Statistics Canada Statistique Canada

SURVEY METHODOLOGY

DECEMBER 1976

VOLUME 2 — NUMBER 2

A JOURNAL PRODUCED BY
STATISTICAL SERVICES FIELD
STATISTICS CANADA

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

December/decembre 1976 Vol. 2 No. 2

A Journal produced by Statistical Services Field, Statistics Canada.

Publié par le secteur des services statistiques, Statistique Canada.

CONTENTS

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

December/decembre 1976 Vol. 2 No. 2

A Journal produced by Statistical Services Field, Statistics Canada.

Publié par le secteur des services statistiques, Statistique Canada.

Editorial Policy:

The objective of the Survey Methodology Journal is to provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology. The Survey Methodology Journal will publish articles dealing with all phases of methodological development in surveys, such as, design problems in the context of practical constraints, data collection techniques and their effect on survey results, non-sampling errors, sampling systems development and application, statistical analysis, interpretation, evaluation and interrelationships among all of these survey phases. The emphasis will be on the development strategy and evaluation of specific survey methodologies as applied to actual surveys. All papers will be refereed, however, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or the Department. Copies of papers in either Official Language will be made available upon request.

Pologique de la rédaction:

La revue Techniques d'enquête veut donner aux personnes qu'intéressent les aspects pratiques de la contuite d'enquetes, la possibilité de publier sur ce sujet dans un cadre canadien. Les textes pourront porter sur toutes les phases de l'élaboration des méthodes d'enquête: les problèmes de conception causé par des restrictions pratiques, les techniques de collecte de données et leur incidence sur les résultats, les erreurs d'observation, l'élaboration et l'application de systèmes d'échantillonnage, l'analyse statistique, l'interprétation, l'évaluation et les liens entre les différentes phases d'une enquête. On s'attachera principalement aux techniques d'élaboration et à l'évaluation de certaines méthodologies appliquées aux enquêtes existantes. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne seront pas nécessairement celles du comité de rédaction ni de Statistique Canada. On pourra se procurer sur demande des exemplaires d'un article dans l'une ou l'autre langue officielle.

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

December/decembre 1976 Vol. 2 No. 2

A Journal produced by Statistical Services Field, Statistics Canada.

Publié par le secteur des services statistiques, Statistique Canada.

Submission of Papers:

The Journal will be issued twice a year. Authors are invited to submit their papers, in either of the two Official Languages, to the Editor, Dr. M.P. Singh, Household Surveys Development Division, Statistics Canada, 10th Floor, Coats Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Two copies of each paper, typed space-and-a-half, are requested. Authors of articles for this journal are free to have their articles published in other statistical journals.

Présentation de documents pour publication:

La revue sera publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division de l'élaboration d'enquêtes ménages, Statistique Canada, 10^{e} étage, Edifice Coats, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Prière d'envoyer deux exemplaires, dactylographiés à interligne et demi. Les auteurs des articles publiés dans cette revue sont libres de les faire paraître dans d'autres revues statistiques.

DOUBLE FRAME ONTARIO PILOT HOG SURVEYS

D. Serrurier and J.E. Phillips
Institutional and Agriculture Survey Methods Division

Three Ontario pilot hog surveys were conducted in 1975 to test a sampling method based on the simultaneous use of two list frames. This paper describes the different aspects of the experience. Particular emphasis is given to the double frame methodology such as discussed by Hartley [11]. Optimal allocation of the sample between frames is considered, with revision for each following survey based on all the accumulated results.

1. INTRODUCTION

The implementation of accurate and robust methods of estimation in the livestock area has always been of particular concern for the agricultural statistician. The fast changes in the management and often short turn-around in the cycle of production, mainly for pigs and poultry, necessitate frequent and timely statistics. To provide these statistics, in the pig area, the Agriculture Division of Statistics Canada runs a quarterly survey. Questionnaires are mailed out to all farmers who reported pigs in the last Census of Agriculture. Returns are paired from one survey to the other and estimates of hog production are produced. Unfortunately the response rate is relatively poor and no sampling errors can be meaningfully associated with the estimates. Also, as the Census mailing list gets older it is becoming increasingly more difficult to keep it up to date.

The purpose of the hog pilot surveys in Ontario was to test alternate procedures based on random sampling. Since there was no reason to expect a higher response rate due to a change of design, one important objective was to keep the sample size sufficiently small in order to make an extensive follow-up possible at a reasonable cost.

First, four pilot surveys, Furrie and Wills [1], were conducted in Ontario from November 1973 to October 1974. All of them were based on replicated stratified simple random sampling using the 1971 Census as a list frame. Non-respondents were followed up by telephone. Although the coefficients of variations were reasonable given the fairly small sample size (less than 200 farms), it was felt that the estimates were too low. This was attributed to the following reasons. First, the Census frame was becoming obsolete and many farms had changed strata. A hog sample robustness study, Serrurier [2] on the evolution of hog farming between the 1966 and 1971 Censuses of Agriculture confirmed the many changes of strata and their damaging effect on any sample design that ignored them. Secondly, the zero stratum composed of farms which had no pigs in 1971, was sampled very lightly, causing possible underestimation of production from those farms which had commenced hog production since 1971. The problem with the zero stratum is that it represents a too vast reservoir (about 2/3 of the population in Ontario) of potential producers for just a few farms that move into hog production. In addition it does not include the completely new hog operations which can have been created from new agricultural activity since the last Census.

The obvious answer to both these problems of changes of strata and detection of new producers is the use of a more up to date list as a sampling frame. Unfortunately such a list does not exist. However, the Ontario Pork Producers Marketing Board (O.P.P.M.B.) maintains a list of individuals who market pigs and makes it available to Statistics Canada. This list offers the great advantage of being updated every year but does not provide a complete coverage of the population in the province. It was thought that the union of this list with the Census list (zero stratum excluded) would provide a good coverage of the population under study without the difficulties associated with the zero stratum. Unfortunately it was not possible to combine the two frames into one as there was not enough common identifying information to make this a feasible low cost operation. However, multiple frame sampling techniques, Hartley [11] allow independent sample selection in each frame, the identification of which selected units belong to one or

other or both frames and thus efficient estimation. Following Hartley (1974) three double frame hog pilot surveys have been conducted in Ontario with reference dates April 1, July 1 and October 1, 1975. These followed previous practice in that they were mail-out/mail-back surveys with telephone follow-up. On each survey occasion stratified simple random samples were drawn independently in each frame. The purpose of this paper is to describe the methods used for these surveys and give the main results of that experience.

2. SAMPLE DESIGN

2.1 The frames

The frames used for these pilot surveys were the 1971 Census of Agriculture and a list of producers obtained from the Ontario Pork Producers Marketing Board (O.P.P.M.B.). In October the 1973 list of the O.P.P.M.B. was replaced by the 1974 list. We proceed to describe these frames.

The Census frame is a list of the 30,626 Ontario farms which had at least one pig in the 1971 Census of Agriculture. The 64,096 Ontario farms without hogs at the time of the Census are not included since, if they have turned into hog businesses since 1971, one would expect to find most of them in the O.P.P.M.B. list. The exclusion of the zero stratum (farms without pigs in 1971) is an important advantage, since the high variability of this group makes it account for a substantial part of the sampling error in surveys based on the complete Census frame. Also the zero stratum has usually a low response rate making follow-up costs higher. The Census list contains the livestock numbers present on the holding at June 1, 1971 and includes the number of pigs by certain main categories. This information can be used for stratification purposes.

Next, we go on to describe the marketing board list frame. The Ontario Pork Producers Marketing Board provided us with a list of all their registered producers. However, the list does not include the few producers located in the north-west counties of the province. Apart from the latter, the list contains all those who marketed through the Board at least one hog that year or in the previous year. When a producer does not market any hogs for two consecutive years he is dropped from the list, but will be registered again if he does start to market hogs in subsequent years.

As was indicated above, the 1973 list was used for the April and July 1975 surveys. The 1974 list became available in time to use it for the October 1975 survey.

Before the O.P.P.M.B. list can be used for sampling, it must be checked for duplicates. For the survey the sampling unit is a farm but the O.P.P.M.B. list contains all those people who marketed hogs. So problems arise when several people market hogs raised on the same farm. To reduce their effect the first step was to drop from the list all those who marketed less than 5 hogs. About 45% of the 1973 list was eliminated by doing this and 43% were eliminated from the 1974 list. Of course, any farm that may have been wrongly eliminated at this stage would most likely be included in the Census frame. What remained of the lists was then sorted by surname and clerically examined to identify all possible duplicates. Possible duplicates were then checked against the 1971 Census list of farmers. If they were all on the Census list then they were all kept on the O.P.P.M.B. list as separate farms. However, if only one was on the Census list then that farm was kept on the O.P.P.M.B. list and the other individuals were deleted from the file, but the total of their hogs marketed for that year was added to the farm that was kept on the list.

When this unduplicating process was completed on the 1973 list we ended up with a list of 15, 698 farms after starting with 31,823 names. The 1974 list originally had 28,573 names but was decreased to 14,991 farms.

2.2. Stratification

Since the surveys were to be conducted by mail with telephone follow-up no travel costs had to be considered and it was decided to use stratified simple random sampling.

The stratification was carried out separately for each frame making use of the available information. However, the number of strata was limited in order to have at least 2 or 3 sample farms in each intersection, stratum x domain.

2.2.1 Stratification of the Census frame

The available information was the number of pigs by main categories at the time of the Census. From the previous pilot surveys, Furrie and Wills [1], it was evident that a cross-stratification by number of sows and hogs was the best compromise for estimations of pigs and sows at the time of the design. Also this cross-stratification was shown to be more robust, Serrurier [2],than others based on pigs alone or sows alone. It was then natural to follow the stratification used for previous surveys, but the number of strata was reduced from 10 to 8.

TABLE 1: Stratification in the Census frame

Stratum	Boundary		Population Size	Total Pigs (1971)
	No. Hogs	No. Sows		
1	> 725	> 25	139	165,523
2	300- 725	> 25	782	345,122
3	26- 299	> 25	1,592	273,292
4	> 100	1- 25	1,852	294,780
5	1- 100	1- 25	13,226	468,463
6	> 600	0	125	122,476
7	150- 600	0	1,277	347,118
8	1- 149	0	11,633	344,893
TOTAL			30,626	2,361,667

2.2.2 Stratification of the O.P.P.M.B. Frame

The information available in the 1973 and 1974 O.P.P.M.B. lists was the total number of pigs and sows marketed through the board during the year for those farms that had marketed five or more pigs. These data were used for stratification in the 1973 list. Seven strata were taken, based only on the number of hogs marketed, since sub-stratification on number of sows was found to be of little benefit.

The same stratum boundaries were used in the 1974 O.P.P.M.B. list.

TABLE 2: Stratification in the O.P.P.M.B. Frame

Stratum	Boundary (No. Hogs Sales)	Population Size	
		1973 List	1974 List
1	1500+	139	164
2	475-1499	1,186	1,254
3	225-474	1,705	1,681
4	90-224	3,232	2,931
5	50-89	2,425	2,123
6	20-49	3,284	3,039
7	5-19	3,727	3,799
Total		15,698	14,991

2.3 Estimation of Design Parameters

Following the notation of Appendix 1, the union of Frame A (Census) and Frame B (OPPMB) defines three domains:

a = Farms in the Census list only
ab = ba = Farms in both lists
b = Farms in the OPPMB list only

Since each frame is stratified independently, strata go across domains. If Ah (Ah = 1 ... K) is the hth stratum in Frame A - it is made of two parts:

aAh = Intersection of stratum Ah with domain a

abAh = Intersection of stratum Ah with domain ab

Accordingly in Frame B, the jth stratum, Bj (Bj = 1 ... L), is made of two parts bBj and baBj.

The optimal design for the estimation of a content item, Y, requires prior knowledge of population variances of Y by stratum and inter-section stratum x domain.

For the April survey the problems were as follows: First, a content item common to both frames had to be chosen. Secondly, the population variances of both frames had to be estimated without the benefit of data from previous surveys. For July the needed survey population parameters could be estimated using April data. In October the estimates could be improved by using all the data accumulated to date. We proceed to describe the estimation of parameters separately for each survey.

First, for the April design, there was no information available about the overlap portion common to both frames. Census contained numbers of pigs on the holding at June 1, 1971, and the O.P.P.M.B. list had sales of pigs and sows for the whole year 1973. The total number of pigs on the holding in 1971 was chosen as the content item. In order to estimate corresponding population variances for each stratum x domain intersection we decided to select stratified pre-samples of size 100, independently in each frame. Each pre-sample was matched against the other frame making it possible to identify all their intersections, stratum x domain. In domains a, ab and ba there was no problem in estimating population variances from these pre-samples, but the 1971 numbers of pigs were not available in domain b. So for the April design the population variances in domain b and for the whole Frame B were assumed to be equal to those for domain ba. In fact the role of the pre-samples was to initiate a process making possible, for each succeeding design, estimation of required parameters from previous surveys.

Then, for the July design, the content item was naturally chosen as the number of pigs at April 1, 1975 and corresponding population variances estimated from the April survey.

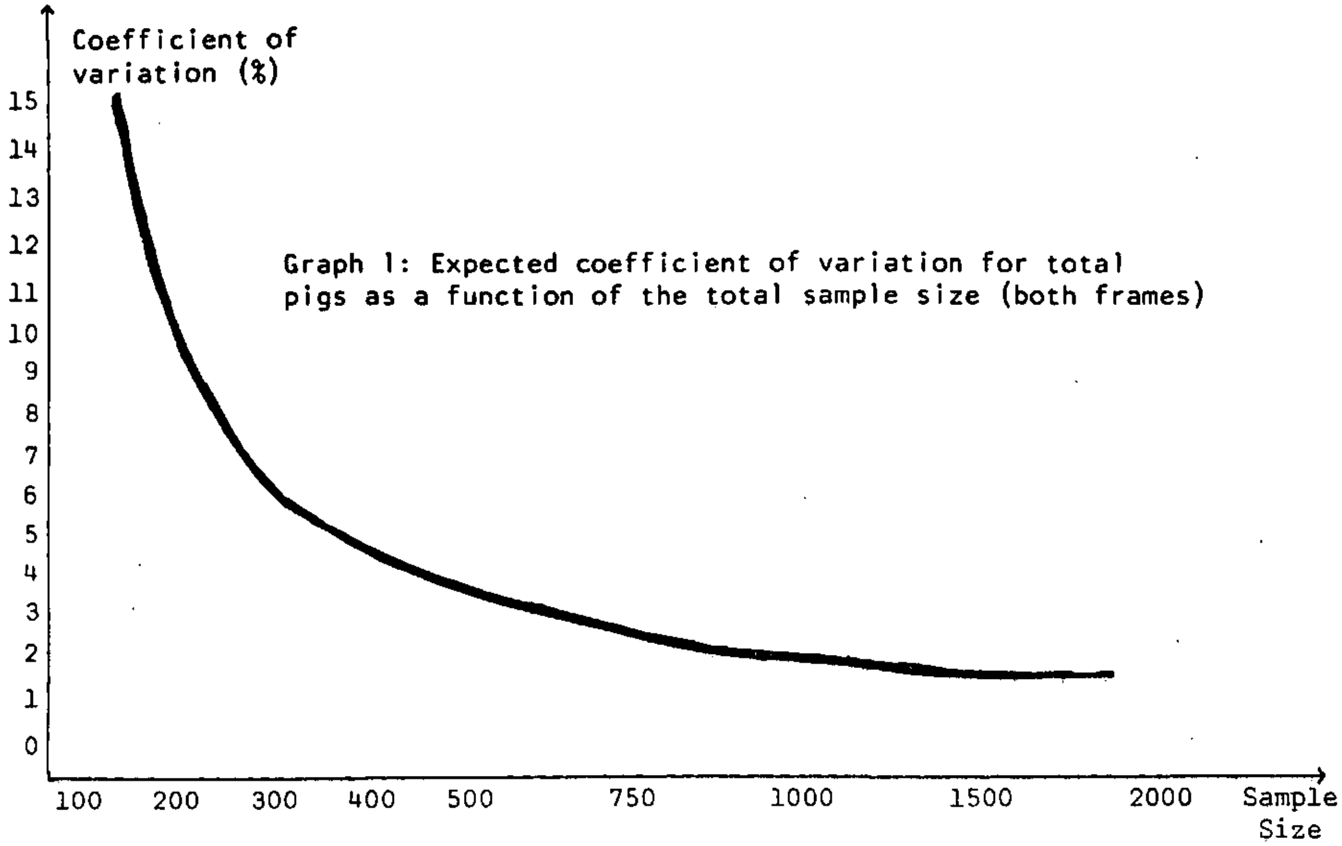
Finally, for the October design, the April and July samples were combined by considering the number of pigs at April 1 and July 1, 1975 as a common content item. It was thought that the population variances would not change very much in three months and, as a consequence, it was worthwhile to make use of more units to estimate them. It must also be noted that, considering the O.P.P.M.B. frame, these estimates came from the 1973 list while the 1974 list was effectively used in October.

2.4 Sample Sizes and Allocation

Once population variances are known the method, as described in Appendix 1, leads to optimal allocation between frames and between strata within each frame. It is also possible to estimate variances and coefficients of variation as a function of the total sample size.

In April, 226 farms were selected. In order to reduce sampling errors the sample size was slightly increased in July (to 250) and substantially in October (to 350). Graph 1 shows the expected coefficient of variation for total pigs (in the October design) as a function of the sample size. It can be seen that beyond 400-500 farms little gain in precision can be expected from a sample increase. On the other hand, the sample size must allow follow-up at a reasonable cost. Considering these constraints it seems that 400 farms is the maximum that can be considered for such a survey.

Table 3 gives sample sizes by frame and stratum for the three surveys. In each survey, the "optimal" sample sizes, as determined by the formulae, were



Graph 1: Expected coefficient of variation for total pigs as a function of the total sample size (both frames)

slightly modified to ensure a minimum number of units in each stratum. The last line of the table gives the value of p used in the design. p, as explained in Appendix 1, is the weight to apply to the estimate of the overlap domain from Frame A. There is a direct relation between p and the proportion, γ, of the total sample to allocate in Frame A. As it can also be seen in the results, p is very sensitive to population variances while γ is more robust. Thus, the pre-samples in the April design gave p = 0.37 for total pigs leading to selection of 46% of the sample in the Census list. Of course, as already noted, these pre-sample estimates of population variances were of poor quality. The April survey provided results somewhat different leading to p = 0.07 for the July design. Finally estimates of population variances from both April and July surveys gave an intermediate p = 0.29 in the October design associated with 38% of the sample allocated in Frame A.

TABLE 3: Sample Sizes by Stratum in the Three Surveys

Stratum	April 1, 1975		July 1, 1975		October 1, 1975	
	Census	1973 OPPMB	Census	1973 OPPMB	Census	1974 OPPMB
1	7	6	4	7	10	12
2	14	23	11	25	13	30
3	5	18	14	29	19	40
4	12	51	4	52	10	54
5	26	9	6	7	29	13
6	3	7	4	21	10	17
7	10	8	19	43	26	52
8	27	-	4	-	15	-
Total by frame ..	104	122	66	184	132	218
Proportion by frame .	46%	54%	26%	74%	38%	62%
Total two frames ..	226		250		350	
p optimal used in the design	0.37		0.07		0.29	

3. ORGANIZATION OF THE SURVEY

3.1 Preparing Sample

Once both the census and O.P.P.M.B. samples were selected several listings were produced. The first of these listings was used for recording the telephone numbers for the selected farms. The telephone number was not captured on the O.P.P.M.B. tape so a listing of the O.P.P.M.B. sample was sent to the Toronto office of the marketing board. They filled in the telephone numbers that they had on file. For the census farms selected, it was necessary to go back to the 1971 Census questionnaires (form 6's) and get the telephone number, if it was reported.

The selected farms have to be identified as to whether they are on both frames or not. So the O.P.P.M.B. sample must be checked against the 1971 Census list and the Census sample must be compared with the O.P.P.M.B. unduplicated list. For the July 1975 Survey using the 1973 O.P.P.M.B. list the overlap of the O.P.P.M.B. sample on Census was 66% and the census overlap on the O.P.P.M.B. was 48%. The October 1975 survey used the 1974 O.P.P.M.B. list and the overlap of the O.P.P.M.B. sample on the Census was 72% while the census sample overlapped the O.P.P.M.B. list by 43%.

Both the O.P.P.M.B. and Census samples were compared with one another to make sure that any farm that had been selected from both frames was not sent two questionnaires. There was one such farm in the April survey and two in both the July and October surveys.

3.2 Telephone Follow-up

The questionnaires were mailed out in time to reach the farmer by the first of the month. The telephone follow-up of non-respondents starts by the middle of the month, allowing 2 weeks for returning the questionnaires. A different procedure was attempted for the July 1975 survey where the respondent was told that he would be contacted by phone and the telephone follow-up was started earlier at the beginning of the month.

The marketing board was able to provide us with 54% of the telephone numbers for the October 1975 sample. About 87% of the telephone numbers were recorded on the census forms.

When we did not have a telephone number for a farmer we first had to get the area code for him and then go through the long distance operator to try and get his number. Our staff had available a set of telephone books and they seemed to have better success than they did going through the operator.

The actual telephoning of the farmers was done during the lunch hour (11.30 a.m. - 1.00 p.m.) and in the evening (5.00 p.m. - 8.00 p.m.). For the October 1975 survey we called throughout the day. Since we were dealing with farmers the season of the year seemed to have an effect on the best time to call. In the summer months the evening was the best time, but in the spring or fall, lunch time was better. In the summer it was difficult to get the farmers at home any time of the day. It was sometimes suggested that we call back early in the morning (7.00 a.m. - 7.30 a.m.) or after 10 p.m. The early morning calls were made but not the ones late at night.

From the April 1975 survey we calculated that 3.7 calls were completed per man-hour. A completed call is one where the farm operator is reached and a questionnaire is completed or a refusal is obtained. This rate seems low but includes the time spent getting the area code or telephone number and also several call backs.

At least 3 attempts were made at contacting a farmer. Since the estimates had to be produced by the end of the month, we only had about two weeks to complete the calls.

For the July 1975 survey we tried a different procedure. The farmers were told that we would be telephoning them and so the telephone follow-up started at the end of the first week. It was decided to try this method because we were telephoning most of the farmers anyway and getting the extra week in for telephoning might be helpful. As was expected, the percentage mailed back was lower (July 31%, vis-a-vis April 44%, October 44%) so more farmers had to be telephoned and also quite a few of the farmers contacted said that they had already sent in the questionnaire or would send it back, and so we did not gain any time by using this procedure.

3.3 Response Rates

The mail-back response rate was the same (44%) for the April and October surveys. The rate was lower (31%) for July because the farmers were told that we would be telephoning them to get the information and it was not necessary to mail back the questionnaires. The percentage of completed questionnaires was the same for April and October (88%) while the July survey was much lower (79%). The lower rate in July could be due to the different procedure that was followed and also the fact that the summer is the busiest time for the farmer and so the non-response rate would be higher.

TABLE 4: Returns by category for the three surveys

	April		July		October	
	Census	OPPMB	Census	OPPMB	Census	OPPMB
Mailed out questionnaires[1]	104	121	66	182	132	215[2]
Completed questionnaires mailed back[3]...........	39	50	17	46	47	93
Post Office Returns ...	} 5	} 6	1	3	5	1
Out of Business			3	7	3	3
TOTAL MAILED BACK	44	56	21	56	55	97
Completed questionnaires by phone	48	49	33	87	51	101
Refusal	3	4	1	7	} 6	} 4
Questionnaire promised but not sent back	1	6	6	6		
No telephone number....	} 8	} 6	4	22	8	6
Unable to contact			1	4	12	7
TOTAL TELEPHONE FOLLOW-UP	60	65	45	126	77	118

[1] Can be different from the sample size because of the duplicates

[2] One farm eliminated because of headquarters outside of the province of Ontario (in Quebec)

[3] July not comparable with April and October because of the special procedure.

The following table gives the response rates for the three surveys

TABLE 5: Response Rates

	April		July		October	
	Census	OPPMB	Census	OPPMB	Census	OPPMB
	%	%	%	%	%	%
Questionnaire mailed back	42	46	32	31	42	45
TOTAL	44		31		44	
Total questionnaires completed .... (mailed back & phoned)	88	86	81	78	80	92
TOTAL	88		79		88	

4. PROCESSING DATA

4.1 Imputation of missing data

First, complete imputations of data were done for those farms that had no phone number or could not be contacted by phone. These records were imputed with zeros for all fields, for the April survey. For the July and October surveys however, these farms were allowed for by not including them in the data file and instead calculating the raising factors for the strata based on the actual number of completed questionnaires.

Secondly, those farms that were refusals or promised to send a questionnaire but did not, were also allowed for by ignoring them and adjusting the raising factor accordingly.

All those farms that reported they had gone out-of-business or whose questionnaire was returned by the post office, were imputed with zeros for all the fields.

Partial imputations had to be done mostly on the last few questions which asked for the number of sows expected to farrow, or on the number of pigs under 3 months. For these partial imputations, we looked for a record which reported close to the same number of "total pigs" - when imputing for "pigs under 3 months", or close to the same number of "sows for breeding" when imputing for "sows expected to farrow". Also the good record had to be in the same frame as the record needing imputation. If several good records were found then the record that was in the same domain (a, b or ab) had preference and if a further criteria was needed then the record in the same stratum was chosen to impute from. Partial imputations were done on less than .1% of the records.

4.2 Estimation

Due to the imputation method which neglects some non-responses, the number of questionnaires processed for estimation does not correspond exactly to the number of farms selected. Of course the estimation procedure takes care of these adjustments, and at that stage "sample size" means in fact "number of questionnaires processed". Table 6 below gives these numbers by domain for each survey.

TABLE 6: Number of questionnaires processed

	CENSUS			O.P.P.M.B.			Total
	Domain a	Domain ab	Total	Domain b	Domain ba	Total	
April	54	46	100	40	72	112	212
July	23	31	54	45	98	143	197
October	63	43	106	52	147	199	305

Once the data have been captured the first stage of the estimation process is to compute estimates of population variances by inter-section, stratum x domain, in each frame.

The results will be discussed in the next section, but here we are concerned with a specific aspect of the method: To what extent the results are sensitive to variations in estimation of population parameters?

The double frame method rests on the fact that the overlap domain is estimated by a weighted combination of two independent estimates. If these estimates are quite different, which can be the case when small sample sizes are involved, a poor choice of the weight can drastically change the results. In theory the weight is calculated to minimize the overall variance, but the calculation by itself depends on the accuracy of input, that is, estimates of population variances. From selected results of the surveys, the impact of population variances on the weight p will be the first considered, then consequences of variation of p will be discussed.

4.2.1 Impact of estimates of population variances on the overlap weight p

For the item being estimated, $p(0 \leq p \leq 1)$ is the weight to apply to the estimate of the overlap domain by the sample from Frame A (Census). The estimate of the overlap domain by the sample from Frame B (OPPMB) is weighted by $q = 1-p$.

