RETROSPECTIVE ASSIGNMENT OF PERMANENT RANDOM NUMBERS FOR OHLSSON'S EXPONENTIAL SAMPLING OVERLAP MAXIMIZATION PROCEDURE FOR DESIGNS WITH MORE THAN ONE SAMPLE UNIT PER STRATUM

Lawrence R. Ernst, Bureau of Labor Statistics
BLS, 2 Massachusetts Ave., N.E., Room 3160, Washington, DC 20212-0001

Key Words: Overlap maximization, Exponential sampling, Permanent random numbers, Variance estimates

1. Introduction

Many procedures have been developed for maximizing the overlap of sample units since Keyfitz's (1951) pioneering work. Ernst (1999) discusses the various overlap procedures. Some of these procedures have been developed for the following application. Units are selected with probability proportional to size (PPS), without replacement, for a survey with a stratified design. Later a new sample is to be selected using a new size measure and generally a different stratification. To reduce costs it may be desirable to maximize the expected number of units common to the two samples while preserving prespecified selection probabilities for the possible sets of sample units in a new stratum. For example, when the units being overlapped are primary sampling units (PSUs), which are geographic areas, an overlap maximization procedure can reduce the costs associated with hiring a new interviewer; when the units are ultimate sampling units, such a procedure can reduce the extra costs of an initiation interview.

Until recently, all of the procedures developed for maximizing the overlap of sample units in the application described, destroy the independence of sampling from stratum to stratum for all but the first sample selected, unless the stratifications are identical or variable sample sizes are allowed. This independence is needed to guarantee the validity of the usual variance estimation procedures. In addition, some overlap procedures, such as those of Kish and Scott (1971), Causey, Cox, and Ernst (1985), and Ernst and Ikeda (1995), preserve the predetermined selection probabilities in the new design, either in theory or in practice, only when the sample units in the initial sample were selected independently from stratum to stratum. Consequently, these procedures cannot be used in two successive redesigns.

Ohlsson (1996, 1999, 2000), however, has developed a simple overlap procedure, applicable to a wide variety of designs with a small number of sample units per stratum, that preserves this independence. Furthermore, he has shown empirically that this procedure, although not optimal, produces a reasonably large overlap in practice. The procedure, which he calls exponential sampling, was originally developed in Ohlsson (1996) for one unit per stratum designs only. Exponential sampling uses transformed permanent random numbers (PRNs) to select each sample. Consequently, it would appear that if the first sample were selected without using exponential sampling, then it would be too late to use this procedure to overlap with the initial sample. However, Ohlsson (1996) has developed a method for retrospectively assigning the PRNs after the initial sample has been drawn, which allows subsequent samples to be selected using these PRNs and exponential sampling, with the resulting selection probabilities and overlap properties the same as if the initial sample had been selected with exponential sampling. We will refer to the case when the PRNs are assigned prior to selection of the first sample as prospective exponential sampling and the case when they are assigned subsequently as retrospective exponential sampling.

Ohlsson (1999) generalizes the results of Ohlsson (1996) to n sample units per stratum, without replacement designs, where $n > 1$. However, this generalization is only for prospective exponential sampling. A key advantage of exponential sampling over other approaches to overlap maximization is the independence of sampling from stratum to stratum. This advantage would be of most interest to survey programs particularly concerned with accurate variance estimates, which would be likely to use sampling designs for which $n > 1$. Since we are unaware of any survey program at present for which the current sample was chosen using exponential sampling, a procedure for retrospectively assigning PRNs is necessary if exponential sampling is to be used in the near future for designs for which it is most attractive, that is designs for which $n > 1$. The major purpose of this paper is to present a retrospective exponential sampling procedure for general n.

In Section 2 we outline Ohlsson's prospective exponential sampling procedure for one unit per stratum designs. In Section 3 we present the generalization of this procedure to n units per stratum designs in a slightly more general form than presented in Ohlsson (1999). In Section 4 we present the procedure for retrospectively assigning the PRNs in the general case of n unit per stratum designs. The assignment depends on the set of sample units selected in the initial sample and the order in which they were

selected. In the case $n=1$, the retrospective assignment reduces to the retrospective assignment in Ohlsson (1996). The proof that the retrospective assignment of PRNs produces the same results as the prospective assignment is given in Sections 5, 6, and 7. In Section 5 we obtain an expression for the joint distribution of the *N* transformed PRNs under prospective sampling conditioned on the set of sample units selected for the initial sample and the order selected. In Section 6 we obtain the analogous expression for the joint distribution under retrospective sampling. Finally, in Section 7 we show that the joint distributions obtained in Sections 5 and Section 6 are identical, which establishes that the sampling and overlap properties are the same for prospective and retrospective exponential sampling.

