Survey Methodology, December 1997
Vol. 23, No. 2, pp. 109-117
Statistics Canada

Sampling and Maintenance of a Stratified Panel of Fixed Size

F. COTTON and C. HESSE[1]

ABSTRACT

Statistical agencies often constitute their business panels by Poisson sampling, or by stratified sampling of fixed size and uniform probabilities in each stratum. This sampling corresponds to algorithms which use permanent numbers following a uniform distribution. Since the characteristics of the units change over time, it is necessary to periodically conduct resamplings while endeavouring to conserve the maximum number of units. The solution by Poisson sampling is the simplest and provides the maximum theoretical coverage, but with the disadvantage of a random sample size. On the other hand, in the case of stratified sampling of fixed size, the changes in strata cause difficulties precisely because of these fixed size constraints. An initial difficulty is that the finer the stratification, the more the coverage is decreased. Indeed, this is likely to occur if births constitute separate strata. We show how this effect can be corrected by rendering the numbers equidistant before resampling. The disadvantage, a fairly minor one, is that in each stratum the sampling is no longer a simple random sampling, which makes the estimation of the variance less rigorous. Another difficulty is reconciling the resampling with an eventual rotation of the units in the sample. We present a type of algorithm which extends after resampling the rotation before resampling. It is based on transformations of the random numbers used for the sampling, so as to return to resampling without rotation. These transformations are particularly simple when they involve equidistant numbers, but can also be carried out with the numbers following a uniform distribution.

KEY WORDS: Panel; Stratified sampling of fixed size; Stratified simple random sampling; Maximum coverage; Sample rotation; Equidistant numbers.

1. INTRODUCTION

We consider the successive selection of samples intended to follow the change over time of sums of variables, more generally functions of sums, in a population. For example, this may be a population of businesses or establishments for which we wish to follow monthly sales trends. The ideal would be to be able to conserve a constant sample, but demographic movements make this impossible and it may not be desirable in light of the survey response burden.

The methods for selecting units presented in this article are subject to the following three constraints:

Firstly, it is necessary to regularly introduce births and to take deaths into account.

Secondly, sampling involves characteristics of units which change over time, such as the size or primary activity of businesses. These characteristics can be used to modulate the probabilities of inclusion. Notably, it is often prudent to increase these probabilities with the size of the units if we estimate sums of variables correlated with this size. In addition, these characteristics may eventually be used as stratification criteria. In this article, a stratum will mean a subset of the population within which the sampling is of fixed size, to the nearest rounded digit. However, the criteria used in the stratification of the first sampling, such as the primary activity of the unit, become "inexact" or become less and less correlated with the variables of interest such as size. This results in a progressive increase in the variance of the estimates. To remedy this, it is appropriate to carry out a resampling of the sample from time to time after updating the stratification and calculating new probabilities of inclusion. This must be done while endeavouring to conserve the maximum number of units. However, fatally, units will be excluded and others will be introduced, mainly because of changes in the probabilities of inclusion, although this would also happen because of the changes of strata, even if the probabilities of inclusion remained constant.

Thirdly, we would like to distribute our survey response burden over a larger number of units. We determined a maximum duration limit for inclusion in the panel. Beyond this limit, the unit is replaced by another unit chosen from those which have never been included, or which have been absent the longest. We call this change of the sample over time rotation. It is generally slow and regular. The various methods for performing this rotation are well known in statistical agencies. They consist mainly in attributing, at the beginning, a permanent random number to each unit of the population. The successive samples are defined by intervals over these numbers or by the ranks induced by these numbers.

We call the chronological sequence of samples resulting from these updating operations a "panel" and the set of updating operations "maintenance" of the panel.

The maintenance scheme presented in this article is analogous to that of Hidiroglou, Choudhry and Lavallée (1991). It corresponds to a frequency of updating of the

[1] F. Cotton, Institut National de la Statistique et des Études Économiques, Département de l'Informatique and C. Hesse, Institut National de la Statistique et des Études Économiques, Département "Système Statistique d'Entreprises", 18 boulevard Adolphe-Pinard, 75675, Paris Cedex 14.

stratification and probabilities which is significantly less than the survey frequency. This is generally the case for surveys with an infra-annual periodicity. The speed of demographic movements is not considered large enough to make it worthwhile to reselect the sample every time. The rotation is carried out without changing the probabilities of inclusion and the strata between two resamplings and it is regularly spread over time to conserve a certain continuity of the quality of the estimators of change over time. This also corresponds to a duration of inclusion of which the expected value is constant. In certain algorithms, we could determine a constant duration between two resamplings; otherwise we could set an upper limit. The speed of rotation represents a compromise between the efficiency of the estimators of change over time, which is greater the lower the rate of renewal, and the concern not to keep a unit in the panel for too long. Note that the quest for maximum coverage in the resampling remains meaningful with the rotation: we first remove the fraction to be renewed as if there were no resampling, then we seek the maximum coverage with the residual portion.

We will examine several methods of panel maintenance, with emphasis on maximizing sample coverage during resamplings. We will distinguish more particularly a process which assigns equidistant numbers to the units before each change of stratum.

The article is divided as follows:

After reviewing definitions and describing a few notations in section 2, we briefly indicate in section 3 how Poisson sampling makes it possible to carry out the previous maintenance scheme simply and perfectly. This sampling has the disadvantage of being of random size, but it serves as a reference for the stratified sampling of fixed size which we then consider.

In most instances, in these samplings, we determined probabilities of inclusion at the outset and used a rounded number to determine an entire sample size in each stratum. This problem, examined in section 4, is not negligible when the strata are small, which can occur for strata of births. In addition, rounding is used in the method which we propose to maximize the coverage after resampling.

