A NOTE ON RANDOM NON-RESPONSE IN SAMPLE SURVEYS

BY

SUKHMINDER SINGH AND RAVINDRA SINGH

Punjab Agricultural University, Ludhiana

(Received : April, 1980)

SUMMARY

The effect of random non-response on the estimators of population total in case of simple random sampling and double sampling has been studied. An empirical study shows that both relative efficiency and relative non-response robustness of the single stage random sampling estimate with respect to the double sampling estimate decrease with the increase in probability of random non-response. However, the rate of decrease in former is decreasing whereas for the latter it is increasing.

INTRODUCTION

Among the various kinds of non-sampling errors, the problem of non-response has received attention of many research workers. For wrok done in this direction, one can refer to Hansen and Hurwitz [1], Srinath [6], Politz and Simmons [3] and Sukhatme and Sukhatme [7]. It is, however, possible to classify the non-response, with respect to it's nature, in two broad categories. In the first category we put all such cases where the data is missing only due to chance factors. It will, therefore, include cases where the enumerator is not able to contact the respondent only by chance and had he been able to contact, the data would have been collected. One can expect such a situation in cases where there does not seem to be any reason for the respondent's refusal to give information. For example in a survey where the information on respodent's credit needs is being collected. The case of crop yield surveys where the sampled plots are harvested before the enumerator's visit also belongs to this category. In a case where the information is kept on the punched cards, the non-response due to the accidental loss of one or more cards, will also be of this type. This type of non-response we shall call by the common name of random non-response. All other cases

of non-response shall be classified as deliberate non-response. For example the non-response in surveys where information is being collected on personal incomes or on certain unsocial habits like drinking, gambling etc., or where answers to some intimate questions are required, will come under this class. This class may also include some percentage of the random non-response. There can also be cases where the non-response of both the types may be present. In this paper we consider the situation where only random non-response is present.

The effects of random-non-response on the estimate of population total in case of double sampling and simple random sampling schemes have been considered in this paper. For double sampling, we have considered two cases. In the first case second phase sample is a subsample of the preliminary sample selected with probability proportional to size (pps) sampling while in the other case it is independently selected from the whole population with the same sampling scheme.

2. The Estimates and Their Variances

Let there be a population $(U_1, U_2, \ldots, U_N)$ of N units from which a sample of size n is to be drawn. Let the study variable be denoted by y and the auxiliary variable by x. Let r $(r=0, 1, \ldots, n-1)$ be the number of units (including repetitions in case of with replacement sampling) on which the information on y could not be collected. The variable r is not supposed to take value n in which case, we do not have information on any unit in the sample and, therefore. the question of building up an estimate does not arise. Also when $r=n-1$, we have information on only one unit in the sample and it enables us to get an unbiased estimate of the population total, although the variance of this estimate cannot be estimated.

2.1 Double Sampling Scheme

When it is desired to select the sample with probability proportional to x and the information on x is not available for all the population units, then this information is usually collected on an initial simple random sample of size n' from which a subsample of size n is selected with probability proportional to x and with replacement. Let $r(r=0, \ldots, n-1)$ be the number of units (including repetitions) on which the information on y could not be collected. In presence of this random non-response, the set of $(n-r)$ units on which the information has been collected could be treated as a

random sample selected without replacement and equal probabilities from the ultimate sample. We then have the following results.

An unbiased estimate of the population total Y is given by

$$\hat{Y}_{DR} = \frac{N}{n'} \frac{x'}{n-r} \sum_{i=1}^{n-r} \frac{y_i}{x_i}, \qquad \ldots(2.1)$$

where y_i and x_i are the value of the variables y and x respectively on the i-th unit in the sample and

$$x' = \sum_{i=1}^{n'} x_i .$$

The variance of the estimate $\hat{Y}_{DR}$ is given by

$$V(\hat{Y}_{DR}) = \frac{N}{N-1} \cdot \frac{n'-1}{n'} \sigma_z^2 \cdot E\left(\frac{1}{n-1}\right) + \frac{N(N-n')}{n'} S_y^2, \qquad \ldots(2.2)$$

where

$$\left.\begin{aligned} \sigma_z^2 &= X \sum_{i=1}^{N} \frac{Y_i^2}{X_i} - Y^2, \\ S_y^2 &= \left(\sum_{i=1}^{N} Y_i^2 - \frac{Y^2}{N} \right) / (N-1), \text{ and} \\ X &= \sum_{i=1}^{N} X_i \end{aligned}\right\} \qquad \ldots(2.3)$$

