Developments in Data Analysis
A. Ferligoj and A. Kramberger (Editors)
Metodološki zvezki, 12, Ljubljana: FDV, 1996

Variable Weights for Unit Non-response

Vasja Vehovar*

Abstract

It is common in multistage samples to treat the unit non-response at the level of primary sampling units. For example, with b_{ri} respondents out of b_i eligible units in the i-th primary sampling unit there is a weight $w_i = b_i/b_{ri}$ attached to the respondents. However, in repeated samples these weights are variable quantities as opposed to fixed weights arising from oversampling strata. In the latter case, a simple formulae for the increase in variance due to weighting can be applied (Kish, 1965). It is shown in the paper that with variable weights the increase in variance due to variable weights is consistently smaller compared to the situation with fixed weights.

1 Introduction

There exist different situations where weights are introduced in sample surveys. Kish (1992) cites the following reasons for the use of weights:

- disproportional sampling fraction;
- disproportional allocation to domains;
- frame problems;
- non-response;
- statistical adjustments;
- adjustments to match controls;
- combining samples.

The weights are generally assumed to be fixed. Whenever a unit is selected in a sample it is thus supposed to receive the same (fixed) weight. For example, if the household is selected proportional to its size, the corresponding weight applied[1] will be the same in all samples that include the household.

*Faculty of Social Sciences, University of Ljubljana, P.O. Box 47, 1 109 Ljubljana, Slovenia

[1]Weight has to compensate the distorted inclusion probability because a larger household has a larger probability of inclusion. The weight is thus proportional to the inverse of the household size.

However, the weights are not always fixed. First, let us observe the poststratification weights.

The postratification weights are constructed for matching the sample with the known population structure. In repeated samples these weights vary together with the sample structure and the sample estimates (Verma, 1993:106). The effect of this variation will be a small increase in sampling variance. However, in certain situations the effect may be quite the opposite. Rust (1987) reports that the increase in variance due to post-stratification weights was five times smaller compared to the increase under the (wrong) assumption of fixed weights. Obviously, the increase in variance cannot be treated by the method developed for the fixed weights.

Another situation with variable weights occurs with non-response weighting in a two stage cluster sample design. There, the non-response weights can be constructed at the level of primary sampling units (clusters) as the inverse of response rate within each cluster[2]. Of course, such weights vary from sample to sample since in repeated surveys the response rate varies within each cluster.

In this paper the increase in sampling variance arising from the above described weights will be discussed for the estimate $\bar{y} = \sum_{i=1}^{n} y_i/n$ of the population average $\bar{Y} = \sum_{i=1}^{N} Y_i/N$. At the beginning the issue of sampling variance (2) and non-response (3) will be introduced. Next, the increase in sampling variance will be treated for variable weights (4) and also under assumption of fixed weights (5). The results will be illustrated with a simulation study (6). After discussion (7) the conclusions (8) will be presented.

2 Sampling variance

In complex designs the sampling variance of the estimator $\hat{\theta}$ for the population parameter θ can be expressed as:

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Deff} * \mathrm{Var}_{SRS}(\hat{\theta}), \tag{1}$$

where Deff (design effect) stands for the effect of the sample design and $\mathrm{Var}_{SRS}(\hat{\theta})$ stands for the sampling variance in the case of simple random sample (SRS) of the same size. In a special case ($\theta \equiv \bar{Y}$) of population average we have[3] a well known expression:

$$\mathrm{Var}_{\mathrm{SRS}}(\bar{y}) \doteq \frac{\sigma_y^2}{n}. \tag{2}$$

In a two stage cluster sample the sampling variance (1) can be written in a much more clear analytical form. With clusters of equal size and with uniform sampling rates at both stages we have the classical example (Kish, 1965: 170; Cochran, 1978: 227):

$$\mathrm{Var}(\bar{y}) = (1 - \frac{a}{A})\frac{1}{a}\sum_{i=1}^{A}\frac{(\bar{Y}_i - \bar{Y})^2}{A-1} + (1 - \frac{b}{B})\frac{1}{ab}\sum_{i=1}^{A}\sum_{j=1}^{B}\frac{(Y_{ij} - \bar{Y}_i)^2}{A(B-1)}, \tag{3}$$

[2]There is an implicit assumption that within each cluster the non-responding units are missing completely at random (MCAR). We will thus assume there is no non-response bias within clusters.

