TWO-PHASE SAMPLE DESIGN FOR TAX DATA

G.H. Choudhry, P. Lavallée, and M.A. Hidiroglou
Statistics Canada, Ottawa (Ontario), Canada K1A 0T6

Keywords: Two-Phase sample design, Optimum sample allocation, Poisson Sampling, Post-Stratification.

1. Introduction

The Business Survey Redesign Project has been a major undertaking at Statistics Canada (STC) over the last three years. The main objectives of this project have been to standardize the concepts, co-ordinate procedures, and reduce cost and response burden. This is being achieved through mandatory use of the Central Frame Data Base (CFDB) for provision of survey frames for all business surveys, and through the application of generalized survey methodology during all phases of survey processing. An overview of the generalized strategy can be found in Colledge (1987), and Colledge and Lussier (1987). In order to reduce cost and response burden, the use of administrative data, both as a basis for building survey frames and as a replacement for direct data collection has also been a crucial part of the strategy (Colledge and Lussier 1985).

The Central Frame Data Base has two components: the integrated portion (IP) which consists of all the large businesses with total operating revenue exceeding certain boundary values, specified separately for each industry within province/territory, and the non-integrated portion (NIP) which consists of all the businesses not in IP. The businesses in IP are divided into statistical units. There is a hierarchical structure of these statistical units which are linked to one another and tracked longitudinally over time. The businesses in NIP are represented by two separate unlinked sets of administrative data. The first set, based on payroll deduction (PD) information from Revenue Canada Taxation (RCT), provides frames for sub-annual surveys. Only account holders with remittances greater than zero are in scope for this set. The second set based on income tax data from RCT, consists of a sample of the individual (T1) and corporate (T2) business tax filers indicating total operating revenue greater than or equal to $25K. The businesses associated with tax filers reporting total operating revenue less than $25K are out of scope for business surveys.

The data on economic production are currently collected by independent survey operations, each covering a particular group of industries. Some of these are, the Annual Survey of Manufacturers, the Census of Construction, the Annual Wholesale Trade Survey, the Motor Carrier Freight Survey, and the Annual Services Programme. In the context of the redesign objectives, annual economic data for the IP units will be obtained by direct census survey using personalized questionnaire, whereas for the NIP units, the data will be obtained largely through the transcription of a sample of tax records to reduce cost and response burden. Non-financial data for the NIP units will be collected from a sub-sample of these units. In the future, the economic production surveys will be based on this sample design, using a common frame extracted from the CFDB, and standardized survey methodology, so that they will be viewed collectively as a single, comprehensive "Annual Programme of Economic Production Statistics (APEPS)".

As mentioned earlier, the frame for the annual economic data will contain a set of statistical units generated automatically from business operating structures or from tax returns within the CFDB. The ideal target population, which consists of all businesses with economic activity during the reference calendar year, is not practical due to timeliness of the data. An approximation to the above conceptual target population is considered which consists of all businesses with their fiscal periods ending during a specified twelve month period. Further details related to frame extraction can be found in Roberge (1988). The data from the respondents will be collected for their fiscal periods to reduce response burden. Subsequently, an adjustment will be made to convert the respondent fiscal period data to calendar year. The adjustment procedures have been described and evaluated by Laniel and Poirier (1989).

The tax records at RCT are coded at 2-digit Standard Industrial Classification (SIC2) level while, the economic statistics are required at the 4-digit Standard Industrial Classification (SIC4) level for each province/territory. In the past, a single phase design has been used to select the sample of tax records which were stratified on the basis of Province/territory × SIC2 (or SIC3) × Size, where size categories were defined on the basis of total operating revenue. The sample within each stratum was selected by Poisson sampling at RCT. The advantages of Poisson sampling are two-fold. First, it ensures that a uniquely identified tax filer can be sampled over time, thereby permitting longitudinality of data. Second, rotation of sampled units can be controlled by properly advancing the sampling interval. With the single phase design, the resulting sample sizes at the SIC4 level were random, subject to sampling variation. Moreover, the expected sample sizes at the SIC4 level depended on the distribution of SIC4's within the SIC2. Thus the reliability of the survey estimates varied considerably from one industry to another.

