PERMANENT AND COLLOCATED RANDOM NUMBER SAMPLING AND THE COVERAGE OF BIRTHS AND DEATHS

Lawrence R. Ernst, Bureau of Labor Statistics, Richard Valliant, Westat, Inc., and Robert J. Casady
Lawrence R. Ernst, BLS, 2 Massachusetts Ave., NE, Room 3160, Washington DC 20212

KEYWORDS: 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 have adopted the use of 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 use 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 different 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 are in such wide use, we feel that a better understanding of their properties is worthwhile.

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 4 gives some numerical results to illustrate the effects of different population sizes, sample sizes, and birth and death rates on the methods of sampling. Section 5 is a conclusion.

This paper is abridged from a longer one that is available from the authors. The longer paper contains more extensive empirical results, discussion of estimation issues, as well as the proofs of Propositions 3 and 4.

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

$$N=10, \qquad n_1 = n_2 = 2, \; n_3 = 4$$

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.

The use of collocated random numbers (Brewer, Early, and Joyce 1972; Brewer, Early, and Hanif 1984) is one solution to this problem. 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 ε is then generated and $u_i = (R_i - \varepsilon)/N_0$ 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.

3. The Effect of Births and Deaths

Let B_1 denote the frame of births at time period 1 and suppose that it contains N_{B1} 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 $F_1 = F_{01} \cup B_1$ and contains $N_1 = N_{01} + N_{B1}$ 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 $P_T = N_{B1}/N_1$. 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. 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 obvious approach 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 $u_i > a_0^*$. This type of sampling is appropriate when the entire sample in a stratum is being rotated. Full-stratum rotation is common in samples where many strata are used and some of the strata are rotated in each time period. An alternative would be to rotate a part of each stratum—a topic not considered here. Analysis for partial rotation is more complex. The problem is that for the intervals corresponding to the portion of the sample that is not rotated we would not get any births, and although there would be deaths in this interval they would not be compensated for in sample size by taking additional units. The full-stratum rotation method, that we do analyze, has a slight selection bias toward births as shown in *Proposition* 1.

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_1 is

$$P_{\text{PRN}} = \frac{N_{B1}}{N_1 - 1}. \tag{1}$$

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_{B1} are in B_1 and, by the nature of PRNs, each of these

$N_1 - 1$ units has a probability $N_{B1}/(N_1 - 1)$ of being in B_1. 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_1 is

$$P_{\text{PRN}} = \frac{N_{B1}}{N_1}\left(1 + \frac{N_{01}}{N_0(N_1 - 1)}\right). \tag{3}$$

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 $N_{B1}/(N_1 - 1)$ of being in B_1. 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_{B1} / N_1 of being in B_1. Since the probability that this final unit is in F_{01} is N_{01} / N_0, we have

$$P_{\text{PRN}} = \frac{N_{01} N_{B1}}{N_0(N_1 - 1)} + \left(1 - \frac{N_{01}}{N_0}\right)\frac{N_{B1}}{N_1},$$

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

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

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_{B1i} noted for each unit. A single $U[0,1]$ random number ε_{B1} is then generated and $u_{B1i} = (R_{B1i} - \varepsilon_{B1})/N_{B1}$ 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 of the complete paper. Assume that $N_{B1} \le N_0$. 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_{B1}, 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

$$P_{\text{CRN}} = \frac{1}{n}\min\left\{\left\lceil \frac{nN_0}{N_1} \right\rceil \frac{N_{B1}}{N_0}, \left\lceil \frac{nN_{B1}}{N_1} \right\rceil\right\}, \tag{5}$$

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

Note that (5) implies that

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

and that

$$\frac{P_{\text{CRN}} - P_{\text{T}}}{P_{\text{T}}} \le \frac{1}{n}\left(\left\lceil \frac{nN_0}{N_1} \right\rceil - \frac{nN_0}{N_1}\right)\frac{N_1}{N_0} \le \frac{N_1}{nN_0}. \tag{7}$$

It follows from (6) and (7) that the CRN misallocation is nonnegative and that the relative misallocation is bounded above by $N_1/(nN_0)$. 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 of the complete paper shows that $|n_{B1} - nN_{B1}/N_1| < 1$ and $|n_{B1}/n - N_{B1}/N_1| < 1/n$. 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.

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 $F_0 \cup B_1$ (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_{B1m} denote the number of units in $S_{1m} \cap B_1$, $N' = N_0 + N_B$, and $n'_{B1m} = \lfloor mN_{B1}/N' \rfloor$. 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:

$$n_{B1m} = n'_{B1m},\ n_{01m} = n - n'_{B1m}, \text{ and } s_{1fm} \in F_{01} \quad (8)$$

$$n_{B1m} = n'_{B1m} + 1,\ n_{01m} = n - n'_{B1m} - 1, \text{ and } s_{1fm} \in B_1 \quad (9)$$

$$n_{B1m} = n'_{B1m} + 1,\ n_{01m} = n - n'_{B1m} - 1, \text{ and } s_{1fm} \in F_{01} \quad (10)$$

