Jour. Ind. Soc. Ag. Statistics
Vol. XXXVIII, No. 1 (1986), pp. 71-81

ON CERTAIN ASPECTS OF RANDOM NON-RESPONSE

RAVINDRA SINGH
Punjab Agricultural University, Ludhiana
(Received : March, 1980)

SUMMARY

Certain concepts have been defined for the estimates in presence of random non-response. A comparison between the estimates of population total for the unequal probability with replacement sampling and the 'Rao, Hartley and Cochran's [15] scheme of sampling' has been made with regard to the above concepts. For this comparison, the sample size has been assumed to be the same in two selection procedures.

Keywords : Non-response robustness, Generalised Binomial Distribution, Relative Efficiency, Equivalence Value.

1. Introduction

Among the various kinds of non-sampling errors that arise in sample surveys and Censuses the problem of non-response has received attention of many research workers. Several approaches to the problem have been followed by these workers. Conceiving the population as consisting of two strata, the first one comprising the units for which the data will be available when selected in the sample and the second consisting of the non-responding units, Birnbaum and Sirken [2] [3] show that either it is impossible or very costly to attain a highly guaranteed precision by increasing the sample size among the respondents only. Deming [4] studied the consequences of the successive call-backs on the units selected in the sample and showed that the bias due to non-response decreases with the repeated calls. The approach of Hansen and Hurwitz [9], suitable for the mail surveys, recommends sub-sampling of the non-respondents and with a major effort all the units in the sub-sample are enume-

rated. For further work in this direction one can refer to Durbin [5], El-Badry [6], Foradori [8], Ericson [7], Srinath [17] and Rao and Ghangurde [14]. In case of interview surveys the technique proposed by Hartley [10] and subsequently developed by Politz and Simmons [12] [13] and Simmons [16], eliminated the need for call-backs and at the same time also reduces the bias in the results of the first call. While computing the estimate of population mean in this approach higher-weight is given to the respondents who are less frequently at home. Further work on this technique has been done by Deming [4], Durbin [5], Bartholomew [1], and Sudman [18].

In interview surveys, however, the non-response is of two types viz., random and deliberate. When the enumerator fails to contact the respondent only by chance and had he been able to contact the respondent the information on the study variable would have been collected, then we classify such a non-response as a random one. The remaining other cases of non-response came under the deliberate category. In the work cited above this distinction in the two types of non-response has not been made. Depending on the nature of the study variable one may have in practice one or both types of non-responses. For example while collecting the information on the family size one is expected to get only the random non-response while in a survey where the information on personal incomes is being collected the non-response encountered will be of both the types. In this paper, the situation where only random non-response is present has been considered. Certain desirable properties of the estimates have been defined and the single stage unequal probability sampling with replacement estimate of population total has been compared with the Rao, Hartley and Cochran's [15] (ref. Murthy : [11]) estimate with regard to these properties. In practice, single stage sampling is possible in small scale surveys although the number of such cases may not be many. Also, the development of theory for one stage sampling is desirable before one proceeds to more commonly employed multi-stage sampling schemes.

2. The Estimates and their Variances

Let there be a population of N units from which a sample of size n ($N = nK$ where K is a positive integer) is to be drawn. Let the study variable be denoted by y and the auxiliary variable by x. The selection probabilities $\{P_i\}$ of the population units are taken to be proportional to the corresponding x-values. Let r ($r = 0, 1, \ldots, n-1$) be the number of units (including repetitions in case of probability proportional to size and with replacement sampling) on which the information on y could not be collected. The variable r is considered not to assume a value n

since otherwise we do not have information on any unit in the sample. In the present paper we propose to investigate the effect of this non-response on the estimates of population total for the single stage *pps* with replacement and Rao, Hartley and Cochran's scheme of sampling. Then we have the following theorem for the Rao, Hartley and Cochran's scheme with equal size ($= N/n$) random groups.