An unbiased estimate of the variance $V(\hat{Y}_{DR})$ is obtained as

$$\begin{aligned} V(\hat{Y}_{DR}) = \frac{N^2}{n'^2} \cdot \frac{x'^2}{(n-r)(n-r-1)} \Bigg[& \sum_{i=1}^{n-r} \frac{y_i^2}{x_i^2} \\ & - \frac{1}{n-r} \left(\sum_{i=1}^{n-r} \frac{y_i}{x_i} \right)^2 \Bigg] + \frac{N(N-n')}{n'(n'-1)(n-r)} \Bigg[x' \sum_{i=1}^{n-r} \frac{y_i^2}{x_i} \\ & - \frac{x'^2}{n'} \frac{1}{(n-r-1)} \left\{ \left(\sum_{i=1}^{n-r} \frac{y_i}{x_i} \right)^2 - \sum_{i=1}^{n-r} \frac{y_i^2}{x_i^2} \right\} \Bigg] \end{aligned} \qquad \ldots(2.4)$$

For $r=0$, the expressions for the estimate $\hat{Y}_{DR}$, variance $V(\hat{Y}_{DR})$ and its estimate $V(\hat{Y}_{DR})$ coincide with those available in the literature.

2.2 Double Sampling Scheme (Independent Samples)

Now, we take the case of independent samples when information on x is available for the population units. This is done when, for instance, information on x for the initial sample is available with one agency and information on both y and x on a small independent sample has been collected by another agency. It is possible to make use of the information collected by both the agencies for improving the estimate of Y. In this case, the first sample of size n' is simple random and selected without replacement, used solely for estimating X, whereas the second sample of size n is independently selected with probability proportional to x with replacement using the procedure given by Lahiri [2] in which case it is not necessary to know X. Then, we have the following there results in presence of random non-response.

The population total Y is unbiasedly estimated by

$$\hat{Y}_{DIR}=\hat{X}\hat{R} \qquad \ldots(2.5)$$

where

$$\left.\begin{aligned}\hat{X}&=\frac{N}{n'}x',\\ \hat{R}&=\frac{1}{n-r}\sum_{i=1}^{n-r}\frac{y_i}{x_i}\end{aligned}\right\} \qquad \ldots(2.6)$$

and x_i=value of x for the i-th unit in the second sample.

The variance of the estimate $\hat{Y}_{DIR}$ is given by

$$V(\hat{Y}_{DIR})=\left(\frac{1}{n'}-\frac{1}{N}\right)N^2R^2S_x^2 + \left[1+\left(\frac{1}{n'}-\frac{1}{N}\right)\frac{S_x^2}{\bar{X}^2}\right]\sigma_z^2 E\left(\frac{1}{n-r}\right), \qquad \ldots(2.7)$$

where

$$\left.\begin{aligned} S_x^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(X_i^2 - N\bar{X}^2), \\ \bar{X} &= \frac{1}{N}\sum_{i=1}^{N} x_i, \end{aligned}\right\} \quad \ldots(2.8)$$

and σ_z^2 as defined earlier.

An unbiased estimate of the variance $V(\hat{Y}_{DIR})$ is obtained as

$$v(\hat{Y}_{DIR}) = \frac{N^2}{(n-r)(n-r-1)} \sum_{i=1}^{n-r}\left(\frac{y_i}{x_i} - \hat{R}\right)^2 \left[\frac{x'^2}{n'} - \left(1-\frac{n'}{N}\right)s_z^2\right] + \hat{R}^2N^2\left(\frac{1}{n'} - \frac{1}{N}\right)s_x^2, \quad \ldots(2.9)$$

where

$$s_x^2 = \frac{1}{n'-1}\sum_{i=1}^{n'}(x_i - x'/n')^2 \quad \ldots(2.10)$$

and $\hat{R}$ as defined earlier.

2.3 Simple Random Sampling Scheme

If the information on auxiliary variable is not used and a single random sample of size n is selected from the population with equal probabilities and without replacement, then for the comparison of this scheme with the double sampling procedure, cost aspect will have to be taken into account. It is because of the fact that in the later scheme, we are also collecting information on auxiliary variable from the preliminary sample.

If c' and c denote the unit costs of collecting information on auxiliary variable x and study variable y, respectively (c' will usually be much samaller than c), the total cost of double sampling procedure would be

$$C = c_o + c'n' + cn \quad \ldots(2.11)$$

where c_o is the overhead cost.

