EFFECT ON MEAN SQUARE ERROR OF DIVIDING INTO ADDITIONAL NONINTERVIEW CELLS

Richard A. Griffin - Bureau of the Census

1. INTRODUCTION

Suppose a simple random sample without replacement is selected in a geographic tabulation area. The units in the tabulation area that would be interviews if selected are fixed. The tabulation area is divided into a collection of sub-areas each of which is geographically contiguous. A survey of the sample cases is conducted and a portion of the sample cases are noninterviews. Two data adjustment alternatives are analyzed in this paper under three specific sets of conditions relating to the distributions of the characteristic of interest and the population interviews across the sub-areas. The first data adjustment alternative is to form an estimator using the simple noninterview adjustment for the entire tabulation area. The second data adjustment alternative is to form an estimator using a simple noninterview adjustment within each sub-area. For each set of conditions, the expected value and the variance of the two alternative estimators are compared.

2. NOTATION

Let

N = population size

n = sample size

n_1 = interviews

n_2 = noninterviews

y_i = value of characteristic Y for sample respondent i

N_I = the size of the subpopulation that would be interviews (the units included in N_I are fixed)

Y_I = the total for characteristic Y for the subpopulation that would be interviews

$\bar{Y}_I = \frac{Y_I}{N_I}$ = the tabulation area mean for the interview population

S^2_I = the variance of characteristic Y for the subpopulation that would be interviews

$W_I = \frac{N_I}{N}$ = overall population interview rate

c = number of sub-areas (cells)

N_j = population size in cell j

N_{j1} = population size in cell j for the subpopulation that would be interviews (the units included in N_{j1} are fixed)

n_{j1} = interviews in cell j

n_{j2} = noninterviews in cell j

$n_j = n_{j1} + n_{j2}$

y_{ji} = value of characteristic Y for sample respondent i in cell j

S^2_{JI} = the variance of characteristic Y for the subpopulation in cell j that would be interviews

$W_j = \frac{N_j}{N}$ = the proportion of the population in cell j

$W_{j1} = \frac{N_{j1}}{N}$ = the proportion of the population that would be interviews in cell j if selected in sample

Y_{jI} = the total of characteristic Y for the subpopulation in cell j that would be interviews

$\bar{Y}_{jI} = \frac{Y_{jI}}{N_{j1}}$ = the cell j mean for the interview population

3. VARIANCE OF ESTIMATORS

Let $\hat{Y}_1$ denote the estimator using the simple noninterview adjustment for the entire tabulation area and let $\hat{Y}_2$ denote the estimator using a simple noninterview adjustment within each sub-area. Table 1 shows each estimator and its appropriate variance. Finite population correction factors are included.

$V(\hat{Y}_1)$ was derived using the fact that the variance of a random variable is the sum of the expected value of the conditional variance and the variance of the conditional expected value. The condition used was the sample interviews in the tabulation area. The $E(1/n_1)$ was approximated using the second order Tayor linearization about $E(n_1)=W_I n$.

$V(\hat{Y}_2)$ was also derived using the sum of the expected value of a conditional variance and the variance of a conditional expected value. The condition was the sample interviews and noninterviews in each sub-area. The $E(n_j^2/n_{j1})$ was approximated using the second order Taylor linearization about $(E(n_j),E(n_{j1}))=(nW_j,nW_{j1})$ ignoring the cross product term. Thus the $COV(n_j,n_{j1})$ is excluded. An approximation for $V(\hat{Y}_2)$ that ignores the finite population correction factors but includes $COV(n_j,n_{j1})$ is given in (1).

Note, in Table 1, that for $V(\hat{Y}_2)$, O_2 is the portion to the right of the last + sign. Thus O_2 is the portion that includes the $\bar{Y}_{jI}$ values.

