Statistics Canada Statistique Canada

SURVEY METHODOLOGY

JUNE 1975

VOLUME 1 — NUMBER 1

A JOURNAL OF
HOUSEHOLD SURVEYS DEVELOPMENT DIVISION
STATISTICAL SERVICES FIELD
STATISTICS CANADA

The first issue of the SURVEY METHODOLOGY JOURNAL is sent to you with the compliments of the Editorial Board. We hope you will find it informative and useful. We plan to bring out two issues a year. We will continuously endeavor to improve the Journal both with regard to the standard and scope of the articles and also its format and printing. Your comments and suggestions in this regard are very welcome.

SURVEY METHODOLOGY

June 1975 Volume 1 Number 1

A Journal of Household Surveys Development Division, Statistical Services Field, Statistics 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.

Submission of Papers:

The Journal will be issued twice a year. Authors are invited to submit their papers to the Editor Dr. M.P. Singh, Household Surveys Development Division, Statistics Canada, 10th Floor, Coats Building, Tunney's Pasture, Ottawa, Ontario - K1A 0T6. The papers should be typed double spaced (2 copies). Authors of articles for this journal are free to have their articles published in other statistical journals.

SURVEY METHODOLOGY

June 1975 Volume 1 Number 1

A Journal of the Household Surveys Development Division, Statistical Services Field, Statistics Canada.

CONTENTS

REINTERVIEW PROGRAMS AND RESPONSE ERRORS

R. Platek and P.F. Timmons

This paper discusses several reinterview techniques and their use in relation to Response Variance, Response Bias, Interviewer Training, and the monitoring of various elements of the interview process. Using the Canadian Labour Force Survey as a case study the article describes how reinterview techniques were developed as the survey evolved and briefly describes the strategy being followed in the present reinterview program.

1. INTRODUCTION

Estimates derived from a sample survey are subject to two main types of error - sampling error and non-sampling error. Sampling error has been the subject of intensive study resulting in the development of a considerable body of applicable theory on sample design, allocation, estimation, variance estimation, etc. Non-sampling error is less susceptible to theoretical development but is of extreme importance to the practitioner. Non-sampling errors can cause biases in the estimates and particularly in the case of large samples should often be of greater concern than the sampling error. The sampling error of course decreases with the increase in sample size, whereas the non-sampling error is apt to increase due to the difficulty of controlling a larger operation.

The study of non-sampling errors is particularly important in a large continuing survey such as the Canadian Labour Force Survey (LFS). In a continuing survey, maintaining control over non-sampling errors is necessary to provide comparability of the estimates over time. Also, a continuing survey presents much greater scope for study and control of non-sampling errors than one time surveys, because of the opportunity to compare performance from one survey period to successive ones over a period of time.

There are many sources of non-sampling errors, such as in data entry or processing, however, this paper will consider only response errors arising out of the interview process, and more specifically those measured by a reinterview technique.

2. THE REINTERVIEW TECHNIQUE

A reinterview can be conducted in a variety of ways. Some of the factors involved in defining a particular reinterview technique are: the person conducting the reinterview, the time interval between the two interviews, the instrument (questionnaire) used in reinterview, the reinterview procedures (ask questions in same way as interviewer or additional probing), independent or dependent reinterview.

The choice of a particular technique depends on the purpose of the reinterview program. Several purposes are possible; monitoring the interview process to avoid distortion of the time series due to deterioration of response accuracy, monitoring the interviewers work for use in retraining, monitoring the interview process to detect problems in questionnaire design, interviewing techniques, etc., estimating response variance, estimating response bias. Of these, the last two, response variance and response bias require the most rigorous techniques. The remainder of the stated objectives usually involve more flexible constraints and can be, to a large extent, accommodated under techniques designed for the variance and bias measures.

2.1 Response Variance

This requires replication, i.e. the reinterview should be carried out under the same conditions as the original. This is, of course, impossible in the strict sense, as some conditioning effect must be assumed as a result of the first interview. It is also difficult to insist that the same person answer as did the first time. The reinterviewer is usually not selected from the same population of interviewers or randomly assigned.

There is an additional memory span for the respondent the extent of which depends on the time of the reinterview. However, the same questionnaire can be used and if it is a highly structured questionnaire such as the Revised Labour Force Survey (RLFS), the same interview technique can be duplicated. By making special efforts, the problems can be reduced, e.g., the memory span can be minimized by careful planning so as not to leave a long period of time between interview and reinterview.

2.2 Response Bias

The reinterviewer strives to discover the "true" response which then serves to measure the "bias" in the original interview. Perfect execution is again not possible, however, there are means of obtaining what should be close to the proper response. Serious consideration must be given to what a "true" response consists of. In a continuing survey, a "true" response could be defined as the response obtained when a knowledgeable respondent is interviewed by a well trained interviewer using a standard questionnaire and interviewing technique. To establish this "true" response, it is possible to have a skilled interviewer reinterview the respondent using the same questionnaire and technique. Subsequently, the reinterviewer can match the original and reinterview questionnaires on the spot, and with the respondent's cooperation, reconcile the differences, if any. This procedure should establish a close approximation to the "true" response.

In other circumstances a "true" response might be defined as one which would have been obtained by using a more refined technique (which might be too expensive to use for the whole survey). Another definition of "true" response is in comparison with other sources such as administrative data.

2.3 Monitoring the Interview Process

Reinterview can be used in a continuing survey as a quality measure to ensure that the quality of response does not change greatly over time. It can also be used to "feed back" information useful in training interviewers, designing questions, improving techniques, writing training materials, etc. The procedure described above in "Response Bias" is of direct use in training as the reinterviewer can contact the interviewer and discuss any errors discovered in the interviewer's work.

2.4 Overall Reinterview Strategy

Reinterview strategy may be influenced by many factors. If interviewer selection, training and supervision have not been carefully controlled, a straight "policing" action may be the prime motivation. If little is known about response problems on a particular survey, the emphasis could be on research and evaluation. In any event, continuing evaluation is a

consideration and in some cases where the survey is reasonably under control, response variation and response bias may be the most interesting factors.

3. THE LABOUR FORCE SURVEY AS A CASE STUDY

The reinterview program for the LFS has been in operation for many years and has been used in most of the ways mentioned above. Over a period of years there have been five distinct phases:

3.1 Phase 1

Early reinterview program concerned mostly with policing of interviewers and operational problems. This was in a period when the survey was in its infancy and the organization and facilities for hiring and control of interviewers was relatively weak. Little documentation remains from this phase, although significant insights into the needs for supervising interviewers were gained.

3.2 Phase 2

Interviewing was more controlled at this period. Research into the interviewing process, the questionnaire, the technique, and anything affecting them was the main objective. Direct feed-back to the interviewers was not attempted, although some residual effect of policing and increased experience for supervisory personnel was acquired. In the LFS, selected households are retained in the survey for six months, then rotated out and replaced. The LFS sample thus consists of six sub-samples or rotation groups, each rotating in a different month. The sampling scheme for reinterview was a sub-sample of each rotation group. Some clustering of reinterview households was utilized to reduce interviewers' travel. The reinterviews were independent of the original interview, i.e. the reinterviewer did not have access to the original completed questionnaires, however, the procedures and questionnaires used in the reinterview were identical to the original ones. The reinterviewers were full-time supervisors from the Regional Offices. The comparison of the matched data for each individual shows the amount of disagreement in what should, theoretically, have been comparable results.

To measure this amount of disagreement for a specific Labour Force category say, "with job at work" which is denoted as "W", the interviewer and the reinterviewer may agree in classifying a person as a worker or a non-worker: But also they may disagree: the interviewer can classify a person as a worker, but the reinterviewer can have the same person as non-worker.

The results of this classification can be presented simply in the following way.

Original Interview

		W	Not W	Total
Reinterview	W	a	b	a + b
	Not W	c	d	c + d
	Total	a + c	b + d	a + b + c + d

where a and d denote agreement between the interviewer and reinterviewer, and the percentage $\frac{b + c}{1/2\,[(a+b) + (a+c)]}$ can be interpreted as a measure of disagreement in response for the "W" category. This measure of disagreement is shown below for an approximate 2-year period ending in 1960.

	Measure of Disagreement	%
W:	persons with job at work	9.0
L:	seeking work	45.5
J:	with job but not at work	72.3
H: S:	keeping house going to school	10.7
U: R: O:	permanently unable to work retired or voluntarily idle other	32.4

It is interesting, though not too surprising, that characteristics indicating a marginal attachment to the Labour Force, e.g., "with a job but not at work"

are a major source of difficulty in classification.

Some examples of the more detailed information that can be derived from this type of reinterview program follow.

Paired Categories: In order to determine which of the Labour Force categories tend to be most easily confused with one another, the number of discrepancies for pairs of labour force categories have been tabulated. Any discrepancy necessarily involves a pair of labour force categories, e.g., the interviewer classifies a respondent as "L" and the reinterviewer classifies the same respondent as "Wp", thus the pair of categories "L" and "Wp" are involved.

For a particular pair of labour force characteristics, Wf[*] and Wp[**], the results can be classified as follows:

Interviewer

L.F. Category	Wf	Wp	Other
Wf	a	b	c
Wp	d	e	f
Other	g	h	i

The number of disagreements within a pair of categories = b + d. The number of disagreements within a pair of categories as a percentage of the number of respondents found by both interviewer and reinterviewer to belong to one or other of the categories in the pair = $\frac{b + d}{a + b + d + e}$ X 100. If the Labour Force categories Wf, Wp, L, J, URO, and HS are considered, there are 15 different pairs possible and these are shown in Table 2.

[*] Wf refers to persons working 35 hours or more during reference week.

[**] Wp refers to persons working, but less than 35 hours during reference week.

Table 1: Disagreements between interviewer and reinterviewer for pairs of Labour Force Categories.

L.F. Categories (1)	Number of Disagreements (2)	% Disagreement within a pair of L.F. Categories (3)
Wf-Wp	736	7.8
Wf-L	86	1.0
Wf-J	146	1.8
Wf-URO	58	.6
Wf-HS	153	1.0
Wp-L	50	2.7
Wp-J	48	3.2
Wp-URO	53	2.1
Wp-HS	300	3.3
L-J	59	6.2
L-URO	81	4.0
L-HS	84	1.0
J-URO	33	2.0
J-HS	27	.3
URO-HS	290	3.1

Two factors are involved in the number of disagreements between two categories:

a) the degree of difficulty in distinguishing between the two categories;

b) the total number of cases which fall into the two categories.

The above data indicated that hours worked were not very well reported as shown by the high rate of confusion between full and part-time work. Even more serious is the difficulty in discriminating between part-time work (Wp), having a job, but not at work (J), and looking for work (L). In any attempt to improve the accuracy of the data through better definitions, questionnaires, or interviewer training, these particular categories require special attention.

Repetition: Another interesting aspect of this study is the effect of repetition of interviews - households remain in the Labour Force Survey sample for six consecutive months. Each month, 1/6 of the households are rotated out of the sample and replaced by a new group of households. As a result, each month the Labour Force Survey sample is composed of six groups, according to the number of months the households have been in the sample. The results as shown in Table 2 of the reinterview of these households, have been studied in order to discover any effect due to the length of time a household has been in the sample.

Table 2: Percentage disagreement between interviewer and reinterviewer by Labour Force category for respondents classified by interviewer number (number of times in sample).

	Wf	Wp	L	J	URO	HS
Int. No.						
1	12.4	74.0	41.5	65.8	30.7	11.2
2	13.8	65.6	47.1	79.9	29.1	10.1
3	15.2	69.9	41.3	70.1	31.2	9.3
4	13.9	63.0	50.9	67.6	33.5	10.0
5	13.7	59.7	50.3	93.9	35.4	11.0
6	14.2	68.6	42.1	56.9	31.7	10.2

On the basis of these results, no effect due to the length of time the household stays in the sample is apparent in the % disagreements.

Age-Sex: Response variability is obviously dependent on a variety of factors. It is interesting to investigate any effect of age and sex of the respondent on the percentage disagreement of major labour force survey classifications (W, L, J, NLF).

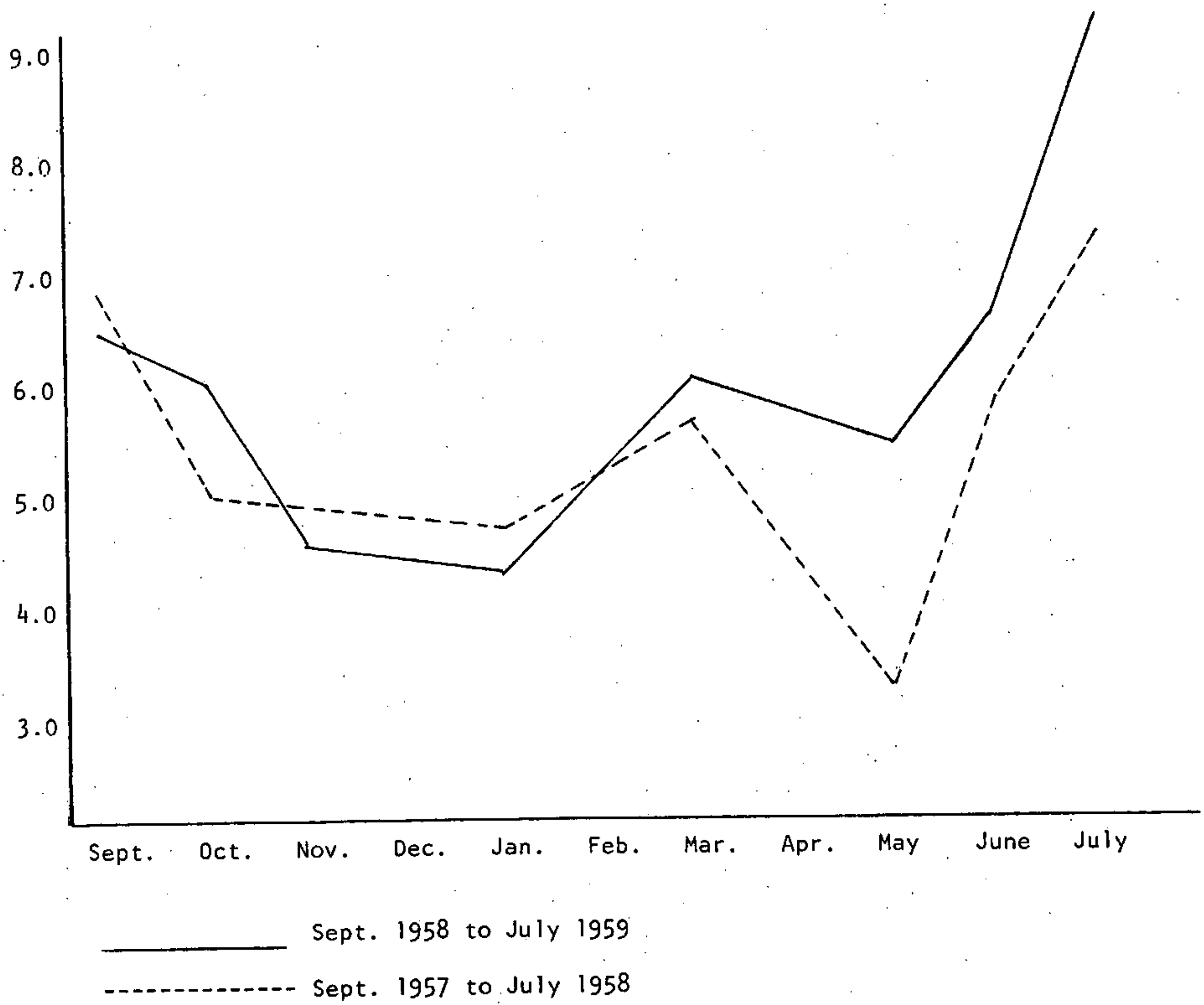
Table 3: Percentage disagreement between interviewer and reinterviewer by age-sex groups (using Labour Force categories W, L, J and NLF):

Age	14-19			20-24			25-44			45-64			65+			Total		
Sex	M	F	M+F	M	F	M+F	M	F	M+F	M	F	M+F	M	F	M+F	M	F	M+F
% Dis-agr.	8.3	6.3	7.3	7.3	4.9	6.0	4.4	4.6	4.5	5.8	7.4	6.6	6.7	2.7	4.6	5.8	5.4	5.6

When only the major classifications of W, L, J, and NLF are considered, the male 14-19 age group and the female 45-64 age group show the highest relative number of discrepancies, while the female 65+ and 25-44 age groups for both males and females show the lowest relative number of discrepancies. The pattern here is quite evident in that the highest rate of disagreement occurs in those groups which are most subject to change in Labour Force status, while the lowest rate is in those groups which tend to be more settled.

Seasonal Effect: Observation of the monthly percentage disagreement over a two year period led to the conclusion that the series was influenced by a seasonal effect. To demonstrate this, Graph 1 was prepared showing the comparison of the monthly percentage of disagreement between the enumeration and the regular re-enumeration for two successive years.

Graph 1



It is interesting to note that the lower percentage disagreements tend to coincide with the more stable employment periods and seem to be influenced by high unemployment periods in the winter and high summer employment periods.

Phase 2 of the reinterview program served a number of purposes. It provided the monitoring function necessary to maintain confidence in the interviewing operations. It enabled the supervisors to gain some insight into interviewing problems. It provided at least a rough measure of the degree of variation in response to be expected under the existing interviewing conditions,

i.e. the measure of disagreement. Some clues to the causes of the differences were also obtained. For example, characteristics of marginal attachment to the LFS were most often misclassified; the age-sex categories most likely to be less stable in relation to LF activities were involved in more misclassification and the seasons of either very high or very low employment seem to produce the greatest number of misclassifications.

In addition, Phase 2 provided data for studying the effect of many factors influencing the reinterview situation such as time lag between interview and reinterview, change of respondent within household, language problems, part-time activities, etc.

3.3 Phase 3

Another phase of the reinterview program was to attempt to study the sources and causes of differences arising between the original and reinterview data and to determine the effect of reconciling the differences during the reinterview. This reconciliation required that the reinterviewer have a copy of the original questionnaires for the households which were to be reinterviewed.

Method I

After the regular reinterview was carried out, the documents were matched in the regional offices, and where a difference occurred in the Labour Force activity, a third interview was carried out to determine which of the two interviews was correct, or if both were in error, and the reason for the difference. This additional information was entered on a separate sheet and attached to the matched questionnaires.

Method II

The reinterview was done in the usual manner. However, when the reinterview was completed, the reinterviewer compared the original and reinterview documents on the spot, to determine if they were exactly the same. If there were differences, the reinterviewer attempted to establish which of the two answers was correct and why the two different answers had been given. This information was entered on a separate sheet and attached to the corresponding

original and reinterview documents. The reinterviewer was instructed not to look at the original schedules before reinterview and it was emphasized that entries on the reinterview document must not be altered when the reconciliation of the two sets of data was made.

In order to remove the effect of different interviewers, the reinterviewer was assigned four households in the selected segments so that he reinterviewed an equal number of households under each method.

In order to investigate the effect of carrying the original document on the results of reinterview, a comparison was made of the number of differences due to the interviewer and the reinterviewer in Methods I and II. In Method I, the source of the difference is determined by a third interview. In Method II, the source of the difference is determined by the reinterviewer after he has completed reinterviewing. Table 24 shows the values of Ee% and Er%, where Ee% is the portion of the measure of disagreement found by reconciliation to be due to errors in the original interview, and Er% is the portion of the measure of disagreement found by reconciliation to be due to errors in the reinterviewing.

Table 4: Values of Ee% and Er% for the two methods of reconciliation by Labour Force category.

	Method I			Method II		
Labour Force Category	Ee%	Er%	Total	Ee%	Er%	Total
W	3.8	2.1	5.9	2.4	0.5	2.9
L	21.3	13.8	35.1	15.8	2.0	17.8
J	23.9	11.8	35.7	21.2	5.5	26.7
NLF	4.7	1.7	6.4	2.8	0.6	3.4
Total	2.8	1.4	4.2	1.8	0.4	2.2

The most obvious observation is that Method II reduced the number of disagreements substantially. Furthermore, most of the reduction resulted from the elimination of reinterviewer errors, presumably due to the presence of

the original questionnaire which could alert the reinterviewer to the possibility of an error. The results also suggest that by using Method II, a more accurate measure of the interviewers performance could be obtained, as the disagreements are reduced to mainly interviewers' errors. In addition, a situation was achieved where the reinterviewer was in the interviewers area and in possession of the original and reconciled questionnaires. This led to the development of Phase 4 of the reinterview program which was aimed at making use of the reinterview program as an instrument in the interviewers' training as well as a quality measure.

3.4 Phase 4

In Phase 4, training purposes were stressed. The sample for reinterview was selected on the basis of interviewers' assignments. When an interviewer's assignment was selected, one-third of the households were reinterviewed. The reinterviewer carried the original questionnaires and reconciled the data on the spot. The reinterviewer (supervisor) also contacted the interviewer while in the area. All of the reinterview households were reviewed with the interviewer and she would be instructed on any weaknesses brought to light by the reinterviewer. In addition, the reinterviewer would file a report on all misclassifications which were tabulated as a quality measure.

The data for this study has been accumulated over a number of years and there are a few convenient points in time for which the analysis from the study is of particular significance. These are the periods when there are changes in the overall Labour Force Survey design. The last major design took place after the 1961 Census of population, and the next redesign which includes major changes in questionnaire design, and mechanization, took place after the 1971 Census of population. The results from the Revised Labour Force Survey will not be available before 1976.

The reinterview program referred to as 1960, included several years of accumulation of data and reinterviews were carried out independently. That is to say, as described under Phase 2. The data from 1967 and later resulted from the procedures described in Phase 4.

The data in the following tables excepting 1960 shows the Measures of Disagreement pertaining to data after reconciliation.

Table 5: Measure of Disagreement (percentages) (Data averaged over 12 months)

Category	1960	1967	1969	1970	1971	1972	1973	1974
W	9.0	3.9	4.4	3.5	3.5	3.7	2.9	3.0
L	45.5	27.1	23.3	17.3	25.0	15.5	20.2	21.0
J	72.3	23.8	24.6	22.2	18.0	17.9	17.9	17.9
URO	32.4	14.5	15.0	13.6	14.5	10.5	11.7	13.0
HS	10.7	5.6	5.9	5.0	5.0	4.6	4.6	5.0

Data from the reinterview studies gives assurance that response errors are under control, though not at most satisfactory levels, and that some gradual reduction has been achieved. This has been through a continuous program of retraining, home exercises, more careful selection of interviewers, observation, reinterview, etc. Other detailed analysis of the reinterview data has indicated, however, that more extensive and intensive questions relating to hours of work, temporary lay-offs, job seeking activities, are necessary to discriminate effectively between certain labour market activities. A new revised questionnaire along with a revised sample and many new procedures, are being introduced for 1976. A new reinterview program is being planned for the revised survey which will serve both analytical and operational purposes. We may refer to it as Phase 5.

3.5 Phase 5

The approach in this program differs considerably from the one which has been, so far, applied in the Canadian Labour Force Survey. The former program rested heavily upon the verification of interviewers' work and upon providing some ideas of the difficulties that interviewers encountered in the field.

The change in the orientation in the program has been dictated by the need for measuring the effects of non-sampling errors on the estimates.

The reinterview program rests upon the following design. Each month, the reinterview sample is to be randomly split into two parts. In the first part, the reinterviews are conducted as independently as possible from the original interviews in order to achieve a repetition of the original survey under similar conditions. In the second part, the reinterviews are conducted in such a way as to detect errors made by the original interviewer. A two-step procedure is to be followed in this second part to achieve this: first, the reinterviewer, who is a senior interviewer responsible for the quality of the work performed by a group of interviewers, carries out an independent interview; second, having completed all interviews within a household, the reinterviewer compares the answers with those appearing on the copies of the original schedules, and tactfully discusses differences with the respondent. He will note the correct responses and code a reason to explain the discrepancies on the reconciliation form.

The accumulation of data from both parts of the sample will enable us to study general aspects of response errors. Thus, simple response variance may be estimated from the first part of the sample, under the assumption of a "repetition of a survey", and response bias may be estimated from the second part of the sample if it is assumed that the reinterviewer provides the "true" answers under the survey conditions. Also, it will be possible to conduct a general study of the variation of interviewers' performance, by post-stratifying according to interviewers' experience, duration of training, working area, season of the year, and so on. Furthermore, the field control aspect is not neglected. The reinterviewers will have the responsibility of determining specific instances where the information originally collected by the interviewer is incomplete, wrong or non-existent with respect to the coverage of households, and persons within, and when reconciliation is to take place.

The periodical analysis of the data produced by the program should permit detection of sources of errors and should lead to a continuous improvement of the overall methodology, since it will have a feedback with respect to various components of the survey design such as questionnaire design, interview technique, interviewer training, and so on.