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

$$P_{\text{CRN}} = \frac{1}{n} \sum_{m \in M} \begin{bmatrix} n'_{B1m} P_{LF_{01}m} + (n'_{B1m} + 1) P_{UB_1m} \\ + (n'_{B1m} + 1) P_{UF_{01}m} \end{bmatrix} \quad (11)$$

where $P_{LF_{01}m}$, P_{UB_1m}, and $P_{UF_{01}m}$ are the probabilities associated with (8), (9), and (10) respectively and are defined as

$$P_{LF_{01}m} = P(n_{B1m} = n'_{B1m}) \frac{N_{01}}{N_0} \times h(n - n'_{B1m} - 1,\ m - n,\ N_{01} - 1,\ N_{00})$$

$$P_{UB_1m} = \left(P(n_{B1m} = n'_{B1m} + 1) - P(n_{B1(m-1)} = n'_{B1m} + 1)\right) \times h(n - n'_{B1m} - 1,\ m - n,\ N_{01},\ N_{00})$$

$$P_{UF_{01}m} = P(n_{B1(m-1)} = n'_{B1m} + 1) \frac{N_{01}}{N_0} \times h(n - n'_{B1m} - 2,\ m - n,\ N_{01} - 1,\ N_{00})$$

where $P(n_{B1m} = n'_{B1m})$, $P(n_{B1m} = n'_{B1m} + 1)$, $P(n_{B1(m-1)} = n'_{B1m} + 1)$ are computed from

$P(n_{B1m} = n'_{B1m} + 1) \equiv P_{Um}$, $P(n_{B1m} = n'_{B1m}) = 1 - P_{Um}$,

$$P_{Um} = \min\left\{ \left\lceil \frac{mN_0}{N'} \right\rceil \frac{N_B}{N_0} - n'_{B1m},\ 1 \right\},$$

$$P(n_{B1(m-1)} = n'_{B1m} + 1) = \begin{cases} P_{U(m-1)} & \text{if } n'_{B1(m-1)} = n'_{B1m} \\ 0 & \text{otherwise} \end{cases}$$

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_{B1m}, 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) is

$$h(n - n'_{B1m} - 1,\ m - n,\ N_{01} - 1,\ N_{00}) = \frac{\binom{N_{01} - 1}{n - n'_{B1m} - 1}\binom{N_{00}}{m - n}}{\binom{N_{01} + N_{00} - 1}{m - n'_{B1m} - 1}}$$

which is the probability of (a) selecting $n - n'_{B1m} - 1$ 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.

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 $(P_{\mathrm{PRN}} - P_{\mathrm{T}})/P_{\mathrm{T}}$ 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 $(P_{\mathrm{CRN}} - P_{\mathrm{T}})/P_{\mathrm{T}}$ 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. 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.

5. Conclusion

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.

There are instances where equal probability PRN or CRN sampling can yield samples that 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 and use of PRNs should be avoided if a proportional allocation is high priority in a survey. For CRN sampling the departures from proportionality occur at small sample sizes.

Whether the achieved sample allocation is proportional or not, use of a post-stratified estimator is prudent. Post-stratification into birth and persistent groups affords protection against the possibility of conditional bias due to births and persistents having different means.

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.

Figure 1. Relative misallocation to births in PRN sampling for four birth rates and various population sizes.

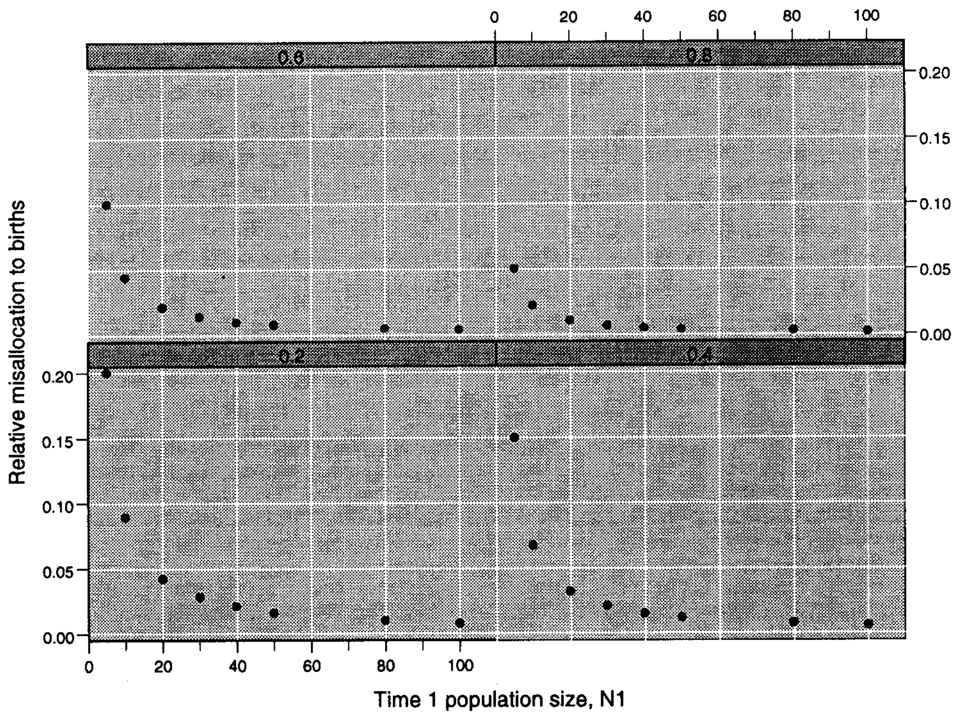

Figure 2. Relative misallocation to births in CRN sampling for four birth rates and various population sizes.