[3]Ignoring finite population correction term.

where a and A are the numbers of clusters in the sample and in the population, b is the sample size of the cluster, B is the population size of the cluster, Y_{ij} denotes population value of the j-th unit in the i-th cluster and $\bar{Y}_i$ stands for the (population) cluster average.

More generally, in two stage cluster designs, the common estimators of the population average - including HT estimator, ratio estimator and PPZ[4] estimator (Cochran, 1978, chp. 10, 11) - can be written in the following form:

$$\mathrm{Var}(\bar{y}) = (1 - \frac{a}{A})\frac{1}{a}U + \frac{1}{Aa}\sum_{i=1}^{A}\Big(1 - \frac{b_i}{B_i}\Big)\frac{1}{b_i}V_i, \tag{4}$$

where U and V_i are quantities depending only on the population values. The label b_i stands for the sample size of the i-th cluster and B_i stands for its population size.

3 Non-response

In this section the sampling variance in the presence of non-response is discussed. Let us denote the initial sample size and the number of responding units with n and n_r respectively. The number of responding units in the i-th cluster is denoted with b_{ri}.

We assume a uniform non-response rate $\bar{R}$ with $\mathrm{E}(n_r) = \bar{R}n$. To obtain n respondents we have to start with a larger initial sample $n^* = n/\bar{R}$. Similarly we have $b_i^* = b_i/\bar{R}$. Only then do we obtain $\mathrm{E}(n_r^*) = n$ and $\mathrm{E}(b_{ri}^*) = b_i$.

Of course, when we compare the sample where there is no non-response with the sample where non-response has occurred, the initial sample sizes must differ. For the purpose of our comparisons the initial sample sizes will be n and n^* respectively.

First, let us consider the simplest case (3). If we start with the initial sample size n^* and apply the uniform non-response ($\bar{R}$), we obtain conditional variance:

$$\mathrm{Var}(\bar{y}_r^*) = (1 - \frac{a}{A})\frac{1}{a}\sum_{i=1}^{A}\frac{(\bar{Y}_{ri} - \bar{Y}_r)^2}{A-1} + \frac{1}{Aa}\sum_{i=1}^{A}\sum_{j=1}^{B_{ri}}\frac{(1 - \frac{b_{ri}^*}{B_{ri}})(Y_{rij} - \bar{Y}_{ri})^2}{(B_{ri}-1)b_{ri}^*}. \tag{5}$$

Here, the additional label "r" refers to the responding units. For example, label B_{ri} stands for the number of responding units in i-th population cluster.

Assuming MCAR property for the non-response mechanism within each cluster the expression (5) differs from the expression (3) in the variable term b_{ri}^* which has moved into the summation symbol.

We will further assume that the term $(1 - a/A)$ is negligible and the terms b_{ri}^* are positive. Due to MCAR assumption the term V_i in (4) remains unchanged when there is a non-response. Thus, with a uniform non-response rate ($\bar{R}$) and initial sample n^*, the expression (4) takes a general form which is conditional on b_{ri}^*:

$$\mathrm{Var}(\bar{y}_r^*) = \frac{1}{a}U + \frac{1}{Aa}\sum_{i=1}^{A}\Big(1 - \frac{b_{ri}^*}{B_{ri}}\Big)\frac{1}{b_{ri}^*}V_i. \tag{6}$$

[4] PPZ (probability proportional to Z) estimator is used when probability of selection for the unit i is proportional to z_i.

When there is a non-response the variance (4) thus takes the conditional variance form (6). The variance (6) is, in fact, the proper population value of the conditional variance when correct inclusion probabilities are taken into account.

To estimate the sampling variance (6) the weights must be attached to the respondents n_r. The weights needed to compensate for the within-cluster non-response are proportional to $w_i = b_i/b^*_{ri}$ with $\mathrm{E}(b^*_{ri}) = b_i$. These weights are based on correct inclusion probabilities. So only[5] with these weights will the variance estimation programmes correctly estimate the corresponding population value of (5) or (6).