A STRATEGY FOR UP-DATING CONTINUOUS SURVEYS

R. Platek and M.P. Singh

The need for regular up-dating of the selection probabilities in continuous surveys is emphasized in this paper. A simple strategy (selection method for the initial sample with the revision procedure) is presented and its application to the Canadian Labour Force Survey is discussed.

1. INTRODUCTION

In designing large scale surveys unequal selection probabilities, based on size measures, are often assigned to sampling units within stratum. The selected units using these initial probabilities may be used over several years in a continuous survey or in some cases, due to cost considerations, these units may be used for surveys with different objectives. But as the time passes the initial size measures used to determine initial selection probabilities become more and more out of date. The difference between the initial and new size measures may be revealed from the latest census or from a field count instituted at some appropriate intervals depending upon anticipated frequency of changes. In many cases the need for current survey may be better served by the new size measures and, therefore, it would be desirable to revise the selection probabilities, using the new size measures, yet retain the initial selections as much as possible since continued use of initial units has several advantages. Such a method of revising (up-dating) the selection probabilities was first presented by Keyfitz [4]. Kish and Scott [5] have discussed this problem in detail and suggested alternative procedures of revising the initial probabilities. However, their treatment is generally restricted to one unit per stratum.

For many complex selection methods involving two or more selections the revision of initial selection probabilities might require complex treatments due to their diverse joint inclusion probabilities [2]. Since the changes in size measures are inherent in any continuous survey the adoptability of the selection method to a simple procedure of revision of the selection probabilities is considered here as a desirable criterion for the choice of a selection method in the initial design of such surveys. Under this

criterion a simple strategy (i.e. selection method with the revision procedure) is presented and its application to the Canadian Labour Force Survey (LFS) is discussed.

2. PROCEDURE OF REVISION

Several methods of unequal probability sampling without replacement have been proposed and comparisons of the stabilities of the estimates of the population total and their variance estimates, and other efficiency comparisons have been made in recent years. However, not much is known about their adoptibility to the revision of selection probabilities. The procedure of revising the selection probabilities developed by Kish and Scott [5] amply demonstrates the suitability of the methods with one unit per stratum for continuous surveys. Two or more selections with replacement also fall into the same category as regards the revision procedures. For situations where two or more selections without replacement are considered necessary the selection method used in randomized grouping is examined and a revision procedure is described below.

In randomized grouping the N units of a stratum are randomized into n groups (where n is the number of units to be sampled). Having n randomized groupings the sample is obtained by selecting one unit with probability proportional to size (PPS) from each of n groups. Two types of groupings can be formed:

Type I: Groups be as equal as possible in the number of units in them (Rao, Hartley and Cochran [6]).

Type II: Groups may be equal as possible in aggregate measure of size (Cochran [1]).

We first discuss the revision procedure in relation to the method of grouping used in Type I when initial size measures X_j has changed to new size measures X_j^* (j = 1,2,...N). Two special features of this selection method are

(a) Each random group is by itself a random sample (with equal probability) from the stratum.

(b) After formation of groups, sampling within a group is done independently and only one unit is selected with PPS from each group.

The initial conditional probability of selecting t^{th} unit from group g (g = 1,2,...n) is

$$P_{gt} = \frac{X_{gt}}{X_g} \quad , \tag{3.1}$$

X_g being the aggregate size measure for the group g, $X_g = \sum_{t=1}^{N_g} X_{gt}$, and N_g is the number of units in group g. For getting maximum efficiency N_g should be as equal as possible [6].

As no size measures are involved in the formation of random groups the initial grouping may be retained for the life of the stratum provided there is no changes in the number of units contained in the initial stratum. The new conditional probabilities corresponding to p_{gt} using the new size measures X_j^* (j=1,2,...N) would then become

$$p_{gt}^* = \frac{X_{gt}^*}{X_g^*} \qquad \begin{array}{l} g = 1,2,...N \\ t = 1,2,...N_g \end{array} \tag{3.2}$$

where $X_g^* = \sum_t X_{gt}^*$ denotes the new aggregate size measure for group g. Note that

$$\sum_t p_{gt} = \sum_t p_{gt}^* = 1 \, .$$

Consider a particular group (say k^{th}) for which the initial and new conditional probabilities are respectively p_{kt} and p_{kt}^* $t = 1,2,...N_k$. Then the selection in this group may be revised as follows:

(i) if $p_{kt}^* \geq p_{kt}$, retain the initially selected unit U_{kt} in the sample as if selected with p_{kt}^*.

(ii) if $p_{kt}^* < p_{kf}$, retain U_{kt} in the sample with probability p_{kt}^*/p_{kt}. This unit is dropped with probability $1 - p_{kt}^*/p_{kt}$.

(iii) if $p^*_{kt} < p_{kt}$ and if the unit is dropped in (ii) then select a unit from those units with $p^*_{kt} > p_{kt}$ with probability proportioned to $(p^*_{kt} - p_{kt})$.

The conditional probability of selecting a unit when $p^*_{kt} < p_{kt}$ is

$$p(U_{kt}) = p_{kt}\ (p^*_{kt}/P_{kt}) = p^*_{kt}\ . \qquad (3.3)$$

And conditional probability of selecting a unit when $p^*_{kt} \geq p_{kt}$ is

$$P(U_{kt}) = p_{kt} + (\underset{1}{\Sigma} p^*_{kt} - \underset{1}{\Sigma} p_{kt})\ \frac{p^*_{kt} - p_{kt}}{\underset{1}{\Sigma}(p^*_{kt} - p_{kt})} = p^*_{kt}\ . \qquad (3.4)$$

where $\underset{1}{\Sigma}$ denotes summation over units with $p^*_{kt} > p_{kt}$. Thus (3.3) and (3.4) show that the required conditional probabilities p^*_{kt} are obtained for all the units (i.e. t = 12...N_k) for this group.

This is essentially a simple version of Keyfitz procedure. It is because of the feature (b) of the selection method that enables revision of the selection probabilities (conditional) within the individual groups.

3. MODIFIED PROCEDURE

Due to independent up-dating of the sample within the individual groups of a stratum the above procedure may be used in all n groups or it may be confined to only those groups and to those units within the group for which the ratio p^*_{gt}/p_{gt} departs greatly from 1. An allowable degree of departure of this ratio may be determined on the basis of the expected gain in efficiency in using the new size measures and the operational constraints in selecting new units or using revised probabilities for the selected units. The procedure is thus quite flexible and consists of following steps.

Set p_{gt} = original probability of selection

p'_{gt} = revised probability of selection

(i) Compute the relative size measures p^*_{gt} and the ratio of relative sizes $R_{gt} = p^*_{gt}/p_{gt}$ for all g and t.

(ii) Determine the groups in which $R_1 \leq R_{gt} \leq R_2$ for all t, where R_1 and R_2 are the lower and upper allowable limits. In each of these groups set

$$p'_{gt} = p_{gt} \quad \text{for } \underline{\text{all}} \ t\varepsilon g$$

implying no up-dating is necessary.

(iii) In each remaining group, devide the units into 2 classes C and $\overline{C}$ such that unit $t\varepsilon C$ if $R_1 \leq R_{gt} \leq R_2$ and $t\varepsilon\overline{C}$ otherwise.

Then set

$$\begin{aligned} p'_{gt} &= p_{gt} && \text{if } t\varepsilon C \\ &= p^*_{gt} + \varepsilon_{gt} && \text{if } t\varepsilon\overline{C}\ , \end{aligned}$$

where $\varepsilon_{gt} = 0$ or whose absolute value is $<<p^*_{gt}$, such that $\sum_{t\varepsilon C} P_{gt} + \sum_{t\varepsilon C} p'_{gt} = 1$ that is

$$\sum_{t\varepsilon C} P_{gt} + \sum_{t\varepsilon\overline{C}} (p^*_{gt} + \varepsilon_{gt}) = 1 \ .$$

And compute the sum

$$\sum_{t\varepsilon\overline{C}} \varepsilon_{gt} = 1 - \sum_{t\varepsilon C} p_{gt} - \sum_{t\varepsilon C} p^*_{gt}$$

$$= \sum_{t\varepsilon C} (p^*_{gt} - p_{gt}) \text{ or } = \sum_{t\varepsilon\overline{C}} (p_{gt} - p^*_{gt}) \ .$$

(iv) Assign the sum $\sum_{t\varepsilon\bar{C}} \varepsilon_{gt}$ to the unit (or the set of units) with ratio(s) showing maximum deviation from 1 to obtain p'_{gt} for units in $\bar{C}$ (method I).

Step (iii) and (iv) may alternatively be performed as follows (method II).

(iii)' In each of the remaining group set $p'_{gt} = p_{gt}$ if $t\varepsilon C$ and for all $t\varepsilon\bar{C}$ compute p'_{gt} as

$$p'_{gt} = p^{*}_{gt} \frac{\underset{t\varepsilon\bar{C}}{} P_{gt}}{\sum\limits_{t\varepsilon\bar{C}} p^{*}_{gt}}$$

$$= p^{*}_{gt} \left(1 + \frac{\sum\limits_{t\varepsilon\bar{C}} \varepsilon_{gt}}{\sum\limits_{t\varepsilon\bar{C}} p_{gt}}\right) .$$

Thus while in (iv) the adjustment will be made to usually the largest unit(s) for getting revised probabilities, in (iii)' the revised probabilities p'_{gt} will have to be computed for all $t\varepsilon\bar{C}$, which in some cases may be quite time consuming particularly when there are large number of groups (strata) involved. Method II is likely to be more efficient than method I due to proportionate distribution of the balance among the units in $\bar{C}$, however the gain may be negligible since $\sum_{t} \varepsilon_{gt}$ is likely to be small in most situations.

Remark 1: Both in the basic procedure and the above modified procedures it is assumed that the number of units contained in the initial stratum remains unchanged. Without loss of generality, if a unit has been omitted from the new stratum, then the new conditioned probability ($p^{*}_{kt'}$) of this particular ($U_{kt'}$) may be considered to be zero and dropped with certainity. Further, the new units of a stratum may be first randomly assigned to the random groups and then the same method of up-dating may be followed to the individual random groups by taking the initial conditional probabilities for such units as zero and computing the new conditional probabilities.

Remark 2: The procedure is also applicable to the classical case of one unit per stratum.

An Example: We consider an example selecting one unit with PPX from stratum consisting of 15 units. The values of X_i, P_i, X_i^*, P_i^* are given in the following Table. The revised probabilities p_i' calculated for methods I and II are given in column (10) and (11) respectively.

Table 1: Example of Calculation

									Values of	p_i'
									Method	Method
Units	X_i	P_i	X_i^*	P_i^*	R_i	Class	p_i'	$p_i-p_i^*$	I	II
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1	27	.0450	30	.0432	.9606	C	.0450		.0450	.0450
2	42	.0700	43	.0620	.8851	$\overline{C}$	$.0620+\varepsilon_2$	.0080	.0620	.0690
3	51	.0850	56	.0807	.9493	C	.0850		.0850	.0850
4	37	.0617	34	.0490	.7942	$\overline{C}$	$.0490+\varepsilon_4$	.0127	.0490	.0482
5	24	.0400	29	.0418	1.0450	C	.0400		.0400	.0400
6	19	.0317	27	.0389	1.2271	$\overline{C}$	$.0389+\varepsilon_6$	- .0072	.0389	.0382
7	36	.0600	42	.0605	1.0083	C	.0600		.0600	.0600
8	52	.0807	57	.0821	.9469	C	.0867		.0867	.0867
9	41	.0683	58	.0836	1.2240	$\overline{C}$	$.0836+\varepsilon_9$	- .0153	.0836	.0822
10	63	.1050	90	.1297	1.2352	$\overline{C}$	$.1297+\varepsilon_{10}$	- .0247	.1213	.1274
11	37	.0617	38	.0548	.8882	$\overline{C}$	$.0548+\varepsilon_{11}$	.0069	.0548	.0539
12	42	.0700	49	.0706	1.0086	C	.0700		.0700	.0700
13	50	.0833	50	.0720	.8643	$\overline{C}$	$.0720+\varepsilon_{13}$	.0113	.0720	.0708
14	39	.0650	42	.0605	.9308	C	.0650		.0650	.0650
15	40	.0667	49	.0706	1.0585	C	.0667		.0667	.0667
	600	1.0001	694	1.0000			1.0084	- .0083	1.0000	1.0000

Supposing that $R_1 = .9$ and $R_2 = 1.1$

Then units 1,3,5,7,8,12,14,15 belong to C and the probabilities of selection $p_i' = p_i$ (unchanged).

The remaining units belong to $\overline{C}$ and the probabilities of selection are altered to $p_i' = p_i^* + \varepsilon_i$ where ε_i must be determined.

$\sum_{t\varepsilon\overline{C}} \varepsilon_{gt} = \sum_{t\varepsilon C} (p_{gt}^* - p_{gt}) = -.0084$ which must be assigned to the units in $\overline{C}$ in some manner by one method of adjustment. Units 6, 9 and 10 show the maximum ratio R_i, deviating from 1 in the opposite direction of $\Sigma\ \varepsilon_{gt}$ so that we would like to adjust the sizes of units. One way would be to decrease .1297 to .1297 - .0084 = .1213.

In Method II,

$$p_i' = p_i^* \left[1 + \frac{\sum_{t\varepsilon\overline{C}} (p_{gt} - p_{gt}^*)}{\sum_{t\varepsilon\overline{C}} p_{gt}^*} \right]$$

$$= p_i^* \left[1 + \frac{(-.0083)}{.4900} \right] = p_i^* \quad .983061 \ .$$

4. APPLICATION IN THE LFS

A detail description of the design used in the Canadian Labour Force Survey (LFS) conducted monthly is given in [3]. We shall briefly discuss the design used for the larger cities (approximately 15,000 population or more) called Self-Representing Units (SRUs). These SRU are first divided into compact strata which are further sub-divided into clusters (mostly city blocks). From each stratum 6 clusters are selected and these are then assigned at random a rotation group number 1 to 6 indicating the months in which the households of the sampled segments will rotate out. The selected households of a cluster remain in the sample for six consecutive months after which they are replaced by another group of households from the same cluster.

Selection of clusters was done systematically with PPS in the old design and difficulties were faced in the past in updating the sample in the SRUs. Recognizing that the problem of up-dating is quite serious and important in larger cities as the population growth in these cities occurs at a much faster rate than in smaller urban centers and rural areas, the method of

randomized groupings is being used to select 6 clusters from each stratum. Further, it is intended to up-date the sample, using the modified procedure, regularly on the basis of the new size measures for all clusters as obtained from an annual field count system instituted for the purpose.

The design of the survey is made self-weighting. An unbiased estimate of the stratum total is

$$\hat{Y} = \sum_g \hat{Y}_g$$

$$= \sum_g \hat{Y}_{gt}/p_{gt} \text{ or } \sum_g \hat{Y}_{gt}/p'_{gt}$$

where t is the selected unit and $\hat{Y}_g$ is the estimate of group and $\hat{Y}_{gt}$ is the estimate of unit t in group g given by

$$\hat{Y}_{gt} = W_{gt} \sum_u Y_{gtu} \quad ,$$

W_{gt} being inverse of the sub-sampling ratio and $\sum_u Y_{gtu}$ the sample total in unit t of group g.

The weight may be made constant from group to group by ensuring that $W_{gt}/p_{gt} = W$

i.e. $W_{gt} = Wp_{gt}$ in original sample

and $= Wp'_{gt}$ in revised sample.

However, with aggregate size measures ($X_{g's}$) being different from group to group in type I grouping, the sample size will be different from group to group within strata. Two possibilities considered for avoiding different sample sizes among groups were (i) to adopt type II grouping or (ii) apply different weights in each group. In adopting type II grouping (i.e. $X_g = X_k$) the sample take will although be the same in the original sample but as the sample is revised the aggregate size measures will become unequal introducing variation in sample sizes among the groups. Further, the computation of

joint inclusion probabilities becomes very complicated in this method since these probabilities become function of t and t', and exact expression for variance and unbiased variance estimates may not be possible.

Further, the use of varying weights over numerous sub-units (strata) is expected to prove cumbersome in the Canadian LFS. Therefore slight different sample sizes in each group was preferred, and attempts are made to balance the sample size over strata by rotation group number.

The authors would like to thank Mr. G.B. Gray for his comments.

REFERENCES

[1] Cochran, W.G., Sampling Technique Second Edition (1963), Ch. 9 Wiley, N.Y.

[2] Fellegi, I.P., "Changing the Probabilities of Selection when Two Units are Selected with PPS without Replacement". Preceedings of Social Statistics Section, American Statistical Association Washington 1966, 434-42.

[3] Fellegi, I.P., G.B. Gray and R. Platek, "The New Design of the Canadian Labour Force Survey"; Journal of American Statistical Association (1967) 62, 421-453.

[4] Keyfitz, N., "Sampling with Probabilities proportional to size - Adjustment for Changes in size; Journal of American Statistical Association, 58 (1961), 183-201.

[5] Kish, L. and Scott, A., "Retaining Units After Changing Strata and Probabilities" Journal of American Statistical Association (1971), 66, 461-470.

[6] Rao, J.N.K., Hartley, H.O. and Cochran, W.G., "On a Simple Procedure of Unequal Probability Sampling Without Replacement" Journal of Royal Statistical Society (1962) 24, 482-491.

COMPONENTS OF VARIANCE MODEL IN MULTI-STAGE STRATIFIED SAMPLES

G.B. GRAY

There are several multi-stage sample designs in various countries, such as the Current Population Survey in U.S.A., Labour Survey in Sweden, and the General Household Survey in United Kingdom. From each survey, estimated totals of Employed, Unemployed, and other characteristics may be obtained.

The Canadian Labour Force Survey is a monthly household survey in which the dwelling is the ultimate unit of sampling requiring two to four stages of selection. Each province is split up into strata and sampling units at various stages so that the sampling variance contains up to four components of variance whose actual formulae and estimation formulae are derived, utilizing those formerly derived by Yates and Grundy [12]. Ratio estimation is employed and the formulas are modified accordingly. To analyze the components of variance, it is necessary to express them in terms of components of sampling ratios and the sizes of sampling units at the various stages at provincial and national levels and approximate variance functions are thus derived.

1. INTRODUCTION

As a typical example of a multi-stage stratified sample, I shall consider the Canadian Labour Force Survey which is a monthly survey used to estimate by a sample of about 30,000 households (ultimate units), total employment, unemployment, and numerous other characteristic totals at provincial and national levels. In each province, there are basically two distinct sample designs (i) self-representing units (SRU) and (ii) non-self-representing units (NSRU) which require separate analyses, although the formulae are somewhat similar for the two designs. Horvitz-Thompson estimators [8], ratio estimators and the variance and variance estimation formulae based on Yates and Grundy [12] have been applied and extended to derive a variance function in terms of weights, numbers of units, and population variances. The Horvitz-Thompson estimators have been employed for obtaining the average weights and sizes of units at various stages, both of which vary quite widely, thus necessitating the use of probability proportional to size sampling in addition to ratio estimation.

2. SAMPLE DESIGN (based on 1961 Census)

The SRU areas comprise strata, denoted by h, in large cities and metropolitan areas aplit up into city blocks or groups of city blocks called segments (k). n_h segments out of N_h segments are selected systematically with probability proportional to size in each stratum. Each segment comprises N_{hk} ultimate units out of which n_{hk} are selected systematically with equal probability in such a way that the overall probability of selection of each ultimate unit is constant for a given province and SRU area. For purposes of components of variance analysis, the segments and listing units are assumed to be in random order prior to selection.

Within each province, the NSRU areas comprise strata denoted by h, each of which is divided up into N_h primary sampling units (denoted by i) and n_h (equalled to 2) are selected with probability proportional to size by a method derived by Fellegi [3]. Most primary sampling units consist of urban areas (outside large cities and metropolitan areas) and rural areas, denoted in general by type of area j. Each type of area within a primary sampling unit is split up into N_{hij} enumeration areas or groups of clusters called segments k of which n_{hij} are selected systematically with probability proportional to size. Most segments are split up into N_{hijk} single or multiple clusters c of which n_{hijk} are selected systematically with probability proportional to size. In each selected cluster containing N_{hijkc} ultimate units, all (in the case of a single cluster) or a random systematic subsample of n_{hijkc} (in the case of a multiple cluster) ultimate units u are selected. The overall probability of selection of any ultimate unit is constant for a given province and NSRU area. As in the case of SRU areas, whenever systematic sampling is applied, the units are assumed to be in random order prior to selection for purposes of components of variance studies.

3. ESTIMATION (NSRU Areas)

Estimates of characteristic totals in a multi-stage stratified sample (eg., Employed or Unemployed in Canadian Labour Force Survey) may be obtained in successive stages (S referring to summation over the sample and Σ referring to summation over the population of units) by Horvitz-Thompson [8] estimators;

as follows:

$$\hat{X}_{hijkc} = \mathop{S}_{u\varepsilon(h,i,j,k,c)} x_{hijkcu}/\pi_{u|hijkc} \text{ estimates } X_{hijkc} ,$$

$$\hat{X}_{hijk} = \mathop{S}_{c\varepsilon(h,i,j,k)} \hat{X}_{hijkc}/\pi_{c|hijk} \text{ estimates } X_{hijk} ,$$

$$\hat{X}_{hij} = \mathop{S}_{k\varepsilon(h,i,j)} \hat{X}_{hijk}/\pi_{k|hij} \text{ estimates } X_{hij} ,$$

$$\hat{X}_{hi} = \sum_{j} \hat{X}_{hij} \text{ estimates } X_{hi} ,$$

$$\hat{X}_{h} = \mathop{S}_{i} \hat{X}_{hi}/\pi_{i|h} \text{ estimates } X_{h} ,$$

and $$\hat{X}_{hj} = \mathop{S}_{i} \hat{X}_{hij}/\pi_{i|h} \text{ estimates } X_{hj} ,$$

where X denotes expected value over all possible samples in the area denoted by the subscripts and if the responses x_{hijkcu} were true responses[1] for listing unit (h,i,j,k,c,u), X would denote true totals of some characteristics in the area denoted by the subscripts.

In general, we shall denote $\underline{i}_r$ as a vector of r th stage units where

$\underline{i}_o = h,$

$\underline{i}_1 = (h,i,j)$ with j being type of area (urban, rural)

$\underline{i}_2 = (h,i,j,k)$

$\underline{i}_3 = (h,i,j,k,c)$

$\underline{i}_4 = (h,i,j,k,c,u)$

and $i_r|\underline{i}_{r-1}$ denotes r th stage unit i_r within (r-1) th stage unit $\underline{i}_{r-1}$. Thus

[1] Actually, x_{hijkcu} is subject to response variance but the variance in response has been omitted from the Components of Variance model since we are discussing sampling variance only in this article.

$\underline{i}_r = (\underline{i}_{r-1}, i_r)$ and $\pi_{i_r|\underline{i}_{r-1}}$ denotes the inclusion probability of unit i_r within $\underline{i}_{r-1}$.

4. COMPONENTS OF VARIANCE

The Yates-Grundy [12] formula for the variance between r^{th} stage units in $\underline{i}_{r-1}$ is given by

$$V_{r|\underline{i}_{r-1}} = \sum_{i_r<i'_r} (\pi_{i_r|\underline{i}_{r-1}} \pi_{i'_r|\underline{i}_{r-1}} - \pi_{i_r i'_r|\underline{i}_{r-1}}) \left(\frac{X_{\underline{i}_{r-1}\, i_r}}{\pi_{i_r|\underline{i}_{r-1}}} - \frac{X_{\underline{i}_{r-1}\, i'_r}}{\pi_{i'_r|\underline{i}_{r-1}}} \right)^2 \qquad (4.1)$$

and for the variance at all subsequent stages within $\underline{i}_{r-1}$,

$$V_{\underline{i}_{r-1}} = V_{r|\underline{i}_{r-1}} + \sum_{\underline{i}_r} \frac{V_{\underline{i}_r}}{\pi_{i_r|\underline{i}_{r-1}}} \qquad (4.2)$$

Here, $\pi_{i_r i'_r|\underline{i}_{r-1}}$ denotes the joint inclusion probability between units i_r and i'_r in (r-1)th stage unit $\underline{i}_{r-1}$.

The true variance (1) written in a different form by Horvitz-Thompson [8] may also be written a third form, in terms of the sampling variance when pps with replacement times a finite population correction and we shall derive the formula here as it is a useful form for analysis and development of variance functions in terms of sampling ratio, numbers of units and population variances. For purposes of analysis and study of the performance of the design under different sample allocations for example, a simple form of the variance must be used if practical results are to be obtained and applied. The Horvitz-Thompson and Yates-Grundy forms of the variance have proven to be unweildy for practical analysis of components of variance study and consequently, a third form has been derived.

