Coordinating the PRN: Combining Sequential and Bernoulli-Type Sampling Schemes in Business Surveys

Ronit Nirel, Aryeh Reiter*, Tzahi Makovky* and Moshe Kelner*
***Central Bureau of Statistics, Israel**
66 Kanfey Nesharim St., Jerusalem 95464, Israel
nirelr@cc.huji.ac.il; areiter@cbs.gov.il; tzahim@cbs.gov.il; moshek@cbs.gov.il.

Abstract

In this paper we use permanent random numbers (PRNs) for the maintenance of a single sample over a long period of time. The sample is stratified by industry and size levels and units are selected with equal probabilities within strata. Two types of updates are carried out: (a) annual updates reflecting all changes in the frame during the preceding year and (b) intra-annual updates accounting for newborns during the year. A sequential PRN scheme is used for the annual sample and a Bernoulli scheme for the intra-annual samples. This design causes two problems. First, because of the differences between the sequential and Bernoulli schemes, the upper bound of PRNs of newborns may differ from that of the old units, thus resulting in different selection probabilities for the two types of units in the next annual sampling. Second, since generally the number of newborns in each update is small, the cumulative sample size over updates may have large deviations from its expected value. Here we propose adjusted PRNs for units included in the intra-annual frames that have a uniform distribution in the unit interval, and a maximal value for the sampled units that is approximately equal to that of the annual sample. We also suggest a sampling scheme that controls the overall sample size of the intra-annual updates, while keeping the expected weights unchanged over time. Finally, we illustrate the proposed method for data from the Israeli Manufacturing Indices Survey.

Keywords: Beta distribution; Collocated sampling; Coordination of samples; Order statistics.

1. Introduction

National Statistical Agencies (NSAs) publish time series of economic indicators based on periodic surveys of business establishments. While efficient estimation of periodic change requires maximal overlap between successive samples, samples should also be continuously updated to account for rapid changes in the business population (e.g. births, changes in size). Other common problems include the need to coordinate between several samples and between units within a single sample, to reduce response burden. Sampling schemes based on Permanent Random Numbers (PRNs) provide simple and flexible solutions to such problems. Ohlsson (1995) reviews the main schemes based on PRNs and describes their application in a number of NSAs. The basic idea is that each unit in the frame is assigned a random number in the interval (0,1) that is permanently associated with that unit. Within strata, units are ordered by their PRNs and the sampling is based on the ordered list. For example, in a "sequential" scheme for selecting a simple random sample without replacement (*srswor*) of n units from a population of size N the first n units in the ordered list are selected. In a Bernoulli scheme, with the same sampling rate, the sample comprises all units with a PRN not larger than n/N.

Here we use PRNs for the maintenance of a single sample over a long period of time (e.g. a decade). The sample is stratified by industry and size levels and units are selected with equal probabilities within strata. Two types of updates are carried out: (a) an annual update reflecting all changes in the frame during the preceding year (e.g. deaths, splits, and changes in industry) and (b) intra-annual updates accounting for newborns during the year. The design of the annual update sample is guided by a requirement for a maximal overlap of successive annual samples. The design of the intra-annual samples is guided by the need to keep the selection probabilities within strata constant throughout the year. To attain the first requirement a sequential PRN scheme is applied each year to an updated frame. Uniformity of the probabilities over the year is achieved by using a Bernoulli scheme for the intra-annual updates. This design causes two problems. First, because of the differences between the sequential and Bernoulli schemes, the upper bound of PRNs of newborns may differ from that of the old units (persistants) resulting in different selection probabilities for the two types of units in the next annual sampling. Second, since generally the number of newborns in each update is small, the cumulative sample size over updates may have a large deviation from its expected value.

