ESTIMATION OF CORRELATED RESPONSE VARIANCE

Karol P. Krótki, Statistics Canada

I. INTRODUCTION

The total variance of a survey estimate incorporates both sampling and response variance. To see this formally one can decompose the total mean squared error into total variance and bias. In turn, total variance can be decomposed into sampling and response variance. Finally, the response variance can be expressed as the sum of simple response variance and the correlated response variance. The first component measures trial-to-trial variability of the response of a given respondent. It is the part of the response variance that is produced by tendencies of individual respondents to commit response errors independently of any other respondents. Correlated response variance, on the other hand, reflects the part of total response variance due to a common influence on a group of respondents.

A more detailed decomposition of total variance is provided in the seminal work by Hansen, Hurwitz and Bershad ([1],[2],[19]). The decomposition of the mean squared error can be expressed algebraically as

$$\text{M.S.E. } (\bar{x}) = \frac{1}{n}\sigma_r^2 + \frac{n-1}{n}\rho_r\sigma_r^2 + \frac{N-n}{N-1}\frac{\sigma_s^2}{n} + \frac{2(n-1)}{n}\sigma_{rs} + B^2$$

where $\bar{x}$ is the mean of variable X, n is the sample size, N is the population size, σ_s^2 is the sampling variance, σ_r^2 is the sample response variance, $\rho_r\sigma_r^2$ is the correlated response variance, σ_{rs} is the sampling-response covariance and B is the bias. A similar result can be obtained through a linear model approach in which the observed value is expressed as a linear combination of the true value, a term reflecting the constant bias of each enumerator and an independent random component [7]. The Census model-type approach was used to develop the theoretical foundation for the measurement of total variance at Statistics Canada ([3],[8],[11]).

Various models have been used to implement the above-mentioned theory. These are described in general treatments of response errors([16],[17]). Some work has also been done on an alternative approach to the Census model by using an indicator function [12]. Finally, attemps have been made to assess the effect of errors for complex statistics but there is still much to be done in this area ([4],[5],[6],[18])

II. 1976 TOTAL VARIANCE STUDY AT STATISTICS CANADA

Total variance results have been calculated for the 1976 Canadian Census. They were based on a sample of 750 enumeration areas (EAs) which were selected from a universe somewhat smaller than all 35,154 EAs in Canada. The sample was a two-stage stratified sample. First, from the universe of 1642 Census Commissioner Districts (CCD), 188 were selected with probability proportional to the number of EAs in the CCD. Within each such primary sampling units, EAs were paired in such a way that both EAs in a pair were adjacent and similar with respect to density, language and enumeration method. In the second stage of selection, two pairs were selected in each CCD.

Whereas sampling variance may be calculated straightforwardly from the sample elements, correlated response variance involves an experimental design to provide multiple observations on each response. Replication of the survey is one way to achieve this. However problems of contamination make this approach subject to criticism. The design used in the 1976 Census to measure total variance and correlated response variance is based on interpenetration of interviewers and respondents. The use of interpenetration in this fashion is due to Mahalanobis [15].

The shortcoming of this approach is that not all components of total variance can be estimated. However, since there was some evidence that simple response variance is a relatively minor component of response variance it was decided to concentrate on measuring the effect of correlated response variance. A detailed derivation of the formulae used can be found in [8] for the case in which both interpenetration and replication are applied. Assuming certain factors negligible, a somewhat shorter derivation applicable to the interpenetrated design of 1976 is presented in [3]. The basic developments of this work, which was the theoretical foundation for the calculation of total variance estimates in 1976, are presented here.

Normally each EA is handled by one Census Representative (CR). However in the case of the total variance EAs this procedure was altered in order to implement the interpenetration of CRs and EAs. Each total variance EA was split in two halves by randomly assigning each household into half one or two. Interpenetration was organized within EA pairs. Each CR was given half of her original EA plus half of the other EA in the pair. The assigning of halves to CRs was done randomly.