Now, if a single sample is taken (without using double sampling procedure) for observing y, the sample size for the same cost will be

$$n_o = n + \frac{c'n'}{c} \qquad \ldots(2.12)$$

With this single random sample of size n_o, when the information on y could not be collecled on r_o ($r_o=0, 1, \ldots, n_o-1$) units, we have the following three obvious results.

The estimate

$$\hat{\bar{Y}}_{SR_o} = \frac{\bar{N}}{n_o - r_o} \sum_{i=1}^{n_o - r_o} y_i \qquad \ldots(2.13)$$

is unbiased for the population totol Y.

The variance $V(\hat{\bar{Y}}_{SR_o})$ of the estimate $\hat{\bar{Y}}_{SR_o}$ is given by

$$V(\hat{\bar{Y}}_{SR_o}) = N^2\left[E\left(\frac{1}{n_o - r_o}\right) - \frac{1}{N}\right]S_y^2 \qquad \ldots(2.14)$$

where S_y^2 as defined earlier.

The variance $V(\hat{\bar{Y}}_{SR_o})$ is unbiasedly estimated by

$$N^2\left[\frac{1}{n_o - r_o} - \frac{1}{N}\right]s_r^2 \qquad \ldots(2.15)$$

where

$$s_r^2 = \frac{1}{n_o - r_o - 1}\left[\sum_{i=1}^{n_o - r_o} y_i^2 - \left(\sum_{i=1}^{n_o - r_o} y_i\right)^2/(n_o - r_o)\right]$$

3. The Relative Efficiency (R.E.)

In order to investigate the relative performance of the strategies proposed in sections 2.1, 2.2 and 2.3 we have the following two obvious results.

The R.E. of the estimate $\hat{Y}_{SR_o}$ with respect to the estimate $\hat{Y}_{DR}$ is given by

$$\text{R.E.}=\frac{\frac{(n'-1)\,\sigma_z^2}{n'(N-1)}E\left(\frac{1}{n-r}\right)+\frac{N-n'}{n'}S_y^2}{N\left[E\left(\frac{1}{n_o-r_o}\right)-\frac{1}{N}\right]S_y^2}. \qquad \ldots(3.1)$$

The R.E. of the estimate $\hat{Y}_{SR_o}$ with respect to the estimate $\hat{Y}_{DIR}$ is given by

$$\text{R.E.}=\frac{\left(\frac{N-n'}{n'}\right)NR^2S_x^2+\left[1+\left(\frac{1}{n'}-\frac{1}{N}\right)\frac{S_x^2}{\bar{X}^2}\right]\sigma_z^2E\left(\frac{1}{n-r}\right)}{N^2E\left[\left(\frac{1}{n_o-r_o}\right)-\frac{1}{N}\right]S_y^2} \qquad \ldots(3.2)$$

4. Non-Responess Robustness (NRR)

A desirable property for the estimate $\hat{\theta}(r)$ (for a given value of r) of the population parameter θ is that it should have smaller variance. 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, in this paper, we shall use variance of $V[\hat{\theta}(r)]$ over r, for this purpose.

Definition 4.1 The *NRR* of an estimate $\hat{\theta}(r)$ of the parameter θ is defined as

$$NRR\,[\hat{\theta}(r)]=[V\{\hat{\theta}(r)\}].$$

It is clear from this definition that $NRR\,[\hat{\theta}(r)]$ will always be positive and the minimum value that it can at least theoretically, take will be zero. Also smaller the value of $NRR[\hat{\theta}(r)]$, more robust the estimate will be. Thus if

$$V[\hat{\theta}(r)]=A+B.\,\alpha(r)$$

where A and B are functions of populotion parameters independent of r and $\alpha(r)$ is a function of r, then

$$NRR\,[\hat{\theta}(r)]=B^2V[\alpha(r)].$$

Thus from (2.2), (2.8) and (2.15) we have

$$NRR(\hat{Y}_{DR})=\left[\frac{N}{N-1}\cdot\frac{n'-1}{n}\sigma_z^2\right]^2 V\left(\frac{1}{n-r}\right), \qquad ...(4.1)$$

$$NRR(\hat{Y}_{DIR})=\left[\left(1+\frac{N-n'}{Nn'}\cdot\frac{S_x^2}{\bar{X}^2}\right)\sigma_z^2\right]^2 V\left(\frac{1}{n-r}\right), \qquad ...(4.2)$$

and

$$NRR(\hat{Y}_{SR_o})=N^4S_y^4V\left(\frac{1}{n_o-r_o}\right). \qquad ...(4.3)$$

5. The Relative Non-Response Robustness (RNRR)

The *RNRR* of the estimate $\hat{Y}_{SR_o}$ with respect to the estimate $\hat{Y}_{DR}$ is given by

$$RNRR\left[\frac{n'-1}{n'}\cdot\frac{\sigma_z^2}{N(N-1)S_y^2}\right]^2\frac{V\left(\frac{1}{n-r}\right)}{V\left(\frac{1}{n_o-r_o}\right)}. \qquad ...(5.1)$$