In this paper we address these problems analytically. In Section 2 we propose a method for combining two frames from which a sequential and a Bernoulli samples were selected, respectively, with the same probabilities. The combined frame is then used to select a new sequential sample that has a sizeable overlap with the two former samples. This problem is related to the issue of "birth bias" discussed by Ernst, Valliant and Casady (2000) and by Ohlsson (1995, p. 166). Both papers are concerned with bias in the selection probability of newborns under sample rotation when the newborns and persistants are sampled together. For specific situations these papers suggest solutions such as a correction factor for the selection probability of newborns, and an adjusted sampling scheme. In our problem, persistants and newborns are sampled separately and the two frames are then combined for a later sampling occasion. Our method is based on an adjustment of the newborns' random numbers. The adjusted numbers have a uniform distribution in the unit interval, and the maximal value for the sampled units is approximately equal to that of the sequential sample.

Section 3 deals with the problem of sample size variability over several intra-annual updates. The PRN literature considers the question of choosing between fixed sample sizes and fixed inclusion probabilities, particularly in the context of Poisson sampling. Brewer, Early and Hanif (1984) suggest collocation of the random numbers so that they are more evenly distributed in the unit interval and thus reduce, but not eliminate, the variation in sample size. Ohlsson (1995) suggests a sequential Poisson scheme that ensures a fixed sample size but results in approximate selection probabilities. The approximation is improved by Saavedra (1995) but no solution is suggested for the calculation of the true selection probabilities. The problem of controlling the cumulative sample size of a sequence of Bernoulli samples imposes an additional challenge. We propose a constructive method that calculates the desired sample size for each update, and ensures that the cumulative sample size does not deviate from its expected value by more than one unit. Furthermore, the selection probabilities are kept at their expected values. The method is based on collocation of the random numbers and on randomization of the conditional inclusion probabilities. In Section 4 we illustrate the suggested method for the Israeli Manufacturing Indices Survey. Finally, Section 5 contains some concluding remarks.

2. A Single Intra-Annual Update

We deal here with annual business samples that are stratified by industry and size levels. Generally there is a take-all stratum and several (e.g. one to five) take-some strata for each industry. To achieve statistical efficiency and reduce fieldwork burden the same base sample is used for several years. This sample is updated periodically (e.g. once in several months) for new establishments (newborns) and once a year for other changes in the frame, including merges and splits as well as changes in industry and size level. Formally, for each stratum h we have a base frame list L_{h0} comprising N_{h0} units. An annual sample S_{h0} of n_{h0} units is selected with equal probabilities ($n_{h0} = N_{h0}$ for the take-all strata). Units are selected using a sequential PRN scheme; that is, each unit i in stratum h is assigned independently a permanent random number x_{hi} from a uniform distribution on (0,1). A random starting point r_{h0} is selected, and we assume here without loss of generality that $r_{h0} = 0$. Hence, units with the n_{h0} smallest random numbers are included in the sample. Let $x_{(n_{h0})}$ be the largest random number in the sample. To abbreviate we drop hereinafter the index h.

We begin with the simple case where the sample is updated for newborns once during a given year. A frame update L_1 comprising N_1 newborn units is obtained, and a PRN is assigned to each unit in L_1. To simplify the estimation procedure it is required that the update selection probabilities are retained at the annual sample value $\pi_0 = n_0 / N_0$. A Bernoulli scheme may be used to select the intra-annual sample update S_1; that is, all units with $x_i \le \pi_0$ are included in the sample. The update sample size n_1 is thus random, with expected value $N_1\pi_0$ and variance $N_1\pi_0(1-\pi_0)$. Collocation of the random numbers may reduce the variability of n_1. For example, we can sort the units in L_1 in a random order and assign to unit k a random number from a uniform distribution on $((k-1)/N_1, k/N_1)$. As before, all units with PRN not greater than π_0 are selected, but the sample size variability is reduced greatly because the random numbers are more equally spaced. In fact, n_1 is equal to $\lfloor N_1\pi_0 \rfloor$ or $\lfloor N_1\pi_0 \rfloor + 1$, with respective probabilities $1 - \langle N_1\pi_0 \rangle$ and $\langle N_1\pi_0 \rangle$, where $\lfloor a \rfloor$ and $\langle a \rangle$ are the integer and fractional part of *a,* respectively. We refer to this modified scheme as a Bernoulli-Type (B-T) scheme.