4. CASE I (See Table 2)

For Case I the total of characteristic Y for the subpopulation that would be interviews is distributed across the sub-areas or cells proportionally to population and each cell has the same population interview rate. These conditions are illustrated in Table 2. Using these conditions the expected value of $\hat{Y}_2$ can be shown to be equal to $N\bar{Y}_I$ which is the expected value of $\hat{Y}_1$.

Furthermore, these conditions result in $O_2=0$ so that $V(\hat{Y}_2)=O_1$.

If the variance of the subpopulation that would be interviews is the same in each cell as the variance for the subpopulation that would be interviews in the entire tabulation area then, as shown in Table 2, $V(\hat{Y}_2)\geq V(\hat{Y}_1)$. Thus, since expected values are equal, it is better to have only one noninterview cell for the tabulation area.

If the variance of the subpopulation that would be interviews in each cell is proportional to the interview population cell size then, as shown in Table 2, $V(\hat{Y}_2)\leq V(\hat{Y}_1)$. Thus, since expected values are equal, it is better to divide the tabulation area into the c noninterview cells.

A practical application for Case I is estimation of a characteristic which can be considered a 0,1 variable, a population unit either has the characteristic or it does not. If the proportion of the population that has the characteristic is thought to be about the same in each sub-area and survey cooperation rates are also likely to be about the same in each sub-area, then Case I with equal cell variances applies.

5. CASE II (See Table 3)

For Case II each cell has the same population interview rate and the cells can be divided into two groups so that k of the c cells have a disproportionately large share of the total of characteristic Y for the subpopulation that would be interviews in the tabulation area. Within each of the two groups the group total of characteristic Y for the subpopulation that would be interviews is proportionally distributed to the group cells. These conditions are illustrated in Table 3. Using these conditions the expected value of $\hat{Y}_2$ can be shown to equal $N\bar{Y}_I$ which is the expected value of $\hat{Y}_1$.

If the variance of the subpopulation that would be interviews is the same in each cell as the variance for the subpopulation that would be interviews in the entire tabulation area then, as shown in Table 3, $V(\hat{Y}_2)\geq V(\hat{Y}_1)$. Thus, since expected values are equal, it is better to have only one noninterview cell for the tabulation area.

If the variance of the subpopulation that would be interviews in each cell is proportional to the interview population cell size then it is better to divide the tabulation area into the noninterview cells if and only if inequality * as shown in Table 3 is true.

A practical application of Case II would be estimating total property value in a tabulation area where it is felt that a portion of the area has higher property values than the other sections. Furthermore, within each of the two components of the tabulation area, property values are felt to be evenly distributed across subareas. Survey cooperation rates are likely to be about the same throughout the entire tabulation area.

6. CASE III (See Table 4)

For Case III the total of characteristic Y for the subpopulation that would be interviews is distributed across the cells proportionally to population and the cells can be divided into two groups so that k of the c cells have a larger interview rate than the remaining cells. Within each of the two groups, each cell has the same population interview rate. These conditions are illustrated in Table 4. Using these conditions the following expression can be derived:

$$E(\hat{Y}_2) = \bar{Y}_I\left[\sum_{j=1}^{k}\frac{N_j}{f_2} + \left(\sum_{j=k+1}^{c}N_j\right)2\Big/\left(N\left(1-f2\left(\sum_{j=1}^{k}\frac{N_j}{N}\right)\right)\right)\right]$$

where

$$\frac{N_{j1}}{N_j} = f_2\frac{N_I}{N} \text{ with } f_2>1, f_2<\frac{N}{N_I}, \text{ and } f_2<\frac{N}{N_j}.$$

Thus, the expected value of Y_2 does not equal the expected value of $Y_1(E\hat{Y}_1=N\bar{Y}_I)$.