Section 5 deals with the maximum coverage of samples of fixed size. First, we review two known methods: that of Kish and Scott (1971) and another based on the attribution to each unit of permanent independent numbers following the uniform distribution. The Kish and Scott method (1971) seems poorly suited to an intermediate rotation between resamplings. The other method, which reproduces simple random sampling in each stratum, does not have this disadvantage, but the coverage is less than with the Kish and Scott method (1971). Finally, we propose that the numbers be equidistant before resampling. We then obtain the same coverage as with the Kish and Scott method (1971), at least in the case of proportional distribution, while facilitating intermediate rotations. However, the coverage remains less than the maximum theoretical coverage which we obtain, for example, with Poisson sampling.

In sections 6 and 7, we present the intermediate phases of updating births and deaths and of rotation.

To conclude the topic of maintenance, we show in section 8 how resampling can take place between two phases of rotation. We present a type of algorithm which extends after resampling the rotation before resampling. It is based on transformations of the random numbers used in the sampling, so as to return to resampling without rotation. These transformations are particularly simple when they involve equidistant numbers, but can also be carried out with the uniform beginning numbers if we wish to continue with simple random sampling.

2. REMINDERS, DEFINITIONS AND NOTATIONS

Let there be a population, or finite set of units $i \in U = \{1, ..., N\}$ where N is the size of the population.

We consider only samples without replacement. A sample is then simply a subset s of U. We call sample size the number n of units which it contains.

A sampling or selection plan is a discrete probability $p(s)$ over the set of samples.

We can generalize to joint sampling of several samples. By limiting ourselves to two samples s_1, s_2, the joint sampling is the probability $p(s_1, s_2)$ over the set of pairs (s_1, s_2).

The first-order probability of inclusion of an individual i is defined by:

$$\pi_i = \sum_{s \ni i} p(s).$$

$E(.)$ being the expected value with respect to the sampling, this yields:

$$E(n) = \sum_{i \in U} \pi_i.$$

In the case of two samples with first-order probabilities of inclusion $\pi_{i,1}, \pi_{i,2}$, we can define the joint probability of inclusion:

$$\pi_{i,1,2} = \sum_{s_1 \ni i, s_2 \ni i} p(s_1, s_2).$$

This yields the constraint:

$$\pi_{i,1,2} \leq \min(\pi_{i,1}, \pi_{i,2}). \tag{2.1}$$

If $i \in s_1$, the probability of reselection in s_2 is $\pi_{i,1,2}/\pi_{i,1} \leq \min(1, \pi_{i,2}/\pi_{i,1})$.

In Poisson sampling, the selection of the units is independent and the sample size is random. Except in section 3, we will instead consider sampling where the size is fixed to the nearest rounded digit.

Simple random sampling (SRS) is sampling of fixed size where the samples are equiprobable. This yields $\pi_i = n/N$.

The population is partitioned into strata U_h, $h = 1, ..., H$ of sizes N_h. In this article, we will call a set of H independent samples of fixed size n_h in each stratum

"stratified sampling of fixed size" and we will limit ourselves to samplings with a uniform first-order probability of inclusion in each stratum. We will then use the notation $f_h = \pi_i$. We will call a stratified sampling of fixed size with simple random sampling in each stratum "stratified simple random sampling" (SSRS).

We will call the number of consecutive surveys where a unit is included in the panel "duration of inclusion of a unit." We will notate it D_i, or D_h in the particular case where it is the same for all units of a stratum h. When $\pi_i \geq 0.5$, this duration cannot be less than $\pi_i/(1 - \pi_i)$. For example, if $\pi_i = 0.7$, the duration of inclusion is at least 3. In practice, we will not rotate units whose π_i exceeds a certain threshold.

In addition, the previous variables are indexed by survey wave t. The population U_t of size N_t and the sample s_t of size n_t vary because of births and deaths, and the sample also varies as a result of the stipulated rotation. Moreover, we will consider samples at particular times $t = t_1$ of the first sampling and $t = t_2$ of the first resampling. For the sake of simplicity, they will be notated s_1, s_2 instead of s_{t_1}, s_{t_2}. The algorithms described for the pair (s_1, s_2) will be valid for the following resampling pairs.

3. SOLUTION BY POISSON SAMPLING

It is enlightening to examine how we can observe the panel maintenance scheme by Poisson sampling. This is the model which we will endeavour to approximate in order to choose a selection method.

We attribute to each unit i, at its birth, a number which is a random number ω_i selected according to the uniform distribution in [0,1). It is implicit in the formulae where these numbers appear that the results of the operations are *modulo* 1.

During the first sampling, at date $t = t_1$, we select the units such that ω_i belongs to the interval $[0, \pi_{i,1})$ where $\pi_{i,1}$ are the probabilities of inclusion given. In the absence of rotation, we keep this interval at the following dates until resampling. Births as well as deaths are distributed at random in this interval. The resampling, at date $t = t_2$ is carried out by selecting the units of the interval $[0, \pi_{i,2})$ where $\pi_{i,2}$ are new probabilities of inclusion. The joint probability of inclusion is equal to the length of the common interval, *i.e.*, $\min(\pi_{i,1}, \pi_{i,2})$ which is the maximum theoretically possible according to the formula (2.1). The expected value of the coverage is therefore itself maximal.

Let us now consider a rotation between the sampling and the resampling. We maintain the probability $\pi_{i,1}$ and we can determine a duration of inclusion $D_{i,1}$, which is variable depending on the units, but fixed until the resampling. This constraint is realized by defining the sample at date $t(t_1 < t < t_2)$ by the interval

$$\omega_i \in [(t - t_1)\pi_{i,1}/D_{i,1}, (t - t_1)\pi_{i,1}/D_{i,1} + \pi_{i,1}).$$

The rate of rotation is a random variable. Its expected value results from $D_{i,1}$. It is equal, for any subset V of the population, to $\sum_{i \in V}(\pi_{i,1}/D_{i,1})/\sum_{i \in V}\pi_{i,1}$.