THEOREM 2.1. *For a given r an unbiased estimate of the population total Y is*

$$\hat{Y}(r) = \frac{n}{n-r} \sum_{i=1}^{n-r} \frac{y_i}{p_i} \pi_i$$

where for the ith unit selected in the sample

y_i = *value of* y,

p_i = *initial probability*

and

π_i = *sum of probabilities for all the units in the random group, from which this unit is selected.*

Proof. Let E_2 denote the expectation for the given set of groups from which the response has been received and E_1 the expectation over all groups. Then we have

$$E\,\hat{Y}(r) = E_1 E_2 (\hat{Y}(r)),$$

$$= \frac{n}{n-r} E_1 \left(\sum_{i=1}^{n-r} \sum_{j=1}^{N/n} Y_{ij} \right),$$

$$= Y$$

since $\left[\frac{N(n-r)}{n}\right]^{-1} \left(\sum_{i=1}^{n-r} \sum_{j=1}^{N/n} Y_{ij} \right)$

denotes the mean of a simple random sample of size $(N/n)(n-r)$ whose expected value is the population mean.

THEOREM 2.2. *The variance of the estimate $\hat{Y}(r)$ is given by*

$$V(\hat{Y}(r)) = \left(1 - \frac{n-1}{N-1}\right) \frac{\sigma^2}{n-r} + \frac{Nr}{n-r} \cdot S^2,$$

where

$$\sigma_z^2 = \sum_{j=1}^{N} \frac{Y_j^2}{p_j} - Y^2$$

and

$$S^2 = \left(\sum_{j=1}^{N} Y_j^2 - N\overline{Y}^2\right)\Big/(N-1).$$

Proof. We have with variances V_1 and V_2 defined in the same manner as the expectations E_1 and E_2

$$V(\hat{Y}(r)) = E_1 V_2(\hat{Y}(r)) + V_1 E_2(\hat{Y}(r)),$$

$$= E_1 \frac{n^2}{(n-r)^2} \sum_{i=1}^{n-r} \left[\sum_{j=1}^{N/n} \frac{Y_{ij}^2}{P_{ij}} \pi_i - Y_i^2\right] + V_1\left[\frac{n}{n-r}\sum_{i=1}^{n-r} Y_i\right]$$

$$= \frac{n^2}{n-r} E_1\left[\sum_{j=1}^{N/n} Y_{ij}^2 + \sum_{j\neq j'=1}^{N/n} \frac{Y_{ij}^2}{P_{ij}} P_{ij'} - Y_i^2\right]$$

$$+ \frac{n^2}{(n-r)^2} V_1\left(\sum_{i=1}^{n-r} Y_i\right)$$

$$= \frac{n^2}{n-r}\left[\frac{1}{n}\sum_{j=1}^{N} Y_j^2 + \frac{N/n\,(N/n-1)}{N(N-1)} \sum_{j\neq j'=1}^{N} \frac{Y_j^2}{P_j} P_j' - \left(\frac{N}{n}\right)^2 \cdot \left[\frac{N-N/n}{N\cdot N/n} S^2 + \overline{Y}^2\right]\right] + \frac{(N-N(n-r)/n)}{(n-r)/n} S^2.$$

$$= \left(1 - \frac{n-1}{N-1}\right)\frac{\sigma_z^2}{n-r} + \frac{Nr}{n-r} S^2$$

This proves the theorem.