To overcome this problem, a two-phase sample design for tax records will be implemented. It will allow more efficient control of the resulting sample sizes at the SIC4 level. The first phase which comprises a relatively large sample of tax records (T1 and T2) will be built over a period of several years. It will be referred to as the master sample. All the businesses associated with the selected tax records will be coded to the SIC4 level. In the second phase, a subsample from the master sample will be selected at the SIC4 level. Operational details about the two-phase sample design are provided by Foy (1987). The two-phase sample design will thus provide the flexibility of controlling the sample sizes at the SIC4 level and hence the reliability of the resulting estimates. The optimum allocation of the sample at the two phases of sampling will be obtained by non-linear optimization techniques.

The master sample with the corresponding SIC4 codes will be maintained on the CFDB. It will be updated with respect to births, deaths, and classification changes such as industry, geography and size. The geography and size review can be done immediately because it is available on a universal basis.

The industry (SIC4) review will be done on a cyclical basis so that each unit gets reviewed at least once over a certain specified period of time. The only exception will be where the SIC2 code has changed, in which case the unit will be reviewed and updated immediately. The second phase sample can be rotated within the first phase sample over time.

The aspects of the redesign strategy that are addressed in this paper include (i) sample design and (ii) weighting and estimation. The weights will be based on a post-stratified ratio estimator for each phase of sampling. The resulting ratio estimator for the domain of interest and the corresponding variance estimator are provided in this paper.

2. Sample Design

2.1 Background

The estimates of economic production are required at the province/territory (Prov) by 4-digit standard industrial classification (SIC4) level and must satisfy prespecified expected coefficients of variation (CV). The SIC codes available from Revenue Canada Taxation (RCT) are generally at the 2-digit level and in some cases at the 3-digit level. Therefore, the stratification by SIC can only be carried out at the 2-digit or 3-digit SIC level. The stratification is also carried out by size to improve the reliability of the estimates where the size catagories are based on the total operating revenue.

The method used for size stratification is given by Hidiroglou (1986), and Latouche and Hidiroglou (1987). The stratification by size is the same across all SIC's and provinces due to operational simplicity, even though there would be some loss in efficiency.

In the past, one-phase sample design with Poisson Sampling was used for the tax sample. The main difficulty with this scheme is that the resulting sample sizes at the SIC4 level are random variables. Moreover, the expected sample sizes at the SIC4 level depend on the SIC4 distribution within the SIC2. It is for this reason that it was decided to select a larger sample of tax filers (T1 and T2) at the first phase and code the businesses associated with the selected tax records to the SIC4 level. At the second phase, the sampling is to be carried out within each Prov × SIC4 × Size level. Thus, the two-phase design gives the flexibility of achieving the desired sample sizes at Prov × SIC4 level. In order to describe the two-phase sample design, we first outline the existing one-phase sample design as it is currently implemented. The proposed two-phase sample design which will be implemented for tax year (TY) 89 will then be described in detail.

2.2 One-Phase Sample Design

Since one cycle of sampling lasts for a period of two years, the information available for allocation is also two years out-of-date. Thus, for a given tax year y, TY(y), the stratum sampling rates for the required level of reliability are obtained from the TY(y-2) data. The sampling rates are adjusted to take into account the growth between TY(y-2) and TY(y). The stratification is carried out by Province × SIC2 (or SIC3) × Size, where size categories are defined on the basis of total operating revenue for the tax entity (TE). Let i denote the TE within a given province where i is the Social Insurance Number (SIN) for the individual (or unincorporated) tax filer (T1-type) and it is the corporation number for the corporate tax filer (T2-type). Also, let e denote the SIC2 (or SIC3) code and g the size group within a province, where e=1, 2, ..., E, and g=1, 2, ..., G. Due to operational reasons, the same size groups are used across all SIC2 (or SIC3) codes and for all provinces, although there would be some loss in efficiency. Thus the size groups are independent of the SIC code and the province. In order to simplify the notation, we shall omit the province subscript.