For the development, we shall omit the subscripts $\underline{i}_{r-1}$ and abbreviate $r|\underline{i}_{r-1}$ by r so that

$$V_r = \sum_{i_r < i'_r} (\pi_{i_r} \pi_{i'_r} - \pi_{i_r \, i'_r}) \left(\frac{X_{i_r}}{\pi_{i_r}} - \frac{X_{i'_r}}{\pi_{i'_r}} \right)^2 .$$

Let us suppose $n_{r|\underline{i}_{r-1}}$ or n_r units out of $N_{r|\underline{i}_{r-1}}$ or N_r units are selected with pps. Then $\pi_{i_r} = n_r \, p_{i_r|\underline{i}_{r-1}}$ or $n_r p_{i_r}$, where p_{i_r} denotes the relative size of unit i_r (or the relative probability of selection when not exactly proportional to some pre-determined size).

5. DEVELOPMENT OF 3RD FORM OF VARIANCE

If sampling with pps with replacement were undertaken, the variance between r^{th} stage units would become

$$V'_r = \frac{N_r^2 \, \sigma_r^2}{n_r} \qquad (5.1)$$

where $$N_r^2 \, \sigma_r^2 = \Sigma \, p_{i_r} \left(\frac{X_{i_r}}{p_{i_r}} - X \right)^2 ,$$

X representing the total over all r^{th} stage units in $\underline{i}_{r-1}$ or $X_{\underline{i}_{r-1}}$.

$$V_r = V'_r \, [1 + (n_r - 1) \, r_{FP:r}] \qquad (5.2)$$

where $r_{FP:r}$ denotes a finite population correlation denoted by:

$$r_{FP:r} = \frac{\sum_{i_r} \sum_{i_r \neq i'_r} \Pi_{i_r \, i'_r} \left(\frac{X_{i_r}}{p_{i_r}} - X\right) \left(\frac{X_{i'_r}}{p_{i'_r}} - X\right) / n_r (n_{r-1})}{\sum_{i_r} p_{i_r} \left(\frac{X_{i_r}}{p_{i_r}} - X\right)^2} \qquad (5.3)$$

σ_r^2 and $r_{FP:r}$ reduce to the classical values $\frac{1}{N} \sum_i (X_i - \overline{X})^2$ and $- 1/(N-1)$ respectively when sampling with equal probability without replacement is undertaken.

$$V_r = \sum_{i_r < i'_r} (n_r^2 p_{i_r} p_{i'_r} - \pi_{i_r i'_r}) [(\frac{X_{i_r}}{n_r p_{i_r}} - \frac{X}{n_r}) - (\frac{X_{i'_r}}{n_r p_{i'_r}} - \frac{X}{n_r})]^2$$

$$= \tfrac{1}{2} \sum_{i_r} \sum_{i'_r \neq i_r} (p_{i_r} p_{i'_r} - \frac{\pi_{i_r i'_r}}{n_r^2})[(\frac{X_{i_r}}{p_{i_r}} - X)^2 + (\frac{X_{i'_r}}{p_{i'_r}} - X)^2$$

$$- 2 (\frac{X_{i_r}}{p_{i_r}} - X) (\frac{X_{i'_r}}{p_{i'_r}} - X)]$$

$$= \sum_{i_r} \sum_{i'_r \neq i_r} (p_{i_r} p_{i'_r} - \frac{\pi_{i_r i'_r}}{n_r^2}) . (\frac{X_{i_r}}{p_{i_r}} - X)^2$$

$$- \sum_{i_r} \sum_{i'_r \neq i_r} (p_{i_r} p_{i'_r} - \frac{\pi_{i_r i'_r}}{n_r^2}) (\frac{X_{i_r}}{p_{i_r}} - X) (\frac{X_{i'_r}}{p_{i'_r}} - X)$$

$$= \sum_{i_r} (p_{i_r} - p_{i_r}^2 - \frac{n_r - 1}{n_r} p_{i_r}) (\frac{X_{i_r}}{p_{i_r}} - X)^2$$

$$- \sum_{i_r} p_{i_r} (\frac{X_{i_r}}{p_{i_r}} - X)[-p_{i_r} (\frac{X_{i_r}}{p_{i_r}} - X)]$$ [1]

[1] This follows from the fact that $\sum_{i_r} p_{i_r} (\frac{X_{i_r}}{p_{i_r}} - X) = 0$ and hence $\sum_{i'_r \neq i_r} p_{i'_r} (\frac{X_{i'_r}}{p_{i'_r}} - X) = -p_{i_r} (\frac{X_{i_r}}{p_{i_r}} - X)$

$$+ \frac{n_r(n_r-1)}{n_r^2} \; r_{FP:r} \; \sum_{i_r} P_{i_r} \left(\frac{X_{i_r}}{P_{i_r}} - X\right)^2$$

$$= \left[\frac{1}{n_r} + \frac{n_r-1}{n_r} r_{FP:r}\right] \sum_{i_r} P_{i_r} \left(\frac{X_{i_r}}{P_{i_r}} - X\right)^2$$

$$= \frac{N_r^2 \; \sigma_r^2}{n_r} \; [1 + (n_r - 1) \; r_{FP:r}] \text{ as required in (5.2)}$$

It should be noted that when sampling with replacement occurs, $r_{FP:r} = 0$ and $V_r = N_r^2 \; \sigma_r^2/n_r$.

6. ESTIMATES OF COMPONENTS OF VARIANCE (by stratum)

By Yates-Grundy formulas [12], the estimation formulas as may be derived as follows:

$$\hat{V}_{\underline{i}_{r-1}} = \underset{i_r<i'_r}{S} \left(\frac{\pi_{i_r} \; \pi_{i'_r}}{\pi_{i_r \, i'_r}} - 1\right) \left(\frac{\hat{X}_{i_r}}{\pi_{i_r}} - \frac{\hat{X}_{i'_r}}{\pi_{i'_r}}\right)^2 + \underset{r}{S} \; \pi_{i_r}^{-1} \; \hat{V}_{i_r} \qquad (6.1)$$

(See also Fellegi [3] p. 185)

and the components of variance derived by subtraction beginning with

$\hat{V}_{4:\underline{i}_3} = \hat{V}_{\underline{i}_3}$ since there are only 4 stages of sampling and consequently $V_{\underline{i}_4} = 0$.

From 4.2 and 6.1,

$$\hat{V}_{\underline{i}_3} = \hat{V}_{4:\underline{i}_3}$$

$$= \underset{i_4<i'_4}{S} \left(\frac{\pi_{i_4|\underline{i}_3} \; \pi_{i_4|\underline{i}_3}}{\pi_{i_4 i'_4|\underline{i}_3}} - 1\right) \left(\frac{\hat{X}_{i_4|\underline{i}_3}}{\pi_{i_4|\underline{i}_3}} - \frac{\hat{X}_{i'_4|\underline{i}_3}}{\pi_{i'_4|\underline{i}_3}}\right)^2 \qquad (6.2)$$

Now $V_{4:\underline{i}_o} = \sum_{i_1 i_2 i_3} (\pi_{i_1|h} \pi_{i_2|\underline{i}_1} \pi_{i_3|\underline{i}_2})^{-1} V_{4:\underline{i}_3}$

and $\hat{V}_{4:i_o} = \underset{i_1 i_2 i_3}{S} (\pi_{i_1|h} \pi_{i_2|\underline{i}_1} \pi_{i_3|\underline{i}_2})^{-2} \hat{V}_{4:\underline{i}_3}$

Also $\hat{V}_{4:\underline{i}_2} = \underset{i_3}{S} \pi_{i_3|\underline{i}_2}^{-2} \hat{V}_{4:\underline{i}_3}$ estimates $V_{4:hijk}$ (6.3)

Now $\hat{V}_{\underline{i}_2} = \hat{V}_{3:\underline{i}_2} + \hat{V}_{4:\underline{i}_2}$ (from 6.1)

$$= \underset{i_3<i_3'}{S} \left(\frac{\pi_{i_3|\underline{i}_2} \pi_{i_3'|\underline{i}_2}}{\pi_{i_3 i_3'} \pi_{\underline{i}_2}} - 1\right) \left(\frac{\hat{X}_{i_3|\underline{i}_2}}{\pi_{i_3|\underline{i}_2}} - \frac{\hat{X}_{i_3'|\underline{i}_2}}{\substack{1 \\ \pi_{i_3|\underline{i}_2}}}\right)^2$$

$$+ \underset{i_3}{S} \frac{\hat{V}_{r:\underline{i}_3}}{\pi_{i_3|\underline{i}_2}} \tag{6.4}$$

and $\hat{V}_{3:\underline{i}_2}$ is thus obtained by subtraction of (6.3) from (6.4) or

$$\hat{V}_{3:\underline{i}_2} = \underset{i_3<i_3'}{S} \left(\frac{\pi_{i_3|\underline{i}_2} \pi_{i_3'|\underline{i}_2}}{\pi_{i_3 i_3'|\underline{i}_2}} - 1\right) \left(\frac{\hat{X}_{i_3|\underline{i}_2}}{\pi_{i_3|\underline{i}_2}} - \frac{\hat{X}_{i_3'|\underline{i}_2}}{\pi_{i_3'|\underline{i}_2}}\right)^2$$

$$- \underset{i_3}{S} \left(\frac{1}{\pi^2_{i_3|\underline{i}_2}} - \frac{1}{\pi_{i_3|\underline{i}_2}}\right) \hat{V}_{4:i_3} \tag{6.5}$$

$$\hat{V}_{3:\underline{i}_1} = \underset{i_2}{S} \pi_{i_2|\underline{i}_1}^{-2} \hat{V}_{3:\underline{i}_2}$$

and $\hat{V}_{3:hj} = \underset{i_1}{S} \pi_{i_1|h}^{-2} \hat{V}_{3:\underline{i}_1}$ (6.6)

Similarly, $\hat{V}_{2:hj} = \underset{i}{S}\ \pi_{i_1|h}\ \hat{V}_{2:\underline{i}_1}$ (6.7)

where
$$\hat{V}_{2:hij} = \underset{i_2<i_2'}{S} \left(\frac{\pi_{i_2|\underline{i}_1}\ \pi_{i_2'|\underline{i}_1}}{\pi_{i_2 i_2'|\underline{i}_1}} - 1\right) \left(\frac{\hat{X}_{i_2|\underline{i}_1}}{\pi_{i_2|\underline{i}_1}} - \frac{\hat{X}_{i_2'|\underline{i}_1}}{\pi_{i_2'|\underline{i}_1}}\right)^2$$

$$- \underset{i_2}{S} \left(\frac{1}{\pi^2_{i_2|\underline{i}_1}} - \frac{1}{\pi_{i_2|\underline{i}_1}}\right) (\hat{V}_{3:\underline{i}_2} + \hat{V}_{4:\underline{i}_2}) \qquad (6.8)$$

and finally
$$\hat{V}_{1:hj} = \underset{i_1<i_1'}{S} \left(\frac{\pi_{i_1|h}\ \pi_{i_1'|h}}{\pi_{i_1 i_1'|h}} - 1\right) \left(\frac{\hat{X}_{i_1 j|h}}{\pi_{i_1|h}} - \frac{\hat{X}_{i_1' j|h}}{\pi_{i_1'|h}}\right)^2$$

$$- \underset{i_1}{S} \left(\frac{1}{\pi^2_{i_1|}} - \frac{1}{\pi_{i_1|h}}\right) (\hat{V}_{2:\underline{i}_1} + \hat{V}_{3:\underline{i}_1} + \hat{V}_{4:\underline{i}_1}) \qquad (6.9)$$

and
$$\hat{V}_{1:h} = \underset{i_1<i_1'}{S} \left(\frac{\pi_{i_1|h}\ \pi_{i_1'|h}}{\pi_{i_1 i_1'|h}} - 1\right) \left(\frac{\hat{X}_{i_1|h}}{\pi_{i_1|h}} - \frac{\hat{X}_{i_1'|h}}{\pi_{i_1'|h}}\right)^2$$

$$- \underset{i_1}{S} \left(\frac{1}{\pi_{i_1|h}} - \frac{1}{\pi_{i_1|h}}\right) \sum_{j} (\hat{V}_{2:\underline{i}_1} + \hat{V}_{3:\underline{i}_1} + \hat{V}_{4:\underline{i}_1}) \qquad (6.10)$$

including a small urban-rural covariance between urban and rural characteristic totals of PSUs which exists because of sampling urban and rural areas together in selected PSUs rather than selecting independent samples in these two types of areas.

At provincial levels for NSRU areas, all above variances are additive over strata and in turn they are additive over stages of sampling to determine total sampling variances.

7. ADAPTATION TO RATIO ESTIMATION

Up to now, we have derived the estimation and variance estimation pertaining to simple blown-up estimates. In multi-stage samples where ratio estimation is applied instead of $\hat{X} = \sum_h \hat{X}_h$, $\hat{Z} = \sum_a P_a \ (\hat{X}_a/\hat{P}_a)$ is obtained where $\hat{X}_a$ and $\hat{P}_a$ are characteristic total and population estimates by age-sex categories or at the provincial level as obtained from the sample using the successive Horvitz-Thompson estimation procedures described at the beginning of the appendix. P_a is the projected census population for province-age-sex cell (as projected from the last census) while not free of mean square error contains no sampling variance so that P_a has been assumed constant in subsequent formulas.

By using the approximate relationship Rel Var (x/y) = Rel Var x - 2 Rel Cov xy + Rel Var y (eg. Cochran [1]), one may replace X by $X - \sum_a R_a P_a$ and $\hat{X}$ by $\hat{X} - \sum \hat{R}_a \hat{P}_a$ (see also [9]) in all of the variance and variance estimation formulas and supply the appropriate subscripts; eg., $\hat{X}_{c|\underline{i}_2}$ would be replaced by $\hat{X}_{c|i_2} - \sum_a \hat{R}_a \hat{P}_{a:c \ i_2}$ in formula (10). Here, $R_a = X_a/P_a$ and $\hat{R}_a = \hat{X}_a/\hat{P}_a$. In these formulas, we assume $\hat{R}_a$ and $\hat{P}_a$ independently distributed, though some small correlation may exist between them. The accuracy of the above 2nd order approximation for the rel-variance of ratios when small populations as often exist in LFS at various stages may also be questionable. We have not investigated the accuracy of the ratio estimate variance approximation for small populations when pps sampling is applied.

8. ESTIMATION OF σ^2 AND r_{FP} FROM THE SAMPLE

It was noted in formula (5.2) that $V_{r|\underline{i}_{r-1}}$ may be factored into two components; viz., $N^2_{r|\underline{i}_{r-1}} \ \sigma^2_{r|\underline{i}_{r-1}} / n_{r|\underline{i}_{r-1}}$ and $[1 + (n_{r|\underline{i}_{r-1}} - 1) \ r_{FP:r|\underline{i}_{r-1}}]$, the first being the rth stage component of variance of the total when the rth stage units in $\underline{i}_{r-1}$ are selected with pps with replacement and $[1 + (n_{r|\underline{i}_{r-1}} - 1) \ r_{FP:r|\underline{i}_{r-1}}]$ being the finite population correction when sampling with pps without replacement. Actual values of $r_{FP:r|\underline{i}_{r-1}}$ depend upon both the sample design and the number of selected units. It will be different, for example, between Fellegi's method [3]

and pps systematic of randomly ordered units (Hartley and Rao [7]). For rough estimates of σ^2 in the discussion of allocation by size and sampling ratio at various stages, one may assume $r_{FP} = -1/N_{r|\underline{i}_{r-1}}$ without much error according to empirical calculations. Otherwise, σ^2 and r_{FP} must be estimated from the sample and we shall derive an estimate of σ^2 and hence of r_{FP} in a particular area.

Noting that in (5.1) $\dfrac{N_r^2\,\sigma_r^2}{n_r} = \sum\limits_r n_r p_{i_r} \left(\dfrac{X_{i_r}}{n_r p_{i_r}} - \dfrac{X}{n_r}\right)^2$, let

us consider the statistic $\dfrac{N_r^2\,s_r^2}{n_r} = \mathop{S}\limits_r \left(\dfrac{\hat{X}_{i_r}}{n_r p_{i_r}} - \dfrac{\hat{X}}{n_r}\right)^2$

and derive its expected value to see if together with $\hat{V}_r$, estimates of σ^2 and r_{FP} may be obtained by the solution of two equations in two unknowns.

$$E\,\frac{N_r^2\,s_r^2}{n_r} = E\,\mathop{S}\limits_r \left(\frac{\hat{X}_{i_r}}{n_r p_{i_r}}\right)^2 - n_r\,E\left(\frac{\hat{X}}{n_r}\right)^2$$

$$= \sum_r n_r p_{i_r}\,\frac{X_{i_r}^2 + V(\hat{X}_{i_r})}{(n_r p_{i_r})^2} - \frac{1}{n_r}\,[X^2 + V(\hat{X})]$$

$$= \sum n_r p_{i_r} \left(\frac{X_{i_r}}{n_r p_{i_r}} - \frac{X}{n_r}\right)^2 + \sum_r \frac{V(\hat{X}_{i_r})}{n_r p_{i_r}} - \frac{V(\hat{X})}{n_r}$$

$$= N_r^2\,\sigma_r^2/n_r + E\,\mathop{S}\limits_r \frac{\hat{V}_{i_r}}{\pi_{i_r}^2} - \frac{1}{n_r}\,E\,\hat{V}_{i_{r-1}},$$

or $$\frac{N_r^2\,\sigma_r^2}{n_r} = E\left[\frac{N_r^2\,s_r^2}{n_r} - \mathop{S}\limits_r \frac{\hat{V}_{i_r}}{\pi_{i_r}^2} + \frac{\hat{V}_{i_{r-1}}}{n_r}\right]$$

$$\text{or} \quad \hat{\sigma}^2_{r|\underline{i}_{r-1}} = s^2_{r|\underline{i}_{r-1}} - \frac{n_{r|\underline{i}_{r-1}}}{N^2_{r|\underline{i}_{r-1}}} \underset{r}{S} \frac{\hat{V}_{i_r}}{\pi^2_{i_r}} - \frac{1}{N^2_{r|\underline{i}_{r-1}}} \hat{V}_{i_{r-1}} \qquad (8.1)$$

When ratio estimation is applied, the parameters X and statistics $\hat{X}$ are replaced as indicated in the section under Adaptation to Ratio Estimates.

9. VARIANCE FUNCTIONS IN TERMS OF AVERAGE σ^2's AND N's

Individual values of the components of variance at small area levels are inaccurate not only for approximations described but also because of the instability of small area data at stratum levels. In order to analyze the variance components at macro levels properly, it is necessary to average the estimated parameters over units as well as merely add the components of variance over strata.

The variance components at province NSRU levels (over several strata) are given by:

$$V_{1j} = \sum_h V_{1:hj} = \sum_h N_{1:h} \; \sigma^2_{1:hj} \; \frac{N_{1:h}}{n_{1:h}} \; [1 + (n_{1:h} - 1) \; r_{FP;1:hj}]$$

$$= L \; \bar{N}_1 \; \bar{\sigma}^2_{1j} \; \bar{W}_{1j} \; [1 + (\frac{\bar{N}_1}{\bar{W}_{1j}} - 1) \; \bar{r}_{FP:1j}] \qquad (9.1)$$

where $\sum_{h=1}^{L} N_{1:h} \, \sigma^2_{1:hj} = L \; \bar{N}_1 \; \bar{\sigma}^2_{1j}$

$\bar{W}_{1j}$ to be obtained later on

$$\bar{N}_1 = \frac{1}{L} \sum_{h=1}^{L} N_{1h}$$

and $\bar{r}_{FP:1j}$ derived by equating the second line to the first.

Similarly, $V_{2j} = \sum_h V_{2:hj} = \sum_h \sum_i \frac{1}{\pi_{i|h}} V_{2:hij}$ est'd by $\sum_h \sum_i \frac{1}{\pi^2_{i|h}} \hat{V}_{2:hij}$

$$= \sum_h \sum_i \frac{1}{\pi_{i|h}} N_{2:hij}\, \sigma^2_{2:hij}\, \frac{N_{2:hij}}{n_{2:hij}} [1 + (n_{2:hij} - 1)\, r_{FP:2:hij}]$$

$$= L\, \bar{N}_1\, \bar{N}_{2j}\, \bar{\sigma}^2_{2j}\, \bar{W}_{1j}\, \bar{W}_{2j}\, [1 + (\frac{\bar{N}_{2j}}{\bar{W}_{2j}} - 1)\, \bar{r}_{FP:2j}] \qquad (9.2)$$

where $L\, \bar{N}_1\, \bar{N}_{2j}\, \bar{\sigma}^2_{2j} = \sum_h \sum_i N_{2:hij}\, \sigma^2_{2:hij}$ est'd by $\Sigma_h\, S_i \frac{1}{\pi_{i|h}} N_{2:hij}\, \hat{\sigma}^2_{2:hij}$

$L\, \bar{N}_1\, \bar{N}_{2j} = \sum_h \sum_i N_{2:hij}$ est'd by $\sum_h \sum_i \frac{1}{\pi_{i|h}} N_{2:hij}$

Similarly, $V_{3j} = L\, \bar{N}_1\, \bar{N}_{2j}\, \bar{N}_{3j}\, \bar{W}_{1j}\, \bar{W}_{2j}\, \bar{W}_{3j}\, \bar{\sigma}^2_{3j}\, [1 + (\frac{\bar{N}_{3j}}{\bar{W}_{3j}} - 1)\, \bar{r}_{FP:3j}] \qquad (9.3)$

and finally, $V_{4j} = L\, \bar{N}_1\, \bar{N}_{2j}\, \bar{N}_{3j}\, \bar{N}_{4j}\, \bar{W}_{1j}\, \bar{W}_{2j}\, \bar{W}_{3j}\, \bar{W}_{4j}\, \bar{\sigma}^2_{4j}$

$$[1 + (\frac{\bar{N}_{4j}}{\bar{W}_{4j}} - 1)\, \bar{r}_{FP:4j}] \qquad (9.4)$$

10. AVERAGE WEIGHTS OVER UNITS

In a self-weighting sample, $(\pi_{i_1|h}\, \pi_{i_2|\underline{i}_1}\, \pi_{i_3|\underline{i}_2}\, \pi_{i_4|\underline{i}_3})^{-1} = W_j$, a constant weight for every selected unit $\underline{i}_4$. If no non-response occurred, it would be desirable to obtain average weights $\bar{W}_{1j}$, $\bar{W}_{2j}$, $\bar{W}_{3j}$, $\bar{W}_{4j}$ so that their product equals W_j.

Average weights are thus defined by:

$$\tilde{W}_{rj} = d_{rj}/d_{r-1:j}, \quad \text{for } r = 4,3,2,1 \text{ in succession, where} \qquad (10.1)$$

$$d_{4j} = \underset{\underline{i}_4}{S'} (\pi_{i_1|h}\, \pi_{i_2|\underline{i}_1}\, \pi_{i_3|\underline{i}_2}\, \pi_{i_4|\underline{i}_3})^{-1},$$

$$d_{3j} = \underset{\underline{i}_3}{S'} (\pi_{i_1|h}\, \pi_{i_2|\underline{i}_1}\, \pi_{i_3|\underline{i}_2})^{-1}\, n_{4:\underline{i}_3},$$

$$d_{2j} = \underset{\underline{i}_2}{S'} (\pi_{i_1|h}\, \pi_{i_2|\underline{i}_1})^{-1}\, n_{4:\underline{i}_2}$$

$$d_{1j} = \underset{\underline{i}_1}{S'} (\pi_{i_1|h})^{-1}\, n_{4:\underline{i}_1},$$

$$d_{0j} = \sum_h n_{4:hj},$$

and $\underset{\underline{i}_4}{S'} = \sum_h \underset{i_1 i_2 i_3 i_4}{S}, \quad \underset{\underline{i}_3}{S'} = \sum_h \underset{i_1 i_2 i_3}{S}$ etc.

and $n_{4:\underline{i}_r}$ = number of sampled 4th stage units within rth stage unit $\underline{i}_r$.

Finally, the product of the four average weights is given by:

$$\tilde{W}_{1j}\, \tilde{W}_{2j}\, \tilde{W}_{3j}\, \tilde{W}_{4j} = \frac{\sum_h \underset{i}{S}\, \underset{k}{S}\, \underset{c}{S}\, \underset{u}{S}\, W_j}{\sum_h n_{4:hj}} = W_j \text{ as required in a self-weighting}$$

sample provided that no non-response has occurred.