In preparation for the selection of the next annual sample, a new frame L_2 is constructed combining living units in L_0 and L_1, and units that were born in-between L_1 and L_2. Generally, a coordination of some degree is desired between subsequent annual samples, ranging from a complete overlap to no overlap. A sequential *srswor* PRN scheme is applied to the new

frame, and the starting point r_2 controls the degree of overlap. In our case, a maximal overlap is desired and we thus assume $r_2 = 0$. However, units in S_1 have generally different likelihood of being included in the new sample as compared to units in S_0, since generally $x_{(n_0)} \neq \pi_0$. The reason is that the order statistic $X_{(k)}$ of a series of N uniform (0,1) random variables has a Beta$(k, N-k+1)$ distribution (e.g. Stone, 1996, p. 178). In particular, $X_{(n_0)}$ has a Beta$(n_0, N_0 - n_0 + 1)$ distribution with expectation $p = n_0/(N_0+1)$ and variance $p(1-p)/(N_0+2)$. Thus, the scales of random numbers in L_0 and L_1 have to be coordinated before sampling from L_2.

We construct now a transformation of the random numbers of units in L_1 so that the transformed numbers have two properties: (a) they have a uniform distribution on the unit interval and (b) the maximal value of the transformed numbers of units in S_1 is approximately equal to $x_{(n_0)}$. Let X_i be the PRN assigned to unit $i \in L_1$ and define a rescaled random variable X_i^* by

$$X_i^* = X_i^*(x_{(n_0)}) = \begin{cases} \dfrac{x_{(n_0)}}{\pi_0} X_i & \text{if } \; i \in S_1 \\ (1 - x_{(n_0)}) \dfrac{X_i - \pi_0}{1 - \pi_0} + x_{(n_0)} & \text{if } \; i \notin S_1. \end{cases} \tag{1}$$

For $i \in S_1$, X_i^* maps X_i from the interval $(0, \pi_0)$ to $(0, x_{(n_0)})$, and for $i \notin S_1$ from $(\pi_0, 1)$ to $(x_{(n_0)}, 1)$. The conditional distribution of $X_i^* \mid X_{(n_0)} = x$ is a mixture of $U(0,x)$ and $U(x,1)$ distributions with respective probabilities π_0 and $1-\pi_0$. The unconditional distribution is given by

$$f_{X_i^*}(t) = \frac{n_0}{n_0 - 1} \Pr\{\text{Beta}(n_0 - 1, N_0 - n_0 + 1) > t\} + \Pr\{\text{Beta}(n_0, N_0 - n_0) \leq t\}, \;\; 0 \leq t \leq 1 \tag{2}$$

(For more details see the Appendix). Analysis of the first and second derivatives of $f_{X_i^*}(t)$ indicates a minimum at $t = \pi_0$ and two turning points $0 < t_1 < \pi_0 < t_2 < 1$, between which the function is convex, and otherwise it is concave. Thus, X_i^* is not uniformly distributed on (0,1). Define the $U(0,1)$ random variable $Y_i = F_{X_i^*}(X_i^*)$, where $F(\cdot)$ is the cumulative distribution function (CDF). The values of Y_i can be computed from

$$y_i(t) = F_{X_i^*}(\text{t}) = \frac{1}{N_0} \left[\frac{n_0}{n_0 - 1} \sum_{i=0}^{n_0 - 2} \Pr\{\text{Beta}(i+1, N_0 - i) \leq t\} + \sum_{i=n_0}^{N_0 - 1} \Pr\{\text{Beta}(i+1, N_0 - i) \leq t\} \right], \;\; 0 \leq t \leq 1 \tag{3}$$

(See the Appendix for details). The transformed variable Y_i has a uniform (0,1) distribution by definition and therefore holds property (a).

