Progressive Random Sampling With Stratification

Plinio A. De los Santos, Jr., Richard J. Burke, and James M. Tien, Fellow, IEEE

Abstract—A number of applications, including claims made under federal social welfare programs, requires retrospective sampling over multiple time periods. A common characteristic of such samples is that population members could appear in multiple time periods. When this occurs, and when the marginal cost of obtaining multiperiod information is minimum for a member appearing in the sample of the period being actively sampled, the progressive random sampling (PRS) method developed by the authors earlier can be applied. This paper enhances the progressive random sampling method by combining it with stratification schemes; the resultant stratified progressive random sampling (SPRS) technique is shown to provide significant improvement over traditional sampling techniques whenever stratification is appropriate. An empirical example based on a data transformation of a real-world application is provided to illustrate the practical application of the technique.

Index Terms—False Claims Act, multiperiod sampling, retrospective sampling, social welfare data estimation methods, stratified sampling.

I. INTRODUCTION

The drawing of periodic retrospective samples is required for many application areas, including expenditure estimation of federally funded welfare programs. Under such an application, a state makes a claim under one of a variety of federally funded welfare programs, but the claim is actually filed retrospectively for the fiscal quarter up to two years prior to the current quarter. Thus, the “claiming period” could be up to eight quarters prior to the current quarter. The actual amount that could be claimed under the specific program could only be determined by an examination of the case file for each recipient and totaling all the eligible expenditures. Given the enormity of the task, the state may wish to present a claim based on a “statistically valid” sample drawn from the population of eligible recipients who received payments in that period.

Sample-based claim values are ordinarily taken to be the lower 90% confidence bound on the estimate of the total claim amount. Sample-based claims must be obtained with considerable care, however, because all claims are subject to audit, and substantial penalties may be applied under the federal False Claims Act [21].

While one could easily construct claims by employing simple random sampling (RS) for each claiming period, the resources required could represent a substantial demand on those responsible for filing such claims. After the sample has been drawn, it may be necessary to physically review case files to determine the value of the eligible claims. As a practical matter, one should observe that in many cases the sample files could also yield claims for fiscal quarters subsequent to the claiming period of interest, and that all of the relevant information for those subsequent periods is obtainable for essentially no additional effort while the review of the files is underway.

This scenario suggests that more effective sampling approaches could be developed, in the sense that either the lower 90% confidence bound could be improved relative to the best estimate of the total or that the level of sampling effort could be reduced without loss to claim recoveries. Such an approach should make use of sample information progressively, saving the information concerning subsequent claiming periods, regardless of which period is being actively sampled.

This paper significantly extends the progressive random sampling approach or (PRS) [3] by developing the stratified progressive random sampling (SPRS) technique for sampling applications in which it is convenient to perform stratified random sampling (SRS). Cochran [2, pp. 87–88] lists some of these situations:

1) when required to specify precision levels for population subgroups;
2) when administrative restrictions demand the use of stratification;
3) when data collection approaches differ considerably for distinct portions of the population;
4) when stratification allows division of a heterogeneous population into subpopulations that are internally homogeneous.

As in the case of PRS, SPRS takes advantage of the fact that some information about subsequent sampling periods is known from previous samples.

This paper is organized into several sections. Section II provides a literature review. Section III presents the SPRS methodology and presents PRS as a special case of SPRS. Section IV provides empirical results obtained using a research dataset (which is based on a data transformation of a real-world application). Section V contains some concluding remarks.

II. LITERATURE REVIEW

Sampling techniques for periodic retrospective estimation problems do not appear in the literature of statistical practice [5]. In addition, traditional sampling techniques directed at obtaining measurements over time do not take into account stratified sampling considerations. As outlined by Kish [11], stratified sampling consists of the following steps:

1) dividing the entire population of sampling units into distinct subpopulations (called strata);
2) selecting a separate sample from the sampling units falling within each stratum;
3) computing a statistic (such as the mean or the total) for each stratum, as well as a combined statistic of the entire population (by appropriately weighting the estimates from each stratum);

4) computing the variation estimate for each stratum, as well as a combined variation estimate of the entire population (by weighting the variability estimates from each stratum).

The literature on sampling designs directed at obtaining measurements over time [1]. [4] includes the following related techniques.

1) Repeated surveys. This design requires obtaining data from samples periodically, but does not seek to include a sampling item in more than one period. The main purpose of this design is for deriving elementary estimates for each analysis period [7]; it uses conventional survey sampling methods.

2) Panel survey (also known as longitudinal studies). In this design, data are obtained from the same sample in different periods. One form of panel surveys, known as cohorts, studies population subgroups that share experiences during the same time period.

3) Rotating panel surveys (also known as repeated surveys with partial overlap). Eckler [6, p. 665] defines the concept of “rotation sampling” as “the process of eliminating some of the old elements from the sample and adding new elements to the same every time a new sample is drawn.” Thus, in this design sample elements are included in the survey and then surveyed a predefined number of times, and subsequently rotated out of the survey.

4) Split panel survey. It combines a panel survey with a repeated or rotating panel survey [12].

The effectiveness of the sampling design is strongly dependent on the underlying survey objectives. For the purpose of this research, the main objective is to provide estimates of population parameters at distinct points in time. Duncan and Kalton [4] make a comparison of the multiperiod sampling designs under such an objective and conclude that the repeated survey is the only method that automatically takes population changes into account. Panel, rotating panel, and split panel designs require a mechanism for taking population changes into account (i.e., introduction of new elements to the population as time passes). However, rotating panel designs have been extensively used for estimation purposes because of their variance reduction properties. Panel surveys (i.e., longitudinal studies) are more appropriate when the survey objective is measuring individual change; its main use has been in the development of causal models.

