The Sample Overlap Problem for Systematic Sampling

Robert E. Fay[1]
[1]Westat, Inc., 1600 Research Blvd., Rockville, MD 20850

Abstract

Within the context of probability-based sampling from a finite population, a number of schemes have been studied to maximize or minimize the overlap between two sample selections while maintaining the required probabilities of selection for each. For example, in redesigning an in-person survey, it may be desirable to overlap the sampling of primary sampling units between the designs. Optimum solutions in general require mathematically and computationally complex approaches, but Ohlsson proposed simpler methods involving permanent random numbers applicable in some situations. Although not optimal, the methods are easily implemented and typically realize much of the gain achieved by the optimal solution. Ernst extended Ohlsson's methods for sequential methods such as Durbin/Brewer method, by a probabilistically correct retrospective assignment of permanent random numbers. This paper presents an extension of the Ernst approach when the first sample was selected by drawing more than one unit per stratum systematically and illustrates its efficiency with a simulation study.

Key Words: Survey redesign, dependent sampling, permanent random numbers

1. Introduction

There are many situations in which it may be advantageous to coordinate sampling for two or more surveys by separately specifying the unconditional probabilities of selection for each survey, and then selecting the samples dependently rather than independently. For example, in the sample overlap problem the goal is typically to select a sample of units according to new probabilities while attempting to retain as much of an older sample as possible. Alternatively, the situation may involve the simultaneous selection of two samples with different probabilities, again with the goal of drawing the same units into both samples as often as possible. In some cases, the probabilities may differ but the same stratification is employed; in others, the stratification may differ as well. Nathan Keyfitz (1951) is recognized for his early contribution to this problem, but decades of work followed to address various situations that would occur in practice. Ernst (1999) provided a careful review of this work.

In the United States, national samples employing in-person interviews have required a multi-stage design with primary sampling units (PSUs) typically formed from counties or groups of counties. The American Community Survey is a notable exception. Many ongoing surveys, such as the Current Population Survey (CPS), draw a new sample of PSUs only once a decade or less often, enabling the survey to retain the same interviewers for an extended period. Retaining interviewers reduces the amount of required training and increases their average experience. For the same reason, recurrent surveys seek to minimize interviewer turnover during redesign periods. Redesigns of the

CPS have used methods to increase the overlap, including incorporating Keyfitz's original method in the 1970 redesign (U.S. Bureau of the Census, 1977). Subsequently, methods developed by Ernst have been used, but the Census Bureau has considered adaptation of methods due to Esbjörn Ohlsson (1996, 1999) for the 2010 redesign (Nguyen and Gerstein, 2011; Rottach and Murphy, 2009).

Two papers of Ohlsson (1996, 1999), although unpublished, represent a key contribution to this literature. Ohlsson demonstrated that the assignment of permanent random numbers could be used to provide a flexible (although typically not fully optimal) solution to the overlap problem. In the simple case of sampling one unit with probability proportional to size (pps) from each stratum, Ohlsson (1996) showed how one could first assign permanent random numbers (PRN) drawn independently from the uniform distribution on [0,1], and then use them to draw a pps sample by a procedure called exponential sampling. The sampling procedure depends on specific properties of the exponential distribution. It is then possible to design a second sample, change the probabilities of selection and stratification, and reuse the PRNs to select the second sample. This approach increases the overlap of the two samples compared to a statistically independent selection of the second sample, in most cases substantially.

In the same paper, Ohlsson (1996) also addressed the likely circumstance that the original survey had been drawn through one-per-stratum methods but without using PRNs originally. In this situation, he provided an approach to generate PRNs for the first sample retrospectively. Specifically, one can start with a given one-per stratum pps sample and retrospectively assign PRNs consistent with the original selection under exponential sampling rules. Furthermore, he showed that for any given unit in the population, the result of applying this retrospective procedure over all possible initial samples would result in a uniform [0,1] random variable, regardless of its measure of size. Unconditionally the retrospective PRNs have the defining properties of prospectively assigned PRNs and perform similarly.

