Permanent and Collocated Random Number Sampling and the Coverage of Births and Deaths

Lawrence R. Ernst, Bureau of Labor Statistics
Richard Valliant, Westat
Robert J. Casady, Consultant

U.S. Bureau of Labor Statistics
Room 3160
2 Massachusetts Avenue NE
Washington DC 20212

December 1998

Abstract

Permanent random number (PRN) and collocated random number (CRN) sampling are practical methods of controlling overlap between different samples. The techniques can be used for overlap control between samples for the same survey selected at different time periods or between different surveys at the same time period. Although the methods are in wide use, their properties, when a population is changing due to births and deaths, have not been studied extensively. Ideally, each technique should produce a sample proportionally allocated to births and persistent units when equal probability sampling is used. We study particular PRN and CRN schemes that produce fixed size samples and involve complete, rather than partial, rotation of units within strata. We present theoretical and empirical results showing the circumstances where proportional allocation is approximately obtained with these particular schemes. We also discuss important cases where PRN and CRN sampling are substantially different in their coverage of birth and persistent units.

Key words: Persistent units; post stratification; sample allocation; sequential simple random sampling.

1. Introduction

The statistical agencies of national governments routinely publish economic statistics based on surveys of business establishments. Often, different surveys use the same frame of establishments for sampling, leading to a need to somehow coordinate sampling for the surveys. Limiting the burden placed on an establishment may be critical to obtaining and maintaining cooperation when a unit is eligible for several surveys. Controlling the length of time that a sample unit is in a particular survey and the number of different surveys that a unit is in are both desirable. Maintaining a frame over time by updating for births and deaths and properly reflecting these changes in each sample are also important issues. Much of Part B, "Sample Design and Selection" in Cox, et. al. (1995), for example, is devoted to these topics.

A number of government agencies either currently use or have in the past used permanent random number (PRN) or collocated random number (CRN) sampling as a way of facilitating sample coordination among surveys and rotation of units within a survey. The general methods are described in Section 2. Statistics Sweden (Ohlsson 1992, 1995), the Institut National de la Statistique et des Etudes Economiques of France (Cotton and Hesse 1992), the Australian Bureau of Statistics (Hinde and Young 1984), and Statistics New Zealand (Templeton 1990) each have used variations of PRN or CRN sampling. Ohlsson (1995) summarizes the methods of the different countries.

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. There has been some recognition, for instance, that certain implementations may have a "birth bias," i.e., births are selected in a sample at more than their proportional rate in the population (see, e.g., Ohlsson 1995, p.166). How serious the bias is and the parameters that effect it are studied in this paper. The calculations are fairly complex, but, since the PRN and CRN methods have seen such wide use, we feel that a better understanding of their properties is worthwhile.

There are a variety of implementations of the methods. Some alternatives lead to random sizes of sample while others produce fixed sample sizes. Different methods also may handle births and sample rotation differently. The theory and empirical results we discuss refer to particular PRN and CRN schemes that (1) yield fixed sample sizes and (2) facilitate rotation of entire samples within strata. This method of complete rotation is useful in some types of surveys but, unlike some other methods, does introduce the possibility of a "birth

Section 2 briefly describes the methods and the reasons why collocated sampling was developed. The third section presents theoretical properties of particular implementations of the methods when births and deaths can occur in the population. Section 3 also describes the particular method of complete rotation we consider and reasons for its use. Section 4 gives some numerical results to illustrate the effects of different population sizes, sample sizes, birth and death rates, and the method of sampling on the relative misallocation of birth units. The empirical results also illustrate the theoretical finding that, for the versions studied here, collocated sampling exercises much tighter control over the achieved sample allocation to persistent and birth units than does PRN sampling. Section 5 is a conclusion where we briefly mention some estimation issues.

2. Description of the Methods

Denote by F_0 the initial (time period 0) frame of N_0 units. In the subsequent sections, we will consider the possibility of births and deaths that occur at later time periods. The methods described in this section are normally applied within strata but, for simplicity, we omit most references to stratification. Denote a random variable that is uniformly distributed on the interval $[0,1]$ by $U[0,1]$.

First, consider equal probability, PRN sampling. A simple random sample of fixed size *n* can be selected from the population of size N_0 by sorting the population in a random order and then picking the first *n* units after some starting point. This can be accomplished as follows:

(P1) independently assign a realization u_i of a $U[0,1]$ random number to each unit in the population,

(P2) sort the units in ascending order based on u_i, and

(P3) beginning at any point $a_0 \in [0,1]$, include the first *n* units with $u_i > a_0$. If *n* units are not obtained in the interval $(a_0, 1]$, then wrap around to 0 and continue.

This method is known as sequential simple random sampling without replacement (*srswor*) and will also be denoted simply as PRN sampling here.

We will consider only fixed sample size plans. These are of interest in survey designs where the budget is fixed and sample size is closely related to cost. An alternative is to use PRNs but sample all units with values of u_i in an interval $[a,b]$. This leads to a fixed sampling fraction but not a fixed sample size, and, thus, makes costs less predictable.