In Section 8 we discuss a different issue. In the case when the units being overlapped are PSUs, which are geographic areas, it may be preferred to select the smaller PSUs independently across samples, while maximizing the overlap for the larger PSUs. In that section we describe how exponential sampling can be modified to achieve this goal. The reason for using such a hybrid type of sample selection is that if a smaller PSU is selected for two successive samples, then the expected number of the ultimate sampling units, such as households or establishments, selected for both samples in the PSU may be undesirably large. Thus respondent burden can be reduced by selecting the smaller PSUs independently across samples instead of maximizing overlap for all PSUs.

2. Prospective Exponential Sampling for One Unit per Stratum Designs

We present here an outline of Ohlsson's prospective exponential sampling procedure for one unit per stratum designs. The sampling is done separately in each stratum. Consider a stratum consisting of *N* units, where p_i is the probability of selection of unit *i*. For each unit *i*, independently generate a random number X_i, where X_i is uniformly distributed on the interval $(0,1)$, and let

$$Y_i = -\log(1-X_i), \tag{1}$$

$$\xi_i = Y_i / p_i . \tag{2}$$

Then the unit with the smallest value of ξ_i is the sample unit.

Following Ohlsson (1996), we introduce the notation $\xi \sim \text{Exp}(\lambda)$ for the fact that ξ is exponentially distributed with mean $1/\lambda$ and note that if X is uniformly distributed on the interval (0,1) and $\xi = -\log(1-X)/\lambda$, then

$$\xi \sim \text{Exp}(\lambda). \tag{3}$$

Ohlsson (1996) observed that for the procedure just described:

$$\xi_i \sim \text{Exp}(p_i),\ i=1,...,N; \tag{4}$$

$$\xi_i,\ i=1,...,N, \text{ are mutually independent;} \tag{5}$$

the sampling is independent from stratum to stratum; (6)

the probability that unit *i* is selected in sample is $p_i,\ i=1,...,N$. (7)

(4) follows from (3). (5) and (6) are direct consequences of the independence of the X_i. (7) follows from (4), (5), and a well-known result in order statistics.

Suppose a second sample is drawn from a design with the same universe but generally a different stratification and different selection probabilities, where now the probability of selection of unit *i* is p_i^*. The same procedure is employed to select the sample unit in each stratum in the new design, except p_i^* replaces p_i. In particular, the same random numbers are used in the second selection, that is X_i is a permanent random number (PRN). Then clearly (4)-(7) hold for the new design with p_i replaced by p_i^*. In addition, Ohlsson (1996) establishes that the probability that unit *i* is included in both samples is greater for exponential sampling than for independent selection of the two samples.

3. Prospective Exponential Sampling Procedure for Designs of More than One Unit per Stratum

We assume an *n* unit per stratum, without replacement design, $n>1$, for which a procedure such as Brewer's or Durbin's (Cochran 1977) in the case $n=2$ or Sampford's (1967) for general *n* is used, in which the *n* units can be selected one at a time, with $p_{i1},\ i=1,...,N$, the probability that unit *i* is the first unit to be selected; and for a unit *i* not among the first $k-1$ selected, $k \geq 2$, probability p_{ik} for its selection as the *k*-th sample unit, where p_{ik} depends on the first $k-1$ units selected and the order in which they are selected. To simplify our notation, we assume, without loss of generality, that the first *n* of the *N* units are selected in order as the *n* sample units and,

consequently, that p_{ik}, $i = k,...,N$, is the conditional probability of selecting unit *i* as sample unit *k*, given that units $1,...,k-1$ were selected in order as the first $k-1$ sample units. Then let

$$\xi_{i1} = Y_i / p_{i1}, \quad i = 1,...,N, \tag{8}$$

where Y_i is as in (1); and for $k = 2,....,n$ recursively define

$$\xi_{ik} = (p_{i(k-1)} / p_{ik})(\xi_{i(k-1)} - \xi_{(k-1)(k-1)}), \quad i = k,...,N, \tag{9}$$

where it is understood that the distribution of ξ_{ik} is conditional on units $1,...,k-1$ having been selected in order as the first $k-1$ sample units. For each *k*, the *k*-th sample unit is the unit with the smallest value of ξ_{ik}, $i = k,...,N$.

As we will demonstrate in Section 5, (23) and (24) hold and, consequently, by (7) the conditional probability of selecting unit *i* as sample unit *k*, given units $1,...,k-1$ were selected in order as the first $k-1$ sample units, is p_{ik}. Thus prospective exponential sampling preserves the selection probability determined by the underlying sampling procedure for any ordered set of *n* units.