Suppose ν_{eg} denotes the sampling rate in the stratum defined by SIC2 code e and size group g. Now, for a TE denoted by i we generate a pseudo-random number (hash number) R_i defined as:

$$R_i = \text{hash}_1(i) \text{ where } 0 \leq R_i < 1.$$

Different hash functions may be used to select samples of T1 and T2 type tax entities. Now, if $TE(i) \varepsilon (e,g)$ and $R_i \leq \nu_{eg}$ then TE(i) is selected. Thus the $TE(i) \varepsilon (e,g)$ for a given province is selected with probability ν_{eg}. Due to Poisson sampling, the selection probabilities of TE's are independent of each other. Although the allocation was carried out using TY(y-2) data, the sampling at RCT is based on TY(y) data. Thus the units get selected with their most up-to-date classification, i.e. province, SIC codes, and size.

2.3 Two-Phase Sample Design

In the one-phase design, the expected sample sizes at the SIC4 level depend on the distribution of SIC4's within the SIC2 codes. In order to satisfy the CV requirements for all estimates at the Prov × SIC4 level, the overall sample size could be very large. However, the CV's for several SIC4 estimates would be smaller as compared to the target CV. The SIC4 estimates are obtained as domain estimates and the sample size at the SIC2 level must be large enough to satisfy the CV requirement for each and every domain.

In the two-phase sample design, a relatively large first-phase sample of TE's is selected and all business entities (BE's) associated with the selected TE's are coded to SIC4 at Statistics Canada (STC). The first-phase sample (or master sample) with SIC4 codes at the BE level is maintained on the CFDB with respect to births, deaths, and classification changes. The second-phase sampling is carried out at STC from the master sample using the Poisson procedure. Although the master sample will be continuously updated, the latest information for updating for TY(y) sample is based on TY(y-2) data. Thus, the population births will have to be sampled at RCT. These births are all the ones that occurred between TY(y-2) and TY(y). The copies of the tax records selected at the second phase and all the births sampled at the first phase are obtained from RCT. The selection probabilities corresponding to the two phases of sampling are as follows.

i) First-Phase Selection Probabilities

For the tax sample, the frame is built up of tax entities (TE's). As mentioned earlier, the frame data available for stratification and selection is 2 years out-of-date. Thus, for TY(y), the frame is based on tax data for TY(y-2). The stratification for first phase is carried out by Prov × SIC2 (or SIC3) × Size, where size category is defined on the basis of the total operating revenue for the TE. Let i denote the TE where i is as defined earlier. Also, let g denote the size group within a province (or territory) and e denote the SIC2 (or SIC3) code, where g=1, 2, ..., G and e=1, 2, ..., E. Suppose ν^1_{eg} denotes the first-phase sampling rate in the stratum denoted by (e,g). The sampling is done

using the Poisson procedure. For a unit denoted by i, we generate a pseudo-random number (hash number) R_i as defined in Section 2.2, with $0 \leq R_i < 1$. Now, if $TE(i) \in (e,g)$ and $R_i \leq \nu'_{eg}$, then TE(i) is selected. The occurrence of $TE(i) \in (e,g)$ is determined on the basis of TY(y-2) information. Note that between TY(y-2) and TY(y), the TE(i) could have changed stratum, say from (e,g) to (e*,g*). Therefore, the probability of selecting TE(i) for TY(y) is given by

$$p1_i = \pi_i(y\text{-}2) + \pi_i(y \mid \text{not } y\text{-}2),$$

where $\pi_i(y\text{-}2)$ is the probability that TE(i) was selected from the frame constructed with TY(y-2) data and $\pi_i(y \mid \text{not } y\text{-}2)$ is the conditional probability that TE(i) was selected from the frame constructed with TY(y) data, given that it was not selected from the frame constructed with the TY(y-2) data.

It is easy to show that $p1_i$ is given by

$$p1_i = \max(\nu'_{eg}, \nu'_{e^*g^*})$$

where $TE(i) \in (e,g)$ for TY(y-2) and $TE(i) \in (e^*,g^*)$ for TY(y) with ν'_{eg} and $\nu'_{e^*g^*}$ being the corresponding first-stage sampling rates. For births that have occured between TY(y-2) and TY(y), ν'_{eg} is equal to zero.