The rotating panel literature assumes that a pattern of overlap of the sampling units, which are involved in a sequence of repeated surveys, can be identified. Then, given the existence of overlapping sampling units and for any particular survey design, it is possible to derive more reliable estimators using the theory of minimum variance and unbiased linear estimators. Of course, if the surveys do not overlap, the covariance matrix becomes diagonal and the estimate for the sampling period becomes independent from estimates obtained in previous periods.

The first attempt at employing overlapping survey information to improve estimates is due to Jessen [10]. He obtains a linear unbiased estimate of the population mean by combining information from sampling in two occasions. A good presentation of Jessen’s findings is provided in [8].

Patterson [17] extends Jessen’s result to estimate the population means for more than two occasions. This result is based on the assumption that the partial correlation coefficients between the same sampling units observed in different occasions decreases geometrically as the time interval between occasions increases. A detailed exposition of Patterson’s findings is presented in [13]. Also, applications of Patterson’s results are contained in [20].

Subsequent research efforts extended Patterson’s result. Eckler [6] points out that Patterson’s result is based on a “one level rotation sampling,” and proposes a generalized sample pattern for the estimation of a linear function of the population means. Gurney and Daly [7] develop a composite estimation method to obtain the weights of the best linear unbiased estimators considering a general correlation structure. Singh [18] proposes a two-stage (nested) multiperiod sampling based on two specific rotation patterns. Huang and Ernst [9] study the performance of Gurney and Daly’s estimation method [7] in the current population survey.

There are at least three drawbacks in the application of rotating panels. First, the improvement in the estimators is highly dependent upon the correlation structure. Second, these methods require prespecifying the population elements’ rotational patterns. Third, the administration of these methods could be quite complicated, especially when dealing with complex rotational patterns.

A nontraditional method for dealing with the multiperiod survey problem is based on time series models. The main advantage of this approach is that it allows using nonoverlapping survey data in undertaking parameter estimation, and it yields models that can be employed in prediction. However, the procedure requires that the population parameters at each time period follow a stochastic underlying phenomenon. A comprehensive discussion of these techniques as applied to the problem of repeated sample surveys is provided in [19]. The key limitation of the available time series survey methods is that their application depends on the stationary conditions of the data. Furthermore, time series modeling may only be undertaken by specially trained personnel.

In sum, the available literature underscores the need to develop cost-effective sampling techniques that improve retrospective multiperiod estimations without depending upon either complex correlation structures or special data assumptions, and within the multiperiod sampling context that do not consider stratified sampling schemas. Indeed, an earlier work [3] and this current extension of that work meet the needs of a cost-effective, multiperiod sampling technique.

III. METHODOLOGY

Utilizing the notation described in Table I and taking into account the scenario described in the Section I, assume that it is necessary to obtain retrospective estimates of the total expenditures in payments to social welfare recipients. Specifically, consider that a member $j$ of a population of size $N_{i,h}$ in a
TABLE I  
NOTATION ASSOCIATED WITH MULTIPERIOD SAMPLING TECHNIQUES