The main objection to using unmodified $U[0,1]$ permanent random numbers in sequential *srswor* is that the PRNs within detailed strata may not be well distributed. If the goal is to coordinate two or more surveys by minimizing the overlap among them, the poor distribution may lead to problems. If the u_i's are, by chance, clumped in one part of the [0,1] interval, the samples for the surveys may overlap unnecessarily. The problem can be especially severe in strata where the population size is small. As an illustration, suppose there are three surveys and that the frame and sample sizes are

$$N{=}10, \qquad n_1 = n_2 = 2\,, \quad n_3 = 4$$

Suppose further that the starting points for the three are

$$a_1 = 0, \quad a_2 = 0.20, \quad a_3 = 0.40$$

and that, by bad luck, the u_i's for all 10 units are in $[0, 0.20)$.

Using PRN sampling, units 1 and 2 in the sorted frame will be in all three surveys because survey 1 takes the first two units starting at $a_1 = 0$, while surveys 2 and 3 wrap around to 0 since there are no $u_i \geq 0.20$. As a result of clumping of the u_i's, only four distinct units are selected even though, with better placement of the u_i's, the samples could be completely non-overlapping.. This example is extreme since the probability of all 10 u_i's being in $[0, 0.20)$ is negligible, but illustrates the general idea that undesirable overlap may occur unless special measures are taken.

The use of collocated random numbers (Brewer, Early, and Joyce 1972; Brewer, Early, and Hanif 1984) is one solution to this problem This technique was originally developed as a way of reducing the randomness of sample size that accompanies Poisson sampling. The assignment of CRNs is accomplished as follows. A $U[0,1]$ random number is assigned to each unit in the frame. These numbers are sorted in ascending order and the rank R_i noted for each. A single $U[0,1]$ random number e is then generated and $u_i = (R_i - e)/N_0$ is calculated for each unit on the frame.

Collocation spaces the random numbers assigned to the population units an equal distance apart and eliminates the clumping that can occur with PRNs.

Note that in the above example there will be no overlap of the three initial samples if CRNs are used. However, if for each succeeding time period each of these samples is completely rotated by choosing as a starting point the CRN of the final unit selected for the survey for the previous time period, and if there are no births or deaths, then for the fourth time period, time period 3, every sample unit for survey 3 will also be in sample for either survey 1 or survey 2. There will never be any overlap of the samples for surveys 1 and 2 because $n_1 = n_2$.

3. The Effect of Births and Deaths

Let B denote the frame of births at time period 1 and suppose that it contains N_B units. Additionally, let F_{01} be the set of units in F_0 that are "nondeaths" or "persistents," and suppose that F_{01} contains N_{01} units. The updated frame at time 1 is $F_1 = F_{01} \cup B$ and contains $N_1 = N_{01} + N_B$ units. The number of deaths is, thus, $N_{00} = N_0 - N_{01}$. The true proportion at time 1 of units that are births is then $P_{\mathrm{T}} = N_B / N_1$. The sample selected from the time 0 frame is S_0 and the time 1 sample is S_1. In this section we give implementations of PRN and CRN sampling for handling births and deaths and examine whether the sample proportion of births, P_{S}, is near P_{T}. If P_{S} differs from P_{T} in expectation, this can be called a "selection bias," but we emphasize that this is different from the bias of an estimator—a topic briefly mentioned in Section 5. To avoid the negative connotations of the word "bias," we will refer to the quantity $E(P_S) - P_T$ as a measure of "misallocation" rather than bias. Misallocation is just a measure of how far the sample departs from being proportionally allocated to births and persistents.

3.1 Permanent Random Numbers

When a frame is periodically updated for population changes, an operationally simple method is desirable for handling births. One option is to create separate strata for births. If the same strata definitions are used for the birth strata and the persistents strata, and many of the strata have few births in the population, then even an allocation of one unit to these birth strata may result in an overall sampling rate for birth units, in comparison with persistent units, that is undesirably high. If, however, broader strata are used for the birth units than the persistents to avoid this problem, this may lead to other undesirable outcomes, such as units with large differences in size having the same selection probability.

Another obvious approach, that we will study, is to repeat for the birth units the procedure used earlier for the old units. For PRN sampling, a $U[0,1]$ random number is assigned independently to each birth unit. Birth units and persistents are then sorted together based on PRN. Let a_0^* be the PRN of the last unit in the time 0 sample and suppose that the time 1 sample consists of the first *n* units with $u_i > a_0^*$. (As can be seen from the proofs of *Propositions* 1 and 2 below, it is this assumption about the time 1 sample that results in the misallocation of birth units.) This approach appears to be quite similar to one used by Statistics Netherlands (Van Huis, Koeijers, and de Ree 1994).

This type of sampling is appropriate when the entire sample in a stratum is being rotated. The Bureau of Labor Statistics (BLS), for example, is currently using this method for its Occupational Employment Statistics survey. Data are collected annually for this survey and BLS promises respondents that they will not be in sample more than once every three years, necessitating full sample rotation annually.

An alternative to full sample rotation that is used many surveys is to rotate a part of each stratum—a topic not considered in detail here. The problem with extending the fixed sample size plan to partial rotation is that if the first n' of the time 0 sample units are replaced at time 1 by the first n' units with $u_i > a_0^*$ and the remaining units at time 0 are retained at time 1, then while the expected proportion of births among the n' new units at time 1 is the same as that given in the propositions below, there is of course no births among the units retained at time 1. In addition, although there would be deaths among the retained units they would not be compensated for in sample size by taking additional units. The PRN shift method, described by Ohlsson (1995), in which a moving fixed-length sampling window is used, avoids this problem with births, but leads to a random sample size.

The full-stratum rotation method, that we do analyze, has a slight selection bias toward births as shown in *Propositions* 1 and 2.