The *RNRR* of the estimate $\hat{Y}_{SR_o}$ with respect to the estimate $\hat{Y}_{DIR}$ is given by

$$RNRR=\frac{\left[1+\frac{N-n'}{Nn'}\cdot\frac{S_x^2}{\bar{X}^2}\right)\sigma_z^2\Big]^2 V\left(\frac{1}{n-r}\right)}{N^4S_y^4V\left(\frac{1}{n_o-r_o}\right)} \qquad ...(5.2)$$

where the symbols have their usual meaning.

6. An Empirical Investigation

For this purpose, we take the most likely distribution of r *i.e.*, the truncated binomial distribution. Thus taking

$$P(r)=\binom{n}{r}p^rq^{n-r}/(1-p^n),\ r=0,\ 1,\ \ldots,\ n-1, \qquad ...(6.1)$$

where p is the probability of random non-response and $q=1-p$.

It can be easily verified that

$$E(r)=np\ (1-p^{n-1})/(1-p^n),\ \text{and} \qquad ...(6.2)$$

$$E(r^2)=np\ (1-p+np-np^{n-1}). \qquad ...(6.3)$$

Now we try to find $E\left(\frac{1}{n-r}\right)$ correct upto the second order of approximation. Thus, we have

$$E\left(\frac{1}{n-r}\right)=\frac{1}{n}\left(1-\frac{r}{n}\right)^{-1},$$

$$=\frac{1}{n}E\left[1+\frac{r}{n}+\frac{r^2}{n^2}+\cdots\right], \qquad \ldots(6.4)$$

since $\frac{r}{n}<1$, and assuming that the trems involving powers of $\frac{r}{n}$ greater than two could be considered to be negligibly small, we get by substituting from (6.2) and (6.3) in (6.4) and after some algebraic samplifications

$$E\left(\frac{1}{n-r}\right)=\frac{1}{n}\left[1+\frac{np(1+p)+p(1-p)-2np^n}{n(1-p^n)}\right]. \qquad \ldots(6.5)$$

Further, to find $V\left(\frac{1}{n-r}\right)$, we can write

$$V\left(\frac{1}{n-1}\right)=E\left(\frac{1}{n-r}\right)^2-\left[E\left(\frac{1}{n-r}\right)\right]^2,$$

$$=\frac{1}{n^4}\left[E(r^2)-(E(r))^2\right].$$

Thus on using (6.2) and (6.3) and after algebraic simplification

$$V\left(\frac{1}{n-r}\right)=\frac{p(1-p)}{n^3(1-p^n)^2}\left[1-p^n-np^{n-1}(1-p)\right]. \qquad \ldots(6.6)$$

We assume further, both $r(r=0,1,\ldots,n-1)$ and $r_o(r_o=0,1,\ldots,n_o-1)$ follow the truncated binomial distribution with the same p. In practice this probability may differ from unit to unit but it will render the theoretical discussion quite complicated. This assumption of equal p-values is, therefore, taken for the sake of simplicity. With this assumption, one can write $E\left(\frac{1}{n_o-r_o}\right)$ and $V\left(\frac{1}{n_o-r_o}\right)$ by changing n to n_o and r to r_o, respectively in (6.5) and (6.6)

TABLE 1

R.E. and RNRR of the estimate $\hat{Y}_{SR_o}$ w.r.t. the estimate $\hat{Y}_{DR}$

Population number	n'	n	n_o	*Probability (P)*							
				.000	*.025*	*.050*		*.100*		*.300*	
				R.E.	*RNRR*	*R.E.*	*RNRR*	*R.E.*	*RNRR*	*R.E.*	*RNRR*
1	12	6	8	.83	.07	.78	.07	.73	.07	.56	.07
2	8	4	6	.98	.09	.91	.09	.84	.09	.62	.08
3	6	3	5	.81	.01	.73	.01	.65	.01	.41	.01
4	6	3	5	.80	.01	.72	.01	.64	.01	.40	.01
5	6	3	5	.84	.01	.75	.01	.57	.01	.43	.01
6	6	3	5	2.51	1.41	2.34	1.40	2.20	1.38	1.64	1.20
7	10	5	7	.53	.00	.48	.00	.43	.00	.28	.00
8	10	5	7	1.35	.34	1.26	.34	1.19	.34	.95	.33
9	10	5	7	.84	.06	.79	.06	.73	.06	.55	.06
10	7	4	6	.79	.00	.71	.00	.63	.00	.39	.00
11	7	4	6	.76	.00	.68	.00	.60	.00	.37	.00
12	12	6	8	1.25	.31	1.19	.31	1.12	.31	.91	.31

