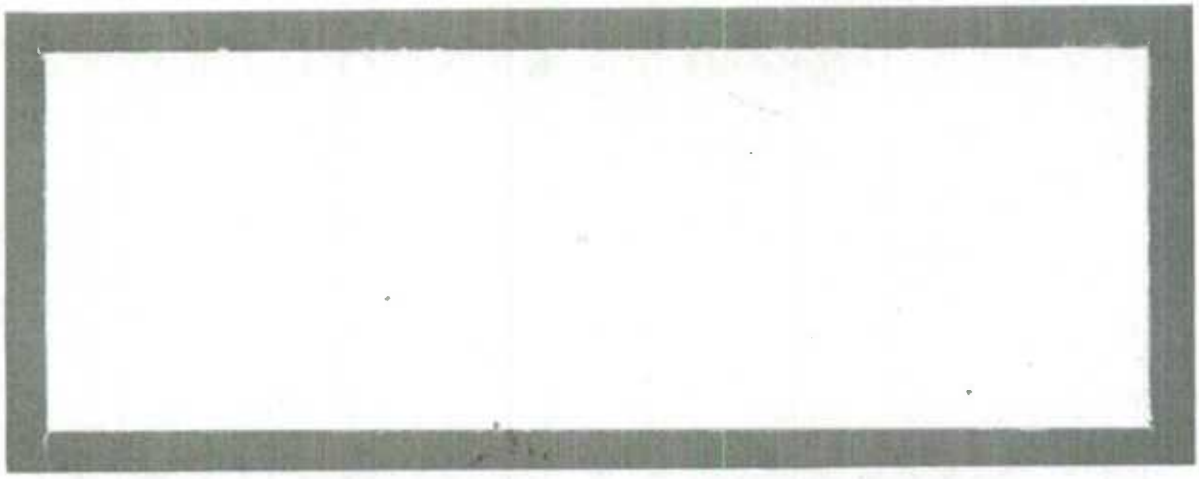
11-617E
no.89-05
c.2

Statistics Canada Statistique Canada



Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes-entreprises

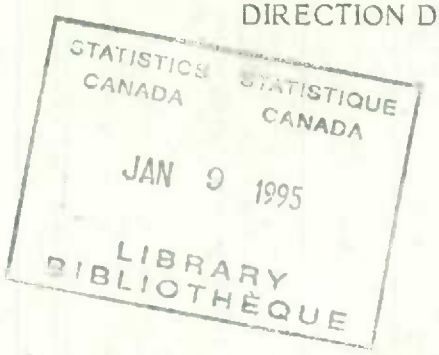
Canada

WORKING PAPER NO. BSMD-89-005E

METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-89-005E

DIRECTION DE LA MÉTHODOLOGIE



GENERALIZED SAMPLING AND ESTIMATION METHODOLOGY FOR SUB-ANNUAL BUSINESS SURVEYS AT STATISTICS CANADA

by

M.A. Hidiroglou and G.H. Choudhry

March 15, 1989

GENERALIZED SAMPLING AND ESTIMATION METHODOLOGY FOR SUB-ANNUAL BUSINESS SURVEYS AT STATISTICS CANADA

M.A. Hidiroglou
G.H. Choudhry
Business Survey Methods Division
Statistics Canada
March 15, 1989.

Méthodes généralisées d'échantillonnage et d'estimation pour les enquêtes entreprises sous-annuelles à Statistique Canada

M.A. Hidiroglou et G.H. Choudhry
Division de méthodes des enquêtes-entreprises

Dans cet article, nous proposons une stratégie pour la sélection initiale, la rotation et l'entretien des enquêtes-entreprises sous-annuelles. Le plan de sondage est stratifié avec échantillonnage par grappe, et la stratification est basée sur l'industrie, la géographie et une mesure de taille. Les méthodes qui effectuent la sélection des naissances, le retrait des morts ainsi que les changements de classification sont esquissées. On explique la rotation de l'échantillon qui a lieu en tenant compte du temps qu'une unité doit passer à l'intérieur et à l'extérieur de l'échantillon.

Un certain nombre d'estimateurs, incluant l'estimateur par valeur dilatée et l'estimateur de Mickey, a été évalué grâce à une étude empirique effectuée sous diverses conditions. Le problème de l'estimation de la variance a été considéré en se servant de la méthode de Taylor ainsi que de la technique du "jackknife".

TABLE OF CONTENTS

1. INTRODUCTION

With the advent of the Business Survey Redesign Project at Statistics Canada, a number of annual and sub-annual business surveys are being redesigned and will be fully integrated with the new Central Frame Data Base (CFDB). The Central Frame Data Base is the new Business Register from which all business surveys at Statistics Canada will draw their universes and samples. The survey processes are to be completely integrated with the CFDB and all functions related to the CFDB will be redesigned, e.g. frame extraction, sample selection and mail-out, etc.

The existing systems, operating off the previous Business Register, need to be redesigned in order to perform the above functions as well as others, such as processing the survey data (which includes edit and imputation, weighting and estimation, tabulation and data dissemination). The redesign of these systems has proceeded on four fronts: i) sampling, ii) data capture and preliminary edit, iii) statistical editing and imputation, and iv) weighting and estimation. The methodology and systems for these four fronts require standardization of methods and concepts, flexibility, parameterization, as well as easy modification and expansion capabilities.

In this paper, the focus will be on the methodology which has been adopted for the sampling, weighting and estimation of sub-annual business surveys. For these surveys, the requirements are that the sample design must be robust, simple and provide valid estimates over time. It must cope with a universe changing on account of births, deaths, splits, mergers, and amalgamations. It must have the capability to reduce response burden, by rotating units in and out of the sample on a regular basis, subject to constraints on the time in and out of sample. It must be able to accommodate re-stratification of the universe and redraw an efficient sample in mid-stream, while maximizing the overlap between the old and new samples. The user must be able to specify the sample sizes or expected coefficients of variation. For the weighting and estimation system, the requirements are as follows. Unbiased (or nearly unbiased) estimates must be produced along with the associated measures of reliability (coefficients of variation). The user must be able to specify domains of interest for which estimates are derived, the domains being defined on the basis of information available on the survey data files. The system should produce unbiased or nearly

unbiased estimates for these domains. It must be able to cope with influential observations which have been validated by the edits and/or re-contact with the respondent but which could distort estimates because of their very large contributions to the estimate.

2. SAMPLING STRATEGY

The stratification of a business universe is usually based on industry, geography and size. The size measure can be univariate (sales, number of employees) or multivariate (revenue and assets). In our context, the stratum will be a cross-classification of industry, geography and size categories. The sample design which has been adopted is that of a simple random sample of randomly formed panels (clusters) within each of the strata. Each panel can represent a group of units or a single unit. All units within a selected panel are included in the sample. Given this design, procedures for the initial selection of the sample, the selection of births, removal of deaths and implementing changes in classification are outlined. The primary objective of the strategy is to ensure that the sample reflects the current structure of the population. Rotation of the sample will take place under certain constraints such as keeping the selected units (businesses) in the sample for a certain period of time and keeping them out of the sample for at least a certain period of time after they have rotated out of the sample.

The methodology which performs the above functions is briefly summarized as follows. For each stratum h, the N_h population units within that stratum are randomly allocated to a predetermined number "P_h" of population panels, so that, initially the number of units between any two panels differs by at most one unit. The number of panels is a function of sampling rates, and time-in and time-out constraints. It may be noted that in order to achieve unbiasedness, the time-in and time-out constraints may have to be sometimes violated. The sample consists of units associated with a subset "p_h" of population panels such that p_h/P_h is approximately equal to the sampling fraction f_h. Rotation of the sample occurs by acquiring an out-of-sample panel and dropping an in-sample panel. Births are randomly allocated to the P_h population panels, one at a time, in a systematic fashion. Deaths are removed from the stratum only if they are detected by a source independent of the survey or if they have been dead for longer than a prespecified period of time. Restratification of the population and the subsequent sample redraw, adopts techniques proposed by Kish and Scott

(1971). The sample redraw maximizes the overlap between the old and the new samples. There are obvious advantages to redrawing the sample in this fashion. First, it minimizes the introduction of too many new units in the sample resulting in a smoother transition from an operational point of view, and minimizes cost. Second, discontinuity the estimates on account of sample redraw is kept to a minimum.

We now describe in detail each of the sampling functions.

2.1 Stratification and Sample Allocation

The primary strata are intersections of the industry and geographic regions for which estimates are desired for the survey. Within these primary strata, further strata are formed by the size of the unit (e.g. sales, number of employees, revenue). Because of the highly skewed nature of businesses, the size stratification within each primary stratum allows for a 'take-all' or complete enumeration stratum and a number of strata for sampling called 'take-some' strata. The boundary for determining the take-all stratum is determined by a method introduced by Hidiroglou (1986). Basically, the method finds the optimum boundary in a primary stratum so as to minimize the overall sample size for a given coefficient of variation. The method is optimal when two strata within the primary stratum are formed: i.e. - a take-all and a take-some stratum. Lavallée and Hidiroglou (1988) provide an iterative procedure when the primary stratum is to be split-up into a take-all stratum and a number of take-some strata. The method yields optimum boundaries so as to minimize the overall sample size for a given coefficient of variation and a chosen allocation scheme for the take-some strata. This method is not implemented in the current system. It will be added later. Further improvements to the Lavallée-Hidiroglou algorithm will include constrained optimization. That is, given the previous conditions and some fixed boundaries which must be respected, the remaining boundaries will be found in an optimal manner so as to satisfy the required conditions. The current system also allows for the prespecification of take-all units. Also, given some input rules, complex structured units are automatically made take-all by the system.

Currently, the boundaries for the take some strata are obtained either using the cumulative square root "f" rule introduced by Dalenius-Hodges (1959) or the cumulative square root "X" rule given by Hansen et al. (1953). Here, X is a key variable of stratification available on the frame. An alternative way to obtain the boundaries

is to simply specify them to the system via a parameter file. This method is not optimal. It must also be noted that the system can compute optimum boundaries given an input boundary. This is necessary because not all key variables (such as sales) are available on the CFDB. The boundary which has to be taken into account is the one which separates the larger units (Integrated Portion) from the smaller units (Non-Integrated Portion).

The sample sizes for each primary stratum can either be entered or computed so as to satisfy reliability criteria, while respecting the required allocation scheme. Given that the take-all sample size has been taken into account, the remaining sample size is allocated to the take-some strata within the primary stratum, proportional to the N^{γ} or X^{γ} where N is the number of units in the stratum, X is the stratum total for the key variable being considered, and the power "γ" is suitably chosen to increase the efficiency of the allocation. The advantages of power allocations are discussed in Bankier (1988). The allocation is prespecified by the user. These allocations can then be adjusted to achieve desired minimum sample sizes and/or maximum weights by stratum. The reliability criteria can be associated to the primary strata in one of two ways. First, the user can prespecify them for each primary stratum. Second, the user can enter a global level of reliability (c.v.) which is then apportioned to the primary strata as follows. The global (national) c.v. is split out so that the c.v.'s for each industry group and geographic region are equal. An iterative process using a raking ratio algorithm is then used to determine the desired c.v.'s within each of the primary strata. These c.v.'s can be controlled by entering a lower and upper bound into the system. The stratification and allocation system is described in greater detail in Latouche and Hidiroglou (1988).

2.2 Initial Selection

The selection procedure for each take-some stratum h consists of the following steps:

i) the sampling fraction, $f_h=n_h/N_h$, is determined where n_h and N_h respectively represent the sample and population sizes in terms of number of units. This will be based on sample size determination procedures adopted for the survey.

ii) The selection procedure requires that the sampling units in stratum h be grouped into a certain number P_h of population panels (clusters). The number of P_h population panels is determined as follows. If there is no "time-out" constraint, then the number of population panels is simply determined by multiplying the inverse of the sampling rate and the number of occasions that a unit must be in sample. This straightforward procedure cannot ensure that the units stay out of the sample for at least a certain period of time after they rotate out of the sample. The algorithm which ensures that this requirement is satisfied is provided in Appendix A. Letting P_h and p_h respectively be the number of population panels and in sample panels for stratum h, we require that the sampling fraction f_h be approximately equal to p_h/P_h. In determining the number of population and sample panels, it is considered that keeping units in the sample for a little longer than the usual period is less of a response burden than allowing them back in the sample before the expiry of a required number of occasions that they must be kept out of the sample.

(iii) The next step is to assign the population units to the panels. Two cases arise:

a) The initial population size N_h is greater than or equal to the predetermined number of population panels P_h. That is, $N_h=m_hP_h+r_h$ when $m_h \geq 1$ and $r_h \geq 0$. Assuming that the population units have been randomized and numbered 1, 2, ..., N_h, they are sequentially and in modulo fashion assigned to the panel numbers 1 through P_h. The ordering 1, 2, ..., P_h of the panel numbers is termed as "assign ordering". Hence, unit 1 is assigned the first panel and so on, the unit P_h going to the P_h-th panel. The unit P_h+1 is again assigned to the first panel and so on. This eventually results in having the first "r_h" assign panels with (m_h+1) units and the next (P_h-r_h) assign panels with m_h units.

b) The initial population size N_h is less than the predetermined number of population panels P_h. That is, $N_h<P_h$. In this case, the N_h units are randomized and sequentially assigned equi-spaced panel numbers ranging from 1 to P_h resulting in N_h panels having one unit each and P_h-N_h empty or dummy panels.

(iv) The units are assigned a "rotation ordering". This is the ordering which will determine which units are in sample.

If $N_h \geq P_h$, then the rotation ordering is a random permutation of the P_h assign ordering. That is, each assign ordering i will be assigned a rotation ordering number r_{hi}, such that $1 \leq r_{hi} \leq P_h$ and $r_{hi} \neq r_{hj}$ for $i \neq j$. The population units acquire the r_{hi}-th rotation ordering if they had been originally assigned the i-th assigned ordering in stratum h.

If $N_h < P_h$, the rotation ordering is a random permutation of the N_h non-dummy panels labelled $t_1, t_2, \ldots, t_{N_h}$ where $1 \leq t_1 < t_2 \ldots < t_{N_h} \leq P_h$. Denoting this permutation as r_{hi}, where $1 \leq r_{hi} \leq P_h$ and $r_{hi} \neq r_{hj}$ for $i \neq j$, the population units acquire the r_{hi}-th rotation ordering if they had been originally assigned the i-th assign ordering in stratum h. The procedure for generating the non-dummy panel labels is provided in Appendix B.

Given that $N_h \geq P_h$ or $N_h < P_h$, the actual population units will be associated with C_h non-dummy panel numbers where $C_h = \min (N_h, P_h)$. Consequently, the initial sample consists of c_h non-dummy in-sample panels where $c_h \leq p_h$. The units belonging to the initial sample are those whose rotation ordering number is included in the closed sampling interval $[1, p_h]$.

2.3 Sample Rotation

On the first survey occasion, panels with rotation orders 1 to p_h are included in the sample. From then on, sample rotation is easily executed by shifting the sampling interval by one panel at each sampling occasion. On the t-th occasion, units in the sample are those population units whose rotation order is contained in the interval $[(t-1) \bmod P_h + 1, (t+p_h-1) \bmod P_h + 1]$ if $(t-1) \bmod P_h \leq (P_h - p_h)$ and in the union of the intervals $[1, (p_h - P_h) + (t-1) \bmod P_h]$ and $[(t-1) \bmod P_h + 1, P_h]$ otherwise. Effectively, rotation occurs by dropping a rotation panel from in-sample and acquiring a rotation panel from out-of-sample in a modular fashion.

2.4 Selection of Births

Births occur as a result of starting a new business activity, or a change of industrial activity of a unit from out-of-scope to in-scope for the survey.

Births will be first stratified according to the particular stratification used by the survey. Next, they will be sequentially given an assign ordering number as follows. Assuming that the last assign ordering number was ℓ_h, where $1 \leq \ell_h \leq C_h$, then the q_h-th birth will be given the assign ordering number $(\ell_h + q_h) \bmod C_h$. Assuming that b_h births have been processed, the new last assigned number to be used on the next occasion is $(\ell_h + b_h) \bmod C_h$. The rotation ordering is then immediately obtained through the one to one correspondence between the rotation and assign ordering. When $N_h < P_h$, births are only assigned to non-dummy panels. This procedure provides unbiased estimates.

2.5 Removal of Deaths

Deaths occur as a result of the termination of business activity for in-scope units or changes of industrial activity from in-scope to out-of-scope to the survey.

Deaths that emanate from a take-all stratum are removed immediately from the sample. Deaths that are part of the take-some sample are assigned a value of zero for estimation. Deaths within the take-some stratum are immediately removed from the frame only if they are identified as such from a source independent of the survey process. Deaths that are not thus removed will be taken off the frame only if they have been in this status for longer than a given time period (currently two years). The assumption behind this procedure is that, beyond this time period, all deaths would have been identified on a universal basis.

Giroux (1988) describes in detail the above functions.

2.6 Changes in Classification and Resampling

The sampling frame changes continuously due to births, deaths and changes of classification to population units, These changes in classification include changes in geography, industry and size. These changes will be detected more rapidly for in-

sample units than for out-of-sample units. Until all units in the population have been reclassified as of a given time period, changes in classification observed in the sample will be handled using domain estimation. That is, the data for a sampled unit will be tabulated in its current stratum using the sampling weight for the original stratum where it was selected. Furthermore, it must be noted that out-of-sample units do not have their classification changed even though changes have occured, until all units in the universe have been re-classified.

There are two procedures which can be used for handling changes in classification. We proceed first to describe the simpler one which can be summarized as follows. When it is known that a universal source has updated all units on the frame as of a given reference time period (which may be at least a year earlier than the current reference time period), all in-scope units to the survey (as of that reference period) are reclassified by industry, geography and size. This classification is next compared to the current one and those units for which the classification is different, are treated as deaths in the stratum where they originated and as births in the strata to which they currently belong. These births will then be selected within their new strata with the new selection probabilities. It must be noted that since not all units might have been re-classified, there will still be misclassified units on the frame. For sampled units, these misclassifications will continue to be handled via domain estimation. Also, although the procedure is quite simple to implement, it has several disadvantages. First, for in-sample units re-assigned as births, the time-in and time-out constraints may be violated, requiring special procedures in the contact and imputation systems which effectively by-pass the contact system and impute data for these units. Second, changes in classification may be severe enough to require the examination of the stratification and subsequent sampling rates. It is for these reasons that a second procedure, adopted from the Kish-Scott (1971) procedure, was introduced by Hidiroglou (1988). The procedure is based on the property that each panel is a simple random sample out of the population units and it is summarized in what follows.

For each new stratum, the required number of population and sample panels are computed using the new sampling rates as well as the time-in and time-out constraints which may also have changed. Each new stratum is broken out into a number of sets of units each of which has units coming from the same old stratum. These sets are mutually exclusive and exhaustive, and each set consists of units that

have the same new and old stratum classifications as defined in the Kish-Scott (1971) terminology. All the following operations occur within these sets. For each set, the sampling interval has proceeded at different rates and it is important to recalibrate all units so that they have the same starting point. This operation, basically a modulo operation, re-assigns the previous panel numbers to all units within each set in such a way that if P_h and p_h are the previous number of population and sample panels, then units in sample will be those whose re-assigned panel belongs to the interval $[1,p_h]$. Given that M_h units are in the origin set h, M_h numbers r_{hj} are generated such that $r_{hj} = (a_h+j-1)/M_h$ for $j=1, \ldots, M_h$ where a_h is a uniform random number in the interval $[0,1)$. The ranked (from low to high) r_{hj} numbers are then associated with each unit in such a way that the M_h population units within the origin set are ranked with their associated non-dummy panel numbers ranging from 1 to P_h. Note that ties, arising on account of having more than one sampling unit associated with a given non-dummy panel, are treated randomly. The r_{hj}'s are used as a bridge to migrate to the new panel numbering system. Next, the old and new sampling rates are compared. If the new sampling rate is greater than or equal to the current sampling rate, no changes occur with the r_{hj}'s. However, if the new sampling fraction is strictly smaller than the old sampling fraction, then the r_{hj}'s must be modified. This modification, which is necessary in order that the new sampling requirements, i.e. time-in and time-out requirements are met, is basically a shifting of the r_{hj}'s by a constant (the difference between the old and new sampling rates). Given that C_m non-dummy panel numbers are required for the m-th destination (new) stratum, C_m disjoint intervals between 0 and 1 are formed. The r_{hj}'s are then assigned the panel number $1, \ldots, C_m$ by observing to which interval they belong to, thus completing the resampling procedure.

This procedure has several advantages. First, it minimizes the impact of re-classification on the estimates, on cost and respondent contact by maximizing the overlap between the new and old sample. Second, the most recent stratification of the universe, new sampling rates and new time-in and time-out constraints are allowed.

3. WEIGHTING AND ESTIMATION STRATEGY

The requirements of the weighting and estimation strategy have been stated earlier in the introduction. The resulting system must be flexible, expandable, menu driven, with the input being provided by parameters (totals, means, ratios), domains of

interest as well as weighting procedures (simple expansion, ratio, Mickey unbiased ratio). The estimation system must also at the very least support the sampling system which has just been described above: i.e. stratified clustered design. Also, since the required tabulations are obtained by summing over many sampling strata, care must be taken to obtain unbiased or nearly unbiased estimators in order to avoid aggregation bias. Choudhry (1988) has investigated the biases and efficiencies of these estimates in an empirical study. Moreover, the influential observations, that is units with relatively high weights and unusually large values as compared to other units within the same stratum, must be detected and treated.

Computation of variance estimates will be made using the "d-values" introduced by Keyfitz (1957) or jackknifing procedures.

Composite estimation can be used to advantage for producing more efficient estimates of totals and trends.

3.1 Notation

As mentioned earlier, the stratum $h (h = 1, 2, \ldots, L)$ is defined at some given level of industry, geography and size. It is at this level that the basic sampling occurs. Domain estimation will be used to generate tables since it is a general procedure that also permits the movement of units between strata (strata jumpers). Domains are defined using existing data on the files (such as industry, geography, size and other characteristics) and their definition can be quite varied. A domain can span across all the sampling strata, be a subset of these strata or be defined within these strata. Examples of such domains, at the geographical level, are aggregations at the Canada level, the provincial level or the sub-provincial level (even though the sampling had occured at a higher level). Consequently, the sum of any domain set must always add up to the domain defined as their union. For this reason, we have opted to have all the estimation carried out separately in each stratum.

Let y_{hij} be the y-value for the j-th unit in panel (cluster) i of stratum h. Let $_d\delta_{hij}$ be an indicator variable defined as 1 if the hij-th observation belongs to domain d, and 0 otherwise. Then, the parameter of interest is the population total $_dY$ given by:

$$_dY = \sum_h \sum_i \sum_j {}_d\delta_{hij}\, y_{hij}$$

$$= \sum_h \sum_i \sum_j {}_dy_{hij}$$

where

$$_dy_{hij} = y_{hij} \text{ if } hij \in d$$

$$= 0 \text{ otherwise.}$$

We will consider a number of alternative estimators $_d\hat{Y}$ for the population parameter $_dY$ and also their variance estimators $v(_d\hat{Y})$. As described earlier, we have a sample of c_h out of C_h panels selected with simple random sampling. Let N_{hi} be the number of units in the i-th sample panel. Without loss of generality, we can assume that the c_h sampled panels are indexed i=1, 2, ..., c_h. Let $_dy_{hi}$ be the total response from the i-th sampled panel, i.e.

$$_dy_{hi} = \sum_{j=1}^{N_{hi}} {}_dy_{hij}.$$

All the estimators considered at the stratum level will be of the form

$$_d\hat{Y}_h = \sum_{i=1}^{c_h} w_{hi}\, {}_dy_{hi}$$

where w_{hi} is the weight of each unit within the i-th sampled panel. It is important to write the estimators in the above form because we can associate a weight with each observation on the output microdata file. Estimators of $_dY$ are obtained by aggregating over strata, that is,

$$_d\hat{Y} = \sum_{h=1}^{L} {}_d\hat{Y}_h$$

and the corresponding variance is $v(_d\hat{Y}) = \sum_{h=1}^{L} v(_d\hat{Y}_h)$. From now on, we will only work at the stratum level. Estimators and variance estimators are summed up from that level.

3.2 Estimators of totals

A. Simple Expansion Estimator

The probability of selecting each panel is c_h/C_h. Therefore, the design weight is $w_{hi}=C_h/c_h$ for $i=1, 2,\ldots,c_h$. We denote the estimator as $\hat{Y}_{d\,h(S)}=C_h/c_h \sum_{i=1}^{c_h} \hat{Y}_{d\,hi}$.

Although this estimator is unbiased, it may not be very efficient because it does not make use of available auxiliary information concerning panel sizes. As the variation in the panel sizes increases, this estimator becomes more and more inefficient over time. The estimated variance of the estimator is given by:

$$v(\hat{Y}_{d\,h(S)}) = (1-f_h) \frac{c_h}{c_h-1} \sum_{i=1}^{c_h} (z_{hi} - \bar{z}_{h.})^2$$

where $z_{hi} = w_{hi}\, {}_d y_{hi}$, $\bar{z}_{h.} = c_h^{-1} \sum_{i=1}^{c_h} z_{hi}$ and $f_h = c_h/C_h$.

Note that covariances between variables can easily be obtained by casting the above formula in its covariance analogue. For more details, see PC CARP (1986).

B. Separate Ratio Estimator

If the correlation between ${}_d y_{hi}$ and panel sizes N_{hi}, (where size is the number of units in each panel) is large, then efficiency gains can be realized through the ratio estimator defined as:

$$\hat{Y}_{d\,h(R)} = \frac{N_h}{\hat{N}_h} \hat{Y}_{d\,h(S)}$$

where

$$\hat{N}_h = \frac{C_h}{c_h} \sum_{i=1}^{c_h} N_{hi} = \frac{C_h}{c_h} n_h.$$

The ratio estimator can also be written as:

$$\hat{Y}_{d\,h(R)} = \frac{N_h}{n_h} \sum_{i=1}^{c_h} {}_d y_{hi}.$$

Hence, the weight N_h/n_h is the inverse of proportion of units in the sample instead of the inverse of the proportion of panels in the sample.

One major drawback of this estimator is that it is subject to the ratio estimation bias. Consequently, if the bias tends to be positive or negative in the majority of the strata, its accumulated effect can be quite significant when aggregating over strata. The estimated variance of $\hat{Y}_{d(R)}$ for large c_h is:

$$v(\hat{Y}_{d(R)}) = (1-f_h) \frac{c_h}{c_h-1} \sum_{i=1}^{c_h} (z_{hi}-\bar{z}_{h.})^2$$

where

$$z_{hi} = w_{hi}[{}_d y_{hi} - N_{hi} c_h^{-1} \sum_{i=1}^{c_h} {}_d y_{hi}]$$

and $\bar{z}_{h.}$ is as defined as in Section 3.2.A.

C. Unbiased and Nearly Unbiased Ratio Estimators

We have considered a number of unbiased (or nearly unbiased) ratio estimators including Quenouille's (1956) jackknife estimator as well as Mickey's (1959) unbiased estimator. The appeal of these estimators is that they greatly reduce (or eliminate) ratio bias, especially when aggregated over strata.

We first start with nearly unbiased estimators. Quenouille's estimator is given by

$$\hat{Y}_{d h(Q)} = N_h c_h (1 - \frac{c_h-1}{c_h}) r_h - N_h (c_h-1)(1-\frac{c_h}{c_h}) \frac{1}{c_h} \sum_{j=1}^{c_h} r_h^{(j)}$$

where r_h is the ratio of the sum of the y-values to the sum of the sizes of the sampled panels and $r_h^{(j)}$ is the ratio over remaining (c_h-1) panels when the j-th panel is dropped. The weight for this estimator is:

$$w_{hi} = N_h c_h (1 - \frac{c_h-1}{c_h})(\frac{1}{n_h} - b_{hi}) + N_h b_{hi}$$

with $$b_{hi} = \frac{1}{c_h} \sum_{j(\neq i)} \frac{1}{n_h - N_{hj}}.$$

The bias of the above estimator is of order $1/c_h 2$ for large c_h. Note that this bias can be further reduced to order $1/c_h 3$ using a Taylor expansion. The resulting estimator is

$$\hat{Y}_{d\,h(J)} = N_h \frac{c_h 2}{2} \left(1 - \frac{c_h - 1}{c_h}\right) \left(1 - \frac{c_h - 2}{c_h}\right) r_h$$

$$- N_h (c_h - 1)^2 \left(1 - \frac{c_h}{C_h}\right)\left(1 - \frac{c_h - 2}{c_h}\right) \frac{1}{c_h} \sum_j r_h^{(j)}$$

$$+ 2N_h \left(\frac{c_h - 2}{2}\right)^2 \left(1 - \frac{c_h}{C_h}\right)\left(1 - \frac{c_h - 1}{c_h}\right) \frac{1}{c_h(c_h - 1)} \sum_{j<k} \sum r_h^{(j,k)}$$

where $r_h^{(j)}$ is as defined above and $r_h^{(j,k)}$ is the ratio based on $(c_h - 2)$ units obtained by deleting two panels (j and k) at a time.

The variance estimators for $\hat{Y}_{d\,h(Q)}$ and $\hat{Y}_{d\,h(J)}$ can be obtained by using the jackknife method.

Next, we introduce Mickey's (1959) unbiased ratio estimator which is given by:

$$\hat{Y}_{d\,h(M)} = \frac{N_h}{c_h} \sum_{j=1}^{c_h} r_h^{(j)} + (C_h - c_h + 1) \left\{ \sum_{i=1}^{c_h} d^y{}_{hi} - \frac{n_h}{c_h} \sum_{j=1}^{c_h} r_h^{(j)} \right\}$$

where $r_h^{(j)}$ is as defined in Quenouille's estimator. The weighted form of the estimator is

$$\hat{Y}_{d\,h(M)} = \sum_{i=1}^{c_h} w_{hi}\, d^y{}_{hi}$$

where $w_{hi} = [N_h - n_h(C_h - c_h + 1)]\, b_{hi} + (C_h - c_h + 1)$

with b_{hi} as defined before.

The corresponding estimator of variance is obtained via a jackknife procedure, leaving out one panel at a time and re-computing Mickey's estimator for the remaining $(c_h - 1)$ panels in the sample. Denote each jackknifed estimator as $\hat{Y}_{d\,h(M)}^{(\ell)}$ for $\ell = 1, 2, \ldots, c_h$ where

$$\hat{d}Y^{(\ell)}_{h(M)} = \sum_{i(\neq \ell)} w^{(\ell)}_{hi} \; d^{y}_{hi}$$

with $$w^{(\ell)}_{hi} = [N_h-(n_h-N_\ell)(C_h-c_h+2)]b^{(\ell)}_{hi} + (C_h-c_h+2)$$

and $$b^{(\ell)}_{hi} = \frac{1}{c_h-1} \sum_{j(\neq i,\ell)} \frac{1}{n_h-N_{hj}-N_{h\ell}}.$$

The jackknife variance estimator is then given by

$$v_J(\hat{d}Y_{h(M)}) = (1-f_h) \frac{c_h-1}{c_h} \sum_{\ell=1}^{c_h} (z_{h\ell}-\bar{z}_{h.})^2,$$

where $$z_{h\ell} = \hat{d}Y^{(\ell)}_{h(M)} \text{ and } \bar{z}_{h.} = c_h^{-1} \sum_{\ell=1}^{c_h} z_{h\ell}.$$

Note that the above variance form can be used for all proposed estimators by suitably defining $w^{(\ell)}_{hi}$. For the simple expansion estimator,

$$w^{(\ell)}_{hi} = C_h/(c_h-1)$$

and for the ratio estimator it is

$$w^{(\ell)}_{hi} = N_h/(n_h-N_{h\ell}).$$

3.3 Estimator of a Ratio

It is often required to estimate the ratio between the totals of two variables, e.g. wages and salaries or a percentage of GBI, and the corresponding variance of the estimated ratio has to be estimated as well. Let x and y be the two variables of interest and the ratio $_dR = {}_dY/{}_dX$ is to be estimated, where

$$d^{Y} = \sum_{h=1}^{L} \sum_{i=1}^{C_h} d^{y}_{hi} \quad \text{and} \quad d^{X} = \sum_{h=1}^{L} \sum_{i=1}^{C_h} d^{x}_{hi}$$

$_dy_{hi}$ and $_dx_{hi}$ are respectively the values of y and x variables for the domain d corresponding to the i-th panel in stratum h. Then the ratio $_dR$ is estimated by $_d\hat{R} = {}_d\hat{Y}/{}_d\hat{X}$, where

$$d\hat{Y} = \sum_{h=1}^{L} \sum_{i=1}^{c_h} w_{hi}\, d^{y}_{hi} \quad \text{and} \quad d\hat{X} = \sum_{h=1}^{L} \sum_{i=1}^{c_h} w_{hi}\, d^{x}_{hi}.$$

The variance of the estimated ratio $d\hat{R}$ will be estimated by the Taylor linearization technique using the jackknife variance estimator.

3.4 Influential Observations

A module will monitor the behaviour of units which dominate or have impact on the estimate within a stratum. The detection of such units will be carried out by basically estimating the weighted sample distribution and isolating units which are too far in the right tail. The impact of such units can be reduced by i) either reducing their weight to one and subsequently modifying the weights of the remaining units in the stratum by ensuring that the sum of the weights over all units add up to the stratum population size (Hidiroglou-Srinath 1981) or by ii) Winsorizing the observations as in Fuller (1970). The Winsorization effectively brings back the values of influential observations to a boundary which is determined from the estimated sample distribution. It must be noted that for the Winsorization procedure, the original data are untouched. They are dampened (reduced) by a deflation factor which is stored for the occasion and computed automatically by the computer in order to produce the required Winsorization. Both of these methods lead to negative bias in the estimates.

The main problems associated with the treatment of influential observations are as follows:

i) At what level of aggregation should their detection occur?

ii) How robust are the estimates to the assumption that they represent unique observations in the population?

iii) How much bias is acceptable?

iv) How continuous (smooth) should be the published results between survey occasions?

3.5 Estimator of Trend

Tam (1984) provided expressions for covariances of repetitive sampling plans of the same finite population over time. Hidiroglou and Laniel (1986), and Laniel (1987)

provided covariance expressions for rotating samples in a changing population. These expressions are necessary in order to compute variances for functions of totals, such as trends. An adaptation of these expressions is provided, in the context of the panel design.

Let $p_h^{(t)}$ and $p_h^{(t+\tau)}$ denote the sampled panel set from stratum h for the two occasions at times t and $t+\tau$. Note that τ is less than both the time-out constraint and p_h. Let $p_h^{(c)}$ denote the intersection of the sampled panel sets for the two given occasions. Furthermore, let $y(t)$ and $y(t+\tau)$ denote the observed values for a given panel at these two time periods.

The estimator for the trend between the two time periods will be defined as:

$$_d\hat{R} = {}_d\hat{Y}^{(t+\tau)} / {}_d\hat{Y}^{(t)}$$

where ${}_d\hat{Y}^{(t+\tau)}$ and ${}_d\hat{Y}^{(t)}$ denote the estimators of population totals. For the two given occasions and the domain of interest, the estimated variance of the trend is given by:

$$v({}_d\hat{R}) = ({}_d\hat{Y}^{(t)})^{-2}[v({}_d\hat{Y}^{(t+\tau)}) + {}_d\hat{R}^2 v({}_d\hat{Y}^{(t)}) - 2\,{}_d\hat{R}\ \mathrm{cov}({}_d\hat{Y}^{(t)}, {}_d\hat{Y}^{(t+\tau)})].$$

The variances of ${}_d\hat{Y}^{(t+\tau)}$ and of ${}_d\hat{Y}^{(t)}$ are estimated as explained earlier. For the covariance, the jackknifing technique can be used by deleting one panel at a time from the common sampled panel sets $p_h^{(c)}$. Since $\tau < c_h$, the set $p_h^{(c)}$ is non-empty. Note that for the simple expansion estimator, the covariance expression is:

$$\mathrm{cov}({}_d\hat{Y}^{(t)}, {}_d\hat{Y}^{(t+\tau)}) = \sum_{h=1}^{L} (1-f_h) \frac{p_h^{(c)}}{p_h^{(c)}-1} \sum_{i\varepsilon p_h^{(c)}} ({}_d\alpha_{hi} - {}_d\bar{\alpha}_{h.})({}_d\beta_{hi} - {}_d\bar{\beta}_{h.})$$

where f_h is the sampling fraction in stratum h, $p_h^{(c)}$ is the number of panels in the common panel set $p_h^{(c)}$ (equal to $c_h - \tau$),

$${}_d\alpha_{hi} = w_{hi}\, {}_d y_{hi}^{(t)},\ {}_d\beta_{hi} = w_{hi}\, {}_d y_{hi}^{(t+\tau)},$$

$${}_d\bar{\alpha}_{h.} = (p_h^{(c)})^{-1} \sum_{i\varepsilon p_h^{(c)}} {}_d\alpha_{hi} \text{ and } {}_d\bar{\beta}_{h.} = (p_h^{(c)})^{-1} \sum_{i\varepsilon p_h^{(c)}} {}_d\beta_{hi}$$

3.6 Composite Estimator

In a rotating sample design, samples overlap between two successive occasions, and there is correlation between the common units over time. The efficiency of the estimator can therefore be improved by making use of the correlation through composite estimation. In the panel design, this correlation may be somewhat weaker as compared to a non-clustered design. However, as noted before, the panel design has the advantages of simple covariance computation and rotation implementation in an unbiased fashion. We assume that the two time periods are t and t+1. Moreover, without loss of generality, we can assume that the sampled panels in stratum h are numbered 1, 2, ..., c_h at time t, and these are 2, 3, ..., c_h+1 at time t+1. Thus there are (c_h-1) common panels, i.e. 2, 3, ..., c_h between the two time periods, where h=1, 2, ..., L. Again using Mickey's estimator, let

$w_{hi}^{(t)}$ = weight for the i-th panel in stratum h at time period t.

Then

$$ {}_d\hat{Y}_{(M)}^{(t)} = \sum_{h=1}^{L} \sum_{i=1}^{c_h} w_{hi}^{(t)} \; {}_d y_{hi}^{(t)} $$

where ${}_d\hat{Y}_{(M)}^{(t)}$ is the estimated total of y for domain d at time t.

Similarly, the estimate of total of y-variable for time (t+1) is given by

$$ {}_d\hat{Y}_{(M)}^{(t+1)} = \sum_{h=1}^{L} \sum_{i=2}^{c_h+1} w_{hi}^{(t+1)} \; {}_d y_{hi}^{(t+1)} . $$

Now using the (c_h-2) common panels, we can estimate the change (or difference) between the two time period, i.e.

$$ {}_d\hat{D}^{(t,t+1)} = \sum_{h=1}^{L} \sum_{i=2}^{c_h} w_{hi}^{(c)} \left\{ {}_d y_{hi}^{(t+1)} - {}_d y_{hi}^{(t)} \right\} $$

where $w_{hi}^{(c)}$ is the Mickey weight based on the (c_h-2) common panels, at time t+1.

Then, we can construct a difference estimator of the total of y-variable for (t+1) as

$$ {}_d\tilde{Y}^{(t+1)} = {}_d\hat{Y}_{(M)}^{(t)} + {}_d\hat{D}^{(t,t+1)} . $$

The composite estimator of the total of y for time (t+1) is given by

$$d\hat{Y}_{(C)}^{(t+1)} = \alpha\, d\hat{Y}_{(M)}^{(t+1)} + (1-\alpha)\, d\tilde{Y}^{(t+1)}$$

where the value of the parameter α lies between zero and one, and the optimum value of α (α_{opt})is obtained such that the variance of $d\hat{Y}_{(C)}^{(t+1)}$ is minimized. Since α_{opt} is not known, it must be estimated from the survey data. However, since this would be difficult in practice, we use a compromise value as suggested by Hansen et al. (1953). We also note that the composite estimator $d\hat{Y}_{(C)}^{(t+1)}$ is unbiased for a fixed α. Furthermore, the variance could be estimated by the jackknife method.

Acknowledgement

The authors would like to thank Judy Clarke and Pat Pariseau for their excellent typing.

APPENDIX A

DETERMINATION OF THE NUMBER OF PANELS

Let N = population size

n = sample size

T_{in} = desired number of occasions a unit should stay in the sample

T_{out} = minimum required number of occasions a unit must stay out of the sample

f = sampling fraction.

If the minimum number of occasions a unit stays out of the sample is a prerequisite, then the following algorithm ensures that this will occur.

Steps a) Compute

$$x = \text{int} \left| T_{in} \frac{1 - f}{f} + 0.5 \right| .$$

b) If $x \geq T_{out}$, then the number of in-sample panels is

$$P_{in} = T_{in}$$

and the number of out-of-sample panels is

$$P_{out} = x$$

c) If $x < T_{out}$, then the number of in-sample panels is

$$P_{in} = \text{int} \left| \frac{f}{1 - f} T_{out} + 0.5 \right|$$

and the number of out-of-sample panels is

$$P_{out} = T_{out} .$$

The number of population panels is $P = P_{in} + P_{out}$ and the number of sample panels in $p = P_{in}$.

Example 1. Suppose $N = 14$, $n = 6$, $T_{in} = 24$ and $T_{out} = 12$, then it can be verified that $P_{in} = 24$ and $P_{out} = 32$ so that $P = 56$. In this case, the unit stays in sample the desired number of occasions but must stay out-of-the sample longer than 12 occasions.

Example 2. Suppose $N = 75$, $n = 30$, $T_{in} = 12$ and $T_{out} = 12$, then it can be verified that $P_{in} = 12$ and $P_{out} = 18$. Consequently there are $P = 30$ population panels.

APPENDIX B

GENERATION OF ROTATION ORDERING WHEN N < P

This algorithm creates a numbering for the N non-dummy panels between 1 and P, in order that they are as equispaced as possible. The steps are as follows:

a) Compute s and q such that

$$P = sN + q \text{ where } q < N \text{ and } s \geq 0.$$

b) Compute d_j (j = 1, 2, ..., N) numbers assuming the values 0 or 1, such that q of them have the value equal to 1 and N - q of them have the value equal to 0. The generation of 0's and 1's must be random.

c) Select a random number "a" between 1 and P inclusively ($1 \leq a \leq P$).

d) Compute $z_1 = | a + d_1 - 1 | \bmod P + 1$.

e) Compute $z_j = | z_{j-1} + s + d_j - 1 | \bmod P + 1$ for $j = 2, \ldots, N$.

f) The ranking of $z_1, \ldots, z_N$ from low to high yields $t_1, \ldots, t_N$.

References

Bankier, M.D. (1988), "Power Allocations: Determining Sample Sizes for Subnational Areas," *The American Statistician*, 42, 174-177.

Cochran, W.G. (1977), "Sampling Techniques," 3rd Edition, John Wiley, New York.

Choudhry, G.H. (1988), "Generalized Estimation and Variance system for sub-annual surveys," Statistics Canada Technical Memorandum, April 1988.

Dalenius, T. and Hodges, Jr., J.L. (1959), "Minimum Variance Stratification," *Journal of the American Statistical Association*, 54, 88-101.

Dumais, J., and Carpenter R. (1988), "Methodology of the Generalized Estimation System for sub-annual Business Surveys," Statistics Canada Technical Memorandum, September 1988.

Dumais, J. (1988), "Generalized Estimation System Interfaces," Statistics Canada Technical Memorandum, December 1988.

Fuller, W. A. (1970), "Simple Estimators for the Mean of Skewed Populations," Technical Report prepared for the U.S. Bureau of the Census, Iowa State University, Dept. of Statistics.,

Fuller, W.A., Kennedy W., Schnell, D., Sullivan, G., and Park, H.J. (1986), "PC CARP," Statistical Laboratory, Iowa State University.

Giroux, S. (1988), "MWRTS Sample Selection System Update," Statistics Canada Technical Memorandum, July 1988.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953), "Sample Survey Methods and Theory," John Wiley and Sons, New York.

Hidiroglou, M. A., and Srinath, K.P. (1981), "Some Estimators of a Population Total From Simple Random Samples Containing Large Units," *Journal of the American Statistical Association*, Vol. 76, No. 375, 690-695.

Hidiroglou, M.A. (1986), "The Construction of a Self-Representing Stratum of Large Units in Survey Design," *The American Statistician*, Vol. 40, 27-31.

Hidiroglou, M. A., and Laniel, N. (1986), "Specifications For the Estimation System of the Monthly Wholesale Retail Trade Survey," Statistics Canada Technical Memorandum, September 1986.

Hidiroglou, M. A. (1987), "Sample Rotation," Statistics Canada Technical Memorandum, March 1987.

Hidiroglou, M.A., and Srinath K.P. (1987), "A General Strategy for Selection, Rotation and Maintenance of Samples For Business Surveys," Statistics Canada Technical Memorandum, September 1987.

Hidiroglou, M.A. (1988), "Redrawing Rotating Samples After Changing Stratification and Sampling Rates," Statistics Canada Technical Memorandum, November 1988.

STATISTICS CANADA LIBRARY
BIBLIOTHEQUE STATISTIQUE CANADA
1010207441