Ohlsson (1999) extended his results to some sampling procedures, such as Durbin's method, that sample more than one unit per stratum without replacement. The paper provided the prospective method only, that is, the results assumed that the PRNs had been assigned before drawing the first sample. In other words, a retrospective method for determining the PRNs was not offered.

Work of Lawrence Ernst on several facets of the sample overlap problem is well known. In one infrequently cited paper, Ernst (2001) slightly extended the Ohlsson prospective procedure to the selection of more than one PSU per stratum for any sequential procedure that selected the PSUs one at a time without replacement. His paper cited as examples Brewer's and Durbin's methods for sampling 2 units per stratum (described by Cochran (1977), and Sampford's (1967) extension for more than 2 units per stratum. He then provided a method to retrospectively assign PRNs for these sequential methods. The prospective and retrospective methods are complementary, in the sense that if the following steps are carried out:

1. One of the sequential methods is used to draw a sample, whether or not PRNs are used to do this;
2. The corresponding retrospective method is applied to the sample from step 1 (ignoring the original PRNs, if any) to assign PRNs;
3. The prospective method is applied to the retrospective PRNs from step 2 to draw a new sample;

then the resulting sample at step 3 is identical to the initial sample at step 1, and in fact the method reproduces the same order of selection as step 1.

Ernst did not specifically comment on how to apply his methods to systematic sampling. As previously noted, in the U.S. most national samples incorporating personal visit interviewing employ a first stage design in which less populous counties are grouped into primary sampling units (PSUs), stratified, and sampled. One-PSU per stratum designs are one approach, but some designers may prefer to form larger strata from which two or three PSUs are selected in order to reduce the relative variability of the stratum sizes. Having made that decision, the designers then may prefer to systematically sample of a sorted list of the PSUs to achieve "implicit stratification." The attractions of this approach are both the control over the relative variation in stratum sizes and maintaining approximately the same degree of stratification as one-per-stratum designs, while sacrificing unbiased variance estimation offered by methods, such as Durbin's, that have non-zero joint inclusion probabilities.

This paper provides a method to retrospectively assign PRNs when systematic sampling was used to select the original sample. An accompanying prospective method recreates the initial sample when the initial strata and probabilities are employed, but the retrospectively assigned PRNs can be used to select a new systematic sample with different strata and probabilities of selection, or they can be used in any other prospective sampling method employing PRNs. Software in R has been developed for the cases of $n = 2$ and $n = 3$.

The paper also extends Ernst's retrospective assignment of PRNs for sequential methods. As presented, his method requires knowledge of the order of selection of units within each strata, not just which units were selected. A method of retrospective assignment is offered for situations in which the order of selection is not known.

The next section reviews features of the exponential distribution. Section 3 summarizes the previous work one unit per stratum due to Ohlsson. Section 4 describes the extension to sequential methods with $n > 1$ and extends retrospective assignment to situations in which the original order is not known. Section 5 describes the application to systematic sampling and is followed in the next section by a simulation study illustrating a hypothetical application. The Section 7 ends the paper with a discussion.

2. The Exponential Distribution

The standard density of the exponential distribution, $f(t;\ \lambda)$, is given by

$$f(t;\ \lambda) = \lambda e^{-\lambda t}$$

for $t \in [0, \infty)$, where λ is often referred to as the rate parameter. The distribution function is

$$F(y;\ \lambda) = \int_0^y \lambda e^{-\lambda t} dt = 1 - e^{-\lambda y}$$

and the inverse of the distribution function is the corresponding quantile function defined for $q \in [0, 1)$

$$F^{-1}(q;\ \lambda) = -\log(1-q)/\lambda\ .$$

The mean of the distribution is $1/\lambda$. It is easily shown that if t has an exponential distribution with density $f(t;\ \lambda)$, then for any constant $c > 0$, ct has an exponential distribution with density $f(t;\ \lambda/c)$.