To verify property (b) we study empirically the relative deviation of $y(p)$ from p, where p is the expected value $X_{(n_0)}$. Table 1 shows the relative deviates $\text{RD} = (y(p) - p)/p$ and the relative standard deviation of $X_{(n_0)}$, $\text{RSTD} = \sqrt{\text{Var}(X_{(n_0)})}/p$, for selected values of n_0 and N_0. It is seen that the values of RD are considerably smaller than the respective values of RSTD. For the least favorable row in the Table, $n_0 = 2$, the ratio RSTD/RD is equal to 4.4, 4.9, and 7 for $N_0 = 20, 50$ and 100, respectively. We conclude that the differences between $x_{(n_0)}$ and π_0 may be substantial and that as, on average, the maximal value of X^* for units is S_1 is p, and as the deviation of the CDF of X^* from p is small, the adjusted PRNs satisfy property (b).

3. Several Intra-Annual Updates

In this section we deal with the more realistic case of several intra-annual updates of the frame (e.g. quarterly). We assume that for each update the number of new business units in a given stratum, if any, is small. As before, it is required that the selection probability in each stratum is kept at its annual sample value, and that the sample size variation for each update and

Table 1. Expected values of $x_{(n_0)}(p)$, values of $y(p)$, relative deviates of $y(p)$ from p (RD), and relative standard deviation of $X_{(n_0)}$ (RSTD) for selected values of N_0 and n_0.

	N_0 =20				N_0 =50				N_0 =100			
n_0	p	$y(p)$	RD	RSTD	p	$y(p)$	RD	RSTD	p	$y(p)$	RD	RSTD
2	0.095	0.109	0.15	0.66	0.039	0.045	0.14	0.69	0.020	0.023	0.14	0.98
5	0.238	0.256	0.08	0.38	0.098	0.105	0.07	0.42	0.050	0.053	0.07	0.61
10	0.476	0.497	0.04	0.22	0.196	0.204	0.04	0.28	0.099	0.103	0.04	0.42
15	0.714	0.738	0.03	0.13	0.294	0.302	0.03	0.21	0.149	0.153	0.03	0.33
20					0.392	0.401	0.02	0.17	0.198	0.202	0.02	0.28
25					0.490	0.499	0.02	0.14	0.248	0.252	0.02	0.24
50									0.495	0.500	0.01	0.14

over updates is minimal. We begin by explaining how to determine the size of the sample for each update and then show how to sample from the updates using PRNs. Finally we describe the adjusted PRNs for the annual update.

3.1 Sample Size Determination

For any take-some stratum let N_j be number of sampling units in the j-*th* update, $j = 1,\ldots,J$, n_j the actual sample size in update j and π_0 the desired inclusion probability. As before, denote by $\lfloor a \rfloor$ the integral part of a, that is $\lfloor a \rfloor \le a < \lfloor a \rfloor + 1$, and by $\langle a \rangle$ the fractional part of a, so that $a = \lfloor a \rfloor + \langle a \rangle$. The expected sample size for the first update (j=1) is $E_1 = N_1\pi_0$. Thus, the actual sample size n_1 is equal to $\lfloor E_1 \rfloor$ or $\lfloor E_1 \rfloor + 1$ with respective probabilities $1 - \langle E_1 \rangle$ and $\langle E_1 \rangle$. For j>1, let $m_j = n_1 + \ldots + n_j$, $E_j = E(n_j) = N_j\pi_0$, and $F_j = E_1 + \ldots + E_j$ be the actual and expected cumulative sample sizes after sampling from the j-th update, respectively. Assume that the cumulative sample size m_{j-1} for update j–1 has the desired expected value with minimal variation, i.e.

$$P\left(m_{j-1} = \lfloor F_{j-1} \rfloor\right) = 1 - \langle F_{j-1} \rangle \text{ and } P\left(m_{j-1} = \lfloor F_{j-1} \rfloor + 1\right) = \langle F_{j-1} \rangle. \quad (4)$$