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>Significance level (a.k.a. Type I error probability).</td>
</tr>
<tr>
<td>( B )</td>
<td>Bernoulli carryover function.</td>
</tr>
<tr>
<td>( b_i(h_0, h) )</td>
<td>Bernoulli probability of carrying-over payment information from period ( i ) and stratum ( h_0 ) to period ( i+1 ) and stratum ( h ).</td>
</tr>
<tr>
<td>( CC )</td>
<td>Cumulative claim after obtaining sampling estimates for ( q ) periods.</td>
</tr>
<tr>
<td>( d_i )</td>
<td>Precision estimate for the total claim amount in period ( i ) under PRS or SPRS.</td>
</tr>
<tr>
<td>( d_i^* )</td>
<td>Precision estimate for the total claim amount in period ( i ) under RS or SRS.</td>
</tr>
<tr>
<td>( E[\cdot] )</td>
<td>Expectation operator, which indicates the expected value of a variable within the brackets.</td>
</tr>
<tr>
<td>( f_{i,h} )</td>
<td>Sampling fraction under SPRS in period ( i ) and stratum ( h ).</td>
</tr>
<tr>
<td>( f_{i,h}^* )</td>
<td>Sampling fraction under SRS in period ( i ) and stratum ( h ).</td>
</tr>
<tr>
<td>( f' )</td>
<td>Target fraction per sampling period in application experiments.</td>
</tr>
<tr>
<td>( G )</td>
<td>Expected claim net growth rate.</td>
</tr>
<tr>
<td>( \text{Int}[\cdot] )</td>
<td>Operator that obtains the integer value of the number within the brackets.</td>
</tr>
<tr>
<td>( L_i )</td>
<td>Total number of strata in period ( i ).</td>
</tr>
<tr>
<td>( L_i^* )</td>
<td>Lower (1-( \alpha ))100% confidence bound for the total claim amount in period ( i ).</td>
</tr>
<tr>
<td>( \text{Max}[\cdot] )</td>
<td>Operator that obtains the maximum of the values within the brackets.</td>
</tr>
<tr>
<td>( m_{i,h} )</td>
<td>Number of carried-over population members in period ( i ) belonging to the set ( Y_{i,h} ).</td>
</tr>
<tr>
<td>( \text{Min}[\cdot] )</td>
<td>Operator that obtains the minimum of the values within the brackets.</td>
</tr>
<tr>
<td>( N_{i,h} )</td>
<td>Population size in a period ( i ) and stratum ( h ). Consequently, the total size of the population ( N_i ) is the summation of the population size in each stratum (i.e., ( N_i = \sum_{h=1}^{h} N_{i,h} )).</td>
</tr>
<tr>
<td>( n_{i,h} )</td>
<td>Size of the random sample drawn from the population associated with stratum ( h ) in period ( i ) under SPRS. As expected, ( n_i = \sum_{h=1}^{h} n_{i,h} ).</td>
</tr>
<tr>
<td>( n_{i,h}^* )</td>
<td>Size of the random sample drawn from the population associated with stratum ( h ) in period ( i ) under SRS. As expected, ( n_i^* = \sum_{h=1}^{h} n_{i,h}^* ).</td>
</tr>
<tr>
<td>( n_i^r )</td>
<td>Target sample size for period ( i ) in application experiments.</td>
</tr>
<tr>
<td>( p_{i,h} )</td>
<td>Carryover proportion associated with stratum ( h ) in period ( i ).</td>
</tr>
<tr>
<td>( \text{PRS} )</td>
<td>Progressive Random Sampling.</td>
</tr>
<tr>
<td>( \text{SRS} )</td>
<td>Stratified Random Sampling.</td>
</tr>
<tr>
<td>( S(h_i) )</td>
<td>Standard error for the total claim amount in period ( i ).</td>
</tr>
<tr>
<td>( \sigma_{i,h}^2 )</td>
<td>Variance of the claim amount per individual in period ( i ) and stratum ( h ).</td>
</tr>
<tr>
<td>( \text{RS} )</td>
<td>Simple Random Sampling.</td>
</tr>
<tr>
<td>( \text{SA} )</td>
<td>Sampling approach.</td>
</tr>
<tr>
<td>( \varepsilon(i,h) )</td>
<td>Standard error estimate for the claim amount per individual in period ( i ) and stratum ( h ).</td>
</tr>
<tr>
<td>( t_{\alpha/2} )</td>
<td>( t ) Student distribution (1-( \alpha )/2)100% percentile.</td>
</tr>
<tr>
<td>( T_i )</td>
<td>Total claim amount in period ( i ).</td>
</tr>
<tr>
<td>( V(M_{i,h}) )</td>
<td>Estimated variance per individual in period ( i ) and stratum ( h ).</td>
</tr>
<tr>
<td>( w_{i,h} )</td>
<td>Sum of the claim amounts for the population members in period ( i ) belonging to the set ( Y_{i,h} ).</td>
</tr>
<tr>
<td>( X_{i,j,h} )</td>
<td>Claim amount in period ( i ) for population member ( j ) in stratum ( h ).</td>
</tr>
<tr>
<td>( \bar{x}_{i,h} )</td>
<td>Average claim amount per individual in period ( i ) and in stratum ( h ).</td>
</tr>
<tr>
<td>( Y_{i,h} )</td>
<td>Set containing only those population members in period ( i ) and stratum ( h ) whose claim amount are known from earlier sampling periods.</td>
</tr>
</tbody>
</table>

Note: In the case of RS and PRS techniques, the stratum sub-index \( h \) is not necessary because under those techniques the total number of strata equals 1. As an example, the sampling fraction under PRS is \( f_i \) instead of \( f_{i,h} \). 

Specific period \( i \) and stratum \( h \) has some claim amount \( X_{i,j,h} \), with variance \( \sigma_{i,h}^2 \) and mean \( \bar{x}_{i,h} \). The purpose of drawing a sample of size \( n_{i,h}^* \) for claiming period \( i \) and for each stratum \( h \) is to estimate the total of the \( x_{i,j,h} \) values over all population members \( j \) contained in all strata \( h \). Member \( j \) may also be present in subsequent periods, and will have a unique label for identification purposes. 
Since the sampling is done retrospectively, the period from which the sample is drawn (i.e., the claiming period) is always a number of periods in the past relative to the period when
the sample is physically gathered (i.e., the sampling period). As a consequence, there is a collection of periods for which information may be obtained while sampling the claiming period. This aggregation of periods is called the sample window, and it begins in the claiming period and ends at the sampling period. To illustrate these concepts, consider the following example: a sample containing information from period $i$ is physically gathered in period $i + k$; under this example, the claiming period is $i$, the sampling period is $i + k$ and the length of the sampling window is $k + 1$.

Suppose that the claim amounts of member $j$ pertaining to subsequent periods within the sample window (if member $j$ is present in the populations of those periods) can be obtained without significant additional expenditure of effort. Then, for any specific claiming period $i$, it is possible to identify some members from a population $N_{i,h}$ whose $x_{i,j,h}$ values were obtained from previous samples of earlier sampling periods. The members present in period $i$ that are known from earlier sampling periods are members of the set $Y_{i,h}$. This set is excluded from the population that is sampled in period $i$ and stratum $h$. The sum of the claim amounts from these “known” members in period $i$ and stratum $h$ is represented by

$$w_{i,h} = \sum_{j \in Y_{i,h}} x_{i,j,h} \quad (1)$$

The typical length of a claiming period is a fiscal quarter, i.e., three months, but the period could be any convenient time interval. Given a specific significance level $\alpha$, a claim for the period $i$ is usually based upon the lower $(1 - \alpha)100\%$ confidence bound $LT_i$ of the total claim amount $T_i$ for that claiming period.

The fundamental concept behind PRS methods is the use of information obtained from previous samples to complement sample data obtained for the current claiming period. This concept requires that members of a population for which information is already on hand from previous samples are to be removed from the population prior to sampling for the current claiming period, since such information is known and is not subject to sampling error in the current claiming period. The statistical estimates derived for the current claiming period are therefore obtained from a current period sampling component that is subject to sampling variation, together with previous periods’ components that are not subject to sampling variation in the current period.