11. AVERAGE SIZE OF UNITS AT VARIOUS STAGES

If we denote average size of r^{th} stage units by $\bar{P}_{rj}$ (number of persons) or $\bar{U}_{rj}$ (number of ultimate units), for type of area j,

$$\text{then}\quad \bar{P}_{rj} = \hat{P}_j/\hat{N}_{rj}$$

$$\text{and}\quad \bar{U}_{rj} = \hat{N}_{4j}/\hat{N}_{rj},$$

$$\text{where}\quad \hat{P}_j = \underset{\underline{i}_4}{S'} W_j P_{4:\underline{i}_3},$$

$$\hat{N}_{rj} = \underset{\underline{i}_r}{S'} (\pi_{i_1|h} \pi_{i_2|\underline{i}_1} \pi_{i_3|\underline{i}_2} \cdots \pi_{i_r|\underline{i}_{r-1}})^{-1} \text{ for } r=1,2,3,4,$$

$$\text{and}\quad P_{4:\underline{1}_3} = \text{number of (enumerable) persons in } \underline{i}_3 .$$

12. JOINT PROBABILITIES OF SELECTION

Joint probabilities of selection are the most difficult parameters to calculate or estimate. For systematic selection with probability proportional to size, exact values for small populations are readily calculated by a method first introduced by W.S. Connor in 1966 [2] and developed with minor modifications by G.B. Gray [6]. Approximate joint inclusion probabilities may be readily calculated either from a large number of random orderings of units, utilizing W.S. Connor's method or by the asymptotic formula in [7]. In all cases, the units must be assumed to be in random order prior to selection.

When Fellegi's method of selection is used (see [3]) as in the case of NSRU primary sampling units, joint probabilities are very easily calculated as indicated in the article.

If the random group method were adopted in place of systematic pps sampling or units in randomized order, the variance and estimated variance formulae must conform to those stated by Rao, Hartley and Cochran [10] and joint inclusion probabilities will not be required.

Systematic sampling with equal probability has been adopted in all cases for ultimate units within clusters or SRU segments. If we assume the units to be randomly ordered prior to selection; the sampling procedure is identical to a simple random selection and for n units out of N units, the joint inclusion probability of any pair of units is simply n(n-1)/[N(N-1)].

13. ACKNOWLEDGEMENT

The author wishes to express his appreciation for the support and comments of R. Platek, Director, Household Surveys Development Staff, and for the comments provided by the referee.

REFERENCES

[1] Cochran, W.G., Sampling Techniques, 2nd ed., New York: John Wiley and Sons, 1963.

[2] Connor, W.S., "An Exact Formula for the Probability that Two Samples Drawn with Unequal Probabilities and Without Replacement", Journal of the American Statistical Association, 61 (1966), 384-390.

[3] Fellegi, I.P., "Sampling with Varying Probabilities Without Replacement: Rotating and Non-Rotating Samples", Journal of the American Statistical Association, 58 (1963), 183-201.

[4] Fellegi, I.P., Gray, G.B., and Platek, R., "The New Design of the Canadian Labour Force Survey", Journal of the American Statistical Association, 62 (1967), 421-453.

[5] Gray, G.B., "Variance Components and Variance Function", Proceedings of the Canadian Conference in Applied Statistics, Statistics '71 Canada, 1971, 119-126.

[6] Gray, G.B., "Joint Probabilities in Systematic Samples", American Statistical Association Proceedings of the Social Statistics Section (1972), 271-276.

[7] Hartley, H.O. and Rao, J.N.K., "Sampling With Unequal Probabilities and Without Replacement", Annals of Mathematical Statistics, 33 (1962), 350-374.

[8] Horvitz, D.G. and Thompson, D.J., "A Generalization of Sampling Without Replacement from a Finite Universe", Journal of the American Statistical Association, 47 (1952), 663-685.

[9] Keyfitz, N., "Estimates of Sampling Variance where Two Units are Selected from Each Stratum", Journal of the American Statistical Association, 52 (1957), 503-510.

[10] Rao, J.N.K., Hartley, H.O., and Cochran, W.G., "On a Simple Procedure of Unequal Probability Sampling Without Replacement", Journal of the Royal Statistical Society, Series B, 27 (1962), 482-491.

[11] Sampling and Survey Research Staff, Dominion Bureau of Statistics (now Statistics Canada): "Methodology, Canadian Labour Force Survey", (1965).

[12] Yates, F. and Grundy, P.M., "Selection Without Replacement from Within Strata with Probability Proportional to Size", Journal of the Royal Statistical Society, Series B, 15 (1953), 253-261.

NON-INTERVIEW PATTERNS IN THE CANADIAN LABOUR FORCE SURVEY

R. Sugavanam

This paper summarizes the results of a project conducted to study non-interviews in the Canadian Labour Force Survey. Temporarily absent (32.7%), no-one-home (31.4%), and refusal (25.5%) are the major components of non-response. The impact of these components to the total non-response in Surveys from July 1972 to June 1973 is discussed in detail.

A detailed analysis of refusal households showed that existing field follow-up procedures were not quite successful in reducing the refusal component. As expected, non-response was found to be related to the length of tenure of households in the sample. Non-response among households enumerated for the first time was generally higher than those households already in the sample.

1. INTRODUCTION

The Canadian Labour Force Survey is a continuing monthly survey whose main function is to provide estimates of employment and unemployment at the national and provincial levels. Interviews are carried out in about 35,000 households (across the country) chosen by area sampling methods. The sample used in the Labour Force Surveys has been designed to represent all persons in the population 14 years of age and over residing in Canada, with the exception of the following: residents of the Yukon and Northwest Territories, Indians living on reserves, inmates of institutions and members of the armed forces. Some non-interviews are virtually certain to occur in each survey whether it is because of operational difficulties, or there is no one at home during the entire enumeration week, or for some other reason. This means that interviewed households have to represent slightly more households than was intended in the design of the survey. In the Labour Force Survey, the final weight attached to each record is adjusted for non-interviews on the assumption that households which have been interviewed represent the characteristics of households which should have been enumerated. However, if this assumption is not true, the estimates will be biased and the bias will increase with a higher rate of non-interview.

Every effort is made in the field to minimize non-interviews. Some of the procedures followed are:

a) a reasonable number of call-backs are made if the reason for non-interview is that there is no one at home at the time of the interviewer's visit,

b) an attempt is made on the Monday following the survey week to interview households which were away during the survey week,

c) the regional office representative attempts to interview households which refuse to provide information to the interviewers.

In addition to the field procedures designed to reduce non-interviews, the design of the survey has a rotation scheme which is conducted every month to replace approximately one-sixth of the households in the sample. A selected household is retained in the sample for six months. The rotation of the sample attempts to reduce the refusal rate which might substantially increase if the same households were required to provide information month after month.

In order to study the non-interview patterns, the response status of households in the Labour Force Surveys during July 1972 to June 1973 were analyzed. This paper provides some new results on the behaviour of the different components of non-interview in the Canadian Labour Force Survey.

2. NOTATION AND TERMINOLOGY

2.1 Household

A household refers to any person or group of persons occupying a dwelling. A dwelling is a set of living quarters which is structurally separate and has a private entrance which can be used without passing through someone else's living quarters.

2.2 Response Status

The response status shows whether the household was interviewed, or the reason if it was not interviewed. The various reasons given for non-interviews (see appendix) can be grouped as follows:

Code	Response Status
I	Interviewed
T	Temporarily Absent
N	No-one-home
R	Refusal
O	Other
V	V-type non-interview

V-type non-interviews include vacant dwellings, vacant seasonal dwellings, dwellings under construction, dwellings occupied by persons not to be interviewed, dwellings that are converted to business use, demolished, etc.

2.3 Rotation Group

There are six rotation groups (1,2,3,4,5 and 6) in the sample. Each rotation group consists of approximately one-sixth of households in the sample. Each month, households belonging to a particular rotation group are replaced by new households. The rotation is conducted in a systematic way such that every month the sample is a probability sample of the population covered by the survey. The design of the Labour Force Survey retains a household in the sample for six consecutive surveys.

In the Labour Force Survey, rotation group 1 is rotated in January and July, rotation group 2 in February and August, rotation group 3 in March and September, ..., and rotation group 6 in June and December. Thus, in any survey, the rotation group numbers enable us to identify the households which have been enumerated once or in more than one survey.

2.4 Response Vector

A response vector for a selected household shows the response status of that household during its tenure of six Labour Force Surveys.

3. SIX MONTH FILE

One of the short-comings of the current Labour Force Survey is that non-interview information is not carried on the tape. Reasons for non-interview are found on survey control documents. In order to accommodate the need of

a data base for studies on response patterns, six month file was created by merging the information on the control documents with the monthly Labour Force Survey tapes.

In this paper, the analysis is restricted to households which have not responded at least once during their tenure in the Labour Force Survey. A response vector was constructed for each household in the sample during July 1972 to June 1973. From this set of vectors, two-way gross-flow tables giving $N(X,X^*,S)$, where $N(X,X^*,S)$ is the number of households with response status X in Survey (S-1) and response status X^* in Survey (S), were prepared. These tables provided the information used to study the non-interview patterns in the Labour Force Survey.

4. MAJOR COMPONENTS OF NON-INTERVIEW

4.1 V-type Non-Interviews

Due to various reasons, the entire sample cannot be enumerated in any survey. Dwellings in the sample can be divided into the following mutually exclusive and exhaustive groups:

A: Dwellings consisting of interviewed households (with response status 1),

B: Dwellings consisting of non-respondent households (with response status T, N, R, or O)

C: V-type non-interview dwellings

It should be noted that dwellings in groups A and B contain households, while dwellings in group C do not contain any household. Table 1 gives the percentage of dwellings in the sample that are in groups A, B, and C during July 1972 to June 1973.

Table 1: Percentage of Dwellings in the sample that are in Groups A, B and C during July 1972 to June 1973.

SURVEY	A Interviewed	B Non-Respondent	C V-type-Non-interview
July 72	77.3	11.3	11.4
Aug. 72	79.9	9.0	11.1
Sept. 72	83.8	5.4	10.8
Oct. 72	84.8	4.6	10.6
Nov. 72	84.6	4.7	10.7
Dec. 72	83.5	5.7	10.8
Jan. 73	82.7	6.4	10.9
Feb. 73	82.8	6.4	10.8
Mar. 73	83.3	6.0	10.7
Apr. 73	81.7	7.0	11.3
May 73	82.7	6.1	11.2
June 73	81.3	7.6	11.1
Average	82.4	6.6	11.0

We notice from Table 1 that the percentage of V-type non-interviews does not vary substantially from month to month in the Labour Force Surveys. The average V-type non-interviews during July 1972 to June 1973 was 11 percent of the total number of dwellings in the sample. V-type dwellings are excluded in the calculation of estimates. Consequently the presence of these dwellings will not result in any bias in the survey estimates, but an excessive proportion of these dwellings will cause an increase in the sampling variance because of the smaller expected dwelling count.

4.2 Temporarily Absent and No-one-home

Table 1 shows that the percentage of non-respondent households varies considerably from one survey to another. Non-response follows a marked seasonal pattern, generally peaking in the summer months and declining in spring and autumn. The seasonal effect is mainly caused by the temporarily absent component which increases sharply during the summer months when people are away on vacation.

Table 2 gives the percentage contribution of components T, N, R and 0 to the total non-response in surveys during July 1972 to June 1973.

Table 2: Percentage Contribution of the Components of Non-Response in Surveys during July 1972 to June 1973.

Survey	Temporarily Absent Response Status T	No one home Response Status N	Refusal Response Status R	Others Response Status 0
July 72	59.0	17.0	20.0	4.0
Aug. 72	50.7	20.7	21.8	6.8
Sept. 72	30.6	30.9	29.8	8.7
Oct. 72	26.6	35.9	27.5	10.0
Nov. 72	24.6	39.0	27.3	9.1
Dec. 72	22.0	36.7	23.5	17.8
Jan. 73	25.5	35.3	24.3	14.9
Feb. 73	31.0	29.3	26.5	13.2
Mar. 73	28.6	30.0	28.1	13.3
Apr. 73	29.4	33.9	25.8	10.9
May 73	24.6	35.8	29.5	10.1
June 73	39.2	32.1	22.1	6.6
Average	32.7	31.4	25.5	10.5
Standard Deviation	11.4	6.6	3.2	3.9
Co-efficient Variation	34.9	20.9	12.4	36.8

The 12-month average contribution of components T and N are almost equal (32.7% and 31.4%), but their contribution to total non-response in any one survey is substantially different. In summer months (June, July and August), temporarily absent non-interviews are higher than no-one-home non-interviews, while the opposite is true in other months. The contribution of these components to non-response in a survey can be divided roughly as shown below.

	Temporarily Absent	No-one-home
Summer months	40-60%	15-35%
Other months	20-30%	30-40%

It is possible to combine T and N non-interviews into a single category, i.e. not-at-home at the time the interviewer visited the dwelling. In the Labour Force Survey, this distinction is made because N's are considered to be controlable non-interviews, while nothing can be done to interview temporarily absent non-interviews.

In July and August, due to the large number of temporarily absent non-interviews, records are imputed for these households, if they have been interviewed in the previous survey. In the Revised Labour Force Survey, there are plans to impute records similarly for no-one-home non-interviews.

4.3 Refusal

A household is classified as refusal (response status R) when a responsible member of the household definitely refuses to provide the survey information. Refusals account for an average of 25.5% of non-response in a survey. Table 2 shows that the percentage contribution of this component to the total non-response does not vary very much from one survey to another. Refusals can significantly affect the survey estimates if refusal households differ significantly from interviewed households. Existing field procedures attempt to gain the co-operation of refusal households by persuasion.

Analyzing the response vectors of households in surveys from July 1972 to June 1973, it was found that on the average 42.2% of households in the sample did not respond at least once during their tenure of six consecutive Labour Force Surveys. Furthermore, 48.3% of households which refused the first survey remained as refusals in the subsequent five surveys, while another 42.6% responded at least once in the next five surveys.

From Table 3, we observe that refusal households in a survey can be divided into the following groups:

i) Hard-core refusals, ie. refused all six surveys 38.1%

ii) Refused in both the previous and this survey, but not accounted for in (i) 26.5%

iii) Responded in the previous survey but refused this survey 15.6%

iv) Other 19.8%

19.8% of refusals in the "other" category includes refusal households (a) that are introduced in this survey but not classified as hard-core refusal and (b) households which were classified as another type of non-interview in the previous survey. Using the facts that the number of households in different rotation groups are approximately of equal size and 48.3% of refusals in the newly introduced rotation group (in this survey) are hard-core refusals, it can be shown that a substantial number of households in the "other" category were classified as non-interviews (other than refusal) in the previous survey. Existing field procedures do not reduce the percentage of refusals, as it can be seen from Table 3 that the net gain of respondents is generally negative, i.e. the number of respondent households which refuse in the subsequent survey is more than the number of refusal households that are persuaded to respond in the subsequent survey.

4.4 Effect of Six Month Tenure on Non-Response

Households in surveys during July 1972 to June 1973 were divided into six groups on the basis of the length of time these households were in the sample. Table 4 (and graph 1) gives the average number of households classified as T,N,R, and O non-response in different groups. Let us denote by group N (N=1 to 6) households which have been enumerated in the previous (N-1) surveys.

Table 4: Average Number of Households classified as Non-Respondents in Groups* 1 to 6 during July 1972 to June 1973 - Canada

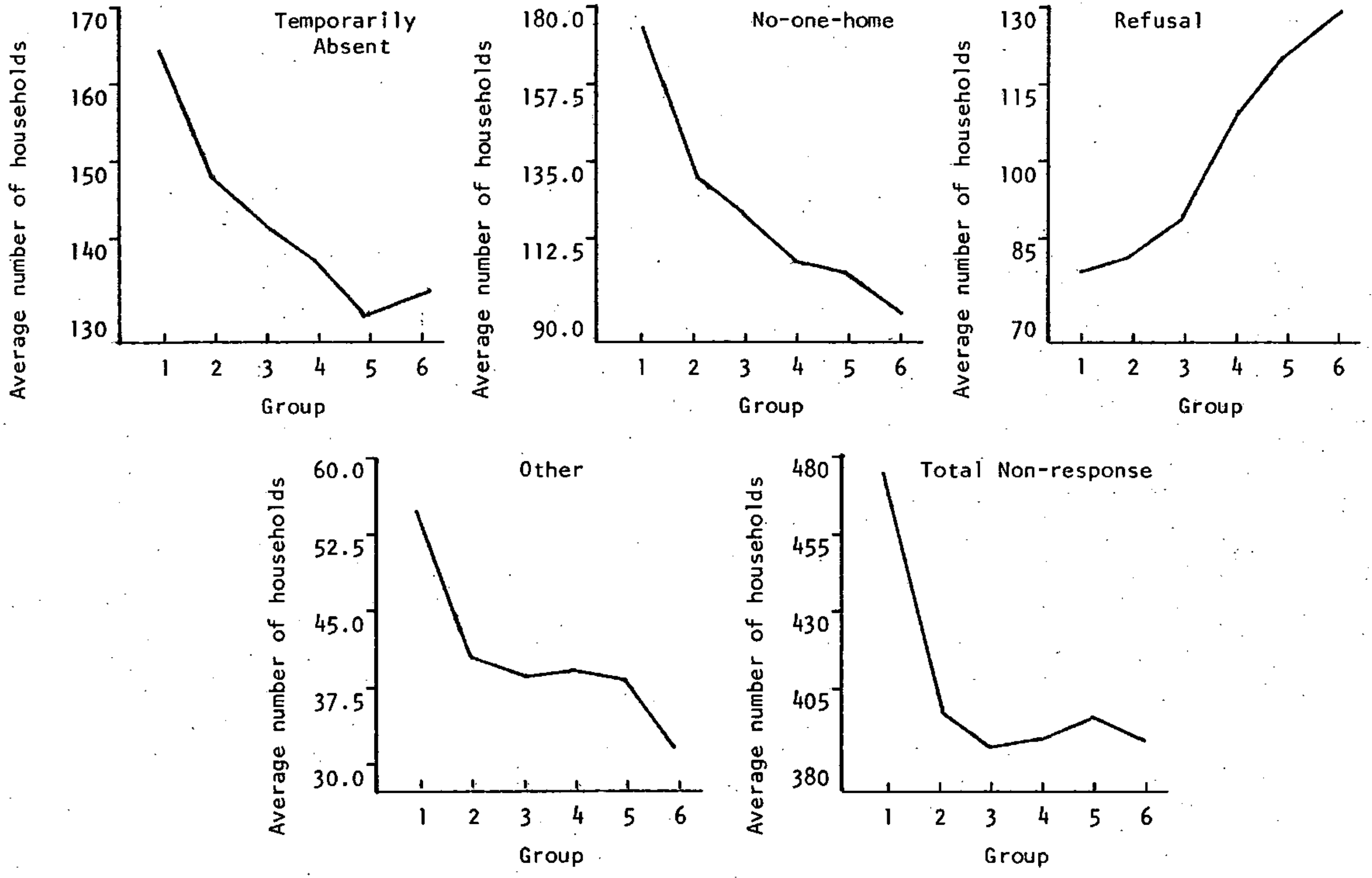
Group	Temporarily Absent	no-one home	Refusal	Other	Total
Group 1	164.4	172.9	79.7	54.0	471.0
Group 2	147.5	128.9	81.8	39.8	397.9
Group 3	141.0	118.2	89.1	38.3	386.6
Group 4	137.6	105.8	106.4	38.8	388.5
Group 5	133.2	104.6	118.1	37.9	393.8
Group 6	134.1	95.2	127.0	31.9	388.2

* Households which belong to Group N have been enumerated in the previous (N-1) surveys.

GRAPH 1

Average Number of households classified as Non-respondents during July 1972 to June 1973

CANADA



Households which belong to group N have been enumerated in the previous (N-1) surveys.

Let $G_N(X)$ represent the average number of households with response status X in group N. From Table 4, we notice the following:

(1) Non-response is related to the length of time the households were in the sample. Total non-response is generally higher in group 1 compared to other groups.

(2) Refusals (R) are the least in group 1. This group consists of rotated in households. One of the purposes of rotation of sample in the Labour Force Survey is clearly accomplished. The increasing trend of refusals present another difficult question, namely what is the optimum period a household should be in the sample. Note that G_1 (R) < G_2 (R) < G_3 (R) G_4 (R) < G_5 (R) < G_6 (R).

(3) Temporarily absent and no-one-home components decrease considerably after the first survey, i.e.

$$G_2 (T) < G_1 (T)$$

$$G_2 (N) < G_1 (N)$$

The decreasing trend is seen in groups 3, 4, and 5. It is not possible to explain why G_6 (T) > G_5 (T).

One expects the no-one-home component to decrease substantially after the first survey and later (i.e. in groups 2 to 6) stabilize as interviewers would have found out the best time to visit. However the decrease of the temporarily absent component in different groups is difficult to explain. Also, the substantial difference between G_1 (O) and G_2 (O) is difficult to explain.

5. CONCLUDING REMARKS

Non-response can be minimized through a better organization to get hold of T and N households. The conclusion that non-response is related to the length of time the households were in the sample suggests that resources should be spent to convert different types of non-interview in the groups discussed in the previous section, i.e. resources may be spent to reduce refusals in group 6, while follow-up procedures in the field should concentrate on no-one-home non-interviews in group 1. It may be useful to develop the profile of households who respond in all the six surveys so that cost and time resources can be efficiently allocated to achieve the goal of minimizing non-response in a survey.

REFERENCES

[1] Methodology: Canadian Labour Force Survey (Dominion Bureau of Statistics, Sampling and Survey Research Staff, 1965; Statistics Canada, Household Surveys Development Staff), catalogue 71-504 (occasional).

[2] Labour Force Quality Report, a monthly publication by Household Surveys Development Staff, Labour Force Survey Division and Field Division, Statistics Canada.

APPENDIX

NON-INTERVIEW CODES USED IN THE CANADIAN LABOUR FORCE SURVEY

One of the most important measures of quality in the Labour Force Survey is the non-response rate. The non-response rate refers to the proportion of households that were not interviewed due to their unavailability to the survey interviewer or to the lack of co-operation on the part of the householder.

In the Labour Force Survey, non-interviews are classified as follows:

Alphanumeric Code	Reasons for Non-Interview
(i) Temporarily Absent	
TA	All members of the household are absent for the entire Enumeration period[1].
(ii) N-type Non-Interview (No Interview)	
N1	No one home after a reasonable number of call-backs.
N2	Refusal - a responsible member of the household definitely refuses to provide the survey information requested
N3	Non-interview due to death, illness, language problems, interviewer's returns lost etc.
N4	No call made - roads impassable.
N5	No enumerator available (sick, resigned etc.).

(iii) V-type Non-Interview (No Household)

V1 Vacant - no persons are living in the dwelling[1].

V2 Vacant seasonal dwelling, vacant summer cottage or vacant trailer parking space in a regular trailer park.

V3 Dwelling under construction.

V4 Dwelling occupied by persons not to be interviewed; that is all persons in the household are not qualified to participate in the Labour Force Survey. Persons not qualified for the survey include those who have a usual place of residence elsewhere, full time members of the Canadian Armed Forces, inmates of institutions, visitors from other countries, etc.

V5 Other types - such as a dwelling which has been demolished or converted to business use, a trailer (not in a regular trailer park) which has been moved away, etc.

The temporarily absent and N-type non-interview classification indicates that the dwelling does contain a household but no interview was completed.

However, the V-type non-interview classification indicates that either no household is contained in the dwelling or all occupants in the dwelling are not qualified to be included in the Labour Force Survey. Moreover, all dwellings classified as V5 are deleted from the sample frame.

A COMPARISON OF SOME BINOMIAL FACTORS FOR THE LABOUR FORCE SURVEY

M. Lawes

Binomial factors (sometimes called design effects) can be used to assess the quality and performance, with respect to sampling variability of survey estimates, of a sample design and estimation procedure relative to assumed simple random sample designs. In this paper four types of binomial factors have been defined and calculated for the monthly Canadian Labour Force Survey. Some results from the analysis of these factors are presented in this paper.

1. INTRODUCTION

The sampling variance of any survey estimate is a function of the following factors: the sample design (including stratification, delineation, and allocation of units), the estimation procedure, the size of the universe and the size of the sample (the sampling ratio), the proportion of the population possessing the characteristic being measured, the distribution of the characteristic being measured in the population, the response rate and the slippage rate.

The Canadian Labour Force Survey (LFS) is a stratified multi-stage area sample of households of Canada ([3]). On the basis of this survey, estimates of major Labour Force characteristics are published on a monthly basis. For the LFS, nearly all of the characteristics being measured are of the qualitative type (i.e., each person either possesses or does not possess a given characteristic). Thus, one might compare the sampling variance estimated from the sample with the corresponding variance of an estimate based on the same number of sampled persons (i.e., the same sampling ratio) assuming a simple random sample (SRS). The sampling variance of individuals based on an SRS with replacement is the binomial variance, and a small correction called the finite population correction is applied to the binomial variance.