COROLLARY 2.1. *For $r = 0$ the variance $V(\hat{Y}(r))$ reduces to the variance when there is no non-response and all the random groups are of equal size.*

THEOREM 2.3. *An unbiased estimate of the variance $V(\hat{Y}(r))$ is :*

$$v(\hat{Y}(r)) = \frac{n}{N(n-r)(n-r-1)} \sum_{i=1}^{n-r} \left(\frac{N-n}{P_i} + Nr\right)\frac{y_i^2}{P_i}\pi_i$$

$$- \frac{N-n+r}{(n-r-1)} \hat{Y}^2(r).$$

Proof. Substituting the expressions for σ_z^2 and S^2 we find after simplification that

$$(N-1)(n-r)\,V(\hat{Y}(r)) = (N-n)\sum_{j=1}^{N}\frac{Y_j^2}{P_j} + Nr\sum_{j=1}^{N}Y_j^2 - (N-n+r)\,Y^2.$$

Thus if $v(\hat{Y}(r))$ denotes the unbiased estimate of the variance $V(\hat{Y}(r))$, then the expression for $v(\hat{Y}(r))$ can be obtained by substituting unbiased estimates for various terms on the right hand side of the above equation. This gives us the following :

$$(N-1)(n-r)\,v(\hat{Y}(r)) = (N-n)\,\frac{n}{n-r}\sum_{i=1}^{n-r}\frac{Y_i^2}{p_i}\pi_i + Nr\frac{n}{n-r}\sum_{i=1}^{n-r}\frac{y_i^2}{p_i}\pi_i - (N-n+r)[\hat{Y}^2(r) - v(\hat{Y}(r))],$$

or

$$N(n-r-1)\,v(\hat{Y}(r)) = \frac{n}{n-r}\sum_{i=1}^{n-r}\left(\frac{N-n}{P_i} + Nr\right)\frac{Y_i^2}{p_i}\pi_i - (N-n+r)\,\hat{Y}^2(r).$$

which proves the theorem.

COROLLARY 2.2. *For $r = 0$, $v(\hat{Y}(r))$ coincides with the one given in literature for equal sized random groups.*

In case of PPS with replacement scheme when the y characteristic cannot be observed on r ($r = 0, 1, 2, \ldots, n-1$) units (including repetitions) selected in a sample of size n, the results stated in the following three theorems are obvious.

THEOREM 2.4. *For a given r an unbiased estimate of the population total Y is*

$$\hat{Y}'(r) = \frac{1}{n-r}\sum_{i=1}^{n-r} y_i/p_i$$

where p_i is as defined in theorem 2.1.

THEOREM 2.5. *The variance* $V(\hat{Y}'(r))$ *of the estimate* $\hat{Y}'(r)$ *is given by*

$$V(\hat{Y}'(r)) = \sigma_z^2 / (n - r),$$

where σ_z^2 *is as defined in theorem 2.2*

THEOREM 2.6. *An unbiased estimate of the variance* $V(\hat{Y}'(r))$ *is obtained as*

$$v(\hat{Y}'(r)) = \frac{1}{(n-r)(n-r-1)} \left[\sum_{i=1}^{n-r} \frac{y_i^2}{p_i} - (n-r)\ \hat{Y}'^2(r) \right]$$

3. The Relative Efficiency

In order to investigate the relative performance of the two estimators considered in this paper, when r of the n selected units could not be observed for the variable y, we define the relative efficiency R.E. (r) below :

DEFINITION 3.1. Let $\hat{\theta}_1(r)$ and $\hat{\theta}_2(r)$ be two unbiased estimators of the same population parameter θ, then the relative efficiency of the estimator $\hat{\theta}_1(r)$ with respect to the estimator $\hat{\theta}_2(r)$ is defined as

$$\text{R.E. } (r) = \frac{V(\hat{\theta}_2(r))}{V(\hat{\theta}_1(r))} .$$

Thus from theorems 2.2 and 2.5 the relative efficiency R.E. (r) of the estimate $\hat{Y}'(r)$ with respect to the estimate $\hat{Y}(r)$ is given by

$$\text{R.E. } (r) = \frac{V(\hat{Y}(r))}{V(\hat{Y}'(r))},$$

$$= \frac{(N-n)\sigma_z^2 + N(N-1)\, r\, S^2}{(N-1)\sigma_z^2} \qquad (3.1)$$

The usual relative efficiency defined in the sampling theory is, therefore, R.E. (0). It is easy to see that the relative efficiency R.E. (r) given in (3.1) increases with r, the lowest value being $(N-n)/(N-1) < 1$ for $r = 0$ and the highest value occuring for $r = n - 1$.