4 Variable weights

The aim of this section is to compare the increase in variance due to non-response weights (6) with the situation where there is no non-response (4).

First, we have to calculate the unconditional variance $\mathrm{Var}_U(\bar{y}^*_r)$. Since $\mathrm{E}(\bar{y}^*_r) = \bar{Y}$ is fixed, we have $\mathrm{VarE}(\bar{y}^*_r) = 0$. Thus, only expected value of (6) need to be considered, i.e. $\mathrm{EVar}(\bar{y}^*_r)$. The expected value of the first term in (6) is a fixed quantity, but the second term varies from sample to sample. Its variation is based on the variability of the actual *take* b^*_{ri} per cluster.

In the simple, but realistic case, where V_i and b^*_{ri} are independent, factor $(1 - b^*_{ri}/B_{ri})$ is negligible or constant, and the non-response mechanism is a uniform Bernoulli mechanism with parameter $\bar{R}$, the increase in expected value of the variance (6) over variance (4) is based on:

$$\mathrm{E}\left(\frac{1}{b^*_{ri}}\right) / \left(\frac{1}{\mathrm{E}(b^*_{ri})}\right) = \mathrm{E}\left(\frac{1}{b^*_{ri}}\right) / \left(\frac{1}{b^*_{ri}}\right) = b_i \mathrm{E}\left(\frac{1}{b^*_{ri}}\right). \tag{7}$$

The ratio (7) compares the population value of the second term in variance (6) of the properly weighted sample of respondents[6] with the corresponding variance (4) of the sample without non-response.

In Table 1, the increase (7) is illustrated. The calculations are based on a truncated[7] hypergeometric distribution for the simplest self-weighted case, $b^*_i = b^*$, with population cluster size $B_i = B = 1000$. In the case of $B = 100$ the figures in Table 1 would be roughly 10% lower. For large B_i, the approximation with binomial distribution and Taylor linearization can be used (Cochran, 1978:135).

The brackets in Table 1 indicate that more than 1% of the clusters were omitted (truncation) because no unit in the cluster responded.

Of course, the above increase refers only to the second component in (6), i.e. the within variance component. The proportion of the within variance component can be expressed as a function of the intracluster correlation ρ and the *actual* size of the cluster (e.g., $\mathrm{E}(b^*_{ri}) = b$). In the special case (3) of equal clusters, constant sampling rates within clusters and sampling without replacement at both stages, we can use the well-known relations (Kish, 1965:166) to obtain approximations in Table 2.

[5] We are speaking, of course, about a two stage sample designs assuming MCAR property within clusters.

[6] Again, the initial sample size here equals n^*.

[7] Truncation was done for clusters with $b^*_{ri} = 0$.

Table 1: Increase in within variance (%) at $B = 1000$

b^*	$\bar{R}$ - non-response rate				
	0.1	**0.2**	**0.3**	**0.4**	**0.5**
3	5.7	11.2	(16.0)	(18.2)	(18.2)
4	3.9	9.2	14.9	(19.4)	(21.7)
5	2.8	6.9	12.3	18.0	(22.4)
10	1.3	2.6	4.7	8.0	13.0
15	0.7	1.6	2.7	4.4	7.1
30	0.3	0.6	1.1	1.7	2.5

Table 2: Proportion of the within variance

b	ρ - intracluster correlation					
	0.005	**0.01**	**0.02**	**0.05**	**0.10**	**0.20**
3	0.99	0.97	0.94	0.86	0.75	0.57
4	0.98	0.96	0.93	0.83	0.70	0.51
5	0.98	0.95	0.91	0.79	0.64	0.44
10	0.95	0.91	0.83	0.66	0.47	0.29
15	0.93	0.87	0.77	0.56	0.38	0.21
30	0.87	0.77	0.62	0.39	0.23	0.12

The increase in sampling variance due to variable weights can be obtained by multiplying the corresponding cells in Tables 1 and 2. It is obvious that, in general, the increase will be relatively small. We can conclude that there may exist some special situations (i.e. small b, small ρ, large $\bar{R}$) with a noticeable increase in variance, but the increase tends to be small.

As an approximation the results from Table 1 and Table 2 can be used also for other sampling strategies within a two-stage sampling scheme.