4. The Retrospective Assignment of the PRNs

We now explain how to retrospectively assign a set of PRNs, X_i', $i = 1,...,N$, after selecting an initial sample for which the units $1,...,n$ were selected in order without the use of PRNs and exponential sampling but with the probabilities described in the previous section. For each unit *i*, associate a temporary random number Z_i uniformly distributed on the interval (0,1), and let:

$$\begin{aligned} X_i' &= 1 - \prod_{j=1}^{i} (1 - Z_j)^{p_{ij}}, \quad i = 1,...,n, \\ &= 1 - \left(\prod_{j=1}^{n} (1 - Z_j)^{p_{ij}} \right) (1 - Z_i), \quad i = n+1,...,N; \end{aligned} \tag{10}$$

$$Y_i' = -\log(1 - X_i'). \tag{11}$$

The selection of subsequent samples with the PRNs assigned retrospectively by (10) is identical to that in the case of prospective exponential sampling except Y_i' replaces Y_i.

In the particular case when $n = 1$, (10) reduces to

$$\begin{aligned} X_1' &= 1 - (1 - Z_1)^{p_{11}}, \\ X_i' &= 1 - (1 - Z_1)^{p_{i1}} (1 - Z_i), \ i = 2,...,N, \end{aligned} \tag{12}$$

which is equivalent to the procedure for retrospectively assigning the PRNs presented in Ohlsson (1996).

5. The Joint Distribution of the Transformed PRNs Conditional on the Initial Sample for Prospective Exponential Sampling

We will show in this section that each Y_i, $i = 1,...,N$, can be expressed as a linear combination of the same *N* random variables, which, conditioned on units $1,...,n$ having been selected in order as the sample units for the initial sample, have exponential distributions and are mutually independent.

For $k = 1,...n$ let

$$\begin{aligned} d_{kk} &= \xi_{kk}, \\ d_{ik} &= \xi_{ik} - \xi_{kk,} \ i = k+1,...,N, \end{aligned} \tag{13}$$

where it is understood that the distribution of d_{ik} is conditional on units $1,...,k$ having been selected in order as the first *k* sample units. Observe that, by (9) and (13), for $k = 2,...,N$,

$$\xi_{ik} = (p_{i(k-1)} / p_{ik}) d_{i(k-1)}, \quad i = k,...,N, \tag{14}$$

$$d_{ik} = (p_{i(k-1)} / p_{ik}) d_{i(k-1)} - d_{kk}, \quad i = k+1,...,N. \tag{15}$$

We proceed to show that each Y_i, $i = 1,...,N$, can be expressed as a linear combination of d_{jj}, $j = 1,...,n$, and d_{jn}, $j = n+1,...,N$, with some zero coefficients. We do this by first establishing by induction on *k* that

$$Y_i = p_{ik} d_{ik} + \sum_{j=1}^{k} p_{ij} d_{jj}, \quad i = k+1,...,N. \tag{16}$$

For $k = 1$, (16) follows from (8) and (13). Furthermore, if (16) holds for $k-1$, that is, if

$$Y_i = p_{i(k-1)} d_{i(k-1)} + \sum_{j=1}^{k-1} p_{ij} d_{jj}, \quad i = k,...,N, \tag{17}$$

then by solving (15) for $d_{i(k-1)}$ and substituting in (17) we obtain that (16) holds for *k*.

For $i = 1,...,n$ we have from (16) with $k = i-1$ that

$$Y_i = p_{i(i-1)} d_{i(i-1)} + \sum_{j=1}^{i-1} p_{ij} d_{jj} \,. \tag{18}$$

Furthermore, from (13) and (14) it follows that $d_{ii} = (p_{i(i-1)} / p_{ii}) d_{i(i-1)}$, which we combine with (18) to obtain

$$Y_i = \sum_{j=1}^{i} p_{ij} d_{jj} \,,\ i = 1,....,n \,. \tag{19}$$

Also, from (16) with $k = n$ we have that

$$Y_i = p_{in} d_{in} + \sum_{j=1}^{n} p_{ij} d_{jj} \,,\ \ i = n+1,...,N \,. \tag{20}$$

The distribution of Y_i in (19) is conditional on units $1,...,i$ having been selected as the first *i* sample units in order or, equivalently, units $1,...,n$ having been selected as the sample units in order. The distribution of Y_i in (20) is conditional on units $1,...,n$ having been selected as the sample units in order.

