A COMPARISON OF CORRELATED RESPONSE VARIANCE ESTIMATES OBTAINED IN THE 1961, 1971 AND 1976 CENSUSES

K.P. Krótki and C.J. Hill[1]

The total variance of a survey estimate incorporates sampling variance, simple response variance and correlated response variance. The last component reflects the part of the total variance due to a common influence on a group of respondents. In the Canadian census, self-enumeration was adopted as the standard method of enumeration in the 1971 Census. One factor in favor of introducing this method was evidence, from the 1961 Census, that correlated response variance made an important contribution to the total variance of census estimates. Based on a study conducted using interpenetration of interviewers, this article compares correlated response variances from the 1961, 1971 and 1976 Censuses. The empirical results demonstrate that although the self-enumeration adopted in the 1971 Census did not completely remove the correlated response variance, this approach has considerably reduced the magnitude of this component of variance for almost all the characteristics examined.

1. 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, the total variance can be decomposed into sampling and response variance. Finally, the response variance can be expressed as the sum of the 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. The correlated response variance, on the other hand, reflects the part of the total variance due to a common influence on a group of respondents.

---

[1] K.P. Krótki and C.J. Hill, Census Survey Methods Division, Statistics Canada.

The decomposition of the total variance is laid out in detail in the seminal work by Hansen, Hurwitz and Bershad [9]. This is often referred to as the Census model ([1], [2], [3]). The decomposition of the mean squared error can be expressed algebraically as:

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

where $\bar{x}$ is the mean variable X, n is the sample size, N is the population size, σ_s^2 is the sampling variance, σ_r^2 is the simple response variance, $\rho_r\sigma_r^2$ is the correlated response variance, σ_{rs} is the sampling-response covariance and B is the bias. Similar results can be obtained through a linear model approach (see [4]) 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. An overall review of total variance can be found in [12].

It has been shown empirically that in the case of canvasser enumeration the correlated response variance is the dominant component of total response variance ([5], p. 1035). In fact, in the case of a Census, for data collected from 100% of the population, there is no sampling variance and the correlated response variance becomes the dominant component of the total variance. This component is due to an effect on the respondents that causes respondents in the same interviewer assignment area to commit similar errors. It is postulated that this homogeneity of errors is due to the effect of the interviewers on the respondents.

In the Canadian census, self-enumeration was adopted as the standard method of enumeration in the 1971 Census. One factor in favor of introducing this method was evidence, from the 1961 Census, that correlated response variance made an important contribution to the total variance of census estimates [5]. In principle, self-enumeration would totally remove this component of variance. In practice, even while using this method, there remains some enumerator-respondent contact. This being the case, it is to be expected that self-enumeration will reduce rather

than completely remove all correlated response variance. This article gives evidence to show that this is indeed the case by making a comparison between the estimates obtained for the 1961 Census with those obtained in 1971 and 1976.

2. THE MODEL AND EXPERIMENTAL DESIGN

Whereas sampling variance may be calculated straightforwardly from the sample elements, the calculation of 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 (lack of independence between the first and second interview) make this approach subject to criticism. The design used in the 1961, 1971 and 1976 Censuses 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 (1949). The theory and experimental designs for investigating response errors and the particular methods applied are discussed in detail elsewhere ([5], [8]). A brief summary is provided here.

An interpenetrated design model was set up for the 1961, 1971 and 1976 Censuses by selecting samples of pairs of adjacent enumeration areas (EAs) in which both enumerators shared the same supervisor and then partitioning the EAs at random into 2 approximately equal parts. Each enumerator of the pair is then responsible for the completion of the paired assignment. In 1961 the study was effected in 96 EAs in the Cornwall area, whereas in 1971 and 1976 a stratified random sample of 376 pairs of EAs from across Canada was used. Only a small number of canvasser EAs in the North, collective EAs and collectives (institutions such as hospitals, prisons, etc.) within EAs was excluded from the scope of this study.

An additional feature of the 1961 study was the use of re-enumeration. Re-enumeration allows the estimation of additional parameters including

the estimation of simple response variance, but introduces considerably increased costs. A drawback of re-enumeration, however, is that it is impossible (in a response error study) to ensure independence between repeated measurements due to the effect of recall. For this reason and the evidence that under canvasser enumeration the simple response variance contributes less to the total variance than does the correlated response variance ([5], p. 1035), the 1971 and 1976 studies were limited to interpenetration.