Considering the variable "total pigs" a value of $p = 0.37$ was deduced from pre-samples for the design of the April survey. The results of this survey gave a value somewhat different with $p = 0.07$. The difference can be explained by the movement of, or change in the variable between 1971 and 1975 and difficulties in estimating population variances from pre-samples for a content item common to both frames (see section 2.3). When comparing estimates of population variances from the two surveys, important discrepancies were noted. It was felt that such important changes in variability for total pigs could not have occured in a three-month period of time. It was then decided to consider pigs in April and July as a common variable and to

group the two samples for the unique purpose of estimating population variances. We thought these population variances should be fairly constant over time and the gain obtained in increasing sample sizes for their estimation should largely compensate the error due to mixing hogs in April and July. This time a value of $p = 0.29$ came out.

Table 7 gives examples of differences in estimates of population variances for each complete frame. At the domain level, differences are even more important.

TABLE 7: Examples of estimates of population variances for total pigs

Stratum	Frame A (Census)				Frame B (OPPMB)			
	Estimate from April	Estimate from July	July-April (%)	April & July Combined	Estimate from April	Estimate from July	July-April (%)	April & July Combined
1	94,240	142,100	+ 51	98,920	268,500	549,300	+105	384,600
2	60,800	78,950	+ 30	65,960	58,390	53,410	- 9	56,630
3	21,190	54,080	+155	42,180	34,460	38,550	+ 12	36,220
4	10,450	2,760	- 74	8,845	31,050	19,540	- 37	25,890
5	3,202	1,075	- 66	2,780	1,017	6,000	+490	3,336
6	10,900	84,270	+673	62,030	4,921	662	- 87	2,055
7	39,580	66,080	+ 67	52,840	17,650	24,180	+ 37	23,250
8	1,031	320	- 69	945	-	-	-	-

When this type of survey was initiated we thought that a common value of p might be used for all items. In fact, due to the sensitivity of p, it was necessary to compute different p's for the estimation of the different items. Thus in the October survey p ranged from 0.12 for "other pigs" to 0.69 for "sows expected to farrow from October to December". One immediate consequence is that the direct estimate of total pigs does not equal the sum of its three components. But this problem is not specific to the method and for the three surveys the two estimates of total pigs were very close.

After the July survey the April and July samples were combined to estimate population variances for the 8 items. July estimates used these variances and April estimates were revised to incorporate them. Also the October design was prepared with these same variances. However, the processing of the October survey was based on population variances estimated from the October sample only. On the one hand the sample increase was expected to provide better estimates of population variances, and on the other the change of O.P.P.M.B. list made previous estimates inaccurate in Frame B.

Another aspect to be considered when determining p is that the expression of the variance to minimize is not the same at the processing stage as at the design stage (see Appendix 1). At the design stage one puts oneself in the situation of finding an optimal design for the estimation of a certain content item. But practical constraints (adjustments, non-responses ... etc.) can alter this optimal design and the item estimated can be quite different from the content item on which the design was based (for example, estimation of boars from a design based on total pigs).

4.2.2 Impact of the overlap weight p on the estimates

Since p is very sensitive to variations in population parameters and to departures from "optimality" in the design it is important to measure its impact on the estimates.

Following notations of Appendix 1 the double frame estimate is given by:

$$\dot{Y} = \hat{Y}_a + \tilde{Y}_b + p\hat{Y}_{ab} + q\tilde{Y}_{ba} \qquad (p+q=1)$$

The importance of p is a function of the difference between the two estimates of the overlap domain $\hat{Y}_{ab}$ and $\tilde{Y}_{ba}$. The larger this difference is, the more important is the role played by p.

The two extremes correspond to p=1 (overlap estimated only from Frame A) and to p=0 (overlap estimated only from Frame B).

$$p = 1: \dot{Y} = \hat{Y}_a + \tilde{Y}_b + \hat{Y}_{ab}$$

$$p = 0: \dot{Y} = \hat{Y}_a + \tilde{Y}_b + \tilde{Y}_{ba}$$

Table 8 gives, as examples, the two overlap estimates for four items in the October survey with the value of p adopted and the corresponding final estimate of the overlap.

TABLE 8: Estimates of the overlap for selected items in the October survey

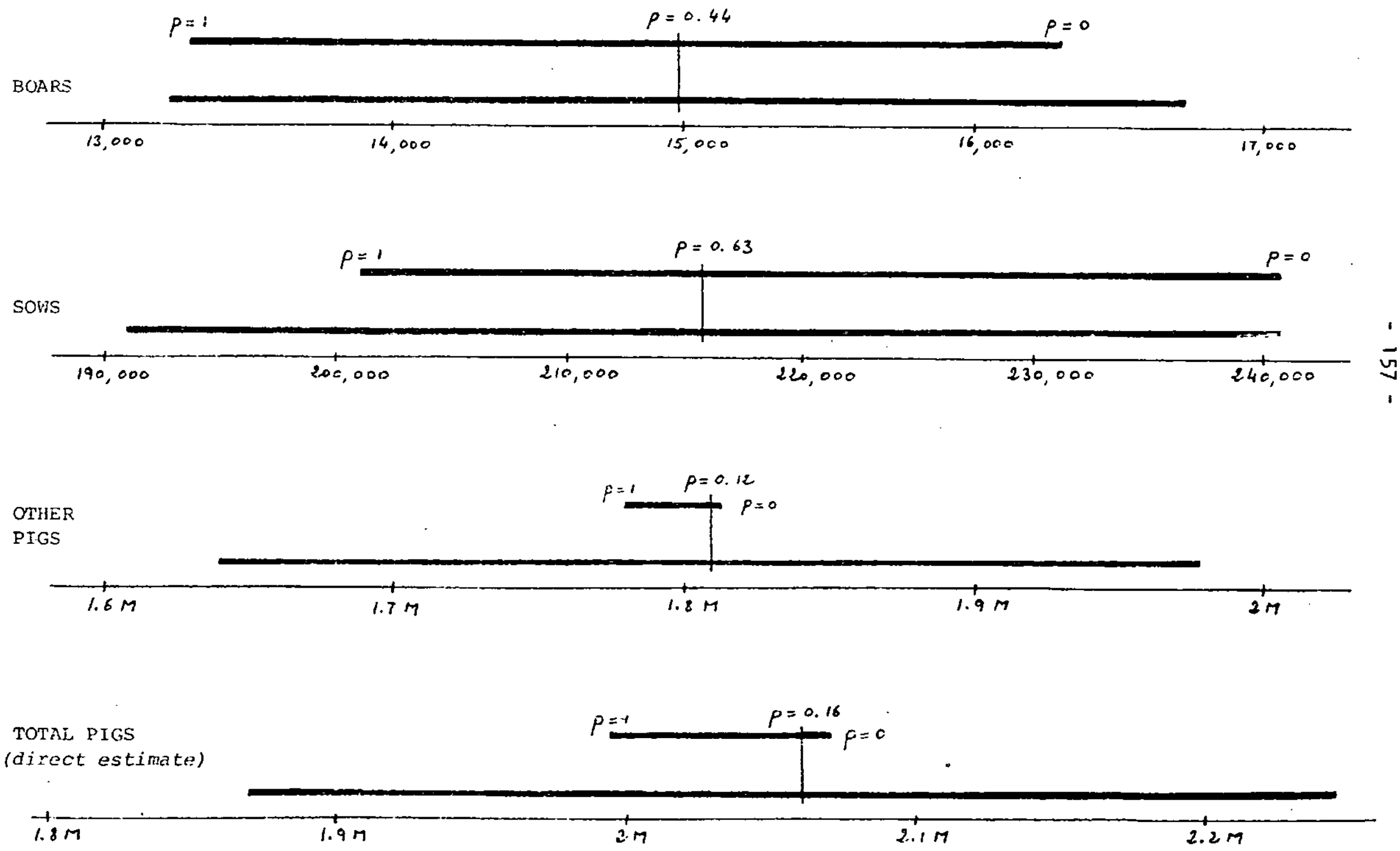
	$\hat{Y}_{ab}$	$\tilde{Y}_{ba}$	p	$p\hat{Y}_{ab} + q\tilde{Y}_{ba}$
Boars	5,214	8,211	0.44	6,892
Sows	76,662	116,322	0.63	91,336
Other pigs	883,620	916,109	0.12	912,210
Total pigs (direct estimate)	965,497	1,040,642	0.16	1,028,619

In the examples of Table 8 the weight p plays a much more important role for Boars and Sows than for Other pigs. It is due to the fact that the two basic estimates of the overlap are relatively closer for Other pigs.

This is illustrated in Graph 2 by comparing the range associated with extreme values of p to the confidence interval at one standard deviation. In the figure, overall double frame estimates are considered, that is including non-overlap domains a and b. For each item the first line gives the range of estimates associated with values of p from 1 to 0, while the second line gives the confidence interval at one standard deviation associated with the double frame estimate with p given in Table 8.

Graph 2: Estimation in the October survey as a function of the overlap weight p

(For each item the second line gives ± 1 standard deviation of the estimate associated with the intermediate value of p)



For boars and sows the possible range of point estimations, depending on p, is to the extent of almost two standard deviations. For other pigs the correct choice of p is much less critical since the possible range is about 1/4 of a standard deviation. For total pigs, since it is made at almost 90% of other pigs, the situation is equivalent to the latter.

It is difficult to draw a conclusion from these contradictory examples except that in some cases a poor choice of p can have a dramatic effect. It seems important in such applications to give much attention to this weighting problem.

5. RESULTS AND COMPARISON WITH OTHER ESTIMATES

The results are shown in Table 9 along with the estimates from Agriculture Division's quarterly hog survey for comparison.

The total pigs estimate for the July pilot was very close to that produced by the quarterly survey. However the April and October estimates were somewhat higher than the quarterly result. Since the quarterly survey uses only the 1971 Census as a frame it is felt that this might have an underestimating effect on the results and so it was not surprising that the pilot estimates were higher. Also looking at the individual estimates the biggest difference between the surveys usually occurs in the "pigs under 3 months" category followed by the "Sows for breeding" category.

The coefficients of variation are high for the pilot estimates but as has been pointed out earlier a large increase in the sample size would be needed in order to reduce the variance.

The variability between the estimates for the three pilot surveys could be explained by the fact that 3 entirely new samples were selected for each survey. If rotation was introduced and part of the sample was the same for all three surveys then the "total pigs" estimates might be closer.

TABLE 9: Results of the 1975 Double Frame Pilot Hog Surveys
(All figures in 1000's)

	April 1, 1975				July 1, 1975				October 1, 1975			
	Pilot Survey		Agr. Div. Est.	%[1] Diff.	Pilot Survey		Agr. Div. Est.	%[1] Diff.	Pilot Survey		Agr. Div. Est.	%[1] Diff.
	Est.	C.V. (%)			Est.	C.V. (%)			Est.	C.V. (%)		
Boars for breeding 6 months old and older	15.6	10.8	15.0	+ 4	15.2	14.7	14.0	+ 8	15.0	11.7	14.0	+ 7
Sows and gilts for breeding 6 months and older	236.8	11.5	205.5	+ 15	228.9	14.1	198.0	+ 16	215.8	11.5	208.0	+ 4
All other pigs	1,840	11.2	1,630	+ 13	1,608	11.4	1,620	- 1	1,809	9.3	1,650	+ 10
Pigs under 3 months	899.7	13.0	740.0	+ 22	940.2	12.8	870.0	+ 8	911	11.4	780	+ 17
Sows having farrowed in the last 3 months	98.4	13.0	91.0	+ 8	114.5	14.8	100.0	+ 14	92.6	11.5	91.0	+ 2
Sows expected to farrow in the next 3 months	107.7	12.9	96.0	+ 12	95.2	15.9	90.0	+ 6	99.0	12.0	90.0	+ 10
Sows expected to farrow in the next 4th to 6th months.	71.7[2]	16.9[2]	93.0	- 23	102.8	15.4	94.0	+ 9	100.3	12.5	94.0	+ 7
TOTAL PIGS[3]..................	2,092	10.8	1,850	+ 13	1,852	11.2	1,832	+ 1	2,040	9.2	1,872	+ 9

[1] This is the difference between the pilot survey estimate and the Agriculture Division estimate as a percentage of the Agriculture Division estimate.

[2] This is not a reliable estimate because the question was missing on some questionnaires.

[3] The direct estimates for "Total pigs", which are the sums of the first three items, were: 2,107 in April, 1,866 in July and 2,058 in October.

6. CONCLUSION

If multiple frame surveys are not new, they apply very often to an area frame associated with a cheaper (but incomplete) list frame. Also in most applications only one estimate of the overlap domain is considered (i.e. p=0 or 1) or a pre-assigned weight is used (e.g. p=0.5 if both frames are of equal cost). To that respect we think that the Ontario Pilot Hog Survey is a fairly new type of double frame application. The main feature of this survey is the deliberate reduction in the complete coverage of the Census frame (by elimination of the zero stratum) and in effect to replace it by the much smaller and updated O.P.P.M.B. list. The advantage of such a procedure is not only to eliminate from the sampling frame about 50,000 farms which are not likely to have any hogs (zero stratum farms still without hogs some years later) but also to take advantage of the updated information available in the O.P.P.M.B. list. The other new aspect in this application is an attempt to optimize the design by allocating the sample between frames according to the double frame theory and revising the weight p every time new information becomes available. In this application it has been discovered that the weight p to apply to the overlap domain is very sensitive to changes in population parameters and it is necessary to pay a great deal of attention to the estimation of these parameters if one wants to make the best use of the multiple frame method.

The survey procedure proved to be workable and we believe that it is a good answer to the old problem of the zero stratum. Unfortunately, at the moment, some questions are unanswered and some results are unsatisfactory. Among the unanswered questions we can note the inability in the April survey, to measure an eventual bias due to poor matching. One expects multiple frame estimates to be biased upwards. Is this what happened here, due to residual duplication, or are the current estimates too low? A partial answer to that question might be obtained by asking selected Census farmers if they had marketed any hogs through the Board. However, one possible

justification of higher estimates in the double frame survey is that some new operations which are found in the O.P.P.M.B. list did not exist at Census time. Another question of interest is the population coverage by the union of the two lists. We have assumed that zero stratum farms which have turned into hog business since 1971 could be found in the O.P.P.M.B. This assumption is probably true for big producers, but to what extent is there undercoverage for small farms? The question is much more difficult to answer since any verification would require sampling the zero stratum with the same trouble as in previous experiences.

Indeed, the main concern with these pilot surveys comes from the relatively high sampling errors, and as a consequence the instability of estimates from one survey to the other. The problem does not come from the double frame approach, which would rather be good for it, but is due to the high variability of the commodity under study. Anyhow the problem is that potential users will probably regard such sampling errors as excessive. We have seen that above 300-400 farms, a slight reduction in coefficients of variation will require a substantial sample size increase. Such an increase is not worth envisaging since, as already mentioned, the advantage of such a survey is to work with reasonable sample sizes which allow extensive follow-up and reduced response burden. The answer to this problem will probably be found with semi-permanent samples. In fact when a solid benchmark is available (the Census for example) one needs to measure movements or changes and that is what is attempted in the regular quarterly survey. It must be noted that for these three double frame pilot hog surveys completely new samples were selected at each occasion, in large part explaining the high instability of estimates. We think that semi-permanent samples associated with the double frame approach should provide more accurate estimates of the evolution of hog production.

Concerning evolutions, a completely permanent sample would eventually look better but is not envisageable because on the one hand it would not be

possible to take advantage of the regular updating of the O.P.P.M.B. list and on the other the burden imposed on the same group of farmers would rapidly have a negative impact on the response rate. A partial replacement of farms would be a compromise between a completely new sample and a completely permanent sample. Of course the optimal scheme of rotation has still to be worked out but it should allow the marrying of the actual advantages of the double frame method with the production of timely and accurate statistics.

The authors wish to thank Dr. M. P. Singh and the referee for some helpful comments.

RESUME

Le but des trois enquêtes pilotes sur les porcs, réalisées en Ontario en 1975, était de tester une méthode d'échantillonnage à partir de deux listes. Le présent exposé décrit les différents aspects de cette expérience. On insiste plus particulièrement sur la méthodologie des doubles bases de sondage telle qu'elle a été décrite par Hartley [11]. On considère une répartition optimale de l'échantillon entre les deux listes, avec révision pour chaque enquête lorsque les résultats précédents peuvent être utilisés.

REFERENCES

[1] Furrie, A.D. & Wills, B.L. (1974), "Ontario Pilot Hog Survey", Census & Institutional Survey Methods Division report, Statistics Canada, February 1974.

[2] Serrurier, D. (1975), "Hog Sample Robustness in Ontario", Census & Institutional Survey Methods Division report, Statistics Canada, January 1975.

[3] Harley, H.O. (1962), "Multiple Frame Surveys", Proceedings of the Social Statistics Section, American Statistical Association meeting, Minneapolis, Minnesota.

[4] Cochran, Robert S. (1964), "Multiple Frame Sample Surveys", Proceedings of the Social Statistics Section, Americal Statistical Association meeting.

[5] Steinberg, Joseph (1965), "A Multiple Frame Survey for Rare Population Elements", Proceedings of the Social Statistics Section, American Statistical Association meeting.

[6] Cochran, Robert S. (1967), "The Estimation of Domain Sizes when Sampling Frames are Interlocking", Proceedings of the Social Statistics Section, American Statistical Association meeting, Washington, D.C.

[7] Lund, Richard E. (1968), "Estimators in Multiple Frame Surveys", Proceedings of the Social Statistics Section, American Statistical Association meeting, Pittsburg, Pennsylvania.

[8] Fuller, W.A. & Burmeister, L.F. (1972), "Estimators for Samples Selected from Two Overlapping Frames", Proceedings of the Social Statistics Section, American Statistical Association meeting, Montreal, Quebec.

[9] Rao, J.N.K. (1972), "Maximum Likelihood Estimation in Multiple Frame Surveys", (unpublished note).

[10] Vogel, Frederic A. (1973), "An Application of a Two-Stage Multiple Frame Sample Design", Proceedings of the Business and Economic Statistics Section, Americal Statistical Association meeting.

[11] Hartley, H.O. (1974), "Multiple Frame Methodology and Selected Applications", Sankhya, Volume 36, Series C, Pt. 3, pp. 99-118.

[12] Burmeister, Leon F. (1974), "Sampling from Two Overlapping Frames when the Domain Sizes are Known", Proceedings of the Social Statistics Section, American Statistical Association meeting, St. Louis, Missouri.

[13] Vogel, F.A. (1975), "Surveys with Overlapping Frames - Problems in Application", Presented at the 135th Annual Meeting of the American Statistical Association, Atlanta, Georgia.

APPENDIX 1: Double Frame Methodology

The method described here is a straightforward application of the general model discussed by Hartley (1974) in [11]. More details and proofs can be found in Hartley's paper.

We consider the special case of overlapping fractions of the same population and when stratified simple random samples of size n_A and n_B are drawn respectively from Frames A and B. Due to the survey procedure the two frames are assumed to be of equal cost per sampling unit and the objective is to optimize the design for a total given sample size $n = n_A + n_B$.

Following Hartley's notations, the two frames A (Census) and B (QPPMB) define three domains:

a = Farms in A but not in B
ab = ba = Farms in both frames
b = Farms in B but not in A

The two frames have not been matched and we are in the situation where the number of population units in the overlap domain is not known.

At the sample level ab and ba refer to the overlap domain estimation from frame A sample and frame B sample respectively.

Let Y be the quantity to estimate (e.g. pigs, or sows ...)

$$Y = Y_a + Y_b + Y_{ab} \text{ and is estimated by:}$$

$$\dot{Y} = \hat{Y}(y_a,\alpha) + \tilde{Y}(y_b,\beta) + p\hat{Y}(y_{ab},\alpha) + q\tilde{Y}(y_{ba},\beta) \qquad (1)$$

$$(p+q=1)$$

$\hat{Y}$ and $\tilde{Y}$ being the single frame estimates respectively for A and B.

FRAME A

Stratum	Pop. Size	Sample		
		Size	Mean	Total
1	N_1	n_1	$\bar{y}_1$	y_1
2	N_2	n_2	$\bar{y}_2$	y_2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
A_h	N_{Ah}	n_{Ah}	$\bar{y}_{Ah}$	y_{Ah}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
K	N_K	n_K	$\bar{y}_K$	y_K
	N_A	n_A		

FRAME B

Stratum	Pop. Size	Sample		
		Size	Mean	Total
1	N_1	n_1	$\bar{y}_1$	y_1
2	N_2	n_2	$\bar{y}_2$	y_2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
B_j	N_{Bj}	n_{Bj}	$\bar{y}_{Bj}$	y_{Bj}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
L	N_L	n_L	$\bar{y}_L$	y_L
	N_B	n_B		

Due to stratified simple random sampling:

$$\hat{Y}(y_A,\alpha) = \sum_{Ah=1}^{K} N_{Ah}\,\bar{y}_{Ah}, \quad \tilde{Y}(y_B,\beta) = \sum_{Bj=1}^{L} N_{Bj}\,\bar{y}_{Bj}$$

Then considering that each domain a, ab, and b goes across strata:

$$\hat{Y}(y_a,\alpha) = \sum_{Ah=1}^{K} \frac{N_{Ah}}{n_{Ah}}\, y_{aAh}, \quad \hat{Y}(y_{ab},\alpha) = \sum_{Ah=1}^{K} \frac{N_{Ah}}{n_{Ah}}\, y_{abAh}$$

$$\tilde{Y}(y_b,\beta) = \sum_{Bj=1}^{L} \frac{N_{Bj}}{n_{Bj}}\, y_{bBj}, \quad \tilde{Y}(y_{ba},\beta) = \sum_{Bj=1}^{L} \frac{N_{Bj}}{n_{Bj}}\, y_{baBj}$$

When reporting these values in (1) we get the estimator $\dot{Y}$[1]. Using the two relations $y_{A.} = y_a + y_{ab}$ and $y_B = y_b + y_{ba}$ to spell out the covariances, the variance of Y is given by:

$$\begin{aligned} \text{Var }(\dot{Y}) = \; & p\, V_A(y_A,\alpha) + q\, V_A(y_a,\alpha) - pq\, V_A(y_{ab},\alpha) \\ & + q\, V_B(y_B,\beta) + p\, V_B(y_b,\beta) - pq\, V_B(y_{ba},\beta) \end{aligned} \qquad (2)$$

$V_A(y_A,\alpha)$ and $V_B(y_B,\beta)$ denoting single frame variance formulae respectively for A and B.

Formula (2) can be written, introducing a simplified notation, in the form:

$$\text{Var }(\dot{Y}) = V_A(\alpha,p) + V_B(\beta,p) \qquad (3)$$

where $V_A(\alpha,p)$ and $V_B(\beta,p)$ denote respectively the three terms expressions on the first and second line of (2).

α and β being the vectors of stratum sample sizes respectively in frame A and B, we have to minimize Var $(\dot{Y})$ subject to a given total sample size n.

The minimization problem will be solved in three stages:

1. $\min\limits_{\alpha,\beta} [V_A(\alpha,p) + V_B(\beta,p)]$ for given p, n_A and n_B

   to yield, in fact, two conditionally minimum variances:

   $V_A(n_A,p)$, $V_B(n_B,p)$

[1] By consideration of duplicated items in the overlap samples, this estimator has been improved by Lund [7], Fuller and Burmeister [8] and Rao [9]. However, there were too few duplicated items to consider these improvements in the Ontario hog surveys.

2. $\min_{n_A, n_B} [V_A(n_A,p) + V_B(n_B,p)]$ for given p and $n = n_A + n_B$

   to yield a minimum variance conditional on p; V(p).

3. $\min_{p} V(p)$ for given n.

STAGE 1

Stage 1 can be split into two disjoint minimization problems:

$\min_{\alpha} V_A (\alpha,p)$ for given p and n_A

and

$\min_{\beta} V_B (\beta,p)$ for given p and n_B

If we denote S^2_{Ah}, S^2_{aAh}, S^2_{abAh} the population variances in stratum Ah respectively for the whole stratum, domain a and domain ab, let Σ^2_{Ah} be the quantity:

$$\Sigma^2_{Ah} = p\ S^2_{Ah} + q\ S^2_{aAh} - pq\ S^2_{abAh} \tag{4}$$

then $$V_A(\alpha,p) = \sum_{Ah=1}^{K} N_{Ah} \frac{N_{Ah}-n_{Ah}}{n_{Ah}} \Sigma^2_{Ah} \tag{5}$$

S^2_{aAh} and S^2_{abAh} can be computed accordingly to domain estimation procedures by using an auxiliary variable taking the given Y value in the domain and zero out of the domain.

Minimization of $V_A(\alpha,p)$ leads to the familiar Neyman's solution:

$$\boxed{n_{Ah} = n_A \frac{N_{Ah}\ \Sigma_{Ah}}{\sum_{Ah=1}^{K} N_{Ah}\ \Sigma_{Ah}}} \tag{6}$$

With analogous notation for frame B, we get:

$$\Sigma_{Bj}^{2} = q\ S_{Bj}^{2} + p\ S_{bBj}^{2} - pq\ S_{baBj}^{2} \tag{7}$$

$$V_{B}(\beta,p) = \sum_{Bj=1}^{L} N_{Bj}\ \frac{N_{Bj} - n_{Bj}}{n_{Bj}}\ \Sigma_{Bj}^{2} \tag{8}$$

and

$$\boxed{n_{Bj} = n_{B}\ \frac{N_{Bj}\ \Sigma_{Bj}}{\sum\limits_{Bj=1}^{L} N_{Bj}\Sigma_{Bj}}} \tag{9}$$

Minimum variances conditional on p, n_A and n_B are given by:

$$\begin{aligned} V_{A}(n_{A},p) &= \frac{\left(\sum\limits_{Ah=1}^{K} N_{Ah}\Sigma_{Ah}\right)^{2}}{n_{A}} - \sum_{Ah=1}^{K} N_{Ah}\ \Sigma_{Ah}^{2} = \frac{A(p)}{n_{A}} - a(p) \\ V_{B}(n_{B},p) &= \frac{\sum\limits_{Bj=1}^{L} N_{Bj}\Sigma_{Bj}}{n_{B}} - \sum_{Bj=1}^{L} N_{Bj}\ \Sigma_{Bj}^{2} = \frac{B(p)}{n_{B}} - b(p) \end{aligned} \tag{10}$$

STAGE 2

The second step is to minimize:

$$V_{A}(n_{A},p) + V_{B}(n_{B},p) \text{ for } n_{A} \text{ and } n_{B} \text{ given p and } n = n_{A} + n_{B}$$

Let γ be the ratio: $= \dfrac{n_A}{n_A + n_B}$, i.e. proportion of the sample drawn in frame A.

$$\min_{n_A, n_B} [V_A(n_A,p) + V_B(n_B,p)] = \min_{\gamma} [\frac{A(p)}{\gamma n} - a(p) + \frac{B(p)}{(1-\gamma)n} - b(p)] \text{ given p and n}$$

The solution is given by:

$$\boxed{\gamma^* = \frac{\sqrt{A(p)}}{\sqrt{A(p)} + \sqrt{B(p)}}} \qquad (11)$$

and the minimum variance conditional on p and n is:

$$V(p) = \frac{1}{n} [\sqrt{A(p)} + \sqrt{B(p)}\,]^2 - a(p) - b(p) \qquad (12)$$

A(p), B(p), a(p), b(p) being defined in relations (10)

STAGE 3

The third step is to minimize V(p) as it appears in (12) for p given n.