The exponential distribution has a distinctive *memoryless* property. For a random variable, t, with exponential distribution $f(t;\ \lambda)$,

$$\Pr(t > S + T | t > S) = \Pr(t > T)$$

for $S,\ T > 0$. In other words, if it is known that $t > S$, then conditionally $t - S$ has an exponential distribution with the same rate parameter, λ.

If $t_1, t_2, \ldots, t_n$ are a set of independent exponentially distributed random variables with rate parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$, then $\min(t_1, t_2, \ldots, t_n)$ has an exponential distribution with rate parameter $\lambda = \sum_{i=1}^{n} \lambda_i$. This follows from

$$\Pr(min(t_1, t_2, \ldots, t_n) > s) = \prod_{i=1}^{n} \Pr(\, t_i > s) = \prod_{i=1}^{n} e^{-\lambda_i s} = 1 - F(s;\ \lambda)$$

Also

$$\begin{aligned}
\Pr(t_k = min(t_1, t_2, \ldots, t_n)) &= \int_0^\infty f(s, \lambda_k) \prod_{i \neq k} \Pr(t_i > s)\ ds \\
&= \int_0^\infty \lambda_k\, e^{-\lambda_k s} \prod_{i \neq k} e^{-\lambda_i s}\ ds \\
&= \lambda_k \int_0^\infty \prod_{i} e^{-\lambda_i s}\, ds = \frac{\lambda_k}{\sum_{i=1}^{n} \lambda_i} = \lambda_k/\lambda
\end{aligned}$$

3. Sampling One Unit per Stratum

Ohlsson (1996) showed that one could first draw a single unit from a stratum of size N by first assigning a PRN $X_i \sim_{ind} Unif(0,1)$ to each element i. For probabilities of selection, $p_i,\ \sum_{i=1}^{N} p_i = 1$, define

$$\xi_i = \frac{-\log(1 - X_i)}{p_i}$$

which applies the quantile function of the exponential distribution with rate parameter, p_i, to X_i, giving a random element from that distribution. Thus, the ξ_i are also independent. From the last result of the previous section,

$$\Pr(\xi_k = min(\xi_1, \xi_2, \ldots, \xi_N)) = \frac{p_k}{\sum_{i=1}^{N} p_i} = p_k$$

so that selecting the sample unit k by finding the minimum of the ξ_k results in the intended probability of selection.

For a new survey, the units from the first sample can be grouped into new strata and assigned new p_i^* . New units can be added to the universe by assigning them new independently drawn PRNs, and some of the original units may leave the universe. Within each new stratum, use the PRNs to compute

$$\xi_i^* = \frac{-\log(1 - X_i)}{p_i^*}$$

and a new sample unit k^* determined by finding $\min(\xi_1^*, \xi_2^*, \ldots, \xi_{N^*}^*)$. As previously remarked, the solution does not typically produce an optimal overlap but in practice is quite effective.

If a single unit k has already been sampled from a stratum of N units with probability p_k without first assigning PRNs, Ohlsson (1996) demonstrated how to assign PRNs X_i retrospectively with the same properties as if they had been used to draw the sample initially. The procedure is first to draw a set of temporary random numbers,

$$Z_i \sim_{ind} Unif(0,1)$$

and to define for sample unit k

$$X_k = 1 - (1 - Z_k)^{p_k}$$

and for non-sampled $j \neq k$

$$X_j = 1 - (1 - Z_k)^{p_j}\left(1 - Z_j\right)$$

If the resulting retrospective PRNs are used in the prospective formulas, then

$$\xi_k = -\log(1 - Z_k)$$

and for non-sampled $j \neq k$

$$\xi_j = -\left(log(1 - Z_k) + \frac{log\left(1 - Z_j\right)}{p_j}\right)$$

ensuring $\xi_j > \xi_k$. Consequently the retrospective assignment is consistent with original sample selection. Ohlsson (1996) showed that this approach to assigning PRNs retrospectively results in independent uniform random variables, as required.

