in the form (5.1) if it is assumed that the large sample mean $\bar{x}_N$ of the auxiliary variate x contains the small sample of x's of size n as a part of it, (c.f. two-phase sampling) i.e. here $\bar{x}_N$ is not independent of $\bar{x}_n$ and can be written as

$$\bar{x}_N = \frac{N-n}{N}\,\bar{x}_{N-n} + \frac{n}{N}\,\bar{x}_n$$

where $\bar{x}_{N-n}$ is independent of $\bar{x}_n$. In such a case

$$Y' = \bar{y}_n + b_n\frac{N-n}{N}\,(\bar{x}_{N-n} - \bar{x}_n) \qquad \ldots (5.2)$$

It is well known that Y' is more accurate than Y. Y' is equivalent to the usual estimate used in two-phase sampling. Now if in double sampling the x's in the small and the large sample are taken from *the same population* in such a way as to ensure homogeneity etc. the x's in the small sample can be used along with the x's in the large sample to get a more precise estimate of population mean of x's. The estimation in double sampling and two-phase sampling, then, become the same.

In this section, we shall derive the variance of such an estimate when there are k auxiliary variables. Thus we would consider two independent samples as follows:

(1) $y, x_1, x_2, \ldots, x_k$ of size n.
(2) $x_1, x_2, \ldots, x_k$ of size $N-n$.

and the estimate Y' of η as

$$Y' = \bar{y}_n + \sum_{i=1}^{k} b_{ni}(\bar{x}_{Ni} - \bar{x}_{ni}) \qquad \ldots \quad (5.3)$$

Obviously $E(Y')=\eta$ and hence the estimate is unbiased. Now (5.3) can be written as

$$Y' = \bar{y}_n + \frac{N-n}{N} \sum_{i=1}^{k} b_{ni}(\bar{x}_{N-n\,i} - \bar{x}_{ni}) \qquad \ldots \quad (5.31)$$

$$V(Y') = \sigma_y^2(1 - R_{y\cdot 12\ldots k}^2)\left[\frac{1}{n} + \frac{k}{n-k-2}\left(\frac{1}{n} - \frac{1}{N}\right)\right] + \frac{\sigma_y^2 R_{y\cdot 12\ldots k}^2}{N} \qquad \ldots \quad (5.4)^*$$

where $R_{y\cdot 12\ldots k}$ denotes the multiple correlation of y on $x_1, x_2, \ldots, x_k$.

In the above derivation b_{ni} has been assumed to be distributed independently of $\bar{x}_{ni}$ which is true for the normal distribution; otherwise it will be valid only in large samples. This formula (5.4) can now be compared with the formula for double sampling with k auxiliary variables (Seal, 1951). It is found that the forms are remarkably similar and (5.4) is always less than the corresponding double sampling variance. Moreover, the cost incurred in this method will be less, as only $(N-n)$ observations, instead of N, of $x_1, x_2, \ldots, x_k$ need be taken in the second sample.

The situations in which the x's in the two samples are not to be combined and the usual double sampling estimation has to be adhered to are obious. Combining of x's is obviously not justified when the expectation of the x's in the two samples are different. Attention may also be drawn to the particular case in which the linear regressions of y on the two sets work out to be the same but the expectations of the two sets of x remain different (Vide C. Bose, 1951). Also when x's in the small sample are assumed to be fixed, combining x's will not always be justified.

I am grateful to Dr. C. R. Rao for his guidance in the course of my investigation; I have also to thank Mr. A. Matthai for suggesting some improvements in the presentation.

REFERENCES

BOSE, C. (1951): Some further results on errors in double sampling technique. *Sankhyá*, 11, 191-194.

GHOSH, B. (1949): A query on double sampling. *Cal. Stat. Assoc. Bull.*, 2, 34-37.

JESSEN, R. J. (1942): Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agric. Exp. Sta. Res. Bull.*, 304.

MATTHAI, A. (1949): Estimation of parameters from incomplete data with application to the design of sample surveys. *Proc. Ind. Sc. Cong.*, 1949.

SEAL, K. C. (1951): On errors of estimates in various types of double sampling procedures. *Sankhyá*, 11, 125-144.

WILKS, S. S. (1932): Moments and distributions of estimates of population parameters from fragmentary samples. *Ann. Math. Stat.*, 3.

Paper received : August, 1951.

* It may be noted that a particular case of (5.4) with $k=1$ was developed by Cochran as given by Jessen (1942).