It is (19) and (20), which we will compare with the corresponding expressions for Y_i' given by (30). In making these comparisons we will need certain distributional information about the d_{ik}, which we proceed to establish. It is proven in Ohlsson (1996, Lemma A.2), that for $k = 1,...,n$, if $\xi_{ik} \sim \text{Exp}(p_{ik}),\ i = k,...,N$, and $\xi_{ik},\ i = k,...,N,$ are mutually independent, then:

$$\begin{aligned} &d_{kk} \sim \text{Exp}(1), \\ &d_{ik} \sim \text{Exp}(p_{ik}),\ i = k+1,...,N; \end{aligned} \tag{21}$$

$$d_{ik},\ i = k,...,N\,,\ \text{are mutually independent.} \tag{22}$$

From (21), (22) it can be established by induction on *k* that for $k = 1,...,n$:

$$\xi_{ik} \sim \text{Exp}(p_{ik}),\ i = k,...,N\,; \tag{23}$$

$$\xi_{ik},\ i = k,...,N,\ \text{are mutually independent;} \tag{24}$$

$$\begin{aligned} &d_{kk} \sim \text{Exp}(1), \\ &d_{ik} \sim \text{Exp}(p_{ik}),\ i = k+1,...,N; \end{aligned} \tag{25}$$

$$\text{the } N \text{ random variables } d_{ii},\ i = 1,...,k\,,\ \text{and } d_{ik}\,,\ i = k+1,...,N\,,\ \text{are mutually independent.} \tag{26}$$

For $k = 1$, (23) and (24) follows from (4) and (5), respectively, while (25) and (26) follow from (21)-(24). If (23)-(26) hold for $k-1$, then they hold for *k*, since: (23) for *k* follows from (25) for $k-1$, and (14); (24) follows from (26) for $k-1$ and (14); and (25) follows from (21),(23), and (24).

Finally, to establish (26) for k, we first observe that it follows from (22)-(24) that $d_{ik},\ i = k,...,N$, are mutually independent. Also, it follows from (26) for $k-1$ that $d_{ii},\ i = 1,...,k-1$, are mutually independent. Consequently, it remains only to show that

$$d_{ik},\ i = k,...,N\ \text{ is independent of } d_{ii},\ i = 1,...,k-1\,. \tag{27}$$

To establish (27), first observe that it follows from (26) for $k-1$ that $d_{i(k-1)},\ i = k,...,N$, is independent of $d_{ii},\ i = 1,...,k-1$, and consequently, by (14), that

$$\xi_{ik},\ i = k,...,N\,,\ \text{is independent of } d_{ii},\ i = 1,...,k-1\,. \tag{28}$$

In addition, by (28), the unit chosen on draw *k* among units $k,...,N$ is independent of $d_{ii},\ i = 1,...,k-1$. Consequently, if unit *k* is chosen on draw *k*, then (21) and (22) hold independently of $d_{ii},\ i = 1,...,k-1$, establishing (27).

6. The Joint Distribution of the Transformed Retrospectively Assigned PRNs

We will show in this section, analogously to the previous section, that each $Y_i',\ i = 1,...,N$, can be expressed as a linear combination of the same *N* random variables, which have exponential distributions and are mutually independent.

Let

$$\begin{aligned} d_i' &= -\log(1-Z_i),\ i = 1,...,n, \\ &= -\frac{\log(1-Z_i)}{p_{in}},\ i = n+1,...,N. \end{aligned} \tag{29}$$

Now by (10), (11), and (29)

$$\begin{aligned} Y_i' &= \sum_{j=1}^{i} p_{ij} d_j',\ i = 1,...,n, \\ &= p_{in} d_i' + \sum_{j=1}^{n} p_{ij} d_j',\ i = n+1,...,N, \end{aligned} \tag{30}$$

and furthermore, by (29) and (3), we have

$$d_i' \sim \text{Exp}(1),\ i = 1,...,n,$$
$$d_i' \sim \text{Exp}(p_{in}),\ i = n+1,...,N. \quad (31)$$

Finally by (29) and the independence of the $Z_i,\ i = 1,...,N$, we have

$d_i',\ i = 1,...,N$, are mutually independent. (32)

It is understood that $Y_i', d_i',\ i = 1,...,N$, are conditional on units $1,...,n$ having been selected as the sample units in order.

7. Comparison of the Distributions of the Previous Two Sections

The distributions of d_{ii} and $d_i',\ i = 1,...,n$, are identical by (25), with $k = i$, and (31); as are the distributions of d_{in} and $d_i',\ i = n+1,...,N$, by (25), with $k = n$, and (31). Furthermore, the set of N random variables, $d_{ii},\ i = 1,...,n$, and $d_{in},\ i = n+1,...,N$, are mutually independent by (26) with $k = n$; as are the set of $d_i',\ i = 1,...,N$, by (32). Finally, by (19), (20), and (30) it follows that Y_i and $Y_i',\ i = 1,...,N$, are the same linear combination of the corresponding random variables. Thus Y_i and $Y_i',\ i = 1,...,N$, have identical joint distributions conditional on units $1,...,n$ having been selected as the sample units in order.