For any given value of r, the estimate corresponding to the Rao, Hartley and Cochran's scheme will be more efficient than the PPS with replacement sampling estimate if

$$V(\hat{Y}(r)) \leqslant V(\hat{Y}'(r)).$$

This condition from (3.1) reduces to

$$\sigma_z^2 \geqslant \frac{N(N-1)\,r}{n-1} \cdot S^2 \tag{3.2}$$

The condition is always satisfied for $r = 0$. It must be mentioned here that in this comparison we have not taken into account the cost aspect of the problem. In case of PPS with replacement scheme of sampling the expected number of distinct units in the sample is less than n and hence the expected cost is also less than that in case of Rao, Hartley and Cochran's scheme. The condition (3.2) may not, however, be satisfied for the values of r more than a certain value, say r_e, which we shall call equivalence value of r.

DEFINITION 3.2. If $\hat{\theta}_1(r)$ and $\hat{\theta}_2(r)$ be two unbiased estimates of the same parameter θ, then the value of $r = r_e$ $(0 \leqslant r \leqslant n-1)$ for which $V(\hat{\theta}_1(r)) = V(\hat{\theta}_2(r))$ is called the equivalence value of r.

For the two estimators considered in this paper the value of r_e is obtained as the solution of the equation

$$\text{R.E.}(r) = 1,$$

or

$$(N-1)\,\sigma_z^2 = (N-n)\,\sigma_z^2 + N(N-1)\,r_e\,S^2.$$

Thus

$$r_e = \frac{(n-1)\,\sigma_z^2}{N(N-1)\,S^2} \tag{3.3}$$

For $0 \leqslant r_e \leqslant n-1$, to hold we get the condition as

$$0 \leqslant \sigma_z^2 \leqslant N(N-1)\,S^2 \tag{3.4}$$

From (3.1) it is clear that R.E. (r) is a monotonically increasing function of r and, therefore, we have, for the two estimates considered in this paper, R.E. $(r) \geqslant 1$ whenever $r \geqslant r_e$.

DEFINITION 3.3 If $\{P(r)\}$ $(r = 0, 1, n-1)$ denotes the probability distribution of r, then expected relative efficiency (E.R.E.) is defined as

$$\text{E.R.E.} = \sum_{r=0}^{n-1} (P(r) \cdot \text{R.E.}(r) \tag{3.5}$$

For the case considered here we have from (3.1) and (3.5)

$$\text{E.R.E.} = \frac{N-n}{N-1} + \frac{NS^2}{\sigma_z^2} \sum_{r=0}^{n-1} rP(r),$$

$$= \frac{N-n}{N-1} + \frac{NS^2}{\sigma_z^2} E(r), \tag{3.6}$$

where $E(r)$ denotes the expected value of r.

In practice the likely distribution that r is expected to follow is the generalized truncated binomial. However, for the discussion in this paper we shall assume it to be truncated Binomial (which implies that the probability of contacting a respondent is same for all respondents which is, of course, a simplification of the actual situation). Thus taking

$$P(r) = \binom{n}{r} p^r q^{n-r}/(1-p^n), \; r = 0, \ldots, n-1 \tag{3.7}$$

where p is the probability of not meeting a person and $q = 1 - p$, it can be easily verified that

$$E(r) = np(1-p^{n-1})/(1-p^n). \tag{3.8}$$

Thus for this case we get the expression for the expected relative efficiency as

$$\text{E.R.E.} = \frac{N-n}{N-1} + \frac{NnS^2}{\sigma_z^2} \cdot \frac{p(1-p^{n-1})}{1-p^n}. \tag{3.9}$$

When the variable r follows the distribution (3.7), we have the following definition :

DEFINITION 3.4. For any given population and sample size the value of $p = p_e$ that satisfies the equation

$$\text{E.R.E.} = 1,$$

is called the equivalence value of p.