An analytic minimization of v(p) is only feasible in special cases. A numerical method is given by Hartley in [11].

AN APPROXIMATION TO THE INVERSE MOMENTS OF THE HYPERGEOMETRIC DISTRIBUTION

M.A. Hidiroglou
Business Survey Methods Division

The negative moments of the positive hypergeometric distribution are often approximated by the inverse of the positive moments of this distribution. In this paper, a suitable approximation to the positive hypergeometric distribution is used to obtain the negative moments.

1. INTRODUCTION

The use of the negative moments of the positive hypergeometric distribution are given in books by Sukhatme and Sukhatme (1970) and papers by Rao (1973). Here, the first negative moment is required to obtain the variances of estimators incorporating post-stratification.

In our case, the need for a good approximation to the inverse moments of the positive hypergeometric distribution arose out of some work on outliers, Hidiroglou (1976) and Srinath (1976). Here, a simple random sample of size n is drawn without replacement from a population $\phi = \{Y_1, Y_2, \ldots, Y_N\}$ of size N which contains T outliers. These outliers are elements of ϕ whose Y value exceeds a given value γ. The sample is found to contain t outliers. The variable t has the hypergeometric distribution

$$P^*(t) = \binom{N-T}{n-t}\binom{T}{t}\bigg/\binom{N}{n}$$

In this type of problem, we will be interested in the mean square error (MSE) of the estimator for the total Y, where the estimator is given by

$$\hat{Y} = \sum_{i=1}^{t} y_i + \frac{N-t}{n-t} \sum_{i=1}^{n-t} y_i \qquad \text{if } t < n \tag{1.1}$$

$$= \sum_{i=1}^{n} y_i \qquad \text{if } t = n$$

The MSE of this estimator involves an expected value of $(n-t)^{-1}$. This is equivalent to obtaining the inverse moment of z, where z = n-t, and z has the hypergeometric distribution given by

$$P^*(z) = \binom{N-T}{z}\binom{T}{n-z}/\binom{N}{n} \tag{1.2}$$

In what follows, we work with the negative moments t rather than those of z. Results associated with t can be applied to z. We use the positive hypergeometric distribution because inverse moments are not defined at t=0. The positive hypergeometric distribution of t is given by

$$P(t) = P^*(t)/(1-c);\ 1 \le t \le T;\ N-T \ge n \tag{1.3}$$

where

$$c = P^*(o) = \prod_{i=o}^{T-1} \left(1 - \frac{n}{N-i}\right)$$

The exact expected value of t^{-k} may be written as

$$E(t^{-k}) = \frac{1}{1-c} \sum_{t=1}^{\min\,(T,n)} t^{-k} \frac{\binom{N-T}{n-t}\binom{T}{t}}{\binom{N}{n}} \tag{1.4}$$

where k is a positive integer. This expected value can be obtained exactly by calculating (1.4) directly. However, for any but small values of T, the computations can become quite tedious. We provide an approximation that is better than the one given by Sukhatme and Sukhatme (1970), particularly for moderate values of T.

2. THE APPROXIMATION TO $E(t^{-1})$ AND $E(t^{-2})$

The most widely used approximation for $E(t^{-1})$ is the inverse of the first moment of the hypergeometric distribution (see e.g. Sukhatme and Sukhatme, 1970, p.29). That is,

$$E(t^{-1}) \doteq E(t)^{-1} = N/nT \qquad (2.1)$$

In the case of the second inverse moment, we might generalize the above to the following approxiamtion,

$$E(t^{-2}) \doteq (Et^{2})^{-1} = [\frac{n}{N} \frac{n-1}{N-1} T(T-1) + \frac{n}{N} T]^{-1} \qquad (2.2)$$

Following Mendenhall and Lehman (1960), we first approximate the distribution of t and then find the exact value of $E(t^{-k})$, k=1,2, for the approximating distribution. The Beta distribution is a good candidate to consider as an approximation to the hypergeometric distribution.

The Beta distribution is given by

$$h(z) = \frac{z^{a-1} (1-z)^{b-1}}{B(a,b)} ;0 \leq z \leq 1; a,b,> 0 \qquad (2.3)$$

where

$$B(a,b) = \frac{\Gamma(a)\ \Gamma(b)}{\Gamma(a+b)}$$

$$\Gamma(g) = \int_{o}^{\infty} t^{g-1}\ e^{-t}\ dt.$$

When a and b are integers, then

$$B(a,b) = \frac{(a-1)!\ (b-1)!}{(a+b-1)!} \quad .$$

Let $z = t/T$ where t is the hypergeometric variate, so that $0 \leq z \leq 1$. The parameters of the approximating Beta function are obtained by equating the first two positive moments of the positive hypergeometric to those of the Beta distribution. We obtain the following two equations:

$$\frac{nT}{N(1-c)} = \frac{Ta}{a+b} \tag{2.4}$$

and

$$\frac{n(n-1)\ T(T-1)}{(1-c)\ N(N-1)} + \frac{nT}{N(1-c)} = \frac{T^2 a(a+1)}{(a+b)(a+b+1)} \tag{2.5}$$

Solving questions (2.4) and (2.5), we get

$$a = \frac{f(1-f)\ (T-1)}{(1-c)(1-f)-c\ fT - \frac{T}{N}(1-f-c)} \tag{2.6}$$

and

$$b = f^{-1}\ (1-f-c)\ a \tag{2.7}$$

where $f = n/N$.

The k^{th} negative moment of the Beta distribution (2.3) is given by

$$E(z^{-k}) = B(a-k,b)/B(a,b)\ . \tag{2.8}$$

In order for the kth negative moments (2.8) to exist, it is necessary that $a > k$ or

$$T \overset{\bullet}{>} \frac{(1-f)(f+k-kc)}{f(1-f+kc) - \frac{k}{N}(1-f-c)} \tag{2.9}$$

We next proceed to provide expressions for the first and second inverse moments. The first inverse moment is approximated by

$$E(t^{-1}) \quad = \quad T^{-1} \ \frac{(a+b-1)}{a-1} \tag{2.10}$$

and the second inverse moment is approximated by

$$E(t^{-2}) \quad = \quad T^{-2} \frac{(a+b-1)(a+b-2)}{(a-1)(a-2)} \tag{2.11}$$

where a and b are found from (2.6) and (2.7). We next proceed to discuss our findings which are tabulated in Tables 2 and 3.

3. COMPARISON OF THE BETA DISTRIBUTION FOR k=1 AND k=2

In calculating the exact inverse moments of the positive hypergeometric distribution, we used an alternate form of Stirling's approximation to compute factorials. The approximation used was

$$n! \quad = \quad (2\Pi)^{\frac{1}{2}} \quad n^{n+\frac{1}{2}} \exp\{-n+ \frac{1}{12n} - \frac{1}{360n^3} \}. \tag{3.1}$$

Table 1

Stirling's Alternate Approximation

n!	Stirling	True
2	$1.999957498\text{x}10^0$	$2.000000000\text{x}10^0$
3	$5.999981842\text{x}10^0$	$6.000000000\text{x}10^0$
4	$2.399998221\text{x}10^1$	$2.400000000\text{x}10^1$
5	$1.199999703\text{x}10^2$	$1.200000000\text{x}10^2$
6	$7.199999258\text{x}10^2$	$7.200000000\text{x}10^2$
7	$5.039999763\text{x}10^3$	$5.040000000\text{x}10^3$
8	$4.031999908\text{x}10^4$	$4.032000000\text{x}10^4$
9	$3.628799927\text{x}10^5$	$3.628800000\text{x}10^5$
10	$3.628799952\text{x}10^6$	$3.628800000\text{x}10^6$

The comparison of the approximation $E(t^{-1})$ with the exact values for various combinations of T and n, given a fixed N, are given in Table 2. The three tabulated values are the exact inverse moment (1.4) the Beta inverse moment (2.10) and the inverse moment obtained using approximation (2.1). We do not list values when n>N-T. We have chosen N=200.

Table 2

Comparison of the Tabulated Value of $E(t^{-1})$, the Beta Approximation and the Approximation given in Equation (2.1) for N=200
Entries: Exact Value (1.4)
Beta Approximation (2.10)
Approximation (2.1)

T	n	10	25	35	80	120	150
2		.989400 1.000000 20.000000	.968892 1.000000 4.000000	.954195 1.000000 6.666666	.876921 1.000000 2.857142	.787307 1.000000 .833333	.701201 1.000000 .666666
3		.977373 1.009769 6.666667	.936683 1.005076 2.666666	.908042 .998464 1.904761	.763719 .916483 .833333	.616258 .746828 .555556	.500181 .574723 .444444
4		.965769 1.013323 5.000000	.905308 .997955 2.000000	.863344 .977884 1.428571	.663069 .782951 .625000	.486383 .540028 .416667	.373046 .392584 .333333
5		.954192 1.017070 4.000000	.873887 .985132 1.599999	.819720 .945869 1.142857	.575107 .653575 .500000	.390670 .411007 .333333	.292299 .299004 .266667
10		.896502 1.019812 2.000000	.727954 .851955 .800000	.624996 .711932 .571429	.300761 .302813 .250000	.180724 .181781 .166667	.138341 .139130 .133333
15		.839422 .972633 1.333333	.600462 .679976 .533333	.472919 .503265 .389952	.188162 .188246 .166667	.116491 .117117 .111111	.090901 .091228 .088889
25		.727954 .851955 .800000	.405628 .422028 .320000	.282478 .283026 .228571	.106221 .106977 .100000	.068340 .068657 .066667	.053968 .054118 .053333
30		.676013 .781998 .666667	.336426 .341042 .266667	.227704 .227946 .190476	.087390 .088050 .083333	.056676 .056911 .055556	.044880 .044980 .044444
35		.624996 .711932 .571429	.282478 .283026 .228571	.189113 .190204 .163265	.074237 .074830 .071429	.048400 .048601 .047619	.038397 .038484 .038095

Table 3

Comparison of the Tabulated Value of $E(t^{-2})$, the Beta Approximation and the Approximation given in Equation (2.2) for N=200

Entries: Exact Value (1.4)
Beta Approximation (2.11)
Approximation (2.2)

T	n	10	25	35	80	120	150
2		.983736 1.027777 9.567315	.953113 1.083333 3.569507	.930847 1.134615 2.440221	.815004 2.000000 .894785	.680638 - .521489	.551433 - .381226
3		.964674 1.054607 6.113673	.905222 1.159329 2.148448	.862963 1.252363 1.419654	.654074 2.714286 .464519	.449392 - .252983	.297327 - .177957
4		.947054 1.080297 4.402655	.859408 1.223781 1.468635	.798518 1.338228 .944472	.519081 2.005244 .285264	.293343 1.085729 .149131	.165851 .303716 .102683
5		.930234 1.175427 3.387235	.814410 1.371091 1.079322	.736866 1.497075 .678892	.408128 1.222332 .193204	.192584 .311014 .098272	.099170 .117850 .066750
10		.847126 1.383247 1.421428	.612085 1.404842 .383615	.477232 1.063786 .225177	.119835 .126820 .054670	.036926 .037409 .026116	.020069 .020503 .017229
15		.765494 1.461541 .816411	.447368 .937601 .198380	.296160 .454556 .112310	.042956 .041759 .025415	.014402 .014640 .011856	.008481 .008586 .007741
25		.612085 1.404842 .383615	.224436 .289957 .082168	.109712 .109277 .044814	.012180 .012387 .009499	.004804 .004872 .004343	.002953 .002976 .002811
30		.542784 1.256846 .288406	.156668 .171236 .059292	.068895 .066063 .031987	.008078 .008255 .006660	.003283 .003326 .003029	.002035 .002050 .001957
35		.477232 1.063786 .225177	.109712 .109277 .044814	.045276 .043904 .023978	.005761 .005898 .004927	.002385 .002415 .002232	.001487 .001497 .001449

Examining Table 2, the Beta approximation is for the most cases better than approximation (2.1). Given a fixed N, the Beta approximation improves over approximation (2.1) as T and n increase. If in addition, the "a" value is examined, we find that the Beta approximation is at its weakest when 1 < a < 2, while approximation (2.1) is better for this range.

The results for the second inverse moment are given in Table 3, for N=200. Again, the Beta approximation is globally better than approximation (2.2). If in addition, the "a" value is examined, we find that the Beta approximation is at its weakest when 2 < a < 3, while approximation (2.2) is better for this range.

RESUME

Les moments de la distribution hypergéométrique positive sont souvent calculés approximativement en prenant l'inverse des moments positifs de cette distribution. Dans cet article, nous avons dévelopé une approximation en évaluant une distribution approximative de la distribution hypergéométrique.

ACKNOWLEDGMENT

The author wishes to thank the referee for helpful suggestions.

REFERENCES

[1] Hidiroglou, M.A. (1976), "An Estimator for Simple Random Sampling When Large Values are Present in the Population", Unpublished B.S.M.D. Technical Report.

[2] Johnson and Kotz (1969), Discrete Distribution, Houghton Miffin Company.

[3] Mendenhall, W. and Lehman, E.H. (1960), "An Approximation to the Negative Moments of the Positive Binomial Useful in Life Testing, Technometrics, 2 233-239.

[4] Rao, J.N.K. (1973), "On Double Sampling For Stratification And Analytical Surveys", Biometrika, 60 1 125-133.

[5] Srinath, K.P. (1976), "A Note On Outliers In Samples", Unpublished B.S.M.D. Technical Report.

[6] Sukhatme, P.V. and Sukhatme, B.V. (1970), "Sampling Theory Of Surveys With Applications", Iowa State University Press.

THE METHODOLOGY OF THE 1971 REVERSE RECORD CHECK

J.-F. Gosselin
Census Survey Methods Division

The 1971 Reverse Record Check is one of the most important studies that were carried out as part of the 1971 Census Evaluation Programme. Its main purpose was to investigate the incidence of under-enumeration in the 1971 Census. To do this, a frame containing all persons who should be enumerated in the 1971 Census was built up from the 1966 Census returns, plus birth and immigrant registrations. A random sample was selected from the frame and each selected person was traced to his current Census address. Current Census returns were then checked to see whether or not the selected person was enumerated. Sample figures were weighted up to the population level to obtain estimates of undercoverage. This paper gives a general description of the methodology of this study, and indicates some of the resulting improvements incorporated for 1976.

1. INTRODUCTION

A full national Census is carried out to provide accurate benchmark figures on which, for example, to base planning decisions and projections, and to provide accurate population statistics for small areas. For these purposes accurate figures are required. But, it has always been acknowledged by statisticians, and is now being more generally recognized by users, that all figures obtained from a survey (sample or Census) are subject to error. In the case of a national Census, the results of which are to be used by many people for many purposes, it is essential to have some measures of the reliability of published figures so that effects of errors can be taken into account.

The Reverse Record Check is designed to measure errors of coverage in the Census of Population and Housing. In most surveys one has a frame available prior to the commencement of the survey. However, in a population Census, the construction of a frame is an integral part

of the survey itself, and the size of the frame is one of the main Census results. Thus an important source of error in the Census arises from failure to include all units (households or persons) in the frame to be enumerated. This error is described as undercoverage.

The main impact of undercoverage on the reliability of Census results is to produce a downward bias in population totals due to omitting units from counts and to introduce a bias into estimates of means or proportions to the extent that persons or households missed in the Census have different characteristics to those enumerated.

The complementary error of overcoverage which would produce an upward bias in population totals is also of importance although probably smaller in size than undercoverage. The Reverse Record Check measures only undercoverage.

Although the measurement of undercoverage is an essential part of the measurement of total error of Census, the Reverse Record Check has a second function. It furnishes data on the types of persons or households that tend to be missed in the Census. This is important in the design of enumeration procedures for future Censuses when extra efforts can be made to cover these previously missed types. Thus as well as measuring the amount of undercoverage in the current Census, the Reverse Record Check also provides information that could lead to improvement in the coverage of future Censuses.

The objectives of the 1971 Reverse Record Check were:

(1) To investigate the incidence of under-enumeration of persons and households in the 1971 Population Census;

(2) To collect and analyze characteristics of persons and households missed in the Census with a view to discovering possible reasons for their being missed.

The sample design and survey methodology of the Reverse Record Check were aimed at these two objectives. As by-products, information was obtained on migration, and on emigration and mortality rates.

Details on the methodology and the results of the study may be found in Brackstcne and Gosselin [1] - [3].

2. CONSTRUCTION OF THE FRAMES

The target population contains all persons resident in Canada on Census Day, 'June 1, 1971'. Hypothetically, one could draw up a list of this population and select a sample from it. No such list is in existence, so a list must be built up from available sources.

The starting point was the 1966 Census returns (Forms 2), i.e., a list of all persons who were enumerated at their usual residence in the 1966 Census. A high proportion of the 1971 Census population will be contained in this group. For sampling purposes, this group was split into five frames according to Enumeration Area (EA) type, i.e., Metropolitan (frame 1), Urban (2), Rural (3), Special Areas (4) such as institutions and Military Bases, and Indian Reserves (5).

Persons in the 1971 Census popualtion who were not enumerated at their usual residence in the 1966 Census can be sub-divided into three groups:

(a) those who were not resident in Canada on June 1, 1966;

(b) those enumerated only at a place other than their usual residence in the 1966 Census;

(c) those who should have been enumerated in the 1966 Census but were missed.

Persons in Group (a) fall into three categories:

(a.1) Births between June 1, 1966 and May 31, 1971;

(a.2) Immigrants between June 1, 1966 and May 31, 1971;

(a.3) Persons other than registered immigrants who entered Canada between June 1, 1966 and May 31, 1971 to take up permanent residence.

Lists for the first two categories are available from birth and immigrant registrations and these lists constitute two more frames (6 and 7). Any births or immigrants not registered or registered after the sample is selected, will have been omitted from these frames. No list is available for the third category.

For Group (b), it was possible to obtain a probability sample of a list of persons in the group, although the list itself was not available. This will be explained when describing sample selection. This group is known as the Forms 3 frame (frame 8).

The 1966 RRC was a project similar to this one in which the 1966 Census population was built up from a set of frames and then sampled. Although there is no list of all persons missed in the 1966 Census, those persons in the 1966 RRC sample who were found to have been missed do constitute a random sample of all persons missed in the 1966 Census. Thus, there exists a random sample from category (c), even though no list exists for the whole category. This category is known as the Missed Persons Frame (Frame 9).

To summarize, the following nine frames were covered by the 1971 RRC.

(1) Persons in metropolitan areas enumerated at their usual place of residence in the 1966 Census.

(2) Persons in urban areas enumerated at their usual place of residence in the 1966 Census.

(3) Persons in rural areas enumerated at their usual place of residence in the 1966 Census.

(4) Persons in special areas enumerated at their usual place of residence in 1966.

(5) Persons in Indian reserves enumerated at their usual place of residence in the 1966 Census.

(6) Registered births between June 1, 1966 and May 31, 1971.

(7) Registered immigrants between June 1, 1966 and May 31, 1971.

(8) Persons enumerated only on a Form 3 (Temporary Residents) in the 1966 Census.

(9) Persons missed in the 1966 Census.

Frames 1-5 are collectively referred to as the Census frame.

Groups that are not covered by the 1971 RRC include

- Births and immigrants not registered, or registered too late to be selected in the 1971 RRC sample.
- Persons other than legal immigrants entering Canada between June 1, 1966 and May 31, 1971, to take up permanent residence.

3. SAMPLE DESIGN

3.1 General

Independent samples were selected from within each frame. The sample design varied from frame to frame depending largely on the format of the list available for each frame.

The sample design for each frame is described in the following sections.

3.2 Sampling the Census Frames

As described earlier, persons enumerated at their usual residence in the 1966 Census were divided into five frames according to the type of EA, (Metropolitan, Urban, Rural, Special Area or Indian Reserve) in which they were enumerated. Each of these five frames was stratified by province (except that, for Special Areas and for Indian Reserves, the four Atlantic provinces were grouped together into one stratum).

Within each stratum a two-stage sample of individuals was selected with replication at the first stage. Within this framework the sample was designed to satisfy the following conditions:

(a) the overall selection probability for each person was to be about 1 in 2,000 in each replicate, except that for persons aged 15-19 in 1966 this probability was to be about 1 in 1,000;

(b) approximately 15 persons were to be selected from within each primary sampling unit (PSU).

The PSU's were EA's with the exception that, if an EA had a population less than 28, it was combined with a neighbouring EA so that each PSU had a population of at least 28 persons. Within each stratum two independent systematic samples of PSU's were selected with probability proportional to 1966 population.

Under this scheme it was quite possible (and this in fact happened in certain strata) for the same PSU to be selected in both replicates. When this occurred, two independent second stage samples were selected from the PSU.

From within each selected PSU, a systematic sample of individuals was selected from a population list ordered by age within sex. Persons aged 15-19 were given twice the chance of selection in this sample by assigning them two consecutive 'listing numbers'. Each person not aged 15-19 was assigned only one 'listing number'.

3.3 Sampling the Birth Frame

All intercensal births were stratified by calendar year of arrival and province of birth. For the years 1966-70, two independent 1 in 2,000 systematic samples were selected from within each stratum. For the year 1971, two independent 1 in 1,000 systematic samples were selected within each stratum. Any selected birth born after May 31, 1971 was not included in the sample.

3.4 Sampling the Immigrant Frame

The population of immigrants arriving in Canada between June 1, 1966 and May 31, 1971 was stratified by calendar year of arrival. Within each stratum two independent 1 in 1,000 systematic samples were selected. Any selected immigrants that arrived before June 1, 1966 or after May 31, 1971 were dropped from the sample. The immigrant frame lists contained

some individuals who were not members of the population to be sampled for the Reverse Record Check. Persons who were refused entry into Canada formed the largest group. These persons were not eliminated from the list before selecting but were dropped from the sample if selected.

3.5 Sampling the Forms 3 Frame

Persons enumerated away from their usual residence on June 1, 1966 were enumerated on an individual Form 3 in the 1966 Census. These Forms 3 were sorted by province of usual residence. From within each province (except P.E.I.) a 1 in 100 systematic sample of Forms 3 was selected (in P.E.I. the ratio was 1 in 50). Each selected Form 3 was matched with the Form 2 at the usual place of residence to see whether the person was also enumerated there.

The above operation formed part of the processing of the 1966 Census. For the 1971 Reverse Record Check, those persons in the sample who were found to have been enumerated only on a Form 3, (i.e., were not found on a Form 2 at their usual residence) were regarded as a sample of all persons enumerated only on a Form 3 in 1966. Within each province this sample was divided at random into two replicates. Thus, the replicates in this frame were not quite independent.

3.6 Sampling the Missed 1966 Frame

Those persons in the 1966 RRC sample found to have been missed in the 1966 Census constitute a random sample of all persons missed in the 1966 Census. These persons were taken as the sample from the Missed Persons frame for the purposes of the 1971 RRC. This sample was divided at random into two replicates.

The probability of selection of a missed person in one replicate was equal to one half of his probability of selection in the 1966 RRC. However, in the analysis of the 1966 RRC, the basic probability of selection was adjusted for non-response and known population totals. This adjusted probability was used in calculating the probability of selection in the 1971 RRC.

3.7 Summary

The sample design was stratified with two independent (or almost independent in frames 8 and 9) replicates being selected within each stratum. For most persons the probability of selection in one replicate was 1 in 2,000. However, for certain groups of persons expected to have high rates of under-enumeration (immigrants, persons aged 20-24 in 1971, young babies) this rate was increased to 1 in 1,000. For frames 8 and 9 the sample was pre-determined as described above. Table 1 gives details of the sample sizes in each frame.

TABLE 1

Frame	Sample Size (persons)
1. Census, Metropolitan	10,119
2. Census, Urban	5,574
3. Census, Rural	5,250
4. Census, Special Areas	379
5. Census, Indian Reserves	169
6. Births	2,147
7. Immigrants	1,792
8. Forms 3	1,243
9. Missed 66	832
TOTAL	27,505

4. TRACING, SEARCHING AND FOLLOW-UP OF SELECTED PERSONS

The purpose of tracing, searching and follow-up was to classify each Selected Person (SP) into one of the final categories:

1 - Enumerated in the 1971 Census
2 - Missed in the 1971 Census
3 - Died before the 1971 Census
4 - Emigrated before the 1971 Census

This was by far the most complex and time consuming operation associated with this study.

Since addresses obtained at the time of selection were generally out-of-date, a Tracing operation was first undertaken to establish the address of each SP on June 1, 1971.

Once a selected person was traced, a Search of Census documents was carried out to determine whether or not he or she was enumerated at that address. For cases not found enumerated, follow-up was undertaken to ascertain the 1971 Census address and to collect information on persons missed in the Census.

For each case classified as missed, it was also established from the searching and follow-up whether or not the SP's household was completely missed in the Census.

Each of these phases will now be described.

4.1 Tracing

The tracing system used in 1971 consisted of a series of individual traces that were carried out sequentially in a pre-determined order (i.e., cases not traced at a given stage underwent the next stage of tracing). The actual tracing methods used varied from frame to frame but included

i) A match in Regional Office (RO) carried out as part of RO Census processing for the Census Frame sample to determine whether or not each SP was enumerated in 1971 at his/her 1966 Census address (Regional Office Match).

ii) Telephone traces from the Regional Offices

iii) Registered mail-out from Head Office

iv) Searches of administrative records such as Unemployment Insurance Commission records, Health and Welfare Family Allowance and Old Age Security records.

v) Intensive telephone and field tracing in the regions (RO Trace).

The overall tracing procedure is outlined in Diagram 1.

Cases found in the RO Match (about 42%) required no further search and were immediately classified as enumerated. In total, over 96% of the original sample was finally traced.

4.2 Searching

Three types of search were carried out:

i) A search of regular Census documents for each SP traced to a potential 1971 Census address to determine whether or not he or she was enumerated at that address.

ii) A search of special Census records for persons reported as serving overseas with the Department of National Defence or External Affairs.

iii) A search of the death register for persons reported as having died prior to June 1, 1971.

Cases not found in the corresponding records were sent to follow-up.

No search was undertaken for persons reported as having emigrated prior to Census since no emigration records exist. Such cases were automatically classified as 'Emigrated'.