Since exponential sampling depends only on these joint distributions, we have shown that conditional on the initial sample, the distributions of the subsequent samples selected by exponential sampling are identical whether Y_i or $Y_i',\ i = 1,...,N$, are used in the selection. Therefore, conditional on the initial sample, the expected number of units in a subsequent sample overlapped with the initial sample is the same whether prospective or retrospective exponential sampling is used. It follows from this result, together with the fact that the probability of selection of any set of n sample units in order for the initial sample does not depend on whether prospective PRN sampling is used to select the initial sample, that the unconditional selection probability for any set of sample units in a subsequent sample is the same for prospective and retrospective exponential sampling and that the unconditional expected number of units overlapped with the initial sample is the same for both approaches to exponential sampling.

8. Combining Overlap Maximization and Independent Selection of Two Samples

We proceed to describe how exponential sampling can be modified so that some units are selected independently across two successive samples while the overlap is maximized for the remaining units. We begin by assuming that an initial sample has been chosen using prospective exponential sampling or that PRNs have been assigned retrospectively, as described in Section 4, after selection of the initial sample. To select a second sample, first partition the N units in a new stratum into two subsets, S and L, consisting, respectively, of those units that to be selected independently of the previous sample and those units for which overlap is to be maximized with the previous sample. For $k = 1,...,n$ let $S_k = S \cap \{j : j \geq k\}$, $L_k = L \cap \{j : j \geq k\}$.

For each k we first determine whether sample unit k is to be selected from S_k or L_k. The selection between these two sets is proportional to size, where the sizes of these two sets are $\sum_{i \in S_k} p_{ik}$ and $\sum_{i \in L_k} p_{ik}$, respectively. If S_k is selected, then the k-th sample unit is chosen from among units i in S_k with probability proportional to p_{ik} independently of the previous sample. If L_k is selected, then the k-th sample unit is chosen using exponential sampling, as described in Sections 2 and 3 except that the selection is restricted to units in L_k.

Subsequent samples after the second are selected similarly to the selection of the second sample. Furthermore, any unit in a subsequent sample may be assigned to S or L regardless of its status in the previous sample. A unit that is moved from L to S is selected in the new sample independently of its selection in the previous sample. A unit moved from S to L is also selected independently of its selection in the previous sample since the selection of the previous sample was independent of the PRN assigned to that unit.

Note that a decision to originally assign a unit to S or L for the second sample or to move a unit in either direction between S and L for subsequent samples must not be based on whether the unit was in the previous sample, but rather on some characteristic of the unit itself, such as size. Generally, the desired unconditional selection probabilities of units in the new sample are not preserved if the decision on which subset to assign units is based on which units were in the previous sample.

9. References

Causey, B. D., Cox, L. H., and Ernst, L. R. (1985). Applications of Transportation Theory to Statistical Problems. Journal of the American Statistical Association, 80, 903-909.

Cochran, W. G. (1977). Sampling Techniques, 3rd ed. New York: John Wiley.

Ernst. L. R. (1999). The Maximization and Minimization of Sample Overlap Problems: A Half Century of Results. International Statistical Institute, Proceedings, Invited Papers, IASS Topics, 168-182.

Ernst, L. R. and Ikeda, M. (1995). A Reduced-Size Transportation Algorithm for Maximizing the Overlap Between Surveys. Survey Methodology, 21, 147-157.

Keyfitz, N. (1951). Sampling with Probabilities Proportionate to Size: Adjustment for Changes in Probabilities. Journal of the American Statistical Association, 46, 105-109.

Kish, L., and Scott, A. (1971). Retaining Units After Changing Strata and Probabilities. Journal of the American Statistical Association, 66, 461-470.

Ohlsson, E. (1996). Methods for PPS Size One Sample Coordination. Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University, No. 194.

Ohlsson, E. (1999). Comparison of PRN Techniques for Small Sample Size PPS Sample Coordination. Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University, No. 210.

Ohlsson, E. (2000). Coordination of PPS Samples over Time. Proceedings of the Second International Conference on Establishment Surveys, Invited Papers, 255-264.

Sampford, M. R. (1967). On Sampling Without Replacement with Unequal Probabilities of Selection. Biometrika, 54, 499-513.

Any opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics.