Survey Methodology, June 1991
Vol. 17, No. 1, pp. 79–90
Statistics Canada

Sample Rotation and Estimation in the Survey of Employment, Payroll and Hours

IOANA SCHIOPU-KRATINA and K.P. SRINATH[1]

ABSTRACT

The current Survey of Employment, Payroll and Hours, conducted by the Labour Division of Statistics Canada is a major monthly survey collecting data from a large sample of business establishments. This paper describes the methodology of the survey. The description of the stratification, sample size determination and allocation procedures is brief, whereas the description of the rotation procedure is more detailed because of its complexity. Some of the possible simplifications of the design are also highlighted.

KEY WORDS: Establishment; Response burden; Sampling frame.

1. INTRODUCTION

1.0 Objectives of the Survey

The Survey of Employment, Payroll and Hours (SEPH) is a monthly establishment based survey conducted by Statistics Canada.

The main objectives of SEPH are:

(i) to provide monthly estimates of the total number of paid employees, average weekly earnings, average hourly earnings, average weekly hours and other related variables at the industry division-province level.

(ii) to provide the above estimates for Canada at the three digit Standard Industrial Classification (SIC) level.

(iii) to provide standard errors of all the estimates produced.

It is also intended to produce estimates at the three digit SIC-province level annually.

The survey covers all industries with the exception of agriculture, fishing and trapping, private household services, religious organizations and military services. For a detailed description of the objectives and uses of SEPH, see Cottrel-Boyd *et al.* (1980).

This article describes the sample selection and rotation as well as the estimation procedure adopted for the survey. Chapter 2 presents the sample selection and rotation procedure in detail. Chapter 3 is devoted to the estimation procedure. Some of the details relating to Chapter 2 are given in the Appendix. The Appendix also presents a simplified estimator of the number of live units.

For a complete description of the SEPH methodology, see Schiopu-Kratina and Srinath (1986).

1.1 Preliminary Definitions

Some of the terms used in this article are defined here for convenience.

(i) Establishment - An establishment is the smallest unit that is a separate operating entity capable of reporting all elements of basic industrial statistics. The establishment is the statistical unit for SEPH. We will use the term unit for establishment.

[1] Ioana Schiopu-Kratina and K.P. Srinath, Business Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6.

(ii) Employment reporting unit (ERU) - For purposes of detailed geographical statistics, the establishment is often sub-divided into reporting units based mainly on location and sometimes on other considerations like payroll, *etc.*

(iii) Standard Industrial Classification (SIC) - 1970. Each establishment is assigned a Standard Industrial Classification (SIC) code according to the nature of its activity. These SIC codes are defined in the SIC Manual (Statistics Canada 1970). For the purpose of the survey, 16 industry divisions (groups) which are groupings of specific three digits SIC's have been created.

In SEPH, the total number of paid employees associated with a unit is the characteristic chosen as a measure of size of that unit.

There are four size groups in SEPH and their boundaries are defined as follows: 0-19 for size group 1, 20-49 for size group 2, 50-199 for size group 3 and 200 or more for size group 4.

(iv) A super-stratum is defined by an industry division, province and size group. With 16 industry divisions, 12 provinces and territories and 4 size groups, there are 768 super-strata.

(v) A stratum is defined as a three digit SIC, province and size group. It is the finest level of detail for which estimates are obtained.

(vi) The take-all portion of the population consists of units which are all included in the sample with certainty. It contains units in size group 4 and pre-specified units of the population. The take-some portion of the population consists of the remaining units which are subject to sampling as described in the following sections.

2. SAMPLE SELECTION AND ROTATION

2.0 Sample Size Determination and Allocation Procedure

In SEPH, the take-some sample size is determined at the industry group - province level based on a designed coefficient of variation of the estimate of the total number of employees for that industry group-province. The required sample size and the sampling fractions are calculated at the super-stratum level, using X-proportional allocation, where X represents the total number of employees. The sampling fractions are held constant from one month to the next. Details about the allocation procedure can be found in Schiopu-Kratina and Srinath (1986).