4.3 Follow-up

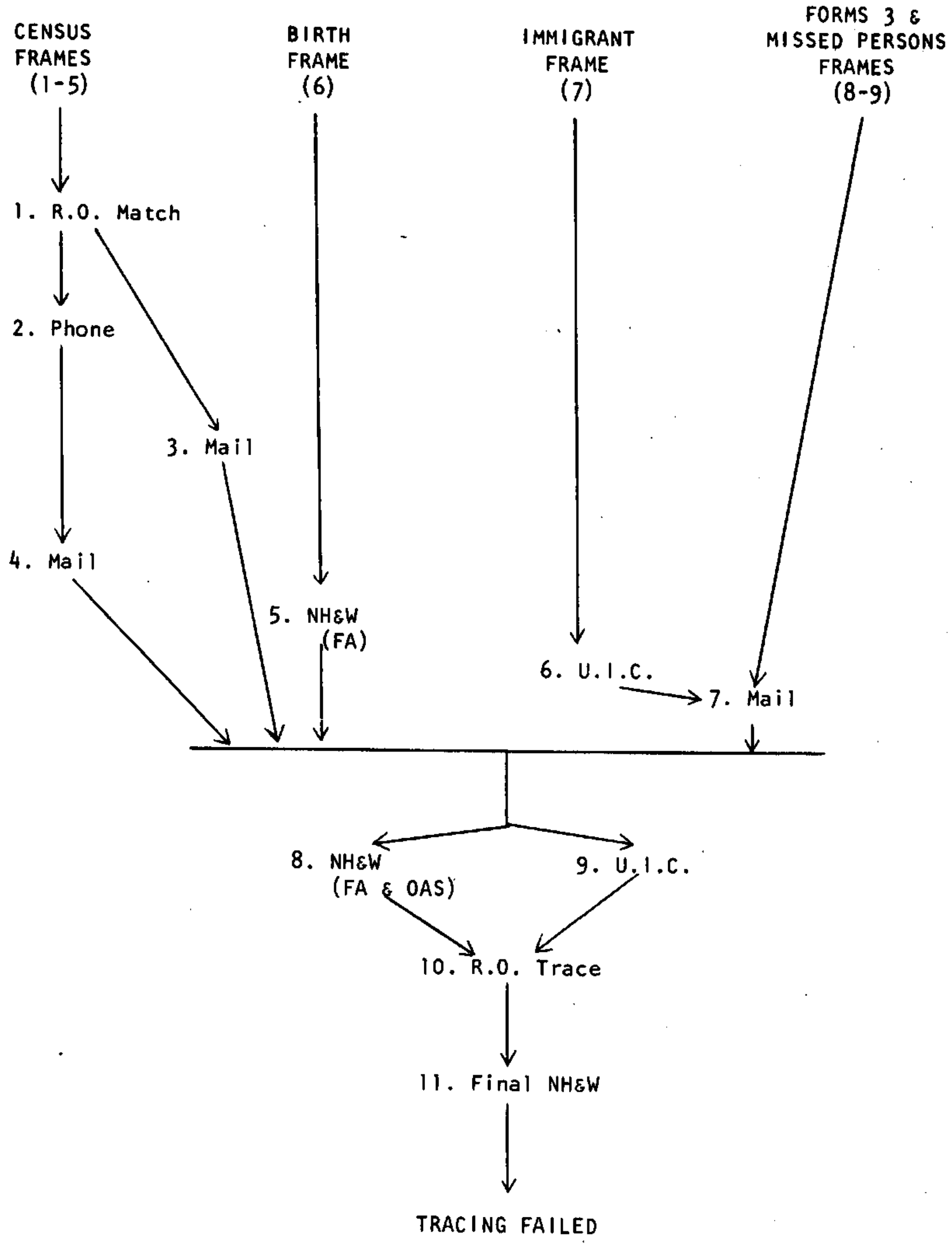
The follow-up was carried out in two phases.

The purpose of Phase 1 was to ascertain the Census address of each SP traced but not found in Census documents and to obtain other addresses where the person might have been enumerated. This began with the mailing out of a short questionnaire. When this was unsuccessful, cases were sent to RO's for a telephone and field follow-up. Further searches were then carried out in Census returns in Head Office and cases were then assigned to a final status category.

The object of the second phase of follow-up was to collect characteristics of persons missed in the 1971 Census. This was carried out entirely by RO staff.

There were some cases where both phases of follow-up were carried out at the same time, mainly for cost reasons.

DIAGRAM 1: 1971 REVERSE RECORD CHECK - TRACING PROCEDURES



KEY

RO:	Regional Office
NH&W:	National Health & Welfare
FA:	Family Allowance
OAS:	Old Age Security
UIC:	Unemployment Insurance Commission

5. ESTIMATION, ANALYSIS, AND RESULTS

Once each SP had been classified with one of the final categories (Enumerated, Missed, Died, Emigrated or Tracing Failed), estimates of undercoverage could be obtained. The purpose of this section is to give an outline of the estimation and analytical methods used as well as a brief overview of the results obtained.

5.1 Estimation Method

The estimation procedure can be subdivided into two parts:

a) the weighting of sample data

b) the production of estimated undercoverage rates and their standard errors.

Three weight adjustments were carried out:

(a.1) A non-response adjustment which consisted of a re-distribution of the weight of cases not traced, to cases traced within certain sub-groups of the sample.

(a.2) Adjustment of weights to ensure consistency with known population totals.

(a.3) A final weight adjustment to take into account the random additions procedures that were carried out as part of Census processing in 1971 for temporary residents enumerated only on a Form 3.

By summing the final adjusted weights, estimates of undercoverage were derived. Standard error estimates were obtained using the built-in replicates.

Estimates of household under-enumeration were also derived but this required a further weight adjustment using the household size at the time of the Census.

5.2 Method of Analysis

Population and household undercoverage rates were obtained for the population as a whole, and for subgroups based on Census variables such as regions, age-sex, marital status, tenure, type of dwelling, etc. These provided basic descriptive measures of the magnitude of coverage errors.

This was supplemented by a detailed analysis of the data collected for missed persons. This involved:

i) applying statistical tests to determine whether or not persons missed have different characteristics than those of the population as a whole,

ii) attempting to identify which variables appear to explain most of the variation in undercoverage,

iii) an analysis of the relationship between population and household undercoverage,

iv) a case by case study where the records on each missed person were examined by experienced staff to identify specific reasons why they were missed.

5.3 Summary of Results

The following points very briefly summarize the results of the 1971 RRC.

i) The overall population and household undercoverage rates were estimated as 1.93% and 1.46% respectively.

ii) With respect to most Census variables, the population of missed persons and missed households appear to be significantly different from the enumerated population. Undercoverage is particularly high for persons not related to the head of household, males aged 20-24, recent immigrants, the unemployed, and for smaller households living in rented dwellings.

iii) Household undercoverage accounts for more than 50% of the total population under-enumeration.

Generally, undercoverage appears to be high in those subgroups of the population that tend to be more mobile.

6. METHODOLOGICAL IMPROVEMENTS FOR 1976

The Reverse Record Check method is again being used to measure undercoverage in the 1976 Census. Essentially, the same methodology is being applied.

However, apart from operational improvements, a certain number of methodological changes have been made:

- A cost-variance analysis was carried out using the 1971 data which was used to better allocate the sample amongst frames.
- The overall sample size was increased to about 33,000 and the sample was allocated to provinces so that reliable provincial estimates could be obtained.
- Since the searching operation in 1971 was carried out using regular Forms 2, that is amongst persons enumerated at their usual place of residence, the sample of persons classified in the final category 'Missed' is representative of all persons not enumerated at their usual place of residence. Since the 1971 Census incorporated a random addition procedure for persons enumerated only on a Form 3 (i.e., as temporary resident away from their usual place of residence), a weight adjustment had to be carried out for the missed sample to 'remove' the effect of these random additions. However, for the purpose of the 1976 RRC, the missed sample (with unadjusted weights) is representative of both the Forms 3 frame and the Missed frame. It will therefore be used as such, thus eliminating the need to include the 71 Forms 3 sample in the 1976 RRC. This is a major methodological improvement since these cases were extremely difficult and costly to trace.
- Some modifications have been made to the tracing system, including the use of the 1974 Taxation records and an extension of the telephone trace for the Census frame to replace the registered mail-out.

As a supplement to the Reverse Record Check, two additional studies are also being carried out to investigate coverage errors in the 1976 Census.

These are the Dwelling Coverage Check and the Vacancy Check which are designed to measure the undercoverage of dwellings and the misclassification of occupied dwellings as vacant in the 1976 Census.

These three studies will provide most of the data on coverage errors in the 1976 Census.

RESUME

La Contre-vérification des dossiers est l'une des plus importantes études de la qualité des données du recensement, au sein du Programme d'évaluation de 1971. Elle vise essentiellement à mesurer le taux de sous-dénombrement de la population lors du recensement de 1971. Pour ce faire, une base de sondage contenant toutes les personnes devant être enumérées au recensement, fût construite à partir des dossiers du recensement de 1966, du registre des naissances et l'immigration, d'où un échantillon a été prélevé au hasard. On a alors procédé à une opération de dépistage dans le but de déterminer l'adresse de chaque personne choisie lors du recensement de 1971. Ceci a par la suite permit d'effectuer une recherche des dossiers de 1971 afin d'établir si chaque personne choisie avait été recensée en 1971. Des estimations du sous-dénombrement furent alors obtenues en pondérant les données de l'échantillon. Cet article a pour but de présenter une description générale de la méthodologie de la Contre-vérification des dossiers de 1971, ainsi que quelques améliorations apportées en 1976.

REFERENCES

[1] Brackstone, G.J., Gosselin, J.-F., 1971 Reverse Record Check, Census Evaluation Programme, Internal Report, Statistics Canada, September 1973.

[2] Brackstone, G.J., Gosselin, J.-F., Results Memorandum, 1971 Reverse Record Check, CDN-E-23(Part 1), Statistics Canada, October 1974.

[3] Brackstone, G.J., Gosselin, J.-F., Results Memorandum, 1971 Reverse Record Check, CDN-E-23 (Part 2),Statistics Canada, January 1975.

THE ESTIMATION OF TOTAL VARIANCE IN THE 1976 CENSUS

G.J. Brackstone and C.J. Hill
Census Survey Methods Division

Published reports for the 1976 Census will include estimates of Total Variance as indicators of the reliability of the figures in these reports. In order to obtain these estimates of Total Variance, an Interpenetrating Design Experiment was incorporated into the collection methods for a sample of enumeration areas. In this paper we derive the formula for Total Variance in terms of variances due to sampling, correlated response and simple response. We then show how the Total Variance, and its components, can be estimated from the design and we give the estimators that will be used for the 1976 Census. The estimates of sampling and correlated response variance are unbiased but the simple response variance estimate is not.

1. INTRODUCTION

The Total Variance study is an integral part of an evaluation program designed to measure the quality of data produced by the 1976 Census of Population and Housing. As the name suggests, its objective is to measure the overall variance of Census estimates including both sampling and non-sampling components. Other studies in the program are designed to measure the bias in Census estimates (particularly due to undercoverage) and to investigate individual sources of error.

The 1976 Census utilizes sampling in that every third private household receives a long form that contains not only the basic (100%) Census questions but also additional (sample) questions on education, labour force status, and migration. In remote areas (accounting for about 2% of the population) and in collective dwellings (hotels, institutions, etc.) all persons are enumerated on a long form. Thus, while all Census estimates are subject to non-sampling variance due to response errors and processing errors, Census estimates for sample characteristics are also subject to sampling variance.

Earlier studies of non-sampling variance in Censuses ([1][2]) have indicated that the correlated component of response variance caused by an enumerator introducing a positive correlation between the errors within his/her assignment was an important, if not over-riding, contributor to the total variance of Census estimates in a canvasser or direct enumeration Census. For this reason the Canadian Census of 1971, unlike its predecessors, utilized self-enumeration in an attempt to reduce the correlated component of response variance. Under self-enumeration, the influence of the enumerator is restricted to those questionnaires that were not returned, or which were returned incomplete, and which therefore required enumerator follow-up. A study of response variance in the 1971 Census [4] indicated that the correlated component of response variance was considerably smaller than in the canvasser Census of 1961 although still a significant contributor to non-sampling variance for some characteristics.

The primary purpose of the 1976 Total Variance study is to produce measures of reliability that can be applied by users of Census data to any published Census figure. The measures that will be produced and published are total standard errors (i.e. the square root of total variance) that take account of the effects of all sources of variance particularly sampling variance and correlated response variance.

In Section 2, we describe briefly the methodology of the Total Variance study while in the remaining sections the estimators of total variance are derived.

2. METHODOLOGY OF THE TOTAL VARIANCE STUDY

To enable an estimate of the correlated response variance to be obtained, the Total Variance study makes use of interpretation of enumerator assignments during the field collection stage of the Census. Using the terminology of Bailar and Dalenius [3], the Total Variance study is an example of interpenetration in both the sample and trial (i.e. enumerator) dimensions.

For the purpose of Census taking, the country is divided up into about 1600 Commissioner Districts (CD's) with each CD containing an average of 20 Enumeration Areas (EA's) with each EA being enumerated by one Census Representative (CR). For the Total Variance study, a stratified random sample of CD's was selected with probability proportional to the number of EA's in the CD. Within each selected CD, the EA's were arranged in pairs so that each pair contained contiguous EA's of the same or similar type (in terms of pay-rates). A simple random sample of two pairs of EA's were then selected from all the pairs in each CD. This resulted in a self-weighting sample of pairs of EA's across Canada.

In selected EA's, questionnaires were dropped off at households in the normal way. After drop-off was complete, the households in each EA were randomly split into two equal halves using tables of random numbers. One random half from one EA in a pair was combined with a random half from the other EA to form a new assignment which would contain approximately the same number of households as each of the original EA's but which would cover twice the geographic area. The other two halves formed a second assignment. These two new assignments were allocated at random to the two original CR's. All subsequent collection operations (i.e. checking of returned questionnaires and follow-up were conducted within these new assignments). Once the collection stage was complete and each of the assignments had separately passed the quality control check, the two assignments were re-sorted back into their original EA's and processed normally through all remaining Census operations. The records of which households were enumerated by which CR were retained so that when the final Census data were available on a data base, the two assignments could be reconstructed for the purpose of applying the estimation formulae derived in the next sections.

3. NOTATION

Assume there are P EA's in Canada and that these are paired into M = P/2 contiguous pairs. A self-weighting sample of m pairs of EA's is selected.

Let the subscript k denote the EA $(k=1,2,\ldots,2M)$.
Let the subscript i denote the half-EA enumerated by one enumerator $(i=1,2)$.
Let the subscript h denote the household.

Let U_{ki} denote the set of households enumerated in the ith half of the kth EA. Let N_{ki} denote the number of households in U_{ki}. Let S_{ki} denote the set of sample households in the ith half of the kth EA. Let n_{ki} denote the number of households in S_{ki}.

The expectation operator E can be divided into four stages $E = E_1 E_2 E_3 E_4$ where:

E_4 indicates expectations over hypothetical replications of the response process (including the assignment of enumerators) for a given household,

E_3 indicates expectations over the random splitting of the EA's given the sets of households S_{ki}, U_{ki},

E_2 indicates expectations over the sampling process that selects S_{ki} from U_{ki} within a given EA,

E_1 indicates expectations over the random selection of the m pairs of EA's from the M pairs in the population.

Similarly V_1, V_2, V_3, V_4 and C_1, C_2, C_3, C_4 indicate the corresponding variance and covariance operators.

Let x_{kh} denote the observed value of a particular Census characteristic for the hth household in the kth EA. If the characteristic is a 100% characteristic x_{kh} is known for all $h \varepsilon U_k$ where $U_k = U_{k1} \cup U_{k2}$, while for sample characteristic x_{kh} is known only for $h \varepsilon S_{k1} \cup S_{k2}$. In the Census application, x_{kh} will generally be either a 0-1 variable indicating absence or presence of a specific household characteristic, or an integer valued variable indicating the number of persons in the household with a specific personal characteristic.

Let $X_{kh} = E_4(x_{kh})$, $\bar{X}_k = \frac{1}{N_k} \sum_{h \varepsilon Uk} X_{kh}$, and $\bar{X}_{Sk} = \frac{1}{n_k} \sum_{h \varepsilon S_k} X_{kh}$.

Let σ_k^2 = Average value of $E_4\ (x_{kh} - X_{kh})^2$ for households in the kth EA and

$$\rho_k\ \sigma_k^2 = \text{Average value of } E_4\ (x_{kh} - X_{kh})(x_{kh'} - X_{kh'})$$

for pairs of households h, h' in the kth EA that were enumerated by the same enumerator.

4. THE VARIANCE OF THE CENSUS ESTIMATOR FOR A SAMPLE VARIABLE

The estimator of the population total for a Census sample characteristic, x, can be written as

$$\hat{X} = \sum_{k=1}^{P} \frac{N_k}{n_k} \sum_{h\varepsilon S_k} x_{kh} \qquad \text{where } N_k = N_{k1} + N_{k2}, \tag{4.1}$$

$$n_k = n_{k1} + n_{k2},$$

$$\text{and} \quad S_k = S_{k1} \cup S_{k2}$$

The variance of $\hat{X}$ is given by

$$V(\hat{X}) = \sum_{k=1}^{P} (\frac{N_k}{n_k})^2 \ V(\sum_{h\varepsilon S_k} x_{kh})$$

$$= \sum_{k=1}^{P} (\frac{N_k}{n_k})^2 \{E_2\ V_4\ (\sum_{h\varepsilon S_k} x_{kh}) + V_2\ E_4\ (\sum_{h\varepsilon S_k} x_{kh})\}$$

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

where

$$S_{xk}^2 = \frac{1}{N_k-1} \sum_{h\varepsilon U_k} (X_{kh} - \bar{X}_k)^2$$

$$\therefore V(\hat{X}) = \sum_{k=1}^{P} N_k^2 \ [\frac{\sigma_k^2}{n_k} (1 + (n_k - 1)\ \ \rho_k) + (1 - \frac{n_k}{N_k})\ \ \frac{S_{xk}^2}{n_k}] \tag{4.2}$$

Note that the above derivation is based on the assumption not only that sampling is carried out independently within different EA's, but also that response errors within different EA's are uncorrelated. Thus, it does not make any allowance for the fact that certain selected pairs of EA's are interpenetrated prior to enumeration.

In the next two sections we consider two estimators, C_k and D_k, defined for the kth selected EA, which will form the basis for estimators of the total variance.

5. THE ESTIMATOR C_k

Let $x_{k(i)} = \sum_{h\varepsilon S_{ki}} x_{kh}$, and let $\bar{x}_{k(i)} = x_{k(i)}/n_{ki}$.

If $X_{k(i)} = \sum_{h\varepsilon S_{ki}} X_{kh}$, $E_4 (\bar{x}_{k(i)}) = X_{k(i)}/n_{ki}$.

Now consider $C_k = \frac{1}{2} [\bar{x}_{k(1)} - \bar{x}_{k(2)}]^2$ (5.1)

$$E(C_k) = \frac{1}{2} E[(\bar{x}_{k(1)} - E_4 (\bar{x}_{k(1)}) - (\bar{x}_{k(2)} - E_4 (\bar{x}_{k(2)}))$$

$$+ (E_4(\bar{x}_{k(1)}) - E_4 (\bar{x}_{k(2)}))]^2$$

$$= \frac{1}{2} [E_2 E_3 V_4 (\bar{x}_{k(1)}) + E_2 E_3 V_4 (\bar{x}_{k(2)}) - 2 E_2 E_3 C_4(\bar{x}_{k(1)},\bar{x}_{k(2)})$$

$$+ E_2 E_3 (\frac{X_{k(1)}}{n_{k1}} - \frac{X_{k(2)}}{n_{k2}})^2] \qquad (5.2)$$

But

$$E_2\, E_3\, V_4(\bar{x}_{k(i)}) \;=\; \frac{1}{n_{ki}}\; \sigma_k^2\, (1 + (n_{ki}-1)\, \rho_k),$$

$E_2E_3C_4(\bar{x}_{k(1)},\bar{x}_{k(2)}) \;=\;$ 0 on the assumption that response errors in different halves of an EA are uncorrelated,

and

$$E_2\, E_3\, (\frac{X_{k(1)}}{n_{k1}} - \frac{X_{k(2)}}{n_{k2}})^2 = E_2\, E_3\, [(\frac{X_{k(1)}}{n_{k1}} - E_3\, \frac{X_{k(1)}}{n_{k1}}) - (\frac{X_{k(1)}}{n_{k2}} - E_3\, \frac{X_{k(2)}}{n_{k2}})]^2$$

(note that $E_3\, (\frac{X_{k(1)}}{n_{k1}}) = \frac{1}{n_k} \sum_{h\varepsilon S_k} X_{kh} = E_3\, (\frac{X_{k(2)}}{n_{k2}}))$

$$= E_2\, V_3\, (\frac{X_{k(1)}}{n_{k1}}) + E_2\, V_3\, (\frac{X_{k(2)}}{n_{k2}}) - 2\, E_2\, C_3\, (\frac{X_{k(1)}}{n_{k1}}\, ,\, \frac{X_{k(2)}}{n_{k2}}).$$

But $V_3\, (\frac{X_{k(i)}}{n_{ki}}) = (1 - \frac{n_{ki}}{n_k})\, \frac{s_{xk}^2}{n_{ki}}$ (where $s_{xk}^2 = \frac{1}{n_k - 1} \sum_{h\varepsilon S_k} (X_{kh} - \bar{X}_{sk})^2$).

Also $V_3\, (\frac{1}{n_k} \sum_{h\varepsilon S_k} X_{kh}) = 0 = V_3\, (\frac{X_{k(1)} + X_{k(2)}}{n_k})$,

$C_3(\frac{X_{k(1)}}{n_k},\frac{X_{k(2)}}{n_k}) = -\frac{1}{2}\, V_3(\frac{X_{k(1)}}{n_k}) - \frac{1}{2}\, V_3(\frac{X_{k(2)}}{n_k})$, and

$$E_2E_3(\frac{X_{k(1)}}{n_{k1}} - \frac{X_{k(2)}}{n_{k2}})^2 = (1 - \frac{n_{k1}}{n_k})\, \frac{S_{xk}^2}{n_{k1}} + (1 - \frac{n_{k2}}{n_k})\, \frac{S_{xk}^2}{n_{k2}} + \frac{n_k^2}{n_{k1}n_{k2}}\, \{((\frac{n_{k1}}{n_k})^2 (1 - \frac{n_{k1}}{n_k})\, \frac{S_{xk}^2}{n_{k1}}$$

$$+ (\frac{n_{k2}}{n_k})^2\, (1 - \frac{n_{k2}}{n_k})\, \frac{S_{xk}^2}{n_{k2}}\, \}$$

$$= \frac{n_k\, S_{xk}^2}{n_{k1}\, n_{k2}}.$$

Thus, substituting in (5.2),

$$E[C_k] = \frac{1}{2} \sum_{i=1}^{2} \frac{\sigma_k^2}{n_{ki}} (1 + (n_{ki} - 1) \rho_k) + \frac{1}{2} \cdot \frac{n_k}{n_{k1} n_{k2}} S_{xk}^2$$

$$= \frac{n_k}{2n_{k1} n_{k2}} [\sigma_k^2 (1 + [\frac{2n_{k1} n_{k2}}{n_k} - 1] \rho_k) + S_{xk}^2]$$

$$= \frac{2\sigma_k^2}{n_k} [1 + (\frac{n_k}{2} - 1) \rho_k] + \frac{2S_{xk}^2}{n_k} \qquad (5.3)$$

when $n_{k1} = n_{k2} = \frac{1}{2} n_k$.

6. THE ESTIMATOR D_k

Consider

$$D_k = \frac{\sum_{i=1} \sum_{h\varepsilon S_{ki}} \frac{(x_{kh} - \bar{x}_{k(i)})^2}{n_{ki}}}{n_{k1} + n_{k2} - 2} \qquad (6.1)$$

$$E \sum_{h\varepsilon S_{ki}} (x_{kh} - \bar{x}_{k(i)})^2 = E \{ \sum_{h\varepsilon S_{ki}} (x_{kh} - X_{kh})^2$$

$$+ \sum_{h\varepsilon S_{ki}} (X_{kh} - \bar{X}_{k(i)})^2 + n_{ki} (\bar{X}_{k(i)} - \bar{x}_{k(i)})^2$$

$$- 2 n_{ki} (\bar{x}_{k(i)} - \bar{X}_{k(i)})^2 \qquad \text{where } \bar{X}_{k(i)} = \frac{1}{n_{ki}} \sum_{h\varepsilon S_{ki}} X_{kh}$$

$$= n_{ki} \sigma_k^2 + (n_{ki} - 1) S_{Xk}^2 - \frac{1}{n_{ki}} E[\sum_{h\varepsilon S_{ki}} (x_{kh} - X_{kh})]^2$$

$$= n_{ki} \sigma_k^2 + (n_{ki} - 1) S_{Xk}^2 - \sigma_k^2 - (n_{ki} - 1) \rho_k \sigma_k^2$$

$$= (n_{ki} - 1) [\sigma_k^2(1 - \rho_k) + S_{Xk}^2]$$

Therefore, $E(D_k) = \frac{1}{n_k - 2} \sum_{i=1}^{2} \frac{n_{ki} - 1}{n_{ki}} [\sigma_k^2 (1 - \rho_k) + S_{Xk}^2]$

$$= \frac{2\ n_{ki}\ n_{k2} - n_k}{(n_k - 2)\ n_{k1}\ n_{k2}}\ [\sigma_k^2\ (1 - \rho_k) + S_{Xk}^2]$$

$$= \frac{2}{n_k}\ [\sigma_k^2\ (1 - \rho_k) + S_{Xk}^2] \qquad (6.2)$$

when $n_{k1} = n_{k2} = n_k/2$.

7. ESTIMATING THE TOTAL VARIANCE FOR A SAMPLE VARIABLE

We next consider how we can utilize the estimator C_k and D_k to obtain an estimator of the Total Variance $V(\hat{X})$ given by (4.2).

Assuming $n_{k1} = n_{k2} = n_k/2$ we have from (5,3)

$$E(\frac{1}{2}\ C_k) = \frac{\sigma_k^2}{n_k}\ [1 + (\frac{n_k}{2} - 1)\ \rho_k] + \frac{S_{xk}^2}{n_k}\ , \qquad (7.1)$$

and from (6.2)

$$E(\frac{1}{2}\ D_k) = \frac{\sigma_k^2}{n_k}\ [1 - \rho_k] + \frac{S_{xk}^2}{n_k}\ , \qquad (7.2)$$

whereas from (4.2)

$$V(\hat{X}) = \sum_{k=1}^{P} N_k^2\ \{ \frac{\sigma_k^2}{n_k}\ [1 + (n_k - 1)\ \rho_k] + (1 - \frac{n_k}{N_k})\ \frac{S_{xk}^2}{n_k}\ \}\ . \qquad (7.3)$$

$$E[\frac{M}{m}\ \frac{1}{2} \sum_{k=1}^{2m} N_k^2\ C_k] = \sum_{k=1}^{P} N_k^2\ \{ \frac{\sigma_k^2}{n_k}\ [1 + (\frac{n_k}{2} - 1)\ \rho_k] + \frac{S_{xk}^2}{n_k}\ \}. \qquad (7.4)$$

This expectation differs from $V(\hat{X})$ in two respects:

(i) the factor $(\frac{n_k}{2} - 1)\ \rho_k$ in place of $(n_k - 1)\ \rho_k$;

(ii) it lacks the finite population correction, $1 - \frac{n_k}{N_k}$.