For the sampled TE's (T1 and T2), all businesses are enumerated and coded to SIC4. For the corporate tax filers (T2-type) there is a one-to-one correspondence between the tax filer and the business. However, for the individual tax filers (T1-type), any given tax filer may own one or more businesses and there may be several partners in any given business. Let j denote the business entity (BE), where a business entity is defined as an economic activity for the purpose of production of goods and/or services. For a more general definition of business entity, see Statistics Canada (1986). Then the statistical entity (SE) will be defined uniquely by the pair of indices (i,j). Let s_1 denote the set of TE's in the first phase sample and J_i^* be the BE's for TE(i) from TY(y) data, then the probability of selecting the SE(i,j) is given by

$$p1_{ij} = p1_i \text{ for all } j \in J_i^*, \; i \in s_1$$

i.e. the probability of selecting any of the statistical entities (SE's) within a TE is the same as that of selecting the TE. Note that the statistical structure given by J_i^* for TE(i) is based on TY(y) information.

In the above sampling scheme, we have used the interval $[0, \nu'_{eg})$ for Poisson sampling. Alternatively we could have generated a random number $\nu_{eg}^{(0)}$ between 0 and 1 for stratum (e,g), and then define the hash interval H_{eg} as:

$$H_{eg} = [\nu_{eg}^{(0)}, \nu_{eg}^{(0)} + \nu'_{eg}) \text{ if } \nu_{eg}^{(0)} + \nu'_{eg} \leq 1$$

$$= [\nu_{eg}^{(0)}, 1) \cup [0, \nu_{eg}^{(0)} + \nu'_{eg} - 1), \text{ otherwise.}$$

$H_{e^*g^*}$ can also be defined in a similar manner. Then $p1_i$ will be given by the size measure of the set defined by $H_{eg} \cup H_{e^*g^*}$. Note that when $\nu_{eg}^{(0)} = \nu_{e^*g^*}^{(0)} = 0$, $H_{eg} = [0, \nu'_{eg})$, and $H_{e^*g^*} = [0, \nu'_{e^*g^*})$, with the size measure of the set $H_{eg} \cup H_{e^*g^*}$ being equal to max $(\nu'_{eg}, \nu'_{e^*g^*})$. Thus in this case, $p1_i$ is equal to max $(\nu'_{eg}, \nu'_{e^*g^*})$, as given above. It should also be noted that the stratification and selection are based on TY(y-2) data and for estimation TY(y) data is used.

ii) Second-Phase Selection Probabilities

In the second phase of sampling, all businesses associated with the selected TE's in the first-phase sample are coded to SIC4. The SE's denoted by (i,j) are stratified by SIC4 within each size group.

Let the SE's in the first phase sample be denoted by (i,j) for $j \in J_i$ and $i \in s_1$, where J_i is the set of businesses owned by TE(i) as known from TY(y-2). The SE's in the first phase sample (or master sample) are stratified by Prov × Size × SIC4. The J_i's are defined at the time of designing (or updating) the master sample. It is planned that the master sample will be updated with respect to births, deaths and classification changes on a cyclical review basis as described later in section 3.

As before, let g be the size group within a province and let h denote the 4-digit SIC code associated with the businesses in the master sample where h=1, 2, ... L. In most cases the SIC4 code (h) will be nested within the SIC2 code (e), but it cannot always be true due to the following reasons:

(1) When a TE(i) is operating more than one business, there will be one dominant SIC2 code for the TE(i). But the associated SE's denoted by (i,j) for $j \in J_i$ will have SIC4 codes which do not necessarily belong to the same SIC2.

(2) SIC2 code is based on TY(y) information and SIC4 code is not always up-to-date. Thus, the SIC4 code may not be consistent with the SIC2 code due to a classification change.