Consider the following example. Suppose that it is required to estimate the total claim amount for period $i$ after already drawing samples in periods $i - 1, i - 2, \ldots, i - k$. In this case and for each stratum $h$ in $i$, the set of elements $Y_{i,h}$ is going to be composed of $m_{i,h}$ members of the population $N_{i,h}$ whose claim amounts were obtained in the samples drawn in periods $i - 1$ up to $i - k$. Then, under progressive random sampling (PRS), the estimate of the total amount for period $i$, $\hat{T}_i$, is composed of two terms:

1) the expected value of the total claim amount obtained from the random sample taken from the adjusted populations in period $i$ and each stratum $h$;

2) the sum of claim amounts of the population elements contained in all $Y_{i,h}$ sets.

More specifically, under stratified progressive random sampling (SPRS) the sampling estimate for the total $\hat{T}_i$ can be represented by the following:

$$\hat{T}_i = \sum_{h=1}^{L_i} \{ (N_{i,h} - m_{i,h}) \bar{x}_{i,h} \} + \sum_{h=1}^{L_i} w_{i,h} \quad (2)$$

where the first term of (2) is the estimate based on the sample from the reduced population in claiming period $i$, and the second term is the previously known value for the population members belonging to the $Y_{i,h}$ sets. Also, under SPRS, the sampling estimate of the standard error of the total claim amount is the standard error of the first term of (2), i.e.,

$$S(T_i) = \sum_{h=1}^{L_i} \sqrt{V(x_{i,h})} \left( \frac{N_{i,h} - m_{i,h} - n_{i,h}}{n_{i,h}} \right) \quad (3)$$

The following remarks should be made concerning the above two equations.

1) The length of the sample window need not be fixed, and may be varied by the decision maker over time. As an example, samples may be made for financial audits on a monthly, quarterly, or annual basis.

2) All of the samples are taken without replacement.

3) Confidence bounds on the total amount are based upon the standard error derived from the random sample drawn for the current claiming period alone.

4) PRS is a special case of SPRS that occurs when the population is not stratified or when $L_i$ equals 1 for all $i$.

5) Simple RS is a special case of PRS that occurs when $m_{i,h}$ equals zero (i.e., when no information from population members is carried over from previous sampling periods).

6) Similarly, SRS is a special case of SPRS that occurs when $m_{i,h}$ equals zero for all $h$ in period $i$.

The PRS techniques can improve RS or SRS results in either of two ways: by demanding smaller sample sizes for confidence intervals of a given width, or by narrowing confidence intervals for equal sample sizes. In the next section, each of these areas of improvement is empirically illustrated.

IV. EMPIRICAL RESULTS

In this section, the empirical results from comparing the following multiperiod sampling methods are summarized for:

1) Sample random sampling (RS);

2) Stratified sample random sampling (SRS);

3) Progressive sample random sampling (PRS);

4) Stratified progressive sample random sampling (SPRS).

To assess the relative effectiveness of these four sampling methods, at different sampling fraction levels, a cumulative claim (CC) performance metric is employed. Such a metric assumes that the total claim is the summation of the lower $(1 - \alpha)100\%$ confidence bound for the total claim amount in each claiming period truncated at zero. For each empirical setting, this metric is calculated as follows:

$$CC = \sum_{i=1}^{q} \text{Max} [\hat{T}_i - t_{\alpha/2}S(T_i), 0] \quad (4)$$
When employing stratified sampling schemas (i.e., SRS or SPRS), the samples are allocated to each stratum using Newman’s optimal allocation of samples as follows:

\[ n_{i,h} = n_i \left( \frac{N_{i,h}S_{i,h}}{\sum_{h=1}^{L_i} N_{i,h}S_{i,h}} \right) \] (5)

When considering (5), this assumes that the unit sampling cost in each stratum in the period \( i \) is the same.

Some important information associated with the empirical experiments is:
1) the payment time aggregation unit employed is quarters (i.e., three payment months);
2) each experimental setting was replicated 100 times;
3) experiments are conducted using four levels of target sampling fraction (i.e., \( f^i = 0.05, 0.10, 0.15, \) and \( 0.20 \)). Such information is necessary to compute a target sample size to draw from in any claiming period \( i \), as follows:

\[ n^i_t = \left\{ \begin{array}{ll}
\text{Int}(f^i N_i), & \text{if } \{ f^i N_i - \text{Int}(f^i N_i) \} = 0 \\
\text{Int}(f^i N_i) + 1, & \text{otherwise}
\end{array} \right. \] (6)

However, the actual sample size \( n_i \) is bounded by the size of the population subject to random variation. This implies that

\[ n_i = \min \left( n^i_t, N_i - \sum_{h=1}^{L_i} m_{i,h} \right) \] (7)

As a consequence, under progressive sampling schemas, the actual sample size is smaller than the target in quarters when the following condition occurs

\[ f^i + \frac{\sum_{h=1}^{L_i} m_{i,h}}{N_i} > 1 \] (8)

4) Experimental results are compared in terms of:
   a) The expected claim net growth rate (\( G \)). This metric is the proportion of growth in the expected cumulative claim amount obtained from the initial to the last sampling fraction under a particular sampling approach (SA). \( G \) is calculated as follows:

\[ G = \frac{E[CC|\text{Max}(f^t), SA] - E[CC|\text{Min}(f^t), SA]}{E[f|\text{Max}(f^t), SA] - E[f|\text{Min}(f^t), SA]} \] (9)

b) The expected claim surplus \( E[\Delta CC|SA, f] \) under a sampling approach SA and expected sampling fraction \( f \). This represents the expected additional claim amount that can be attributed to the use of a particular sampling approach for a given expected sampling fraction. It can be computed as follows:

\[ E[\Delta CC|SA, f] = E[CC|SA, f] - \text{Min}_{SA,f}(E[CC]) \] (10)

5) The size of the sampling window length is set to 8 quarters. In this sense, the first 8 quarters are called “carryover growth quarters” because during this period the number of quarters carrying-over payment information increases. The quarters beginning at the quarter number 9 and ending in the last claiming quarter are called “carryover maturity quarters” because during this period the number of quarters carrying-over information stays approximately constant.