We can obtain a separate estimator of the term in $\rho_k\ \sigma_k^2$ by noting that

$$E[\frac{1}{2}\ \frac{M}{m} \sum_{k=1}^{2m} N_k^2\ (C_k - D_k)] = \sum_{k=1}^{P} N_k^2\ \frac{1}{2}\ \rho_k\ \sigma_k^2\ . \qquad (7.5)$$

From (7.4) and (7.5) it follows that the estimator

$$\hat{V}_1(\hat{X}) = \frac{1}{2}\,\frac{M}{m}\sum_{k=1}^{2m} N_k^2\,(C_k + C_k - D_k)$$

$$= \frac{M}{m}\sum_{k=1}^{2m} N_k^2\,(C_k - \frac{1}{2}D_k) \qquad (7.6)$$

would have a bias of $\sum_k N_k\, S_{xk}^2$ in estimating $V(\hat{X})$. $V_1(\hat{X})$ would be a suitable estimator of $V(\hat{X})$ in cases where the finite population correction was negligible.

In the case of the Census sample, the finite population correction is 2/3 and cannot be ignored. Returning to (7.4), we have

$$E[\frac{1}{2}\,\frac{M}{m}\sum_{k=1}^{2m} N_k^2\,(1 - \frac{n_k}{N_k})\,C_k] = \sum_{k=1}^{P} N_k^2\,\{\frac{\sigma_k^2}{n_k}\,(1 - \frac{n_k}{N_k})(1 + (\frac{n_k}{2} - 1)\,\rho_k)$$

$$+ (1 - \frac{n_k}{N_k})\,\frac{S_{xk}^2}{n_k}\}\ .$$

We can now adjust this estimator by the appropriate multiple of (7.5) to obtain the correct coefficient for the term $\rho_k\,\sigma_k^2$. Thus, the estimator

$$\hat{V}_2(\hat{X}) = \frac{1}{2}\,\frac{M}{m}\sum_{k=1}^{2m} N_k^2\,[(1 - \frac{n_k}{N_k})\,C_k + \frac{N_k + n_k - 2}{N_k}\,(C_k - D_k)]$$

$$= \frac{M}{m}\sum_{k=1}^{2m} N_k^2\,[\frac{N_k - 1}{N_k}\,C_k - \frac{N_k + n_k - 2}{2N_k}\,D_k] \qquad (7.7)$$

is biased only in the term involving σ_k^2. In fact

$$\text{Bias } (\hat{V}_2(\hat{X})) = -\sum_{k=1}^{P} N_k \frac{\sigma_k^2}{n_k} (n_k - 1).$$

8. ESTIMATING THE TOTAL VARIANCE FOR A 100% VARIABLE

For a 100% variable, $\hat{X} = \sum_{k=1}^{P} \sum_{h \varepsilon U_k} x_{kh}$, (8.1)

and $V(X) = \sum_{k=1}^{P} N_k \sigma_k^2 (1 + (N_k - 1) \rho_k)$. (8.2)

Redefining $x_{k(i)} = \sum_{h \varepsilon U_{ki}} x_{kh}$

and $\bar{x}_{k(i)} = x_{k(i)}/N_{ki}$ in C_k, we have

$$E(C_k) = \frac{1}{2} \sum_{i=1}^{2} \frac{\sigma_k^2}{N_{ki}} (1 + (N_{ki} - 1) \rho_k) + \frac{1}{2} \frac{N_k}{N_{k1} N_{k2}} S_{xk}^2$$

$$= \frac{2\sigma_k^2}{N_k} [1 + (\frac{N_k}{2} - 1) \rho_k] + \frac{2S_{xk}^2}{N_k} \qquad (8.3)$$

when $N_{k1} = N_{k2} = N_k/2$.

Redefining

$$D_k = \frac{\sum_{i=1}^{2} \sum_{h \varepsilon U_{ki}} (x_{kh} - \bar{x}_{k(i)})^2}{N_{k1} + N_{k2} - 2} \qquad (8.4)$$

$$E(D_k) = \frac{1}{N_k-2} \sum_{i=1}^{2} \frac{N_{ki}-1}{N_{ki}} [\sigma_k^2 (1-\rho_k) + S_{xk}^2]$$

$$= \frac{2}{N_k} [\sigma_k^2 (1-\rho_k) + S_{xk}^2] \qquad (8.5)$$

when $N_{k1} = N_{k2} = N_k/2$.

Thus $E[\frac{M}{m} \sum_{k=1}^{2m} N_k^2 (C_k - D_k)] = \sum_{k=1}^{P} N_k^2 \rho_k \sigma_k^2$.

Therefore, the estimator

$$V_3(X) = \frac{M}{m} \sum_{k=1}^{2m} N_k(N_k-1)(C_k - D_k) \qquad (8.6)$$

is biased only in the term involving σ_k^2. In fact

$$\text{Bias } (\hat{V}_3(\hat{X})) = -\sum_{k=1}^{P} N_k \sigma_k^2.$$

9. CONCLUSION

The total variance of the Census estimator for a sample variable involves a simple response variance term (SRV), a correlated response variance term (CRV), and a sampling variance term (SV). It is possible to obtain a total variance estimator that is unbiased in any two of these three components but not in all three. On the grounds that SRV is likely to be the smallest of the three components, the preferred total variance estimator is given by $\hat{V}_2(\hat{X})$ in (7.7) and is biased only in the SRV term.

For 100% variables, the total variance of the Census estimator involves a SRV term and a CRV term. Only one of these two terms can have the correct coefficient in the expected value of a total variance estimator.

On the grounds that SRV is likely to be the smaller term, the preferred total variance estimator is given by $\hat{V}_3(\hat{X})$ in (8.6) which is biased only in the SRV term.

The above derivations for sample estimators assume an estimator of the form (4.1)

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

In fact, the estimator used in practise is more complex involving the use of raking ratio estimation. It can be written in the form

$$\hat{X} = \sum_{k=1}^{P} \sum_{h \epsilon S_k} W_{kh}\, x_{kh} \tag{9.1}$$

where W_{kh} is a weight calculated using the raking-ratio procedure [5][6]. W_{kh} is itself a random variable since it depends on the values of the 100% characteristics of sample members. Some of the properties of raking-ratio estimators, in the absence of response errors have been examined in [6]. The above derivations can be extended to the estimator (9.1) by redefining σ_k^2, $\rho_k\ \sigma_k^2$ and S_{xk}^2 in terms of a variable $W_{kh}\ x_{kh}$ in place of x_{kh}, and by utilizing a more complex expression for the sampling variance terms when the expectation E_2 is taken. In effect this would treat W_{kh} as a constant when taking expectations E_3 and E_4 and would thus neglect the effect of response errors in the 100% variables used in raking-ratio estimation. These extensions will not be considered further here.

RESUME

Les publications du recensement de 1976 contiendront des estimations de la variance totale. Les estimations visent à donner une indication de la fiabilité des chiffres présentés dans ces publications. Ces estimations de la variance totale sont obtenues d'un plan d'expérience mis sur pied lors du recensement de 1976, comportant une interpénétration au niveau de la collecte des données pour un échantillon de secteurs de dénombrement. Dans cet article, la variance totale est exprimée en fonction des variances dues à l'échantillonage, aux réponses corrélées et aux réponses simples. On montre ensuite comment la variance totale, ainsi que ses composantes, peuvent être obtenues à partir de l'échantillon. On donne aussi les estimateurs qui seront utilisés pour le recensement de 1976. Contrairement à l'estimateur de la variance aux réponses simples, les estimateurs de la variance due à l'échantillonage et aux réponses corrélées sont sans biais.

REFERENCES

[1] Fellegi, I.P., "Response Variance and Its Estimation", JASA 59 (1964).

[2] U.S. Bureau of the Census, Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960, Effects of Interviewers and Crew Leaders Series ER60, No. 7, Washington, D.C. 1968.

[3] Bailar, B.A. and Dalenius, T., "Estimating the Response Variance Components of the U.S. Bureau of the Census' Survey Model", SANKYA B, Vol. 31, December 1969.

[4] Hill, C., "A Comparison of the 1961 Census and 1971 Census Correlated Response Variance Estimates", Internal Report, Statistics Canada, (Dec.76).

[5] Brackstone, G.J. and Rao, J.N.K., "Raking Ratio Estimators", Survey Methodology, Vol. 2, Number 1, June 1976.

[6] Arora, H.R. and Brackstone, G.J., "An Investigation of the Properties of Raking Ratio Estimators With Simple Random Sampling", Internal Report, Statistics Canada, (Dec.76).

1974 SURVEY OF HOUSING UNITS

H. Hofmann
Household Surveys Development Division

The 1974 Survey of Housing Units was carried out by Statistics Canada on behalf of the Central Mortgage and Housing Corporation during the autumn of 1974. Statistics Canada's responsibilities on this project included the design and implementation of all phases of the survey up to and including the production of "clean" micro data tapes. The sponsoring department was in turn responsible for the specification of objectives and data requirements and for the analysis of the resulting data.

This report, which is a modification of the summary report produced by the project team at the conclusion of the project, provides a general description of the survey and the work done by Statistics Canada on the survey.

1. OBJECTIVES OF THE SURVEY

To further its understanding of specific urban housing markets, research objectives of Central Mortgage and Housing (CMHC) included the following:

(i) a description of housing needs by levels of needs;

(ii) the identification of the reasons why households changed their consumption of housing units and, in changing, what determines their choice of a particular unit;

(iii) an indication of the process of deterioration or revitalization in units and neighbourhoods;

(iv) a description of dwelling unit characteristics, costs and carrying charges, over time.

The objective for Statistics Canada was to provide CMHC with the information required to accomplish the specified research objectives. Since the particular set of data required was not available from any existing source, or combination of sources, a survey was to be conducted at intervals using the same basic sample of dwelling units. In order to provide longitudinal information on the dwelling unit at the time of the first cycle of

the survey, the basic sample was to be selected from the 1971 Census of Canada and specific data items for selected units collected in 1971 were to be linked to the information collected in 1974.

Again subsequent to the conduct of the survey, CMHC decided to use the information collected in 1974 to produce a publication on housing statistics.

2. SAMPLE DESIGN

This section provides a general description of the sample design used for the first cycle of this survey.

2.1 General

The sample design of this survey was based upon the following four requirements specified by CMHC.

(A) Data relating to household and dwelling characteristics which is statistically reliable at the level of specific urban housing markets was required at specific points in time. This data would provide for intercycle cross-sectional analysis.

(B) Notwithstanding (A), detailed analysis of data gathered about these concepts was to be done of low income households to a greater extent than that of middle and upper income households.

(C) Specific subsets of this data, statistically reliable at the level of specific urban areas, were required for comparison over time. These subsets would provide for intercycle longitudinal analysis.

(D) At the first cycle, a secondary set of data (preferably the 1971 Census) relating to household and dwelling characteristics was required. This data would provide for intercycle longitudinal analysis.

These general requirements were synthesized jointly by Statistics Canada and CMHC into a sample design embodying the following features:

(i) A household survey would be conducted three times, the first cycle being undertaken in the fall of 1974, and collecting information on the same set of variables at each cycle on a

common set of households, with the sampling being heaviest at the low end of the household income scale.

(ii) At each cycle after the first, the sample of dwellings would be updated by a supplementary sample of dwellings constructed since the time of the previous cycle. This would ensure that the total sample at each cycle would be representative of the population of dwellings in existence at that cycle.

(iii) Requirement (D) would be met by linking the household and dwelling characteristics measured at the first cycle with household and dwelling characteristics measured by the 1971 Census for the same set of dwellings.

2.2 Population

The population of interest in this household survey was the private dwellings in existence during the reference period (the autumn of 1974) and located within designated municipalities in the 23 largest metropolitan areas.

2.3 Frames

A frame is the list of units in the population at a given point in time from which a sample can be selected. In light of the requirements specified by the sponsor, three frames were used for this survey. These were:

(1) The 1971 Census of Canada file of occupied (on June 1, 1971) private dwellings.

(2) The 1971 Census of Canada file of vacant (on June 1, 1971) private dwellings.

(3) Statistics Canada's summary records of issued building permits.

Frames 1 and 2 provide a base for this and future cycles of this survey since they are a list of all private dwellings in existence prior to June 1, 1971. Since these files were created and stored separately it was necessary for various reasons to consider them as two independent frames. These frames covered approximately 85% of the population of private dwellings in existence at the first cycle. Frame 3 allowed for

inclusion in the sample the dwellings constructed between June 1, 1971 and the reference period, the autumn of 1974.

The 1971 Census of Canada (2B) file of occupied private dwellings is essentially a computer maintained list of the private dwellings enumerated in the 1971 Census of Canada which were occupied at the time of enumeration. Detailed dwelling and household characteristics are available for individual units on the frame on a one-third sample basis and each unit is geographically identified to the census enumeration area (EA) level.

The 1971 Census of Canada file of vacant private dwellings lists all private dwellings enumerated in the 1971 Census of Canada which were vacant at the time of enumeration. The frame is essentially stored in the census visitation records (VRs). Because of the vacancy, no detailed characteristics are available for the units in the frame. Each unit can be geographically identified down to the EA level.

At the end of each month all the municipalities considered in the population for this survey (and others) submit to Statistics Canada a record of all building permits issued within that municipality during that month. This provides a source for obtaining the location of all dwellings constructed since the 1971 Census of Canada. For each permit, the type and location of the intended structure as well as the number of dwelling units that it will contain is reported. All units on the frame can be geographically identified to the municipality level.

2.4 Reliability

As mentioned in 2.1, statistical estimates produced from each cycle of this survey (cross-sectional estimates) were to be meaningful (reliable) only at the level of each of the 23 survey areas. This required that all

important estimates produced for each of the survey regions had a variability which was not greater than a certain pre-specified value; and that estimates for any lower (or higher) geographic level were not required.

The measure most suitable for specifying the reliability criteria successfully is the coefficient of variation (α) of an estimate ($\hat{X}$) which expresses the standard error of the estimate as a fraction or percentage of the true value.

2.5 Sampling

2.5.1 Stratification

To provide for a more efficient design with regard to the variables likely to be highly correlated with the concepts to be measured, the units on frame 1 were stratified as follows into 40 strata by (i) tenure, (ii) total income of head of household (and spouse) and (iii) age of head.

Tenure	Income of Head (and spouse) (in dollars)	Age of Head (in years)
owned (or being bought)	Under 5,000	Under 25
rented	5,000 - 6,999	25 - 44
	7,000 - 8,999	45 - 64
	9,000 -10,999	65 and over
	11,000 and over	

Because no relevant stratification variables were available, frame 2 (and the sample selected from it) was not stratified.

In frame 3, the units were stratified as follows into 4 strata by (i) type of structure and (ii) period of issuance of the building permit.

Stratum	Type Of Structure	Period Of Issuance
11	Apartment buildings and row housing	October 1969 - May 1971
12	Apartment buildings and row housing	June 1971 - December 1973
21	All other types of housing	June 1970 - May 1971
22	All other types of housing	June 1971 - December 1973

These strata were chosen primarily for operational purposes, to permit the selection of units from frame 3 independently of the selection from frames 1 and 2.

2.5.2 Sample Size

The necessary sample sizes were calculated independently for each survey area to provide a coefficient of variation of 6% or less for characteristics representing 10% or more of the population at the survey area level.

The basic size was determined by solving for n_o the relationship

$$\alpha(\hat{X}) = \frac{Q}{\sqrt{n_o P}} \quad \text{where } P = .10$$

For an α of 6%, this gives $n_o = 2{,}500$ for each survey area.

This basic size was then adjusted for

(i) the finite population correction at time t (≡1978) for each survey area

(ii) the attrition rate for sampled dwellings

(iii) an estimated non-response rate for each survey area

(iv) an additional non-sampling error rate for each survey area.

The application of these four factors gave rise to a sample size for survey area r, $n^{(r)}$, which was then allocated proportionally to each of the survey areas for an α of 6%, with P = .10

Table 1: Sample Size and Allocation for α = .06 when P = .10

Survey Area (r)	Survey Area Name	Precision: α = .06, P = .10			
		$n^{(r)}$	Frame 1 (f=1)	Frame 2 (f=2)	Frame 3 (f=3)
01	Calgary	3380	2540	162	678
02	Charlottetown	1324	1026	32	266
03	Chicoutimi	3238	2512	78	648
04	Edmonton	3388	2600	108	680
05	Halifax	3330	2558	106	666
06	Hamilton	3374	2618	80	676
07	Kitchener	3338	2536	134	668
08	London	3350	2520	160	670
09	Montreal	3432	2524	220	688
10	Ottawa	3388	2656	54	678
11	Quebec	3402	2558	164	680
12	Regina	3324	2526	132	666
13	St. Catharines	3348	2544	134	670
14	Saint John	3238	2538	52	648
15	St. John's	3210	2518	50	642
16	Saskatoon	3294	2424	210	660
17	Sudbury	3286	2602	26	658
18	Thunder Bay	3220	2498	78	644
19	Toronto	3410	2618	110	682
20	Vancouver	3396	2608	108	680
21	Victoria	3328	2528	134	666
22	Windsor	3330	2584	80	666
23	Winnipeg	3398	2608	108	682
TOTAL		74,726	57,294	2,520	14,962

2.5.3 Sample Allocation

Within each frame, the sample of units was allocated to strata and selected independently from each stratum in the following manner.

The 1971 Census of Canada (2B) File of Occupied Private Dwellings

The sample selected from this frame was stratified according to the variables outlined in 2.5.1 in the following manner. For a given value of P (and Q), calculate $n^{(r)}$. Then determine n_{oh}, such that when it is adjusted by the finite population correction within each stratum (h = 1, ..., 40) according to

$$n_h^{(r)} = \frac{n_{oh}^{(r)}}{1 + \frac{n_{oh}^{(r)}}{N_h^{(r)}}}$$

the value of $\left| \sum_{h=1}^{40} n_h^{(r)} - n^{(r)} \right|$ is a minimum, where $\alpha^{(r)}$ is the coefficient of variation r; $n^{(r)}$ is the sample size for survey area r; $n_h^{(r)}$ is the sample size of stratum h in area r; $N_h^{(r)}$ is the population size of stratum h in area r.

This procedure ensured that a uniform coefficient of variation could be expected for all stratum estimates in a given survey area, although the coefficient of variation would vary from area to area.

The 1971 Census of Canada File of Vacant Private Dwellings

Since no stratification of the units in this frame was done, no sample allocation was necessary.

Statistics Canada's Records of Issued Building Permits

The sample for the frame was allocated proportionally.

2.5.4 Sample Selection

In frame 1, an automated generalized sample selection program was used to select a stratified simple random sample (without replacement) of dwelling identification codes from the census 2B file. These identification codes were then matched against the VRs to retrieve the address. For approximately 50% of the sample addresses, this matching had already been done previously as part of the National Address Register (NAR) project undertaken by the Census Field of Statistics Canada. For those addresses, it was necessary only to retrieve the address from that file. In frame 2, a

systematic random sample (without replacement) was manually selected directly from the VRs. In frame 3, a systematic random sample (without replacement) was manually selected from the building permits summary reports.

2.5.5 Sample Processing

To facilitate the control and updating of selected dwellings for all cycles of the survey, a control list of all sample units was created. This central list contains for each dwelling: the survey identifier (survey area code plus household number), the frame and strata codes, the census identifiers (for frames 1 and 2), the civic address.

To create this central list, all selected units were transcribed on preprinted forms from which the information was then captured using the type and scan technique. This was followed by an editing and correction phase designed specifically for this control list.

2.5.6 Estimation

All estimates derived from the data collected in this survey are based upon weighted records; the sampling weights are in turn comprised of the inverse of the basic sampling frame adjusted for complete non-response; no external variables are available to further refine these estimates.

Coefficients of variation are also calculated to provide estimates of the variability of estimated totals, means, and proportion. In addition, crude sampling variability tables have been constructed as indicators of approximate sampling variability.

3. DATA COLLECTION

The data collection phase took place during the months of October, November and December 1974 in the 23 survey areas. This section of the report highlights some of the main features of this collection phase.

3.1 Enumeration

Enumeration was carried out by personal (face to face) interviewing in each survey area. Enumerators were instructed to attempt to obtain an interview with the head of the household and/or his/her spouse. Only if it could be determined that both head and spouse would be absent during the entire survey reference period (2 1/2 months) would interviews with another member of the household (18 years of age or older) be allowed.

3.2 Call-Backs

A reasonable number of call-backs were suggested for attempting to establish initial contact and/or completing a previously partially-completed questionnaire. All follow-ups to obtain missing or additional information were to be done in person unless the enumerator had explicit permission from the particular respondent concerned to obtain the information by telephone. Follow-ups were particularly important for income data and for room dimension data.

Personal income information was to be obtained by personal interview from each income recipient in the household. If this was not possible because members could not be interviewed directly, proxy response was to be allowed for all members of the household. In any event, income data for the head and spouse were to be considered more important than income data for the other members of the household.

The information on room dimensions was to be gathered by having the respondent supply the information from blueprints or other documents or by measuring the length and width of each room. Tape measures were supplied for that purpose, if required. In instances where the respondent refused or was unable to perform this task, at the time of interview, the enumerator dropped off a document on which the respondent could enter the measurements at his/her leisure. The interviewer would follow-up (either in person or by telephone) to retrieve this data.

3.3 Non-Response Procedures

If, after the maximum number of call-backs, the questionnaire had not been completed, the interviewers were instructed to record the last call-back, showing date, time and completion status on the front page of the questionnaire and the reason for the non-completed interview. In instances where no information had been obtained, interviewers were also instructed to complete a "Non-Interview" form containing such items as the address of the household, and if known, the name and telephone number, the interviewers name, the dates and times of all call-backs, and the reason for the non-interview. This form was then attached to the appropriate questionnaire and sent to their respective supervisors. The interviewer supervisors were instructed to verify each non-interview received and attempt to convert it to a successful interview.

If all failed, the documents were forwarded to the respective Regional Offices. Usually, at this point, further action would not have resulted in a completed interview; therefore, these cases were finalized as non-interviews.

3.4 Quality Control Procedures

To maintain an adequate level of quality in the data obtained during the collection phase, a sample of all completed questionnaires was reviewed weekly by senior regional office personnel. In each questionnaire, key items were checked and if responses were found to be improperly or illegibly recorded, the interviewer concerned was contacted and asked to improve the quality of work and all his/her questionnaires were examined and corrected where necessary.

In addition, a 3% sample of all interviewed households was recontacted by telephone by regional office personnel who asked a certain minimum of questions to determine if, in fact, an interviewer had called upon the household.

4. DATA PROCESSING

The output from this survey was a set of "clean" microdata files containing data items pre-specified by the sponsor. The system which generated these files consisted of series of manuals and automated data processing steps. The basic philosophy behind the design of this system was to build in sufficient generalities as to ensure a long life system for anticipated additional cycles of this survey. Certain techniques and methods such as OCR data entry, data base, data dictionaries, and an automated error detection and correction procedure (GEISHA)[1] were used. This section presents a brief outline of each of the processing steps.

4.1 Questionnaire Preparation

Upon receipt from the regional offices, all documents were verified against the master control sheets and grouped into batches. All documents were subjected to a limited quality control check to ensure all required entries were in a form that could be easily captured. Corrections were made as necessary. The questionnaires were then forwarded, by batch, to the data capture area.

4.2 Data Capture

The "type and scan" technique with the string keying method was employed to capture both the questionnaires and updates. This method requires the typing of the data to be captured on special forms in a continuous string. The typed data is then scanned by the IBM 1288 Optical Character Reader (OCR) and transformed into machine readable form. To ensure that the typists did not exceed a five percent error rate, each batch of typed documents were sample verified using the key-edit facilities. This procedure required the re-keying of a specified number of documents for each batch via keyboard to disc equipment. If the number of errors encountered exceeded the

[1] GEISHA is the acronym for the generalized edit and imputation system using the hotdeck approach developed by the Statistical Services Field of Statistics Canada.

maximum allowable for that batch, all documents in the batch were verified and all errors corrected. To maintain an error rate of less than five percent, verification of twenty-eight percent of the documents was required.

4.3 Basic Edit/Update

This sub-system for the extraction and validation of the raw data was designed so that external control could be exercised over all types and levels of editing.

The basic validation of the keyed data was controlled via the central master list (only valid questionnaire identifications were accepted) and table driven edit specifications (to ensure only valid data was captured). Each household record created from the validation procedures was then processed through a structure editing step which checked the basic consistency of the logical data flow present. Inconsistencies were listed with the validation errors and processed through a manual interface for correction via a turnaround document. Corrections were applied through the existing modules until all records were free of inconsistent or invalid data items to the extent detectable.

4.4 Processing of the Present Dwelling and Household Data

After the completion of the basic edit step, the existing data files were split into two portions, (i) those containing the variables relating to present dwelling and household characteristics and (ii) those containing the variables relating to previous dwelling and household characteristics and the mobility data. This split was necessary because the two portions were to be processed in different manners - the former with imputation, the latter without imputation, and because separate files were to be produced for each. The following additional processing steps refer to the present dwelling and household data processing only.

4.5 Editing and Imputation

The editing and imputation phase of the data processing itself consisted of three steps. The first involved editing and imputing, through the use of a tailor-made module, a limited set of key variables to which most of the other variables were logically connected. These few variables were corrected separately because of their great influence over the response pattern in the remaining variables. Following this the remaining variables were edited and imputed using a generalized package employing a hot-deck imputation routine (GEISHA). Finally, due to technical limitations of this generalized package, it was necessary to perform some further imputations on quantitative variables.

4.6 Family Formations

Following the editing/imputation stage, the demographic variables were combined to create a number of derived variables relating to size and types of families.

4.7 Weighting

All records were assigned a weight based upon the record's probability of inclusion in the sample and adjusted for complete non-response.

4.8 Confidentiality Masking

Certain variables on certain records were thought to be sufficiently unique that they might divulge information about the identity of individual respondents and thereby violate confidentiality safeguards. To overcome this, procedures and programs were adopted for masking those data items (see section 8).

Present Dwelling File Creation

Finally, micro-data files containing existing derived variables pre-specified by the sponsor were created.

The previous dwelling and household data underwent many fewer steps because (i) it was not to be imputed and (ii) the weights to be used for it were those supplied with the present dwelling and household data.

Mirco-data files containing these variables were created, matched and merged with the previous dwelling and household data files creating one record for each questionnaire.

Certain variables from the census data were also masked to meet the confidentiality requirements.

5. PUBLIC RELATIONS

Introductory letters were mailed to respondents approximately one week prior to the interview and,at the time of the interview, brochures were handed out. "Thank you" letters were mailed to all responding households.

Local CMHC offices and chiefs of police were informed of the conduct of the survey.

6. PRE-TESTS

In the spring of 1973, a pre-test on 550 households was conducted in Toronto, Ontario and Hull, Quebec. This study was intended to test some of the concepts to be measured in the main survey and to determine what difficulties might be encountered by asking for the recall of information up to three years old.

The pre-test sample was selected from the 1968 Survey of Consumer Finances and the survey was conducted by face-to-face interviewing. A debriefing of enumerators took place at the completion of the interviews; analysis of results and evaluation of the pre-test was done manually with the use of the actual questionnaires.