At the second phase of sampling, the TE's are sampled via SE's, i.e. a TE will be sampled if at least one of the SE's belonging to the TE is selected. It should be noted that the second-phase tax sample consists of two types of records. First, those which have already been identified on the master sample and have been sampled at STC. Second, those which are births to the master sample and are selected at RCT for the first time. We first deal with the birth TE's in the master sample. The birth in the master sample should be distinguished from a birth in the universe. The reason for this distinction is that a universe birth will not always be added to the master sample. A master sample birth can occur in one of three ways: (1) a universe birth has been sampled because the hash number falls within the sampling interval, (2) the TE has changed stratum due to a change in size or SIC2, and has been sampled in the new stratum, and (3) due to changes in sampling requirements, i.e. the sampling rate was increased and the unit was sampled. In any case, if the TE is a birth in the master sample, it is automatically selected for the second phase sample.

For the other cases, the TE's are sampled via the SE's, i.e. TE will be selected if any one of the SE's belonging to the TE is selected. Let ν''_{gh} be the second phase sampling fraction for the stratum denoted by (g,h). Let $SE(i,j) \in (g,h)$, then define $R_{ij} = R_i$ for all $j \in J_i$, $i \in s_1$ where R_i could be as defined in the first-phase. Alternatively, it could be defined as $R_i = \text{hash}_2(i)$ where $0 \leq R_i < 1$, where the hash_2 function is independent of the hash_1 function. We opt for the

$hash_2$ function, as it affords the flexibility of master sample size increase and decrease. Furthermore it can be used to rotate the second-phase sample within the master sample. Also, as in the first phase, different hash functions may be used for the selection of T1 and T2 type tax entities at the second-phase.

The $SE(i,j)\varepsilon(g,h)$ is selected if $R_{ij} \leq \nu''_{gh}$ and the probability of selecting $SE(i,j)$ is given by

$$\pi_{ij} = \nu''_{gh} \text{ if } SE(i,j)\varepsilon(g,h).$$

The $TE(i)$ and the corresponding SE's denoted by (i,j) for $j\varepsilon J_i^*$ are selected if at least one of the SE's is selected. Therefore the probability of selecting $TE(i)$ at the second phase is given by $p2_i = 1 - \prod_{j\varepsilon J_i} (1-\pi_{ij})$.

Note that J_i is the statistical structure for $TE(i)$ at the time of designing or updating the master sample. Thus the second phase selection probability $p2_i$ is equal to 1 if $TE(i)$ is a birth to the master sample and is equal to $1-\prod_{j\varepsilon J_i} (1-\pi_{ij})$ for all other TE's. However, J_i^* is the statistical structure for $TE(i)$ at the time of sampling, and the probability of selecting the $SE(i,j)$ is given by $p2_{ij} = p2_i$ for $j\varepsilon J_i^*$, $i\varepsilon s_2$

where J_i^* is the set of BE's for $TE(i)$ as observed from $TY(y)$ and s_2 is the set of TE's selected in the second phase sample. It should be noted that for the purpose of computing selection probabilities at the second phase, the SE structure at the time of designing (or updating) the master sample is used. Therefore, it will be necessary to keep the master sample up to date by an SIC review process.

3. Implementation of the Sample Design

The implementation of this sample design requires a careful interplay between STC and RCT, with respect to the files sent by STC to RCT and the resulting data sent by RCT to STC. The files sent by STC control all the sampling of tax returns which takes place at RCT. These files contain sampling parameters as well as lists of inclusion and exclusion tax entities. Inclusion units are those which are sampled at STC for the second-phase sample and are to be transcribed. Exclusion units are the ones which are (i) out-of-scope for STC purposes, (ii) in IP for which data will be collected by direct survey, and (iii) the master sample units not included in the second-phase sample. Sampling parameters include strata definition along with their associated sampling intervals. In order to carry out this mandate, a sequence of steps must be closely adhered to. These are as follows.

STC sends RCT the first-phase sampling rates through parameter files. All births are sampled within their sampling strata using the first-phase sampling intervals. The physical selection of the first-phase sample is embedded into RCT's tax assessing system. The selection occurs using a hashing procedure which basically transforms the tax filer's number (SIN or T2) to a unique hash number between 0 and 1. The units whose hash number falls within the sampling interval are automatically selected. The selected tax returns are photocopied or microfilmed, and sent to STC where data capture of the required data items occurs. These records, along with their associated data are then added to the master sample residing at Statistics Canada. Note that two years between sample selection and the end of the data capture would have elapsed: that is, every annual cycle of tax sampling lasts for two years. This sampling is continuous and transcription occurs as the records are sampled at RCT. Note that the births that occurred between $TY(y-2)$ and $TY(y)$ are sampled only at the first-phase, and these are selected at the second-phase with certainty. They could have been sampled at the second-phase using appropriate sampling rates. However this is not done, since the first-phase tax records have to be examined for the purpose of SIC4 coding on the master sample and that the additional cost of transcribing the required data items is not excessive.

The master sample is also used as a frame for the selection of the second-phase units. This selection is carried out at Statistics Canada using hashing functions. The resulting sample is sent to RCT by means of an inclusion list (historical file). This list indicates to RCT which tax returns must be pulled out for STC transcription. RCT sends STC both the first-phase sample births for $TY(y)$ and the second-phase records pulled from the master sample referring to $TY(y-2)$. STC also sends RCT an exclusion list as described earlier. Furthermore, in addition to sending STC the above sampled tax returns, RCT sends a universe file which is coded to the SIC2 level, and has a very fine geographical classification as well as an exact measure of size (the total operating revenue). This universe file is later used to provide the population counts in the post-stratification for estimation purposes.

The quality of the stratification data on the master sample will deteriorate over time. Note that geography and size can be updated from the universe file on a regular basis. However, as noted earlier, SIC4 information is not available on a universal basis. Therefore, we must update this SIC4 information on the master sample on a regular basis. This will be done in one of two ways, one being through direct contact and the other being tax record examination. Direct contact is costly and increases the response burden. Hence, SIC4 codes will be updated using tax data on a cyclical basis. However when a change in SIC2 coding is noted, the record is automatically identified for SIC4 coding. The remaining units are coded on a cyclical basis so that each unit is updated at least once during a pre-specified period.

The second-phase sample will rotate through the master sample over time. The rotation scheme is implemented by incrementing the hash interval in a modular fashion within the master sample. Rotation permits the use of composite estimation, thereby realizing efficiency gains.

4. Weighting and Estimation

The weighting and estimation strategy must take into account partnerships. Partnerships arise quite frequently in the unincorporated universe, as a result of several tax filers pooling ressources in joint business ventures. A partnership can be as simple as having two or more tax filers owning a business or as complex as owning several businesses. Each tax filer belonging to the partnership reports for each business that he co-owns, indicating his share of profits (or losses) in the venture. The partnership factor must be taken into account when estimation occurs, because otherwise, overestimation would result.

In this section we discuss in detail the handling of partnerships, the weighting as well as the estimation.

4.1 Adjustment for Partnership Share

In the case of corporate (T2) tax filers, there is one-to-one correspondence between TE's and BE's. However, this is not always true for individual (T1) tax filers. An individual tax filer can own one or more unincorporated businesses and there could be several partners in a business.

Let I_j be the set of TE's which are partners in a business denoted by j,BE(j). First, we define

$$I = \bigcup_j I_j$$

where I is the set of all TE's in the universe (i=1,2, ... I). Let δ_{ij} be the partnership share of TE(i) for BE(j). Then

$$\sum_i \delta_{ij} = 1 \text{ for all } j.$$

Note that the δ_{ij}'s are only known for TE's in the sample. For the units in the first phase sample, the partnership factor is obtained at the time of designing (or updating) the master sample. For the units in the second phase sample, the factor is available on a current basis. However, there is no information available about other partners in the business from a particular TE in the sample. Therefore it is not possible to obtain the overall selection probabilities for all the business entities. But the data reported by the sampled TE is for the entire business activity. In order to estimate the number of businesses and other related statistics unbiasedly, a pseudo business corresponding to each SE(i,j) is created and the data values are adjusted by multiplying by the partnership factor. Thus the SE(i,j) in the master sample represents δ_{ij} businesses for $j \in J_i$ at the time of designing (or updating) the master sample where $\delta_{ij} \leq 1$. Similarly, SE(i,j) in the second phase sample represents, say δ^*_{ij} ($\delta^*_{ij} \leq 1$) businesses for $j \in J^*_i$ at the current time period.

Alternatively, the partnership factor could have been applied as a multiplicative factor to the design weight. This solution is not appropriate because the partnership share is a data item obtained from the sampled tax records and should be viewed as such. Hence the selection probabilities and the design weights should not depend on this factor.

4.2 Weighting

Suppose that an auxiliary variable x is available for all TE's in the universe. Denote by x_i, the x-value for TE(i). Let U={u} define a set of post-strata for first phase weighting. Then, we have

$$\hat{X}_u = \sum_{i \in s_1 \cap u} \frac{x_i}{p1_i}$$

and

$$X_u = \sum_{i \in u} x_i ,$$

where $\sum_{i \in s_1 \cap u}$ denotes the summation over the first phase sampled TE's from post-stratum u and $\sum_{i \in u}$ is the summation over all the TE's in the post-stratum u.

Then the first phase weight $W1_i$ for TE(i) is given by

$$W1_i = \frac{X_u}{\hat{X}_u} \cdot \frac{1}{p1_i} \text{ for } i \in s_1 \cap u.$$

For the one-phase design, the weight $W1_i$ is used for all SE's (i,j) for $j \in J^*_i$, $i \in s_1$ i.e.

$$W1_{ij} = W1_i \text{ for } j \in J^*_i,$$

where $i \in s_1 \cap u$.

For the two-phase design, we assume that another auxiliary variable z is available for all units in the first-phase sample.

We have used the symbol "^" for estimates based on design weights and will use the symbol "~" for the estimates based on post-stratified weights.

Let z_i be the z-variable value for TE(i), available for all $i \in s_1$ and V = {v} be the set of post-strata for second phase weighting. Then

$$\hat{\hat{Z}}_v = \sum_{i \in s_2 \cap v} \frac{W1_i z_i}{p2_i}$$

and

$$\tilde{Z}_v = \sum_{i \in s_1 \cap v} W1_i z_i$$

where $\sum_{i \in s_2 \cap v}$ is the summation over the second-phase sampled units in the post-stratum v and $\sum_{i \in s_1 \cap v}$ is the summation over the first-phase sampled units in the post-stratum v. Then the second-phase weight $W2_i$ for TE(i) is given by

$$W_{2i} = \frac{\tilde{Z}_v}{\hat{\hat{Z}}_v} \frac{1}{p2_i} \text{ for } TE(i) \in s_2 \cap v.$$

The overall weight W_i for TE(i) is the product of first and second phase weights, i.e.

$$W_i = W1_i W2_i \text{ for } TE(i) \in s_2 \cap u \cap v,$$

and the weight for SE(i,j) is given by

$$W_{ij} = W_i \text{ for } j \in J^*_i, \quad TE(i) \in s_2 \cap u \cap v.$$

4.3 Estimation

Let δ^*_{ij} be the partnership factor for $j \in J^*_i$, $i \in s_2$ and y_{ij} be the y-value reported by TE(i) for BE(j). Let $_dY$ be the sum of y-values for all businesses belonging to the domain of interest d. Then $_dY$ is estimated by

$$\tilde{\tilde{{}_dY}} = \sum_{i \in s_2} \sum_{j \in J^*_i} W_{ij} \delta^*_{ij} {}_dy_{ij}$$

$$= \sum_{i \in s_2} W_i \sum_{j \in J^*_i} \delta^*_{ij} {}_dy_{ij}$$

where

$${}_dy_{ij} = y_{ij} \text{ if } SE(i,j) \in d$$
$$= 0 \quad \text{otherwise.}$$

The above estimate can also be written as

$$\tilde{\tilde{d^Y}} = \sum_{i \in s_2} W_i \, d^{y_i} \quad \text{where} \quad d^{y_i} = \sum_{j \in J_i^*} \delta^*_{ij} \, d^{y_{ij}}.$$

The variance of $\tilde{\tilde{d^Y}}$ is approximately given by

$$\text{Var}(\tilde{\tilde{d^Y}}) \doteq \sum_u \sum_{i \in u} \frac{(1-p1_i)}{p1_i} \{d^{y_i} - d^{R_u} x_i\}^2$$

$$+ \sum_v \sum_{i \in v} \frac{(1-p_{2i})}{p1_i \; p2_i} \{d^{y_i} - d^{R_v} z_i\}^2$$

where

$$d^{R_u} = \frac{d^{Y_u}}{X_u} \quad \text{and} \quad d^{R_v} = \frac{d^{Y_v}}{Z_v}$$

d^{Y_u} is the y-value total for the portion of domain d belonging to the post-stratum u and d^{Y_v} is similarly defined for the post-stratum v. The above variance is estimated by

$$\text{var}(\tilde{d^Y}) = \sum_v \sum_u \left(\frac{\tilde{Z}_v}{\hat{Z}_v} \frac{X_u}{\hat{X}_u}\right)^2 \sum_{i \in s_2 \cap u \cap v} \frac{1}{p1_i \; p2_i} \frac{1-p1_i}{p1_i} (d^{y_i} - \tilde{d^{R_u}} x_i)^2$$

$$+ \sum_v \sum_u \left(\frac{\tilde{Z}_v}{\hat{Z}_v} \frac{X_u}{\hat{X}_u}\right)^2 \sum_{i \in s_2 \cap u \cap v} \frac{1-p2_i}{(p1_i \; p2_i)^2} (d^{y_i} - \tilde{d^{R_v}} x_i)^2$$

where $\tilde{\tilde{d^{R_u}}}$ and $\tilde{\tilde{d^{R_v}}}$ are respectively the estimates for d^{R_u} and d^{R_v}.

5. Conclusions

The objectives of the Business Survey Redesign Project will be met through the use of multipurpose sample design across all industrial sectors. This will facilitate data comparability, and achieve cost savings in terms of survey development and implementation.

The proposed two-phase sample design with Poisson sampling offers the flexibility of controlling the reliability of the estimates for any given SIC4. It is also flexible to accommodate new industrial as well as geographical demands (e.g. subprovincial data). Also, the rotation of the sample is simple to implement, thus enabling the control of sample overlap between successive survey ocasions.

References

Colledge, M. (1987), "The Business Survey Redesign Project: Implementation of a New Strategy at Statistics Canada, "Presented at the Bureau of the Census 3rd Annual Research Conference, Washington.

Colledge, M., and Lussier, R. (1985), "A strategy for the Provision of Frame Data and Use of Tax Data for Economic Surveys, "Proceedings of the Section on Survey Methods, American Statistical Association, Washington.

Colledge, M., and Lussier, R. (1987), "A Generalized Methodology for Economic Surveys," Proceedings of the Section on Survey Methods, American Statistical Association, San Francisco.

Foy, P. (1987), "A Two-Phase Sampling Design for Estimation on the Basis of Tax Data within the context of the Annual Surveys of Economic Production, "Internal Technical Report, Business Survey Methods Division, Statistics Canada, Ottawa.

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

Laniel, N., and Poirier, C. (1989), "Calendar Year Adjustment of Annual Business Survey Data, "Working Paper Series, Business Survey Methods Division, Statistics Canada, Ottawa.

Latouche, M., and Hidiroglou, M.A. (1987), "Sample Size Determination and Allocation for the Monthly WRTS," Internal Technical Report, Business Survey methods Division, Statistics Canada, Ottawa.

Roberge, D. (1988)," Annual Programme of Economic Production Statistics (APEPS) Target Population: Definition, Identification, and Extraction," Internal Technical Report, Business Survey Methods Division, Statistics Canada, Ottawa.

Statistics Canada (1986), "Business Statistics Survey Frame Model," Statistics Canada, Ottawa.