Jour. Ind. Soc. Ag. Statistics
Vol. XLI, No. 2 (1989), pp. 149-156

ESTIMATION FROM INCOMPLETE DATA IN A TWO-STAGE SAMPLING DESIGN FOR STUDY OF TWO CHARACTERS

RANDHIR SINGH
Indian Agricultural Statistics Research Institute, New Delhi-12
(Received : February, 1984)

SUMMARY

In the present investigation suitable estimators have been proposed for estimating the population mean for study of two characters from a two-stage sampling design when data are missing for one or both the characters from some primary stage units (psus) as well as some second stage units (ssus).

Keywords : Incomplete data, Random non-response, Two-stage sampling design.

Introduction

In many of the large scale sample surveys a multi-stage sampling design is generally used for selection of a sample and data are collected for several items. But, quite often, some of the sample units fail to provide the information for one or more characters. Wilks [4] obtained unbiased estimators from incomplete data for two characters from a unistage sampling design. Singh and Singh [3] considered the study of two characters when both these characters are not enumerated for the entire sample and instead these are observed only on sub-samples. In the case of a two-stage sampling design for study of two characters, the data for one or both the characters may be missing either for some of the psus or some ssus.

2. Statement of the Problem

Consider a finite bivariate population and let x and y denote the two

characters under study. Let N and M denote respectively the number of psus and ssus within each psu in the population and let n and m be corresponding sample sizes selected by 'srswor' at both the stages. Due to non-response, information may be missing for some of the psus as well as some ssus, with respect to one or both the characters under study. We consider here the simple situation where a fixed number of ssus as well as psus are found missing at random for both the characters. When information is assumed to be missing at random the available sample can be considered as a random sample from the population and therefore the bias due to non-response is not serious and can be ignored. The missing information may be random in situations where non-response is mainly due to non-availability of the respondents or the failure of the enumerator to contact the respondent in time. For example, in milk yield surveys, milk yield for some particular day may be missing because the calf might have sucked the milk. Similarly crop yield for some plot may be missing because of the failure of the enumerator to reach on the day of harvest.

Now we define

n_x, m_{ix} = number of psu's and number of ssu's within ith psu respectively providing data for x.

n_x', m_{ix}' = number of psu's and ssu's within ith psu for which data for x are missing;

n_x'', m_{ix}'' = number of corresponding units providing data for only x.

Similarly, we may define n_y, n_y', n_y'', m_{iy}, m_{iy}' and m_{iy}'' for y, and

n_c, m_{ic} = number of psu's and ssu's within ith psu for which data are available for both x and y.

Then, the following relationships hold :

$$n_x = n_c + n_y' = n_c + n_x'' \qquad m_{ix} = m_{ic} + m_{iy}' = m_{ic} + m_{ix}''$$

$$n_y = n_c + n_x' = n_c + n_y'' \qquad m_{iy} = m_{ic} + m_{ix}' = m_{ic} + m_{iy}''$$

$$n = n_x + n_x' = n_y + n_y' \qquad m = m_{ix} + m_i' = m_{iy} + m_{iy}''$$

Let x_{ij} and y_{ij} denote the values for the two characters x and y for the jth ssu in the ith psu.

Now the sample will consist of

(i) $\sum^{n_0} m_{i0}$ sampling units for which both x and y are observed. Let $\bar{x}'$ and $\bar{y}'$ be the sample means of two characters based on these units.

(ii) $\sum^{n_x''} m_{ix} + \sum^{n_0} m_{ix}$ sampling units for which only x is observed. Let $\bar{x}''$ be the mean of x based on these units.

(iii) $\sum^{n_y''} m_{iy} + \sum^{n_0} m_{iy}''$ sampling units for which only y is observed. Let $\bar{y}''$ be the mean of y based on these units.

3. Estimate of Population Mean

An unbiased estimator for the population mean $\bar{y}$ based on all the observations available for y may be given by

$$\bar{y} = \sum^{n_y} \sum^{m_{iy}} y_{ij}/m_{iy}\, n_y \tag{1}$$

The variance of $\bar{y}$, for given n_y and m_{iy}, is given by

$$V(\bar{y}) = \left(\frac{1}{n_y} - \frac{1}{N}\right) S_{by}^2 + \frac{1}{n_y N} \sum^{N} \left(\frac{1}{m_{iy}} - \frac{1}{M}\right) S_{iy}^2 \tag{2}$$

where S_b^2 and S_i^2 are the usual between and within components of variance for a two-stage sampling design. Similarly an unbiased estimator of $\overline{X}$ based on all the observations available for x may be given by

$$\bar{x} = \sum^{n_x} \sum^{m_{ix}} x_{ij}/m_{ix}\, n_x, \quad \text{and} \tag{3}$$

$$V(\bar{x}) = \left(\frac{1}{n_x} - \frac{1}{N}\right) S_{bx}^2 + \frac{1}{n_x N} \sum^{N} \left(\frac{1}{m_{ix}} - \frac{1}{M}\right) S_{ix}^2. \tag{4}$$

Now before proceeding to obtain improved estimators of $\overline{Y}$ and $\overline{X}$, we prove the following Lemma.