The following four types of binomial variances are considered in this paper:

Type 1: $$BV_1 \, (\hat{X}_p) = (W_p - 1) \, X_p \, (1 - \frac{X_p}{P_p})$$

$$\text{Type 2:} \quad BV_2\ (\hat{X}_p) \;=\; \sum_{t\varepsilon p} (W_{pt} - 1)\ X_{pt}\ (1 - \frac{X_{pt}}{P_{pt}})$$

$$\text{Type 3:} \quad BV_3\ (\hat{X}_p) \;=\; \sum_{h\varepsilon p} (W_{ph} - 1)\ X_{ph}\ (1 - \frac{X_{ph}}{P_{ph}})$$

$$\text{Type 4:} \quad BV_4\ (\hat{X}_p) \;=\; (W_p - 1) \sum_a X_{pa}\ (1 - \frac{X_{Pa}}{P_{pa}}) \qquad (1.1)$$

where p denotes province

t denotes stratum

a denotes age-sex group (used for post-stratification and ratio estimation)

X and P denote characteristic and population totals respectively in the appropriate area and category determined by the subscripts

W denotes the theoretical sampling weight for the area identified by the subscripts

The BVs of types 1, 2 and 3 differ from each other in the levels at which the SRSs are assumed to have been applied, that is, province level, type of area levels within provinces and stratum levels, respectively. For these BVs it is assumed that simple blow-up estimates had been calculated. For Type 4 the level at which the SRS was applied is the same as for Type 1 but ratio estimates by age-sex groups are assumed to have been calculated. The binomial factors (BF) are calculated as

$$BF_i\ (\hat{X}_p) = \frac{Var\ (\hat{X}_p)}{BV_i\,(\hat{X}_p)} \quad , \quad i = 1, 2, 3, 4 \qquad (1.2)$$

where Var $(\hat{X}_p)$ is the variance of the estimate $\hat{X}_p$. The binomial factors essentially normalize a variance estimate by eliminating, either partially or completely, various effects which the population total, the characteristic level, the sample design or the estimation procedures have on the estimated variance. Generally the higher the above factors are, the worse the sample design relative to a simple random sample variance for the characteristic in

question. The restrictions on the sample design due to cost considerations, however, may result in a high variance estimate but the sample design considering the restrictions may be good. Hence, an overall evaluation of the sample design would include costs per unit as well as binomial factors.

The relative magnitude of two binomial factors can be used to assess the effect of particular aspects of the sampling scheme or estimation procedure. The ratio of BF1 to BF3 provides a rough stratification index which will measure the gain (reduction in variance) due to stratification and the examination of this ratio over time will indicate possible deterioration of the stratification. The ratio of BF1 to BF4 will indicate an approximate gain (reduction in variance) due to post-stratification by age-sex groups and subsequent ratio estimation within each post-stratum.

2. ESTIMATION OF THE BINOMIAL FACTORS

For an estimated total number of persons possessing a characteristic at the provincial level, the variance estimate of this total can be expressed as a sum of contributions by strata and subunits. From the Labour Force Survey, an estimate $\hat{X}_p$ of X_p is calculated. The estimated sampling variance is expressed as

$$\hat{V}ar\ (\hat{X}_p) = \sum_{h\epsilon p} \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} D^2_{phi}\ (\hat{X}_p) \qquad (2.1)$$

where $D_{phi}\ (\hat{X}_p) = \hat{X}_{phi} - \sum_{a} \frac{\hat{X}_{pa}}{\hat{P}_{pa}} \hat{P}_{phia} - \frac{1}{n_h} \sum_{i=1}^{n_h} (\hat{X}_{phi} - \sum_{a} \frac{\hat{X}_{pa}}{\hat{P}_{pa}} \hat{P}_{phia})$,

where $\sum_{h\epsilon p}$ signifies summation over all strata and subunits in province P and n_h units are selected from stratum h. This variance estimate was essentially developed by Keyfitz in [2].

The binomial variances introduced in Section 1, formulae (1.1) can be rewritten as

$$BV_i\ (\hat{X}_p) = \sum_{t\epsilon p} (W_{pt} - 1)\ X_{pt}\ (1 - \frac{X_{pt}}{P_{pt}}) \qquad (2.2)$$

where
for $i = 1$ t denotes the province level
for $i = 2$ t denotes type of area (NSRU and SRU) level within the province
for $i = 3$ t denotes the stratum (subunit) area level within the province
for $i = 4$ t denotes the age-sex group within the province

To obtain an estimate of these binomial variances, an estimate of the quantity $X_{pt}\ (1 - \frac{X_{pt}}{P_{pt}})$ is required for each value of t. The estimates of $X_{pt}\ (1 - \frac{X_{pt}}{P_{pt}})$ are given by

$$\hat{X}_{pt}\ (1 - \frac{\hat{X}_{pt}}{\hat{P}_{pt}}) + \frac{\hat{V}\ (\hat{X}_{pt})}{\hat{P}_{pt}} \text{ for each value of t.} \qquad (2.3)$$

These formulae were initially developed by G.B. Gray in [1]. In most cases, the term $\hat{V}\ (\hat{X}_{pt})/\hat{P}_{pt}$ is small compared with $\hat{X}_{pt}\ (1 - \frac{\hat{X}_{pt}}{\hat{P}_{pt}})$ and is usually neglected.

The relationship between binomial variances calculated under the assumption of SRS at differing area levels can be expressed as follows. Let W_p and W_{pr} denote the theoretical weights at the province level and at subprovincial area (r) levels, respectively.

$$P_r = \frac{X_{pr}}{P_{pr}}, \quad P = \frac{\sum_{r\epsilon p} X_{pr}}{\sum_{r\epsilon p} P_{pr}} \qquad W_p = \frac{P}{\sum_{r\epsilon p} \frac{P_{pr}}{W_{pr}}}$$

$$\begin{aligned}
BV_p\ (X_p) &= (W - 1)\ P_{p(1-p)} \\
&= (W - 1) \sum_{r\varepsilon p} P_{pr} p_r\ (1 - p_r + p_r - p) \\
&= \sum_{r\varepsilon p} (W_{pr} - 1)\ P_{pr} p_r\ (1 - p_r) + \sum_{r\varepsilon p} (W_p - W_{pr})\ P_{pr} p_r\ (1 - p_r) \\
&\quad + (W_p - 1) \sum_r P_r\ p_r\ (p_r - p) \\
&= \underbrace{\sum_{r\varepsilon p} BV_r\ (X_p)}_{\text{the sum of BVs for the r sub-provincial areas}} + \underbrace{\sum_r (W_p - W_{pr})\ P_{pr}\ p_r\ (1 - p_r)}_{\text{the effect of the variation in weights between the r sub-provincial areas}} + \underbrace{(W_p - 1)\ P_p Var^*(p_r)}_{\text{the effect of the variation in the proportion of persons possessing the characteristic between the r sub-provincial areas}} \qquad (2.4)
\end{aligned}$$

where $$Var^*\ (p_r) = \sum_{r\varepsilon p} \frac{P_{pr}}{P_p}\ p_r^2 - \left(\sum_{r\varepsilon p} \frac{P_{pr}}{P_p}\ p_r\right)^2$$

BV_p is the binomial variance assuming a simple random sample at province p level; BV_r is the binomial variance assuming a simple random sample at area r level. If r refers to NSRU and SRU areas, then the above equation expresses a relationship between BV1 and BV2. If p = NSRU area or SRU areas (or both combined) and each area r is a single stratum or subunit, then the above equation expresses the relationship between BV2 and BV3 (or between BV1 and BV3). It should be noted that the middle term, viz, $\sum_r (W_p - W_{pr})\ P_{pr}\ p_r\ (1 - p_r)$ could be positive or negative, while the other terms in (2.3) are always positive.

3. SPECIAL PROBLEM AREAS

High binomial factors do indicate for which characteristic estimates the variance estimates are high and by probing deeper into the subprovincial contributions to each large variance estimate, it is often possible to isolate one or more subprovincial areas which contribute to the high variance. On the basis of these identified subprovincial areas, an adjusted binomial factor can be calculated which can be used to determine whether these

subprovincial areas are the primary cause of the high variance estimate. If certain areas are consistently identified as "problem areas", then this information may be useful in a subsequent partial or complete redesign for a potential improvement in the present sample design as far as reduced variance are concerned.

3.1 Identification of Problem Areas

For each stratum or subunit, an actual percentage contribution and a desired percentage contribution are calculated and these are used to identify subprovincial areas in which the actual variance contribution is deemed to be significantly greater than the desired variance contribution.

The actual percentage contribution by stratum or subunit h of province p is defined by:

$$\text{Act.}(h:p) = \frac{\frac{n_h}{n_{h-1}} \sum_{i=1}^{n_h} D_{phi}^2}{\sum_{h\varepsilon p} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} D_{phi}^2} \times 100\text{, note that } \sum_{h\varepsilon p} \text{Act.}(h:p) = 100 \qquad (3.1)$$

A desired percentage contribution is calculated under the assumption that the binomial factor and the proportion of the population possessing the characteristic is constant over the province. In this case the desired contribution by a subprovincial area (a stratum or a subunit) is proportional to $(W_{ph} - 1) \hat{P}_{ph}$ and the variance estimate at the provincial level is proportional to $(W_{p1} - 1) \hat{P}_{p1} + (W_{p2} - 1) \hat{P}_{p2}$ (1 denotes NSRU areas, 2 denotes SRU areas) with the same factor of proportionality as for subprovincial areas. Thus, a desired percentage contribution by stratum or subunit h to the variance estimate of the provincial estimated total is defined by:

$$\text{Des.}(h:p) = \frac{(W_{ph} - 1) \hat{P}_{ph}}{(W_{p1} - 1) \hat{P}_{p1} + (W_{p2} - 1) \hat{P}_{p2}} \qquad (3.2)$$

The fact that both Act(h:p) and Des(h:p) are obtained from the results of the survey means that both are subject to sampling variability. However, if

Act(h:p) is "sufficiently greater" than Des(h:p), then there is reason to believe that stratum (or subunit) h contributes disproportionately to the variance of the provincial estimate. An attempt must now be made to elucidate the term "sufficiently greater" mentioned in the preceding sentence.

For the current Labour Force Survey in which two PSUs are selected per stratum, the contribution to the variance of the provincial estimate can be expressed as $D^2_{ph} = [D_{ph1} - D_{ph2}]^2$. On the assumption that D_{ph1} and D_{ph2} are independent, the variance of this difference can be approximated by

$$\hat{\text{Var}}\ (D_{ph1} - D_{ph2}) = \frac{(W_{ph} - 1)\ \hat{P}_{ph}}{(W_{p1} - 1)\ \hat{P}_{p1} + (W_{p2} - 1)\ P_{p2}} \sum_{h \epsilon p} D^2_{ph} \qquad (3.3)$$

Now if a normal distribution with common mean is assumed to be valid for D_{ph1} and D_{ph2} the statistic

$$\text{Test (ph)} = \frac{D_{ph1} - D_{ph2}}{\sqrt{\hat{\text{Var}}\ (D_{ph1} - D_{ph2})}}$$

has an approximate t-distribution with the number of degrees of freedom equal to the number of strata and subunits in the province.

(3.4)

Thus, at the 90% confidence level we reject the hypothesis that $E(D_{ph1}\ (\hat{X}_p) - D_{ph2}\ (\hat{X}_p)) = 0$ if the value of the t-statistic $|\text{Test (ph)}| > 1.645$. It should be noted that $\text{Test}^2\ (\text{ph}) = \frac{\text{Act (h:p)}}{\text{Des (h:p)}}$ and hence if the ratio of the actual contribution to the desired contribution is greater than 2.706, then the subprovincial area is identified as a "problem area", i.e. that the actual contribution to the variance estimate significantly exceeded the desired contribution for this subprovincial area.

3.2 Adjusted Binomial Factor

Having determined the set of strata and/or subunits for which the actual percentage contributions to the variance of the provincial estimate were deemed to be significantly in excess of the desired percentage contributions, it is possible to calculate an adjusted binomial factor which essentially reduces the large contributions by these areas to a more desirable contribution based on the contribution per element as calculated from the portions of the province where the actual contribution was more or less equal to the desired contribution.

To clarify the calculation of this adjusted binomial factor, consider the following development.

Let A denote the collection of subprovincial areas for which the actual contribution to the variance significantly exceeded the desired contribution to the variance. P\A denotes those areas of the province in which the actual and desired contributions to the variance of the provincial estimate were not significantly different (i.e. the residual portion of province p).

The contribution to the variance of the provincial estimate by area A, denoted by $\widehat{\mathrm{Var}}_A\ (\hat{X}_p) = \sum_{h\varepsilon A} \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} D^2_{phi}$ is deemed to be greater than the desired contribution according to the above test. On the other hand, the contribution to the variance of the provincial estimate by area P\A, denoted by $\widehat{\mathrm{Var}}_{p\setminus A}\ (\hat{X}_p) = \sum_{h\varepsilon p\setminus A} \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} D^2_{phi}$ is considered to be more or less what the desired contribution should be. The contribution by area A is replaced by a contribution which has the same variance per estimated person (adjusted to account for the differing sampling ratios) as area P\A, i.e.

$$\sum_{h\varepsilon A} \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} D^2_{phi} \quad \text{is replaced by} \quad \hat{P}'_A \frac{\sum_{h\varepsilon p\setminus A} \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} D^2_{phi}}{\hat{P}'_{p\setminus A}}$$

and

$$\hat{P}'_A \frac{\sum_{h\varepsilon p\setminus A} \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} D^2_{phi}}{\hat{P}'_{p\setminus A}} = \hat{P}'_A \frac{\mathrm{Act}_{p\setminus A}\ (\hat{X}_p)\ \widehat{\mathrm{Var}}\ (\hat{X}_p)}{\hat{P}'_{p\setminus A}}$$

where $\hat{P}'_R = \sum_{h\varepsilon R} (W_{ph} - 1)\ \hat{P}'_h$, for any area R

and $\mathrm{Act}_{p\setminus A}\ (\hat{X}_p) = \sum_{h\varepsilon p\setminus A} \mathrm{Act}\ (h{:}p)/100.$

while $\mathrm{Des}_{p\setminus A}\ (\hat{X}_p) = \hat{P}'_{p\setminus A}/\hat{P}'_p$

Thus the adjusted variance estimate is

$$\text{Adjusted } \hat{\text{Var}}\ (\hat{X}_p) = \hat{P}'_A \frac{\text{Act}_{p\setminus A}\ (\hat{X}_p)\ \hat{\text{Var}}\ (\hat{X}_p)}{\hat{P}'_{p\setminus A}} + \text{Act}_{p\setminus A}\ (\hat{X}_p)\ \hat{\text{Var}}\ (\hat{X}_p)$$

$$= \frac{\hat{P}'_A + \hat{P}'_{p\setminus A}}{\hat{P}'_{p\setminus A}}\ \text{Act}_{p\setminus A}\ (\hat{X}_p)\ \hat{\text{Var}}\ (\hat{X}_p)$$

$$= \frac{\text{Act}_{p\setminus A}\ (\hat{X}_p)}{\text{Des}_{p\setminus A}\ (\hat{X}_p)}\ \hat{\text{Var}}\ (\hat{X}_p) \qquad (3.5)$$

What the above formula tells us is that if the critical areas A had contributed the desired variance to the province, then automatically the actual contribution by $P\setminus A$ would have equalled the desired variance and there would have been no adjustment. However, if $\text{Act}_A\ (\hat{X}_p)$ was greatly in excess of $\text{Des}_A\ (\hat{X}_p)$, then the complement $\text{Act}_{p\setminus A}\ (\hat{X}_p) = 1 - \text{Act}_A\ (\hat{X}_p)$ would be greatly deficient compared with $\text{Des}_{P\setminus A}\ (\hat{X}_p)$ and it turns out that the adjusted variance would be reduced by a factor $\text{Act}_{P\setminus A}\ (\hat{X}_p)/\text{Des}_{P\setminus A}\ (\hat{X}_p)$ to bring it in line with a more realistic value.

The binomial variance as calculated, assuming a simple random sample within NSRU and SRU areas, remains unchanged, and thus the adjusted binomial factor can be expressed as

$$\text{Adj.BF}(\hat{X}_p) = \frac{\dfrac{\text{Act}_{p\setminus A}\ (\hat{X}_p)}{\text{Des}_{p\setminus A}\ (\hat{X}_p)}\ \text{Var}\ (\hat{X}_p)}{\text{BV}\ (\hat{X}_p)} = \frac{\text{Act}_{p\setminus A}\ (\hat{X}_p)}{\text{Des}_{p\setminus A}\ (\hat{X}_p)}\ \text{BF}(\hat{X}_p) \qquad (3.6)$$

By examining this adjusted binomial factor in comparison with corresponding binomial factors for previous surveys, it is often possible to make an assessment of whether the identified subprovincial areas were the predominant cause of the high variance estimate. These areas are often studied in greater detail to determine any unusual features which may have caused the large

4) On the basis of the binomial factor of type 2, a decision is made as to whether or not a detailed analysis of the subprovincial contribution to the variance of the provincial estimates of Employed, Unemployed and In Labour Force totals should be carried out for a particular province.

Pertaining to the time during which this analysis has been carried out (from July 1973 to the present time), the following comments can be stated:

i) The analysis is carried out much more frequently for some characteristics than others due to a more erratic behaviour for the associated binomial factors. In this category fall the characteristics Unemployed in Newfoundland, Unemployed in New Brunswick and Unemployed in British Columbia for which a detailed analysis of the subprovincial contributions to the variance was carried out in 9, 12 and 9 months, respectively, out of the total of 15 months. Detailed analyses were carried out less frequently for other characteristics and for some cases these analyses were never carried out.

ii) The subprovincial analysis identified some subprovincial areas which consistently appeared as "problem areas". The following subprovincial areas were identified as "problem areas" in 1/2 or more of the surveys for which an analysis of subprovincial contributions was carried out.

Province	Characteristic	Identification	Location
Newfoundland	Unemployed	04021 & 04025	- western part of Nfld.
New Brunswick	Unemployed	30002 & 30004	- southeast corner of N.B.
New Brunswick	Unemployed	33003 & 33005	- northeast part of N.B.
Quebec	Unemployed	41004 & 41013	- northeastern part of the Gaspe Peninsula
B.C.	Unemployed	92003 & 92013	- southern part of B.C. the Okanagan district

The remaining subprovincial areas identified as "problem areas" did not appear consistently from survey to survey and the problem areas tended not to recur for Employed and In Labour Force.

iii) From time to time a study to determine the causes of the excessive contribution by an identified "problem area" was carried out. This study generally examined weighted estimates and unweighted counts of Labour Force status by industry for half-stratum totals. In some cases there was an unequal distribution (on the basis of sample results) of persons associated with a given industry classification between the two PSUs indicating poor PSU delineation or a deterioration of the "equality" of PSUs over the time since the design of the current Labour Force Survey caused by changes in the composition of one or both PSUs. This was particularly the case for the subprovincial areas PSUs 30002 & 30004 and PSUs 33003 & 33005 in New Brunswick and PSUs 41004 & 41013 in Quebec. High Unemployment within an industry (due perhaps to seasonal factors) would then cause an excessive contribution by the pair of PSUs. For other subprovincial areas, as for example, PSUs 92003 & 92013 in British Columbia, the distribution by industry appeared relatively equal between the two PSUs but for cases examined there was, nonetheless, a tendency for the unemployment to be clustered in one of the PSUs.

5) Examination of the relationships between binomial factors BF1 and BF2 revealed several characteristics for which BF2 was less than BF1. This means that the binomial variance assuming a simple random sample by type of area is larger than the binomial variance calculated on the assumption of a simple random sample over the entire province, which intuitively does not seem reasonable. Formula (2.3) is satisfied for these binomial variances. The effect of differing weights between types of areas has a greater effect on the magnitude of BV1 than was initially anticipated. However, since the weights are the same in all of the strata within NSRU areas and in all of the subunits within SRU areas, BV3 < BV2 always and consequently BF3 > BF2 although the increase is often very slight.

6) Estimates of the four types of binomial factors for five selected characteristics at the province and Canada levels for the September 1974 survey are presented in Table 1 in the Appendix. Similar tables with estimates of the binomial factors for some 55 characteristics are available beginning with the July 1973 survey and for each successive survey.

5. ACKNOWLEDGEMENT

The author acknowledges the assistance of G.B. Gray in the preparation and writing of this article.

REFERENCES

[1] Gray, G.B. "Variance-Covariance Analysis in LFS - Report No. 1", An Internal Report, Statistics Canada.

[2] Keyfitz, N. "Estimates of Sampling Variance Where Two Units Are Selected from Each Stratum", Journal of the American Statistical Association, 52 (1957), 503-510.

[3] Sampling and Survey Research Staff, Dominion Bureau of Statistics (now Statistics Canada): "Methodology, Canadian Labour Force Survey", (1965). Catalogue No. 71-504 (occasional).

APPENDIX

Table I: Binomial Factors for Selected Characteristics for the Sept. 1974 Survey

PROVINCE	CHARACTERISTIC	BF1	BF2	BF3	BF4
Nfld.	Employed	1.82	1.75	1.81	2.67
	Unemployed	2.08	2.05	2.12	2.18
	Not in LF	1.73	1.67	1.71	2.85
	Emp. Ag.	1.32	1.30	1.32	1.18
	Emp. Non Ag.	1.85	1.79	1.85	2.71
P.E.I.	Employed	1.32	1.30	1.33	1.93
	Unemployed	2.36	2.30	2.32	2.36
	Not in LF	0.78	0.76	0.78	1.17
	Emp. Ag.	6.94	7.74	7.76	7.60
	Emp. Non Ag.	4.34	4.38	4.47	5.56
N.S.	Employed	1.57	1.52	1.56	2.34
	Unemployed	1.69	1.58	1.60	1.73
	Not in LF	1.39	1.35	1.38	2.13
	Emp. Ag.	2.02	2.43	2.47	2.04
	Emp. Non Ag.	1.39	1.34	1.38	2.03
N.B.	Employed	1.43	1.37	1.39	2.19
	Unemployed	1.91	1.83	1.86	1.94
	Not in LF	1.39	1.35	1.36	2.24
	Emp. Ag.	0.83	0.98	1.01	0.84
	Emp. Non Ag.	1.41	1.36	1.38	2.11
Que.	Employed	0.97	0.94	0.96	1.52
	Unemployed	1.57	1.51	1.54	1.60
	Not in LF	0.92	0.84	0.91	1.54
	Emp. Ag.	3.23	4.27	4.35	3.27
	Emp. Non Ag.	1.02	0.99	1.02	1.54
Ont.	Employed	0.93	0.90	0.92	1.41
	Unemployed	1.54	1.47	1.50	1.55
	Not in LF	0.85	0.83	0.85	1.34
	Emp. Ag.	1.69	2.23	2.34	1.71
	Emp. Non Ag.	0.93	0.91	0.94	1.38

PROVINCE	CHARACTERISTIC	BF1	BF2	BF3	BF4
Man.	Employed	0.86	0.83	0.85	1.31
	Unemployed	1.53	1.41	1.45	1.54
	Not in LF	0.94	0.91	0.93	1.45
	Emp. Ag.	2.29	3.39	3.50	2.42
	Emp. Non Ag.	1.60	1.63	1.67	2.16
Sask.	Employed	1.91	1.83	1.88	2.88
	Unemployed	1.69	1.45	1.47	1.71
	Not in LF	1.99	1.92	1.97	3.05
	Emp. Ag.	2.14	2.77	2.87	2.43
	Emp. Non Ag.	2.26	2.27	2.35	2.74
Alta.	Employed	1.48	1.43	1.46	2.19
	Unemployed	1.88	1.76	1.79	1.89
	Not in LF	1.32	1.28	1.30	1.97
	Emp. Ag.	3.17	4.53	4.73	3.32
	Emp. Non Ag.	1.79	1.81	1.86	2.36
B.C.	Employed	1.17	1.13	1.15	1.76
	Unemployed	1.55	1.53	1.56	1.59
	Not in LF	1.01	0.97	1.00	1.61
	Emp. Ag.	2.46	2.81	2.90	2.47
	Emp. Non Ag.	1.27	1.23	1.26	1.88
Can.	Employed	1.04	1.00	1.03	1.59
	Unemployed	1.58	1.51	1.54	1.60
	Not in LF	0.96	0.94	0.96	1.54
	Emp. Ag.	2.44	3.25	3.37	2.52
	Emp. Non Ag.	1.11	1.09	1.12	1.64

SOME ESTIMATORS FOR DOMAIN TOTALS

M.P. Singh and R. Tessier

A major concern in large scale surveys is the problem of sub-population estimation (domain estimation). This paper presents a study of four estimators for estimating domain totals. The domain considered in the study is an area type of domain, that is, a domain consisting of a combination of a certain number of area units belonging to different strata. This paper uses some actual data and some fictitious data to compare variances and mean square errors of the four estimators.