3. FORMULAE

A detailed derivation of the formulae for total and correlated response variance is given in [5] 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 1971 and 1976 is presented in [3]. The basic developments of this work are presented here.

The following notation will be used in the formulae.

P	is the number of EAs in Canada
k	is the subscript denoting the EA
h	is the subscript denoting the household within an EA
$k(i)$	denotes the ith half of EA k
N_k, n_k	denotes the total number of households in the population and in the sample respectively for the kth EA
n_{ki}	is the number of households in the sample for the ith half of the kth EA
X_{kh}	denotes the observed characteristic for household h in EA k
σ_k^2	is a measure of response variance
$\rho_k\sigma_k^2$	is a measure of correlated response variance
S_{xk}^2	is the sampling variance.

The estimate of the population total for a Census sample characteristic can be written as

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

where S_k is the set of sample households in EA k,

and its total variance is then given by

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

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.

$$C_k = \frac{1}{2} [\bar{x}_{k(1)} - \bar{x}_{k(2)}]^2$$

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

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

$$D_k = \frac{\sum_{i=1}^{2} \sum_{h \epsilon S_{ki}} (x_{kh} - \bar{x}_{k(i)})^2}{n_{k1} + n_{k2} - 2}$$

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

Finally, it can be shown that, with a certain bias,

$E(C_k - D_k) = \rho_k \sigma_k^2$ a measure of correlated response variance for EA k.

In order to make the 1971 and 1976 results comparable to the 1961 results, a weighted average of the $\rho_k \sigma_k^2$ over all EAs in the project was calculated. Weights reflected the size of the EA in terms of number of people.

It should be pointed out that for several reasons the results presented in this paper do not correspond to the published total variance results for 1971 and 1976 publications. One reason is that the publications contain estimates of total variance which differ from the estimates of correlated response variance. Secondly, the publications present results inflated to the Canada level. The formulae used to calculate these total variance results at national and sub-national levels can be found in [3]. Finally, for purposes of publication, for any EA in which $C_k - D_k$ was negative this quantity was set to zero. Looking at the results, for several characteristics there is a sufficient number of negative $C_k - D_k$ values to make the overall average value of the estimate of the correlated response variance negative. A discussion of this problem can be found in [13] in which a statistical explanation is given for negative estimates of the correlated response variance.

4. LIMITATIONS IN COMPARING 1961, 1971 AND 1976 CORRELATED RESPONSE VARIANCE

The major limitations in comparing the 1961, 1971 and 1976 correlated response variance estimates are the differences in scope and design of the projects. As has been indicated above the 1961 study included both replication and interpenetration in its design but was only applied in the Cornwall area, whereas the 1971 and 1976 studies, which only included interpenetration, were applied to a Canada wide sample. These differences reflect different objectives for the studies. The 1961 study attempted a detailed investigation of the factors contributing to total variance. The 1971 study accepted as given the 1961 result that the correlated response variance dominates the simple response variance. Thus, rather than provide more detail on the components of total variance,

the 1971 study sought to give a reliable measure of variance at a Canada wide level to accompany the published Census estimates.

The difference in the domain of study, i.e. Cornwall as opposed to Canada may not be of importance for the majority of variables. There is no reason to believe that the enumerators in Cornwall were more or less inclined to influence responses than anywhere else in Canada. However, in certain crucial respects Cornwall is atypical of the whole of Canada. It is a boundary area between French speaking and English speaking regions containing a substantial proportion of both language groups. Clearly if 100% of the persons in an enumeration area have the same characteristics the likelihood of there being response errors for any particular characteristic is very low, whereas in a heterogeneous area errors are to be expected, if for no other reason than that there exists a proportion of persons for whom the 'true' response is ambiguous. In effect, therefore, for variance estimates of either English or French, the comparison is between an area with potentially high correlated response errors and all of Canada that includes both areas of this nature and other areas where response errors will be low.

There are two additional limitations to be considered in making the comparisons.

(1) The 1961 Census asked all questions of the entire population, whereas in 1971 some questions were asked only of a 1/3 sample. Comparisons are therefore sometimes between 100% questions and at other times between 100% questions and sample questions. It is not clear whether or not this difference is critical.

(2) The wording of questions has often changed from one Census to the next, indeed in some cases there is a slight change in the concept covered by the questions.

5. THE RESULTS

The results are given in two tables. The first provides 1961, 1971 and 1976 results for those characteristics that were included in [5]. The second table is an extension of Table 1 in that it gives 1961, 1971 and 1976 comparisons for characteristics not published in [5].

