Maximizing Overlap of Large Primary Sampling Units in Repeated Sampling: A comparison of Ernst's Method with Ohlsson's Method

Reid Rottach and Padraic Murphy[1]

U.S. Census Bureau
4600 Silver Hill Road, Washington DC 20233
padraic.a.murphy@census.gov, reid.a.rottach@census.gov

Abstract

Many large repeated or continuous demographic surveys employ a multi-stage design where large geographic areas (such as counties or clusters of contiguous counties) are sampled in the first or primary stage. Usually, a new sample of these primary sample units (PSUs) is selected periodically in order to account for changes in population, survey objectives, or other considerations. But because hiring and training new interviewers can be expensive, and replacing experienced interviewers with inexperienced ones may have an adverse effect on data quality, there is often a strong incentive to retain as many as possible of the PSUs from the "old" sample design when selecting the "new" PSU sample. At the same time, one wishes to also retain the advantages of having a probability sample. Various methods have been proposed to coordinate repeated samples with these two considerations in mind. This paper discusses and compares two such methods. The first method, due to Ernst (1986,) has been used for demographic surveys at the U.S. Census Bureau. This method does not require independent sampling between strata in the previous design, and is cast as a constrained optimization problem, so in some respect the solution is optimal. The second method, due to Ohlsson (1996, 2001,) uses exponential sampling, and does have the requirement of independent sampling; but it may be used repeatedly because it does not destroy independence in the current design.

Key Words: Repeated Sampling, Coordinated Sampling, Maximizing Overlap, Exponential Sampling, Permanent Random Numbers (PRNs)

1. Introduction

The Census Bureau is currently in the research phase of a sample redesign for several major demographic surveys. Sample will be selected following the 2010 Census. One of the areas of research is that of maximum overlap of its PSUs. We define a method of "maximum overlap" as one that increases the probability of reselecting PSUs already in sample compared to independent selections, while maintaining unconditional probability proportional to size (pps) sampling. We are interested in comparing overlap procedures that would be suitable given the constraint that they can be used repeatedly across multiple designs. The method of Ernst (1986) was first used at the Census Bureau following the 1980 redesign, and has been used in the 1990 and 2000 redesigns as well. The method of Ohlsson (1996, 2001) was an important development since it appears to be the only method that does not lead to dependent selections in the current design. This is at the heart of how Ohlsson's method satisfies the requirement for repeated use, whereas Ernst's method satisfies the requirement by not requiring independent sampling from stratum to stratum in the old design. Our interest in presenting a direct comparison between the two methods comes from this feature they have in common, and from the lack of a direct numerical comparison of the two methods in statistical literature. Ernst (1999) discusses several different features of several methods of overlap, although he does not include their expected overlaps. Ohlsson (1996) compares the expected

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical issues are those of the author and not necessarily those of the U.S. Census Bureau.

overlaps of several different methods, but he does not include Ernst, saying that in part it was to avoid linear programming.

For our numerical comparisons we use data from the previous two redesigns of the Current Population Survey (CPS), in which we formed and restratified PSUs following the 1990 and 2000 Censuses.

2. PSU creation, stratification, and probabilities of selection

The primary motivation for PSU creation is to form areas that allow manageable interviewer workloads. Many PSUs are single counties, although they may be formed from any number of contiguous counties, or in some cases, county-equivalents. The PSUs are then stratified into like groups, such as by choosing the stratification that minimizes a sampling variance.

PSUs are then assigned probabilities of selection that are proportional to size. For surveys that select one PSU per stratum, this is the measure of size of the PSU divided by the measure of size of the stratum. Otherwise, for the selection of two PSUs, the probability of selection is twice the measure of size of the PSU divided by the measure of size of the stratum; this is appropriate for without replacement sampling. Within each stratum, the joint selection probabilities for selecting pairs of PSUs are controlled using Durbin's formula (1967).

When selecting PSUs, we restrict ourselves to pps sampling, but do not necessarily constrain joint probabilities of selection of PSUs in different strata. In fact, we may follow an approach that leads to unknown joint probabilities of selection.

3. Overlap

We define overlap to be an indicator of whether a PSU, or some portion of the PSU, was in sample in two consecutive designs. For the current design, the sum of these indicator variables is the number of PSUs that were sampled in the previous design.

Expected overlap is the expected value of the number of PSUs selected in both designs. This variable is defined at the stratum level for the current design. It does not depend on any realization in the old or new designs, but integrates over all possible outcomes. From this, we may present the expected number of continuing PSUs (those sampled in both designs), which would be the sum of expected overlaps, or similarly, an average expected overlap.