The actual selection is made at the stratum level. Due to the minimum sample requirements at that level, the number of sampled units is larger than the required sample size at the industry group - province level (see (2.2)).

2.1 Sample Selection

Let us now consider a specific stratum. Let N be the size of the take-some portion of the population and n the size of the take-some sample in this stratum.

Whereas the allocation of the sample to the super-strata is X-proportional, the allocation at the stratum level within each super-stratum is essentially proportional to the total number of units in the take-some portion of that stratum.

The sampling fraction in each stratum is given by the formula:

$$f = \max\left(f', \frac{1}{100}\right), \tag{2.1}$$

where f' is calculated at the corresponding super-stratum level. In order to reduce the instability of the estimates caused by small values of the sampling fraction, it was decided to set 1/100 as a minimum sampling fraction for all strata.

The detailed calculations of the sample size at stratum level are given in section 2.3 (see the derivation of the formula (2.8)). A systematic sample is drawn from each stratum.

2.2 The Rotation Scheme

The sample rotation (partial periodic replacement of the sample) in SEPH is designed primarily to reduce the response burden. From previous surveys, it appeared that the average response rate in strata in which there was no rotation was significantly lower than the average response rate in strata in which there was rotation. Also, the existence of a large portion of units common to the sample for two consecutive months improves the reliability of the estimates of month-to-month change relative to the estimate of change based on two independent monthly samples. Rotation of the sample in each stratum has to be done under certain constraints such as keeping the units out of the sample for a certain period of time after they rotate out of sample.

The monthly sample consists of 14 groups numbered from 0 to 13. Group 0 contains the take-all units of the stratum. Groups 1 to 13 are called rotation groups. The labels 1 to 12 on rotation groups indicate the month in which the units other than births rotated into the sample. For example, rotation group 1 contains mostly units which entered the sample in January and births, rotation group 2 contains mostly units which entered the sample in February and births, *etc.* Rotation group 13 contains units which have completed 12 months in the sample. These units are the oldest in terms of the time spent in the sample and are eligible to be rotated out. Each month, births are selected and allocated at random to the rotation groups.

At the time of the monthly selection and rotation, all units in the reference month are transferred to rotation group 13. In February, for example, all units in the rotation group 2 are transferred to rotation group 13. A replacement group is selected from "eligible for selection" units, and newly recorded units (births). The units of the replacement group (with the exception of 11/12 of all births) are then placed in rotation group 2, and they are not eligible to rotate out for at least 12 months. If sufficient units are available for a replacement group, the contents of group 13 are removed from the sample and are not eligible for reselection for 12 months. Otherwise, some units in group 13 are retained in the sample until such time that there are enough available units outside the sample to form a replacement group. This is done in order to maintain the minimum sample size or attain a sample size large enough to provide estimates with prespecified reliability. This way, in general at least 11/12 of the units stay in the take-some portion of the sample for two consecutive months.

The units that have left the sample are assigned to a waiting group which is divided into subgroups. A subgroup consists of units which were all removed from the sample in the same month. The waiting group contains 12 subgroups in every stratum. The time each unit has spent outside the sample is thus recorded to ensure that the units will not be reselected for at least 12 months. The units that have spent the required amount of time in the "not eligible for selection" group are transferred to the "eligible for selection" group and are thus assigned a positive probability of reselection.

To summarize, the entire take-some population at any given time consists of four groups of units. These are:

(i) units that are in the sample for that month;

(ii) units that are eligible for selection (E.F.S.);

(iii) units in the waiting group which have rotated out of the sample less than 12 months ago and which are not eligible for selection (N.E.F.S.);

(iv) births, *i.e.* units that have not been previously recorded on the frame.

The process of monthly selection and rotation involves an exchange of units among these groups. Some units leave group (i) for group (iii) and new units enter group (i) from group (ii), after some selected births from group (iv) have been transferred to group (ii). The remainder of the births are allocated to groups (ii) and (iii) after selection. This is done in order to insure that the sample is representative of the population in any given month.