Clearly $E\left(m_{j-1}\right) = F_{j-1}$. We show now how to determine n_j so that m_j has the same optimal distribution as m_{j-1}, that is, with the index j replacing the index j–1 in (4). Consider first the case $\lfloor F_j \rfloor > \lfloor F_{j-1} \rfloor$. Given m_{j-1}, n_j is selected at random from the following conditional distribution:

$$P\left(n_j = \lfloor F_j \rfloor - \lfloor F_{j-1} \rfloor + 1 \mid m_{j-1} = \lfloor F_{j-1} \rfloor\right) = \langle F_j \rangle,\ P\left(n_j = \lfloor F_j \rfloor - \lfloor F_{j-1} \rfloor \mid m_{j-1} = \lfloor F_{j-1} \rfloor\right) = 1 - \langle F_j \rangle,$$

$$P\left(n_j = \lfloor F_j \rfloor - \lfloor F_{j-1} \rfloor \mid m_{j-1} = \lfloor F_{j-1} \rfloor + 1\right) = \langle F_j \rangle, \text{ and } P\left(n_j = \lfloor F_j \rfloor - \lfloor F_{j-1} \rfloor - 1 \mid m_{j-1} = \lfloor F_{j-1} \rfloor + 1\right) = 1 - \langle F_j \rangle.$$

This ensures that the unconditional distribution of $m_j = m_{j-1} + n_j$ is the desired one. If $\lfloor F_j \rfloor = \lfloor F_{j-1} \rfloor$ then $n_j = 1$ with conditional probabilities

$$P\left(n_j = 1 \mid m_{j-1} = \lfloor F_{j-1} \rfloor\right) = E_j \Big/ \left(1 - \langle F_{j-1} \rangle\right) \text{ and } P\left(n_j = 1 \mid m_{j-1} = \lfloor F_{j-1} \rfloor + 1\right) = 0,$$

and $n_j = 0$ otherwise. The expectation of n_j is thus equal to E_j.

3.2 Sampling for the Intra-Annual Updates

Once the sample size is determined for each stratum within update we sample using the Bernoulli-Type method. For the first update, the units $1,\ldots,N_1$ are randomly permuted and a random number between $(i-1)/N_1$ and i/N_1 is assigned to unit i in the permuted list. Units with numbers not greater than π_0 are sampled. The random numbers are U(0, 1) distributed and the sample size n_1 is as desired.

For the j-th update we assign the random numbers as for the first update, with N_j replacing N_1. The sampling process depends, again, on the relationship between $\lfloor F_j \rfloor$ and $\lfloor F_{j-1} \rfloor$. If $\lfloor F_j \rfloor > \lfloor F_{j-1} \rfloor$ then for every value of m_{j-1} we have

$E\left(n_j \mid m_{j-1}\right) = F_j - m_{j-1}$. Denote by $\pi_j = \left(F_j - m_{j-1}\right)/N_j$ so that $E\left(n_j \mid m_{j-1}\right) = N_j \pi_j$. Units with random numbers not greater than π_j are sampled. Note that here we have a sampling scheme with a random sampling probability and a fixed expected probability. The sample size variability is minimized, with a maximal deviation of one unit between the actual and expected cumulative sample sizes. The values of π_j are $\left(F_j - \left[F_{j-1}\right]\right)/N_j$ and $\left(F_j - \left[F_{j-1}\right] - 1\right)/N_j$ with probabilities $1 - \left\langle F_{j-1} \right\rangle$ and $\left\langle F_{j-1} \right\rangle$, respectively. These values are non-negative, and in rare cases may be equal to zero, or be equal or greater than one. Such a case happens when $m_{j-1} = \left[F_{j-1}\right] + 1 = F_j$, and then $\pi_j = 0$ and $n_j = 0$. If $\pi_j > 1$, we truncate it to one and sample $n_j = N_j$ units. If $\left[F_j\right] = \left[F_{j-1}\right]$ we basically proceed as before, except for the case $m_{j-1} = \left[F_{j-1}\right] + 1$ where we take $\pi_j = 0$ and $n_j = 0$, and the case when $m_{j-1} = \left[F_{j-1}\right]$ where we take $\pi_j = E_j \Big/ \left\{N_j \left(1 - \left\langle F_{j-1} \right\rangle\right)\right\}$. We refer to this sampling scheme over updates as a conditional Bernoulli-Type (CB-T) scheme.