The major conclusions coming out of this study were that (i) there seemed to be no obvious difficulties with respondents adequately understanding the concepts asked about; (ii) there seemed to be no hostility towards this subject on the part of respondents and; (iii) recall seemed to pose no exceptionally high non-response rates.

In the winter of 1974, as part of the development of the questionnaire for the 1974 survey, a second pre-test was conducted on 300 households in the Ottawa area with the intention of determining the feasibility of obtaining room dimension data through an interview situation. Again, the study was carried out using face-to-face interviewing and was followed by a debriefing of enumerators and a manual analysis and evaluation of data collected. The major recommendation coming out of this pre-test was that, in order to obtain reliable room dimension data, respondents be asked to measure the rooms in their dwellings and that the interviewer either wait while this was done or call-back by telephone to obtain the information.

7. DATA AVAILABILITY

As mentioned earlier, the 1974 Survey of Housing Units data was split into two portions during the processing phase and each has been stored in a different manner. Within Statistics Canada, the present household and dwelling data are stored individually for each of the 23 survey areas on a direct access data base both in the imputed and unimputed state. It is accessible for tabulations via STATPAK as well as being available on sequential tape files created according to CMHC's format specifications[1].

CMHC also has created a data base for this data and have provided an interactive retrieval system for tabulation.

The sequential tape files should be available for use by third party users in 1977[2].

[1] These tapes do not include the records screened out due to confidentiality restrictions.

[2] Further information concerning the cost and availability of these files can be obtained by contacting the Special Surveys Co-ordination Division, Statistics Canada.

The previous household and dwelling data and the data retrieval from the 1971 Census for linkage purposes is restricted to use within Statistics Canada and within CMHC only in the sequential tape mode in the unimputed state.

These files will not be made available for use by third party users.

8. MICRO-DATA RELEASE SCREEN

In order to preserve confidentiality, the data on the micro-data files released from this survey has been screened to remove the possibility of identification of individuals. The following steps have been taken:

I. 1974 Data

1. all records with household income greater than $100,000 have been deleted from the files,
2. all ages of individuals greater than 75 are coded 76 on the file,
3. all principals outstanding on mortgages $75,000 and greater, are coded $75,000 on the file,
4. all market values $150,000 and greater, are coded $150,000 on the file.

For each survey area, the following summary of the results of the micro-data release screen is available:

(i) Total weighted amount of income deleted.

(ii) Total number of records deleted.

(iii) Total record weight deleted.

(iv) Total weighted amount of principals outstanding on mortgages greater than $75,000, i.e. the weighted difference between the actual amount and $75,000.

(v) Total number of records with principals outstanding on the mortgage $75,000 or greater.

(vi) Total weighted amount of market values greater than $150,000 i.e. the weighted difference between the actual value and $150,000.

(vii) Total number of records with market values $150,000 or greater.

II. Previous Dwelling Data

1. all information for records with household income greater than $100,000 has been deleted from the files.
2. all ages of individuals greater than 75 are coded 76 on the file.
3. all principals outstanding on mortgages $75,000 and greater, are coded $75,000 on the file.
4. all selling prices $150,000 and greater are coded $150,000 on the file.

III. Census Data

1. all ages of individuals greater than 75 are coded 76 on the file.
2. all census data for records with household income greater than $100,000 have been dropped.
3. the income of all heads of households $75,000 or greater is coded $75,000 on the file with the exception of female heads of households in survey areas in the Atlantic Region where all income $50,000 or greater is coded $50,000 on the files.

ACKNOWLEDGEMENT

The author wishes to acknowledge the contribution of project team members R. Haddow, S. Porter, C. Scott and G. Smith in the preparation and writing of the report upon which this paper is based.

RESUME

L'enquête de 1974 sur les logements menée par Statistique Canada au cours de l'automne de 1974 était parrainée par la Société centrale d'hypothèques et de logement. Statistique Canada était chargé d'élaborer et de mettre en oeuvre toutes les phases de l'enquête, y compris la production de bandes magnétiques de données "sans erreur". Pour sa part, la Société était responsable de l'établissement des objectifs et des besoins en termes de données, ainsi que de l'analyse des résultats.

Ce rapport est une modification du sommaire préparé par l'équipe responsable, une fois le projet terminé; il consiste en un exposé général de l'enquête et de l'apport de Statistique Canada.

APPENDIX

Details of Estimation Procedures

1. Estimates

Each of the frames used in this survey is different from the others with respect to the information it contains about the units on it and with respect to the manner in which it is stored. For these reasons a different design was used in selecting a sample of units from each frame. This in turn results in three different sets of estimation techniques being required.

This section presents formulae for the calculation of weights, estimates of domain values and their estimated coefficient of variation for each of the three subpopulations.

Although the sampling unit used in this survey is the dwelling unit, the unit for which estimates are desired is the occupied dwelling (household), a domain of each of the three subpopulations. Consequently, all estimates discussed in this section refer to this domain.

2. The 1971 Census of Canada File of Occupied Private Dwellings

The units in this frame were stratified into 40 strata using the variables: tenure of dwelling, income of head (and spouse), age of head. Within each stratum a simple random sample of units without replacement was selected for each survey area.

<u>Notation</u>

h: subscript denoting the hth stratum (h=1,...,40)

i: subscript denoting the ith unit (record)

N: number of subpopulation units (i.e. number of occupied private dwellings on June 1, 1971)

N' : (unknown) number of subpopulation units in domain of interest (i.e. number of dwellings occupied during reference period in subpopulation)

$\hat{N}'$: estimated number of subpopulation units in domain of interest (i.e. number of dwellings occupied during reference period in subpopulation)

N_h: number of subpopulation units in stratum h

N_h': number of subpopulation units in stratum h in the subsample (i.e. on the Census 2B file)

Corresponding sample values are expressed by replacing N by n.

X: domain total for characteristic X

$\bar{X}$: domain mean for characteristic X

x_{hi}: value of characteristic X taken on by the ith sample unit in stratum h

$\bar{x}_h$: sample mean in domain of interest for characteristic x for stratum h

$V(\hat{X})$: variance of estimated domain total

$V(\hat{\bar{X}})$: variance of estimated domain mean

$\alpha(\hat{X})$: coefficient of variation of estimated domain total

$\alpha(\hat{\bar{X}})$ coefficient of variation of estimated domain mean

<u>Weights</u>

Three factors must be taken into account in calculating weights for units from this subpopulation.

i) the subsampling factor (census household weight) associated with each unit on the frame

ii) the basic selection probability for each unit selected from the frame

iii) the non-response factor applied to each selected unit as a result of non-response to the survey.

The first weight is necessitated by the fact that the census file used represents only approximately 1/3 of all occupied private dwellings. So each record on the file has associated with it a weight with a value of 3. This first weight is denoted by W_{h_1}.

The second and third weights are calculated separately at the stratum level and are applied to each record in that stratum. The second weight is accounted for by:

$$W_{h_2} = \frac{N_h'}{n_h}, \ h = 1, \ldots, 40.$$

The third weight is accounted by:

$$W_{h_3} = \begin{cases} \dfrac{n_h}{n_h'} & \text{if } n_h' \neq 0 \\ 0 & \text{if } n_h' = 0 \end{cases}$$

The combination of W_{h_1}, W_{h_2} and W_{h_3} results in a basic stratum weight

$$W_h = W_{h_1} W_{h_2} W_{h_3} = \begin{cases} \dfrac{N_h}{n_h'} & \text{if } n_h' \neq 0 \\ 0 & \text{if } n_h' = 0 \end{cases}$$

Estimates

Domain total: $\hat{X} = \sum_{h=1}^{40} W_h \sum_{i=1}^{n_h''} x_{hi}$

Domain mean : $\hat{\bar{X}} = \dfrac{\hat{X}}{\hat{N}_d}$ where $\hat{N}_d = \sum_{h=1}^{40} W_h n_h''$

Coefficient of Variation

To calculate the estimated coefficient of variation of an estimate, it is first necessary to calculate the estimated variation. This is given by

For domain total: $\hat{V}(\hat{X}) = \sum_{h=1}^{40} N_h (N_h - n_h') \dfrac{S_{ha}^2}{n_h'}$

where $$s_{ha}^2 = \frac{1}{n_h' - 1} \left[\sum_{i=1}^{n_h''} x_{hi}^2 - \frac{\left(\sum_{i=1}^{n_h''} x_{hi}\right)^2}{n_h'} \right]$$

For domain mean: $$\hat{V}(\hat{\bar{X}}) = \frac{1}{(\hat{N}')^2} \left[\sum_{h=1}^{40} N_h \, (N_h - n_h') \, \frac{s_{hb}^2}{n_h'} \right]$$
$$+ \frac{1}{(\hat{N}')^2} \left[\sum_{h=1}^{40} \frac{N_h \, (N_h - n_h')}{n_h' \, (n_h' - 1)} \quad n_h'' \, (1 - \frac{n_h''}{n_h'}) \, (\bar{x}_h - \hat{\bar{X}})^2 \right]$$

where $$\hat{N}' = \sum_{h=1}^{40} \frac{N_h}{n_h'} n_h'' = \sum_{h=1}^{40} W_h \, n_h''$$

and $$s_{hb}^2 = \frac{1}{n_h' - 1} \left[\sum_{i=1}^{n_h''} x_{hi}^2 - \frac{\left(\sum_{i=1}^{n''} x_{hi}\right)^2}{n_h''} \right]$$

This leads to estimated coefficients of variation as follows.

For domain total: $\hat{\alpha}(\hat{X}) = \sqrt{\hat{V}(\hat{X})/\hat{X}}$

For domain mean: $\hat{\alpha}(\hat{\bar{X}}) = \sqrt{\hat{V}(\hat{\bar{X}})/\hat{\bar{X}}}$

3. The 1971 Census of Canada File of Vacant Private Dwellings

Since none of the units on this frame have dwelling or household characteristics associated with them, no stratification was performed on the units. In each survey area, one systematic random sample of units was selected without replacement from the 1971 Census of Canada Visitation Records.

Consequently, the notation and formulae in this sub-section are a special case of those in sub-section 2 with h=1 with the one following exception.

In the calculation of the estimated variance for a domain mean,

$$s_b^2 = \frac{1}{n'' - 1} \left[\sum_{i=1}^{n''} x_i^2 - \frac{\left(\sum_{i=1}^{n''} x_i \right)^2}{n''} \right]$$

4. Statistics Canada's Records Of Issued Building Permits

Because of the problems associated with the use of this frame (see 1974 Survey of Housing Units/A Report On the Sample Design), the selection probabilities of sample units are very difficult to calculate. To simplify procedures, the sample can be weighted, not according to building permit issuance counts, but rather according to dwelling completion counts.

Since dwelling completion counts are not available at the municipality level by type of dwelling, the stratification used in the selection procedure will not be employed in the estimation procedures. Consequently, this sample will be considered to have been selected by systematic random sampling without replacement and without replication from an unrestricted population in each survey area.

Thus the notation and formulae for estimates obtained from this frame are identical to those in sub-section 3.

5. Population Estimates

As mentioned in section 1, the assumption is made here that the subpopulations are independent of each other. The validity of this assumption will not be known until an evaluation of the survey design and procedures has been done.

By treating the subpopulations as independent of one another, the estimates for the domain of interest over the entire population can be obtained by treating the units of each of frames 2 and 3 as belonging to two independent strata; consequently all summations are over 42 strata.

Estimates

Domain total: $\hat{X} = \sum_{h=1}^{42} \hat{X}_h$

Domain mean: $\hat{\bar{X}} = \dfrac{\sum_{h=1}^{42} \hat{X}_h}{\sum_{h=1}^{42} \hat{N}'_h}$

Coefficient of Variation

To calculate the estimates coefficient of variation of an estimate, it is first necessary to calculate the estimated variances. This is given by:

For domain total: $\hat{V}(\hat{X}) = \sum_{h=1}^{42} \hat{V}(\hat{X}_h)$

For domain mean:

$$\hat{V}(\hat{\bar{X}}) = \frac{1}{(\hat{N}')^2} \left[\sum_{h=1}^{42} N_h (N_h - n'_h) \frac{s^2_{hb}}{n'_h} \right.$$

$$+ \frac{1}{(\hat{N}')^2} \left[\sum_{h=1}^{42} \frac{N_h (N_h - n'_h)}{n'_h (n'_h - 1)} n''_h \left(1 - \frac{n''_h}{n'_h}\right) (\bar{x}_h - \hat{\bar{X}})^2 \right.$$

where, for h = 41, 42: $s^2_{hb} = \dfrac{1}{n'_h - 1} \left[\sum_{i=1}^{n''_h} x^2_{hi} - \dfrac{\sum_{i=1}^{n''_h} x^2_i}{n''_h} \right.$

and for h = 1, ..., 40, s^2_{hb} is as in sub-section 2.

STRATIFICATION INDEX: METHODOLOGY AND ANALYSIS

G.B. Gray
Household Surveys Development Division

To obtain estimates of means or totals for a universe, a sample of units is often drawn to represent the universe and these units are then surveyed. One of the most important procedures used in the selection of the units is that of stratification, whereby the universe is split up into strata and independent samples of units are drawn from each stratum. A stratification index is developed to indicate the approximate fractional reduction in the sampling variance from that which would result if no stratification were undertaken. Also the methodology is extended to examine the effect of stratification on the sampling variance at different levels of stratification through the concept of a summary index. The stratification index is also extended to the case of ratio estimates using independent source data to re-weight the sample data. The index has been applied to the Canadian Labour Force Survey (LFS), a typical multi-stage stratified sample where ratio estimation, using projected age-sex population estimates is applied and empirical data are presented and analyzed.

1. INTRODUCTION

Stratification basically belongs to one of two categories: a) administrative, such as province, city, or other necessary area for which estimates are needed, and b) optimal for the purpose of maximizing the mean square errors between strata so as to derive an estimate with as low a variance as possible for the available resources. In the case of optimal stratification, the strata so delineated rarely conform to well-defined administrative areas for which estimates may be needed. The sole purpose of these strata is to reduce the sampling variance of estimates of a given area as much as possible rather than obtain estimates for the individual strata. Estimates can be obtained for an area that consists of partial as well as complete strata with usually a considerable loss of efficiency over those estimates that would result if the delineated strata had honoured the stratum boundaries. Thus, when estimates are required for certain domains, administrative strata are usually delineated to permit estimates in these domains.

To determine whether stratification has been effective in reducing the sampling variance from that which would have resulted if there had been no stratification, a stratification index is developed in the context of multi-stage sample designs, employing the simple estimation procedure in Section 2. The summary indexes over separate areas are so developed in Section 3. In Section 4, the LFS design and estimation procedure is described and the application of the methodology of stratification indexes to LFS is considered. In Section 5, the adoption of the index to ratio estimates is considered, again with reference to LFS. Some results are anticipated in Section 6 on the basis of intuition while the empirical results pertaining to a 10-month period of LFS (Mar.-Dec., 1975) are presented and analyzed in Section 7.

2. DEFINITION OF STRATIFICATION INDEX

To develop the stratification index[1] in the context of multi-stage stratified designs, let us consider an area A consisting of L strata. In stratum h, suppose that N_h primary sampling units (PSU's) have been delineated, and that n_h PSU's have been selected with probability proportional to size (pps) with or without replacement.

Let p_h be the relative size of stratum h in area A and $p_{i|h}$ be the relative size of PSU i in stratum h.

The estimate of the characteristic total X for area A is given by

$$\hat{X}_s = \sum_h \hat{X}_h = \sum_h \sum_i \hat{X}_{hi}/(n_h\ p_{i|h}) \quad \text{when stratification is} \qquad (2.1)$$

undertaken and

$$\hat{X}_{\bar{s}} = \sum_h \sum_i \hat{X}_{hi}/(np_h\ p_{i|h}) \text{ when } n = \sum_h n_h \text{ PSU's are} \qquad (2.2$$

selected with or without replacement in area A, ignoring stratification.

[1] Some preliminary work in this connection had been undertaken by Fellegi [1].

$\hat{X}_{hi}$ estimates X_{hi}, whatever the sample design undertaken in (h,i) and we shall assume here that the sampling procedure within each PSU is the same whether stratification is undertaken or not although in self-weighting designs, there could be slight changes in the sampling fractions within the selected PSU's without stratification from those with stratification.

The stratification index is then defined by

$$I \;=\; [V(\hat{X}_{\bar{s}}) - V(\hat{X}_s)]/V(\hat{X}_{\bar{s}}), \text{ where} \qquad (2.3)$$

$V(\hat{X}_s)$ and $V(\hat{X}_{\bar{s}})$ are sampling variances of $\hat{X}_s$ and $\hat{X}_{\bar{s}}$ respectively.

The stratification index as defined in (2.3) includes the effect of varying sizes of strata and numbers of primary sampling units between strata as well as the variance between strata.

To account solely for the variance between strata, another stratification index is defined by

$$I' \;=\; (L^2 \sigma^2_{BS}/n)/[V(\hat{X}_s) + L^2 \sigma^2_{BS}/n] \qquad (2.4)$$

where $n = \sum_h n_h$ primary sampling units have been selected in the L strata, and σ^2_{BS} is the population variance between strata (algebraically defined in the Appendix). The reason for using $L^2 \sigma^2_{BS}/n$ along with $V(\hat{X}_s)$ in I' is made clear in the appendix. In this paper, the analysis deals solely with the index as defined by 2.4 rather than 2.3.

The indexes I and I' are fully developed and discussed in the appendix.

3. SUMMARY INDEX

In most sample survey designs, the question may not simply arise as to the merits of stratification as opposed to no stratification but rather stratification at one level or two levels, where stratification at the second level simply implies deeper stratification within first level strata, i.e., once strata at the first level have been assumed or delineated, smaller strata at the second level may be delineated within.

Summary indexes for both I and I' may be obtained, as follows:

$$I_A = \frac{V_{\bar{s}} - V_s}{V_{\bar{s}}}$$

and to distinguish area A, we shall redefine this as follows:

$$I_A = \frac{V_{\bar{s}A} - V_{sA}}{V_{\bar{s}A}}$$

Similarly

$$I'_A = \frac{L_A^2\ \sigma^2_{BS:A}/n_A}{V_{sA} + L_A^2\ \sigma^2_{BS:A}/n_A}$$

If A denotes a stratum at the 1st level while h denotes one at the second level, then summary indexes may be defined, as follows:

$$\bar{I} = \sum_A (V_{\bar{s}A} - V_{sA})/\sum_A V_{\bar{s}A} \tag{3.1}$$

and

$$\bar{I}' = (\sum_A L_A^2\ \sigma^2_{BS:A}/n_A) \div [\sum_A (V_{sA} + L_A^2\ \sigma^2_{BS:A}/n_A)] \tag{3.2}$$

and these may be readily estimated by summing the estimates of the numerators and denominators over areas A. $\bar{I}$ and $\bar{I}'$ may be written as follows:

$$\bar{I} = [\sum_A V_{sA}\ I_A/(1-I_A)]/[\sum_A V_{sA}/(1-I_A)] \tag{3.3}$$

and

$$\bar{I}' = [\sum_A V_{sA}\ I'_A/(1-I'_A)]/[\sum_A V_{sA}/(1-I'_A)]\ . \tag{3.4}$$

If I_A's are readily available but individual V_{sA}'s are not, such as may be the case if averages over several months are calculated, one may obtain approximate summary indexes by multiplying $I_A/(1-I_A)$ and $1/(1-I_A)$ by $X_A(1-X_A/P_A)(W_A-1)$, where

V_{sA} = $F_A\ \hat{X}_A(1-\hat{X}_A/\hat{P}_A)(W_A-1)$ Lawes [5], where

F_A = design effect for area A

$\hat{X}_A$ = estimate of characteristic total in area A

$\hat{P}_A$ = population estimate from the sample in area A

W_A = theoretical inverse sampling fraction in area A

In the process of obtaining $\bar{I}$ and $\bar{I}^1$, changes in F_A may be ignored if they are not readily available.

4. LABOUR FORCE SURVEY DESIGN

The design of the Canadian Labour Force Survey (LFS) is a multi-stage stratified sample with strata at several levels and two to four stages of sampling. The primary strata are the ten provinces, the secondary level are type of area (self-representing units or SRU and non-self-representing units or NSRU). The third level strata in the NSRU areas are the economic regions [6] and the fourth levels are strata delineated within economic regions. In the SRU areas,the second level of strata are the metropolitan areas or large cities called individual SRU's while the third level are subunits delineated within SRU's. In each NSRU stratum, up to 19 primary sampling units (PSU's) are delineated out of which 2 to 6 are selected with pps systematic. Sub-sampling is undertaken in each selected PSU in two or three more stages. In each subunit of each SRU, the random group method of selection with the PSU's being clusters of dwellings and the sub-sampling units dwellings. Further details on the stratification and selection procedure may be obtained in [6].

Two types of estimates of totals of characteristics are produced in the LFS; viz, the subweighted estimates which are simple inflated totals and final weighted estimates which are multiple ratio estimates based on post-stratification by 20 age-sex cells within each province. The final weights are the subweights adjusted so that population estimates tally to projected values. Further details on the weighting may again be obtained in [6]. Since all sampling variances and their estimates may be obtained for ratio estimates so stratification indexes may also be obtained for ratio estimates and the adjustments in the statistics are presented in the next section.

5. ADAPTATION TO RATIO ESTIMATES

Significant gains in the efficiency of the statistics may be affected by ratio estimation over simple estimation as for example, in the LFS where the characteristics are often highly correlated with age-sex populations. In order to determine whether or not further gains have been accomplished as a result of stratification, the statistics used in the variance and variance estimates must be appropriately adjusted. The following adjustments to the formulas of section 2 are given below.

$$\hat{X}'_s = \sum_a (P_a/\hat{P}_a)\ \hat{X}_{Aas} \qquad (5.1)$$

where ' refers to ratio estimate for area A. P_a and $\hat{P}_a$ are the projected and simple population estimates for category a (eg., age-sex groups in LFS) at the province level.

$\hat{X}_{Aas}$ = estimate of characteristic total obtained by 2.1 for area A. At province level, $\hat{X}_{Aa} = \hat{X}_a$.

Similarly $$\hat{X}'_{\bar{s}} = \sum_a (P_a/\hat{P}_a)\ \hat{X}_{Aa\bar{s}} \qquad (5.2)$$

When area A is the province, all variance and variance estimates are identical in the abridged notation with σ^2's and r_{FP}'s. However σ^2's and r_{FP}'s in all cases must be re-defined with

$$\begin{aligned}
X_{hi} &\text{ replaced by } X_{hi} - \sum_a (X_a/P_a)\, P_{hia}, \\
\hat{X}_{hi} &\text{ replaced by } \hat{X}_{hi} - \sum_a (\hat{X}_a/\hat{P}_a)\, \hat{P}_{hia}, \\
X_h &\text{ replaced by } X_h - \sum_a (X_a/P_a)\, P_{ha}, \text{ etc.}
\end{aligned} \qquad (5.3)$$

When summary indexes at province level are desired, the individual indexes at subprovincial domains as applied to ratio estimates may be calculated, using the substitutions of 5.3 but the individual variances and hence the indexes at subprovincial domains do not refer to the gain in efficiency of subprovincial estimates for these domains but instead to the gain in the efficiency of provincial ratio estimates contributed by the subprovincial domain.

In order to assess the gain in efficiency of $\hat{X}'_s$ over $\hat{X}'_{\bar{s}}$ in a subprovincial area A, $V(\hat{X}'_s)$ must be estimated by a more complex formula given by Gray and Ghangurde [4]. $V(\hat{X}'_{\bar{s}})$ is smaller in form to $V(\hat{X}'_s)$ except for the lack of stratum breakdowns in A. However, an estimate $\hat{V}(\hat{X}'_{\bar{s}})$ from the stratified sample s in A remains to be worked out.

6. ANTICIPATION OF THE RESULTS

Before studying any empirical results, it would be interesting to anticipate some possible results. The NSRU Economic Regions were stratified on the basis of "important" industry classifications, while SRU's were delineated into subunits on the basis of counts of blocks and block faces honouring to a great extent, census tracts, but disregarding the LF characteristics. Thus, one would expect the indexes to be higher in NSRU areas than in SRU areas, and in NSRU areas, one would in turn expect the indexes to be higher among the industry components than among the more general characteristics such as Employed and Unemployed.

It should be noted that stratification indexes are very unstable, even though 10-month averages were obtained. Successive observations in any given period, however, are highly correlated, especially among the industry breakdowns because of the large fraction of commonly sampled individuals month to month as well as the high measure of homogeneity for Employed by Industry (see [3]). The observed index as only an estimate of a so-called theoretical index could deviate far from the theoretical value because of its instability, and as mentioned before, may even be negative.

One may anticipate higher indexes for NSRU areas completely than for economic regions, since the index for NSRU areas completely compares the current variance with the variance if there were no stratification at all in the NSRU areas (not even by economic region), while the province NSRU summary index defined by the weighted average index over the economic region NSRU areas compares the current variance with that resulting from defining ER's as strata, but performing no further stratification within ER's.

7. TABLES OF STRATIFICATION INDEXES

For 8 characteristics, Employed (Emp.), Unemployed (Unemp.), Employed Agriculture (Emp. Ag.), Employed Non-Agriculture (Non-Ag.), Employed Manufacturing (Manuf.), Employed Construction (Constr.), Employed, Transportation and Public Utilities (TPU), and Employed Trade (Trade), the following stratification indexes were obtained for the ten month period (March-Dec., 1975) in the old LFS and averaged over the whole period.

Table 1: I'_{1p} or the stratification index pertaining to type of area T, province p and the summary index $\bar{I}'_T$ over the provinces. (T=1 denotes SRU areas and T=2 denotes NSRU areas).

Table 2: $\bar{I}'_{2p}$ or the summary index over all economic regions of the NSRU portion of each province p and $\bar{\bar{I}}'_2$, the summary index over all NSRU portions of each province and over the provinces $\bar{\bar{I}}'_2$ is compared with $\bar{I}'_2$.

Table 3: I'_M or the stratification index for each of 10 metropolitan areas and $\bar{I}'_{\underline{M}}$ is the summary index over the 10 metropolitan areas.

All indices above apply to the ratio estimates so that they refer to the reduction in the variances at sub-provincial domains only as the domains are portions of the province since all ratios used in ratio estimates are obtained at provincial levels only.

Corresponding indices for simple estimates for the above areas would have been very interesting and important for the cost-benefit study of the extensive work in delineating the strata but they are not available for the time being.

8. ANALYSIS OF TABLE 4

The summary indexes $\bar{I}'_T$ or the weighted average index by type of area over the 10 provinces (I'_{Tp}) tend to be higher for NSRU areas than for SRU areas as one had anticipated, as they are higher for 7 or 8 characteristics. Estimated reductions of over 40% were realized for Employed Manufacturing and Employed Agriculture in NSRU areas, while for Employed, the reduction was less than 30%. For Unemployed, the reduction decreased to about 13%. In SRU areas, except for Employed Manufacturing with a reduction of 17.3%, the indexes indicated insignificant reductions there between 2 and 10%.