LEMMA

$$\text{Cov}\,(\bar{x},\ \bar{y}) = \left(\frac{n_o}{n_x\ n_y} - \frac{1}{N}\right) S_{bxy} + \frac{n_o}{n_x\, n_y\, N} \sum_{\Sigma}^{N} \left(\frac{m_{io}}{m_{ix}\, m_{iy}} - \frac{1}{N}\right) S_{ixy} \qquad (5)$$

Proof. Making use of the relationship between the available and the missing observations on x and y, we may write

$$\text{Cov}\,(\bar{x},\ \bar{y}) = \text{Cov}\left(\sum_{\Sigma}^{n_o + n'_y} \frac{m_{io} + m'_{iy}}{\Sigma} \frac{x_{ij}}{m_{ix} n_x},\ \sum_{\Sigma}^{n_o + n'_x} \frac{m_{io} + m'_{ix}}{\Sigma} \frac{y_{ij}}{m_{iy}\, n_y}\right)$$

$$= \text{Cov}\left(\sum_{\Sigma}^{n_o} \frac{m_{io} + m'_{iy}}{\Sigma} \frac{x_{ij}}{m_{ix}\, n_x},\ \sum_{\Sigma}^{n_o} \frac{m_{io} + m'_{ix}}{\Sigma} \frac{y_{ij}}{m_{iy}\, n_y}\right)$$

+ three other similar covariance terms.

Now we have

$$\text{Cov}\left(\sum_{\Sigma}^{n_o} \frac{\bar{X}_i}{n_x},\ \sum_{\Sigma}^{n_x} \frac{Y_i}{n_y}\right) = -\frac{n_o\, n_x}{n_x\, n_y\, N} S_{bxy}$$

where $\bar{X}_i$ and $\bar{Y}_i$ are the population means for the ith psu and

$$\text{Cov}\left[\left(\sum_{\Sigma}^{m_{io}} \frac{x_{ij}}{m_{ix}},\ \sum_{\Sigma}^{m_{ix}} \frac{y_{ij}}{m_{iy}}\right)\Big/ i\right] = -\frac{m_{io}\, m_{ix}}{m_{ix}\, m_{iy}} S_{ixy}$$

Similar type of relations hold for other such covariance terms in the Cov $(\bar{x},\ )$. After obtaining all these terms and simplifying we get the required result.

From the lemma it may be written

$$\text{Cov}\,(\bar{y}',\ \bar{x}) = \left(\frac{1}{n_x} - \frac{1}{N}\right) S_{bxy} + \frac{1}{n_x\, N} \sum_{\Sigma}^{N} \left(\frac{1}{m_{ix}} - \frac{1}{M}\right) S_{ixy}$$

and

$$\text{Cov}\,(\bar{y}',\ \bar{x}') = \left(\frac{1}{n_o} - \frac{1}{N}\right) S_{bxy} + \frac{1}{n_o N} \sum_{\Sigma}^{N} \left(\frac{1}{m_{io}} - \frac{1}{M}\right) S_{ixy}$$

Also putting $x = y$ or $y = x$ we may easily obtain the expressions for $V(\bar{y})$, $V(\bar{x})$, $V(\bar{y}')$ and $V(\bar{x}')$.