3.3 Sampling for the Annual Update

For the annual update a new frame based on $L_0, L_1, \ldots, L_J$ and additional newborns is constructed. As in Section 2, in preparation for the next annual sampling we adjust the random numbers assigned to units in the update frames. First, let X^*_{ji} be the random variable defined for the *i*-th unit in the *j*-th update frame as in (1), with π_j and S_j replacing π_0 and S_1, respectively. The variable X^*_{ji} is well defined when $0 < \pi_j < 1$. When $\pi_j = 0$, no units are sampled for the *j*-th update and hence X^*_{ji} is computed only for units $i \notin S_j$. Similarly when $\pi_j = 1$, X^*_{ji} is computed only for units $i \in S_j$.

As we have seen in the single update scenario of Section 2, X^*_{ji} does not have a U(0,1) distribution and thus we further adjust it using $Y_{ji} = F_{X^*_{ji}}(X^*_{ji})$. To extract $F_{X^*_{ji}}(t)$, we first note that the conditional distribution of $X^*_{ji} \mid X_{(n_0)}, \pi_j$ is, as before, a mixture of U(0, $x_{(n_0)}$) and U($x_{(n_0)}$, 1) distributions, but with respective probabilities π_j and $1 - \pi_j$. Using the independence of X_{ji}, π_j and $X_{(n_0)}$, we write their joint distribution as a product of $f_{X^*_{ji} \mid X_{(n_0)}, \pi_j}$ and the marginal distributions of $X_{(n_0)}$ and π_j. Integrating over the values of π_j it is seen in the Appendix that the unconditional distribution of X^*_{ji} is as in (2), because $\mathrm{E}(\pi_j) = \pi_0$. Therefore, the CDF of X^*_{ji} is as in Equation (3).

4. An Illustrative Example: The Manufacturing Indices Survey

The proposed method is implemented for the monthly Manufacturing Indices Survey. The survey provides measurements of the development in manufacturing and in the economy of Israel in general by estimating the total number of employees, labor cost, turnover and other economic indicators, by industry and sector.

The sampling frame is extracted from the Israeli Business Register (BR). Data for each sampling unit include the industry classification, the annual turnover (*T*) and the annual average number of employees (*E*). To determine the unit size, a regression model of *T* on *E* is fitted (in square root scale) for each industry. The unit size is equal to the observed annual turnover T_i, except for cases with a negative Studentized residual with absolute value larger than a predefined value. In these cases the size is set to the model prediction plus a random noise. For the 2004 base sample, units were stratified by 77 manufacturing industries and by 2 to 4 size strata. The number of size strata depends on the number of units in each industry, where the top size group is a take-all stratum. The Lavallée-Hidiroglou (1988) stratification algorithm with a uniform coefficient of variation (CV) across industries is used to set the strata boundaries and to allocate the sample to the take-some strata. The allocation may then be adjusted to meet additional methodological constraints, such as a minimal number of sampled cases in a stratum, a maximal sampling weight and a maximal CV of the estimated number of employees. In each take-some stratum a *srswor* is drawn using a sequential scheme. The total number of sampled units for the 2004 base sample was 2,200. The sample is updated for newborns once in two months, using the conditional B-T (CB-T) scheme described in Section 3.2.