For the case of the variance of the subpopulation that would be interviews the same in each cell as the variance for the subpopulation that would be interviews in the entire tabulation area $V(\hat{Y}_2)$ was compared with $V(\hat{Y}_1)$ for a tabulation area population of 1,000,000, sampling fractions of .01, .05, and .09 in the tabulation area, overall interview rates of .7, .8, and .9, the proportion of the tabulation area in those cells which have a larger interview rate equal to .25, .5, .75, .85, .90, and .95, and various possible interview rates in the cells with the larger interview rate. As presented in Table 4, for all these comparisons $V(\hat{Y}_2)$ was greater than $V(\hat{Y}_1)$. Thus, the variance of an estimate using one noninterview cell for the tabulation area is likely to be less than the variance of an estimate formed by dividing the tabulation area into noninterview cells. Since the expected values are not equal, consideration must be given to mean square error.

A practical application for Case III would be estimating a 0,1 characteristic where the proportion of the population that has the characteristic is thought to be the same in each sub-area. However, it is felt that a portion of the area will have a higher cooperation rate than the rest of the tabulation area. Within each of these components of the tabulation area, cooperation rates are likely to be the same in each sub-area.

7. VARIANCE ASSUMPTIONS

Nonequal variances within each noninterview cell was examined for the model S^2_{jI} proportional to the cell interviewed population size. This might be expected when there are forces that exert a similar influence on elements close together. (For example the estimated number of households with a poverty level income.) Equal variances within each noninterview cell is more likely close to reality when all sub-areas are large.

REFERENCES

(1) Kalton, G. (December, 1980), "Compensating for Missing Survey Data," Income Survey Development Program, Survey Development Research Center in Nonresponse and Imputation, Contract No. HEW 100-79-0127, draft interim report.

Table 1

Estimator	Variance
$\hat{Y}_1 = \sum_{i=1}^{n_1} \left[\frac{N}{n}\right] \left[\frac{n_1 + n_2}{n_1}\right] y_i$	$V(\hat{Y}_1) \cong \left[\frac{N^2 S_I^2}{W_I}\right] \left[\frac{1}{n} - \frac{1}{N} + \frac{1-W_I}{n^2 W_I}\right]$
$\hat{Y}_2 = \sum_{j=1}^{c} \sum_{i=1}^{n_{j1}} \left[\frac{N}{n}\right] \left[\frac{n_{j1} + n_{j2}}{n_{j1}}\right] y_{ji}$	$V(\hat{Y}_2) \cong \left[\frac{N^2}{n^2}\right] \sum_{j=1}^{c} S_{jI}^2 X$ $\left[\frac{n W_j^2}{W_{j1}} + \frac{W_j (1-W_j)}{W_{j1}} + \frac{W_j^2 (1-W_{j1})}{W_{j1}^2} - \frac{1}{N_{j1}} \left[n^2 W_J \left[\frac{1-W_j}{n} - \frac{1-W_j}{N} + W_j\right]\right]\right]$ $+ \frac{N^2 (1-\frac{n}{N})}{n} \left[\sum_{j=1}^{c} W_j \left[\bar{Y}_{jI} - \sum_{J=1}^{c} \bar{Y}_{jI} W_j\right]^2\right]$ $= O_1 + O_2$

Table 2 - Case I

Conditions	$Y_{jI} = \frac{N_j \, Y_I}{N}$ for all j $N_{jI} = \frac{N_I \, N_j}{N}$ for all j
Expected Values	$E(\hat{Y}_1) = E\,\hat{Y}_2 = N\,\overline{Y}_I$
Variances	$0_2 = 0$ $V(\hat{Y}_2) = 0_1$
	1. If $S^2_{jI} = S^2_{kI} = S^2_I$ for all j, k, then $$V(\hat{Y}_2) \geq V(\hat{Y}_1) + \frac{(1.75)N^3 \, S^2_I(c-1)}{n^2 \, N_I} \geq V(\hat{Y}_1)$$
	2. If $S^2_{jI} = A\,N_{jI} = \frac{A\,N_I\,N_j}{N}$ for all j with A a constant that does not depend on N_{jI}, then $V(\hat{Y}_2) \cong V(\hat{Y}_1) + \mathcal{Z}_1 + \mathcal{Z}_2$ where $\mathcal{Z}_1 = \left[\frac{N^3 \, A}{n^2}\right](2-n) < 0$ if $n > 2$ $\mathcal{Z}_2 = \left[\frac{N\,A}{n^2}\right](n-2)\left[\sum_{j=1}^{c} N_j^2\right] > 0$ if $n > 2$, and $\mathcal{Z}_1 + \mathcal{Z}_2 = -\left[\frac{N\,A}{n^2}\right](n-2)\left[\sum_{i \neq j} N_i \, N_j\right] < 0$ so $V(\hat{Y}_2) < V(\hat{Y}_1)$