2.3 Determination of the Sample Size and Weights

2.3.1 Monthly Updates

The sampling frame contains a large number of units which are inactive, out of business, out of scope *etc.* Apart from the burden of retaining an increasing number of inactive units on the frame, the estimators based on samples drawn from such a population are likely to have a high variance, due to the fact that the sample contains a high proportion of zero observations. Ideally, all such units should be eliminated from the sampling frame before the monthly sample is drawn. The frame is updated each month, after a monthly selection and rotation and prior to the next. For this reason, the indices we use to denote births and deaths on the frame are one unit higher than those used for the sample size in the sample selection preceeding the update. For example, after the initial sample selection, say $n(0)$ units are in the sample, of which $d(1)$ units are subsequently found to be dead units. Then $D(1)$ denotes the number of dead units in the out-of-sample portion of the population and $B(1)$ the number of units registered as births that month. In calculating the required sample size for the following month $n(1)$, one must take into account these updates (see (2.3)) as well as the size of the population at the time of the first sample selection $N(0)$.

2.3.2 Determination of Sample Size

The population of a given take-some stratum is a function of time and it will be denoted by $N(t)$ say, where t is a positive integer which increases by one unit from one month to the next. The required sample size is, for each month:

$$n'(t) = [fN(t) + 0.5]. \tag{2.2}$$

Here $[a]$ is the largest integer number which is not greater than a. The constant 0.5 is used for a better approximation in the rounding off procedure.

Suppose $d(t)$ units are eliminated from the sample (in-sample deaths) and $D(t)$ from the rest of the population of the stratum (out-of-sample deaths). Also let $B(t)$ new units be recorded during the same time interval (births).

As a result, the size of the population of the cell at the time of the t^{th} selection is:

$$N(t) = N(t-1) - d(t) - D(t) + B(t). \tag{2.3}$$

Since the updates are not exhaustive, undetected inactive (dead) units are expected to exist in the population.

Let $n_\ell(t)$ be the number of live units left in the previous month sample (after the updates) *i.e.*:

$$n_\ell(t) = n(t-1) - d(t). \tag{2.4}$$

We assume that there are no undetected dead units in the sample at this point. We can think of the population of a stratum as consisting of two domains: the domain of live units and the domain of dead units. The size of the dead domain is not known, but an estimate $\hat{U}_d(t)$ can be calculated based on the information given by the sample and the updates (see Appendix). Let $\hat{U}_d(t)$ be an estimate of the number of undetected dead units in the population at the time of the t^{th} monthly selection. Then:

$$N(t) = \hat{U}_\ell(t) + \hat{U}_d(t), \tag{2.5}$$

where $\hat{U}_\ell(t)$ is the estimate of the number of live units.

The probability of choosing a dead unit when selecting a unit at random from the out-of-sample units is:

$$\hat{P}_d(t) = \min\left\{\frac{\hat{U}_d(t)}{N(t) - n_\ell(t)}, 1\right\}. \tag{2.6}$$

The required number of live units in the sample is:

$$n_\ell'(t) = f\hat{U}_\ell(t). \tag{2.7}$$

The replacement sample size is calculated in such a way as to ensure that the expected number of live units in the sample after selection is $n_\ell'(t)$.

Now assume that at the time of the t^{th} sample selection and rotation, of the $n_\ell(t)$ live units in the sample, $n_o(t)$ units are eligible to rotate out.

Since there are $n_\ell(t) - n_o(t)$ live units left in the sample, $n_\ell'(t) - n_\ell(t) + n_o(t)$ more live units are required in the sample for the t^{th} month.

In order to represent the births in the sample adequately, $b(t) = fB(t)$ births should be selected at random and included in the sample.

Therefore:

$$\ell(t) = \max(n_\ell'(t) - n_\ell(t) + n_o(t) - b(t), 0),$$

live units should be selected from the eligible for selection group and added to the sample along with the selected births. Taking into account the existence of an unknown number of inactive units in the population and integerizing, it is required that:

$$n_i(t) = \min\left(\left|\frac{\ell(t)}{1 - \hat{P}_d(t)} + 0.5\right|, n''(t)\right),$$

more units rotate into the same sample, with $\hat{P}_d(t)$ given by (2.6) and $n''(t) = N(t) - n_o(t)$.