5 Fixed weights

The above discussed increase in variance differs from a more common situation with fixed weights where the approximation:

$$\widehat{\text{VIF}} = \sum_{i=1}^{n} w_k^2 \Big/ \Big(\sum_{i=1}^{n} w_k\Big)^2 = 1 + \text{relvar}(w) = 1 + \widehat{\text{CV}}_w^2 \tag{8}$$

can be used (Kish, 1965:427). The expression refers to the weights which typically arise from oversampling strata. There, each unit receives a fixed weight which is the same in the repeated samples. The label CV_w^2 in expression (8) stands for the square of the element coefficient of variation of the weights ($CV_w = \sigma_w/\bar{w}$).

Table 3 illustrates the increase in variance based on expression (8). Again, the weights here discussed arise from the same non-response adjustments as in the previous section, but they were treated as fixed weights. The figures from Table 3 can thus be compared with the increase based on Table 1 and Table 2. Since different principles are used, the results also differ.

Table 3: Increase (%) in variance based on (VIF-1)

	$\bar{R}$ - non-response rate				
b	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
3	10.2	17.6	(19.4)	(17.6)	(16.0)
4	6.8	16.8	22.1	(24.0)	(22.0)
5	4.0	12.5	21.1	27.0	(26.5)
10	1.5	3.6	7.3	14.4	24.0
15	0.1	2.0	4.0	6.8	12.5
30	0.0	01.0	1.7	2.7	4.0

We can observe that the increase in Table 3 is higher compared to the increase based on Tables 1 and 2. However, in the majority of practical situations the increase is small in both cases, the differences thus being negligible.

6 Simulations

The above results were studied in a simulation study.[8] The study variable has a normal distribution $N : (700, 300^2)$ with population average $\bar{Y} = 700$ and population variance $S^2 = 90,000$. We assumed[9] an intracluster correlation $\rho = 0.07$ and $B_i = B = 1000$.

The basic design used in simulation was close to the sample design for a national social survey ($n = 2,100$) with $a = 140$ primary sampling units. Of course, with non-response $\bar{R} = 3/18 = 0.17$ the initial sample size in simulation must be $140 \times 18 = 2,520$, which is comparable (after non-response rate $\bar{R} = 3/18$) to the design $140 \times 15 = 2,100$ with no non-response.

Also used in simulation was an extreme design with $a = 700$, $b^* = 5$ (e.g.: $700 \times 5 = 3,500$) and $\bar{R} = 0.4$ resulting (approximately) in a sample of $700 \times 3 = 2,100$.

The simulations were performed in the following three steps:

[8]The programme was written in S-PLUS 3.1 software.

[9]This is close to the monthly income of the working population (variable Y) in two stage cluster sample of the national General Social Survey.

1. The initial sample was generated by the use of standard formulae for a two-stage cluster sample (Kish, 1965: 167). First, the primary sampling units (clusters) were generated using the between variance component $S_b^2 \approx (S^2 * \rho)$ and $\bar{y}_i : N(\bar{Y}, S_b^2/a)$. Next, the individual values were generated within each cluster. In each cluster, the parameters from the first step were used together with the within variance component: $S_w^2 \approx (S^2 - S_b^2)$ and $y_{ij} : N(\bar{y}_i, S_w^2/b_i^*)$ or $y_{ij} : N(\bar{y}_i, S_w^2/b_i)$.

2. The missing data were generated as a uniform Bernoulli mechanism with the parameter $\bar{R} = 0.17$ or $\bar{R} = 0.40$.

3. The non-response weights were attached at the cluster level as the inverse of the response rate within clusters.

Within a simulation 5,000 samples were generated. Each time the (ratio) estimate of the population average was calculated. The simulation was performed for the following two situations:

- no non-response and no adjustments,
- the non-response weighting adjustment at the cluster level.

The sampling variance based on 5,000 samples was calculated for each simulation. To observe the stability of the results, each simulation was performed three times (Table 4).

Table 4: Sampling variance for the design 140×15, $\bar{R} = 0.17$

Procedure	Simulations		
	1	2	3
No non-response	84.6	85.2	84.1
Weighting	86.6	87.1	86.2

It can be verified that with no non-response the sampling variance based on analytical expression (3) equals $\mathrm{Var}(\bar{y}) \approx S_b^2/140 + S_w^2/2100 = 84.8$. We can also confirm the theoretical results expressed in Tables 1 and 2. We thus conclude that with variable weights, the sampling variance is only slightly higher compared to the situation where there is no non-response.