Data from the 1976 Canadian Census are based on both a 100% enumeration of the population of Canada and a 1/3 sample. The total variance formulae developed here pertain to the sample data since this case is the more general one. The extension of formulae for 100% characteristics is straightforward since only the sample size must be changed and the sampling variance component of total variance eliminated. The estimate of population total for a Census sample characteristic can be written as:

$$\hat{X} = \sum_{k=1}^{P} \frac{N_k}{n_k} \sum_{h \varepsilon S_k} x_{kh}$$

where k indexes EAs, P is the total number of EAs in Canada, N_k is the size of EA k (total number of persons, families or households), n_k is the sample size (about $N_k/3$), h indexes households within an EA, S_k is the set of sample households in EA k and x_{kh} is the value for some characteristic for household h in EA k. The total variance of this estimate is:

$$\hat{V}(X)=\sum_{k=1}^{P} N_k^2 \frac{\sigma_k^2}{n_k}(1+(n_k-1)\rho_k)+(1-\frac{n_k}{N_k})\frac{S_{xk}^2}{n_k}$$

where S_{xk}^2 is the sampling variance for characteristic x in EA k, σ_k^2 is the simple response variance and $\sigma_k\rho_k$ is the correlated response variance.

From the experimental design two estimators can be obtained. The first is the between enumerator variance, C_k, that is a measure of variance for EA k.

$$C_k=\frac{1}{2}(\bar{x}_{k(1)}-\bar{x}_{k(2)})^2$$ ($\bar{x}_{k(i)}$ is the mean value of x for half i of EA k)

and $$E(C_k)=\frac{2\sigma_k^2}{n_k}(1+(\frac{n_k}{2}-1)\rho_k)+\frac{2S_{xk}^2}{n_k}.$$

The second is D_k, a within enumerator variance for EA k.

$$D_k=\frac{\sum_{i=1}^{2}\sum_{h\varepsilon S_{ki}}(x_{kh}-\bar{x}_{k(i)})^2}{n_k-2}$$ (S_{ki} is the set of sample households in half i of EA k)

and $$E(D_k)=\frac{2}{n_k}[\sigma_k^2(1-\rho_k)+S_{xk}^2].$$

Finally, it can be shown that $E(C_k-D_k)=\rho_k\sigma_k^2$, a measure of correlated response variance for EA k.

For characteristics based on sample data the following two formulae were used to estimate the correlated response variance and total variance, resepctively.

$$CRV=\frac{P}{p}\sum_{k=1}^{p} N_k^2 [(\frac{n_k-1}{n_k}(C_k-D_k]$$

$$TV=\frac{P}{p}\sum_{k=1}^{p} N_k [\frac{(N_{k-1})}{N_k}C_k-(\frac{N_k+n_k-2}{2N_k})D_k]$$

where P is the total number of EAs in Canada and p is the number of EAs in the total variance sample.

The estimate of correlated response variance is unbiased whereas the estimate of total variance has a slight negative bias. In the case of variables based on the entire Census, there is no sampling variance and thus the unbiased estimate of the correlated response variance serves as an estimate of the total variance, albeit biased in the simple response variance term.

Before the empirical results are described, a brief treatment of some problems that were encountered will be given. The formulae were developed using the EA as the unit of observation, since it was assumed that there is a one-to-one correspondence between EAs and CRs. It is the CR assignment that is of central importance but the EA is the geographical unit by which data are collected. However, for quite a high proportion of the EAs, a CR in fact handled more than one EA. This problem was handled by substituting for P in the above formulae, not the total number of EAs in Canada but the total number of CRs.

The sample did not yield equal probability of selection because the number of pairs in an EA was not always half of the number of EAs. It was decided that the benefits were too small to justify the complicated task of including adjustment weights. There was some displeasure on the part of field personnel who were assigned to the project and who found that their average travelling distance had doubled. CRs were deliberately not informed as to which EAs were to be in the total variance project until after the households in the EAs had been listed by the CR assigned to the EA. For some subclasses the number of units in an EA was very small. In cases where either half of the EA had less than two units possessing a given characteristic this EA was excluded from the calculations. Finally, because the basic building block of these estimates is the quantity (C_k-D_k) there is no guarantee that the resulting estimate will be positive. In fact it has been shown in [19] that the variance of D_k tends to be larger than the variance of C_k often resulting in negative estimates for individual EAs.

Results are presented in table 1. The choice of characteristics was quite subjective with an aim of providing a wide variety both with respect to type of characteristic and size of total variance. The ratio of the correlated response variance to the total variance is provided rather than the ratio of the correlated response variance to the sampling variance as in [2]. The reason is that the sampling variance has not been calculated. To calculate sampling variance accurately, at least in the case of person and family variables, stratification and clustering would have to be accounted for. The Census 1/3 sample is stratified by enumeration area and clustered by household. Various alternative solutions are available, the most immediate one of which is to use the simple random sample formulae. The accuracy of this approach depends on the design effect. The sampling variance could also be estimated based on the total variance sample. This has its drawbacks since the results may not be compatible with the results obtained by subtracting the correlated response variance from the total variance.

It is not clear how negative estimates are to be treated. For official publication purposes, EAs which contributed negative values to the overall correlated response variance were considered to contribute a value of zero. However, the avoidance of this rule leads to a number of overall negative variances. Turning to the positive results, it is difficult to build a story around the relative sizes of the variances. If indeed correlated response variance measures interviewer effect, then we would be hard pressed to explain why, for example, the estimate for Italian is higher than for French or English unless it is a question of poor communication between the interviewer and the respondent. The questionnaire is available in English and French only. Interviewers also tend to belong to these two language groups.

Similar results have been prepared for characteristics based on 100% data. As would be expected the total variance estimates are larger for the sample data. Many of the relationships evident in table 1 are also present for 100% data. For example, the results for Italian are larger than those for other mother tongues. There are far more negative results for 100% data. This suggests it is more than just a question of variance results generally being of the same absolute size but sometimes having positive and sometimes negative signs. It seems that there is a continuum whose lower end extends below zero.

III A COMPARISON OF 1961, 1971 and 1976 CORRELATED RESPONSE VARIANCE ESTIMATES

The correlated response variance reflects the part of total response variance due to a common influence on a group of respondents. This common influence could be the interviewer and thus the correlated response variance is often interpreted as the interviewer effect. The 1961 Census of Canada was carried out in the traditional canvasser method using interviewers. However the 1971 and 1976 Censuses were carried out almost entirely using self-enumeration, in which the interviewer has less influence on the respondent than in the canvasser system. Thus, if correlated response variance indeed measures interviewer effect, then we would suspect that the estimates for 1971 and 1976 would be lower than those for 1961.

Empirical results show that, in general, this is the case. The 1961 estimate used in the comparison is the quantity $\frac{1}{N}[C_1-F_1]$, using the notation in [8]. For 1971 and 1976 the estimate is calculated as the weighted average of C_k-D_k. The weights reflect the relative sizes of the EAs.

Most characteristics with positive results in 1961 give lower estimates in 1971 and 1976 including some that are negative. Some estimates that are not lower in 1971 are some age groups and French as the official language spoken. A substantial decline occurs for the ethnic group French. In some cases the decline of the correlated response variance can be observed across all three time points (e.g. mother tongue English). It is also interesting to note that for all three Censuses the result for Mother Tongue English is larger than that for Mother Tongue French. The above findings are qualified by a number of considerations concerning the variance and accuracy of the estimators and the problem of preparing results derived under different circumstances. Further details about this and other issues can be found in a more complete report [13].

IV. AN ALTERNATIVE METHOD OF CALCULATING CORRELATED RESPONSE VARIANCE

An empirical investigation of an alternative method of calculating the correlated response variance is being carried out at Statistics Canada. 'New method' shall refer to the method outlined in [9] and 'old method' shall refer to the method used to calculate the estimates in table 1. The new method calls for supplementing the interpenetrated EAs with EAs that are not interpenetrated. For the interpenetrated EAs, the sum of the differences between EAs in a pair is calculated. The same is done for the non-interpenetrated EAs which have been paired in a manner similar to that used for the interpenetrated EAs. The difference between these two differences is the core of the new estimate of correlated response variance. Reasoning intuitively, this is so because in the interpenetrated EAs the difference between the EAs in a pair should be less than for the normal EAs. The extent to which this is true reflects the magnitude of the effect of the interviewer.

For the pusposes of this investigation, 564 non-interpenetrated pairs of EAs were selected to supplement the 375 pairs in the original total variance sample. The formulae in [9] were changed slightly to make the results using the old method. The main problem in this exercise lay in the fact that the new method uses the EA pair as the unit of analysis while the old method uses the individual EA. The formulae used to calculate the correlated response variance are given in [14].

Results using the old and new methods are presented in table 2. To avoid detracting from the main point of this table only characteristics with positive total variance results are included. It is clear from these results that the new method produces a higher estimate than does the old method. The variance of each method was calculated using the balanced repeated replication method. The results are not consistent across characteristics and no conclusion concerning the relative merit of the two methods can be drawn on the basis of these data.

There were a number of problems in carrying out this piece of research and there are still several unanswered questions. It was suspected that outliers may be affecting the results. By outliers is meant those EA pairs in which the difference between the two EAs is large. The elimination of these outliers resulted in estimates that were generally closer to zero whether the original estimates were negative or positive. The new method is defined in terms of EA totals (t_{mi} in the notation of the formula) and it is presumed that this was developed under the assumption of equal EA size. In fact EAs vary considerably in size and to test the effect of this departure from the assumption the formula was altered to handle means rather than totals. The resulting calculations did not produce results that were very different from those produced using totals. It was also postulated that maybe the number of non-interpenetrated EAs has some effect on the results. This was tested by repeating the calculations after deleting a varying number of EA pairs. The results did not indicate any regular relation between sample size and size of the estimate. However it was interesting to note that the results were not only irregular with respect to size but also in several cases their sign oscillated randomly back and forth between positive and negative. This suggests that the negativity phenomenon may be the product of a very unstable estimator. A more detailed and complete report of this investigation is available [14].

V. FUTURE RESEARCH

It is important to establish a connection between the mathematical developments and substantive interpretation. It might be obvious in the case of canvasser methodology that correlated response variance measures interviewer effect, but in the case of self-enumeration the issue is not so clear. Mathematically, total variance can be decomposed into its various components, one of which is sampling variance. However this model is only valid under the assumption of simple random sampling. The question arises as to how the complexity of the sample design (stratification, clustering) can be incorporated into the mathematical model. As a somewhat related topic, it must also be realized that the current formulae apply to simple statistics. But in fact at Statistics Canada Census data are weighted before distribution. This weighting, using the weighted raking ratio technique, makes the resulting estimate very complicated mathematically. The problem is to find appropriate total variance estimators for such weighted results. The phenomenon of negative variances could stand some more scrutiny. The extent of this problem and its intractability very often interferes with such basic considerations as to whether a variance is zero and whether one variance is larger than another one. Finally it has been noted that there is a large body of literature in the area of response variance and there has been little attempt to unify it. Certainly it would be useful to review this research with a view to identifying the main trends, define a common notation and to suggest where research is needed.

The author gratefully acknowledges the assistance of G. Brackstone, C.Hill, A.MacLeod and C. Maurice.

REFERENCES

(1) Bailar, B., T. Dalenius (1969), "Estimating the response variance components of the U.S. Bureau of the Census' survey model", Sankhya B 31 (parts 3 & 4): 341-60 Dec.

(2) Bailey, L., T.F. Moore, B. Bailar (1978), "An interviewer variance study for the eight impact cities of the National Crime Survey cities sample", Journal of the Americal Statistical Association, 73(361):16-23 March.

(3) Brackstone, G.J., C.J. Hill (1976), "The estimation of total variance in the 1976 Census", Survey Methodology 2(2):195-208, Dec.

(4) Chai, J.J. (1971), "Correlated measurement errors and the least squares estimator of the regression coefficient", Journal of the American Statistical Association, 66(335):478-83, Sept.

(5) Cochran, W.G. (1968),"Errors in measurement in statistics", Technometrics, 10,637-66

(6) Cochran, W.G. (1970), "Some effects of errors of measurement on multiple correlation", Journal of the American Statistical Association, 65(329):22-34, March

(7) Dodds, D.J., T.M.F. Smith (1973), "Estimation of correlated response variance under a linear additive model", presented at the IASS conference, August

(8) Fellegi, I.P. (1964), "Response variance and its estimation", Journal of the American Statistical Association, 59(308):1016-41, Dec.

(9) Fellegi, I.P. (1974), "An improved method of estimating the correlated response variance", Journal of the American Statistical Association, 69(346):496-501, June.

(10) Hansen, M.H., W.N. Hurwitz, M.A. Bershad (1961), "Measurement errors in censuses and surveys", Bulletin of the ISI, 38(2):359-74

(11) Hill, C.J. (1976), "1971 Census evaluation programme, 1971 response variance project, methodology report and results", internal report, Census Survey Methods Division, Statistics Canada.

(12) Koch, G.G. (1973), "An alternative approach to multivariate response error models for sample survey data with applications to estimators involving subclass means", Journal of the American Statistical Association, 68(344): 906-13, Dec.

(13) Krótki, K.P., C.J. Hill (1978), " A comparison of correlated response variance estimates obtained in the 1961,1971 and 1976 Censuses", Survey Methodology, in print

(14) MacLeod, A.D. (1978) "Two methods of measuring correlated response variance", internal report, C.S.M.D., Statistics Canada.

(15) Mahalanobis, P.C. (1946), "Recent experiments in statistical sampling in the Indian Statistical Institute", Journal of the Royal Statistical Society, 109:325-370.

(16) Nisselson, H., B.A. Bailar (1976), "Measurement, analysis and reporting of nonsampling errors in surveys", Proceedings of the 9th international Biometric Conf., invited papers, vol.2:301-22, Boston, 22-27 August

(17) O'Muircheartaigh, C.A., C. Payne (eds) (1977), The Analysis of Survey Data, Vol. 2, Model Fitting, John Wiley and Sons

(18) Rao, A.V. (1971) "Response and ratio estimators", Technical Report nó. 5, Project SU-618, Research Triangle Institute, U. of North Carolina, July.

(19) U.S. Bureau of the Census (1968), "Evaluation and research program of the U.S. Census of Population and Housing, 1960"Series ER60 (7).

TABLE 1

Estimates of Total Variance and Correlated Response Variance for Selected Characteristics Based on Sample Data From the 1976 Census of Canada

Characteristic	Estimated Population Count (in millions)	Estimate of Total Variance (TV)	Coefficient of Variation $(x10^4)$	Ratio[3] of Correlated Response Variance to TV
Persons				
Age: 21	0.43	-532,094	-	-
25-29	1.98	4,881,500	11.15	-
45-64	2.63	9,003,866	11.43	.254
≥25	12.51	7,095,674	2.13	-
Marital Status: Married	10.78	14,609,671	3.54	.181
Labour Force Status: Employed	9.55	19,094,175	4.57	.176
Sex: Female	11.49	17,230,450	3.61	.091
Mobility Status: Migrant[1]	5.17	30,356,840	10.66	.127
Mother Tongue: English	13.88	46,626,984	4.92	.172
French	5.99	31,517,949	9.37	-
Italian	0.53	10,447,183	61.27	-
Polish	0.11	-3,080,193	-	-
Education:[2] No University	14.28	20,891,310	3.20	.046
Grade 9-10	3.35	9,453,881	9.18	.159
Some University	0.92	1,549,849	13.58	-
Households				
Size: 4 Persons	1.34	1,888,540	10.22	-
Sex of Head: Female	1.42	2,688,765	11.52	.084
Marital Status of Head: Married	5.57	2,733,997	2.97	.064
Type: One Family	5.71	2,937,072	3.00	.193
Non-Family	1.49	2,902,981	11.43	.142
Families				
Type: Husband-Wife	5.33	3,679,841	3.60	.183
Lone Parent	0.57	513,247	12.55	-

[1] Did not live in the same municipality five years age

[2] Highest level reached

[3] This ratio is not calculated for characteristics for which either the total variance or correlated response variance is negative

TABLE 2

Variance Results Using the Old and New Methods for Selected Characteristics From the 1976 Census of Canada

Characteristics	Estimated Population Count (in millions)	$\frac{\sqrt{\text{Correlated Res. Var.}}}{\text{Pop. Count}}$		Coefficient of Variation of Estimate of Correlated Response Variance	
		Old	New	Old	New
Age: 25-29	1.98	2.90	8.94	1.18	0.89
Highest Degree Received:					
H.S. Certificate	3.36	4.56	10.00	0.45	0.44
Highest Grade: <5	0.86	11.20	13.71	2.02	1.46
Not attending school	14.33	1.43	2.29	0.76	5.05
Attending school full time	1.68	4.20	4.69	1.53	3.50
Employed	9.55	1.92	6.74	0.79	0.67
Mover[2]	9.93	1.71	14.30	3.16	0.40
Non-Migrant[2]	4.76	2.51	10.40	1.71	0.59
Migrant from[2] outside Canada	0.60	34.00	47.00	0.20	0.76

1. All results, except those for age, are based on sample data

2. A mover is one who in 1971 lived in a different household. A migrant is a mover who lived in a different municipality in 1971.