In calculating $n_i(t)$, we made the assumption that there are no inactive units among the births, so the expansion factor $[1 - \hat{P}_d(t)]^{-1}$ is applied only to the "older" units in the E.F.S. group.

The sample size $n(t)$ for the t^{th} month is:

$$n(t) = \max\{n_\ell(t) - n_o(t) + n_i(t) + b(t), m\}. \tag{2.8}$$

In (2.8), m represents the minimum required sample size for a stratum, which is presently set at 3. This additional requirement increases the sample size by 3,000 units in all strata, of which 1,800 are expected to be in the sample for a considerable length of time.

Of the $n_i(t)$ units which rotate in, $\hat{n}_d(t) = \hat{P}_d(t)n_i(t)$ are expected to be found inactive and $\tilde{n}_\ell(t) = n_i(t) - \hat{n}_d(t)$ active. Thus, of the $n(t)$ units in the sample after the t^{th} monthly selection and rotation, $\hat{n}_\ell(t) = n_\ell(t) - n_o(t) + \tilde{n}_\ell(t) + b(t)$ are expected to be alive and they represent the $\hat{U}_\ell(t)$ units of the live domain at the proper rate f (see (2.7) - (2.8)), when $n(t) > m$ in (2.8).

2.3.3 Determination of Weights

The weight $w(t)$ used for estimation for the t^{th} month is expressed in terms of the size of the population and sample at time t. However, the use of $N(t)/n(t)$ as weight for estimation could lead to an overestimation of the live units in the population. Indeed, $n(t)$ in formula (2.8) was chosen so that the expected number of live units in the sample equals the required sample size. The number of dead units in the sample drawn as described above may not represent the size of the dead domain at the proper rate. In (2.8), n_i is drawn from the general population and is thus expected to preserve the proportion between the dead and the live domain. No deaths are expected to be found among births and thus $b(t)$ properly represents the birth subgroup of the population. There are, however, $n_\ell(t) - n_o(t)$ units left in the sample from a previous selection, after rotation and the updates on the sample. The proportion of deaths among them is likely to be much smaller than the corresponding proportion in the general population, in spite of the fact that the updates are based on information from sources other than the survey. Then the value of $N(t)/n(t)$ should be adjusted for the underrepresentation of dead units in the sample. This gives, when $n(t) > m$,

$$\frac{N(t)}{n(t) + \hat{u}(t)} = \frac{1}{f}, \tag{2.9}$$

where $\hat{u}(t)$ will be determined subsequently (see (2.10)). The value of $\hat{u}(t)$ represents the "deaths" that have to be added to the sample to correctly represent the dead units in the population. Notice that when the first sample is drawn or if a redraw takes place, such an adjustment is not needed, *i.e.* $\hat{u}(0) = 0$.

In order to find a formula for $\hat{u}(t)$, we use (2.5) in the numerator of (2.9) and (2.8) in the denominator. By (2.7) - (2.8), we must also have:

$$\frac{\hat{U}_d'(t)}{\hat{n}_d(t) + \hat{u}(t)} = \frac{1}{f}.$$

With $\hat{n}_d(t) = \hat{P}_d(t)n_i(t)$, we obtain from above

$$\hat{U}_d'(t) = \frac{1}{f}[\hat{n}_d(t) + \hat{u}(t)] \quad \text{or} \quad \hat{u}(t) = f\hat{U}_d'(t) - \hat{n}_d(t). \tag{2.10}$$

The death adjustment is given by:

$$v(t) = \begin{cases} \hat{u}(t) & \text{if} \quad \hat{u}(t) \geq -\hat{n}_d(t) \\ 0 & \text{if} \quad \hat{u}(t) < -\hat{n}_d(t) \end{cases} \tag{2.11}$$

and the weight used in estimation is:

$$w(t) = \frac{N(t)}{n(t) + \hat{v}(t)}. \tag{2.12}$$

Note that the weight in (2.12) is defined using an estimate and so it is a random variable.