Finding of the admissible value of p_e by algebraically solving the equation

$$\frac{N-n}{N-1} + \frac{NnS^2}{\sigma_z^2} \frac{p(1-p^{n-1})}{1-p^n} = 1, \tag{3.10}$$

is somewhat difficult as the equation is of nth degree in p. The solution can, however, be found by using interpolation methods.

4. Non Response Robustness

A desirable property for any estimate $\hat{\theta}(r)$ of the parameter θ is that it should have smaller mean square error. In case of random non-response another desirable property that the estimate should possess is the non-response robustness. An estimate, the variance of which increases less with increase in the value of r shall be more non-response robust. Although several measures could be proposed for the non-response robustness but in this paper we shall use the variance of $V(\hat{\theta}(r))$ as a measure for this property (since the estimates considered are unbiased).

DEFINITION 4.1. A non-response robustness measure of an estimate $\hat{\theta}(r)$ of the parameter θ is defined as :

$$\text{NRR}(\hat{\theta}(r)) = V(V(\hat{\theta}(r))).$$

It is clear from the definition 4.1, that NRR $(\hat{\theta}(r))$ will always be positive and the minimum value that it can possibly take will be zero. Also smaller the value of NRR $(\hat{\theta}(r))$ more robust the estimate will be.

DEFINITION 4.2. The relative non-response robustness of an estimate $\hat{\theta}_1(r)$ with respect to another estimate $\hat{\theta}_2(r)$ of the same parameter is defined as

$$\text{RNRR} = \text{NRR}(\theta_2(r))/\text{NRR}(\hat{\theta}_1(r)).$$

It can be easily verified that

$$\text{NRR}(\hat{Y}(r)) = \left[\frac{N-n}{N-1}\sigma_z^2 + NnS^2\right]^2 \cdot V\left(\frac{1}{n-r}\right) \tag{4.1}$$

and

$$\text{NRR}(\hat{Y}'(r)) = \sigma_z^4 \cdot V\left(\frac{1}{n-r}\right) \tag{4.2}$$

Hence we get the relative robustness of the estimate $\hat{Y}'(r)$ with respect to the estimate $\hat{Y}(r)$ as

$$\text{RNRR} = \left[\frac{N-n}{N-1} + \frac{NnS^2}{\sigma_z^2}\right]^2.$$

5. Numerical Illustration

For the numerical illustration the data collected in a survey conducted by the department of Economics and Sociology Punjab Agricultural

University, Ludhiana in 1975-76 has been used. In this survey a sample of 70 farmer loanees of the Bank of India, Clock Tower, Ludhiana branch was selected. Information on several characters was collected in this survey. Here we have taken the study variable (y) as the loan requirement for the year 1976-77 and the farm size as the auxiliary variable (x). For the purpose of this illustration we shall treat this sample as population and the variable x is treated as the size variable. Now let us take $n = 15$ whereas $N = 70$. For this population we have $\sigma_z^2 = 38460456.52$ and $S^2 = 16072.13$. It can be easily seen from (3.1) that R.E. (0) = 0.7971 while R.E. (7) = 1.0019. Thus when the non-response is zero the PPS with replacement estimate $\hat{Y}'(0)$ is less efficient in comparison to the Rao, Hartley and Cochran's estimate $\hat{Y}(0)$ and continues to be so till $r = 6$. But for $r = 7$ the estimate $\hat{Y}'(7)$ becomes more efficient than $\hat{Y}(7)$ and the relative efficiency is 1.0019. The value of $r = 7$ is of course unlikely in actual practice and this is reflected by the value 0.8190 for the expected relative efficiency (E.R.E) when p is taken as 0.05 (Equation 3.9). The equivalence ualue of r i.e. r_e is equal to 6.936 (equation 3.3). The equivalence value for p i.e. p_e is seen to be 0.4624 (equation 3.10). The relative non-response robustness of the estimate $Y'(r)$ with respect to the estimate $\hat{Y}(r)$ is seen to be (equation 4.3) R.N.R.R. = 1.5277. The estimate $\hat{Y}'(r)$ is, therefore, more non-response robust in comparison to the estimate $\hat{Y}(r)$.