To make the exposition clearer, we have separated the case of no deaths (*Proposition* 1) from one having both births and deaths (*Proposition* 2). This separation will be especially useful when considering CRN sampling in Section 3.2.

***Proposition* 1.** Assume that $n < N_1$ and that there are no deaths, i.e., $N_0 = N_{01}$. Using the PRN method of sampling described above, the expected proportion of the time 1 sample that is in *B* is

$$P_{\text{PRN}} = \frac{N_B}{N_1 - 1}. \tag{1}$$

Proof: The final unit selected for the S_0 sample is a unit on the F_0 frame. This unit is not among the first $N_1 - 1$ units that can be selected for the S_1 sample. Consequently, among these $N_1 - 1$ units, exactly N_B are in *B* and, by the nature of PRNs, each of these $N_1 - 1$ units has a probability $N_B/(N_1 - 1)$ of being in *B*. This establishes (1). ■

The relative misallocation in the proportion of birth units in the S_1 sample for PRN sampling is

$$\frac{P_{\text{PRN}} - P_{\text{T}}}{P_{\text{T}}} = \frac{1}{(N_1 - 1)}. \tag{2}$$

Thus, the relative misallocation does not depend on *n* and is small when the population size N_1 is large.

***Proposition* 2.** Assume that $n < N_1$ and that there may be deaths, that is $N_{01} \le N_0$. The expected proportion of the time 1 sample that is in *B* is

$$P_{\text{PRN}} = \frac{N_B}{N_1}\left(1 + \frac{N_{01}}{N_0(N_1 - 1)}\right). \tag{3}$$

Proof. As in the first proof, if the final unit selected for the S_0 sample is in F_{01}, then each of the first $N_1 - 1$ units that can be selected for the S_1 sample has a probability $N_B/(N_1 - 1)$ of being in *B*. If, however, this final unit is a death, and hence not in F_1, then each unit in the frame F_1 has a probability N_B / N_1 of being in *B*. Since the probability that this final unit is in F_{01} is N_{01} / N_0, we have

$$P_{\text{PRN}} = \frac{N_{01} N_B}{N_0 (N_1 - 1)} + \left(1 - \frac{N_{01}}{N_0}\right) \frac{N_B}{N_1}$$

from which (3) follows after simplification. ■

The relative misallocation in the proportion of birth units in the S_1 sample for PRN in the general case is

$$\frac{P_{\text{PRN}} - P_{\text{T}}}{P_{\text{T}}} = \frac{N_{01}}{N_0 (N_1 - 1)} \leq \frac{1}{N_1 - 1}. \tag{4}$$

As in the case when there are no deaths, the relative misallocation does not depend on *n* and is small for large N_1. The relative misallocation also decreases as the death rate, $1 - N_{01} / N_1$, increases.

3.2 Collocated Random Numbers

Assigning collocated random numbers has the advantage of spreading the numbers evenly across the unit interval, but the analysis becomes quite complicated. The CRN method can also lead to some unexpected results for small populations as we show in this section. Assume that the births are handled as the original units were. A $U[0,1]$ random number is assigned to each birth. These numbers are sorted in ascending order and the rank R_{Bi} noted for each unit. A single $U[0,1]$ random number e_B is then generated and

$u_{Bi} = (R_{Bi} - e_B)/N_B$ is calculated for every birth unit on the frame. The original CRNs and the new birth CRNs are then sorted together.

The results for collocated random number sampling are considerably more complicated to derive, and we have placed proofs in the Appendix. Assume that $N_B \leq N_0$. We first consider the case $N_{01} = N_0$, i.e., there are no deaths. *Example* 1 illustrates a disconcerting phenomenon that occurs when the birth rate is extremely high, and the sample size is small.

***Example* 1**. Suppose that (N_0, N_B, n) = (4,4,1). Let the rounded CRNs for the N_0=4 old units be (0.20, 0.45, 0.70, 0.95) and the sample at time 0 be the first unit—the one with CRN=0.20. The CRN assigned to the first birth unit will be in $[0, 0.25)$. If it is less than 0.20, then the next birth unit will receive a CRN somewhere in the interval $[0.20, 0.45)$. If the CRN for the first birth is greater than 0.20, it will have to be in $(0.20, 0.25)$. In either case, the sample unit at time period 1 must be a birth. In fact, this forced selection of a birth holds regardless of the particular CRNs used.

The general result for the expected proportion of births is given in *Proposition* 3, which shows that the problem disappears when the sample size is large.

***Proposition* 3.** Assume that $N_{01} = N_0$ and $n < N_1$. For CRN sampling, the expected proportion, denoted P_{CRN}, of the sample that is birth units is

$$P_{\text{CRN}} = \frac{1}{n} \min\left\{ \left\lceil \frac{nN_0}{N_1} \right\rceil \frac{N_B}{N_0}, \left\lceil \frac{nN_B}{N_1} \right\rceil \right\} \tag{5}$$

where $\lceil x \rceil$ is the smallest integer $\geq x$.