The use of the weight defined by (2.12) implies that the estimate of the number of live units in the population, defined by $\hat{U}_\ell(t) = w(t)\hat{n}_\ell(t)$ does not exceed $N(t)$, the size of the population at the time of the t^{th} sample selection.

Let us define:

$$\hat{U}_d(t) = w(t)\,[\hat{v}(t) + \hat{n}_d(t)].$$

By (2.10) - (2.11), it follows that $\hat{U}_d(t) \geq 0$ and its minimum value is 0 when $\hat{v}(t) - \hat{n}_d(t) = 0$. By (2.5), the maximum value of $\hat{U}_\ell(t)$ is then $N(t)$.

The restriction that the estimator of live units be truncated at $N(t)$ has implications on estimation which will be discussed in section 3.1. The estimator $\hat{U}_d(t)$ is calculated recursively (see the Appendix) using 2.10 - 2.11 and the fact that $\hat{v}(0) = \hat{u}(0) = 0$.

It has to be noted that the formula (2.11) is slightly different from the formula giving the death adjustment in SEPH. Firstly, for the sake of simplicity, we did not consider here the cases when the minimum sample size m has to be used. In such cases, the use of the sampling fraction f in (2.10) is not appropriate. In SEPH, the previous month weight is used in (2.10) in lieu of f in all instances. Secondly, the death adjustment in SEPH is always taken to be positive. Formula (2.11) shows that it could be negative, as long as it is larger than $-\hat{n}_d(t)$. The actual instances in which this happens, or, more generally, when $\hat{u} \leq 0$ are very rare.

The Appendix presents a formula for the estimator of the live units which does not require the use of the death adjustment.

2.4 Sampling of Births

As mentioned previously, every month new units are added to the frame. Since it is believed that these new units (births) may differ from the "old" units, a special birth strategy was designed, aimed at adequately representing the births in the sample.

Ideally, if B births are added during the current month and if f is the stratum sampling fraction, then $b = fB$ births should be selected in the sample during that month. The selected births are randomly assigned to the rotation groups described in section 2.2. This ensures the same probability of rotating out of the sample for births as for the "old" units, so that the age distribution of the in-sample units is the same as the out-of-sample units.

Using the notation of the previous section, let n_i be the number of units required to rotate in at the time of the monthly sample selection excluding births and N' the number of units in the E.F.S. group (group (ii) of section 2.2). The birth strategy consists of a two-phase selection procedure. This procedure involves the formation of a common pool of births and "older" units from which a sample is then drawn. This was thought necessary because usually, the birth group is too small for sampling births separately each time. There are two ways of forming the common pool depending on the sizes of the birth group and the E.F.S. group. If:

$$\frac{n_i}{N'} > f, \tag{2.13}$$

then b' births are preselected from the birth group and a common pool of size $(N' + b')$ is formed where:

$$b' = \frac{b\,N'}{n_i}. \tag{2.14}$$

Inequality (2.13) ensures that $b' \leq B$ which means that the birth group is large enough and the preselection can take place. From the common pool of $N' + b'$ units, $n_i + b$ units are selected next and added to the sample.

The choice of b' as given by (2.14) ensures that the expected number of births in the sample is the desired one, since the probability of selecting one birth from the pool is $b' / (b' + N')$ and therefore the expected number of births when $n_i + b$ units are selected without replacement from this pool is $(n_i + b)\, b' / (b' + N')$ which by (2.14) equals b. Similarly, it is easy to see that the expected number of "older" units is n_i.

In the complementary situation, when (2.13) does not hold, a common pool of size $(n' + B)$ units is formed where:

$$n' = n_i / f \leq N'. \tag{2.15}$$

Then n' "older" units are selected from the E.F.S. group. Let us note that in this situation b' as given by (2.14) is larger than B and so the first procedure cannot be applied.