The empirical experiments employ quarterly payment information from a real-life research dataset of social welfare claims. It should be noted that the research dataset has been partially transformed to safeguard its confidentiality and anonymity. This analysis requires the summation, on a quarterly basis, of all the payments associated with a recipient considering the actual payment transaction date.

Given that recipient payment information came from various sources, the stratification variable is chosen to be the quarterly primary payment information source. This variable contains the source of payment that provides the greatest amount to the recipient in a particular quarter. For example, as illustrated in Table II, the recipient \( j \) shows different primary payment sources in each of the observed quarters: in quarter \( i \), the primary payment source is \#1; in quarter \( i + 1 \), the primary payment source is \#3; and in quarter \( i + 2 \), the primary payment source is \#2.

As presented in Table III, in spite of the fact that each stratum shows some similarity in mean and median values, the data also contain a large number of outliers. This situation causes high variability in each stratum and degrades the potential improvement in the reliability of the sampling estimates that could be obtained using stratification techniques. Consequently, the degree of improvement from using stratified sampling techniques is not expected to be large.

The summarized experimental findings are as follows.

1) As presented in Table IV, progressive sampling approaches (PRS and SPRS) yield significantly different expected cumulative claim amounts as compared to simple and stratified random sampling approaches (RS and SRS).
TABLE III
DESCRIPTIVE PAYMENT INFORMATION PER STRATUM IN RESEARCH DATASET EXPERIMENTS

| Box-Plot for Amounts Paid per Quarter to Recipients |

<table>
<thead>
<tr>
<th>Descriptive Statistic</th>
<th>Stratum #1</th>
<th>Stratum #2</th>
<th>Stratum #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3,840.44</td>
<td>1,589.67</td>
<td>3,289.47</td>
</tr>
<tr>
<td>Median</td>
<td>3,519.04</td>
<td>1,479.36</td>
<td>3,224.13</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>2,202.67</td>
<td>1,304.97</td>
<td>1,565.41</td>
</tr>
<tr>
<td>Minimum</td>
<td>35.67</td>
<td>0.60</td>
<td>113.16</td>
</tr>
<tr>
<td>Maximum</td>
<td>75,823.93</td>
<td>94,017.00</td>
<td>10,434.06</td>
</tr>
<tr>
<td>Number of observations</td>
<td>10,354</td>
<td>24,447</td>
<td>591</td>
</tr>
</tbody>
</table>

TABLE IV
KRUSKAL–WALLIS RANK SUM TEST IN MULTIPERIOD SAMPLING EXPERIMENTS USING THE RESEARCH DATASET

<table>
<thead>
<tr>
<th>Type of Sampling Technique</th>
<th>Chi-Square Value</th>
<th>Degrees of Freedom</th>
<th>Significant (at α = 0.05)?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple (RS and PRS)</td>
<td>272.1042</td>
<td>1</td>
<td>Yes</td>
</tr>
<tr>
<td>Stratified (SRS and SPRS)</td>
<td>317.6916</td>
<td>1</td>
<td>Yes</td>
</tr>
<tr>
<td>Overall</td>
<td>631.2037</td>
<td>3</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note: the response variable was the cumulative claim amount.

TABLE V
COMPARING EXPECTED CLAIM NET GROWTH RATES IN RESEARCH DATASET EXPERIMENTS

<table>
<thead>
<tr>
<th>Type of Sampling Schema</th>
<th>Sampling Technique</th>
<th>Expected Claim Net Growth Rate (G, in US$ Millions per percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>RS</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>PRS</td>
<td>0.80</td>
</tr>
<tr>
<td>Stratified</td>
<td>SRS</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>SPRS</td>
<td>0.62</td>
</tr>
</tbody>
</table>

* Notation in Exhibit 1.

2) The net growth rate associated with the expected cumulative claim is higher under progressive sampling conditions. That is, as presented in Table V, PRS yields a higher $G$ than RS, and SPRS yields a higher $G$ than SRS. This implies that progressive sampling approaches use random samples more efficiently than the nonprogressive sampling approaches.

3) As observed in Fig. 1, progressive sampling approaches improve the reliability of multiperiod sampling estimates as compared to simple and stratified random sampling approaches (RS and SRS). Also, due to the poor performance
of the employed stratification variable, even PRS outperforms SRS. However, SPRS yields more reliable sampling estimates than PRS at any of the employed sampling fractions.

This improvement in the reliability of the sampling estimates is evident by the existence of larger claim amounts under the progressive sampling schemes. As an example, with expected sampling fraction of a 0.10.

1) PRS is expected to collect $4.4 million (or 6%) more than RS and $3.1 million (4%) more than SRS.

2) SPRS is expected to collect $5.2 million (or 7%) more than RS and $3.9 million more (or 5%) than SRS.

As observed in Fig. 1, progressive sampling approaches reduce the sampling effort at each claim collection level. Some examples are:

1) To obtain the same expected collection amount as PRS with an expected sampling fraction of a 0.05, RS requires an expected sampling fraction of about 0.10 and SRS requires an expected sampling fraction of about 0.07.

2) Similarly, in order to obtain the same expected collection amount as SPRS with an expected sampling fraction of a 0.05, RS requires an expected sampling fraction of about 0.12 and SRS requires an expected sampling fraction of about 0.15.

By comparing the aforementioned examples, it is possible to observe that SPRS makes a more efficient use of the sampling effort as compared to PRS.