Individual indexes at province/type of area levels were very spotty in both types of areas with the large gains due to stratification occurring in Ontario and Quebec, smaller gains in B.C. and the Atlantic Provinces, but surprisingly, negligible gains in the Prairie Provinces with only a few isolated cases of gains exceeding 20%, mainly in Employed Manufacturing.

In Saskatchewan NSRU, an odd result can be seen which indicates little stratification benefit for Employed (Index of .086), and for Employed Agriculture (Index of only .018), but an

index of .282 for Employed Non-Agriculture. No explanation for this strange phenomenon could be obtained. Apparently, Employed, Agriculture is so wide-spread in the Prairies NSRU that stratification does little to decrease the sampling variance of Agriculture estimates there.

Negative observed indexes tend to occur in the Prairies for Employed, and in the Atlantic region for Employed: Trade. A few negative observed indexes also occur among Employed: Transportation and Public Utilities. Apart from these cases, only a few negative indexes are observed.

9. ANALYSIS OF TABLE 5

Two summary indices, at the Canada NSRU level; viz., $\bar{\bar{I}}'_2$ and $\bar{I}'_2$ may be readily compared. $\bar{I}'_2$ (based on no stratification within NSRU areas of each province and also recorded on Table 1) is higher for 5 of 8 characteristics than $\bar{\bar{I}}'_2$ (based on stratification by ER's but no deeper stratification within), and the indices are higher among those characteristics for which the stratification is effective (Employed, Employed Agriculture, Employed Manufacturing). For Employed Non-Agriculture, Construction and Trade there appeared to be no gain, due to stratification within ER's over stratification down only to the ER level in NSRU areas.

10. ANALYSIS OF TABLE 6

Two summary indices $\bar{I}'_1$ and $\bar{I}'_M$ at the Canada SRU level may be readily compared. $\bar{I}'_1$ (index for stratification vs. no stratification within SRU areas of each province and also recorded on Table 1) do not differ much from $\bar{I}'_{\underline{M}}$ (index for deeper stratification vs. no deeper stratification within metropolitan areas, averaged over the met areas). The comparisons are somewhat muddy, since $\bar{I}'_1$ applies to all Canada SRU, while $\bar{I}'_{\underline{M}}$ applies only to the 10 major metropolitan areas denoted by $\underline{M}$.

11. CONCLUSIONS AND RECOMMENDATIONS

In the NSRU areas, stratification indices $\bar{\bar{I}}'_2$ indicates that deeper stratification within ER's removed 8% of the sampling variance of Employed

at the Canada NSRU level, 6% for Unemployed, but 22% for Employed Agriculture and 26% for Employed Manufacturing. Smaller reductions between 3% and 17% were accomplished for other characteristics. The overall reductions as a result of stratification by ER as well as Deeper Stratification within ER's (given by $\bar{I}_2'$) are even more striking: 29% for Employed, 13% for Unemployed and over 40% for Employed Agriculture and Employed Manufacturing.

In the SRU areas, the results were not so striking as the reductions caused by stratification by delineation of subunits of metropolitan areas (indicated by $\bar{I}_M'$) were only between 3% and 10%. Nor was the overall reduction as a result of stratification by city and by delineation into subunits (given by $\bar{I}_1'$) very startling for any characteristics. One must realize, however, that the stratification by individual cities is largely administrative rather than optimal according to our original definitions. Despite the small reductions in the variance as a result of subunit delineation, the procedure remains a necessary one for Sample Control and Assignment Control purposes.

Only old LFS survey data was used in the analysis here. Similar analysis of the recently revised LFS data should be undertaken in a similar way utilizing 1976 data. The indexes should be calculated for the same characteristics for the same areas, or as closely as possible the same areas to determine if there is any increase in the stratification index.

It should be emphasized that stratification indexes are very unstable statistics, even more so than between PSU components of variance, since relatively few degrees of freedom exist for estimates of between and within stratum MSE's, and the approximate stratum effect in the MSE's must be derived by subtraction in much the same manner as individual variance components. Consequently, as in this paper, it will be necessary to average the indexes continuously beginning January 1976. It is not recommended to use 1975 data because of the unstable results anticipated as a result of the random drop of 1/4 of the NSRU PSU's across Canada.

ACKNOWLEDGEMENTS

The author acknowledges the work of B. Bogart (Computer Systems Development Division), M. Lawes (Methodology and Research Sub-Division) and R. Boyer (Senior Mathematical Advisory Group) in making the empirical data possible. The comments of the editor and the referee are also appreciated, especially in connection with the application of the Sukhatme reference to the above article.

RESUME

Dans le but d'obtenir des estimations de moyennes ou de totaux se rapportant à un univers donné, il arrive souvent que l'on choisisse un échantillon et enquête les unités de cet échantillon afin de représenter l'univers en question. Lors de la sélection des unités, l'une des techniques les plus utilisées est la stratification qui consiste à diviser l'univers en strates et à choisir des échantillons indépendants de chacune des strates. Ici, un indice de stratification est developpé afin de mesurer la réduction fractionnelle approximative de la variance échantillonnale imputable à la stratification. Une extension de la méthode permet d'étudier l'effet de la stratification sur la variance échantillonnale en considerant différents niveaux de stratification; ceci s'obtient en utilisant le concept d'un indice sommaire. L'indice de stratification est généralisé au cas de l'estimation par quotient où des données auxiliaires indépendantes sont utilisés pour repondérer les données de l'échantillon. L'enquête canadienne sur la population active sert d'illustration à l'application d'un tel indice et à son analyse; cette enquête, on le sait, est basé sur un échantillon stratifié, à plusieurs degrés avec estimation par quotient utilisant les estimations projetées de la taille de groupes par âge et par sexe.

REFERENCES

[1] Fellegi, I.P., "Some Problems and Possibilities in Connection With A Design For A Rotating Sample Involving The Selection Of Two PSU's Per Stratum With PPS Without Replacement", Unpublished notes prepared around 1961 prior to 1961 Census Redesign of LFS.

[2] Gray, G.B., "Definition and Analysis of Stratification Indexes (Old LFS Design)", HSDS Technical Memorandum, February 2, 1976, pp. 23, 24.

[3] Gray, G.B., "Components of Variance Model in Multi-Stage Stratified Samples", HSDS Survey Methodology Journal, Vol. 1 No. 1 (June, 1975), pp. 27-43.

[4] Gray, G.B., and Ghangurde, P.D., "On A Ratio Estimate With Post-Stratified Weighting", HSDS Survey Methodology Journal, Vol. 1 No. 2 (Dec., 1975), pp. 134-143.

[5] Lawes, M., "A Comparison Of Some Binomial Factors For The Canadian Labour Force Survey", HSDS Survey Methodology Journal, Vol. 1 No. 1 (June, 1975), pp. 59-73.

[6] Platek, R., and Singh, M.P., "Methodology Of Canadian Labour Force Survey", Household Surveys Development Division, 1976.

[7] Sukhatme, P.V., "Sampling Theory Of Surveys With Applications", pp. 389-395, Iowa State University Press, 1954.

Table 1: Average Index of Stratification by Characteristic and Type of Area (I'_{Tp} and $\bar{I}'_T$) for the period March-December 1975 (old LFS design)

	Characteristics							
Province (type of area)	Employed	Unemp	Emp Ag	Emp Non-Ag	Emp Manuf	Emp Const	Empl Transp Pub Utilit	Emp Trade
Nfld SRU	.112	.110	.022	.101	.205	.079	.263	.090
Nfld NSRU	.212	.177	.016	.180	.244	.351	-.068	-.081
PEI SRU	-.013	-.006	.079	.007	.054	.032	-.031	-.010
PEI NSRU	.250	.392	-.152	.000	.327	.071	.061	-.229
NS SRU	.146	.030	.056	.196	.172	.032	.123	.085
NS NSRU	.354	.192	.247	.130	.114	.063	.215	-.067
NB SRU	.020	.063	.053	.019	.160	.009	.136	-.004
NB NSRU	.546	.572	.437	.350	.163	.162	.365	-.119
Que SRU	.053	.044	.066	.053	.112	.023	.048	.017
Que NSRU	.336	.217	.415	.246	.534	.055	-.058	-.060
Ont SRU	.067	.060	.146	.076	.225	.071	.075	.022
Ont NSRU	.319	.005	.541	.017	.431	.056	.296	.005
Man SRU	-.014	.024	.069	.002	.117	-.003	.074	.088
Man NSRU	.154	.023	.019	.096	.308	.096	-.214	.313
Sask SRU	-.011	.014	.032	.048	.022	.017	-.034	.054
Sask NSRU	.086	-.022	.018	.282	.077	.208	-.097	.139
Alta SRU	-.024	.016	.058	-.001	.041	.024	.006	.046
Alta NSRU	-.044	.144	-.013	.050	.463	.292	.041	-.028
BC SRU	.047	.046	.018	.063	.167	.004	.078	.066
BC NSRU	.114	.018	.507	-.106	-.016	.026	.225	.029
Can SRU	.055	.050	.102	.063	.173	.045	.065	.028
Can NSRU	.287	.129	.437	.135	.429	.093	.142	.000

eg., Nfld SRU (index = I'_{10}); Nfld NSRU (index = I'_{20}); Canada NSRU (index = $\bar{I}'_2$), Canada SRU (index = $\bar{I}'_1$)

Table 2: Average Summary Indexes of Stratification by Characteristic $\bar{I}'_{2p}$ and I'_{2p} at Province and National Levels for the Period March-December 1975 (Old LFS Design)

Province (type of area)		Employed	Unemp	Emp Ag	Emp Non-Ag	Emp Manuf	Emp Const	Empl Transp Pub Util it	Emp Trade
		Characteristics							
Nfld NSRU	$\bar{I}'_{20}$	.175	.215	-.063	.105	.167	.124	-.090	.012
Nfld NSRU	I'_{20}	.212	.177	.016	.180	.244	.351	-.068	-.081
PEI NSRU	$\bar{I}'_{21}$	.250	.392	-.152	.000	.327	.071	.061	-.229
NS NSRU	$\bar{I}'_{22}$	.174	.033	.019	.103	.160	.026	.111	.015
NS NSRU	I'_{22}	.354	.192	.247	.130	.114	.063	.215	-.067
NB NSR	$\bar{I}'_{23}$	.355	.283	.149	.449	.164	.191	.300	.053
NB NSRU	I'_{23}	.546	.572	.437	.350	.163	.162	.365	-.119
Que NSRU	$\bar{I}'_{24}$	.152	.103	.110	.122	.318	.088	-.071	.156
Que NSRU	I'_{24}	.336	.217	.247	.246	.534	.055	-.058	-.060
Ont NSRU	$\bar{I}'_{25}$	.243	.002	.325	.224	.324	-.001	.110	.045
Ont NSRU	I'_{25}	.319	.005	.541	.017	.431	.056	.296	.005
Man NSRU	$\bar{I}'_{26}$	.196	-.008	.086	.109	-.077	-.055	-.007	.115
Man NSRU	I'_{26}	.154	.023	.019	.096	.308	.096	-.214	.313
Sask NSRU	$\bar{I}'_{27}$	.134	.044	.097	.377	.085	.290	-.108	.271
Sask NSRU	I'_{27}	.086	-.022	.018	.282	.077	.208	-.097	.139
Alta NSRU	$\bar{I}'_{28}$	-.079	.045	.059	.041	.275	.198	.031	.142
Alta NSRU	I'_{28}	-.044	.144	-.013	.050	.463	.292	.041	-.028
BC NSRU	$\bar{I}'_{29}$	.148	.063	.400	-.062	-.052	.213	-.014	.006
BC NSRU	I'_{29}	.114	.018	.507	-.106	-.016	.026	-.225	.029
Can NSRU	$\bar{\bar{I}}'_{2}$	.287	.129	.436	.135	.429	.093	.142	.000
	$\bar{I}'_{2}$	.181	.064	.224	.168	.265	.095	.025	.104

Table 3: Average Index of Stratification, I'_M by Characteristic (Mar-Dec 1975) for Specified Met Areas

	Characteristic							
Met Area	Employed	Unemp	Emp Ag	Emp Non-Ag	Emp Manuf	Emp Const	Emp Transp Pub Utilit	Emp Trade
Halifax	.050	-.028	.140	.064	.058	.090	.047	.047
Quebec City	.179	.169	.234	.301	.269	.049	.100	.000
Montreal	.014	.060	.070	.032	.031	.028	.049	.031
Ottawa	.017	.108	.238	.033	.315	-.023	-.032	.002
Toronto	.048	.065	.209	.057	.108	.156	.071	.016
Hamilton	.003	.028	.085	-.002	.150	.036	-.046	.047
Winnipeg	.022	.029	.166	.038	.129	-.029	.061	.073
Calgary	-.040	.039	.204	-.001	.049	.053	-.028	.067
Edmonton	.009	.011	.035	.031	.041	.017	.003	.029
Vancouver	.028	.034	.020	.047	.063	.006	.020	.077
Can SRU $\bar{I}'_1$	.055	.050	.102	.063	.173	.045	.065	.028
Summary Index $\bar{I}'_{\underline{M}}$	.036	.064	.142	.059	.107	.069	.045	.031

$\bar{I}'_1$ is the fractional reduction in the variance as a result of stratification at both levels of large cities and subunits within, averaged over the provinces.

$\bar{I}'_{\underline{M}}$ is the fractional reduction in the variance as a result of delineating subunits in the 10 metropolitan areas vs. stratification at no lower level than the large cities, averaged over the 10 metropolitan areas, perhaps comprising about 2/3 of the SRU population of Canada.

APPENDIX

DEVELOPMENT OF STRATIFICATION INDEXES

The stratification index is defined by

$$I = [V(X_{\bar{s}}) - V(X_s)]/V(X_{\bar{s}})$$

as in 2.3 and the modified stratification index, accounting for the population variance between strata alone is discussed later on in the appendix. We shall deal with the index I first.

In the case of multi-stage stratified design,

$$V(\hat{X}_s) = \sum_h V(\hat{X}_h)$$

$$= \sum_h \{N_h^2 \sigma_h^2/n_h \ [1+(n_h-1) \ r_{FP:h}] + \sum_{i=1}^{N_h} \sigma_{hi}^2/(n_h \ p_{i|h})\} \qquad (A.1)$$

while in the case of an unstratified design,

$$V(\hat{X}_{\bar{s}}) = N^2 \sigma^2/n.[1+(n-1) \ r_{FP}] + \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sigma_{hi}^2/(mp_h \ p_{i|h}) \qquad (A.2)$$

σ_h^2 is the population variance between PSU's of stratum h, weighted by sizes $p_{i|h}$, or

$$N_h^2 \sigma_h^2 = \sum_i p_{i|h} \ (X_{hi}/p_{i|h} - X_h)^2 \qquad (A,3)$$

$r_{FP:h}$ is the finite population correlation, which occurs when sampling without replacement in stratum h and would equal $-1/(N_h-1)$ if sampling

with equal probability were undertaken. σ^2 and r_{FP} are similar parameters applied to the whole area A when it is not delineated into several strata. More complete developments of the variance and definitions of the above symbols are given by Gray [3].

If sampling is undertaken with replacement, $r_{FP:h} = r_{FP} = 0$

To obtain $V(\hat{X}_{\bar{s}}) - V(\hat{X}_s)$, it is necessary to obtain σ^2 in terms of σ^2_{BS}, the population variance between strata and σ^2_h's.

Adopting the algebraic definition of $N^2_h\ \sigma^2_h$ to area A, we find that:

$$N^2\ \sigma^2 = \sum_h \sum_i p_h\ p_{i|h}\ (X_{hi}/p_h\ p_{i|h} - X)^2 \qquad (A.4)$$

and by employing the algebra in a similar manner as in Sukhatme [7], we find that:

$$N^2\ \sigma^2 = L^2\ \sigma^2_{BS} + \sum_{h=1}^{L} N^2_h\ \sigma^2_h/p_h, \text{ where} \qquad (A.5)$$

$$L^2\ \sigma^2_{BS} = \sum_{h=1}^{L} p_h(X_h/p_h - X)^2 \qquad (A.6)$$

Hence,

$$V(\hat{X}_{\bar{s}}) = \frac{1}{n} N^2\ \sigma^2\ [1 + (n-1)\ r_{FP}] + \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sigma^2_{hi}/(np_h\ p_{i|h})$$

$$= \frac{1}{n} L^2\ \sigma^2_{BS}\ [1 + (n-1)\ r_{FP}]$$

$$+ \frac{1}{n} \sum_{h=1}^{L} N^2_h\ \sigma^2_h/p_h\ [1 + (n-1)\ r_{FP}]$$

$$+ \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sigma^2_{hi}/(np_h\ p_{i|h}) \qquad (A.7)$$

$$= \frac{L^2 \sigma_{BS}^2}{n} \; [1 + (n-1) \; r_{FP}]$$

$$+ \sum_{h=1}^{L} \frac{N_h^2 \sigma_h^2}{n_h} \; \frac{n_h}{np_h} \; [1 + (n_h-1) \; r_{FP:h} + (n-1) \; r_{FP} - (n_h-1) \; r_{FP:h}]$$

$$+ \sum_{h=1}^{L} \sum_{i=1}^{N_h} \frac{\sigma_{hi}^2}{n_h \; p_{i|h}} \; \frac{n_h}{np_h}$$

$$= \frac{L^2 \sigma_{BS}^2}{n} \; [1 + (n-1) \; r_{FP}]$$

$$+ \sum_{h=1}^{L} \frac{n_h}{np_h} \quad V(\hat{X}_h)$$

$$+ \sum_{h=1}^{L} \frac{N_h^2 \sigma_h^2}{np_h} \; [(n-1) \; r_{FP} - (n_h-1) \; r_{FP:h}] \qquad \text{(A.8)}$$

so that $V(\hat{X}_{\bar{s}}) - V(\hat{X}_s)$ may be split up as follows, noting that

$$V(\hat{X}_s) = \sum_{h=1}^{L} V(\hat{X}_h)$$

$$T_1 = L^2 \; \sigma_{BS}^2/n \; [1 + (n-1) \; r_{FP}],$$

$$T_2 = \sum_{h=1}^{L} V(\hat{X}_h)(n_h/np_h-1),$$

and $$T_3 = \sum_{h=1}^{L} N_h^2 \; \sigma_h^2/(np_h) \cdot [(n-1) \; r_{FP} - (n_h-1) \; r_{FP:h}]. \qquad \text{(A.9)}$$

T_1 = effect due to the population variance between strata,

T_2 = effect due to the different size of strata and/or the different number of selected PSU's per stratum,

and T_3 = effect due to the different finite population corrections between strata.

One would usually expect the main contribution to $V(\hat{X}_{\bar{s}}) - V(\hat{X}_s)$ in (A.9) to be T_1 involving the population variance between strata. The other terms may be positive or negative among the strata. If sampling is undertaken with replacement with or without stratification, the difference simplifies to:

$$\{[V(\hat{X}_{\bar{s}}) - V(\hat{X}_s)] \,|\, r_{FP} = r_{FP:h} = 0\} = L^2\, \sigma^2_{BS}/n + \sum_{h=1}^{L} V(\hat{X}_h)(n_h/np_h - 1). \qquad (A.10)$$

To obtain an estimate of $V(\hat{X}_{\bar{s}}) - V(\hat{X}_s)$, one could obtain estimates $\hat{\sigma}^2_h$, $\hat{\sigma}^2_{hi}$, $\hat{r}_{FP:h}$ in the manner described by Gray [3] and substitute in (A.8). It remains to derive an estimate $\hat{\sigma}^2_{BS}$ and finally to derive an approximate estimate of the difference $V(\hat{X}_{\bar{s}}) - V(\hat{X}_s)$ under certain assumptions. An estimate of r_{FP} would be very difficult in most pps sample designs.

Consider the statistic $s^2 = \sum_{h=1}^{L} p_h(\hat{X}_h/p_h - \hat{X})^2$

$$Es^2 = E \sum_{h=1}^{L} \hat{X}^2_h/p_h - E\hat{X}^2$$

$$= \sum_{h=1}^{L} X^2_h/p_h + \sum_{h=1}^{L} V(\hat{X}_h)/p_h - X^2 - \sum_{h=1}^{L} V(\hat{X}_h)$$

$$= L^2\, \sigma^2_s + \sum_{h=1}^{L} (1/p_h - 1)\, \hat{V}(\hat{X}_h)$$

or an estimate of $L^2\, \sigma^2_s/n$ is given by:

$$L^2 \hat{\sigma}_s^2/n = [s^2 - \sum_{h=1}^{L} (1/p_h - 1) \; V(X_h)]/n \qquad (A.11)$$

To account for the population variance between strata alone, one may define a stratification index for area A by:

$$I'_A = L^2 \sigma_{BS}^2/n \div [V(\hat{X}_s) + L^2 \sigma_{BS}^2/n]$$

and an estimate of it, neglecting the ratio estimate bias, is given by

$$\hat{I}'_A = \frac{\frac{1}{n} [s^2 - \sum_{h=1}^{L} (\frac{1}{p_h} - 1) \; \hat{V}(\hat{X}_h)]}{\hat{V}(\hat{X}) + \frac{1}{n} [s^2 - \sum_{h=1}^{L} (\frac{1}{p_h} - 1) \; \hat{V}(\hat{X}_h)]} \qquad (A.12)$$

An estimate of $I_A = [V(\hat{X}_{\bar{s}}) - V(\hat{X}_s)]/V(\hat{X}_{\bar{s}})$, assuming $r_{FP} = r_{FP:h} = 0$, is readily available. For, with this assumption, referring to A.10 and A.11, we find that

$$\hat{V}(\hat{X}_{\bar{s}}) - \hat{V}(\hat{X}_s) = \frac{1}{n} [s^2 - \sum_{h=1}^{L} (\frac{1}{p_h} - 1) \; \hat{V}(\hat{X}_h) + \sum_{h=1}^{L} (\frac{n_h}{p_h} - n) \; \hat{V}(\hat{X}_h)]$$

so that

$$\hat{I}_A = \frac{\frac{1}{n} [s^2 - \sum_{h=1}^{L} (\frac{1}{p_h} - 1 + n - \frac{n_h}{p_h}) \; \hat{V}(\hat{X}_h)]}{\hat{V}(\hat{X}_s) + \frac{1}{n} [s^2 - \sum_{h=1}^{L} (\frac{1}{p_h} - 1 + n - \frac{n_h}{p_h}) \; \hat{V}(\hat{X}_h)]} \qquad (A.13)$$

When $r_{FP:h}$ is assumed to be 0, $\hat{V}(\hat{X}_h) = [n_h/(n_h - 1)] \sum_{i=1}^{n_h} (\hat{X}_{hi} - \hat{X}_h/n_h)^2$ (A.14)

and $\hat{V}(\hat{X}_s) = \sum_h \hat{V}(\hat{X}_h)$.

Now it can readily be shown that when sampling is done without replacement with pps so that $r_{FP:h} \neq 0$ and usually < 0,

$$E\hat{V}(\hat{X}_h) = V(\hat{X}_h) - N_h^2 \sigma_h^2 r_{FP:h}$$

Apart from the ratio estimate bias,

$$E\hat{I}'_A = \frac{\dfrac{L^2 \sigma_{BS}^2}{n} + [\dfrac{1}{n} \sum_{h=1}^{L} (\dfrac{1}{p_h} - 1) N_h^2 \sigma_h^2 r_{FP:h}]}{V(\hat{X}_s) + \dfrac{L^2 \sigma_{BS}^2}{n} + [\dfrac{1}{n} \sum_{h=1}^{L} (\dfrac{1}{p_h} - 1 - n) N_h^2 \sigma_h^2 r_{FP:h}]} \tag{A.15}$$

To obtain $E\hat{I}_A$, we shall write it as $\hat{T}_A/\hat{B}_A$, where $\hat{T}_A$ and $\hat{B}_A$ are the numberator and denominator, respectively of $\hat{I}_A$ as stated in (A.13).

$$\begin{aligned} E\hat{T}_A &= L^2 \sigma_{BS}^2/n + \sum_{h=1}^{L} (1/p_h - 1) N_h^2 \sigma_h^2 r_{FP:h}/n \\ &+ \sum_{h=1}^{L} [n_h/np_h - 1][V(\hat{X}_h) - N_h^2 \sigma_h^2 r_{FP:h}] \\ &= (T_1 + T_2) + [\sum_{h=1}^{L} (1/p_h - 1 + n - n_h/p_h) N_h^2 \sigma_h^2 r_{FP:h}/n \\ &- L^2 \sigma_{BS}^2 (n-1) r_{FP}/n] \end{aligned} \tag{A.16}$$

$$\begin{aligned} E\hat{B}_A &= L^2 \sigma_{BS}^2/n + \sum_{h=1}^{L} (1/p_h - 1) N_h^2 \sigma_h^2 r_{FP:h}/n \\ &+ \sum_{h=1}^{L} (n_h/np_h) \cdot [V(\hat{X}_h) - N_h^2 \sigma_h^2 r_{FP:h}] \end{aligned}$$

$$= V(\hat{X}_s) + L^2 \sigma_{BS}^2/n + \sum_{h=1}^{L} (n_h/np_h - 1) V(\hat{X}_h)$$

$$+ \sum_{h=1}^{L} (1/p_h - 1 - n_h/p_h) N_h^2 \sigma_h^2 r_{FP:h}/n$$

$$= V(\hat{X}_s) + T_1 + T_2 + [\sum_{h=1}^{L} (1/p_h - 1 - n_h/p_h) N_h^2 \sigma_h^2 r_{FP:h}/n$$

$$- L^2 \sigma_{BS}^2 (n-1) r_{FP}/n] \qquad (A.17)$$

Neglecting T_3 of (A.9) , we find that the biases in the estimation of the numerators and denominators of I_A' and I_A are given by the expressions in squared brackets on the right side in all cases.

In $\hat{I}_A'$, the bias in the denominator is most likely negative since $1/p_h-1$ is positive and $r_{FP:h}$ is most likely negative. The bias in the denominator; however, is most likely positive since($1/p_h-1-n$)and $r_{FP:h}$ are both most likely negative; ($1/p_h-1-n$) is approximately $-(1+L)$ when $n=2L$. Consequently, $\hat{I}_A'$ under-estimate I_A'.

In $\hat{I}_A$, the bias in the numerator $\hat{T}_A$ is most likely negative but with lower absolute value than the numerator of $\hat{I}_A'$ since $-L^2 \sigma_{BS}^2/n.(n-1) r_{FP}$ is most likely positive. However, the bias in the denominator, $\hat{B}_A$, is most likely positive since($1/p_h- 1 - n_h/p_h$)and $r_{FP:h}$ are most likely both negative and $-r_{FP}$ is positive. Consequently, $\hat{I}_A$ under-estimates I_A though probably not to the same extent that $\hat{I}_A'$ under-estimates I_A'.

ROTATION GROUP BIASES IN THE OLD AND NEW LABOUR FORCE SURVEY

R. Tessier
Household Surveys Development Division

This paper presents results on rotation group biases in the Canadian Labour Force Survey (LFS). The biases are studied in detail by decomposition into components responsible for the biases. Also, a comparison between the old and the new LFS is done on the basis of 1975 parallel run and differences are analyzed. Some conclusions are drawn and recommendations for other studies presented.