1. INTRODUCTION

In large scale nationwide surveys, estimates are often required for certain 'domains' in addition to the overall estimates at the national and provincial levels. Domains may be specified by classification characteristics (such as employed by age group) or by geography such as groups of primary sampling units (PSUs) in an area frame. In general no new theory is needed for domain estimation (see for example Cochran [1], Murthy [2]) and the situation is similar to sampling from a universe (the domain) known to include extra units not belonging to the universe under consideration. The basic principle used in domain estimation is that the probability sample taken from the entire universe would also serve as a probability sample taken from the domain provided that the units in the sample not belonging to the domain are assumed to have zero value for the character under study. The estimator T_1 given in the following section is the usual estimator of total for the domain.

The technique of post-stratification may also be applied for estimating domain totals when the actual number of units belonging to the domain is known. This additional information would usually be available for domains defined on the basis of geography where the units under consideration are area units. Using this technique, some alternative estimators are presented in this paper. Efficiency comparison of these estimators are made using data from the Canadian Labour Force Survey, and also from a set of fictitious data.

2. ALTERNATIVE ESTIMATORS AND THEIR VARIANCES

2.1 Notations

D: domain

D_I: the area belonging to D which falls in stratum i

y_{ij}: total of the study variable y for j th sampling unit of i th stratum

x_{ij}: known size measure, corresponding to y_{ij}

N_i: total number of sampling units in stratum i

n_i: total number of selected sampling units

Throughout this paper we shall assume that the domain D is composed of complete sampling units (SU).

Further, $X_i = \sum_{j=1}^{N_i} x_{ij}$, $p_{ij} = x_{ij}/X_i$ and the new variable

$$y'_{ij} = \begin{cases} y_{ij} & \text{if } j \in D_i \\ 0 & \text{otherwise;} \end{cases} \qquad (2.1)$$

we define Y_{Di} = total for characteristic y corresponding to area D_i and

$$Y_D = \Sigma_i \, Y_{Di} = \Sigma_i \sum_{j=1}^{N_i} y'_{ij} \, .$$

Note that D_i may be complete stratum i or a portion of stratum i. Using these notations we give the estimators in the following sub-sections. We assume that SUs are selected with PPS and with replacement.

2.2 Estimator T_1

An unbiased estimator used in domain estimation for the area D_i is given by

$$T_{1i} = \sum_{j=1}^{n_i} y'_{ij}/n_i p_{ij} \qquad (2.2)$$

and summing over i the unbiased estimator for the entire domain D is

$$T_1 = \Sigma_i T_{1i} \quad . \tag{2.3}$$

Note that, if all SUs of stratum i are included in the domain D, then D_i is the complete stratum i; furthermore, the sampling in two different domains is independent. It is easy to see that T_1 is an unbiased estimator for the domain total Y_D with variance

$$V(T_1) = \Sigma_i \; V(T_{1i}) \quad , \tag{2.4}$$

where

$$V(T_{1i}) = (\sum_{j=1}^{N_i} y'^2_{ij}/P_{ij} - Y^2_{Di})/n_i \quad . \tag{2.5}$$

An unbiased variance estimator is

$$\hat{V}(T_1) = \Sigma_i \; \hat{V}(T_{1i}) \quad , \tag{2.6}$$

where

$$\hat{V}(T_{1i}) = \{ \sum_{j=1}^{n_i} (y'_{ij}/p_{ij})^2 - n_i \; T^2_{1i} \}/n_i (n_i - 1) \quad . \tag{2.7}$$

Note that the expressions for $V(T_{1i})$ and $\hat{V}(T_{1i})$ contain actual numbers of SUs selected from stratum i (which is n_i) and not the number of sampled SUs belonging to D_i.

2.3 Estimator T_2

Since the SUs belonging to domain D_i can be identified, the number of SUs belonging to the domain D_i is known (N_{Di} , say), and the number of SUs belonging to the complete domain D is $\Sigma_i N_{Di} = N_D$ (say).

Let us define the variable δ_{ij} as

$$\delta_{ij} = \begin{cases} 1 \text{ if } j \in D_i \\ \\ 0 \text{ otherwise .} \end{cases} \tag{2.8}$$

Then, we have that

$$N_{Di} = \sum_{j=1}^{N_i} \delta_{ij} .$$

An unbiased estimator of N_{Di} is given by

$$\hat{N}_{Di} = \sum_{j=1}^{n_i} \delta_{ij}/n_i \, p_{ij} \tag{2.9}$$

and summing over i we get an unbiased estimator of the number of SUs belonging to the domain D,

$$\hat{N}_D = \Sigma_i \hat{N}_{Di} . \tag{2.10}$$

With this information, we define a combined ratio estimator

$$T_2 = T_1 N_D/\hat{N}_D \tag{2.11}$$

where T_1 is the usual unbiased domain estimator defined in (2.3).

This estimator has the usual ratio estimator bias which is approximately given by

$$B_2 = Y_D \{V(\hat{N}_D)/N_D^2 - \mathrm{Cov}\ (T_1, \hat{N}_D)/Y_D\ N_D\} .$$

Note that

$$V(\hat{N}_D) = \Sigma_i\ V(\hat{N}_{Di}) \tag{2.12}$$

with

$$V(\hat{N}_{Di}) = (\sum_{j=1}^{N_i} \delta_{ij}/p_{ij} - N_{Di}^2)/n_i \tag{2.13}$$

and

$$\mathrm{Cov}\ (T_1, \hat{N}_D) = \Sigma_i\ \mathrm{Cov}\ (T_{1i}, \hat{N}_{Di}) \tag{2.14}$$

with

$$\text{Cov }(T_{1i}, \hat{N}_{Di}) = \sum_{j=1}^{N_i} (y'_{ij}/p_{ij} - Y_{Di}\, N_{Di})/n_i \qquad (2.15)$$

and thus B_2 can be expressed as

$$B_2 = (-1/N_D)\; \Sigma_i \; \{\sum_{j=1}^{N_i} (y'_{ij} - \overline{Y}_D \delta_{ij})/P_{ij} - N_{Di}(Y_{Di} - N_{Di}\overline{Y}_D)\}/n_i , \qquad (2.16)$$

where

$$\bar{Y}_D = Y_D/N_D \; . \qquad (2.17)$$

To find an expression for the variance of the estimator we use the usual approximation

$$V(T_2) = V(T_1) + \bar{Y}_D^2 \; V(\hat{N}_D) - 2\bar{Y}_D \; \text{Cov }(T_1\, \hat{N}_D) \; .$$

Thus from equation (2.4) and (2.12) to (2.15)

$$V(T_2) = \Sigma_i \; \{\sum_{j=1}^{N_i} (y'_{ij} - \bar{Y}_D\, \delta_{ij})^2/P_{ij} - (Y_{Di} - N_{Di}\, \bar{Y}_D)^2\}/n_i \; . \qquad (2.18)$$

The variance estimator can be written as

$$\hat{V}(T_2) = \Sigma_i \; \{ \sum_{j=1}^{n_i} (y'_{ij} - \delta_{ij}\; T_1/\hat{N}_D)^2/n_i \; p_{ij}^2 \qquad (2.19)$$

$$- (T_{1i} - \hat{N}_{Di}\; T_1/\hat{N}_D)^2\}/(n_i - 1) \; .$$

For the sake of comparing T_2 with other estimators, we consider its mean square error (MSE).

$$\text{MSE}(T_2) = V(T_2) + B_2^2 \qquad (2.20)$$

It may be pointed out that if the sample sizes (n_i) in the strata are large enough such that the usual ratio estimator approximations are valid at the stratum level, then the following estimator can be used

$$T_2' = \Sigma_i \; T_{2i}' \qquad (2.21)$$

where

$$T'_{2i} = \begin{cases} T_{1i} \; N_{Di}/\hat{N}_{Di} & \text{if } n_{Di} \geq 1 \\ 0 & \text{if } n_{Di} = 0 \end{cases}$$

T_{1i} and $\hat{N}_{Di}$ being defined as in equation (2.2) and (2.9) respectively, and n_{Di} being the number of sampled SUs of stratum i falling in domain D_i.

Again, using the approximate expression of the bias given for the ratio estimator we find that the bias of T'_{2i} is

$$B'_{2i} = (-1/n_i \; N_{Di}) \sum_{j=1}^{N_i} (y'_{ij} - \bar{Y}'_{Di} \; \delta_{ij})/n_i P_{ij} \tag{2.22}$$

where

$$\bar{Y}_{Di} = Y_{Di}/N_{Di} \tag{2.23}$$

and the bias for T'_2 is thus

$$B'_2 = \Sigma_i \; B'_{2i}. \tag{2.24}$$

In this case

$$V(T'_2) = \Sigma_i \; V(T'_{2i}) \tag{2.25}$$

where $$V(T'_{2i}) = \sum_{j=1}^{N_i} (y'_{ij} - \bar{Y}_{Di} \; \delta_{ij})^2/n_i \; p_{ij}, \tag{2.26}$$

from equation (2.5), (2.13) and (2.15) and $\bar{Y}_{Di}$ being defined in (2.23).

The estimator of the variance is given by

$$\hat{V}(T'_2) = \Sigma_i \; \hat{V}(T'_{2i}) \tag{2.27}$$

where $$\hat{V}(T'_{2i}) = \sum_{j=1}^{n_i} (y'_{ij} - \delta_{ij} \; T_{1i}/\hat{N}_{Di})^2/n_i(n_i-1) \; p_{ij}^2 \tag{2.28}$$

and the mean square error is given by

$$\mathrm{MSE}(T_2^{'}) \quad = \quad V(T_2^{'}) \quad + \quad B_2^{'2} \tag{2.29}$$

Estimator $T_2^{'}$ is not compared to the other estimators in section 3, Empirical Study.

2.4 Estimator T_3

This is a post stratified type estimator where the weights for the post strata are assumed to be known. In case of estimators T_1, T_2 and $T_2^{'}$, the contribution from the domain D_i that does not contain any sampled SUs is zero (ie. $T_{1i} = T_{2i}^{'} = 0$ if no sampled SU belongs to D_i). This may not be a desirable situation particularly when D_i contains large numbers of SUs of stratum i but not those in the sample. The following estimator T_3 (and also T_4) avoids zero contribution from any D_i (unless however the Y_{Di} is itself zero) and depends upon the stratum aggregate estimates.

T_3 is defined as

$$T_3 \quad = \quad \Sigma_i \, T_{3i} \;, \tag{2.30}$$

where

$$T_{3i} \quad = \quad T_i X_{Di}/X_i \;, \tag{2.31}$$

$$X_{Di} \quad = \quad \sum_{j=1}^{N_i} \delta_{ij} \, x_{ij}, \; X_i \quad = \quad \sum_{j=1}^{N_i} x_{ij}$$

and

$$T_i \quad = \quad \sum_{j=1}^{n_i} y_{ij}/n_i^{'} \, p_{ij} \;, \tag{2.32}$$

T_i being the usual PPS with replacement estimator of stratum i. Thus the estimator T_{3i} is actually obtained by deflating the estimator T_i according to the proportion (in terms of the size variable) by which the stratum is included in D_i.

This estimator is biased, its bias being

$$B_3 = E(T_3) - Y_D = \Sigma_i \quad (Y_i X_{Di}/X_i - Y_{Di}) \tag{2.33}$$

and Y_i is the stratum total.

By looking at the relative bias

$$(E(T_3) - Y_D)/Y_D = \Sigma_i \ (X_{Di} \ Y_i/Y_{Di} \ X_i - 1) \tag{2.34}$$

we see that the bias is small if the size variable is highly correlated with the variable under study.

The variance of the estimator is

$$V(T_3) = \Sigma_i \ (X_{Di}/X_i)^2 \ V(T_i) \tag{2.35}$$

since size variables are known constants. In the same way, the estimated variance is found by using the estimated variance at the stratum level. The mean square error of estimator T_3 is given by

$$MSE(T_3) = V(T_3) + B_3^2 \tag{2.36}$$

2.5 Estimator T_4

This estimator is a special case of T_3. If the size of the SUs do not differ very much from one to another, then, one can replace X_{Di}/X_i by N_{Di}/N_i in T_3 which means that the stratum total is deflated using the proportion of SUs that are in the domain D_i.

The estimator becomes

$$T_{4i} = T_i \ N_{Di}/N_i \ , \tag{2.37}$$

where T_i is defined in (2.32),

and

$$T_4 = \Sigma_i \ T_{4i}. \tag{2.38}$$

The bias of the estimator has the same form as for T_3 and is given by

$$B_4 = E(T_4) - Y_D = \Sigma_i \ (Y_i N_{Di}/N_i - Y_{Di}). \tag{2.39}$$

The bias will be small if the variable under study is more or less uniform from one SU to another and if the size of the SUs do not differ very much from one to another. In such a case, the advantage of T_4 over T_3 is at the computation level.

The variance is given by

$$V(T_4) = \Sigma_i \ (N_{Di}/N_i)^2 \ V(T_i) \tag{2.40}$$

and the estimated variance is found by using the estimated variance at the stratum level. The mean square error of the estimator is given by

$$MSE(T_4) = V(T_4) + B_4^2 \tag{2.41}$$

Remarks

A more general estimator of which T_3 and T_4 are particular cases may be written as

$$T = \Sigma_i \ W_i \ T_i$$

where W_i is any known suitable deflating factor.

3. EMPIRICAL STUDY

3.1 Example 1: Labour Force Data

From the province of New Brunswick, the following four domains were formed for the purpose of this study:

Domain No. 1: Western half of the province,
Domain No. 2: Southern half of the province,

Domain No. 3: Along Chaleur Bay and the Gulf of St. Lawrence,
Domain No. 4: Center of the province.

Table 1 gives the number of complete strata, the number of incomplete strata and the number of selected PSUs for each domain of the frame used in the Labour Force Survey (LFS). It may be noted that in LFS two PSUs are selected from each stratum and that sub-sampling is done within selected PSUs.

(Table 1)

Using the data from survey 274 (April 1973) at the PSU level, the estimate of the three main characteristics of the Labour Force (unemployed, employed and not in the LF) were calculated for each domain using the four different estimators. Table 2 gives the estimated coefficient of variation (C.V.) in percent of the three main characteristics.

(Table 2)

The following conclusion may be drawn from Table 2:

1. T_2, T_3 and T_4 have much smaller C.V. than T_1 for all domains. The comparison is however not fair since for T_2, T_3 and T_4 only variances have been considered instead of the MSE in obtaining their coefficient of variation. Size of bias would be necessary to have fair comparisons (see example 2).

2. T_3 and T_4 follow each other very closely, as expected, since PSU sizes do not vary much (the average C.V. of the mean of the size variable at the stratum level is 0.97%).

3. C.V. of T_3 and T_4 is smaller than that of T_2 except for two cases (domain 4, unemployed and not in LF).

3.2 Example 2: Fictitious Data

For a more realistic comparison between estimators, it is necessary to have an idea of the bias of T_2, T_3 and T_4. To achieve this a random population was generated with the help of a table of random numbers. Four strata were considered with the number of sampling units (SU) as given in Table 3. In each stratum, the size of the first SU was randomly selected between 10 and 90, the size of the other SUs in the stratum were obtained by using random

numbers that were within plus or minus 5% of the size of the first SU (this gives an expected coefficient of variation at the stratum level of approximately 1%). Three study variables were formed assuming linear regression with the size variable: the first variable (noted as y in Table 4) was randomly selected to have an expected correlation of 0.75 with the size variable. The second variable (noted as z in Table 4) was randomly selected to have an expected correlation of 0.50 with the size variable. Finally, the third variable (noted as u in Table 4) was randomly selected to have an expected correlation of 0.25 with the size variable. Then, the SUs of the four strata were allocated at random to two mutually exclusive domains. The total number of SUs in that part of the domain under study falling in each stratum (D_i) are also given in Table 3.

(Table 3)

True variance, bias and mean square error were calculated for each estimator assuming a sample of size two in each stratum. Table 4 gives, as a percent of the domain total, the bias (Rel Bias), the standard deviation (C.V.) and the square root of the mean square error (Rel Error) of the four estimators, for each study variable. Domain totals are also given on the last line of the table.

(Table 4)

The following points may be noted from Table 4:

1. T_4 does not have a significantly different Rel Error than T_3 for any of the three variables.
2. T_2 has lower Rel Errors than T_1, and both T_3 and T_4 have lower Rel Errors than T_2, with a maximum gain for variable Y which has the highest correlation with the size variable.
3. Comparing biases of T_2, T_3 and T_4, T_2 has smallest bias for variable u, T_3 for variable y and T_4 for variable 2.
4. Bias of T_3 decreases with increase in correlation between size variable and study variable. This trend is not evident for T_2 and T_4, possibly because of the use of variable $\hat{N}_D$ (N_D) instead of X_D, as used in T_3.

Table 1: Domain Composition

Domain Number	Complete Strata	Incomplete Strata	Selected PSUs
1	5	3	13
2	2	5	11
3	2	4	10
4	0	4	2

Table 2: Percent Coefficients of Variation

Estimator	Characteristics	Domain 1	Domain 2	Domain 3	Domain 4
1	Unemployed	33.65	31.35	17.62	71.80
	Employed	15.19	19.16	15.21	82.39
	Not in LF	14.39	18.60	19.25	70.79
2	Unemployed	25.71	22.25	12.01	7.04
	Employed	8.24	6.48	13.98	29.25
	Not in LF	6.06	5.17	11.51	4.25
3	Unemployed	12.61	20.46	9.82	24.31
	Employed	5.98	3.73	8.13	12.35
	Not in LF	4.99	2.83	7.24	5.65
4	Unemployed	12.57	21.23	9.88	23.82
	Employed	5.98	3.56	8.39	12.09
	Not in LF	4.93	2.58	7.39	5.40

Table 3: Fictitious Population

Stratum Number	No. of SUs In Stratum	No. of SUs in D_i
1	6	3
2	8	2
3	8	7
4	9	4

Table 4: Relative Errors and Components[a]

Estimator	Parameters	Variable Y	Variable Z	Variable U
1	Rel Bias	0.00	0.00	0.00
	C.V.	28.57	28.71	29.64
	Rel Error	28.57	28.71	29.64
2	Rel Bias	0.75	0.76	0.50
	C.V.	5.94	6.84	6.98
	Rel Error	5.99	6.88	7.00
3	Rel Bias	-0.02	0.93	-1.18
	C.V.	2.14	3.09	4.75
	Rel Error	2.14	3.22	4.89
4	Rel Bias	-0.50	0.44	-1.66
	C.V.	2.13	3.07	4.72
	Rel Error	2.19	3.10	5.00
	Total	1173	1217	1151

a: all figures except totals are in percentages.

REFERENCES

[1] Cochran, W.G., "Sampling Techniques", Second Edition, New York: John Wiley and Sons, Inc., 1963, pp. 33-38.

[2] Murthy, M.N., "Sampling Theory and Methods", Calcutta: Statistical Publishing Society, 1967, pp. 77-78.

SAMPLE DESIGN OF THE FAMILY EXPENDITURE SURVEY (1974)

M. Lawes and G.B. Gray

In order to monitor changes in expenditure patterns and, if necessary, provide information for a reweighting of the Consumer Price Index, family expenditure surveys have been carried out at approximately two year intervals since 1953.

While all of the Family Expenditure Surveys have utilized the Canadian Labour Force Survey [1] frame, the particular survey in 1974 was designed somewhat differently from earlier surveys in that segments or city blocks were specially selected for the survey and there was strict control on the sample size not adhered to in earlier surveys.

The sample design, from the considerations based on the broad requirements of the survey to the details of the sampling procedures, is described in this article.

1. INTRODUCTION

The Family Expenditure Survey program is designed to collect information on the money transactions of families and unattached individuals living in private dwellings within the cities covered by the sample. It consists of two phases: the collection, by means of monthly record-keeping surveys throughout the reference year, of detailed information on family food expenditures (and selected non-food items); and the collection of information by recall of all family expenditure, income and change in assets and liabilities for the reference year, in a survey carried out at the beginning of the following year. These programs have been carried out by the Family Expenditure Section, Consumer Income and Expenditure Division at approximately two year intervals over the past twenty years, but the record-keeping phase has not featured in all the survey programs. The sample frame and sample selection has always been based on the Canadian Labour Force Survey [1]. Until a few years ago, a sub-sample of rotated out households was used for the survey provided that they had been rotated out for at least 6 months to a year. Increased pressure on methodologists to alleviate response burden has meant searching for alternative sampling procedures that avoid using households that have been in LFS. In one procedure, future random starts in selected areas that would have yielded LFS households were used for the Family

Expenditure Survey and these random starts were by-passed in LFS as the sample for LFS rotated every 6 months in selected areas. The chief drawback of this procedure was that extra listing of sample areas is required as a result of the areas rotating out earlier. Also, with revisions in self-representing units to compensate for size measures being out of data (introduced about 1967), there has been considerable variation in expected takes between strata, although the revisions have removed much of the variance of sample take within strata. The uneven growth rates within cities resulted in some difficulties in controlling the total sample size, especially when a sample of sub-units was employed. Because data was required at individual city level, fairly strict control of the sample size at city levels was required in order to avoid the risk of extreme deviations from the desired sample size. In order to control the sample size strictly at the city level, sub-sampling of samples based on reserved random starts would have been required because of different sampling rates at the segment or city block level. It was not feasible to select systematic samples at different rates to those used in LFS without clashing with LFS sampled households and there would have been an extensive field problem to select a specified quota from residual households in a list. With these difficulties noted, the 1974 Family Expenditure Survey sample was derived by selecting entirely new segments and selecting a pre-determined number of households within, using the most up-to-date household count as the size measures. Fourteen cities across Canada were represented in the sample. Some economies of training and supervising were realized by utilizing 8 Regional Office centres in the group of 14 cities.

2. DESCRIPTION OF SURVEY

The 1974 program consisted of two parts, viz., a diary survey conducted in each month of 1974 and a Recall Survey conducted early in 1975, but referring to the calendar year 1974, to complement the data derived from the diary survey. In the same segments where the diary and recall samples were selected, two other samples of the same size were drawn by utilizing different random starts for systematic selection. Families and unattached individuals in these samples were then screened on the basis of family size and income in a first phase interview and units selected were asked to complete diaries or

questionnaires in the second phase identical to those used in the main survey. These samples were thus used to over-sample families and unattached individuals with specific characteristics in both diary and recall surveys.

For the diary survey, respondents from the regular sample, and those selected from the special sample, were asked to complete diaries of items purchased for a two week period. In the recall survey, similar groups of families and unattached individuals from both the regular and screened samples provided details of their expenditures, incomes and changes in assets and debt in 1974 during a lengthy interview.

3. CONSIDERATIONS FOR SAMPLE DESIGN

For the diary survey annual data was much more important than monthly or even quarterly data although some comparisons were to be made between quarters. For this reason the sample was spread out over two 6-month periods by interviewing in different segments each month of each period. It is believed that correlations between data within the same area 6 months apart are considerably lower than they are 1 month apart, although there has been no empirical demonstration of this. Therefore, neighbouring households were avoided in the sample up to 5 months apart. However, a duplicate sample for "the screening survey" required comparisons with the regular sample but with independent estimates so that households in the same segments were used in the same months. Since integration of the results from the Recall Survey and the Diary Survey were to be made, it was desirable for the two samples to be in the same segments. For the most efficient linkage between data of the two samples, identical households should have been used, but response burden rules out the use of sampled households more than once. The use of households in other surveys such as Revised Labour Force Survey could be eliminated to remove response burden, but because of the difficulty of handling the sample control in two distinct sample frames, it would be less trouble to allow the duplicate sampling in the few cases that do occur. If there is a danger of complaints, the households could be eliminated from interviewing in either survey without any undue effect on the overall response.

3.1 Sample Size

For the diary sample, a monthly sample of 585 households was distributed among the 8 Regional Office cities and 6 other cities, viz., St. John, Quebec City, Thunder Bay, Regina, Saskatoon and Calgary. The size of the Recall Survey was about 7,000 households or about 12 times the monthly diary sample. The first phase sample for the screening survey was also 585 households per month for the diary and 7,000 for the recall.