TABLE 2

R.E. and RNRR of the estimate $\hat{Y}_{SR_0}$ w.r.t. the estimate $\hat{Y}_{DIR}$

Population Number	n'	n	n_0	*Probabilily (P)*							
				.000	*.025*	*.050*		*.100*		*.300*	
				R.E.	*RNRR*	*R.E.*	*RNRR*	*R.E.*	*RNNR*	*R.E.*	*RNNR*
1	12	6	8	.77	.08	.72	.08	.68	.08	.52	.08
2	8	4	6	1.13	.10	1.05	.10	.07	.10	.70	.09
3	6	3	5	.59	.01	.54	.01	.48	.01	.32	.01
4	6	3	5	.84	.01	.75	.01	.68	.01	.42	.01
5	6	3	5	.70	.01	.63	.01	.57	.01	.37	.01
6	6	3	5	2.51	1.65	2.35	1.64	2.22	1.61	1.68	1.40
7	10	5	7	.53	,00	.48	.00	.44	.00	.29	.00
8	10	5	7	1.48	.41	1.39	.41	1.31	.41	1.04	.40
9	10	5	7	1.03	.08	.96	.07	.89	.07	.65	.07
10	7	4	6	.69	.00	.62	.00	.55	.00	.35	.00
11	7	4	6	.83	.00	.74	.00	.66	.00	.40	.00
12	12	6	8	1.02	.34	.98	.34	.93	.34	.72	.34

or the empirical study we substitute the value of $E\left(\frac{1}{1-u}\right)$ and $E\left(\frac{1}{n_o-r_o}\right)$ in (3.1) and (3.2) and $V\left(\frac{1}{n-r}\right)$ and $V\left(\frac{1}{n_o-r_o}\right)$ in (5.1) and (5.2) for the calculation of *R.E.* and *RNRR*. For this purpose we have used a wide variety of 25 natural populations considered by Rao and Singh [5]. For these populations results of the first 12 populations are given here which represent the whole study. N (the size of the population) ranges from 10 to 20. Preliminary sample size (n') and second smaller sample size (n) are taken 60 percent and 30 percent of N respectively. When a single random sample is taken for observing y, we assume that the cost components are such that they give the sample size (n_o) for the same cost as $n+2$. The probability (p) of random non-response ranges between 0 to 0.3.

Table 1 given below gives *R.E.* and *RNRR* of the estimate $\hat{Y}_{SR_o}$ with respect to the estimate $\hat{Y}_{DR}$. We observe that both *R.E.* and *RNRR* decrease with the increase in p. The rate of decrease in *R.E.* is decreasing whereas the rate of decrease in *RNRR* is increasing. For the other comparison, when the second smaller sample is independently selected from the population, Table 2 yields the same results. Thus to summarize, one can say that the double sampling estimates are less effected with increase in the random non-response than the estimate based on single random sample selected for observing y.

ACKNOWLEDGEMENT

The authors are grateful to the referees for their suggestions which considerably helped in improving the earlier draft of the paper.

REFERENCES

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

[2] Lahiri, D.B. (1951) : A method of sample selection providing unbiased ratio estimates ; *Bull. Inter. Stat. Inst.* **33(2)**, 133-140.

[3] Politz, A. and Simmons, W. (1949) : An attempt to get 'not-at-homes' into the sample without call-backs ; *Jour. Amer. Stat. Assoc.* **44**, 8-31.

[4] Raj, Des. (1968) : *Sampling Theory* Tata McGraw Hill Publishing Co., Bombay, New Delhi.

[5] Rao, J.N.K. and Singh, M.P. (1973) : On the choice of estimator in survey sampling ; *Aust. Jour. Stat.* **15(2)**, 95-104.

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

[7] Sukhatme, P.V. and Sukhatme, B.V. (1970) : *Sampling Theory of Surveys with Applications*, Indian Society of Agricultural Statistics New Delhi.