Note that (5) implies that

$$P_{\text{CRN}} - P_{\text{T}} \geq 0 \tag{6}$$

and that

$$\frac{P_{\text{CRN}} - P_{\text{T}}}{P_{\text{T}}} = \frac{1}{n}\min\left\{\left(\left\lceil \frac{nN_0}{N_1} \right\rceil - \frac{nN_0}{N_1}\right)\frac{N_1}{N_0}, \left(\left\lceil \frac{nN_B}{N_1} \right\rceil - \frac{nN_B}{N_1}\right)\frac{N_1}{N_B}\right\} \leq \frac{1}{n}\min\left\{\frac{N_1}{N_0}, \frac{N_1}{N_B}\right\} \leq \frac{2}{n}. \tag{7}$$

It follows from (6) and (7) that the CRN misallocation is nonnegative and that the relative misallocation is bounded above by $2/n$. As n varies, the expected number of excess birth units in sample fluctuates within these bounds, but the general trend in the misallocation is downward as n increases and is small for large n.

The proof of *Proposition* 3 in the Appendix shows (see expressions A.2 - A.4) that $|n_B - nN_B/N_1| < 1$ and $|n_B/n - N_B/N_1| < 1/n$. In other words, the realized number of births selected with CRN will be within 1 unit of the expected number. Consequently, for large n, being off from the expectation by 1 unit is nothing to worry about. On the other hand, when n is small, being off by 1 may be a large percentage misallocation. For *Example* 1 we have, $P_{\text{CRN}} = \min\{\lceil 4/8 \rceil, \lceil 4/8 \rceil\} = 1$, reflecting the fact that, in this extreme case, we have no choice but to select a birth at time 1. Note that if PRN sampling was used, then *Proposition* 1 implies that $P_{\text{PRN}} = 4/7$ compared to the proportion of births in the population which is $1/2$. Thus, the degree of misallocation is less for PRNs.

An advantage of CRN sampling is that it offers tighter control over the sample allocation than PRN because of the way the CRNS are spaced in the interval. That is, while the realized number of births for CRN sampling is always within 1 of the expected number when there are no deaths, the only restrictions on n_B for PRN sampling are that $\max\{n - N_0 + 1, 0\} \le n_B \le \min\{N_B, n\}$.

The following two examples illustrate the ideas behind the proof of Proposition 3. In particular, they illustrate the key results (A.4) and (A.5) with the first expression following "min" in (A.5) applicable in *Example* 2 and the second expression applicable in *Example* 3.

***Example* 2.** Suppose that (N_0 , N_B , n) = (5,4,3) and the CRN of the last sample unit in S_0 is $a_0^* = .27$. The smallest interval of the form $(.27, x]$ for which there must be CRNs of at least $n-1$ units in the interval is $(.27, .52]$. The CRN of 1 unit in F_0 and 1 unit in *B* is in this interval. The third unit to be selected for S_1 is in *B* if and only if there is a unit in *B* with CRN in the interval (.52, .67) since the CRN of the first unit in F_0 with CRN >.52 is .67. The probability that there is such a unit in *B* is .6. Hence $n_B = 1$ or 2 and $P(n_B = 2) = .6$

***Example* 3.** The only change from *Example* 2 is that $n = 4$. Then with *x* defined as in *Example* 2, we have $x = .67$, since there are in $(.27, .67]$ the CRNS of 2 units in F_0 and the CRNs of either 1 or 2 units in *B*. Furthermore, there will be 2 units in *B* in S_1 if and only if there is at least 1 unit in *B* with CRN in the interval (.52, 87), which is always the case. Hence $P(n_B = 2) = 1$.

We next consider the general case for CRNs, that is $N_{01} \le N_0$. We proceed to derive an expression for P_{CRN}, which is much more complex than for the case $N_{01} = N_0$.

For each positive integer *m* let S_{1m} denote the first *m* units in $F_0 \cup B$ (that is including deaths) following the last unit in S_0. CRN sampling begins at the first unit after the last one in the time 0 sample and marches through the updated frame until the desired sample of size *n* is obtained, skipping over a death whenever one is encountered. In symbols, we seek the smallest *m* such that $S_{1m} \cap F_1$ has exactly *n* elements, and hence $S_1 = S_{1m} \cap F_1$. The number of deaths between times 0 and 1 is $N_{00} = N_0 - N_{01}$. The range of *m* is given by the set $M = \{m : n \le m \le n + N_{00}\}$ since, with 0 deaths, we have to traverse only *n* units to obtain the sample, but with deaths, we may need to skip over all N_{00} of them before getting a sample of *n*.

In *Proposition* 4 below $h(x,t,a,b)$ denotes the hypergeometric probability of *x* successes in *x*+*t* trials when there are *a* successes and *b* failures in the population, i.e., $h(x,t,a,b) = \binom{a}{x}\binom{b}{t} \Big/ \binom{a+b}{x+t}$.

***Proposition* 4**. Let n_{Bm} denote the number of units in $S_{1m} \cap B$, $N' = N_0 + N_B$, and $n'_{Bm} = \lfloor mN_B / N' \rfloor$. Next, let n_{0m}, n_{01m} denote the number of elements in $S_{1m} \cap F_0$, $S_{1m} \cap F_{01}$, respectively, and s_{1fm} denote the final sample unit in S_{1m}. For each *m* there are at most three different ways that *m* can be the smallest integer for which $S_{1m} \cap F_1$ has exactly *n* elements, namely:

$$n_{Bm} = n'_{Bm},\ n_{01m} = n - n'_{Bm}, \text{ and } s_{1fm} \in F_{01} \tag{8}$$

$$n_{Bm} = n'_{Bm} + 1,\ n_{01m} = n - n'_{Bm} - 1, \text{ and } s_{1fm} \in B \tag{9}$$

$$n_{Bm} = n'_{Bm} + 1,\ n_{01m} = n - n'_{Bm} - 1, \text{ and } s_{1fm} \in F_{01}. \tag{10}$$

Then, the expected proportion of a sample of size *n* that is birth units is the sum of the number of births in the events (8), (9), and (10) times their respective probabilities of occurrence divided by the total sample size. Symbolically, this is

$$P_{\text{CRN}} = \frac{1}{n} \sum_{m \in M} \left[n'_{Bm} P_{LF_{01}m} + \left(n'_{Bm} + 1 \right) P_{UBm} + \left(n'_{Bm} + 1 \right) P_{UF_{01}m} \right] \tag{11}$$

where $P_{LF_{01}m}$, P_{UBm}, and $P_{UF_{01}m}$ are the probabilities associated with (8), (9), and (10) respectively and are shown to be

$$P_{LF_{01}m} = P\left(n_{Bm} = n'_{Bm} \right) \frac{N_{01}}{N_0} h\left(n - n'_{Bm} - 1,\ m - n,\ N_{01} - 1,\ N_{00} \right) \tag{12}$$

$$P_{UB_1m} = \left[P\left(n_{Bm} = n'_{Bm} + 1 \right) - P\left(n_{B(m-1)} = n'_{Bm} + 1 \right) \right] h\left(n - n'_{Bm} - 1,\ m - n,\ N_{01},\ N_{00} \right) \tag{13}$$

$$P_{UF_{01}m} = P\left(n_{B(m-1)} = n'_{Bm} + 1 \right) \frac{N_{01}}{N_0} h\left(n - n'_{Bm} - 2,\ m - n,\ N_{01} - 1,\ N_{00} \right) \tag{14}$$

where $P\left(n_{Bm} = n'_{Bm} \right)$, $P\left(n_{Bm} = n'_{Bm} + 1 \right)$, $P\left(n_{B(m-1)} = n'_{Bm} + 1 \right)$ are computed from (A.8), (A.9) and (A.14).

Proposition 4 can be interpreted as follows. At time 1 we update the frame with births but, for the moment, we just note which units are deaths without removing them. To select the time 1 sample, we start with the first unit beyond where the time 0 sample left off. If we go some arbitrary number *m* of units further on the list (including deaths) and the number of births, n_{Bm}, in this sample plus the number of persistents, n_{01m}, equals the desired sample size *n* (after throwing away deaths), then this sample is a possibility for being the one with the smallest *m*. Because of the random ordering of the collocated units, a probability is associated with each possible value of *m*. Depending on the last unit in the S_{1m} sample, the probability of obtaining a particular number of persistents and passing over a particular number of deaths is hypergeometric. For instance, associated with (8) and (12) is

$$h\left(n-n'_{Bm}-1,\, m-n,\, N_{01}-1,\, N_{00}\right)=\frac{\binom{N_{01}-1}{n-n'_{Bm}-1}\binom{N_{00}}{m-n}}{\binom{N_{01}+N_{00}-1}{m-n'_{Bm}-1}}$$

which is the probability of (a) selecting $n-n'_{Bm}-1$ persistents from the $N_{01}-1$ population persistents (given that the last unit in S_{1m} is a persistent) and (b) having to pass over $m-n$ deaths from the N_{00} population deaths. The remaining two terms in (12) are obtained as follows. It is shown in the proof in the Appendix that if $n_{Bm}=n'_{Bm}$ then $s_{1fm}\in F_0$ and hence $P(s_{1fm}\in F_{01}|n_{Bm}=n'_{Bm})=N_{01}/N_0$. Finally $P(n_{Bm}=n'_{Bm})$, which is given in (A.8) and (A.9), is obtained from (A.4) and (A.5) in the proof of *Proposition* 3 and the fact that the distribution of n_{Bm} is independent of the set of deaths among units in F_0. The remainder of the proof of *Proposition* 4 uses similar ideas.

4. Numerical Comparisons

Because the effects of different parameters on the expected proportions of births are difficult to discern in some of the earlier formulas, we present some numerical results in this section. First, we calculated the relative misallocation for PRN sampling in (4) using various population sizes ranging from 5 to 100. Equal birth and death rates, from 0.2 to 0.8, were used so that the population was stable ($N_0=N_1$). The relative misallocation $\left(P_{\mathrm{PRN}}-P_{\mathrm{T}}\right)/P_{\mathrm{T}}$ is plotted in Figure 1 versus the N_1 population size. The four panels show the different birth rates. The relative misallocation, which is independent of sample size, can be as large as 0.20 for $N_1 = 5$ but decreases rapidly as the population size increases.

Figure 2 shows the relative misallocations for CRN sampling plotted versus the sample size for the same four birth rates. Equal birth and death rates were again used and relative misallocations were computed as $(P_{\text{CRN}} - P_{\text{T}})/P_{\text{T}}$ with P_{CRN} computed from (11). Population sizes of N_0=5, 10, 50, 100, and 200 were used. Expression (11) was evaluated for samples of n=1, 3, 5, 20, 35, and 50 in cases where $n < N_0$. The results for the different population sizes are shown in Figure 2 with different shades of gray. The points are jittered slightly to minimize overplotting. For a given sample size, the shading goes from dark gray for the smallest value of N_0 to light gray for the largest. For example, for n = 5, there are four population sizes having $n < N_0$: N_0 = 10, 50, 100, 200. The darkest gray dot is for N_0=10, the lightest gray dot is for N_0=200, while N_0=50 and 100 are intermediate shades. As the figure shows, the main determinant of misallocation is the sample size with population size much less important. For samples of size 1 the relative misallocation can be as much as 50%, but decreases rapidly as n increases.

Since the earlier analytic work was confined to two time periods, we conducted a simulation study to examine the performance of PRN and CRN sampling over three periods (t=0, 1, and 2). An initial population of N_0=200 was used and equal birth and death rates of 0.2 were assumed to generate the populations at t=1, 2. Persistents at t=1 were identified by generating a Bernoulli random variable for each of the N_0=200 time 0 units. If the random variable was greater then 0.2, then the unit was a persistent; otherwise, it was a death. To generate the number of births at t=1, a realization from a Poisson distribution was generated with parameter $0.2N_0$. For the t=2 population the procedure was repeated: each t=1 persistent was given a 0.2 chance of death and a Poisson number of births generated with parameter $0.2N_1$. PRNs and CRNs were assigned to original units and births as described in Sections 2 and 3. At both times, samples of n=1, 3, 5, 20, 35, and 50 were selected in cases where $n < N_0$ (and N_1). At time $t+1$ $(t = 0, 1)$ the sample consisted of the first n units with PRNs (or CRNs) greater than the u_i associated with the last sample unit at time t. This procedure of population generation and sample selection was repeated 10,000 times for every sample size.

Relative misallocations like those above were then computed. Let N_{B1} be the number of births in the t=1 population, N_{012} the number of units that persist through t=0, 1, and 2, N_{B12} be the number of time 1 birth units that persist at t=2, and let N_{B2} equal the number of births at time 2. Further, let n_{B1}, n_{012}, n_{B12}, and n_{B2} be the corresponding numbers of units in a sample of size n. The relative misallocations in the simulations were computed as $\left(\overline{P}_S - \overline{P}_T\right)/\overline{P}_T$, where $\overline{P}_S = \sum P_{Si} / 10{,}000$ and P_{Si} is sample proportion of units of a specified type (births or persistents) in sample i, and the summation is across the 10,000 simulations. The average population proportion was calculated as $\overline{P}_T = \sum P_{Ti} / 10{,}000$ where P_{Ti} is the true population proportion in simulation i. Due to the way that the number of births and deaths were randomly generated in the simulations, these population proportions can vary among the runs.

Figure 3 is a dotchart of the relative misallocations for n_{B1}, n_{B12}, and n_{B2}. A panel for n_{012} is omitted since $n = n_{012} + n_{B12} + n_{B2}$. Note that n_{B1} corresponds to $t=1$, and

n_{B12} and n_{B2} to $t = 2$. For sample sizes of *n*=1, 3, and 5, the misallocation is much less for PRN sampling than for CRN. The CRN technique tends to over-allocate the new births (n_{B1} and n_{B2}) at both $t = 1$ and 2 for the small sample sizes. For sample sizes of 20 and larger the discrepancy between PRN and CRN sampling disappears since the relative misallocations approach 0 for both techniques. Note that for *n*=3 and 5 in the panel for n_{B1}, the relative misallocations for PRN sampling are slightly negative although the theoretical expected value in (4) is positive. However, both are well within simulation error of (4).

CRN sampling offers the possibility of tighter control over the sample allocation than PRN because of the way the CRNs are spaced on the unit interval. To investigate this, we calculated, for both the PRN and CRN simulations, a relative misallocation for simulation run *i* as $\left(P_{Si} - P_{Ti}\right)/P_{Ti}$ where the proportions are for the four types of units mentioned above subscripted by *B*1, 012, *B*12, and *B*2. Figure 4 gives box plots of these quantities for samples of sizes 20, 35, and 50 for the 10,000 simulation runs. The box plots for the combination (PRN, $n = 20$), for example, are labeled on the horizontal axis as PRN20. Other combinations use the same convention. The whiskers in the plot extend to the extreme values of the data or a distance 1.5 times the interquartile range from the center, whichever is smaller. The horizontal white line across each box is at the median and outlying points are shown as dots. The key point to note is that the distributions of relative misallocations are much tighter for the CRN samples than for PRN. PRN sampling produces noticeably larger interquartile ranges and generates more extreme misallocations for all four types of units.

5. Conclusions

Permanent random number sampling and collocated random number sampling are appealing methods because they are simple to execute and offer practical ways of controlling sample overlap between different surveys and between time periods for a single survey. CRN sampling was developed to eliminate the clumping that can occur with PRNs and to provide more control over sample allocations. Although intuitively reasonable, the CRN method leads to much more complicated theoretical analysis than does PRN sampling.

We have studied particular implementations of PRN and CRN sampling that yield fixed sample sizes and rotate entire stratum samples at once. There are instances where equal probability PRN or CRN sampling can yield samples that in expected value are far from proportionally allocated to births and persistent units. The closeness of the PRN allocation to proportionality, for example, depends on the size of the population. The creation of small strata combined with the use of a fixed sample size PRN method with complete sample rotation should be avoided if a proportional allocation is high priority in a survey. For CRN sampling the large departures from proportionality occur at small sample sizes.

If at time 1 all units are incorrectly assumed to have a selection probability of n/N, and hence weighted by N/n, a biased estimator of total will generally result when using the PRN and CRN implementations considered here. This bias can be avoided by using the Horvitz-Thompson estimator, which differentially weights the birth and persistent units. However, calculation of the selection probabilities for the births and persistents requires the use of either *Proposition* 1, 2, 3, or 4, which, particularly in the case of *Proposition* 4, is cumbersome. An alternative, easily calculable, estimator is the post-stratified estimator with

F_{01} and *B* the two post-strata. However, if it is possible that either $n_{01} = 0$ or $n_B = 0$, then this post-stratified estimator is not unconditionally unbiased without modification.

Finally, we note that different variants of PRN and CRN sampling have been used by different countries. The most commonly used procedures appear to be ones that allow the sample size to be random, perhaps because of a realization that statistical properties of these methods are easier to derive. Each variant may require separate theory to describe its properties. We hope that the methods presented here will be useful in analyzing other alternatives that are in use.

Appendix

***Proof of Proposition* 3.** Let n_0, n_B denote the random number of sample units in a CRN sample of size *n* that are in F_0, B, respectively. Note that to establish (5) it is sufficient to show that

$$E(n_B) = \min\left\{\left\lceil \frac{nN_0}{N_1} \right\rceil \frac{N_B}{N_0}, \left\lceil \frac{nN_B}{N_1} \right\rceil\right\}. \tag{A.1}$$

Define

$$n_0' = \lfloor nN_0 / N_1 \rfloor, \;\; n_B' = \lfloor nN_B / N_1 \rfloor \tag{A.2}$$

where $\lfloor x \rfloor$ is the greatest integer $\leq x$. Since there are no deaths, $N_1 = N_0 + N_B$. Note that if nN_B/N_1 is an integer, then so is nN_0/N_1 and also $n_0' + n_B' = n$. Otherwise, $n_0' + n_B' = n-1$. We will show that

$$\text{if } nN_B/N_1 \text{ is an integer then } n_B = n_B'; \tag{A.3}$$

$$\text{and if } nN_B/N_1 \text{ is not an integer then } n_B = n_B' \text{ or } n_B = n_B' + 1, \tag{A.4}$$

and

$$P(n_B = n_B' + 1) = \min\left\{\left\lceil \frac{nN_0}{N_1} \right\rceil \frac{N_B}{N_0} - n_B', 1\right\}. \tag{A.5}$$

Observe that (A.2) and (A.3) establish (A.1) in the integer case and that (A.2), (A.4) and (A.5) establish (A.1) in the noninteger case. To establish (A.3), (A.4), and (A.5) let:

$$\ell_0' = n_0' / N_0, \;\; \ell_B' = n_B' / N_B, \;\; \ell_0'' = (n_0' + 1) / N_0, \;\; \ell_B'' = (n_B' + 1) / N_B \tag{A.6}$$

and for any $\ell > 0$ let $I(\ell) = (a_0^*, a_0^* + \ell]$, where a_0^* is the CRN for the last sample unit in S_0.

Now if nN_B/N_1 is an integer then $\ell_B' = \ell_0'$ by (A.6), and hence $I(\ell_B')$ contains n_0' CRNs from F_0 and n_B' CRNs from B and, since $n_0' + n_B' = n$, (A.3) follows.

To establish (A.4), let $\ell' = \max\{\ell'_0, \ell'_B\}$ and observe the following. $I(\ell')$ contains at least n'_0 CRNs from F_0 and n'_B CRNs from B by the definitions of ℓ'_0, ℓ'_B. Furthermore, $I(\ell')$ contains no more than n'_0 CRNs from F_0, since $I(\ell''_0)$ is the smallest interval of the form $I(\ell)$ containing $n'_0 + 1$ CRNs from F_0 and $\ell' < n / N_1 < \ell''_0$. Similarly, $I(\ell')$ contains no more than $n'_B + 1$ CRNs from B since $\ell' < \ell''_B$ (Note that while $I(\ell''_B)$ contains $n'_B + 1$ CRNs from B, it is not necessarily the smallest interval of the form $I(\ell)$ to do so, which is why it is possible for $I(\ell')$ to contain $n'_B + 1$ CRNs from B.) Thus $I(\ell')$ contains no more than $n'_0 + n'_B + 1 = n$ CRNs from $F_0 \cup B$, and (A.4) follows.

To obtain (A.5), we observe that since $I(\ell'_B)$ contains n'_B CRNs from B and since $I(\ell''_0)$ is the smallest interval of the form $I(\ell)$ containing $n'_0 + 1$ CRNs from F_0, then $P(n_B = n'_B + 1)$ is the probability that $I(\ell''_0) \sim I(\ell'_B)$ contains at least 1 CRN from B. (The notation $I(\ell_a) \sim I(\ell_b)$ means the interval ℓ_a excluding ℓ_b.) However, since the length of $I(\ell''_0) \sim I(\ell'_B)$ is $\left\lceil \frac{nN_0}{N_1} \right\rceil \frac{1}{N_0} - \frac{n'_B}{N_B}$ and there is a distance of $1/N_B$ between CRNs in B, (A.5) follows by taking the quotient of the last two expressions. ■

***Proof of Proposition* 4.** As in the statement of *Proposition* 4, n_{Bm} is the number of units in $S_{1m} \cap B$, $N' = N_0 + N_B$,

$$n'_{Bm} = \lfloor mN_B / N' \rfloor \tag{A.7}$$

and

$$P_{Lm} = P(n_{Bm} = n'_{Bm}), \quad P_{Um} = P(n_{Bm} = n'_{Bm} + 1). \tag{A.8}$$

