Manoj Kr. Chaudhary, V. K. Singh, Rajesh Singh
Department of Statistics, Banaras Hindu University
Varanasi-221005, INDIA

Florentin Smarandache
Department of Mathematics, University of New Mexico, Gallup, USA

On Some New Allocation Schemes in Stratified Random Sampling under Non-Response

Published in:
Rajesh Singh, F. Smarandache (Editors)
STUDIES IN SAMPLING TECHNIQUES AND TIME SERIES ANALYSIS
Zip Publishing, Columbus, USA, 2011
ISBN 978-1-59973-159-9
pp. 40 - 54

Abstract

This chapter presents the detailed discussion on the effect of non-response on the estimator of population mean in a frequently used design, namely, stratified random sampling. In this chapter, our aim is to discuss the existing allocation schemes in presence of non-response and to suggest some new allocation schemes utilizing the knowledge of response and non-response rates of different strata. The effects of proposed schemes on the sampling variance of the estimator have been discussed and compared with the usual allocation schemes, namely, proportional allocation and Neyman allocation in presence of non-response. The empirical study has also been carried out in support of the results.

Keywords: Stratified random sampling, Allocation schemes, Non-response, Mean squares, Empirical Study.

1. Introduction

Sukhatme (1935) has shown that by effectively using the optimum allocation in stratified sampling, estimates of the strata variances obtained in a previous survey or in a specially planned pilot survey based even on samples of moderate sample size would be adequate for increasing the precision of the estimator. Evans (1951) has also considered the problem of allocation based on estimates of strata variances obtained in earlier survey. According to literature of sampling theory, various efforts have been made to reduce the error which arises because of taking a part of the population, *i.e.*, sampling error. Besides the sampling error there are also several non-sampling errors which take place from time to time due to a number of factors such as faulty method of selection and estimation, incomplete coverage, difference in interviewers, lack of proper supervision, etc. Incompleteness or non-response in the form of absence, censoring or grouping is a troubling issue of many data sets.

In choosing the sample sizes from the different strata in stratified random sampling one can select it in such a way that it is either exclusively proportional to the strata sizes or proportional to strata sizes along with the variation in the strata under proportional allocation or Neyman allocation respectively. If non-response is inherent in the entire population and so are in all the strata, obviously it would be quite impossible to adopt Neyman allocation because then the knowledge of stratum variability will not be available, rather the knowledge of response rate of different strata might be easily available or might be easily estimated from the sample selected from each stratum. Thus, it is quite reasonable to utilize the response rate (or non-response rate) while allocating samples to stratum instead of Neyman allocation in presence of non-response error.

In the present chapter, we have proposed some new allocation schemes in selecting the samples from different strata based on response (non-response) rates of the strata in presence of non-response. We have compared them with Neyman and proportional allocations. The results have been shown with a numerical example.

2. Sampling Strategy and Estimation Procedure

In the study of non-response, according to one deterministic response model, it is generally assumed that the population is dichotomized in two strata; a response stratum considering of all units for which measurements would be obtained if the units happened to fall in the sample and a non-response stratum of units for which no measurement would be obtained. However, this division into two strata is, of course, an oversimplification of the problem. The theory involved in HH technique, is as given below:

Let us consider a sample of size n is drawn from a finite population of size N. Let n_1 units in the sample responded and n_2 units did not respond, so that $n_1 + n_2 = n$.

The n_1 units may be regarded as a sample from the response class and n_2 units as a sample from the non-response class belonging to the population. Let us assume that N_1 and N_2 be the number of units in the response stratum and non-response stratum respectively in the population. Obviously, N_1 and N_2 are not known but their unbiased estimates can be obtained from the sample as

$$\hat{N}_1 = n_1 N / n ; \;\; \hat{N}_2 = n_2 N / n .$$

Let m be the size of the sub-sample from n_2 non-respondents to be interviewed. Hansen and Hurwitz (1946) proposed an estimator to estimate the population mean $\overline{X}_0$ of the study variable X_0 as

$$T_{0HH} = \frac{n_1 \overline{x}_{01} + n_2 \overline{x}_{0m}}{n} , \tag{2.1}$$

which is unbiased for $\overline{X}_0$, whereas $\overline{x}_{01}$ and $\overline{x}_{0m}$ are sample means based on samples of sizes n_1 and m respectively for the study variable X_0.

The variance of T_{0HH} is given by

$$V(T_{0HH}) = \left[\frac{1}{n} - \frac{1}{N}\right] S_0^2 + \frac{L-1}{n} W_2 S_{02}^2 , \tag{2.2}$$

where $L = \frac{n_2}{m}$, $W_2 = \frac{N_2}{N}$, S_0^2 and S_{02}^2 are the mean squares of entire group and non-response group respectively in the population.

Let us consider a population consisting of N units divided into k strata. Let the size of i^{th} stratum is N_i, ($i = 1,2,...,k$) and we decide to select a sample of size n from the entire population in such a way that n_i units are selected from the i^{th} stratum. Thus, we have $\sum_{i=1}^{k} n_i = n$.

Let the non-response occurs in each stratum. Then using Hansen and Hurwitz procedure we select a sample of size m_i units out of n_{i2} non-respondent units in the i^{th} stratum with the help of simple random sampling without replacement (SRSWOR) such that $n_{i2} = L_i m_i$, $L_i \geq 1$ and the information are observed on all the m_i units by interview method.

The Hansen-Hurwitz estimator of population mean $\overline{X}_{0i}$ for the i^{th} stratum will be

$$T_{0i}^{*} = \frac{n_{i1}\overline{x}_{0i1} + n_{i2}\overline{x}_{0mi}}{n_i}, \qquad (i = 1,2,...,k) \tag{2.3}$$

where $\overline{x}_{0i1}$ and $\overline{x}_{0mi}$ are the sample means based on n_{i1} respondent units and m_i non-respondent units respectively in the i^{th} stratum.

Obviously T_{0i}^{*} is an unbiased estimator of $\overline{X}_{0i}$. Combining the estimators over all strata we get the estimator of population mean $\overline{X}_0$, given by

$$T_{0st}^{*} = \sum_{i=1}^{k} p_i T_{0i}^{*} \tag{2.4}$$

where $p_i = \frac{N_i}{N}$.

Obviously, we have

$$\mathrm{E}\left[T_{0st}^{*}\right] = \overline{X}_0. \tag{2.5}$$

The variance of T_{0st}^{*} is given by

$$V\left[T_{0st}^{*}\right] = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)p_i^{\,2} S_{0i}^{\;2} + \sum_{i=1}^{k}\frac{(L_i - 1)}{n_i} W_{i2}\, p_i^{\,2} S_{0i2}^{2} \tag{2.6}$$

where $W_{i2} = \dfrac{N_{i2}}{N_i}$, ${S_{0i}}^2$ and S_{0i2}^2 are the mean squares of entire group and non-response group respectively in the i^{th} stratum.

It is easy to see that under 'proportional allocation' (PA), that is, when $n_i = np_i$ for $i = 1,2,...,k$, $V\left[T_{0st}^*\right]$ is obtained as

$$V\left[T_{0st}^*\right]_{PA} = \sum_{i=1}^{k}\left(\frac{1}{n} - \frac{1}{N}\right)p_i {S_{0i}}^2 + \frac{1}{n}\sum_{i=1}^{k}(L_i - 1)W_{i2}p_i S_{0i2}^2, \tag{2.7}$$

whereas under the 'Neyman allocation' (NA), with $n_i = \dfrac{np_i S_{0i}}{\sum\limits_{i=1}^{k} p_i S_{0i}}$ $\left(i = 1,2,...,k\right)$, it is equal to

$$V\left[T_{0st}^*\right]_{NA} = \frac{1}{n}\left(\sum_{i=1}^{k} p_i S_{0i}\right)^2 - \frac{1}{N}\sum_{i=1}^{k} p_i {S_{0i}}^2 + \frac{1}{n}\left(\sum_{i=1}^{k}(L_i - 1)W_{i2}p_i \frac{S_{0i2}^2}{S_{0i}}\right)\left(\sum_{i=1}^{k} p_i S_{0i}\right). \tag{2.8}$$

It is important to mention here that the last terms in the expressions (2.7) and (2.8) arise due to non-response in the population. Further, in presence of non-response in the population, Neyman allocation may or may not be efficient than the proportional allocation, a situation which is quite contrary to the usual case when population is free from non-response. This can be understood from the following:

We have

$$V\left[T_{0st}^*\right]_{PA} - V\left[T_{0st}^*\right]_{NA} = \frac{1}{n}\sum_{i=1}^{k} p_i\left(S_{0i} - \overline{S}_w\right)^2 + \frac{1}{n}\sum_{i=1}^{k}\left(L_i - 1\right)p_i W_{i2} S_{0i2}^2\left(1 - \frac{\overline{S}_w}{S_{0i}}\right) \tag{2.9}$$

$$\overline{S}_w = \sum_{i=1}^{k} p_i S_{0i}\ .$$

Whole the first term in the above expression is necessarily positive, the second term may be negative and greater than the first term in magnitude depending upon the sign and magnitude of the term $\left(1-\frac{\overline{S}_w}{S_{0i}}\right)$for all i. Thus, in presence of non-response in the stratified population, Neyman allocation does not always guarantee a better result as it is case when the population is free from non-response error.

3. Some New Allocation Schemes

It is a well known fact that in case the stratified population does not have non-response error and strata mean squares, $S_{0i}^2 \; (i=1,2,...,k)$, are known, it is always advisable to prefer Neyman allocation scheme as compared to proportional allocation scheme in order to increase the precision of the estimator. But, if the population is affected by non-response, Neyman allocation is not always a better proposition. This has been highlighted under the section 2 above. Moreover, in case non-response is present in strata, knowledge on strata mean squares,S_{0i}^2, are impossible to collect, rather direct estimates of S_{0i1}^2 and S_{0i2}^2 may be had from the sample. Under these circumstances, it is, therefore, practically difficult to adopt Neyman allocation if non-response is inherent in the population. However, proportional allocation does not demand the knowledge of strata mean squares and rests only upon the strata sizes, hence it is well applicable even in the presence of non-response.

As discussed in the section 2, unbiased estimates of response and non-response rates in the population are readily available and hence it seems quite reasonable to think for developing allocation schemes which involve the knowledge of population response (non-response) rates in each stratum. If such allocation schemes yield précised estimates as compared to proportional allocation, these would be advisable to adopt instead of Neyman allocation due to the reasons mentioned above.

In this section, we have, therefore, proposed some new allocation schemes which utilize the knowledge of response (non-response) rates in subpopulations. While some of the proposed schemes do not utilize the knowledge ofS_{0i}^2, some others are proposed based on the knowledge of S_{0i}^2 just in order to make a comparison of them with Neyman allocation under the presence of non-response. In addition to the assumptions of proportional and Neyman allocations, we have further assume it logical to allocate larger sample from a stratum having larger number of respondents and vice-versa when proposing the new schemes of allocations.

Scheme-1[OA (1)]:

Let us assume that larger size sample is selected from a larger size stratum and with larger response rate, that is,

$$n_i \propto p_i W_{i1} \qquad \text{for} \qquad i = 1,2,...,k\,.$$

Then we have

$$n_i = K p_i W_{i1} \quad \text{where } K \text{ is a constant.}$$

The value of K will be

$$K = \frac{n}{\sum_{i=1}^{k} p_i W_{i1}}\,.$$

Thus we have

$$n_i = \frac{n p_i W_{i1}}{\sum_{i=1}^{k} p_i W_{i1}}\,. \tag{3.1}$$

Putting this value of n_i in expression (2.6), we get

$$V\left[T_{0st}^{*}\right]_1 = \frac{1}{n}\left[\sum_{i=1}^{k} p_i W_{i1}\right]\left[\sum_{i=1}^{k}\left\{\frac{p_i S_{0i}^2}{W_{i1}} + \frac{(L_i - 1)}{W_{i1}} W_{i2} p_i S_{0i2}^2\right\}\right] - \frac{1}{N}\sum_{i=1}^{k} p_i S_{0i}^2 \tag{3.2}$$

Scheme-2[OA (2)]:

Let us assume that

$$n_i \propto p_i W_{i1} S_{0i} .$$

Then, we have

$$n_i = \frac{n p_i W_{i1} S_{0i}}{\sum_{i=1}^{k} p_i W_{i1} S_{0i}} \tag{3.3}$$

and hence the expression (2.6) becomes

$$V\left[T_{0st}^*\right]_2 = \frac{1}{n}\left[\sum_{i=1}^{k} p_i W_{i1} S_{0i}\right]\left[\sum_{i=1}^{k}\left\{\frac{p_i S_{0i}}{W_{i1}} + \frac{(L_i - 1) W_{i2} p_i}{W_{i1}} \frac{S_{0i2}^2}{S_{0i}}\right\}\right] - \frac{1}{N}\sum_{i=1}^{k} p_i S_{0i}^2 . \tag{3.4}$$

Scheme-3[OA (3)]:

Let us select larger size sample from a larger size stratum but smaller size sample if the non-response rate is high. That is,

$$n_i \propto \frac{p_i}{W_{i2}} .$$

Then

$$n_i = \frac{n p_i}{W_{i2} \sum_{i=1}^{k} \frac{p_i}{W_{i2}}} \tag{3.5}$$

and the expression of $V\left[T_{0st}^*\right]$ reduces to

$$V\left[T_{0st}^*\right]_3 = \frac{1}{n}\left[\sum_{i=1}^{k} \frac{p_i}{W_{i2}}\right]\left[\sum_{i=1}^{k}\left\{p_i W_{i2} S_{0i}^2 + \left(L_i - 1\right) W_{i2}^2 p_i S_{0i2}^2\right\}\right] - \frac{1}{N}\sum_{i=1}^{k} p_i S_{0i}^2 . \tag{3.6}$$

Scheme-4[OA (4)]:

Let

$$n_i \propto \frac{p_i S_{0i}}{W_{i2}}, \text{ then}$$

$$n_i = \frac{n p_i S_{0i}}{W_{i2} \sum_{i=1}^{k} \frac{p_i S_{0i}}{W_{i2}}}. \tag{3.7}$$

The corresponding expression of $V\left[T_{0st}^{*}\right]$ is

$$V\left[T_{0st}^{*}\right]_4 = \frac{1}{n}\left[\sum_{i=1}^{k} \frac{p_i S_{0i}}{W_{i2}}\right]\left[\sum_{i=1}^{k}\left\{p_i W_{i2} S_{0i} + \left(L_i - 1\right) W_{i2}^2 p_i \frac{S_{0i2}^2}{S_{0i}}\right\}\right] - \frac{1}{N}\sum_{i=1}^{k} p_i S_{0i}^2 . \tag{3.8}$$

Scheme-5[OA (5)]:

Let

$$n_i \propto \frac{p_i W_{i1}}{W_{i2}},$$

then

$$n_i = \frac{n p_i W_{i1}}{W_{i2} \sum_{i=1}^{k} \frac{p_i W_{i1}}{W_{i2}}}. \tag{3.9}$$

The expression (2.6) gives

$$V\left[T_{0st}^{*}\right]_5 = \frac{1}{n}\left[\sum_{i=1}^{k} \frac{p_i W_{i1}}{W_{i2}}\right]\left[\sum_{i=1}^{k}\left\{\frac{p_i W_{i2} S_{0i}^2}{W_{i1}} + \frac{\left(L_i - 1\right) W_{i2}^2 p_i S_{0i2}^2}{W_{i1}}\right\}\right] - \frac{1}{N}\sum_{i=1}^{k} p_i S_{0i}^2 . \tag{3.10}$$

Scheme-6[OA (6)]:

If $\quad n_i \propto \dfrac{p_i W_{i1} S_{0i}}{W_{i2}}$,

then, we have

$$n_i = \frac{n p_i W_{i1} S_{0i}}{W_{i2} \sum_{i=1}^{k} \frac{p_i W_{i1} S_{0i}}{W_{i2}}}. \tag{3.11}$$

In this case, $V\left[T_{0st}^{*}\right]$ becomes

$$V\left[T_{0st}^{*}\right]_6 = \frac{1}{n}\left[\sum_{i=1}^{k} \frac{p_i W_{i1} S_{0i}}{W_{i2}}\right]\left[\sum_{i=1}^{k}\left\{\frac{p_i W_{i2} S_{0i}}{W_{i1}} + \frac{(L_i - 1) W_{i2}^{2} p_i S_{0i2}^{2}}{W_{i1} S_{0i}}\right\}\right] - \frac{1}{N}\sum_{i=1}^{k} p_i S_{0i}^{2}. \tag{3.12}$$

Remark 1: It is to be mentioned here that if response rate assumes same value in all the strata, that is $W_{i1} = W$ (say), then schemes 1, 3 and 5 reduces to 'proportional allocation', while the schemes 2, 4 and 6 reduces to 'Neyman allocation'. The corresponding expressions, $V\left[T_{0st}^{*}\right]_r$, $(r = 1,3,5)$ are then similar to $V\left[T_{0st}^{*}\right]_{PA}$ and $V\left[T_{0st}^{*}\right]_r$, $(r = 2,4,6)$ reduce to $V\left[T_{0st}^{*}\right]_{NA}$.

Remark 2: Although the theoretical comparison of expressions of $V\left[T_{0st}^{*}\right]_r$, $(r = 1,3,5)$ and $V\left[T_{0st}^{*}\right]_r$, $(r = 2,4,6)$ with $V\left[T_{0st}^{*}\right]_{PA}$ and $V\left[T_{0st}^{*}\right]_{NA}$ respectively is required in order to understand the suitability of the proposed schemes, but such comparisons do not yield explicit solutions in general. The suitability of a scheme does depend upon the parametric values of the population. We have, therefore, illustrated the results with the help of some empirical data.

4. Empirical Study

In order to investigate the efficiency of the estimator T_{0st}^{*} under proposed allocation schemes, based on response (non-response) rates, we have considered here an empirical data set.

We have taken the data available in Sarndal et. al. (1992) given in Appendix B. The data refer to 284 municipalities in Sweden, varying considerably in size and other characteristics. The population consisting of the 284 municipalities is referred to as the MU284 population.

For the purpose of illustration, we have randomly divided the 284 municipalities into four strata consisting of 73, 70, 97 and 44 municipalities. The 1985 population (in thousands) has been considered as the study variable, X_0.

On the basis of the data, the following values of parameters were obtained:

Table 1 : Particulars of the Data

$(N = 284)$

Stratum (i)	**Size** (N_i)	**Stratum Mean** $(\overline{X}_{0i})$	**Stratum Mean Square** (S_{0i}^2)	**Mean Square of the Non-response Group** $(S_{0i2}^2) = \frac{4}{5} S_{0i}^2$
1	**73**	**40.85**	**6369.10**	**5095.28**
2	**70**	**27.83**	**1051.07**	**840.86**
3	**97**	**25.78**	**2014.97**	**1611.97**
4	**44**	**20.64**	**538.47**	**430.78**

We have taken sample size, $n = 60$.

Tables 2 depicts the values of sample sizes, n_i $(i=1,2,3,4)$ and values of $V\left[T_{0st}^{*}\right]$ under PA, NA and proposed schemes OA(1) to OA(6) for different selections of the values of L_i and W_{i2} $(i=1,2,3,4)$.

Table 2

Sample Sizes and Variance of T^*_{0st} under Different Allocation Schemes

(L_i =2.0, 2.5, 1.5, 3.5 for i = 1, 2, 3, 4 respectively)

Stratum	Non-response Rate $\left(W_{i2}\right)$ (Percent)	Sample Size $\left(n_i\right)$ and $V\left[T^*_{0st}\right]$ under															
		PA		NA		OA(1)		OA(2)		OA(3)		OA(4)		OA(5)		OA(6)	
		n_i	$V\left[T^*_{0st}\right]$	n_i	$V\left[T^*_{0st}\right]$	n_i	$V\left[T^*_{0st}\right]$	n_i	$V\left[T^*_{0st}\right]$	n_i	$V\left[T^*_{0st}\right]$	n_i	$V\left[T^*_{0st}\right]$	n_i	$V\left[T^*_{0st}\right]$	n_i	$V\left[T^*_{0st}\right]$
1	20	15	43.08	26	36.04	17	41.02	28	116.59	20	38.43	31	38.43	22	37.85	33	40.25
2	25	15		10		15		10		15		10		15		10	
3	30	21		19		20		18		18		16		17		14	
4	35	9		5		8		4		7		3		6		3	
1	35	15	45.97	26	37.27	14	49.17	24	117.37	12	55.41	21	39.07	10	60.76	19	40.72
2	30	15		10		14		10		13		10		13		10	
3	25	21		19		21		21		22		22		23		24	
4	20	9		5		10		5		13		7		14		7	
1	25	15	43.91	26	36.30	16	43.40	27	116.54	16	44.15	27	37.76	16	44.69	27	38.94
2	20	15		10		16		11		19		13		21		14	
3	30	21		19		20		18		18		17		17		16	
4	35	9		5		8		4		7		3		6		3	
1	20	15	43.17	26	35.99	17	41.32	28	115.40	20	39.45	32	38.82	22	39.73	34	41.30
2	25	15		10		15		10		16		10		16		10	
3	35	21		19		19		17		16		14		14		12	
4	30	9		5		9		5		8		4		8		4	

5. Concluding Remarks

In the present chapter, our aim was to accommodate the non-response error inherent in the stratified population during the estimation procedure and hence to suggest some new allocation schemes which utilize the knowledge of response (non-response) rates of strata. As discussed in different sub-sections, Neyman allocation may sometimes produce less précised estimates of population mean in comparison to proportional allocation if non-response is present in the population. Moreover, Neyman allocation is sometimes impractical in such situation, since then neither the knowledge of S_{0i} $(i = 1,2,3,4)$, the mean squares of the strata, will be available, nor these could be estimated easily from the sample. In contrast to this, what might be easily known or could be estimated from the sample are response (non-response) rates of different strata. It was, therefore, thought to propose some new allocation schemes depending upon response (non-response) rates.

A look of Table 2 reveals that in most of the situations (under different combinations of W_{i2} and L_i), allocation schemes OA (1), OA (3) and OA (5), depending solely upon the knowledge of p_i and W_{i2} (or W_{i1}), produce more précised estimates as compared to PA. Further, as for as a comparative study of schemes OA (1), OA (3) and OA (5) is concerned, no doubt, all these schemes are more or less similar in terms of their efficiency. Thus, in addition to the knowledge of strata sizes, p_i, the knowledge of response (non-response) rates, W_{i1} (or W_{i2}), while allocating sample to different strata; certainly adds to the precision of the estimate.

It is also evident from the table that the additional information on the mean squares of strata certainly adds to the precision of the estimate, but this contribution is not very much significant in comparison to NA. Scheme OA (2) is throughout worse than any other scheme.

References

Evans, W. D. (1951) : On stratification and optimum allocation. Journ. of The Amer. Stat. Assoc., 46, 95-104.

Hansen, M. H. and Hurwitz, W. N. (1946) : The problem of non-response in sample Surveys. Journ. of The Amer. Stat. Assoc., 41, 517-529

Sarndal, C. E., Swensson, B. and Wretman, J. (1992) : Model Assisted Survey Sampling. Springer-Verlag, New York, Inc.

Sukhatme, P. V. (1935) : Contributions to the theory of the representative Method. Journ. of the Royal Stat. Soc., 2, 253-268.