At the first resampling at date $t = t_2$, we could define the sample by

$$\omega_i \in [(t_2 - t_1)\pi_{i,1}/D_{i,1}, (t_2 - t_1)\pi_{i,1}/D_{i,1} + \pi_{i,2}).$$

However, we encounter a difficulty for units such that

$$\pi_{i,2} < \pi_{i,1}\left(1 - \frac{1}{D_{i,1}}\right),$$

and if ω_i belongs to the interval

$$\left[(t_2 - t_1)\pi_{i,1}/D_{i,1} + \pi_{i,2}, (t_2 - t_1)\pi_{i,1}/D_{i,1} + \pi_{i,1}\left(1 - \frac{1}{D_{i,1}}\right)\right).$$

These units, which were previously in the sample, are excluded but will be reincluded in a future rotation. If we wish to avoid this, we must make the limit of the new interval coincide with that of the old interval, and the sample at date $t = t_2$ is finally defined by:

$$\omega_i \in [a_{i,1}, a_{i,1} + \pi_{i,2}),$$

where:

$$a_{i,1} = (t_2 - t_1)\pi_{i,1}/D_{i,1} + \max\left[0, \pi_{i,1}\left(1 - \frac{1}{D_{i,1}}\right) - \pi_{i,2}\right].$$

The joint probability of inclusion is equal to the length of the common interval, *i.e.*,

$$\min\left(\pi_{i1}\left(1 - \frac{1}{D_{i,1}}\right), \pi_{i,2}\right).$$

This is also the maximum compatible with the rotation.

If we continue the rotation with durations of inclusion $D_{i,2}$ the interval at date $t > t_2$ is:

$$\left[a_{i,1} + (t - t_2)\pi_{i,2}/D_{i,2}, a_{i,1} + (t - t_2)\pi_{i,2}/D_{i,2} + \pi_{i,2}\right).$$

Poisson sampling controls exactly the duration of inclusion and maximizes, as an expected value, the coverage during resampling but with the disadvantage of a random sample size, regardless of the subpopulation. In the following pages, we will endeavour to devise algorithms similar to those just described for Poisson sampling in order to apply them to stratified sampling of fixed size. We will try to control the duration of inclusion in the rotation, as for Poisson sampling, and to approximate the same rate of coverage during resampling. We will begin with the problem of coverage during resampling in section 5, but first, it is useful to clarify certain concepts concerning the rounding of sample sizes by stratum.

4. ROUNDING OF SAMPLE SIZES BY STRATUM

This problem is related to the estimation formulae. These formulae use the first-order probabilities of inclusion, either in the unbiased Horvitz-Thompson estimator or in adjusted estimators. Let f_h be the probability of inclusion by stratum, and let $v_h = N_h f_h$. We must have a whole number n_h per stratum. An initial method for accomplishing this consists in restricting the choice of the f_h in such a way that v_h is an integer. In each stratum where we would have had $v_h < 1$, we must take $v_h = 1$ so that $f_h > 0$. However, if the stratification is very fine vis-à-vis the sample size, this occurs in numerous strata. This makes it necessary either to increase the sample size or to decrease the sampling rate in the other strata, to the detriment of efficiency.

We will use a second method, which consists in linking the probability f_h more loosely to n_h. We apply a rounding process such that $E(n_h) = v_h$, where v_h is no longer necessarily an integer.

Let us assume that $I(.)$ is the integer part function. We must have

$$\Pr[n_h = I(v_h) + 1] = \varphi_h,$$

$$\Pr[n_h = I(v_h)] = 1 - \varphi_h,$$

where $\varphi_h = v_h - I(v_h)$.

It is then no longer necessary that $n_h > 0$ in order for $f_h > 0$. Note that the first method can be considered a particular case of the second. This rounding can be done independently by stratum, in a linked way by systematic rounding or by the Cox method (1987). We describe only systematic rounding.

Let us first order all of the strata, and index them by their rank. Let $c_0 = 0$ and $c_h = \sum_{j=1}^{h} \varphi_j$; we select a number θ in the interval [0, 1), according to the uniform distribution and we take $n_h = I(v_h) + 1$ in the strata such that $c_{h-1} \le m - 1 + \theta < c_h$ for m entirely.

This implies that

$$|\,(n_{j_1} + \ldots + n_{j_2}) - (v_{j_1} + \ldots + v_{j_2})\,| < 1,$$

for any j_1, j_2 such as $1 \le j_1 \le j_2 \le H$.

In particular, the global size differs by less than one unit from its expected value. This is obviously not the case with independent roundings.

5. ALGORITHMS FOR THE MAXIMUM COVERAGE OF SAMPLES OF FIXED SIZE

The maintenance algorithms which we propose are based on the attribution of equidistant numbers. This is not necessary during the first sampling, nor in the rotation, but is used to maximize the coverage during updates of the stratification. That is why we examine this maintenance phase first.

Let us begin by describing all the notations and making a few useful observations.

We select a first sample s_1 stratified according to criterion h_1. After a certain time has elapsed, we select a new sample s_2 with an updated stratification h_2. The first-order probabilities of inclusion are respectively f_{h_1}, f_{h_2} and the sample sizes required by stratum are respectively n_{h_1}, n_{h_2}. It is sufficient to consider what happens in any stratum $h_2 = g$. Let $s_{g,1}$ be the part of the first sample s_1 in this new stratum, of which the size $n_{g,1}$ is generally random. Let $s_{g,2}$ be the part of the second sample s_2 in this new stratum, of which the size is fixed to the nearest rounded digit. The size $n_{g,1,2}$ of the coverage cannot exceed the limit $\overset{+}{n}_{g,1,2} = \min(n_{g,1}, n_{g,2})$. We can hope to devise $s_{g,2}$ a resampling process with a uniform first-order probability of inclusion in $s_{g,1}$ which makes it possible to attain this limit, at least when the first-order probabilities of inclusion in are also equal to a single value $f_{h_1} = f_1$. Note that, even if this limit is attained, the fixed size constraints decrease the coverage. The finer the stratification, the greater this effect. In fact, the smaller the population of stratum g, the greater the likelihood that the coefficient of variation of $n_{g,1}$ will be large, as well as the proportion of units not reselected in the case $n_{g,1} > n_{g,2}$.

There is an obvious way of attaining the limit $\overset{+}{n}_{g,1,2}$. Let us assume first of all that the first-order probabilities of inclusion in $s_{g,1}$ are uniform. If $n_{g,1} < n_{g,2}$, we add $n_{g,2} - n_{g,1}$ units to $s_{g,1}$ selected at random in the complement of $s_{g,1}$. If $n_{g,1} > n_{g,2}$, we remove $n_{g,2} - n_{g,1}$ units from $s_{g,1}$ selected at random. By construction this yields $s_{g,2} \subseteq s_{g,1}$ or $s_{g,2} \supseteq s_{g,1}$, and $n_{g,1,2} = \overset{+}{n}_{g,1,2}$. If the first-order probabilities of inclusion in $s_{1,g}$ are not uniform, we apply the same method within subsets where these probabilities are uniform. This is the method proposed by Kish and Scott (1971) on page 468 of their article. They do not stipulate the procedure for random selection, but we assume that it is SRS.

As Kish and Scott point out, the second-order probabilities of inclusion are not uniform and if the first sampling is a SSRS, the second sampling no longer meets this definition. The first-order probability of inclusion, itself, is not strictly uniform when includes elements of strata from the previous sampling: see an example in the appendix. However, there is another method which verifies this condition. It is well known to statistical agencies which practise coordination of samples. For the sake of convenience, we will call it "method 1".

Method 1: Use of independent numbers following the uniform distribution

We attribute to the units, at their birth, ω_i numbers which follow the uniform distribution in [0, 1) and are independent, as in Poisson sampling. The first sample s_1 is obtained by selecting, for example, the n_{h_1} units of lower rank according to ω_i in each stratum. With this algorithm, the maximum coverage is also obtained by selecting the n_{h_2}

units of lower rank according to ω_i in each stratum h_2. Moreover, it is obvious that these two samplings are SSRS.

It is also obvious that we cannot obtain greater coverage with this algorithm. In addition, we conjecture that it is not possible to do better, for SSRS, regardless of the algorithm.

On the other hand, the coverage is poorer as an expected value than with the Kish and Scott method (1971), at least in the particular case where the first-order probabilities of inclusion in s_1 are uniform. In fact, at that point the relations $g\, s_{g,2} \subseteq s_{g,1}$ or $s_{g,2} \supseteq s_{g,1}, n_{g,1,2} = n^{+}_{g,1,2}$, are not necessarily true and the loss of coverage is greater, the smaller the strata during the first sampling.

We shall demonstrate this, again in the particular case of a uniform probability of inclusion f_1 in s_1. Let us assume that ω_{h_1} is the greatest value of ω_i for the units of s_1 in stratum h_1, and ω_g the greatest value of ω_i for the units of s_2 in stratum g. Let $\omega_1^- = \min(\omega_{h_1})$ and $\omega_1^+ = \max(\omega_{h_1})$. If $\omega_g \le \omega_1^-$ then $s_{g,2} \subseteq s_{g,1}$ and if $\omega_g \ge \omega_1^+$, then $s_{g,2} \supseteq s_{g,1}$. In both cases $n_{g,1,2} = n_{g,1,2}$. The risk of not attaining the limit exists only if $\omega_1^- \le \omega_g \le \omega_1^+$. In this case, the relation $s_{g,2} \subseteq s_{g,1}$ or $s_{g,2} \supseteq s_{g,1}$ is no longer necessarily true: see Figure 1, where we considered only 2 strata h_1. The loss of coverage is greater where the quantity $\omega_1^+ - \omega_1$ is greater as an expected value, and therefore where the strata h_1 are smaller.

Method 2: Use of equidistant numbers

If we accept not to conserve a SSRS, how can we modify the previous method to obtain the same coverage as the Kish and Scott method (1971), at least when we have the uniform probability of inclusion f_1 in s_1 ? We have seen that the loss of coverage was the result of the deviation between the ω_{h_1}. It is sufficient to transform the ω_i into new numbers $\rho_{i,1}$ in such a way that the ρ_{h_1} which correspond to the ω_{h_1} are as close as possible to a common value, *i.e.*, f_1. More specifically, we would like to have the equivalence:

$$\{ i \in s_1 \Leftrightarrow R_{h_1}(i) \in [1, ..., n_{h_1}]\} \Leftrightarrow \rho_{i,1} \in [0, f_{h_1}),$$

where $R_{h_1}(i)$ is the rank according to ω_i in h_1 of unit i. A solution is given by the transformation:

$$\rho_{i,1} = \frac{R_{h_1}(i) - 1 + \theta_{h_1}}{N_{h_1}} \tag{5.1}$$

where θ_{h_1} is a real number which verifies:

$$\begin{cases} \theta_{h_1} \in [0, \varphi_{h_1}), n_{h_1} = I(v_{h_1}) + 1, \\ \theta_{h_1} \in [\varphi_{h_1}, 1), n_{h_1} = I(v_{h_1}). \end{cases}$$

The transformation therefore involves the rounded number of the v_{h_1} examined in section 4. The sampling of s_2 is carried out like that of s_1 except that the $\rho_{i,1}$ now play the role of the ω_i: in each new stratum g we define rounded sizes $n_{g,2}$ and we select the $n_{g,2}$ units of lower rank according to $\rho_{i,1}$. Note that these ranks are different from those induced by ω_i.

Let us assume that the probability of inclusion in s_1 is still uniform. Let ρ_g be the value of $\rho_{i,1}$ for the unit of rank $n_{g,2}$ in g. If $\rho_g \in [0, f_1)$, then $s_{g,2} \subseteq s_{g,1}$. Otherwise $s_{g,2} \supseteq s_{g,1}$. In this particular case, we therefore attain the maximum coverage $n^{+}_{g,1,2}$ as in the Kish and Scott method (1971), and unlike method 1. We illustrate in Figures 1 and 2 how the transformation into equidistant numbers makes it possible to increase the coverage compared to method 1.

We apply the same algorithm when the probabilities of inclusion in s_1 are not uniform. Unlike the Kish and Scott method (1971), we do not need to fix the size of the new sample within subsets where these probabilities are uniform. This is another advantage and we think that it increases the coverage.

Nonetheless, the coverage obtained by this algorithm remains lower, as an expected value, than that of a Poisson sampling with the same probabilities of inclusion. In order to have, as an expected value, the same coverage as with Poisson sampling, it would be sufficient to define $s_{g,2}$ by $\rho_{i,1} \in [0, f_g)$. In fact, we would then have $\Pr(i \in s_i \cap s_2) = \min(f_{h_1}, f_g)$, but the sampling so obtained would no longer be of fixed size.

The following resamplings, after new updates, are carried out by repeating the process. For example, before selecting s_3 we calculate equidistant numbers $\rho_{i,2}$ based on $\rho_{i,1}$ (and not ω_i) in each stratum h_2.

The resulting sampling plan in the new strata is no longer a SRS. In particular, the probabilities of inclusion of the pairs of units vary generally as a function of the former strata. In other words, the resampling keeps a "trace" of the stratification of the first sampling. Moreover, the probabilities of inclusion of the units in $s_{g,2}$ are not exactly equivalent to f_g, except for the sample defined by $\rho_{i,1} \in [0, f_g)$. For the sample of fixed size $n_{g,2}$ this probability varies as a function of the size of the former strata. As in the Kish and Scott method (1971), we do not strictly control these probabilities. However, the deviation between f_g and the true probability becomes negligible when $n_{g,2}$ is sufficiently large.

Note 1. The transformation of numbers which independently follow the uniform distribution in equidistant numbers was proposed by Brewer, Early and Hanif (1984) as a way of rotating samples in the same manner as Poisson sampling, with the advantage of a smaller variance of the sample size. However, this transformation is performed by taking the set of the population, and therefore they did not address the problem of maximum coverage during changes of stratum. The numbers change only when births and deaths are updated, according to a procedure which is also quite different from that which we propose for changes of stratum.

Note 2. In the demonstration we just provided, it is not necessary that the numbers be completely equidistant. It is sufficient that the n_{h_1} units of s_1 and the $N_{h_1} - n_{h_1}$ complementary units have their new numbers respectively in $[0, f_{h_1})$, $[f_{h_1}, 1)$. We could attribute these new numbers

in such a way that they independently follow the uniform distribution in these intervals.

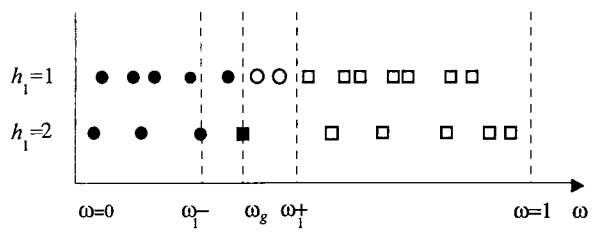

Figure 1. Coverage with method 1 (numbers following the uniform distribution).

We have represented the units in g according to the value of the number ω (on the abscissa) and the stratum h_1 of the first sampling (on the ordinate). We assume that there are only two strata. The circles correspond to $s_{g,1}$ and the squares to the complementary part. The solids correspond to $s_{g,2}$ and the blanks to the complementary part. The size of $s_{g,2}$ was fixed at 9 which defines ω_g. In this example, we see that two units are not reselected (in $h_1 = 1$) and that another is new (in $h_1 = 2$). The size of the coverage is 8, while the Kish and Scott method would make it possible to reselect the 9 units in $s_{g,1}$.

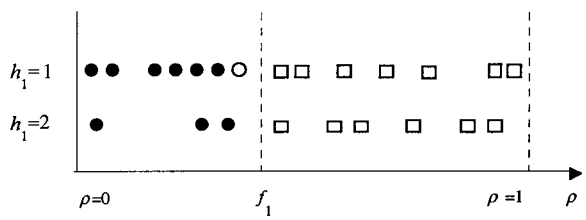

Figure 2. Coverage with method 2 (equidistant numbers).

We are in the same situation as in Figure (1), but this time the equidistant numbers ρ serve as the abscissa of the units. This equidistance is defined in each of the whole strata h_1 and the gaps we see in the sequence of numbers correspond to the units which are not in g. The first sample $s_{g,1}$ is composed of the units for which this number is less than the probability of inclusion f_1, regardless of the stratum. The second sample $s_{g,2}$ is composed of the 9 units with the smallest ρ and the coverage is 9, as with the Kish and Scott method (1971).

6. UPDATING BIRTHS AND DEATHS WITHIN STRATA

In this section and the following one, we consider the stratification (h) without reference to the period. The updating of births and deaths within strata is essentially a particular case of change of the strata of units. It is exactly as if the births entered the strata and the deaths left. We can therefore apply the previous methods. Let us take a look, in particular, at method 2.

In a stratum, the population $U_{h,t}$ of size $N_{h,t}$ varies with each updating carried out at time t. We will notate the births as $B_{h,t+1}$ and the deaths $D_{h,t+1}$ between t and as $t + 1$, this yields $U_{h,t+1} = U_{h,t} + B_{h,t+1} - D_{h,t+1}$.

We consider the simple case where the probabilities of inclusion f_h remain uniform in $U_{h,t}$ and constant. The size $n_{h,t}$ of the sample $s_{h,t}$ is a rounded number to the integer of $N_{h,t} f_h$. The numbers $\rho_{i,t}$ change with each updating. Just before updating $s_{h,t}$, leading to $s_{h,t+1}$:

a) we make equidistant the numbers $\rho_{i,t-1}$ in $U_{h,t}$.

b) we attribute equidistant numbers to the units of $B_{h,t+1}$.

Let $\rho_{i,t}$ be the number so obtained. An initial solution would consist in selecting the $n_{h,t+1}$ units of $U_{h,t+1}$ with the smallest $\rho_{i,t}$. Note that these are no longer equidistant because we removed the deaths situated at random.

However, units with numbers close to f_h can leave the sample and then return on a future occasion. We remedy this by a rightward shift of the selection interval. Let ρ_{h,d_t} be the number of the beginning unit of the selection interval for $s_{h,t}$ and ρ_{h,e_t} that of the unit immediately following the end unit of this interval in $U_{h,t}$ In other words, the sample $s_{h,t}$ consists of the interval closed to the left and open to the right $[\rho_{h,d_t}, \rho_{h,e_t})$. Between t and $t + 1$, the number of units of $U_{h,t+1}$ belonging to this interval becomes $m_{h,t+1}$. If $n_{h,t+1} \geq m_{h,t+1}$, the beginning of the interval for $s_{h,t+1}$ is fixed to the unit of number ρ_{h,d_t}, otherwise we shift the interval in such a way that its end is the unit of number ρ_{h,e_t}. We therefore have a slight involuntary rotation.

7. ROTATION BETWEEN TWO RESAMPLINGS

7.1 Rotation Without Updating of Births and Deaths

We can then stipulate a time of inclusion D_h whole and constant in the stratum. We have two variants, depending on whether we keep the same rounded number or vary it.

7.1.1 Fixed Rounded Number

We therefore have a size n_h strictly fixed during the rotation. We divide n_h into D_h whole numbers $n_{h,l}$, ($l = 1, ..., D_h$) such that $|n_{h,l} - n_h/D_h| < 1$. Let q_h be the quotient and r_h the remainder of the division of $t - t_1$ by D_h and let $n_{h,0} = 0$. The sample at time t includes the units ranging from rank $1 + q_h n_h + \sum_{l=0}^{r_h} n_{h,l}$ to rank $(q_h + 1)n_h + \sum_{l=0}^{r_h} n_{h,l}$.

If $D_h = D$, we can stipulate in addition

$$|\sum_{h=1}^{H} n_{h,l} - \frac{n}{D}| < 1, l = 1, ..., D_h.$$

The variance of the rate of rotation is then practically nil.

However, the duration of inclusion is not controlled when $v_h < 1$: this yields $n_h = 0$ or $n_h = 1$. In the first case, there is no rotation, and in the second case, on the contrary, the time of exclusion can be considered too short. The following method makes it possible to obtain a rotation which corresponds to v_h.

7.1.2 Variable Rounded Number

The sample $s_{h,t}$ is defined based on the numbers rendered equidistant:

$$i \in s_{h,t} \Leftrightarrow \rho_{i,1} \in \left[f_h \frac{t - t_1}{D_h}, f_h \frac{t - t_1}{D_h} + f_h \right).$$

The sample size varies between $I(v_h)$ and $I(v_h) + 1$ in the stratum, and it is independent of the sizes in the other strata. This shows us what the result would be of the sample rotation advocated by Brewer *et al.* (1984) in the case of stratified sampling of fixed sized and uniform probability in each stratum.

7.2 Rotation With Updating of Births and Deaths

To simplify, we assume that each new survey wave is accompanied by the introduction of the births since the previous wave and a rotation. The method bifurcates into two procedures depending on whether or not we wish to respect exactly the durations of inclusion D_h between two resamplings.

7.2.1 Procedure A

The births are isolated in separate strata, and we wait for the resampling before subtracting the deaths. In this case each wave of births is dealt with exactly like an initial sampling after attributing the numbers ω_i. The sampling is carried out by stratifying with the same nomenclature (h), or with another more scattered or more confined. To simplify the notations, but without loss of generality, we assume that this is the same nomenclature. The index of stratification can then be written (b, h), where b crossed with h indicates the wave of births with a particular modality $b = 1$ corresponding to the units already existing during the first sampling or a previous resampling. This brings us back to the case of section 7.1 in each stratum (b, h) and the duration of inclusion is respected exactly.

The number of strata, and therefore of rounded numbers, is multiplied by the number of waves of births. The sample size can become fairly random with independent roundings (but less so than with Poisson sampling). It may therefore be worthwhile to link, at least partially, the rounded numbers. For example, we carry out a systematic rounding in the dimension h for each b or the reverse. We then keep these roundings and this is the 7.1.1 method which then applies rather than the 7.1.2 method.

7.2.2 Procedure B

In procedure B, we subtract the deaths at each survey wave. This is the type of updating presented in section 6. We would prefer a fixed duration of inclusion, but that is made difficult by the random number of deaths. At most, we can try to control a maximum duration of inclusion DM_h. We may also wish to prevent the units which have just left the sample from returning on a future occasion, which can occur if the rotation is slow. The idea is to get back to the algorithm described in section 6 by removing first of all from $s_{h,t}$ the units of which the previous duration of inclusion in $s_{h,t}$ attained DM_h. They are found the farthest to the left of the interval $[\rho_{h,d_t}, \rho_{h,e_t})$ and are mixed with the births too recent to have attained DM_h. However, these must still be removed in order for the distribution of the sample according to the generations to be correct. For that, it is sufficient to attribute to the births a fictitious previous duration of inclusion which falls between 1 and DM_h, just after defining the sample. For example, after defining $s_{h,t}$, we assign to each unit of $B_{h,t}$ belonging to the sample the same previous duration of inclusion in the sample as that of the unit of $U_{h,t-1}$ situated immediately to the left. Then let R_{h,d_t} be the highest rank among the ranks according to $\rho_{i,t}$ of the units of the interval associated with $s_{h,t}$ which have been included DM_h times in the sample; we discard the first units of $s_{h,t}$ up to and including rank R_{h,d_t}. Finally, this brings us back to the algorithm described in section 6 with, for ρ_{h,d_t}, the number of the unit of rank $R_{h,t} + 1$, ρ_{h,e_t} remaining that of the unit which follows the unit of last rank in $s_{h,t}$.

8. RESAMPLING AFTER ROTATION

We now reselect the indices of strata h_1, h_2. We define the stratification $\mathbf{h}_1$ as a function of the procedure used for the updates of the births. With procedure A, we place the births in separate strata, this is the stratification defined by crossing the waves of births b with the nomenclature h_1. With procedure B, $\mathbf{h}_1$ is identical to h_1 However, we keep the notations of the independent quantities of b as f_{h_1}, D_{h_1}.

The selection of the new sample s_2, in a new stratification h_2 must be carried out at period $t = t_2$.

We begin by removing from the previous sample (at period $t = t_2 - 1$) the units which have attained the maximum authorized duration of inclusion. There remains a sample s'_1 of size $n'_{\mathbf{h}_1}$ of which we would like to conserve the maximum number of units in the resampling.

In the case without rotation examined in section 5, it was easy to define the resampling because the sample s_1 was composed of the units of lower rank according to ω_i in each stratum after a real number independent of the ω_i. In this instance, this number is 0. The resampling took place in the same manner by selecting the units of lower rank according to $\rho_{i,1}$, after this number, in the new strata.

After rotation this no longer works: there is no longer any real independent of the numbers such that the sample s'_1 is composed of units of lower rank after it. This is true even in the case where $f_{h_1} = f_1$. The problem is obviously aggravated with f_{h_1} varying by stratum. The idea which then comes to mind is to first carry out a transformation of the numbers in such a way that those from s'_1 find themselves at the beginning of [0, 1). This will then bring us back to the case without rotation. This is the same kind of idea which is presented by Hidiroglou, Choudhry and Lavallée (1991).

This transformation is fairly immediate in the particular case where the updates are done with procedure A and with the variable rounded number from section 7.1.2. Without

resampling, the selection interval at time t_2 would have been:

$$\rho_{i,1} \in \left[(t_2 - t_1) f_{h_1}/D_{h_1}, (t_2 - t_1) f_{h_1}/D_{h_1} + f_{h_1}\right).$$

The resampling results in new strata with probabilities f_{h_2}. These include the creations of units between the dates $t_2 - 1$ and t_2, to which we attribute equidistant numbers $\rho_{i,1}$, in each stratum h_2, independently of the survivors. They still contain units whose death has occurred since the previous sampling. It is possible to define a new sample s_2 in the same way as for Poisson sampling, by the interval, *i.e.*,

$$\rho_{i,1} \in \left[a_{h_1}, a_{h_1} + f_{h_2}\right),$$

where:

$$a_{h_1} = (t_2 - t_1) f_{h_1}/D_{h_1} + \max\left[0, f_{h_1}\left(1 - \frac{1}{D_{h_1}}\right) - f_{h_2}\right].$$

Let us recall that we shift from the supplementary quantity

$$f_{h_1}(1 - \frac{1}{D_{h_1}}) - f_{h_2}, \quad \text{if} \quad f_{h_1}(1 - \frac{1}{D_{h_1}}) - f_{h_2} > 0,$$

to prevent the units which have just left the sample from returning too quickly.

As for Poisson sampling, the probability of a survivor being in the old and the new sample is then the maximum possible, namely:

$$\min\left(f_{h_1}\left(1 - \frac{1}{D_{h_1}}\right), f_{h_2}\right).$$

However the size n'_{h_2} of this sample is random, whereas we want a sample of fixed size n_{h_2}. We obtain it by selecting, in each new stratum h_2, after having removed the deaths, the n_{h_2} units of lower rank according to $\eta_{i,1} = \rho_{i,1} - a_{h_1}$. This number therefore plays, for the resampling, the same role that ω_i played during the first sampling.

If, on the other hand, we chose procedure A with a fixed rounded number in the rotation or if we chose procedure B, we must begin again with the rank of the units of $\mathbf{h}_1$ during the last updating. This is the rank according to ω_i with procedure A or the rank according to $\rho_{t_2} - 1$ with procedure B. Let us assume that $N_{\mathbf{h}_1}$ is the size of the population at date $t_2 - 1$. Let $R_{\mathbf{h}_1,d}$ be the rank of the unit preceding the one of lower rank in s'_1 and $R_{\mathbf{h}_1}(i)$ the rank of unit i. The number used to classify the units in the new strata becomes:

$$\eta_{i,1} = \frac{R_{\mathbf{h}_1}(i) - 1 - a_{\mathbf{h}_1} + \delta_{\mathbf{h}_1}}{N_{\mathbf{h}_1}} \text{ modulo } 1,$$

where:

$$a_{\mathbf{h}_1} = R_{\mathbf{h}_1,d} + \max\left(0, n'_{\mathbf{h}_1}/N_{\mathbf{h}_1} - f_{h_2}\right).$$

With procedure A we can keep $\delta_{\mathbf{h}_1} = \theta_{\mathbf{h}_1}$ while we make a choice of $\delta_{\mathbf{h}_1}$ consistent with the last rounded number if procedure B is applied. However, because of the rotation, this choice has a minor impact on the coverage and it would be almost as well to select at random in [0, 1).

9. CONCLUSION

Algorithms based on equidistant numbers do not produce SRS. The first-order probabilities of inclusion are not exactly controlled and the second-order probabilities are unknown. During the changes of stratum, there remains a "trace" of the former strata in the new strata. The application of the SRS formulae to estimate the variance leads to biased results, generally in the direction of over-estimation. However, we think that the improvement in coverage during resamplings provided by the algorithms based on equidistant numbers outweighs the disadvantage of biased estimation of the variance and of the confidence intervals. According to section 5, the finer the stratification the greater this advantage. In particular, the use of equidistant numbers seems to be quite indicated with procedure A where the strata (b, h) are likely to be very small for the waves of births $(b > 1)$. The advantage of equidistant numbers is not as great with procedure B. However, making the numbers of births equidistant renders both the number of survivors reselected at each updating of the sample and the duration of inclusion less random.

However, let's take a quick look at what would change in the maintenance if we wanted to conserve SSRS. At each stage we must conserve the independent and uniform distribution of the ω_i. First of all, the phases of updating the births and of rotation between resamplings described in sections 6 and 7 apply while still conserving the same ω_i and the procedure is even simpler. The most delicate part is the resampling after the intermediate phase of rotation. The objective is to obtain not only a SSRS but also, if possible, the same coverage as for method 1 in section 5.