3.2 Overview of Sample Design

To obtain a representative sample of the city inhabitants and to possess a further degree of control over the sample sizes, the dwellings were stratified by type of area - hard core, fringe, and apartment dwellings, with the selection procedure within each type of area essentially the same. The sample frame was the Labour Force Survey frame which was drawn up subsequent to the 1961 census with segment dwelling counts updated periodically. The most recent dwelling counts were used to obtain the various measures of size required to select the sample. Segments selected for the Family Expenditure Food Diary Survey, with a very few exceptions, neither had been selected for the Labour Force Survey in the recent past, nor would have been selected for the Labour Force Survey or any other survey in the future.

4. GENERAL DESCRIPTION OF THE SAMPLE DESIGN

The description pertains to the selection of the months for the regular diary sample, however, the same segments were used in the Recall Survey and the screened sample simply by using different random starts in each selected segment for the different surveys. Whenever random starts were used up, new segments were selected.

Segments selected for the first 6 months of 1974 were used for the second 6 months of 1974, i.e., households for the July survey were selected within the same segments as for the January survey simply by utilizing different random starts. The same segments that were used in the February survey were also used in the August survey, and so on.

In order to explain the sampling procedure, it is necessary to describe briefly the self-representing unit frame of the Labour Force Survey. Each city has been subdivided into subunits, which are really contiguous strata of city blocks or groups of city blocks called segments. The segments are divided into 6 or a multiple of 6 groups denoting the months of household rotation. These groups could be hard core (little potential for growth except through demolition and urban renewal), or fringe (potential for growth by urban development). Most subunits are of one type or another but some possess both types. Many cities also have apartments subunits comprising segments defined by large apartment blocks. Smaller apartment buildings, however, are included in the regular hard core and fringe subunits. More details are presented in [1].

The interview groups within the subunits which may be denoted as subunit-groups were divided into 3 strata in each city - hard core, fringe and apartment. The selection of the households was undertaken in 3 stages, as follows:

1) Selection of the required number of subunit-groups.
2) Selection of 1 segment, (or 2 or more segments if the subunit-group was selected twice or more) from within each selected subunit-group.
3) Selection of 2 systematic samples of households from within each selected segment; one sample for the first month the sample is introduced, the other sample to be introduced 6 months later. Additional systematic samples were obtained from the segments for the Recall Survey and Low Income probe; ultimately, 8 random starts were used. Occasionally all the random starts were used up in a segment and it had to be replaced.

4.1 Selection of the Subunit-Groups

Due to differing growth rates and sometimes deterioration, revisions have been made to subunits to reflect these changes. This has resulted in various expected takes for subunits and rotation groups within such units. The selection of subunit-groups must incorporate these differences. In each type of area, the sizes for subunit-groups were tallied and from these measures, estimates of the number of dwelling units by type of area within a city were made.

A proportional allocation by type of area within each city was obtained on the basis of the total pre-determined monthly sample size. The number of subunit-groups to be selected equalled the number of segments to be selected as indicated above except when a subunit-group was selected more than once. The number of subunit-groups counting repetitions was determined by assuming a preliminary density factor (number of dwellings to be selected per segment) of 3 in the hard core, 2 in the fringe areas and 10/3 in the apartment segments. 10/3 was the density factor used in LFS apartment frame but density factors varied considerably between 1.5 and 6 for the fringe and hard core subunit-groups, while a somewhat constant density factor was desired for FEX and the density factors respectively of 2 and 3 were judged to be the most appropriate for fringe and hard core segments. Since a fixed integral number of segments was to be selected, the density factors were adjusted to ensure that the sample size would correspond to the size according to the proportional allocation. This procedure also ensured a self-weighting sample ignoring the effect of adjustments for non-response. More details of the procedure as well as the actual sample allocation by city and type of area are provided later on.

The total number of subunit-groups (which is equal to the number of segments) required for the year was equal to 6 times the number required for a particular month. These subunit-groups were selected systematically with probability proportional to size (number of dwellings) with the subunit-groups randomly assigned to the samples for one pair of months, i.e., for January and July, or February and August, etc. In some cases subunit-groups were selected for 2 or more pairs of months and in this case 2 or more segments were selected at the next stage. The dwelling unit counts were recorded for each subunit-group and the selection was undertaken manually.

4.2 Selection of the Segments

One segment was selected in each subunit-group with probability proportional to size (most recent dwelling count). If a subunit-group was selected twice, then two segments were selected with probability proportional to size. It was necessary to avoid LFS segments as much as possible. For LFS, one segment is selected with probability proportional to size (again, most recent dwelling count and re-apportioned so that the total size for each subunit-

group equals the theoretical weight for the province, e.g., 300 in Quebec s.r.u.). The random start pertaining to segments as of a particular survey in LFS may be determined from Sample Control and used as a guide for avoidance of the LFS segment in FEX samples. In our case, the random start pertaining to Survey 280 (Oct. 1973) was used as the reference point. The FEX segment corresponded to the random start for Survey 280 + 1/3 (theoretical weight). Wherever two segments were required in a subunit-group, random start + 1/3 (theoretical weight) and random start + 2/3 (theoretical weight) were calculated to select the two segments. For more than 2 segments, other fractions of the theoretical weight such as 1/5, 2/5, 3/5, 4/5 were employed in the case of 4 segments or in the case of 3 segments where one had to be rejected because it clashed with LFS or another survey. Occasionally, a segment is so large that no choice was left but to select it in both FEX and LFS. In such a case, as we shall see in the next section, a set of households was selected for FEX that were distinct from LFS and never to be used in LFS.

4.3 Selection of the Households

Using the most recent dwelling count of selected segments and factors determined by the allocation formula, the step intervals were determined and random starts obtained for the first month the segments were in the sample. The random starts for the second month the segments were in the sample were simply the random starts for the first month, plus 2.

The segments selected for the Family Expenditure Food Diary Survey which had also been selected for the Labour Force Survey, were sub-sampled so that the expected number of households to be sampled in the segment was equal to the value determined previously. To accomplish this, a random start within the Labour Force segment was reserved strictly for the FEX and on the basis of the density factor, subsampling of this systematic sample yielded the sample for the Family Expenditure Food Diary Survey.

5. FURTHER DETAILS OF THE SAMPLE DESIGN

5.1 Allocations of the Sample

Suppose that in a particular city n dwelling units would be desired every month for a FEX sample. The problem remained to allocate the sample among

1) hard core, 2) fringe, and 3) apartment samples and maintain a self-weighting sample in each city.

Let D_j = number of dwelling units in type of area j (1, 2 or 3 for hard core, fringe, and apartment areas, respectively) as obtained from the list of subunits and interview groups containing all the segments.

The theoretical weight (monthly) for FEX survey is given by $(D_1 + D_2 + D_3)/n$; i.e., the sum of the counts of the dwelling units by type of area divided by the monthly sample size n (arbitrarily set by Family Expenditure Section in consultation with Household Surveys Development Staff).

The sample size was allocated among the types of areas proportional to the estimated dwelling unit count or $n \dfrac{D_j}{D_1 + D_2 + D_3} = n_j$ so that the weight D_j/n_j was the same for each type of area in a given city.

Approximate density factors G_j were assigned by type of area and these were 3, 2, and 10/3 in the hard core, fringe, and apartment areas, respectively. The number of segments to be selected in type of area j was calculated by $m_j = [n_j/G_j + .5]$, where [ ] denotes integral value and if $m_j = 0$ by the formula, then m_j was adjusted to the value 1.

5.2 Selection of Subunit-group (detailed description)

Subunit-groups were selected in the following manner:

1) D_{hg} = dwelling unit count estimated for subunit h and interview group g (g = 1 to 6 and sometimes repeated if 12 segments were selected in a particular subunit).

2) In each city and type of area, a list of subunit-groups with D_{hj} was prepared in the following manner:

Subunit-groups	D.U. Count	Accumulated D.U. Count	Selection of Subunit-group
XXXXX-X	XXXX	XXXXX	

3) For systematic selection, the dwelling unit counts were accumulated.

4) Six times the number of subunit-groups required per month, say 6m, were selected systematically with probability proportional to size by taking a random number between 1 and $\sum_h \sum_g D_{hg}/6m$ and adding $\sum_h \sum_g D_{hg}/6m$ in the usual manner of systematic selection. Repeated selections were permitted and if a subunit-group was selected twice, two segments were selected from the subunit-group as described in the next section. If selected three times, three segments were selected. The 6m subunit-groups were then systematically assigned to one of the 6 samples of subunit-groups required for the monthly surveys.

5.3 Selection of Segments

One (and sometimes two) segment(s) was (were) selected by estimating the accumulated size relevant to Survey 280 for LFS and adding to this value 1/3 (LFS theoretical weight, e.g., 300 in Quebec s.r.u.), and 2/3 (LFS theoretical weight) and selecting the appropriate segment(s).

The accumulated size for Survey 280 was estimated by adding to the accumulated size of the previous segment the number of partial or complete half-years between the introductory survey data of the segment and Survey 280. The selection is explained in more detail below.

Let S be the selected segment of a particular subunit-group, C_{s-1} be the accumulated size (accumulated inverse sampling ratios), up to and including segment s-1, and M_s be the survey when segment s was introduced. Here, segment s-1 was in LFS for the last time in survey M_{s-1} and segment s was sampled for the first time in survey M_s.

Now let C = accumulated size relevant to Survey 280

$C = C_{s-1} + [\frac{280 - M_s}{6}] + 1$, where [ ] denotes integral portion of calculated value.

Suppose now that k segments were required in the subunit-group, then segments i_1, i_2, i_3, ... i_k were selected such that

$$C_{i_1} - 1 < C + [\frac{W}{k+1} + .5] \pmod{W} \le C_{i_1}$$

$$C_{i_2} - 1 < C + 2 [\frac{W}{k+1} + .5] \pmod{W} \le C_{i_2}$$

$$C_{i_k} - 1 < C + k [\frac{W}{k+1} + .5] \pmod{W} \le C_{i_k}$$

where W = theoretical weight in LFS for SRU areas of a particular province.

A value of K + 1 instead of K was required since essentially we are selecting (K+1) segments including the current segment in LFS and rejecting the one that is in the LFS. "3" was used for both K = 1 and 2 because these were the most common values and speeded up the selection procedure with a simple program. Also in the case such as K = 1, rejection of a segment was facilitated when it clashed with other surveys (eg. Basic or MTP). If K > 1 and rejection was required, then K + 1 was replaced by K + 2 and the selection repeated with the possible rejection undertaken.

If a segment was so large that it clashed with LFS and other surveys, then random starts were reserved and a subsample of the expected LFS take was made. The problem usually occurred in large segments with density factors in LFS so that subsampling was always possible. The reserved random starts were multiples of W/(k+1) from the random start used in LFS.

The procedure was readily adapted to a program on a small calculator so that clerks could make the selection.

5.4 Selection of Households Within Segments

Step interval for systematic sampling = $[\frac{D \text{ (dwelling segment count)}}{G \text{ (density factor)}} + 5]$
i.e., D/G rounded to the nearest integer.

Random start = $[[\frac{D}{G} + .5] \times \frac{R}{10^4} + 1]$ with decimal portion ignored.

R = 4 digit random number between 0000 and 9999.

The random start for systematic sampling of households within a selected segment was independent of the accumulated size for segment selection since there is no bias of segment rotation with the small number of random starts to be used in the various FEX surveys. This differs from LFS, a continuing survey of long duration, where the random start within segment must be related to the accumulated sizes for segment selection to avoid a long term bias toward large segments.

Acknowledgements

The authors acknowledge the comments and criticisms of G. Chauvin in the original preparation of this article. Also to be commended for his contribution to the descriptive portions is H. Champion, Chief, Family Expenditure Division, Consumer Income & Expenditure Division.

REFERENCES

[1] Methodology: Canadian Labour Force Survey (Dominion Bureau of Statistics, Sampling and Survey Research Staff, 1965; Statistics Canada, Household Surveys Development Staff), catalogue 71-504 (occasional).

Table A: Sample Allocation for FEX Food Diary per Month

City	Estimated Dwelling Counts*				Monthly Sample Size	Theoretical Weight	Sample Allocation (No. of segments and density factor by type of area)					
							hard core		fringe		apartment	
	Hard Core	Fringe	Apartments	Entire City			No. Seg.	Density Factor	No. Seg.	Density Factor	No. Seg.	Density Factor
St. John's	13,409	10,600	0	24,009	30	800.3	6	2.79	10	1.32	0	-
Halifax	33,897	17,435	6,056	57,388	35	1,639.7	7	2.95	8	1.33	2	1.85
Montreal	614,625	89,286	69,000	772,911	80	9,661.4	20	3.18	8	1.15	3	2.38
Ottawa	82,411	22,082	30,194	134,687	35	3,848.2	7	3.06	4	1.43	3	2.62
Toronto	531,980	40,045	232,000	804,025	75	10,720.3	16	3.10	3	1.25	7	3.09
Winnipeg	131,535	15,452	19,414	166,401	45	3,697.8	12	2.96	3	1.39	2	2.63
Edmonton	116,259	14,407	15,231	145,897	40	3,648.9	10	3.19	3	1.32	2	2.09
Vancouver	252,594	40,281	31,849	324,724	55	5,904.1	14	3.06	5	1.36	2	2.70
St. John, N.B.	20,291	7,101	0	27,392	30	913.1	7	3.17	6	1.30	0	-
Quebec City	80,276	27,079	6,421	113,776	35	3,250.7	8	3.09	6	1.39	1	1.98
Thunder Bay	23,566	3,749	0	27,315	30	910.5	8	3.24	3	1.37	0	-
Regina	38,828	0	0	38,828	30	1,294.3	9	3.33	0	-	0	-
Saskatoon	32,356	5,038	0	37,394	30	1,246.5	8	3.24	3	1.35	0	-
Calgary	108,116	6,411	12,655	127,182	35	3,633.8	9	3.31	1	1.76	2	1.74

* based on dwelling units in the most recent count by subunit

A COMPUTER ALGORITHM FOR JOINT PROBABILITIES OF SELECTION

(SYSTEMATIC PPS SAMPLING)

M.A. Hidiroglou and G.B. Gray

In 1962, Hartley and Rao derived an asymptotic formula for the joint probability selection for samples selected with unequal probability sampling. In 1966, Connor, derived an exact formula for this joint probability, however, his formulae were very involved. In the present paper the authors, using a modification of Connor's formula derive the exact joint probabilities using a specially designed computer algorithm.

1. INTRODUCTION

We present a computer algorithm which will be used to compute joint probabilities of selection of units in systematic samples given that the units in the population are randomly ordered prior to sample selection. H.O. Hartley and J.N.K. Rao [3] have derived an asymptotic formula for joint probability selection which holds approximately true for pairs of units selected from a large population. The present method is based on G.B. Gray's [2] algorithm. We present a systematic method of computing joint probabilities directly applicable to computer programming. This will permit us to compare Hartley and Rao's asymptotic results to our exact results for small sample and population sizes.

2. NOTATION

Let the units in the population be denoted by U_i, $i = 1, 2, \ldots, N$. Associated with U_i is an assigned probability p_i (usually a relative size measure of unit U_i). We assume that a sample of size $n \geq 2$ is selected systematically and without replacement with each included element chosen proportionally to p_i. We assume that $np_i \leq 1$ for $i = 1, 2, \ldots, N$.

The joint inclusion probability Π_{ij} of units U_i and U_j is of interest. In the actual selection of U_i and U_j in the sample, these two units may have been separated by $k = 0, 1, 2, \ldots, N-2$ units in the population. Due to the symmetry introduced by the sampling design, we only need consider:

$$\sum_{k=0}^{\frac{N-3}{2}} \binom{N-2}{k} = 2^{N-3} \qquad \text{(N odd)} \qquad (2.1)$$

distinct arrangements of units between i and j for k = 0, 1, 2,..., $\frac{N-3}{2}$ and

$$\sum_{k=0}^{\frac{N-4}{2}} \binom{N-2}{k} + 1/2 \binom{N-2}{\frac{N-2}{2}} = 2^{N-3} \qquad \text{(N even)} \qquad (2.2)$$

distinct arrangements of units between i and j for k = 0, 1, 2,..., $\frac{N-2}{2}$. Given units i and j, an (N-2) dimensional vector consisting of ones and zeroes must be constructed. The position of the ones in this vector will indicate which units are to be included in the calculation for joint probability. Since we have 2^{N-3} distinct arrangements, we construct a 2^{N-3} x (N-2) matrix. The following algorithm provides a way to generate these elements.

3. ALGORITHM

The algorithm used is based on the decomposition of the numbers 0, 1, 2,..., 2^{N-3} - 1 into an (N-2) dimensional vector composed of digits 0 or 1 of the binary equivalent. In this fashion, 2^{N-3} such vectors are generated. Denote by A the 2^{N-3} x (N-2) matrix made up of such vectors. The rows of matrix A provide us with the number of units between i and j and their exact address in terms of the population units. For example the null vector (0,0,0,0) indicates that no units are to be taken between i and j, whereas (1,0,1,0) indicates that two units are to be taken. We introduce a pseudo-complement of A which we will call $\tilde{A}$ defined as follows. Let a_{kj} and $\tilde{a}_{kj}$ be the (k,j)th elements of A and $\tilde{A}$ respectively, then,

$$\tilde{a}_{kj} = \begin{cases} a_{kj} & \text{if } r \leq [\frac{N-2}{2}] \\ |1 - a_{kj}| & \text{otherwise} \end{cases} \qquad (3.1)$$

where $r = \sum_{j=1}^{N-2} a_{kj}$ = number of 1's in a given row

and [n] = integer part of n.

We re-arrange the population units $(U_1, U_2, \ldots, U_i, \ldots, U_j, \ldots, U_N)$ as $(U_i, U_j, U_1', U_2', \ldots, U_{N-2}')$ and define the corresponding vector $S = (np_i, np_j, np_1', \ldots, np_{N-2}')'$ where without loss of generality $np_i \leq np_j$. Decompose S into S_1 and S_2 where $S_1 = (np_i, np_j)'$ and $S_2 = (np_1', np_2', \ldots, np_{N-2}')'$. Define the 2^{N-3} dimensional vectors W_2 and V_2 where

$W_2 = \tilde{A}\, S_2$ (summing the probabilities in each set),

$V_2 = \tilde{A}\, 1$ (counting the number of 1's in each set),

$1' = (1, 1, \ldots, 1)$ (a row-vector of 2^{N-3} 1's).

We define a "selecting" 2^{N-3} dimensional vector $Z(k)$ whose j^{th} element is

$$Z_j(k) = \begin{cases} 1 & \text{if } V_{2j} = k \\ 0 & \text{otherwise} \end{cases} \qquad (3.2)$$

V_{2j} being the j^{th} element of V_2 .

The 2^{N-3} dimensional vector W_2 is transformed through a linear transformation to a 2^{N-3} dimensional vector b_2. Denoting the m^{th} element of W_2 as W_{2m}, define

$$d_{2m} = W_{3m} - [W_{3m}], \; m = 1, 2, \ldots, 2^{N-3} \qquad (3.3)$$

where $$W_{3m} = W_{2m} + np_i + np_j \; . \qquad (3.4)$$

Then the elements of the 2^{N-3} dimensional vector b_2 are obtained by using the rules given in tables 1 and 2.

Table 1: Value of b_{2m} $(np_i + np_j \leq 1)$

Range of d_{2m}	Value of b_{2m}
$[0, np_i)$	d_{2m}
$[np_i, np_j)$	np_i
$[np_j, np_i + np_j)$	$np_i + np_j - d_{2m}$
$[np_i + np_j, 1)$	0

Table 2: Value of b_{2m} $(np_i + np_j > 1)$

Range of d_{2m}	Value of b_{2m}
$[0, np_i + np_j - 1)$	$np_i + np_j - 1$
$[np_i + np_j - 1, np_i)$	d_{2m}
$[np_i, np_j)$	np_i
$[np_j, 1)$	$np_i + np_j - d_{2m}$

Define $\ell(k) = b_2' Z(k)$ and $m(k) = \binom{N-2}{k}$

$b_2' = \{b_{2m}\}$, b_{2m} denoting the conditional joint probability of selection of i and j for a given selection of units between i and j, associated with the m^{th} row of $\tilde{A}$.

Then the joint probability π_{ij} is simply,

$$\pi_{ij} = \frac{2}{N-1} \sum_{k=0}^{M} \frac{\ell(k)}{m(k)} \tag{3.5}$$

$$\text{where } M = \begin{cases} \left[\frac{N-3}{2} \right] & \text{if N is odd} \\ \left[\frac{N-2}{2} \right] & \text{if N is even} \end{cases}$$

Rao [3]'s formula for calculating the joint probability of including units i and j in the sample is,

$$\begin{aligned} \pi_{ij} = n(n-1)p_i p_j \; [1 + (p_i + p_j) - S_2 + 2(p_i + p_j)^2 \\ - 2\, p_i p_j - 3(p_i + p_j)\, S_2 \\ + 3S_2^2 - 2S_3] \end{aligned} \tag{3.6}$$

where $S_2 = \sum_{j=1}^{N} p_i^2$ and $S_3 = \sum_{i=1}^{N} p_i^3$.

4. EXAMPLES

To illustrate the algorithm, we present two examples: one taken from Gray [2], and the other from Connor [1]. We compare the effect of increasing population size on the joint probabilities calculated using the exact formula and using Rao's asymptotic formula.

In our first example we use as input probabilities the values used by Gray [2].

Table 3: Units and associated np_i

Unit No	$2p_i$
2	.28
5	.38
1	.20
3	.34
4	.36
6	.44

The population size is N = 6 and the sample size is n = 2. Hence in this case, k = 0, 1 or 2 and $2^{N-3} = 8$ for N = 6. We calculate the joint probability of units 2 and 5 being selected in the sample. The (8 x 4) A and $\tilde{A}$ matrices are:

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad \tilde{A} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Next, we list the required vectors and scalars involved in the calculation of π_{25}:

$$\begin{aligned}
S_2 &= (.20,\ .34,\ .36,\ .44)' \\
1 &= (1,\ 1,\ 1,\ 1)' \\
W_2 &= (0.0,\ 0.44,\ 0.36,\ 0.80,\ 0.34,\ 0.78,\ 0.70,\ 0.20)' \\
V_2 &= (0,\ 1,\ 1,\ 2,\ 1,\ 2,\ 2,\ 1)' \\
Z(0) &= (1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0)' \\
Z(1) &= (0,\ 1,\ 1,\ 0,\ 1,\ 0,\ 0,\ 1)' \\
Z(2) &= (0,\ 0,\ 0,\ 1,\ 0,\ 1,\ 1,\ 0)'
\end{aligned}$$

$$\begin{aligned}
W_3 &= (0.66,\ 1.10,\ 1.02,\ 1.46,\ 1.00,\ 1.44,\ 1.36,\ 0.86)' \\
d_2 &= (0.66,\ 0.10,\ 0.02,\ 0.46,\ 0.00,\ 0.44,\ 0.36,\ 0.86)' \\
b_2 &= (0.0,\ 0.10,\ 0.02,\ 0.20,\ 0.0,\ 0.22,\ 0.28,\ 0.0)'
\end{aligned}$$

$$\begin{aligned}
\ell(0) &= 0.00 \quad , \quad m(0) = 1 \\
\ell(1) &= 0.12 \quad , \quad m(1) = 4 \\
\ell(2) &= 0.70 \quad , \quad m(2) = 6 \ .
\end{aligned}$$

Hence,

$$\begin{aligned}
\pi_{25} &= \frac{2}{5}\left(\frac{0.0}{1} + \frac{0.12}{4} + \frac{0.70}{6} \right) \\
&= 0.0586
\end{aligned}$$

The calculated joint probabilities using formulae (3.5) and (3.6) are presented in table 4.

Table 4: Joint Probability for each pair of units. (Gray)

UNIT j	1	2	3	4	5	6	TOTAL
1	-	0.0387* 0.0295** -23.77***	0.0387 0.0371 -4.13	0.0387 0.0398 2.84	0.0420 0.0426 1.43	0.0420 0.0513 22.14	0.2001 0.2003
2	0.0387 0.0295 -23.77	-	0.0487 0.0545 11.91	0.0553 0.0584 5.61	0.0587 0.0624 6.30	0.0787 0.0752 -4.45	0.2801 0.2800
3	0.0387 0.0371 -4.13	0.0487 0.0545 11.91	-	0.0753 0.0736 -2.26	0.0787 0.0787 0.00	0.0987 0.0948 -3.95	0.3401 0.3387
4	0.0387 0.0398 2.84	0.0553 0.0584 5.61	0.0753 0.0736 -2.26	-	0.0853 0.0844 -1.05	0.1053 0.1016 -3.51	0.3599 0.3578
5	0.0420 0.0426 1.43	0.0587 0.0624 6.30	0.0787 0.0787 0.00	0.0853 0.0844 -1.05	-	0.1153 0.1087 -5.72	0.3800 0.3768
6	0.0420 0.0513 22.14	0.0787 0.0752 -4.45	0.0987 0.0948 -3.95	0.1053 0.1016 -3.51	0.1153 0.1087 -5.72	0	0.4400 0.4316

* : Exact joint probability
** : Rao's joint probability
*** : % difference

To see the effect of increased population size on the joint probabilities, we use the following example taken from Connor [1], with N = 10 and n = 2. Let the π's be as follows:

π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8	π_9	π_{10}
.05	.08	.10	.15	.18	.20	.20	.30	.34	.40

The joint probabilities are calculated using formulae (3.5) and (3.6): the results are given in table 5.