Conclusion	If the interview population variance of characteristic Y is equal to the tabulation area interview population variance for each cell, then it is better to have only one noninterview cell for the tabulation area. If the interview population variances are proportional to the interview population cell sizes, it is better to divide the tabulation area into noninterview cells.

Table 3 - Case II

Conditions	1. Divide cells into two groups; one with k cells the other with c-k cells. 2. The k cells have a disproportionately large share of the total of characteristic Y for the interviewed population in the tabulation area. 3. For both groups, the group total of characteristic Y for the interviewed population is proportionately distributed to the group cells. 4. $N_{j1} = \frac{N_I \, N_j}{N}$ for all j
Expected Values	$E(\hat{Y}_1) = E(\hat{Y}_2) = N \, \overline{Y}_I$
Variances	1. If $S^2_{jI} = S^2_{kI} = S^2_I$ for all j, k, then $V(\hat{Y}_2) \geq V(\hat{Y}_1) + \frac{(1.75) \, N^3 \, S_I^2 \, (c-1)}{n^2 \, N_I} + O_2 \geq V(\hat{Y}_1)$
	2. If $S^2_{jI} = A \, N_{j1} = \frac{A \, N_I \, N_J}{N}$ for all j with A a constant that does not depend on N_{j1}, then $V(\hat{Y}_2) \leq V(\hat{Y}_1)$ if and only if $\left(\frac{N \, A}{n^2}\right) (n-2) \left(\sum_{i \neq j}^{c} N_i \, N_j\right) > O_2$ *

Conclusion	If the interview population variance of characteristic Y is equal to the tabulation area interview population variance for each cell, then it is better to have one noninterview cell for the tabulation area. If the interview population variances are proportional to the interview population cell sizes, then it is better to divide the tabulation area into the noninterview cells if and only if inequality * above is true.

Table 4 - Case III

Conditions	1. Divide cells into two groups; one with k cells the other with c-k cells. 2. The k cells have a larger interview rate than the remaining cells. 3. Within each group, each cell has the same interview rate. 4. $Y_{jI} = \frac{N_j \ Y_I}{N}$ for all J
Expected Values	$E(\hat{Y}_1) = N \ \overline{Y}_I$ $E(\hat{Y}_2) \neq E(\hat{Y}_1)$
Variances	For $S^2_{jI} = S^2_{kI} = S^2_I$ for all j, k, $V(\hat{Y}_2)$ was compared with $V(\hat{Y}_1)$ for $N = 1{,}000{,}000$, $\frac{n}{N} = .01, .05,$ and $.09$, $\frac{N_I}{N} = .7, .8,$ and $.9$, $\frac{\sum_{j=1}^{k} N_j}{N} = .25, .5, .75, .85, .90, .95,$ and various possible interview rates in the group of k cells. For all these comparisons $V(\hat{Y}_2)$ was greater than $V(\hat{Y}_1)$.
Conclusion	If the interview population variance of characteristic Y is equal to the tabulation area interview population variance for each cell, then the variance of an estimate using one noninterview cell for the tabulation area is likely to be less than the variance of an estimate formed by dividing the tabulation area into noninterview cells. Since the expected values are not equal, consideration must be given to mean square error.