Two estimates of correlated response variance for 1961 are available. The first is calculated as $\delta_{21}\,\sigma_{r1}^2$, using the notation from [5]. In fact, the estimate of this quantity involves several other terms (see formulae 39 and 41 in [5]) which are assumed to be negligible. Furthermore, the estimate of this version of the correlated response variance is based on both the interpenetrated and replicated aspects of the 1961 survey. Thus this estimate is not directly comparable to the estimate used in 1971 and 1976. The estimate from [5] that is comparable to the estimate used in 1971 and 1976 is $\frac{1}{n}(C_1 - F_1)$ where n is the average enumerator assignment size, C_1 is the between enumerator variance for the first survey and F_1 is the within enumerator variance for the first survey ([5], p. 1033). This estimate is based only on the interpenetrated part of the study. It is this estimate that is used in comparing 1961, 1971 and 1976 results.

Comparing 1961 results with those from 1971 and 1976 in Table 1, it is evident that the estimate of correlated response variance is in general considerably reduced. All the characteristics with positive values in 1961 give lower estimates in 1971 and 1976 including two that are negative. The only estimate that is not lower in 1971 is for Age = 5 which in any case is scarcely different from zero. The most substantial difference occurs for Ethnic Group = French.

Table 2 comparing the unpublished estimates from 1961 with the 1971 and 1976 figures gives essentially the same results. Official language spoken = French Only gives a higher estimate, but all the other results are either lower or negligible. The main reservation in interpreting these estimates is the presence of negative results that give numerically large values.

The possible explanations for these results are that (1) errors are introduced which actually reduce total variance and (2) the variance of the correlated response variance is high particularly for those variables that were sample variables in 1971 and 1976. The variance of the correlated response variance may be further inflated by the fact that these variables are clustered and only apply to persons over 15.

In some cases (e.g. Mother Tongue = English), a decline of the correlated response variance can be observed across all three time points. Calculation of variances for more characteristics in which the results are not negative could shed more light on the prevalence of this situation. It is also interesting to note that for all three Censuses the values for Mother Tongue = English are larger than those for Mother Tongue = French. Little can be said about the implications of this result until more is known about the languages of both interviewers and respondents.

6. CONCLUSION

The comparison of the 1961, 1971 and 1976 correlated response variance estimates gives empirical evidence of a reduction in the variance between 1961 and 1971. This reduction can presumably be attributed to the change from canvasser to self-enumeration collection of data. The reduction of variance is seen to persist through to the 1976 Census. In fact, in some cases, the variance is even further reduced.

These findings are, however, qualified by a number of considerations concerning the variance and accuracy of the estimators and the problems of preparing results derived under different circumstances. Research concerning the variance and accuracy of the estimators is now being carried out on two fronts. First, a new method of calculating the response variance [6] is being investigated with data from the 1971 and 1976 Censuses.

Some preliminary results are already available ([9], [11]). Second, the old estimator is being studied in depth to shed further light on its theoretical and empirical properties.

RESUME

La variance totale d'un estimateur dans une enquête comprend la variance due à l'échantillonage, la variance due aux réponses simples et la variance due aux réponses correlées. Ce dernier composant reflète la partie de la variance totale causée par une influence commune sur un groupe de répondants. Dans le cas du recensement canadien, on a adopté l'auto-énumération comme méthode générale d'énumération pour le recensement de 1971. Un facteur en faveur de l'introduction de cette méthode était l'évidence, dans le recensement de 1961, que la variance due aux réponses correlées apportait une contribution importante à la variance totale des estimations du recensement. Cet article, basé sur une étude faite en utilisant l'interpénétration des interviewers, compare les variances dues aux réponses correlées des recensements de 1961, 1971 et 1976. Les resultats démontrent que, bien que la méthode d'auto-énumération adoptée pour le recensement de 1971 n'ait pas enlevé complètement la variance due aux réponses correlées, cette approche a considérablement réduit l'importante de cette composante de la variance pour presque toutes les caractéristiques examinées.

REFERENCES

[1] Bailar, B. and Dalenius, T. (1969), "Estimating the Response Variance Components of the U.S. Bureau of the Census' Survey Model", Sankhya B 31 (parts 3 & 4):341-60, December.

[2] Bailey, L., Moore, T.F. and Bailar, B. (1978), "An Interviewer Variance Study for the Eight Impact Cities of the National Crime Survey Cities Sample", Journal of the American Statistical Association, 73(361):16-23, March.

[3] Brackstone, G.J. and Hill, C.J. (1976), "The Estimation of Total Variance in the 1976 Census", Survey Methodology 2(2):195-208, December.

[4] Dodds, D.J. and Smith, T.M.F. (1973), "Estimation of Correlated Response Variance Under a Linear Additive Model", presented at the IASS Conference, August.

[5] Fellegi, I.P. (1964), "Response Variance and its Estimation", Journal of the American Statistical Association, 59(308):1016-41, December.

[6] Fellegi, I.P. (1974), "An Improved Method of Estimating the Correlated Response Variance", Journal of the American Statistical Association, 69(346):496-501, June.

[7] Hansen, M.H., Hurwitz, W.N. and Bershad, M.A. (1961), "Measurement Errors in Censuses and Surveys", Bulletin of the ISI, 38(2):359-74.

[8] Hill, C.J. (1976), "1971 Census Evaluation Programme, 1971 Response Variance Project, Methodology Report and Results", internal report, Census Survey Methods Division, Statistics Canada.

[9] MacLeod, A.D. (1978), "Investigation of Fellegi's 'Improved' Method of Estimating Correlated Response Variance", internal report, Census Survey Methods Division, Statistics Canada, April.

[10] Mahalanobis, P.C. (1946), "Recent Experiments in Statistical Sampling in the Indian Statistical Institute", Journal of the Royal Statistical Society, 109:325-70.

[11] Maurice, C., (1978), "Un nouvel estimateur de la variance corrélée de réponse", internal report, Census Survey Methods Division, Statistics Canada, August 1977, revised April.

[12] Nisselson, H. and Bailar, B.A. (1976), "Measurement, Analysis and Reporting of Non-sampling Errors in Surveys", Proceedings of the 9th International Biometric Conference, invited paper, Vol. 2:301-22, Boston, 22-27 August.

[13] United States Bureau of the Census (1968), "Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Effects of Interviewers and Crew Leaders", Series ER60, no. 7, U.S. Govt. Printing Office, Washington, D.C.

Table 1

A Comparison of 1961, 1971 and 1976 Correlated Response Variance Estimates for Characteristics Published in Fellegi (1964)

	1961 estimate x 10^4 comparable to 1971	1971 estimate x 10^4	1976 estimate x 10^4 (4)
Sex: Male	13.0	0.2	- 3.4
Age: 5	- 0.9	0.4	- 0.6
Ethnic Group: French	1688.0	151.1	-
Highest Grade of School Attended: High School, Grade 5 or Univ. (2)	23.3	9.8	-
Persons Looking For Work Last Week	40.8	5.6	-
Persons Who Usually Work 40 Hours a Week (3)	83.1	48.9	-
Industry: Manufacturing	6.7	- 27.0	-
Industry: Trade	18.5	- 3.4	-

NOTES: (1) Age, Sex and Ethnic Group were 100% variables in all three Censuses. The other variables were 100% in 1961 but sample variables in 1971 and 1976

(2) 1961 Wording. In 1971 and 1976 the estimate is for Grade 12, 13 or University.

(3) 1961 Wording. In 1971 the estimate is for Persons Who Usually Work 40-44 Hours a Week.

(4) Entries marked with - are unavailable for 1976.

Table 2

A Comparison of 1961, 1971 and 1976 Correlated Response Variance Estimates for Characteristics not Published in Fellegi (1964)

Characteristic	1961 estimate x 10^4 comparable to 1971	1971 estimate x 10^4	1976 estimate x 10^4 (1)
Official Language Spoken English only	129.2	112.5	-
Official Language Spoken French only	193.8	681.5	-
Mother Tongue English	228.2	12.0	9.8
Mother Tongue French	97.2	- 0.7	2.2
EDUCATION (Highest Grade)			
Elementary Only	93.6	- 196.0	-
High School (Grade 1 or 2)	77.0	- 52.9	-
High School (Grade 3 or 4)	6.1	- 69.3	-
AGE			
4	- 0.6	- 0.4	- 0.2
5	- 0.9	0.4	- 0.6
6	0.8	- 0.3	- 0.7
64	0.2	- 0.1	- 0.7
65	0.5	- 0.2	- 0.6
66	- 0.5	0.2	- 0.5
OTHER CHARACTERISTICS			
Relation to Head = Son	16.1	7.4	0.7
Marital Status = Married	1.6	0.8	- 1.4

(1) Entries marked with - are unavailable.