Our working definition of maximum overlap is a method of sampling PSUs that:

- Is a probability sample; that is, has known selection probabilities
- Has a higher average expected overlap than sampling independently from the previous design

4. Sampling PSUs

4.1 *Notation*

In this paper we will identify PSUs as though their definition had not changed across designs, although in fact that will not be the case. The PSUs that changed definition were divided into pieces, and these pieces were treated as PSUs for the sake of overlap. For a given stratum in the new design:

i represents a PSU

π_i is its probability of selection in the new design

p_i was its probability of selection in the old design
Sums indexed by i are over all PSUs in the new design stratum

4.2 *Independent Sampling (A Lower Bound for Expected Overlap)*

The overlap procedures we examine will perform at least as well as independent sampling in each stratum, so the expected overlap of independent sampling is an obvious lower bound. Furthermore, we would like to consider the possibility of using this approach if we can't show there are real benefits to using maximum overlap procedures. With independent selection, we ignore the outcome of the previous design when selecting PSUs in the new design, so for each PSU the probability it is in both designs is the product of their probabilities. For each new design stratum, the expected overlap for independent sampling is:

$$overlap_{ind} = \sum_i p_i \pi_i$$

4.3 *Poisson Sampling (An Upper Bound for Expected Overlap)*

If we allowed variable sample sizes, we could implement a Poisson sampling procedure that would achieve an expected overlap higher than the procedures we are considering. Poisson sampling refers to an approach in which each PSU is selected independently of every other PSU in the stratum. That is, the PSUs are subjected to independent Bernoulli trials, in which the expected number of PSUs selected is a sum of the probabilities of selection. So, for example, if we were to select an expected one PSU per stratum, we may end up with some strata with no PSUs in sample, as well as strata with multiple PSUs.

Brewer, Early, and Joyce (1972) discuss an approach to sampling in which a PRN from a uniform [0,1] distribution is assigned to every PSU, and the PSU is selected if the PRN is less than the target number of PSUs times the probability of selecting that PSU. Using these PRN's in the next design will result in a maximum overlap approach to sampling, and one that is in fact optimal. Following an approach other than Poisson sampling, in which we add the constraint of a fixed sample size, will lead to an expected overlap no greater than that of the Poisson approach. We discuss Poisson sampling only as an upper bound for the expected overlap of the methods we will consider.

For each new design stratum, the expected overlap for Poisson sampling is:

$$overlap_{poi} = \sum_i \min(p_i, \pi_i)$$

4.4 *Ernst's Method*

Ernst's method is a variant of an approach outlined in Causey, Cox, and Ernst (CCE, 1985). These authors address the problem of constrained optimization directly, in which the expected overlap is maximized using numerical techniques subject to the constraints on sample size and probabilities of selection. So, CCE is truly optimal, but has the drawback that it can only be used once for pps sampling since it requires the knowledge of joint probabilities of selection. These are difficult enough to determine after it's implemented that we consider them effectively unknown.

The way in which Ernst (1986) avoids the requirement of independent sampling from stratum to stratum is by selecting only one stratum from the old design to overlap with, similar to an earlier method described in Perkins (1970). Essentially, the expected overlap is optimized given the requirement that we will select just one stratum in the old design to overlap with. It is superior to Perkins' procedure in this respect, but it is not necessarily optimal among a broader class of overlapping algorithms. Using Ernst's method, the old design stratum is chosen probabilistically, with the probabilities determined via the optimization procedure. The expected overlap for Ernst's method is determined by the optimization procedure and does not have a closed form. The expected overlap is the value of the objective function we will maximize by linear programming (PROC LP in SAS).

4.5 *Ohlsson's Method*

As with Poisson sampling, Ohlsson's method uses PRNs. For a one-PSU per stratum design, the approach is to transform the uniformly distributed PRNs and select the PSU with the smallest assigned value. In particular, for a given PSU with PRN equal to X_i, the transformed number is $\xi_i = \frac{-\log(1-X_i)}{\pi_i}$. It is very simple to implement, and correlates with the selections in the old design only through the PRN. Although not immediately apparent, it can be shown to be a method of maximizing overlap; it satisfies our constraint of being a probability sample that increases the expected overlap when compared to sampling independently from the old design.

For each new design stratum, the expected overlap for Ohlsson's method is

$$overlap_{ohl} = \sum_i \frac{p_i \pi_i}{\pi_i \sum_{j \in A_i} p_j + p_i \sum_{j \in A'_i} \pi_j}$$

Where for each *i*, define the following: D_i is the set of PSUs *{j}* in the same old and new strata as unit *i*, and satisfy $\pi_j p_i > \pi_i p_j$. A_i is the set of PSUs in the same old design stratum as *i*, but not in D_i. A_i' is the set of PSUs in the same new stratum as unit i, except those units in A_i.

This approach has been expanded to the selection of n>1 PSUs per stratum (Ohlsson, 1999), but that case will not be considered here.