4. Sampling More than One Unit per Stratum

Ohlsson (1999) provided an extension to sampling schemes selecting more than one unit per stratum, without replacement. Examples of these methods include Durbin's and Brewer's methods for $n = 2$ (described by Cochran, 1977) and Sampford (1967) for $n = 3$ or more. These methods may be called sequential in the sense that they select one unit at a time and modify the probabilities of selection at each step. As in the previous section, we consider sampling in one stratum. By notational convention, each element has an associated probability p_i with $\sum_i p_i = 1$, and unconditional probability of selection in the sample, $\Pr(i \in s) = np_i$.

For example, in Durbin's procedure, the probability of selecting unit i as the first unit is p_i. Suppose unit k_1 is first selected. Then the probability of selecting unit $j = k_2 \neq k_1$ is

$$\Pr(k_2 = j) = p_{j1} = \frac{p_j\left(1/\left(1-2p_j\right) + 1/\left(1-2p_{k_1}\right)\right)}{1+\sum_i p_i/(1-2p_i)}$$

where the notation p_{j1} will represent the conditional probability of selecting j on the second draw conditional on selecting k_1 on the first draw.

The selected sample, s, is usually understood as the unordered set $\{k_1, k_2\}$, but the notation $< k_1, k_2 >$ will be used here to denote an ordered selection.

The general prospective procedure is as follows. To select the first unit, for each i, using PRN X_i, compute

$$\xi_{i1} = \frac{-\log(1-X_i)}{p_{i1}}$$

and find the minimum ξ_{i1} to determine k_1. Recursively for $m = 2,..n$, compute

$$\xi_{im} = \left(p_{i(m-1)}/p_{im}\right)\left(\xi_{i(m-1)} - \xi_{(m-1)(m-1)}\right)$$

for $i \notin < k_1,\ldots, k_{(m-1)} >$ and find the minimum ξ_{im} to determine sample unit k_m.

Ernst (2001) extended the results by providing a method of retrospective assignment of PRNs, X_i, again using a set of temporary random numbers, $Z_i \sim_{ind} Unif(0,1)$. For the ordered selection $< k_1, \ldots, k_n >$, the X_i are defined for the sample units, $m = 1,..n$

$$X_{k_m} = 1 - \prod_{j=1}^{m}(1-Z_{k_j})^{p_{k_j j}}$$

and for $i \notin < k_1,\ldots, k_n >$

$$X_i = 1 - \left(\prod_{j=1}^{n}\left(1-Z_{k_j}\right)^{p_{k_j j}}\right)(1-Z_i)$$

The Ernst procedure of retrospective assignment requires the specific order of selection rather than simply the unordered sample s. In applications where this information is lost or cannot be readily recreated, it is possible to extend the Ernst procedure by adding an additional level of randomization.

For a given sample s drawn by a given sequential procedure, let the set of permutations of s be denoted

$$perm(s) = \{< k_1, \ldots, k_n >, where\ k_1 \neq \cdots \neq k_n, and\ k_1, \ldots, k_n \in s\}.$$

Let $s^* \in perm(s)$. Given the probabilities assigned to the units, the sequential procedures each provide an explicit specification for $\Pr(s^*)$. Consequently, an explicit

calculation can be made of the conditional probability $\Pr(s^*|s)$ of s^* given s as $\Pr(s^*)/\Pr(s)$. Ernst's procedure can be extended by

1. drawing $s^* \in perm(s)$ according to $\Pr(s^*|s)$, and
2. assigning PRNs retrospectively based on s^*.

Then a new sample, s_2, drawn using the retrospective PRNs assigned in this manner will have the same conditional distribution given the original sample as if the PRNs had been assigned based on the known order of selection.

$$\Pr(s_2|s) = \sum_{s^* \in perm(s)} \Pr(s^*|s) \Pr(s_2|s^*)$$

Although this result applies to all sequential selection procedures, it is particularly relevant to the extension to systematic sampling.