Let us assume that $\alpha_{\mathbf{h}_1}(j)$ is the number ω of the unit of rank j in a former stratum $\mathbf{h}_1$.

Let us assume first of all that, in a former stratum, all the units are such that $f_{h_2} \geq n'_{\mathbf{h}_1}/N_{\mathbf{h}_1}$. In particular, this occurs in all the strata for a sampling with a single rate in the sampled part, if we do not lower this rate. We then endeavour to find a transformation such that the numbers of the units of the sample are at the beginning of [0, 1). The simplest is the permutation:

$$\begin{cases} \beta_{\mathbf{h}_1}(j) = \alpha_{\mathbf{h}_1}(j + N_{\mathbf{h}_1} - R_{\mathbf{h}_1,d}), & j \leq R_{\mathbf{h}_1,d}, \\ \beta_{\mathbf{h}_1}(j) = \alpha_{\mathbf{h}_1}(j - R_{\mathbf{h}_1,d}), & j > R_{\mathbf{h}_1,d}. \end{cases}$$

However, a less costly transformation is:

$$\begin{cases} \beta_{h_1}(j) = \alpha_{h_1}(j) + \alpha_{h_1}(N_{h_1}) - \alpha_{h_1}(R_{h_1,d}), & j \le R_{h_1,d}, \\ \beta_{h_1}(j) = \alpha_{h_1}(j) - \alpha_{h_1}(R_{h_1,d}), & j > R_{h_1,d}. \end{cases}$$

It is sufficient to find the result of $\alpha_{h_1}(R_{h_1,d})$ and $\alpha_{h_1}(N_{h_1})$, after which a simple sequential calculation makes it possible to deduct β from α.

The Jacobian of the transformation is equal to 1 and consequently the numbers conserve their uniform distribution. Moreover, the joint distribution $p(s_1, s_2)$ is the same as if there had been no rotation. The demonstration is provided in Cotton and Hesse (1992, page 55). We therefore have the maximum coverage of SSRS.

If this yields units with $f_{h_2} < n'_{h_1}/N_{h_1}$ in the stratum and we apply the transformation, the units whose rank falls approximately between $N_{h_1} f_{h_2}$ and n'_{h_1} are not reselected during the resampling but will be reintroduced during a future rotation. It is therefore preferable to use, for these units, a transformation which is situated just before f_{h_2} the new numbers. We must proceed by subsets according to the value of f_{h_2}. However, that tends to decrease the coverage.

ACKNOWLEDGEMENTS

The starting point of our work is an internal document from the Business Survey Methods Division of Statistics Canada: Hidiroglou, M.A., Srinath, K.P. (1990), Methods of integrated sampling for sub-annual business surveys.

We would like to thank a co-writer and an anonymous referee for their assistance in the drafting of this article.

Some of the methods proposed have been applied to the INSEE, but the opinions expressed are solely those of the authors.

APPENDIX
Probabilities of Inclusion in the Kish and Scott Method (1971)

Let us consider an example where the first-order probability of inclusion is not strictly controlled.

The population is divided into three parts A, B and C of equal size N. The first sampling is a SRS of $2a$ units in $A + B$ and a SRS of a units in C. During the second sampling, we wish to select a units in A and $2a$ units in $B + C$, while retaining the maximum number of units from the first sample and with uniform probability of inclusion a/N. The Kish and Scott method consists in adding or removing by SRS the appropriate number of units separately in A and in $B + C$. In A, the second marginal sampling is a SRS and the probability of inclusion is quite uniform. We will show that this is not the case in $B + C$. Let n_1 and n_2 be the sizes of the two successive samples in B. By symmetry, the probability of inclusion during the second sampling is uniform in B. It is equal to:

$$\begin{aligned} E(n_2)/N &= [E(n_1) + E(n_2 - n_1)]/N \\ &= a/N + E(n_2 - n_1)/N. \end{aligned}$$

If $n_1 = a, n_2 - n_1 = 0$; otherwise the expected value of $n_2 - n_1$ conditional on n_1 differs depending on the sign of $a - n_1$:

If $a - n_1 > 0, E[(n_2 - n_1)] \mid n_1] = (a - n_1)(N - n_1)/(2N - n_1 - a)$.

If $a - n_1 < 0, E[(n_2 - n_1) \mid n_1] = (a - n_1)n_1/(n_1 + a)$.

Note $p(n_1)$ the probability that the first sample will have the size n_1 in B. This yields:

$$E(n_2 - n_1) = \sum_{n_1} p(n_1) E[(n_2 - n_1) \mid n_1].$$

Since the sizes of A and B are equal, $p(n_1) = p(2a - n_1)$, therefore:

$$\begin{aligned} &E(n_2 - n_1) \\ &= \sum_{n_1 < a} p(n_1)\{E[(n_2 - n_1) \mid n_1] + E[(n_2 - n_1) \mid (2a - n_1)]\} \\ &= \sum_{n_1 < a} p(n_1)(a - n_1)[(N - n_1)/(2N - n_1 - a) - (2a - n_1)/(3a - n_1)] \\ &= (2a - N) \sum_{n_1 < a} p(n_1)(a - n_1)^2/[(2N - n_1 - a)(3a - n_1)] \\ &= (2a - N)K, K > 0. \end{aligned}$$

Except in the case $2a - N = 0$, $E(n_2 - n_1)$ is not nil and $E(n_2)/N$ is different from a/N. The probability of inclusion is therefore not uniform in $B + C$.

REFERENCES

BREWER, K.R.W., EARLY, L.J., and HANIF, M. (1984). Poisson, modified Poisson and collocated sampling. *Journal of Statistical Planning and Inference*, 10, 15-30.

COTTON, F., and HESSE C. (1992). Tirages coordonnés d'échantillons. INSEE working paper E9206.

COX, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.

HIDIROGLOU, M.A., CHOUDHRY, G.H., and LAVALLÉE, P. (1991). A sampling and estimation methodology for sub-annual business surveys. *Survey Methodology*, 17, 195-210.

KISH, L., and SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.