Table 2 compares the relative standard deviations $\text{RSTD} = \sqrt{\text{Var}(m_5)} / F_5$ of the cumulative sample size obtained up to and including the fifth update, for selected industries and three sampling schemes: Bernoulli (B), Bernoulli-Type (B-T) and the proposed conditional B-T (CB-T) scheme. The variances of the cumulative sample size m_5 are computed as follows:

$$\text{Var}_{\text{B}}\left(m_5\right) = \sum_{j=1}^{5} N_j \pi_0 \left(1-\pi_0\right),\ \text{Var}_{\text{B-T}}\left(m_5\right) = \sum_{j=1}^{5} \left\langle E_j \right\rangle \left(1-\left\langle E_j \right\rangle\right),\ \text{Var}_{\text{CB-T}}\left(m_5\right) = \left\langle F_5 \right\rangle \left(1-\left\langle F_5 \right\rangle\right).$$

We present industries that had a substantial number of newborns throughout 2004. It is seen that on average using the Bernoulli scheme yields sample sizes with RSTDs ranging between about 50 to 120 percent for the selected industries, as compared to 10-85% for the CB-T method. Over all take-some strata, the B-T scheme reduces the RSTD by about 17 percent, and the CB-T by further 30 percent. Equality between values for the B-T and CB-T, as seen for industry 265 stratum 2, occurs when all newborn units are in the same update. Equality between the Bernoulli and B-T RSTD values (industry 360, stratum 3) occurs when there is at most one newborn unit per update.

5. Concluding Remarks

Revisions of monthly business surveys that account for new units are common practice. Frequently, however, it is desired that these sample revisions interfere as little as possible with the ongoing main sample. The technique suggested by Ohlsson (1995, Section 9.1.1) for coordination over time is based on recurring standard sequential sampling from an updated main frame. This scheme may affect persistants and change the selection probabilities. We have seen that it is possible to sample newborns separately from the main sample, while preserving fixed selection probabilities throughout the year as well as minimal sample size variability. We have also shown how to combine correctly sequential and Bernoulli samples to form a subsequent sample with a desired overlap. Once all units have random numbers that are U(0,1) and have a comparable scale for all previously sampled (and non sampled) units, we can continue with the standard sequential scheme.

Ernst, Valliant and Casady (2000) discuss various PRN-based techniques in the context of births and deaths and argue that

> "Though the methods are in common use, there appears to be a limited literature on their properties, particularly regarding the treatment of population changes due to births and deaths"

In this paper we have suggested a workable scheme that is based on exact distributional considerations. The deviate $y(p)$-p is not defined analytically and we leave this question for future work. It would also be interesting to study the bias induced by using X or X^* for sampling, rather than using Y.

Table 2. Relative standard deviations (RSTD) of the cumulative sample size for take-some strata in selected industries using the Bernoulli, Bernoulli-Type (B-T) and Conditional B-T (CB-T) schemes. Data from five updates, the Manufacturing Indices Survey 2004.

Industry	Stratum	Cumulative number of newborns	Expected cumulative sample size	RSTD		
				Bernoulli	B-T	CB-T
180	1	59	2.91	0.57	0.34	0.10
	2	17	3.31	0.49	0.27	0.14
202	1	24	0.85	1.06	0.96	0.41
	2	4	0.59	1.21	1.15	0.84
265	1	34	0.93	1.02	0.88	0.28
	2	6	0.91	0.97	0.32	0.32
360	1	135	1.94	0.71	0.51	0.13
	2	36	1.86	0.71	0.50	0.19
	3	4	0.84	0.97	0.97	0.43
All take-some strata		1144	67.68	0.12	0.10	0.07

Appendix: Density and Cumulative Distribution Function of X^*

The density of $X_{(n_0)}$ is

$$g_{X_{(n_0)}}(x) = \frac{N_0!}{(n_0-1)!(N_0-n_0)!} x^{n_0-1}(1-x)^{N_0-n_0},$$

and the conditional density of $X_i^* \mid X_{(n_0)} = x$ is

$$f_{X_i^*}(t \mid X_{(n_0)} = x) = \begin{cases} \dfrac{\pi_0}{x} & \text{if } \ 0 \le t \le x \\ \dfrac{1-\pi_0}{1-x} & \text{if } \ x<t \le 1 . \end{cases} \tag{5}$$