5. Systematic Sampling

The extension to systematic sampling is particularly relevant to situations in which n is small, typically 2 or 3, and the application is to the selection of PSUs at the first stage of a multi-stage design. Consider the task of sampling two units (PSUs) from the frame in Figure 1. A common approach to systematic sampling in this situation is to pick a uniform random number between 0 and 0.5 as a start, select the unit according to the cumulative distribution. Then 0.5 is added to the random start and the corresponding unit selected as second sample unit. For example, a random start of 0.17 would select PSU 1 and PSU 4.

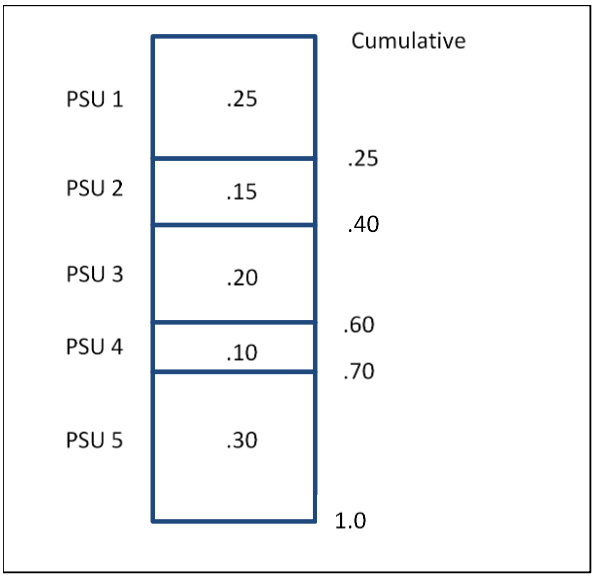

Figure 1: Illustration of a possible frame prepared for systematic sampling of two units from a stratum.

With this view of systematic sampling, some units have no chance of selection on the first draw and others no chance on the second. Ernst's method cannot be applied in a straightforward manner to reflect the zero probabilities. Cases with zero probabilities can be excluded from the calculation at a given step and a selection made among units with positive probabilities, but the method does not provide for the excluded units to come back into the process at the next step.

The approach is to re-envision systematic sampling so that units have positive probability as long as possible. For $n = 2$, the systematic sample can be drawn by selecting a uniform random number from 0 to 1, and then moving down or up by .5 as appropriate. For example, random starts of 0.17 and 0.67 result in the same sampled PSUs, but in a different order. When PSU 1 is selected first, the conditional probability of selecting PSU 3 on the second draw is 0.10/(0.10+0.10+0.05) = 0.40, as is the probability of selecting PSU 4, and the probability of selecting PSU 5 is 0.20; these values are used to determine the only 3 positive values of p_{i1} given an initial selection of PSU 1. If PSU 4 is selected first, the conditional probability of selecting PSU 1 is 1 and the rest are 0. Calculation of the conditional probabilities on the second step is tedious but easily accomplished algorithmically.

For $n = 3$, the first draw is again using a random start between 0 and 1. The second draw is performed by first selecting randomly moving forward or backward by 1/3, circularly if a boundary is reached. The second draw follows the principle of maintaining positive probabilities as much as possible. The third draw proceeds in the same direction as the second, again by a distance of 1/3.

A generalization for any n begins with a random permutation of the numbers 1, 2,…, n, $< i_1, \ldots, i_n >$ and a random uniform number u between 0 and 1. The first draw should be found from the cumulative distribution at the point $(i_1 + u - 1)/n$, the second at $(i_2 + u - 1)/n$, and so forth. Probabilities for draw m should be computed based on the $m - 1$ units selected thus far, but not conditioned on the specific permutation.