Now an improved unbiased estimator of $\bar{Y}$ may be given by

$$\bar{y}_{1r} = \bar{y}' + \beta_2 (\bar{x} - \bar{x}') \tag{5}$$

where β_2 is chosen such that the variance of $\bar{y}_{1r}$ is minimum. Similarly an improved estimator of $\bar{X}$ may be written as

$$\bar{x}_{1r} = \bar{x}' + \beta_1 (\bar{y} - \bar{y}') \tag{6}$$

The values of β_2 and β_1 which minimize the variance of $\bar{y}_{1r}$ and $\bar{x}_{1r}$ respectively are given by

$$\beta_2 = \frac{\text{Cov}(\bar{y}'\bar{x}') - \text{Cov}(\bar{y}', \bar{x})}{V(\bar{x}') - V(\bar{x})}, \quad \text{and}$$

$$\beta_1 = \frac{\text{Cov}(\bar{x}, \bar{y}') - \text{Cov}(\bar{x}', \bar{y})}{V(\bar{y}') - V(\bar{y})}$$

The variances of $\bar{y}_{1r}$ and $\bar{x}_{1r}$ are given by

$$V(\bar{y}_{1r}) = V(\bar{y}')\left[1 - \frac{[\text{Cov}(\bar{y}', \bar{x}') - \text{Cov}(\bar{y}', \bar{x}]^2}{V(\bar{y}')\,V(\bar{x}') - V(\bar{y})\,V(\bar{x})}\right] \tag{7}$$

$$V(\bar{x}_{1r}) = V(\bar{x}')\left[1 - \frac{[\text{Cov}(\bar{x}', \bar{y}') - \text{Cov}(\bar{x}', \bar{y})]^2}{V(\bar{x})\,V(\bar{y}') - V(\bar{x}')\,V(\bar{y})}\right] \tag{8}$$

4. Estimate of Variance

In order to obtain an estimate of variance of $\bar{y}_{1r}$ and $\bar{x}_{1r}$, we obtain the estimates of the between and within components of variances and covariance, viz., S_{bx}^2, S_{by}^2, S_{bxy}, S_{ix}^2 S_{iy}^2 and S_{ixy} respectively.

Consider s_{bxy} given below as an estimator of S_{bxy}

$$s_{bxy} = \frac{n_x \quad n_y}{n_c\,(n_c - n + n_x\,n_y)} \sum^{n_c} (\bar{y}_{im_{iy}} - \bar{y})(\bar{x}_{im_{ix}} - \bar{x}) \tag{9}$$

For fixed n_c, n_x and n_y, we have

$$E \sum^{n_c} (\bar{y}_{im_{iy}} - \bar{y})(\bar{x}_{im_{ix}} - \bar{x}) = E \Big(\sum^{n_c} \bar{y}_{im_{iy}} \bar{x}_{im_{ix}} - \bar{x} \sum^{n_c} \bar{y}_{im_{iy}} - \bar{y} \sum^{n_c} \bar{x}_{im_{ix}} + n_c \bar{x} \bar{y} \Big)$$

Now we may write

$$\bar{x} = \frac{\sum^{n_c} \bar{x}_{im_{ix}} + \sum^{n'_y} \bar{x}_{im_{ix}}}{n_x}, \quad \bar{y} = \frac{\sum^{n_c} \bar{y}_{im_{iy}} + \sum^{n'_x} \bar{y}_{im_{iy}}}{n_y}$$

Using these values of $\bar{x}$ and $\bar{y}$ we obtain, after some simplification,

$$E \sum^{n_c} (\bar{y}_{im_{iy}} - \bar{y})(\bar{x}_{im_{ix}} - \bar{x}) =$$

$$\frac{n_c (n_c - n + n_x n_y)}{n_x n_y} \left[S_{bxy} + \frac{1}{N} \sum^{N} \left(\frac{m_{ic}}{m_{ix} m_{iy}} - \frac{1}{M} \right) S_{ixy} \right]$$

Thus we may write

$$\text{Est. } S_{bxy} = s_{bxy} - \text{Est. } \frac{1}{N} \sum^{N} \left(\frac{m_{ic}}{m_{ix} m_{iy}} - \frac{1}{M} \right) S_{ixy} \qquad (10)$$

Similarly we may obtain s_{ixy} as an estimate of S_{ixy} given by

$$s_{ixy} = \frac{m_{ix}\, m_{iy}}{m_{ic} (m_{ic} - m + m_{ix} m_{iy})} \sum^{m_{ic}} (y_{ij} - \bar{y}_{im_{iy}})(x_{ij} - \bar{x}_{im_{ix}}) \qquad (11)$$

such that $E(s_{ixy}) = S_{ixy}$.