4.6 *A Hybrid Approach*

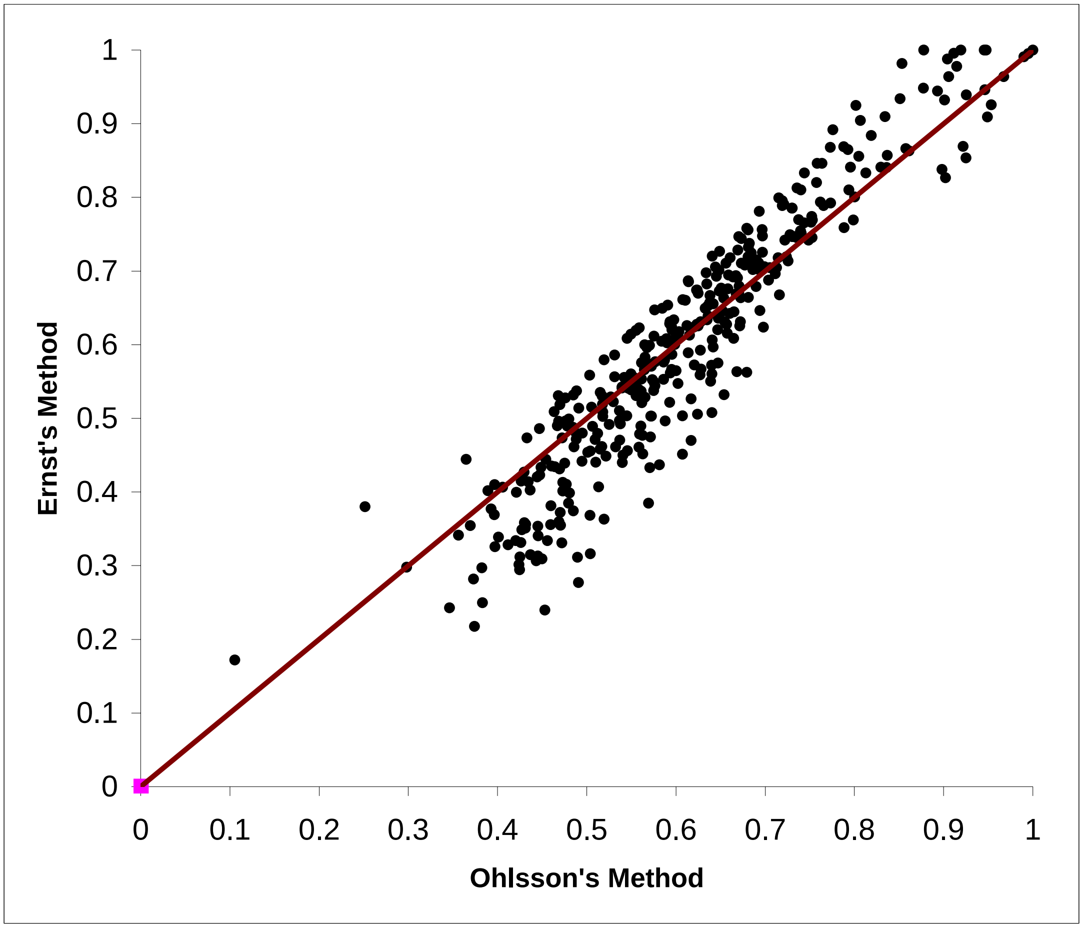
As already discussed, if we are to maintain probability sampling we cannot use Ohlsson's method without first selecting independently. One option would be to phase out Ernst's method and phase in Ohlsson's across multiple designs, by selecting independently first in some states. For example, if half the states were selected using Ernst's method, and the other half independently, then the average expected overlap would be approximately halfway between that of the two methods.

5. Results

Table 1. Average Expected Overlap

Method	Average Expected Overlap
Ernst	60%
Ohlsson	61%
Independent (Lower Bound)	35%
Poisson (Upper Bound)	81%

Figure 1. Expected Overlap For 374 Non-self-representing Strata



The average expected overlaps of Ernst and Ohlsson were very close, at 60% and 61%, respectively. Independent selection was 35% on average, and the upper bound of Poisson sampling resulted in 81%.

It is interesting to note the differing distributions of expected overlap in Ernst and Ohlsson, as shown in Figure 1. The diagonal line represents equality of the two axes. Ohlsson's method seems to perform better at the lower end of the scale, while Ernst's method seems to perform better at the higher end. Lower expected overlaps may suggest larger strata relative to the PSU sizes, which may also be related to the number of strata that are overlapped with. A possible reason for Ohlsson's performing better at the lower end is that the method uses information from all strata overlapped in the old design, rather than having to select just one to overlap with. Ernst's method will be optimal when stratum definitions do not change, and it seems that in general the method will work better when there are fewer old design strata that overlap, which may explain why it performs better at the higher end.

References

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.

Durbin, J. (1967). Design of Multi-Stage Surveys for the Estimation of Sampling Errors. Applied Statistics, 16, 152-164

Ernst, L.R. (1986). Maximizing the Overlap Between Surveys When Information is Incomplete. European Journal of Operational Research, 27, 192-200.

Ernst. Lawrence 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, Lawrence R. (2000). Discussion Paper - Session 31: Coordinating Sampling Between and Within Surveys. The Second International Conference on Establishment Surveys. Alexandria VA: American Statistical Association, 265-267.

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. The Second International Conference on Establishment Surveys. Alexandria VA: American Statistical Association, 255-264.

Perkins, W.M. (1970). 1970 CPS Redesign: Proposed Method for Deriving Sample PSU Selection Probabilities Within 1970 NSR Strata. Memorandum to Joseph Waksberg, U.S. Bureau of the Census.