ACKNOWLEDGEMENT

The author is grateful to the referee for his helpful suggestions to improve the draft of the paper. He is also grateful to Mr. K. G. Mehta and Mr. Amar Singh for the excellent typing of the manuscript.

REFERENCES

[1] Bartholomew, D. J. (1961) : A method of allowing for 'not-at home' Bias in sample surveys, *Appl. Stat.*, **10** : 52-59.

[2] Birnbaum, Z. W. and Sirken, M. G. (1950 a) : Bias due to non-availability in sampling surveys, *Jour. Amer. Stat. Assoc.*, **45** : 98-111.

[3] Birnbaum, Z. W. and Sirken, M. G. (1950 b) : On the total error due to non-interview and to random sampling, *Int. Jour. Opinion and Attitude Res.*, **4** : 179-191.

[4] Deming, W. E. (1953) : On a probability mechanism to attain an economic balance between the resultant error of non-response and the bias of non-response, *Jour. Amer. Stat. Assoc.*, **48** : 743-772.

[5] Durbin, J. (1954) : Non-response and call-backs in surveys, *Bull. Int. Stat. Inst.*, **34**(2) : 72-86.

[6] El-Badry, M. A. (1956) : A sampling procedure for mailed questionaries, *Jour. Amer. Stat. Assoc.*, **51** : 209-227.

[7] Ericson, W. A. (1967) : Optimum sample design with non-response, *Jour. Amer. Stat. Assoc.*, **62** : 63-78.

[8] Foradori, G. T. (1961) : *Some Non-response Sampling Theory for Two Stage Designs*. North Carelina State College, Raleigh, USA.

[9] Hansen, M. H. and Hurwitz, W. N. (1946) : The problem of non-response in sample surveys, *Jour. Amer. Stat. Assoc.*, **41** : 517-529.

[10] Hartley, H. O. (1946) : Discussion of paper by F. Yates, *Jour. Roy. Stat. Soc.*, 109-47.

[11] Murthy, M. N. (1967) : *Sampling Theory and Methods*. Satistical Publishing Society, Calcutta.

[12] Politz, A. and Simmons, W. (1949) : An attempt to get 'not at home' into the sample without call-backs, *Jour. Amer. Stat. Assoc.*, **44** : 9-31.

[13] Politz, A. and Simmons, W. (1950) : Note on attempts to get'not at homes' into the sample without call-backs, *Jour. Amer. Stat. Assoc.*, **45** : 136-137.

[14] Rao, J. N. K. and Ghangurde, P. D. (1972) : Bayesian optimization in sampling finite populations, *Jour. Amer. Stat. Assoc.*, **67** : 439-443.

[15] Rao, J. N. K., Hartley, H. O. and Cochran, W. (1962) : A simple procedure of unequal probability sampling without replacement, *Jour. Roy. Stat. Soc.* (*B*), **24** : 482-491.

[16] Simmons, W. (1954) : A plan to account for 'not at homes' by combining weighting and call-backs, *Jour. Marketing*, **19** : 42-54.

[17] Srinath, K. P. (1971) : Multiphase sampling in non-response problems, *Jour. Amer. Stat. Assoc.*, **66** : 583-586.

[18] Sudman, S. (1966) : Probability sampling with quotas, *Jour. Amer. Stat. Assoc.*, **61** : 752.

[19] Sukhatme, P. V. and Sukhatme, B. V. (1970) : *Sampling Theory of Surveys with Applications*. Asia Publishing House, New Delhi.