1) As presented in Fig. 2, the carryover proportion resulting from the application of progressive random sampling approaches increases as the target-sampling fraction increases. Furthermore, as observed in Fig. 3, the expected carryover proportion reaches its limit in the last quarter of the carryover growth quarters, and it stays relatively constant during the “carryover maturity quarters.” This is explained by the following facts: a) during the carryover growth quarters, all the population members carrying-over information enter the progressive estimation process and stay active during the sampling window length; and b) during the carryover maturity quarters, the carry-over component of the population is progressively updated.

2) In any claiming period, where the summation of the expected carry-over proportion and the target sampling fraction is equal to or greater than 1, the expected results are equivalent to conducting a census. In these cases, the sampling variation is zero and all the expenditures in payments to social welfare recipients are claimed. This is observed under PRS and SPRS with target sampling proportion greater than 0.10.

V. CONCLUSION

As demonstrated in Section IV and using quarterly payment information from a transformation of a real-world dataset of social welfare claims, the progressive sampling approaches can improve the reliability of the sampling estimates when compared to traditional approaches. As a consequence, the proposed sampling methods can be employed either for increasing the precision of multiperiod sampling estimates or for reducing sample size requirements in retrospective sampling conditions.

On the other hand, as outlined by the environmental protection agency [22], stratified random sampling methods require reliable a priori knowledge of the population in order to define the strata and allocate the sample sizes, including a prior knowledge of: 1) the strength of the relationship between the stratification variable(s) and the variable of interest and 2) knowledge of the variability within the strata.

For the case of the stratified progressive random sampling (SPRS) approach, the nature (or strength) of the relationship between the stratification variable(s) and the variable of interest could change over time. In addition, the expected gain from using this approach is directly associated with the amount of information that the population members carry over in subsequent sampling periods. Although the potential impact of these two factors should be considered in future research of the SPRS approach, it is expected that the proposed methodology can yield significant improvements in the reliability of the sampling estimates (as demonstrated in Section IV with a real-life application).
Future work for extending the SPRS method could focus on:
1) making it adaptive and (potentially) robust to changes in the stratification and information carryover patterns over time;
2) integrating the progressive sampling concepts into more complex sampling designs, including cluster and multiple stage sampling; and
3) combining progressive sampling techniques with sampling-related resource allocation decisions, including the development of methods for sequencing multiple sampling projects and for the efficient management of sampling resources.

REFERENCES

[21] United States Code, False Claims, Title no. 31, Subtitle 3, Ch. 37, Subch. 3, Section 3729.

Plinio A. De los Santos, Jr., received the B.S. I.E. degree in industrial engineering from the Instituto Tecnológico de Santo Domingo, Santo Domingo, Dominican Republic, in 1986, the Master in Business Administration (MBA) degree from the University of Puerto Rico, Rio Piedras, in 1992, and the M.S.I.M.E. degree in industrial management engineering, and the Ph.D. degree in decision sciences and engineering systems from the Rensselaer Polytechnic Institute, Troy, NY, in 1996 and 1998, respectively.
He has experience in the development and application of statistical methods in pharmaceutical, social services, and software development environments.
Dr. De los Santos is a member of the IEEE Computer Society and the American Statistical Association.

Richard J. Burke received the B.Eng. degree in naval architecture from Maritime College of the State University of New York, in 1972, the S.M. degree in naval architecture and marine engineering from the Massachusetts Institute of Technology, Cambridge, in 1974, and the Ph.D. degree in industrial engineering and operations research from the University of Massachusetts, Amherst, in 1990.
He was a Statistician with the Knolls Atomic Power Laboratory, Lockheed-Martin, New York. He was also with the University of Massachusetts, Amherst and the Rensselaer Polytechnic Institute, Troy, NY, and is currently a Professor and Chairman of Engineering at Maritime College. His current research interests include applying statistical techniques in industrial settings, especially the designs of effective studies and experiments.

James M. Tien (M’78–SM’82–F’91) received the B.E.E. degree in electrical engineering from Rensselaer Polytechnic Institute, Troy, NY, in 1966, and the S.M.E.E. degree in electrical and systems engineering and the Ph.D. degree in systems engineering and operations research from the Massachusetts Institute of Technology, Cambridge, in 1967 and 1972, respectively.
He was with Bell Telephone Laboratories, the Rand Corporation, and the Structured Decisions Corporation. In 1977, he joined the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, where he has been the Acting Chair, the Founding Chair in the Department of Decision Sciences and Engineering Systems, and the Acting Dean of Engineering. His current research interests include the development and application of computer and systems analysis techniques to information and decision systems. He is the author or coauthor of several publications.
Dr. Tien is a Fellow of the Institute for Operations Research and the Management Sciences (INFORMS) and the American Association for the Advancement of Science. He is also an elected member of the U.S. National Academy of Engineering. He is the recipient of the IEEE Joseph G. Wohl Outstanding Career Award, the IEEE Major Educational Innovation Award, the IBM Norbert Wiener Award, and the IBM Faculty Award.
Progressive Random Sampling With Stratification

Plinio A. De los Santos, Jr., Richard J. Burke, and James M. Tien, Fellow, IEEE

Abstract—A number of applications, including claims made under federal social welfare programs, requires retrospective sampling over multiple time periods. A common characteristic of such samples is that population members could appear in multiple time periods. When this occurs, and when the marginal cost of obtaining multiperiod information is minimum for a member appearing in the sample of the period being actively sampled, the progressive random sampling (PRS) method developed by the authors earlier can be applied. This paper enhances the progressive random sampling method by combining it with stratification schemes; the resultant stratified progressive random sampling (SPRS) technique is shown to provide significant improvement over traditional sampling techniques whenever stratification is appropriate. An empirical example based on a data transformation of a real-world application is provided to illustrate the practical application of the technique.