It may be checked here that for $n_c = n_x = n_y = n$ and $m_{ic} = m_{ix} = m_{iy} = m$, the estimates of S_{bxy} and S_{ixy} reduce to standard forms in two stage sampling. Thus we obtain

$$\hat{\text{Cov}}(\bar{x}, \bar{y}) = \left(\frac{n_c}{n_x n_y} - \frac{1}{N} \right) s_{bxy} + \frac{1}{n_c N} \sum^{n_c} \left(\frac{m_{ic}}{m_{ix} m_{iy}} - \frac{1}{M} \right) s_{ixy} \qquad (12)$$

By putting $n_c = n_x$ or n_y, $m_{ic} = m_{ix}$ or m_{iy}, and $y = x$ or $x = y$ we can obtain the estimates of Cov $(\bar{y}', \bar{x})$, Cov $(\bar{x}, \bar{y}')$, $V(\bar{x})$, $V(\bar{y})$ etc. Substituting these values we obtain the required estimator of the variance of $\bar{y}_{1r}$ as

$$V(\bar{y}_{1r}) = \left(\frac{1}{n_o} - \frac{1}{N}\right) s_{by}^2 + \frac{1}{n_o N} \sum_{i}^{n_o} \left(\frac{1}{m_{io}} - \frac{1}{M}\right) s_{iy}^2$$

$$- \frac{\left[\left(\frac{1}{n_o} - \frac{1}{n_x}\right) s_{bxy} + \frac{1}{n_o N} \sum_{i}^{n_o} \left(\frac{1}{m_{io}} - \frac{1}{m_{ix}}\right) s_{ixy}\right]^2}{\left(\frac{1}{n_o} - \frac{1}{n_x}\right) s_{bx}^2 + \frac{1}{n_o N} \sum_{i}^{n_o} \left(\frac{1}{m_{io}} - \frac{1}{m_{ix}}\right) s_{ix}^2} \qquad (13)$$

Expression for variance of $\bar{x}_{1r}$ may also be obtained easily by changing y to x in (13).

If we do not use the entire information from the incomplete samples by using $\bar{y}_{1r}$ and $\bar{x}_{1r}$ we will be making use of the estimators $\bar{y}'$, and $\bar{x}'$, for estimating $\bar{Y}$ and $\bar{X}$ respectively. It can be easily seen that the estimators $\bar{y}_{1r}$ and $\bar{x}_{1r}$ are always more efficient than the estimators $\bar{y}'$, and $\bar{x}'$.

5. Empirical Illustration

In order to illustrate the usefulness of the suggested procedure, data from an actual survey conducted by the Indian Agricultural Statistics Research Institute, New Delhi (Raut *et al.* [2]) to study the Economics of Raising Cattle during the years 1977-1980 in the district Nadia, West Bengal have been utilised. In this survey data were collected for two characters : (i) milk yield per milch cross bred cattle (x), and (ii) maintenance cost per milch cross bred cattle (y).

Stratified two-stage sampling was utilized where Police Stations or group of Police Stations formed the strata. The cluster of villages formed the psus and households within psus formed ssus. From each stratum a sample of 8 clusters was selected and from each selected cluster a sample of 12 households having cross bred milch cattle was selected for data collection. In the present investigation, the data for the first stratum in which only 7 selected clusters having the cross bred milch cattle could be enumerated ware used. Similarly in some households the data on one or both the characters could not be recorded for different reasons.

The extent of data for the two characters is as follows:

Cluster No.	1	2	3	4	5	6	7
m_{ix} (households on milk yield)	7	12	8	8	12	12	10
m_{iy} (households on maintenance cost)	7	12	10	10	12	12	12

The four estimators $\bar{x}'$, $\bar{y}'$, $\bar{x}_{1r}$ and $\bar{y}_{1r}$ and their estimated variances are obtained as given in the table :

Estimator	*Estimated variance*	*Estimated Efficiency %*
$\bar{x}' = 2.57$ kg	.0835	100
$\bar{x}_{1r} = 2.89$ kg	.0317	263.4
$\bar{y}' =$ Rs. 5.96	1.6752	100
$\bar{y}_{1r} =$ Rs. 6.01	.5070	360.4

It is seen from the results in the table that the efficiency of $\bar{x}_{1r}$ and $\bar{y}_{1r}$ as compared to $\bar{x}'$ and y' respectively is quite high.

REFERENCES

[1] Grewal, G. K. (1970). Variance estimation under deep stratification and estimation from incomplete data. Unpublished Ph.D. thesis submitted to IARI, New Delhi.

[2] Raut, K. C. Singh, S. and Rustagi R, L. (1982). Economics of raising cattle in a rural area of West Bengal, Project Report, IASRI.

[3] Singh, R. and Singh, D. (1983) : Sampling with partial enumeration from bivariate populations. *J.S.P.I.*, 7 : 343-351.

[4] Wilks, S. S. (1932): Distribution of estimates from fragmentary samples. A.M.S