The design[10] 700×3 and $\bar{R} = 0.4$ is presented in Table 5. There, the analytical result in case of no non-response equals $\mathrm{Var}(\bar{y}) = 48.7$. We can observe that with small clusters and large non-response rates the differences between procedures become larger.

Again, the simulation results in Table 5 support the theoretical conclusions based on on Table 1 ($b^* = 5$, $\bar{R} = 0.4$) and Table 2 ($b = 3$, $\rho \approx 0.07$) where we obtain the

[10]Here, the initial sample was $n^* = 3,500$.

Table 5: Sampling variance for the design $n = 700 \times 3$, $\bar{R} = 0.40$

Procedure	Simulations		
	1	2	3
No non-response	48.8	48.3	49.3
Weighting	57.9	57.5	58.1

approximation of the same increase by multiplying $0.18 * 0.82 = 0.15$.

Let us compare the above results with the Kish's formulae (8) which is not, of course, the correct procedure to evaluate the increase in variance due to variable non-response weights. We should remember that Kish's formulae is based on assumption of fixed weights. However, it is true that in practice, the approximation (8) is almost uniformly used for all types of weights.

The calculations of the VIF factor are based on Table 3. Let us observe the discrepancy:

- with the design 140×15 and $\bar{R} = 0.17$ there is a negligible difference between the two methods;
- with the design 700×3 and $\bar{R} = 0.40$ the increase in variance based on the (wrong) assumption of fixed weights equals 27% (Table 3) which can be compared with the proper result 15% based on variable weights approach (Tables 1 and 2, or, simulation results in Table 5).

7 Discussion

The results offer the following conclusions:

- The increase in sampling variance due to variable non-response weights is generally small. In practice, it will rarely exceed 5%. Only with both - small clusters (smaller than $b = 5$) and high non-response rates (over $\bar{R} = 40\%$) - can it reach up to 20%. With primary sampling units (clusters) larger than $b = 30$ or non-response rate smaller than $\bar{R} = 0.10$ the corresponding increase in variance is negligible.
- The increase is additionally reduced by the fact that it arises only from the within component of the variance. Its proportion can be small, especially where both, the intracluster correlation and the average cluster size, are large.
- When the variable non-response weights are treated as fixed, the increase in variance will be overestimated. Since the fixed weights formulae (8) does not depend on intracluster correlation, the level of overestimation varies from variable to variable. In general, the danger of this overestimation is relatively

small, since in practice the primary clusters are often large. However, there are some important exceptions, for example the two-stage telephone sample design Waksberg-Mitofsky or the sample designs for variables with large intracluster correlations (e.g., Family Budget Surveys).

It seems that, generally, we do not need to worry about the increase in variance arising from the variable non-response weights. This somehow counters the common fear among practitioners that such weighting may significantly increase the sampling variance. The opinion is partially based on the wrong assumption that variable weights create the same increase in sampling variance as the fixed ones. The other source of misconception is overlooking of the fact that the variability of the non-response weights cannot (not even theoretically) exceed the coefficient of variation $CV_w = 0.50$. This can result only in a limited increase in sampling variance compared to the possible variation from other sources. For example, when combining two samples, the corresponding weights can be in a proportion of $10 : 1$ which leads to a much higher increase in sampling variance.[11] A similar ratio is not so rare in the case of oversampling strata.

There are five points that should be carefully considered when the above results are interpreted:

1) We should repeat that the assumption of the Bernoulli non-response mechanism means a uniform responding mechanism for each unit selected in the survey. However, this is not true when there are differences in the level of response rate across different domains or strata. In these situations, our results hold only for the homogeneous partitions of the sample (e.g., strata). Since the differences in response rates between strata are fixed quantities, the corresponding component of the weights should also be treated as fixed, and handled by Kish's approximation (8).
Ideally, of course, one should separate the fixed and the variable component of the non-response weights. Each component should then be treated separately. This, of course, is not needed when the variable component is negligible. However, when the component of the fixed weights is small (e.g., we have an almost uniform response rate in the whole sample) and the variable weight component is important (small clusters, small response rate, small ρ), the two contributions to the increase in sampling variance should be separated. Otherwise, overestimation of the sampling variance occurs.