We now calculate the expected number of "old" units in the sample. Since the probability of selecting one "old" unit is now $n' / (n' + B)$ and $n_i + b$ units are drawn from the pool of $n' + B$ units and placed in the sample, the expected number of "old" units is $(n_i + b)\, n' / (n' + B)$ which equals n_i. The expected number of births is b.

It has to be noticed that an underrepresentation of births may occur in some situations. For example, if in some stratum no units are required to rotate in, then for the month in question the births will not be represented in the sample taken from that stratum. However this situation usually arises when the population size is small and in the long run, the representation of births can be expected to average out correctly.

The b births actually selected are randomly assigned to the rotation groups 1-12 in the sample, resulting, on the average, in $b/12$ births assigned to each of these rotation groups. This ensures that the probability of a birth rotating out is the same as that of any other unit.

In order to keep the age distribution of the units in the groups (i) – (iii) (see section 2.1) constant, the non-selected births will be allocated to the E.F.S. and the N.E.F.S. group at random. That is, if N' is the number of units in the E.F.S. group and N'' is the number of units in the N.E.F.S. group, then $N' (B - b) / (N' + N'')$ unselected births will be assigned to the E.F.S. group and $N'' (B - b) / (N' + N'')$ to the N.E.F.S. group.

3. ESTIMATION

3.0 Introduction

In this chapter we describe the procedure for estimating the characteristic "the total number of paid employees".

As indicated in section 2.0, only the reliability of the estimates of the total employment at the industry group-province level of aggregation is prespecified. The estimates of characteristics other than the total employment have varying degrees of reliability. For example, estimates of average weekly earnings are expected to have higher reliability than the total employment.

The finest level of aggregation at which the estimates are published is the SIC-province level, but the basic "building blocks" for producing the estimates are the strata.

An outlier in SEPH is an observation in the take-some portion of the sample which is larger than a prespecified value.

The weight of each stratum is first calculated as in section 2.3, then it is adjusted for outliers in that stratum. Estimates of the total employment are computed for each stratum using the adjusted weights. There is no adjustment for nonresponse, as values for the nonrespondents are imputed. The estimate of the total employment for each stratum is obtained by adding the following totals:

(i) the total employment for take-all units

(ii) the total employment for outlier units

(iii) the sum of the weighted values of employment for take-some units, excluding outliers.

Since the weight assigned to each outlier is one, outliers are treated as take-all units for the purpose of estimation and are therefore not used in the variance estimation.

3.1 Estimation of the Total of a Characteristic

Let us consider a specific stratum for a given month of the survey. Let N be the size of the take-some portion of the population of the stratum and n the number of take-some units in the sample for that month. If $\hat{v}$ is the death adjustment (see (2.11)), then the original weight assigned to each unit in the sample for the purpose of estimation is (see (2.12)):

$$w = \frac{N}{n + \hat{v}}. \tag{3.1}$$

However, if t outliers are present in the sample with $t \geq 1$ then this weight is modified by giving each outlier a weight of 1 and assigning to the remaining units in the take-some portion of the sample the weight:

$$w' = \frac{N - t}{n + \hat{v} - t}. \tag{3.2}$$

Let S represent the set of in-sample units. If $Y(u)$ represents the value of employment corresponding to the unit u in the sample, then the estimate of the total employment in the stratum is:

$$\hat{Y} = \sum_{u \in S} w(u) Y(u). \tag{3.3}$$

Where $w(u) = 1$ if u is an outlier or a take-all unit and $w(u) = w'$ for all other units in the sample (see (3.2)).

Estimates of totals at any level of aggregation higher than the stratum are obtained adding the estimates of the stratum totals, for all strata in the level of aggregation considered.

Since for the purpose of estimation the outliers may be considered take-all units, we may replace, for the sake of simplicity w' by w, with w given by (3.1). Then N and n should be modified accordingly (see (3.2)).

Let N_ℓ be the size of the live domain in the population of the stratum and n_ℓ the number of live units in the sample. Let $\bar{Y}_\ell$ represent the average employment of the live units in the sample.

Since only the live units contribute to the total employment, an estimate of the cell total is:

$$\hat{Y}_\ell = \hat{U}_\ell' \bar{Y}_\ell. \tag{3.4}$$

We consider the case $fN > m$.

In (3.4), $\hat{U}_\ell'$ is an estimate of N_ℓ, the number of live units in the population and is given by:

$$\hat{U}_\ell' = \frac{1}{f} n_\ell. \tag{3.5}$$

In (3.5), f is the stratum sampling fraction which is held fixed. The estimator based on the values of $\hat{U}_\ell'$ is unbiased, that is:

$$E(\hat{U}_\ell') = N_\ell. \tag{3.6}$$

As a consequence of the unbiasedness of the estimator based on (3.5), $\hat{U}_\ell'$ may exceed N in some instances, as low possible outcomes of $\hat{U}_\ell'$ compensate for it.

In SEPH, the estimate of live units $\hat{U}_\ell$ is defined by:

$$\hat{U}_\ell = w n_\ell \tag{3.7}$$

with the weight given by (3.1).

The definition of the weight in SEPH (see the end of section 2.3) implies that:

$$\hat{U}_\ell = \begin{cases} \hat{U}_\ell' & \text{if } \mathrm{n}_\ell \leq fN \\ N & \text{if } n_\ell > fN. \end{cases} \tag{3.8}$$

The estimate of the total employment in SEPH is defined by:

$$\hat{Y}_\ell = \hat{U}_\ell \bar{Y}_\ell. \tag{3.9}$$

An estimate of the total employment based on (3.5) is:

$$\hat{Y}_\ell' = \hat{U}_\ell' \bar{Y}_\ell. \tag{3.10}$$

The estimator based on (3.8) is biased and consequently the estimator of the total employment in SEPH is also biased.

However, the mean square error of the estimator based on (3.9) conditioned on n_ℓ is smaller than the mean square error of the estimator based on (3.10), conditioned on n_ℓ. We now sketch a proof of this claim.

It is not difficult to see that, for each particular outcome, the bias B_ℓ of the estimator based on (3.9) and conditioned on n_ℓ is given by:

$$B_\ell = (\hat{U}_\ell - N_\ell)\bar{Y}_\ell. \tag{3.11}$$

Similarly, we obtain for the estimator based on (3.10):

$$B_\ell' = (\hat{U}_\ell' - N_\ell)\bar{Y}_\ell. \tag{3.12}$$

We show that the conditional mean square error of the estimator (3.9) is smaller than the conditional mean square error of the estimator (3.10). The same result then holds for the unconditional mean square errors. We condition on the realized sample size of live units n_ℓ. From (3.8) - (3.10), $\mathrm{Var}[\hat{U}_\ell' \bar{Y}_\ell \mid n_\ell] - \mathrm{Var}[\hat{U}_\ell \bar{Y}_\ell \mid n_\ell] = [n_\ell^2 f^{-2} - N^2] 1\{n_\ell > fN\} \mathrm{Var}[\bar{Y}_\ell \mid n_\ell]$. Notice that $n_\ell^2 f^{-2} - N^2 > 0$ on the set $\{n_\ell > fN\}$. We now compare $[B_\ell']^2$ and B_ℓ^2:

$$[B_\ell']^2 - B_\ell^2 = 1\{n_\ell f^{-1} > N\}\{[\hat{U}_\ell' - \mathrm{N}_\ell]^2 \bar{Y}_\ell^2 - (N - N_\ell)^2 \bar{Y}_\ell^2\}.$$

But $\hat{U}_\ell' - N_\ell = f^{-1} n_\ell - N_\ell > N - N_\ell$ if $n_\ell f^{-1} > N$.