Index Terms—False Claims Act, multiperiod sampling, retrospective sampling, social welfare data estimation methods, stratified sampling.

I. INTRODUCTION

The drawing of periodic retrospective samples is required for many application areas, including expenditure estimation of federally funded welfare programs. Under such an application, a state makes a claim under one of a variety of federally funded welfare programs, but the claim is actually filed retrospectively for the fiscal quarter up to two years prior to the current quarter. Thus, the “claiming period” could be up to eight quarters prior to the current quarter. The actual amount that could be claimed under the specific program could only be determined by an examination of the case file for each recipient and totaling all the eligible expenditures. Given the enormity of the task, the state may wish to present a claim based on a “statistically valid” sample drawn from the population of eligible recipients who received payments in that period.

Sample-based claim values are ordinarily taken to be the lower 90% confidence bound on the estimate of the total claim amount. Sample-based claims must be obtained with considerable care, however, because all claims are subject to audit, and substantial penalties may be applied under the federal False Claims Act [21].

While one could easily construct claims by employing simple random sampling (RS) for each claiming period, the resources required could represent a substantial demand on those responsible for filing such claims. After the sample has been drawn, it may be necessary to physically review case files to determine the value of the eligible claims. As a practical matter, one should observe that in many cases the sample files could also yield claims for fiscal quarters subsequent to the claiming period of interest, and that all of the relevant information for those subsequent periods is obtainable for essentially no additional effort while the review of the files is underway.

This scenario suggests that more effective sampling approaches could be developed, in the sense that either the lower 90% confidence bound could be improved relative to the best estimate of the total or that the level of sampling effort could be reduced without loss to claim recoveries. Such an approach should make use of sample information progressively, saving the information concerning subsequent claiming periods, regardless of which period is being actively sampled.

This paper significantly extends the progressive random sampling approach or (PRS) [3] by developing the stratified progressive random sampling (SPRS) technique for sampling applications in which it is convenient to perform stratified random sampling (SRS). Cochran [2, pp. 87–88] lists some of these situations:

1) when required to specify precision levels for population subgroups;
2) when administrative restrictions demand the use of stratification;
3) when data collection approaches differ considerably for distinct portions of the population;
4) when stratification allows division of a heterogeneous population into subpopulations that are internally homogeneous.

As in the case of PRS, SPRS takes advantage of the fact that some information about subsequent sampling periods is known from previous samples.

This paper is organized into several sections. Section II provides a literature review. Section III presents the SPRS methodology and presents PRS as a special case of SPRS. Section IV provides empirical results obtained using a research dataset (which is based on a data transformation of a real-world application). Section V contains some concluding remarks.

II. LITERATURE REVIEW

Sampling techniques for periodic retrospective estimation problems do not appear in the literature of statistical practice [5]. In addition, traditional sampling techniques directed at obtaining measurements over time do not take into account stratified sampling considerations. As outlined by Kish [11], stratified sampling consists of the following steps:

1) dividing the entire population of sampling units into distinct subpopulations (called strata);
2) selecting a separate sample from the sampling units falling within each stratum;
3) computing a statistic (such as the mean or the total) for each stratum, as well as a combined statistic of the entire population (by appropriately weighting the estimates from each stratum);
4) computing the variation estimate for each stratum, as well as a combined variation estimate of the entire population (by weighting the variability estimates from each stratum).

The literature on sampling designs directed at obtaining measurements over time [1]. [4] includes the following related techniques.

1) Repeated surveys. This design requires obtaining data from samples periodically, but does not seek to include a sampling item in more than one period. The main purpose of this design is for deriving elementary estimates for each analysis period [7]; it uses conventional survey sampling methods.

2) Panel survey (also known as longitudinal studies). In this design, data are obtained from the same sample in different periods. One form of panel surveys, known as cohorts, studies population subgroups that share experiences during the same time period.

3) Rotating panel surveys (also known as repeated surveys with partial overlap). Eckler [6, p. 665] defines the concept of “rotation sampling” as “the process of eliminating some of the old elements from the sample and adding new elements to the same every time a new sample is drawn.” Thus, in this design sample elements are included in the survey and then surveyed a predefined number of times, and subsequently rotated out of the survey.

4) Split panel survey. It combines a panel survey with a repeated or rotating panel survey [12].

The effectiveness of the sampling design is strongly dependent on the underlying survey objectives. For the purpose of this research, the main objective is to provide estimates of population parameters at distinct points in time. Duncan and Kalton [4] make a comparison of the multiperiod sampling designs under such an objective and conclude that the repeated survey is the only method that automatically takes population changes into account. Panel, rotating panel, and split panel designs require a mechanism for taking population changes into account (i.e., introduction of new elements to the population as time passes). However, rotating panel designs have been extensively used for estimation purposes because of its variance reduction properties. Panel surveys (i.e., longitudinal studies) are more appropriate when the survey objective is measuring individual change; its main use has been in the development of causal models.

The rotating panel literature assumes that a pattern of overlap of the sampling units, which are involved in a sequence of repeated surveys, can be identified. Then, given the existence of overlapping sampling units and for any particular survey design, it is possible to derive more reliable estimators using the theory of minimum variance and unbiased linear estimators. Of course, if the surveys do not overlap, the covariance matrix becomes diagonal and the estimate for the sampling period becomes independent from estimates obtained in previous periods.

The first attempt at employing overlapping survey information to improve estimates is due to Jessen [10]. He obtains a linear unbiased estimate of the population mean by combining information from sampling in two occasions. A good presentation of Jessen’s findings is provided in [8].