5. CONCLUSIONS

Table 5 gives the joint probabilities using formulae (3.5) and (3.6). From table 4, the highest % deviation between the joint probability values obtained using formulae (3.5) and (3.6) is 24%, while in table 5 it is 10%. Moreover, the mean of these % deviations is .09 (± 9.94) in table 4 while it is -0.16 (± 2.06) in table 5. From these observations, we conclude that Rao's formula is not very good for $N \leq 10$. However, due to its asymptotic properties its precision will increase as N gets larger. Moreover, using Rao's formula, joint probabilities are easy to calculate, whereas using formula (3.5), 2^{N-3} calculations must be made for each joint probability.

Table 5: Joint Probability for each pair of units (Connor)

UNIT j	1	2	3	4	5	6	7	8	9	10
1	-	0.00190* 0.00185** -2.63	0.00248 0.00233 -6.05	0.00348 0.00359 3.16	0.00440 0.00438 -0.45	0.00483 0.00492 1.86	0.00483 0.00492 1.86	0.00793 0.00781 -1.51	0.00907 0.00908 0.11	0.01107 0.01110 0.27
2			0.00419 0.00379 -9.55	0.00590 0.00583 -1.19	0.00683 0.00711 4.10	0.00793 0.00798 0.63	0.00793 0.00798 0.63	0.01279 0.01259 -0.78	0.01469 0.01475 0.41	0.01783 0.01803 1.12
3				0.00733 0.00736 0.41	0.00893 0.00897 0.45	0.01002 0.01008 0.60	0.01002 0.01008 0.60	0.01545 0.01602 3.69	0.01879 0.01862 -0.90	0.02279 0.02278 0.04
4					0.01360 0.01381 1.54	0.01536 0.01552 1.04	0.01536 0.01552 1.04	0.02460 0.02469 0.37	0.02860 0.02870 0.35	0.03579 0.03512 -1.87
5						0.01895 0.01893 -0.11	0.01895 0.01893 -0.11	0.03014 0.03014 0.00	0.03481 0.03503 0.63	0.04338 0.04286 -1.20
6							0.02110 0.02127 0.81	0.03381 0.03387 0.18	0.03924 0.03937 0.33	0.04876 0.04818 -1.15
7								0.03381 0.03387 0.18	0.03924 0.03937 0.33	0.04876 0.04818 -1.19
8									0.06271 0.06276 0.08	0.07876 0.07681 -2.48
9										0.09286 0.08929 -3.84

* : Exact joint probability
** : Rao's joint probability
*** : % difference

REFERENCES

[1] Connor, W.S., 1966. An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement. J. Amer. Statist. Assoc. 61: 384-390.

[2] Gray, G.B., 1971. Joint Probabilities of Selection of Units in Systematic Samples. Proceedings of Amer. Statist. Assoc.: 271-276.

[3] Hartley, H.O. and Rao, J.N.K., 1962. Sampling with unequal probabilities without replacement from a finite universe. J. Amer. Statist. Assoc. 47: 663-685.

THE DEVELOPMENT OF AN AUTOMATED ESTIMATION SYSTEM

A. Satin and A. Harley

Although a survey is designed to satisfy a specific set of survey constraints, some steps involved in designing a survey, such as stratification, sample allocation and sample selection are common to all surveys. The steps involved in the creation of survey design systems are to identify, develop and implement common methods and procedures for such stages which, when taken together, constitute a survey design.

The paper describes some methodological considerations in the development of an automated system for three methods of ratio estimation.

1. INTRODUCTION

A survey is designed to conform to a set of defined survey objectives, and is affected by the method of enumeration, cost and time constraints, data processing requirements and a variety of other considerations. Although a survey design is developed to satisfy a specific set of survey constraints, the various stages involved in designing a survey, however, such as stratification, sample allocation and sample selection are common to all surveys. The steps involved in the creation of survey design systems are to identify, develop and implement common methods and procedures for such stages which, when taken together, constitute a survey design.

Within Statistics Canada, it was felt that there existed a need to co-ordinate and unify methodological and system requirements in this area of work. In order to achieve this, a General Survey System (GSS) Design Group composed of survey methodologists and systems analysts was formed within the Statistical Services Field of Statistics Canada. The function of this group is to develop methods and prepare the general algorithms necessary to mechanize survey design stages.

The group has concentrated its effort on the design phase, as other phases of a survey system (for example, generalized editing and tabulation systems) have already received wide attention by other groups in Statistics Canada.

Flexibility is being incorporated into the design systems to accommodate various methods of selection and estimation with a view to meet the requirements of a variety of surveys undertaken by Statistics Canada.

In the case of sample selection, the following four sample selection methods were automated by the group.

a) Simple random sampling without replacement.
b) Systematic sampling with equal probability.
c) Systematic sampling with probability proportional to size.
d) Sampling with probability proportional to total size.

A computer system was implemented which enabled the user to specify any of the above sample selection methods. The system enables a stratified sample to be drawn in several replicates and provides the corresponding sampling weights (inverse probabilities of selection). The system is designed so as to permit other methods of sample selection to be incorporated should the need arise. It can be used whenever an appropriate frame of the population under study is available.

This paper describes some methodological considerations in the development of an automated system for estimation.

1.1 Why a General Survey System

Systems for processing survey and census data form a major part of systems development in Statistics Canada. Within this context, systems can mean both manual and automated. However, at this point, we are concerned more with automated systems with their user/methodologists interfaces. The characteristics of General Survey Design Systems are summarized below.

a) The system can be adapted readily to the users application, that is, the system can be "tailor-made" to suit the problem.
b) The system covers a broad scope of frequently encountered methodologies.
c) Repeat runs produce identical results, a useful feature if files, sample lists, etc. are mislaid.

d) Decisions made during the designing and processing of surveys are made objectively, i.e. in a given set of circumstances the same decision is made irrespective of any personal biases.

d) By utilizing the power of the computer, it is possible to implement more complex methodology leading to a better survey design. For this purpose guidelines should be available outlining the statistical procedures and framework, to ensure user understanding of the methodology and their implicit assumptions.

f) Implementation of the survey designer's application is achieved in less time, thus allowing more effort to be concentrated on evaluation of the methodological decisions.

1.2 Philosophy of Systems Implementation

The goals of generalized survey design systems can be divided into three time frames.

a) Short term goals

The immediate concern of the Design Group is to produce programs and systems which will be of immediate use in a wide number of applications. That is, the first implementation of the system should concern itself with providing programs which will operate independently from other systems, which cover the most currently used methodologies and which are easy to use. Given these characteristics, programs developed in this way serve the majority of users.

b) Medium term goals

To provide a framework in which additional methodologies can be readily implemented. This can arise if new or more sophisticated ideas are deemed desirable to deal with everchanging survey processing needs. This process can only be achieved by designing systems in such a way that components of programs are functionally self-contained, are as independent from each other as is practicable, and that the dependencies are well defined. In this way functional components can be added, updated, replaced or deleted with a minimum of disruption to the overall functioning of the system.

c) Long term goals

The ultimate goal in survey processing environments is one in which the survey designers and systems analysts can concentrate their efforts in the statistical design processes, without undue resources being expended in the areas of data transformations, data control and integrity.

To achieve this end it is necessary to have a data management supervisor system whose functions can be defined as

1. To act as the interface between the data files and the survey processing algorithms.
2. To control the sequence of processing for each unit of data in the system.
3. To control the integrity of data by validating the authority of persons applying transactions to the data, by checking previous transactions, by storing data in both its processed and unprocessed states.

Computer tools to achieve these ends include such things as, data dictionaries to store the logical and physical structure of the data, data base descriptors to store the physical linkage of files and execution controllers to maintain proper sequencing. Such tools have been implemented or are under active consideration by Statistics Canada, for use in their future survey processing systems.

The GSS design group first identified the steps involved in a survey system from the design to the evaluation stage. The three stages of survey design, stratification, sample selection, and estimation were considered to have immediate application as independent modules and were to be implemented with a view to later integrate them with sample size determination and sample allocation so as to comprise one major area of survey design. Of these stages, a computer system has been implemented for sample selection and another is currently being implemented for estimation.

2. AUTOMATED ESTIMATION SYSTEM

2.1 Estimation Procedures

The estimation system consists of adjusting the sampling weights of a survey file by auxiliary variables such as known age-sex or income distributions of the population under study. A weight adjustment factor (A_i) is applied to each weight (W_i) of a survey file so as to give rise to the ratio estimator which may be expressed as follows:

$$\hat{Y}_p = \sum_{i=1}^{n} (A_i^{(t)} W_i) \, y_i \, \delta_i^{(p)}$$

where $\hat{Y}_p$: Estimated total of characteristic y for sub-population p.

Y_i : 'y' value as reported by unit i. For the special case of estimated counts y_i = 1 if unit i has characteristic y
= 0 otherwise.

$\delta_i^{(p)}$: Indicator variable referring to those records belonging to sub-population p. $(\delta_i^{(p)} = 1 \text{ or } 0)$.

t : Refers to type of adjustment which is performed when more than one auxiliary variable is used.

n : Total number of records in survey file.

When the weights are to be adjusted by more than one auxiliary variable, the weight adjustment factor may be determined in one of the following three ways:

<u>Separate Adjustment Factor $(A_i^{(1)})$</u>

$$A_i^{(1)} = \sum_{k=1}^{S} \S_k \sum_{\ell} \frac{X_k^{(\ell)}}{\hat{X}_k^{(\ell)}} \, \delta_{x_k, i}^{(\ell)}$$

where $X_k^{(\ell)}$: External value of the ℓth category of variable x_k

$\hat{X}_k^{(\ell)}$: Estimate of X_k based on the sample.

$\delta_{x_k,i}^{(\ell)}$: Indicator variable referring to those records belonging to the ℓth category of variable x_k.

$\S_k$: Relative weight assigned to variable x_k

$$\left(\sum_{k=1}^{S} \S_k = 1 \right)$$

S : Number of Auxiliary Variables

$A_i^{(1)}$ is a weighted average adjustment based on the entire set of auxiliary variables $x_1, x_2, \ldots, x_s$.

Sequential Adjustment Factor ($A_i^{(2)}$)

$$A_i^{(2)} = \frac{X_1^{(\ell_1)}}{\hat{X}_1'^{(\ell_1)}} \; \frac{X_2^{(\ell_2)}}{\hat{X}_2'^{(\ell_2)}} \; \cdots \; \frac{X_S^{(\ell_s)}}{\hat{X}_S'^{(\ell_s)}} \qquad i \varepsilon x_1^{(\ell_1)} , x_2^{(\ell_2)} \ldots x_S^{(\ell_s)}$$

where $\hat{X}_k'^{(\ell)}$: Estimate of $X_k^{(\ell)}$ based on the sample adjusted for variables $x_1, x_2, \ldots x_{k-1}$. This estimate is obtained by summing the product of the reported $x_k^{(\ell)}$ value and the adjusted weights of all records belonging to ℓth category of variable x_k

Note: The sequence of adjustment of the weights by the auxiliary variables is important in the sense that there is an exact fit of the sample estimates to the corresponding population distribution only for the last auxiliary variable in the sequence. The auxiliary variables may be sequenced in terms of their correlation with the variable(s) of interest.

Iterative Sequential Adjustment Factor ($A_i^{(3)}$)

The sequence of adjustment by the auxiliary variables described for the sequential adjustment is iterated a specified number of times. The resulting estimator becomes the raking ratio estimator which, for the case of two auxiliary variables, is known to converge quite rapidly [4]. The set of

weights at the start of an iteration (W_i) are taken to be the set of adjusted weights at the end of the previous iteration ($A_i^{(2)}W_i$).

The system can be used to adjust weights for non-response based on a defined variable. The system will also have the capability of providing variance estimates for stratified replicated designs or based on the method of pseudo replication. Consideration may then be given to providing variance estimates on the basis of balanced repeated replication described in Section 3.

2.2 Input Requirements for the Estimation System

i) Record file of respondents. Each record contains a sampling weight (W_i) and fields corresponding to the variables of interest (y), variable x to be used for non-response adjustment (option), and the auxiliary variables x_k ($k = 1, 2, \ldots S$).

Note: The estimation system will have the capability of accepting through a high level specification language, alternative definitions of categories or even of defining new variables and new categories from existing ones, eg. Creation of 'age-sex' variable from the 'age' and 'sex' variables, creation of age category 10-25, from age categories 10-15, 15-25, etc.

ii) A = 0 Weights are not to be adjusted

A = 1 Weights are to be adjusted for non-response
→ Specify (a) Variable x and categories ℓ for which the non-response adjustment will be carried out.
(b) External values of $X^{(\ell)}$.

A = 2 Weights are to be adjusted by auxiliary variables
→ Specify (a) The set of auxiliary variables (x_k), and the levels ℓ for which the adjustments will be carried out, ($\ell = 1, 2, \ldots r_k$; $k = 1, 2, \ldots S$).
(b) External values $X_k^{(\ell)}$ ($\ell = 1, 2, \ldots r_k$; $k = 1, 2, \ldots S$).

iii) A = 2   B = 1 - Separate Ratio Adjustment

→ Specify the relative weight $\S_k$ of the auxiliary variables

B = 2 - Sequential Ratio Adjustments

→ Specify the order of the auxiliary variables

B = 3 - Iterative Sequential Ratio Adjustments

→ Specify the order of the auxiliary variables and the required number of iterations.

2.3 Some Applications of the Estimation System

(1) Estimates of Proportions and Percentages

The proportion of individuals having characteristic y in sub-population p ($\hat{P}_{y_p}$) can be obtained as the ratio of two adjusted estimates as follows:

$$\hat{P}_{y_p} = \frac{\hat{N}_{y_p}}{\hat{N}_p}$$

where

$$\hat{N}_{y_p} = \sum_{i=1}^{n} (A_i^{(t)} W_i)\, y_i\, \delta_i^{(p)}$$

y_i = 1 if unit i has characteristic y

= 0 otherwise

$$\hat{N}_p = \sum_{i=1}^{n} (A_i^{(t)} W_i)\, \delta_i^{(p)}$$

$\hat{P}_{y_p}$ x 100% is the corresponding percentage.

(2) Stratified Sampling

$$\hat{Y}_p = \sum_{h=1}^{H} \hat{Y}_{ph}$$

where $\hat{Y}_{ph}$ is the estimated total of y for sub-population p in stratum h (h = 1, 2, ... H).

(3) Replicated Sampling

$$\hat{Y}_p = \sum_{r=1}^{R} \hat{Y}_{pr}$$

where $R\ \hat{Y}_{pr}$ is the estimated total of y for sub-population p based on the rth replicate (r = 1, 2, ... R).

(4) Multi-Phase Sampling

For multi-phase survey designs, where the sample at a given phase is sub-selected from a prior phase, the values of the variables which are required for ratio estimation at a given phase may be estimated from a prior phase.

3. VARIANCE ESTIMATION

3.1 Variance Estimation for Replicated Designs

Let $R\ \hat{Y}_{phr}$ denote the estimated total of y for sub-population p within stratum h (h = 1, 2, ..., H) based on replicate r (r = 1, 2, ..., R)

$$\hat{Y}_{phr} = \sum_{i=1}^{n} (A_i^{(t)}\ W_i)\ y_i\ \delta_i^{(phr)}$$

$$\hat{Y}_{ph} = \sum_{r=1}^{R} \hat{Y}_{phr}$$

$$\hat{Y}_p = \sum_{h=1}^{H} \hat{Y}_{ph}$$

The variance estimate of the estimated total $\hat{Y}_p$ may be calculated as follows:

$$V_{REP}\ (\hat{Y}_p) = \sum_{h=1}^{H} \sum_{r=1}^{R} \frac{(R\ \hat{Y}_{phr} - \hat{Y}_{ph})^2}{R(R-1)} \qquad (3.1)$$

The method of pseudo replication consists of splitting the sample randomly into R replicates and then applying formula (1).

3.2 Variance Estimation Based on Balanced Repeated Replication (BRR)

The basic framework for the BRR method is the selection (with replacement), of two replicates (PSUs) from each of H strata. Selections can be made randomly or with probability proportional to size. The PSUs may also be selected without replacement for the case of simple random sampling. Estimates of the variance are obtained using the estimates derived from the repetitions that can be generated with alternate combinations of two PSUs from H strata. The BRR method is a technique used to reduce the required number of repetitions from 2^H to a set of orthogonally balanced repetitions [1].

3.2.1 Determination of the Set of Balanced Repeated Replications

The required number of balanced repeated replications R, is the smallest integer multiple of 4 greater or equal to the number of strata H. For each of the H strata, one of the selected PSUs is designated +, the other -. A repetition involves the selection of one of the two PSUs from each stratum. This selection pattern corresponds to a R x R matrix of + and - signs, whose columns are orthogonal and R is a multiple of 4. A method for constructing such matrices has been worked out by Plackett and Burman (1943 - 1946), and is described in Biometrika 33, 305-325 [2].

The situation is similar when R is any integral multiple of 4 and the number of strata is H = R-1. If H = R-2 or H = R-3, orthogonal balance may be obtained by omitting any one or two columns respectively. If H = R, orthogonal balance may be obtained by writing a whole column of - for the last stratum, using the sample replicate from it for every repetition. Although this does not disturb the variance estimates, it should be noted that H = R does sacrifice the symmetrical use of all replicates and therefore the estimate obtained by averaging over all repetitions does not equal the estimate obtained from the entire sample.

3.2.2 BRR Variance Estimation Procedures

Let $2\hat{Y}_{phj}$ denote the estimated total (unadjusted) of characteristic y for sub-population p within stratum h (h = 1, 2, ..., H) based on PSU j (j = 1,2)

$$\hat{Y}_{phj} = \sum_{i=1}^{n} W_i \; y_i \; \delta_i^{(phj)}$$

$$\hat{Y}_{ph} = \hat{Y}_{ph1} + \hat{Y}_{ph2}$$

$$\hat{Y}_p = \sum_{h=1}^{H} \hat{Y}_{ph}$$

CASE 1 - Unadjusted Estimate (SRSWR, PPSWR)

Determine for replicate r (r = 1, 2,..., R):

$$\hat{Y}_{phr} = 2\,\hat{Y}_{ph1}\,\delta_{h1}^{(r)} + 2\,\hat{Y}_{ph2}\,\delta_{h2}^{(r)} \qquad (3.2)$$

$$\hat{Y}_{pr} = \sum_{h=1}^{H} \hat{Y}_{phr} \qquad (3.3)$$

where $\delta_{hj}^{(r)}$ = 1 if PSU j is selected in stratum h for replicate r
= 0 otherwise
$\delta_{h1}^{(r)} = 1 - \delta_{h2}^{(r)}$

The variance estimate is calculated as follows:

$$V_{BRR}\,(\hat{Y}_p) = \sum_{r=1}^{R} \frac{(\hat{Y}_{pr} - \hat{Y}_p)^2}{R}$$

CASE 2 - Unadjusted Estimate (SRSWOR)

Determine for replicate r (r = 1, 2,..., R)

$$\hat{Y}'_{phr} = \sqrt{1 - \frac{2}{N_h}}\;\hat{Y}_{phr}$$

$$\hat{Y}'_{ph} = \sqrt{1 - \frac{2}{N_h}}\;\hat{Y}_{ph}$$

where $\hat{Y}_{phr}$ is determined as in (2).

The variance estimate is then calculated as follows:

$$V_{BRR}(\hat{Y}'_p) = \sum_{r=1}^{R} \sum_{h=1}^{H} \frac{(\hat{Y}'_{phr} - \hat{Y}'_{ph})^2}{R}$$

CASE 3 - Ratio of two Unadjusted Estimates (SRSWR, PPSWR)

Let $R_{p_1p_2}$ denote the ratio $\frac{Y_{p_1}}{X_{p_2}}$ and $\hat{R}_{p_1p_2}$ the estimate of $R_{p_1p_2}$.

Determine $\hat{Y}_{p_1r}$ and $\hat{X}_{p_2r}$ as in (3) and $\hat{R}_{p_1p_2r} = \frac{\hat{Y}_{p_1r}}{\hat{X}_{p_2r}}$ (3.4)

The variance estimate is calculated as follows:

$$V_{BRR}(\hat{R}_{p_1p_2}) = \sum_{r=1}^{R} \frac{(\hat{R}_{p_1p_2r} - \hat{R}_{p_1p_2})^2}{R}$$

CASE 4 - Ratio Estimate Using One Auxiliary Variable (SRSWR, PPSWR)

Let $X^{(\ell)}$ denote the estimated total (unadjusted) of the auxiliary variable x for category ℓ and $\hat{Y}_{p\ell}$ the estimated total (unadjusted) of characteristic y for sub-population p belonging to category ℓ.

Let $\hat{Y}^*_{p\ell} = \frac{\hat{Y}_{p\ell}}{\hat{X}^{(\ell)}} X^{(\ell)}$

$$\hat{Y}^*_p = \sum_{\ell \varepsilon x} Y^*_{p\ell}$$

Determine $\hat{R}_{p\ell r} = \frac{\hat{Y}_{p\ell r}}{\hat{X}^{(\ell)}_r}$ as in (4)

$$\hat{Y}^*_{p\ell r} = X^{(\ell)} \hat{R}_{p\ell r}$$

The variance estimate is calculated as follows:

$$V_{BRR}\,(\hat{Y}^*_p) = \sum_{r=1}^{R} \sum_{\ell\epsilon x} \frac{(\hat{Y}^*_{p\ell r} - \hat{Y}^*_{p\ell})^2}{R} \qquad (3.5)$$

<u>CASE 5</u> - Ratio Estimate Using Several Auxiliary Variables (SRSWR, PPSWR)

$$\text{Let } \hat{Y}^{**}_p = \sum_{k=1}^{S} \S_k \sum_{\ell\epsilon x_k} \frac{\hat{Y}_{p\ell}}{\hat{X}^{(\ell)}_k} X^{(\ell)}_k$$

Determine for each auxiliary variable x_k, $(k = 1, 2, \ldots, S)$, $V_{BRR}\,(\hat{Y}^*_{pk})$ as in (5).

The variance estimate is calculated as follows:

$$V_{BRR}(\hat{Y}^{**}_p) = \sum_{k=1}^{S} \S^2_k\, V_{BRR}(\hat{Y}^*_{pk}) + \sum_{r=1}^{R} \sum_{k\neq k'}^{S} \S_k \S_{k'} \frac{(\hat{Y}^*_{pkr} - \hat{Y}^*_{pk})\,(\hat{Y}^*_{pk'r} - \hat{Y}^*_{pk'})}{R}$$

3.2.3 Extensions of BRR (SRSWR)

The BRR method for increasing the precision of variance estimates exploits the repetition that can be generated with alternate combinations of two or more replicates from several strata. The basic framework within which the method of BRR operates, ie. two replicate selections for each stratum, may be extended and modified as illustrated below [1].

(1) Reduction of the Number of Strata

If the number of strata is too large for operational convenience, pseudo-strata may be formed by "combining" replicates across the strata. If one can chose the number of 'computing', (ie. collapsed) strata, they should be made one less than the number of repetitions, which is a multiple of 4. One should if possible, avoid strata of grossly unequal sizes because these would have unfavorable effects on the variance.

(2) Several PSUs / Stratum

If the number of PSUs is even, they can be randomly combined into two halves. BRR can then be applied to the replications based on the pairs. If the number of PSUs is odd, combining PSUs across strata might be considered.

REFERENCES

[1] Kish, L. and Frankel, M.R. (1970) "Balanced Repeated Replications for Standard Errors", J. Amer. Statist. Assoc. Vol. 65, Pg. 1071-1094.

[2] Plackett, R.L. and Burman, J. (1946), "The Design of Multifactorial Experiments", Biometrika - Vol. XXXII, Pg. 305-325.

[3] Nargundkar, M.S. and Arora, H. (1971). "The Raking Ratio Estimation Procedure for the 1971 Census", Census Division Memorandum, Statistics Canada.

[4] Rao, J.N.K., (1974) "Raking Ratio Estimators", Statistics Canada.

[5] Murthy, M.N., "Sampling Theory and Methods".