6. Simulation Study

A simulation study was performed to illustrate the application of these results. A set of functions were programmed in R to implement sampling based on prospective assignment, retrospective assignment with known order, and the extension to retrospective assignment with unknown order, for Durbin's and Brewer's methods for $n = 2$ and Sampford's for $n = 3$. Functions for prospective and retrospective assignment for systematic sampling using the approach of the previous section were also programmed for $n = 2$ and $n = 3$.

For a separate study, county-level data on the nursing home population and on the population in group homes intended for adults were used to form PSUs and 34 strata of non-self-representing PSUs for a small study. The sources of this publicly available data were the Medicare/Medicaid program (CMS) and the 2010 Census, respectively. For purposes of illustration, the 34 strata were collapsed into 17 strata for the simulation. The PSU-level totals of the nursing home population were used as the measure of size for a simulated first sample, and the counts of adults in group homes intended for adults as the measure of size in the simulated second sample.

The overlap, that is, the average number of PSUs sampled for the first sample that were retained in the second, was measure for a simulation of $n = 10{,}000$ samples, for various combinations of sampling methods used in the first selection and the second. An upper bound on the overlap that was possible was derived by the usual procedure of summing the lower of the PSU probabilities of selection in the first and second samples. A Table 1 reports these results, as well as the expected overlap if the two samples were independently sampled.

Table 1: Theoretical Bounds and Simulation Results for Expected Number of Overlapping PSUs in Two Samples

Theoretical	*First sample*	*Second sample*	*Average Overlap*
Upper bound			25.9
	Durbin	Durbin	24.8
	Brewer	Brewer	24.8
	Systematic	Systematic	18.4
	Durbin	Systematic	18.6
	Systematic	Durbin	17.8
Independent			6.2

The results of Durbin/Durbin are identical to those of Brewer/Brewer, which is not surprising given that the two methods yield identical probabilities for the unordered pairs s of sample units, although they generally disagree in the probabilities of ordered pairs. The achieved overlap is close to the optimal. The systematic/systematic, Durbin/systematic, and systematic\Durbin combinations give similar, if not identical results, less than the first two methods, but considerably better than independent sampling.

7. Discussion

As previously noted, the use of systematic sampling provides for somewhat greater implicit stratification than the methods permitting an unbiased estimate of variance. Applications of systematic sampling to strata with $n = 2$ and $n = 3$ can provide more flexibility to control the relative sizes of the strata than does $n = 1$. Thus, these results may find occasional future application for surveys involving personal visit, particularly when retention of interviewing staff is desired.

Investigating of the performance of the PRN methods for systematic sampling for the combination of revised measures of size and strata would be of interest.

Acknowledgements

I wish to thank Graham Kalton for past conversations on this topic.

References

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," *Proceedings, Invited Papers, IASS Topics,* International Statistical Institute, pp. 168-182.

_____ (2001), "Retrospective Assignment of Permanent Random Numbers for Ohlsson's Exponential Sampling Overlap Maximizing Procedure for Designs with More than One Sample Unit per Stratum." *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA.

Keyfitz, N. (1951), "Sampling with Probabilities Proportionate to Size: Adjustment for Changes in Probabilities," *Journal of the American Statistical Association*, 52, 105-109.

Nguyen, T.T. and Gerstein, A. (2011) "Sample Design Research in the 2010 Sample Redesign," *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 4359-4372.

Ohlsson, E. (1996). Methods for PPS size one sample coordination. Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University, No. 194.

_____ (1999). Comparisons of PRN techniques for small sample size PPS sample coordination. Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University, No. 210.

Rottach, R and Murphy, P. (2009), "Maximizing Overlap of Large Primary Sampling Units in Repeated Sampling: A Comparison of Ernst's Method with Ohlsson's Method," 2009 Research Conference of the Federal Committee on Statistical Methodology, fcsm.sites.usa.gov/files/2014/05/2009FCSM_Rottach_IX-C.pdf

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

U.S. Census Bureau (1978), *The Current Population Survey, Design and Methodology,* Technical Paper 40, U.S. Department of Commerce, Washington, DC.