2) Another important issue arises with respect to the estimation. In principle, the estimation of the increase due to variable weights is a straight-forward process. However, it involves estimation of the components of the variance. This might be slightly complicated in a technical sense, because certain circumstances may pose severe practical problems.

[11] In the case of an equal sized subsample (strata) the ratio 1:10 leads to the increase in variance VIF=3 (Kish, 1965:431).

3) Whenever we compare the increase in variance due to weighting we face a conceptual problem: 'What are we comparing the increase with?' When the non-response weights are used in sample surveys the following situations must be clearly distinguished:

(a) the sampling variance that would occur in the case with no non-response (initial sample size n);

(b) the sampling variance that arises from the properly weighted sample (initial sample size n^*);

(c) the sampling variance (and additional bias) in the situation where the weights are not applied though they should be constructed (initial sample size n^*).

For proper evaluation of the weighting, the mean squared errors of the procedures (b) and (c) should be compared. This is especially important when we face the dilemma of whether to use the weights or not. However, the proper estimation of the variance (c) is much more complicated than the estimation of the variance (b). It is true that, in practice, the unweighted sample (c) is often (in fact, almost generally) treated as the sample with no non-response (a). Thus, the variance is easily "estimated". With fixed weights, the error from this oversimplification is small; however, with variable weights, the problem become much more serious.

4) It is of great importance to notice that different estimators behave differently when variable weights are used. It can be shown that as a general tendency, variance (c) is larger than (b) and variance (a) is, of course, the smallest. However, with ratio estimator in situation (c) some special effects may overrules this general statement.

5) A generalization for the designs with three or more stages should be carried out with some care. Generally, the discrepancy in the increase in sampling variance between fixed and variable weights will become larger when non-response adjustments are performed at the lower levels of cluster. The non-response adjustment at the second, third,..., stage clusters affects the increase in sampling variance to a smaller extent compared to the adjustment at the primary sampling units. The reason for this arises from a certain "cancelation" of the lower level weights at the level of primary sampling units. On the other hand, with fixed weights the same increase (8) occurs regardless of the level of adjustment.

8 Conclusion

We can summarize the following:

- The variable non-response weights in a two-stage cluster design create a smaller increase in sampling variance compared to the situation with fixed weights. As a general tendency the increase is about two times smaller.

- The corresponding increase in sampling variance is sizable only in the case of small clusters ($b < 10$) and large non-response rates ($\bar{R} > 0.20$). If clusters are large or non-response rate is small, the increase becomes negligible.
- Care should be taken whenever the component of the variable weights is sizable:
  - the variable component of the increase in sampling variance due to weighting should be separated from the fixed component;
  - comparison should be made to the situation with no non-response rather to the situation of the non-weighted sample of respondents;
  - it should be noted that the sampling variance of the unweighted sample of respondents - if correctly estimated - may dominate the variance of the properly weighted sample.

References

[1] Groves R.M. (1989): *Survey Errors And Survey Costs.* New York: John Wiley & Sons.

[2] Cochran W.G. (1978): *Sampling Techniques.* New York: John Wiley & Sons.

[3] Kish L. and Frankel M. (1970): Balanced Repeated Replication of Standard Errors. *Journal of the American Statistical Association*, **65**, 1071-1093.

[4] Kish L. (1992): Weighting For Unequal P_i. *Journal of Official Statistics*, **8**, 2, 183-201.

[5] Kish L. (1965): *Survey Sampling.* New York: John Wiley & Sons.

[6] Rust K. (1987): Practical Problems in Sample Error Estimation.*Bulletin of the International Statistical Institute*, **52**, 39-56.

[7] Verma V. (1993): *Sampling Errors in Household Surveys.* New York: United Nations.

[8] Vehovar V. (1995): Field Substitutions - A Neglected Option. *Proceedings of the Section on Survey Research Methods*, Vol I, 49th Annual Meeting of the American Statistical Association, Toronto, August 13-18, 1994, Alexandria: ASA, 589-594.