Then it follows from (A.2), (A.3), (A.4), (A.5), (A.7) and (A.8) that

$$P_{Um} = \min\left\{\left\lceil \frac{mN_0}{N'} \right\rceil \frac{N_B}{N_0} - n'_{Bm},\, 1\right\} \text{ and } P_{Lm} = 1 - P_{Um}\,. \tag{A.9}$$

Recall that n_{0m}, n_{01m} denote the number of elements in $S_{1m} \cap F_0$, $S_{1m} \cap F_{01}$, respectively, and s_{1fm} denotes the final sample unit in S_{1m}. For each *m* the three different ways that *m* can be the smallest integer for which $S_{1m} \cap F_1$ has exactly *n* elements were given in (8), (9), and (10). Note that it is not possible to have $n_{Bm} = n'_{Bm}$ and $s_{1fm} \in B$. This is because if $s_{1fm} \in B$ and ℓ is the length of an interval with left end point the CRN for the last sample unit in S_0 and right end point the CRN of s_{1fm}, then $n_{0m} / N_0 < \ell < n_{Bm} / N_B$. Consequently,

$$n_{Bm} = N_B \frac{n_{Bm}}{N_B} > N_B \frac{n_{Bm} + n_{0m}}{N_B + N_0} > n'_{Bm}\,.$$

To compute the probability of (8), $P_{LF_{01}m}$, first note that

$$P(s_{1fm} \in F_{01} \mid n_{Bm} = n'_{Bm}) = N_{01} / N_0 \tag{A.10}$$

since $s_{1fm} \in B$ from the above discussion. Next observe that,

$$P(n_{01m} = n - n'_{Bm} \mid n_{Bm} = n'_{Bm},\ s_{1fm} \in F_{01}) = h(n - n'_{Bm} - 1,\ m - n,\ N_{01} - 1, N_{00})\,. \tag{A.11}$$

Combining (8), (A.8), (A.10) and (A.11), we obtain that

$$P_{LF_{01}m} = P_{Lm}(N_{01} / N_0) h(n - n'_{Bm} - 1,\ m - n,\ N_{01} - 1,\ N_{00})\,. \tag{A.12}$$

To obtain the probability of (9), P_{UBm}, we observe that

$$\begin{aligned} P(n_{Bm} = n'_{Bm} + 1 \text{ and } s_{1fm} \in B) &= P_{Um} - P(n_{Bm} = n'_{Bm} + 1 \text{ and } s_{1fm} \notin B) \\ &= P_{Um} - P(n_{B(m-1)} = n'_{Bm} + 1) \end{aligned} \tag{A.13}$$

and

$$P(n_{B(m-1)} = n'_{Bm} + 1) = \begin{cases} P_{U(m-1)} & \text{if } n'_{B(m-1)} = n'_{Bm} \\ 0 & \text{otherwise} \end{cases}. \quad \text{(A.14)}$$

We then combine (A.13) with

$$P(n_{01m} = n - n'_{Bm} - 1 \mid n_{Bm} = n'_{Bm} + 1,\ s_{1fm} \in B) = h(n - n'_{Bm} - 1,\ m - n,\ N_{01},\ N_{00}) \quad \text{(A.15)}$$

to obtain

$$P_{UBm} = [P_{Um} - P(n_{B(m-1)} = n'_{Bm} + 1)]h(n - n'_{Bm} - 1,\ m - n,\ N_{01},\ N_{00}). \quad \text{(A.16)}$$

Similarly we obtain that

$$P_{UF_{01}m} = P(n_{B(m-1)} = n'_{Bm} + 1)(N_{01} / N_0)h(n - n'_{Bm} - 2,\ m - n,\ N_{01} - 1,\ N_{00}). \quad \text{(A.17)}$$

We finally combine all of the above to conclude.

$$P_{\text{CRN}} = \frac{1}{n} \sum_{m \in M} [n'_{Bm} P_{LF_{01}m} + (n'_{Bm} + 1)P_{UBm} + (n'_{Bm} + 1)P_{UF_{01}m}] \quad \text{(A.18)}$$

where $M = \{m : n \le m \le n + N_{00}\}$. ■

Authors' Note

Much of this research was conducted while the second and third authors were employed by the U.S. Bureau of Labor Statistics (BLS). Any opinions expressed are those of the authors and do not constitute policy of the BLS.

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.

Brewer, K.R.W., Early, L.J., and Joyce, S.F. (1972). Selecting Several Samples from a Single Population. Australian Journal of Statistics, 14, 231-239.

Cotton, F. and Hesse, C. (1992). Co-ordinated Selection of Stratified Samples. Proceedings of Statistics Canada Symposium 92. Ottawa: Statistics Canada. 47-54.

Cox, B.G., Binder, D.A., Chinnappa, D.N., Christianson, A., Colledge, M.J., and Kott, P.S. (1995). *Business Survey Methods*. New York: John Wiley & Sons, Inc.

Hinde, R. and Young, D. (1984). Synchronised Sampling and Overlap Control Manual. Belconnen: Australian Bureau of Statistics.

Ohlsson, E. (1992). SAMU—The System for Coordination of Samples from the Business Register at Statistics Sweden—A Methodological Summary. R&D Report 1992:18. Stockholm: Statistics Sweden.

Ohlsson, E. (1995). Coordination of Samples Using Permanent Random Numbers. In *Business Survey Methods*, eds. B.G. Cox, D.A. Binder, D.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott. New York: John Wiley & Sons, Inc., 153-169.

Templeton, R. (1990). Poisson Meets the New Zealand Business Directory. The New Zealand Statistician, 25, 2-9.