Therefore $[B_\ell']^2 - B_\ell^2 \geq 0$. Since MSE $[\hat{U}_\ell' \bar{Y}_\ell \mid n_\ell] = \mathrm{Var}[\hat{U}_\ell' \bar{Y}_\ell \mid n_\ell] + [B_\ell']^2$ and MSE $[\hat{U}_\ell \bar{Y}_\ell \mid n_\ell] = \mathrm{Var}[\hat{U}_\ell \bar{Y}_\ell \mid n_\ell] + [B_\ell]^2$, the term-by-term comparison leads to the conclusion that:

$$\mathrm{MSE}[\hat{U}_\ell' \bar{Y}_\ell \mid n_\ell] \geq \mathrm{MSE}[\hat{U}_\ell \bar{Y}_\ell \mid n_\ell].$$

This important property motivates the choice of the estimator (3.9) over (3.10) for the total employment in SEPH.

APPENDIX

In this Appendix, we use the notation of section 2.3.2.

We first derive the formula for $\hat{U}_d(t)$, the size of the dead domain used in SEPH for the t^{th} selection.

Recall $\hat{U}_d(t) = w(t)[\hat{v}(t) + \hat{n}_d(t)]$. At the time of the t^{th} update, $d(t + 1)$ dead units are found in the sample. We can replace therefore $\hat{n}_d(t)$, the estimated number of dead units in the sample by $d(t + 1)$ in order to obtain an update of $\hat{U}_d(t)$, namely $\hat{N}_d(t + 1)$:

$$\hat{N}_d(t + 1) = w(t)[\hat{v}(t) + d(t + 1)]. \tag{1.1}$$

Formula (1.1) uses the death adjustment from the previous month. The initial value of $\hat{v}$ is $\hat{v}(0) = 0$ (see the remark after (2.9) and the definition of $\hat{u}$) and $w(0) = f^{-1}$. Now the estimate of the size of the dead domain for the $(t + 1)^{\text{th}}$ monthly selection is:

$$\hat{U}_d(t + 1) = \max(\hat{N}_d(t + 1) - D(t + 1) - d(t + 1), 0). \tag{1.2}$$

Notice that $\hat{U}_\ell(t + 1)$ can be calculated from (2.5) when $\hat{U}_d(t + 1)$ is known and vice versa. An alternative form for $\hat{U}_\ell(t + 1)$ is obtained recursively as follows. Let us assume that $\hat{U}_\ell(t)$ is known before the $t + 1^{\text{th}}$ selection of th sample (recall that $t = 0$ is used for the first, or original selection). Then $\hat{U}_d(t)$ is also known and can be used to calculate $\hat{P}_d(t)$, the probability of selecting a dead unit from the out-of-sample units (see formula 2.6). This probability is then used to calculate the required number of units which should rotate in as described in 2.3.2, as well as the expected number of live units in the sample at the time of the $(t + 1)^{\text{th}}$ selection, $\tilde{n}_\ell(t)$.

Then the weight used in estimation for the next selection is $\hat{U}_\ell(t) / \tilde{n}_\ell$. After selection, the $(t + 1)^{\text{th}}$ update takes place and the actual number of live units in the sample is found to be $n_\ell(t + 1)$. The estimate of the size of the live domain for the following selection can be calculated

$$\hat{U}_\ell(t + 1) = \min\left\{\frac{\hat{U}_\ell(t)}{\tilde{n}_\ell(t)} n_\ell(t + 1) + B(t + 1), N(t + 1)\right\}$$

and so forth. To initiate the process, note that the weight used in the first estimation is $w(0) = f^{-1}$ and after the first update,

$$\hat{U}_\ell(1) = \min\{w(0) \times n_\ell(1) + B(1), N(1)\}.$$

REFERENCES

COTTREL-BOYD, T.M., DUNN, M.R., HUNTER, G.E., and SRINATH, K.P. (1980). Development of the redesign of the Canadian establishment based employment surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 8-15.

SCHIOPU-KRATINA, I., and SRINATH, K.P. (1986). The methodology of the Survey of Employment, Payroll and Hours. Working Paper No. BSMD-86-010E, Statistics Canada.

STATISTICS CANADA (1970). *Standard Industrial Classification Manual.* Catalogue 12-501, Ottawa: Statistics Canada.