1. INTRODUCTION

In large scale periodic surveys, such as the Canadian Labour Force Survey, repetitive interviewing of the same respondents is a common practice. It has the advantage of reducing cost and improving the precision of the estimates of month to month changes. On the other hand, it is well known, Barbara Bailar, and Williams and Mallows, ([1] and [4]) that repetitive interviewing of the same respondents affect the estimates due to the introduction of conditioning effects and possible systematic changes in response probabilities.

The Labour Force Survey (LFS) sample is composed of six rotation groups out of which one is replaced each month; therefore, respondents are exposed to the survey for six consecutive months. By design, the expected sample size is the same for all rotation groups. This feature of the LFS sample was kept from the old survey, which was operational until December 1975.

In this paper, analysis of rotation group biases is done, first by comparing biases between the old and the new survey using the parallel run of one year conducted in 1975. In a second step, an attempt is made to decompose the biases into components indicating that part of the biases due to the estimation procedure itself and that part due to the respondents. Finally,

estimates are adjusted in order to eliminate rotation group biases due to the estimation procedure and respondent biases, as a whole, are isolated. No breakdown, though, of respondent biases have been attempted in order, for example, to isolate that part of the biases due to conditioning effects on that part due to changes in response probabilities.

2. THE OLD AND THE NEW SURVEY

A thorough redesign of the LFS was undertaken a few years ago where all aspects of the methodology, reporting procedure, data processing, etc. were looked into and updated. A feature, though, that was left unchanged from the old survey is the rotation plan; that is, the sample is composed of six rotation groups of the same expected size of which one is replaced each month. Before the publication of data from the revised survey started, it was judged necessary to conduct both the old and the new survey at the same time for a period of one year on two independent samples. This parallel run took place in 1975 and the data from the two surveys are used here to compare rotation group biases.

The difference between the expected value of a characteristic based on a particular rotation group (respondents being interviewed for a given number of times) and the expected value based on all rotation groups is called the rotation group bias. Table 1, rotation group biases are emphasized by showing separate estimates based on the respondent's number of exposure to the survey. The estimates are presented relative to the average estimate using all respondents, multiplied by 100 (see appendix 1). Therefore, a relative estimate of 100.0 means that the estimate for that group is the same as the average estimate using all groups, a relative estimate of 95.0 means that the estimate for that group is 5.0% lower than the average estimate and a relative estimate of 105.0 means that the estimate is 5.0% larger than the average estimate. In order to safeguard against possible seasonal patterns in the biases, estimates used are averages over the year. Also, comparison is made between the old Labour Force Survey (LFS) and the new Labour Force Survey (RLFS). Note, though, that the target population for the LFS is the civilian population 14+ years old, while for the RLFS it is the civilian population 15+ years old. Further,

though it is intended to use year averages as estimates, we have only ten months averages for RLFS estimates since data for July and September were not made available. Average number of responding persons is approximatively 75,600 in the LFS and 56,300 in the RLFS.

Variance estimates for the relative estimates are not available but it was found that a conservative upper bound for the variances could be given by the coefficient of variation (C.V.) of the monthly estimates. Therefore, noticing that for the LFS, monthly C.V.s for the characteristics employed, unemployed and not in LF (both sexes) are less than 0.4%, 2.7% and 0.5% respectively we have that relative estimates of the three characteristics are at more than two standard deviation (s.d.) from 100.0 if the estimate is lower than 99.2 or higher than 100.8 for the characteristic employed, if the estimate is lower than 94.6 or higher than 105.4 for unemployed and lower than 99.0 or higher than 101.0 for Not in LF (sex breakdowns would have different bounds).

We find from Table 1 that the difference between the relative estimates and 100.0 in absolute value is in most cases larger for the RLFS than for the LFS. The first month interview bias is much stronger in the RLFS while differences between the two survey estimates are much less important for the other interviews. If we look at the characteristics employed and not in LF (both sexes) using the above mentioned bounds, we find that the first interview provides estimates that are more than two s.d. away from 100.0 for both surveys while other interviews yield estimates slightly larger than two s.d. from 100.0 in only some cases. This indicates a strong first month effect that is gradually adjusted in the five other interviews. As for the characteristic unemployed both sexes, all estimates for both surveys are within two s.d. from 100.0.

On the other hand, we must notice that the characteristic total population is also strongly affected by rotation group biases in both surveys. This indicates that the biases are not only attributable to the behaviour of the respondent, as is generally thought of, but need to be explained by some other phenomena. This will be done in section 3.

Table 1: Relative Estimates by Rotation Group for Selected Labour Force Characteristics

Comparison Old (LFS) and New (RLFS) Survey 1975 Data

1975 Data

Characteristic	Survey	First Interv.	Second Interv.	Third Interv.	Fourth Interv.	Fifth Interv.	Sixth Interv.
Total Population	LFS	98.1	100.3	100.8	100.7	100.3	99.8
	RLFS	96.2	100.9	101.0	101.0	100.7	100.1
Employed Male	LFS	97.6	99.8	100.7	101.0	100.4	100.5
	RLFS	95.5	101.0	100.7	101.0	101.3	100.6
Female	LFS	98.8	100.2	101.2	100.9	100.2	98.8
	RLFS	94.3	101.0	101.1	101.6	101.1	101.0
Both Sexes	LFS	98.0	100.0	100.9	101.0	100.3	99.9
	RLFS	95.1	101.0	100.9	101.2	101.2	100.7
Unemployed Male	LFS	101.2	101.9	99.8	99.5	99.9	97.7
	RLFS	99.5	102.1	102.9	100.1	98.6	96.8
Female	LFS	103.4	100.0	98.5	99.1	100.2	96.8
	RLFS	105.5	99.9	102.7	102.6	96.5	92.9
Both Sexes	LFS	101.9	101.3	99.4	99.4	100.6	97.4
	RLFS	102.1	101.2	102.8	101.1	97.7	95.1
In Labour Force	LFS	98.3	100.1	100.8	100.8	100.3	99.7
	RLFS	95.6	101.0	101.0	101.2	101.0	100.3
Not in Labour Force	LFS	97.9	100.7	100.8	100.5	100.2	100.0
	RLFS	97.3	100.8	101.1	100.6	100.4	99.9
Unemployment Rate	LFS	103.5	101.1	98.7	98.6	100.3	97.7
	RLFS	106.7	100.1	101.7	99.9	96.7	94.7
Participation Rate	LFS	100.2	99.7	100.0	100.1	100.1	100.1
	RLFS	99.3	100.1	100.0	100.2	100.2	100.2

Finally, we can see from Table 1 that among the characteristics studied, it is the estimation of the unemployment rate that is mostly affected by the rotation plan and that of the participation rate that is less affected.

3. COMPONENTS OF THE BIAS

As mentioned in section 2, the characteristic "total population" is also affected by the rotation plan, which indicates that explanation of the biases on this characteristic will most probably not be given by the respondent's behaviour to the survey. Looking into the Labour Force Survey estimation procedure, we find that the biases can be attributable to the two following possible causes:

a) non-response rates are different from one rotation group to another although the adjustment factor for non-response is the same for all rotation groups;

b) coverage rates are different from one rotation group to another although age-sex correction factors are the same for all rotation groups.

In order to isolate each of the two possible causes, use was made of non-response rates by rotation group available for the twelve months of 1973. Monthly estimates were adjusted by a factor of the form

$$W^{(r)} = \frac{\text{expected number of persons in rotation group } r}{\text{number of persons interviewed in rotation group } r}$$

to replace the current non-response correction factor which is of the form

$$W = \frac{\text{expected number of households in all rotation groups}}{\text{number of households interviewed in all rotation groups}}$$

Notice that in $W^{(r)}$ use is made of persons rather than households; this was done by adjusting the weight by the ratio of overall coverage household size to average responding household size by rotation group made available in [2]. It was judged more accurate to adjust for non-response at the

individual level rather than the household level since it was found from [2] that average non-responding household sizes are different from one rotation group to another and also they are different from that of responding households.

Table 2 gives relative estimates for the characteristic total population using three different estimation procedures: T_1 is an estimator where adjustment is made for differences in coverage rates from one rotation group to another (that is, it contains only rotation group biases due to differences in non-response rates from one rotation group to another). T_2 is an estimator where adjustment is made for differences in non-response rates, as mentioned above (that is, it contains only rotation group biases due to differences in coverage rates). Finally, T_3 includes no adjustment and, therefore, contains both sources of rotation group biases for this characteristic. Note that T_1 has been obtained by derivation from T_2 and T_3 (see appendix 1)).

Table 2: Relative Estimates by Rotation Group for the Characteristic "Total Population" by Type of Estimator - 1973 Data

Estimator	First Interv.	Second Interv.	Third Interv.	Fourth Interv.	Fifth Interv.	Sixth Interv.
T_1	99.4	100.4	100.4	100.1	99.7	99.9
T_2	99.7	100.5	100.2	100.1	99.8	99.7
T_3	99.1	100.9	100.7	100.2	99.6	99.6

We see from Table 2 that the total bias (in T_3) is approximately the sum of the bias in T_1 and that in T_2 (the other term in the formula has a negligible contribution). For example, using respondents being interviewed for the first time, we find that T_3 has a bias of -0.9 (=99.1 - 100.0)

which is the sum of the bias in T_1, -0.6 (=99.4 - 100.0), and that in T_2, -0.3 (=99.7 - 100.0). Further, we see that the two components have similar trends, that is, the first, fifth and sixth interviews yield a negative bias while the second, third and fourth interviews yield positive biases with a maximum peak near the second interview. Also, magnitude of the two components of the biases is of the same order. Appendix 2 provides a graph of the biases permitting easy visualization of the trends.

4. ADJUSTED ROTATION GROUP BIASES

We have seen in section 3 that the biases in the estimation of the characteristic total population are due to the estimation procedure itself; therefore, all labour force characteristics are affected by the same biases. In order to isolate the component of the biases, in the estimation of labour force characteristics, that is due to the fact that a respondent is classified differently depending on the number of times he is exposed to the survey, estimates were adjusted to eliminate components of biases due to different non-response and coverage rates from one rotation group to another. Table 3 provides relative estimates using adjusted and unadjusted estimates based on year averages of 1973 data.

We find from Table 3 that the component of rotation group biases for the characteristics employed (male, female, both sexes), in labour force and not in labour force due to the estimation procedure are at least as important as the component due to response error. Further, we find that for the characteristic unemployed, the unadjusted estimates tend to minimize the importance of the first interview bias due to response error. As for the rates, we have that the adjusted and the unadjusted estimates provide the same biases which is due to the particular nature of the adjustment procedure (see appendix 1). Finally, we may mention that the participation rate is almost unaffected by the rotation plan.

Table 3: Relative Estimates by Rotation Group
for Selected Labour Force Characteristics
Adjusted and Unadjusted Estimates

1973 Data

Characteristic	Estimator	First Interv.	Second Interv.	Third Interv.	Fourth Interv.	Fifth Interv.	Sixth Interv.
Total Population	Unadj.	99.1	100.9	100.7	100.2	99.6	99.6
	Adj.	100.0	100.0	100.0	100.0	100.0	100.0
Employed Male	Unadj.	98.9	101.0	100.5	100.2	99.7	99.7
	Adj.	99.4	99.8	99.8	100.1	100.4	100.5
Female	Unadj.	98.7	100.8	101.1	100.6	99.9	98.9
	Adj.	99.9	100.2	100.5	100.3	100.0	99.1
Both Sexes	Unadj.	98.8	100.9	100.7	100.3	99.8	99.5
	Adj.	99.7	100.1	100.0	100.1	100.2	99.9
Unemployed Male	Unadj.	101.7	101.1	101.0	101.6	97.4	97.2
	Adj.	102.2	99.9	99.7	101.5	98.1	97.9
Female	Unadj.	106.3	98.6	100.4	98.3	100.2	96.3
	Adj.	107.6	97.9	99.7	98.0	100.4	96.4
Both Sexes	Unadj.	103.1	100.3	100.8	100.6	98.3	96.9
	Adj.	104.0	99.4	100.2	100.4	98.7	97.3
In Labour Force	Unadj.	99.1	100.9	100.7	100.3	99.7	99.3
	Adj.	99.9	100.0	100.0	100.1	100.1	99.8
Not in Labour Force	Unadj.	99.2	100.9	100.6	100.0	99.4	99.9
	Adj.	100.1	100.0	100.0	99.8	99.8	100.3
Unemployment Rate	-	104.1	99.4	100.1	100.2	98.6	97.6
Participation Rate	-	99.9	100.0	100.0	100.1	100.1	99.8

Since we find that the estimation procedure has an effect on the rotation group biases, it is of interest to compare the LFS and the RLFS on the basis of adjusted estimates. Though it is impossible at the present time to produce breakdowns of biases as presented in Table 2 for 1975 LFS data (some necessary data is not available), it is nevertheless possible to eliminate the overall contribution of biases due to the estimation procedures from the two series of data. Table 4 presents the adjusted data for both surveys which permits comparison of rotation group biases due to response errors in both surveys. Note that the estimation procedures are slightly different for both surveys; for example, adjustment for non-response in the LFS is done by means of a weight applied to the current month data while in the RLFS, if some special conditions are satisified, preceding month data are imputed in the current month. This difference in non-response adjustment affects differently the data due to the fact that in the LFS non-responding households size is assumed to be the same as that of responding households (see [2]) while in the RLFS non-responding household sizes are exact, unless composition has changed for households where imputation is done.

Table 4 reveals that, except for the characteristic employed female, relative estimates of both surveys follow the same trend over the six interviews. Also, magnitude of biases on RLFS data is at least as large as that on LFS data except for the characteristics unemployed female and both sexes where RLFS data is more subjected to rotation group biases. Finally, if we compare Table 4 with Table 1, we find that conclusions drawn from Table 3 still hold, but more important, we find that first interview bias is stronger on RLFS data for unadjusted data while it is of the same magnitude for both surveys for adjusted data (except for unemployed female and both sexes). This permits to conclude that the RLFS estimation procedure yield stronger rotation group biases than the LFS estimation procedure while rotation group biases due to response errors are similar for both surveys though the two questionnaires are quite different.

Table 4: Relative Estimates by Rotation Group for Selected Labour Force Characteristics

Comparison Old (LFS) and New (RLFS) Survey Adjusted 1975 Data

Characteristic	Survey	First Interv.	Second Interv.	Third Interv.	Fourth Interv.	Fifth Interv.	Sixth Interv.
Employed Male	LFS	99.7	99.5	99.9	100.2	100.2	100.6
	RLFS	99.3	100.0	100.0	100.0	100.5	100.3
Female	LFS	100.5	99.9	100.4	100.3	99.9	99.0
	RLFS	98.0	100.1	99.8	100.7	100.4	100.8
Both Sexes	LFS	99.9	99.7	100.1	100.2	100.1	100.1
	RLFS	98.8	100.1	99.8	100.2	100.5	100.5
Unemployed Male	LFS	103.3	101.6	99.1	98.8	99.8	97.9
	RLFS	103.4	101.2	102.0	99.0	97.8	96.6
Female	LFS	105.4	99.6	97.7	98.5	101.9	96.9
	RLFS	109.5	99.2	101.4	101.7	96.1	93.0
Both Sexes	LFS	103.9	101.0	98.8	98.8	100.5	97.8
	RLFS	106.1	100.2	101.6	100.2	97.0	95.1
In Labour Force	LFS	100.2	99.8	100.0	100.1	100.1	99.9
	RLFS	99.3	100.1	100.0	100.2	100.2	100.2
Not in Labour Force	LFS	99.7	100.4	100.1	99.8	99.9	100.2
	RLFS	101.1	99.9	100.1	99.6	99.7	99.8

5. CONCLUSION

From the present study we may stress the following points:

a) Rotation group biases in the estimates can be attributed to three sources: difference in non-response rates from one rotation group to another, difference in coverage rates and difference in response errors.

b) The component of bias due to the difference in non-response and coverage rates taken together is at least as important as that of response errors.

c) The characteristic mostly affected by rotation group biases is the number of unemployed and the unemployment rate, with extreme values in the first and sixth interviews.

d) Comparison of the LFS and RLFS data using the 1975 parallel run indicates that the two sets of data yield rotation group biases is slightly larger on RLFS data in some instances. Except for the characteristic unemployed, it is mainly the first interview biases that are larger in the RLFS than the LFS.

e) Decomposition of biases into components permits to say that differences in rotation group biases between the two surveys seems to be due to a difference in estimation procedures since the component of biases due to response error are of the same order of magnitude, in spite of the fact that the two questionnaires are quite different.

Note that the conclusion drawn from the 1975 data must be taken with caution since RLFS was producing its first year data and therefore may not be perfectly stabilized. Further, RLFS data used in this study are averages over ten months only since July and September data were not made available.

6. ACKNOWLEDGEMENT

The author would like to express his sincere thanks to R. Platek and other professionals of Household Surveys Development Staff for very useful discussions. Also, thanks are expressed to members of the Research Section of the Labour Force Survey Division for providing most of the data.

RESUME

Cet article présente l'importance relative des biais dûs aux groupes de rotation dans l'enquête canadienne sur la population active. Une étude détaillée est faite en décomposant les biais en ses différentes composantes. De plus, une étude comparative de l'ancienne et la nouvelle enquête est faite en utilisant les données de 1975 produites simultanément pour les deux enquêtes. Certaines conclusions sont tirées et des études plus élaborées sont récommendées.

REFERENCES

[1] Bailar, Barbara, A. (1975), "The Effects of Rotation Group Bias On Estimates from Panel Surveys", JASA, Vol. 70, pp. 23-30.

[2] Lawes, M. (1975), "A Note On Average Household Sizes For Respondent And Non-Respondent Households As Calculated From The Six Month Data File", Household Surveys Development Staff, memorandum, Statistics Canada.

[3] Tessier, R. and Tremblay, V. (1976), "A Study Of Rotation Group Biases In The LFS", Technical Memorandum, Household Surveys Development Staff, Statistics Canada.

[4] Williams, W.H. and Mallows, C.L. (1970), "Systematic Biases In Panel Surveys Due To Differential Non-response", JASA, Vol. 65, pp. 1338-49.

APPENDIX 1

Let $X^{(r)}$ be the expected number of persons having characteristic x over all possible samples for respondents being interviewed for the rth time and let X be the correct corresponding value, then $X^{(r)} - X$ is the bias of the rth rotation group for estimating X. We may then write

$$X^{(r)} = X(1 + B_x^{(r)}), \tag{A.1}$$

where $B_x^{(r)}$ is the relative bias in the estimation of X when using the rth rotation group. A suitable indicator of the relative magnitude of the bias is therefore given by the relative value of the characteristic x for respondents being interviewed for the rth time multiplied by 100, Rel $X^{(r)}$ say, which is

$$\text{Rel } X^{(r)} = 100\ X^{(r)}/X = 100\ (1 + B_x^{(r)}). \tag{A.2}$$

Let $\hat{X}^{(r)}$ be an unbiased estimate of $X^{(r)}$, that is, $\hat{X}^{(r)}$ is an estimate of characteristic x using data from only those respondents being interviewed for the rth time and let

$$\hat{X} = \sum_{r=1}^{6} \hat{X}^{(r)}/6.$$

If we can assume that $\hat{X}$ is an unbiased estimate of X, that is, the average of estimates over all rotation groups is unbiased for estimating X, then the relative estimate is

$$\text{Rel } \hat{X}^{(r)} = 100\ \hat{X}^{(r)}/\hat{X}. \tag{A.3}$$

For the characteristic total population, T say, we have

$$\mathrm{Rel}\ T^{(r)} = 100\ T^{(r)}/T \qquad (A.4)$$

and

$$\mathrm{Rel}\ \hat{T}^{(r)} = 100\ \hat{T}^{(r)}/T. \qquad (A.5)$$

Note that in (5), the denominator is the true value rather than the estimate since it is the census projection, which we assume to be exact. The adjusted value, $X_a^{(r)}$ say, is

$$X_a^{(r)} = X^{(r)}\ T/T^{(r)}, \qquad (A.6)$$

with estimate

$$\hat{X}_a^{(r)} = \hat{X}^{(r)}\ T/\hat{T}^{(r)}. \qquad (A.7)$$

Therefore, the adjusted relative value multiplied by 100, Rel $X_a^{(r)}$ say, is

$$\mathrm{Rel}\ X_a^{(r)} = 100\ X_a^{(r)}/X \qquad (A.8)$$

with estimate

$$\mathrm{Rel}\ \hat{X}_a^{(r)} = 100\ \hat{X}_a^{(r)}/\hat{X}. \qquad (A.9)$$

Note that the above adjustment has been done separately for male, female and both sexes on the previous data since census projections are available at that level.

To decompose the relative estimate of the characteristic total population into its two components, one containing only bias due to difference in coverage rates from one rotation group to another and the other containing

only bias due to difference in non-response rates, we make use of the following adjustment factor

$$K^{(r)} = W^{(r)}/W \qquad (A.10)$$

where $W^{(r)} = \dfrac{\text{Expected number of persons in rotation group r}}{\text{Number of persons interviewed in rotation group r}}$

and $W = \dfrac{\text{Expected number of households in all rotation groups}}{\text{Number of households interviewed in all rotation groups}}$

By using the above correction factor in the monthly estimates before averaging over the year, we obtain an estimate, $*\hat{T}^{(r)}$ say, which contains only coverage biases, with relative estimates given by

$$\text{Rel } *\hat{T}^{(r)} = 100 *\hat{T}^{(r)}/T \qquad (A.11)$$

and by combining (5) and (11) we can demonstrate that the relative estimate containing only non-response bias, Rel $'T^{(r)}$ say, is given by

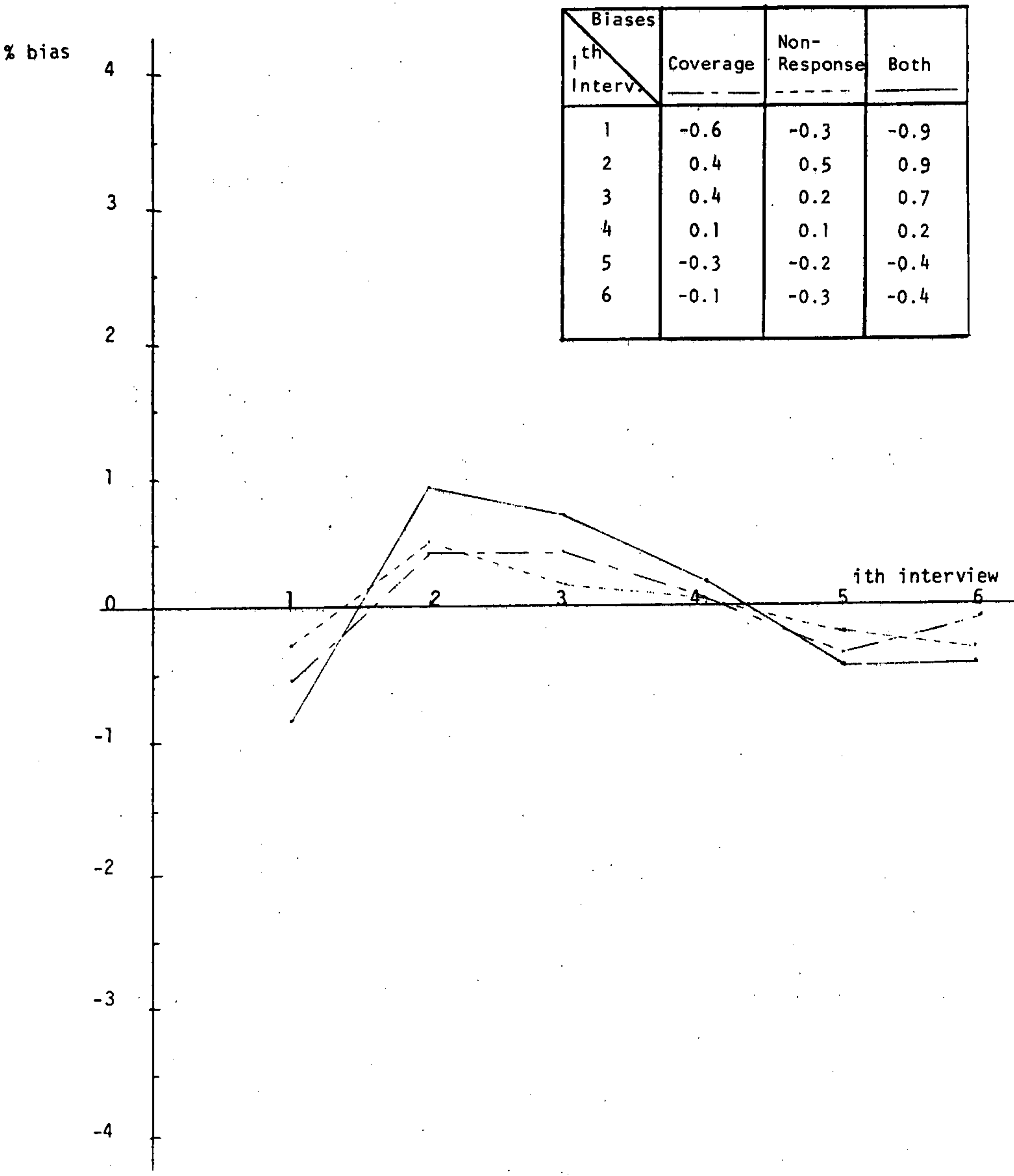
$$\text{Rel } '\hat{T}^{(r)} = 100\ \hat{T}^{(r)}/*\hat{T}^{(r)}.$$

Note that the details concerning the above formulae can be found in [3]. Further, [3] presents graphs similar to that in Appendix 2 for all the characteristics presented in this paper and for both the adjusted and the unadjusted estimates.

APPENDIX 2

Rotation Group Biases on Total Population With Components

1973 Data



Biases i^{th} Interv.	Coverage — — —	Non-Response - - - - -	Both ———
1	-0.6	-0.3	-0.9
2	0.4	0.5	0.9
3	0.4	0.2	0.7
4	0.1	0.1	0.2
5	-0.3	-0.2	-0.4
6	-0.1	-0.3	-0.4

LIST OF REFEREES/LISTE DES CRITIQUES

The Editorial Board wish to thank the following persons who have served as referees during the past year.

Le comité de rédaction désire remercier les personnes suivantes, qui ont bien voulu faire la critique des articles présentés au cours de l'année dernière.

G. Brackstone
M. Cairns
G. Chauvin
N. Chinnappa
D. Dodds
P. Ghangurde
G.B. Gray
M. Hidiroglou
G. Hole
M. Lawes
M. Rahman
A. Satin
K. Shrinath
M.P. Singh
P.F. Timmons
V. Tremblay

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

June/juin 1976 Vol. 2 No. 1

A Journal produced by Statistical Services Field, Statistics Canada.

Publié par le secteur des services statistiques, Statistique Canada.

CONTENTS

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

December/decembre 1976 Vol. 2 No. 2

A Journal produced by Statistical Services Field, Statistics Canada.

Publiê par le secteur des services statistiques, Statistique Canada.

CONTENTS