Noting that the density function of a Beta(α,β) variable on (0,1) with integer parameters is $\dfrac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!} x^{\alpha-1}(1-x)^{\beta-1}$, the unconditional density is equal to

$$\begin{aligned} f_{X_i^*}(t) &= \int_0^1 f(t \mid x_{(n_0)} = x) g(x) \mathrm{d}x \\ &= \frac{N_0!}{(n_0-1)!(N_0-n_0)!} \left\{ \pi_0 \int_t^1 x^{n_0-2}(1-x)^{N_0-n_0} \mathrm{d}x + (1-\pi_0) \int_0^t x^{n_0-1}(1-x)^{N_0-n_0-1} \mathrm{d}x \right\} \\ &= \frac{n_0}{(n_0-1)} \Pr\{\text{Beta}(n_0-1, N_0-n_0+1) > t\} + \Pr\{\text{Beta}(n_0, N_0-n_0) \le t\} \end{aligned}$$

The random variable X_{ji}^* has the same density. To see this, recall that the conditional density of $X_{ji}^* \mid X_{(n_0)} = x, \pi_j = \pi$ is as in (5) with π replacing π_0. Let $h(\pi)$ denote the probability function of π_j. The joint distribution of X_{ji}^*, $X_{(n_0)}$ and π_j is

$$f_{X_{ji}^*, X_{(n_0)}, \pi_j}(t, x, \pi) = \left\{ \frac{\pi}{x} I_{[0,x]}(t) + \frac{1-\pi}{1-x} I_{(x,1]}(t) \right\} g(x) h(\pi),$$

where $I_A(t)$ is the indicator function of a set A at t. Integrating the joint distribution over π yields $f_{X_i^*}\left(t \mid X_{(n_0)} = x\right) g(x)$ since $\mathrm{E}(\pi_j) = \pi_0$. To extract the CDF it is useful to recall that $\Pr\{\text{Beta}(k, m-k+1) \le t\} = \Pr\{\text{Bin}(m,t) \ge k\}$. Hence,

$$\begin{aligned} F_{X_i^*}(t) &= \int_0^t f_{X_i^*}(u) \mathrm{d}u \\ &= \frac{n_0}{(n_0-1)} \int_0^t \Pr\{\text{Bin}(N_0-1, u) < n_0 - 1\} \mathrm{d}u + \int_0^t \Pr\{\text{Bin}(N_0-1, u) \ge n_0\} \mathrm{d}u \\ &= \frac{n_0}{(n_0-1)} \int_0^t \sum_{i=0}^{n_0-2} \binom{N_0-1}{i} u^i (1-u)^{N_0-1-i} \mathrm{d}u + \int_0^t \sum_{i=n_0}^{N_0-1} \binom{N_0-1}{i} u^i (1-u)^{N_0-1-i} \mathrm{d}u \\ &= \frac{1}{N_0} \left[\frac{n_0}{n_0-1} \sum_{i=0}^{n_0-2} \Pr\{\text{Beta}(i+1, N_0-i) \le t\} + \sum_{i=n_0}^{N_0-1} \Pr\{\text{Beta}(i+1, N_0-i) \le t\} \right]. \end{aligned}$$

References

Ernst, L. R., Valliant, R. and Casady, R. J. (2000). Permanent and collocated random number sampling and the coverage of births and deaths. *Journal of Official Statistics*, **16**, 211-228 .

Lavalée, P. and Hidiroglou, M. A. (1988). On the stratification of skewed populations. *Survey Methodology*, **14**, 33-43.

Ohlsson, E. (1995) Coordination of samples using permanent random numbers. In *Business Survey Methods* Ed B. G. Cox, D. A. Binder, D. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott. New York: John Wiley, 153-169.

Saavedra, P. J. (1995). Fixed sample size PPS approximations with a permanent random number. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Orlando . http://www.amstat.org/sections/srms/Proceedings/papers/1995_120.pdf
Stone, C. J. (1996). *A Course in Probability and Statistics*. Belmont: Duxbury Press.