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 coordinate 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 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:

Separate Adjustment Factor ($A_i^{(1)}$)

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

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.

$\delta^{(\ell)}_{x_k,i}$: 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 ($A_i^{(2)}$)

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

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, ... 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, (ℓ = 1, 2, ... r_k; k = 1, 2, ... S).

(b) External values $X_k^{(\ell)}$ (ℓ = 1, 2, ... r_k; k = 1, 2, ... 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

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

$\hat{P}_{y_p} \times 100\%$ 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 $\hat{Y}^*_{p\ell} = \frac{\hat{Y}_{p\ell}}{\hat{X}^{(\ell)}} X^{(\ell)}$

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

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:

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

CASE 5 - 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:

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

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".