Patterson [17] extends Jessen’s result to estimate the population means for more than two occasions. This result is based on the assumption that the partial correlation coefficients between the same sampling units observed in different occasions decreases geometrically as the time interval between occasions increases. A detailed exposition of Patterson’s findings is presented in [13]. Also, applications of Patterson’s results are contained in [20].

Subsequent research efforts extended Patterson’s result. Eckler [6] points out that Patterson’s result is based on a “one level rotation sampling,” and proposes a generalized sample pattern for the estimation of a linear function of the population means. Gurney and Daly [7] develop a composite estimation method to obtain the weights of the best linear unbiased estimators considering a general correlation structure. Singh [18] proposes a two-stage (nested) multiperiod sampling based on two specific rotation patterns. Huang and Ernst [9] study the performance of Gurney and Daly’s estimation method [7] in the current population survey.

There are at least three drawbacks in the application of rotating panels. First, the improvement in the estimators is highly dependent upon the correlation structure. Second, these methods require prespecifying the population elements’ rotational patterns. Third, the administration of these methods could be quite complicated, especially when dealing with complex rotational patterns.

A nontraditional method for dealing with the multiperiod survey problem is based on time series models. The main advantage of this approach is that it allows using nonoverlapping survey data in undertaking parameter estimation, and it yields models that can be employed in prediction. However, the procedure requires that the population parameters at each time period follow a stochastic underlying phenomenon. A comprehensive discussion of these techniques as applied to the problem of repeated sample surveys is provided in [19]. The key limitation of the available time series survey methods is that their application depends on the stationarity conditions of the data. Furthermore, time series modeling may only be undertaken by specially trained personnel.

In sum, the available literature underscores the need to develop cost-effective sampling techniques that improve retrospective multiperiod estimations without depending upon either complex correlation structures or special data assumptions, and within the multiperiod sampling context that do not consider stratified sampling schemas. Indeed, an earlier work [3] and this current extension of that work meet the needs of a cost-effective, multiperiod sampling technique.

III. METHODOLOGY

Utilizing the notation described in Table I and taking into account the scenario described in the Section I, assume that it is necessary to obtain retrospective estimates of the total expenditures in payments to social welfare recipients. Specifically, consider that a member \( j \) of a population of size \( N_{i,h} \) in a
TABLE I
NOTATION ASSOCIATED WITH MULTIPERIOD SAMPLING TECHNIQUES

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>Significance level (a.k.a. Type I error probability).</td>
</tr>
<tr>
<td>B</td>
<td>Bernoulli carryover function.</td>
</tr>
<tr>
<td>( b_{i}(n_{0}, h) )</td>
<td>Bernoulli probability of carrying-over payment information from period ( i ) and stratum ( h ) to period ( i+1 ) and stratum ( h ).</td>
</tr>
<tr>
<td>CC</td>
<td>Cumulative claim after obtaining sampling estimates for ( q ) periods.</td>
</tr>
<tr>
<td>( d_{i} )</td>
<td>Precision estimate for the total claim amount in period ( i ) under PRS or SPRS.</td>
</tr>
<tr>
<td>( d_{i}^{*} )</td>
<td>Precision estimate for the total claim amount in period ( i ) under RS or SRS.</td>
</tr>
<tr>
<td>( E[] )</td>
<td>Expectation operator, which indicates the expected value of a variable within the brackets.</td>
</tr>
<tr>
<td>( f_{i,h}^{*} )</td>
<td>Sampling fraction under SPRS in period ( i ) and stratum ( h ).</td>
</tr>
<tr>
<td>( f_{i,h} )</td>
<td>Sampling fraction under SRS in period ( i ) and stratum ( h ).</td>
</tr>
<tr>
<td>( f' )</td>
<td>Target fraction per sampling period in application experiments.</td>
</tr>
<tr>
<td>G</td>
<td>Expected claim net growth rate.</td>
</tr>
<tr>
<td>( \text{Int}[] )</td>
<td>Operator that obtains the integer value of the number within the brackets.</td>
</tr>
<tr>
<td>( L_{i} )</td>
<td>Total number of strata in period ( i ).</td>
</tr>
<tr>
<td>( U_{H} )</td>
<td>Lower (1-( \alpha ))100% confidence bound for the total claim amount in period ( i ).</td>
</tr>
<tr>
<td>( \text{Max}[] )</td>
<td>Operator that obtains the maximum of the values within the brackets.</td>
</tr>
<tr>
<td>( m_{i,h} )</td>
<td>Number of carried-over population members in period ( i ) belonging to the set ( Y_{i,h} ).</td>
</tr>
<tr>
<td>( \text{Min}[] )</td>
<td>Operator that obtains the minimum of the values within the brackets.</td>
</tr>
<tr>
<td>( N_{i,h} )</td>
<td>Population size in a period ( i ) and stratum ( h ). Consequently, the total size of the population ( N_{i} ) is the summation of the population size in each stratum (i.e., ( N_{i} = \sum_{h} N_{i,h} )).</td>
</tr>
<tr>
<td>( n_{i,h} )</td>
<td>Size of the random sample drawn from the population associated with stratum ( h ) in period ( i ) under SPRS. As expected, ( n_{i} = \sum_{h} n_{i,h} ).</td>
</tr>
<tr>
<td>( n_{i,h}^{\ast} )</td>
<td>Size of the random sample drawn from the population associated with stratum ( h ) in period ( i ) under SRS. As expected, ( n_{i}^{\ast} = \sum_{h} n_{i,h}^{\ast} ).</td>
</tr>
<tr>
<td>( n_{i}^{\ast} )</td>
<td>Target sample size for period ( i ) in application experiments.</td>
</tr>
<tr>
<td>( p_{i,h} )</td>
<td>Carryover proportion associated with stratum ( h ) in period ( i ).</td>
</tr>
<tr>
<td>PRS</td>
<td>Progressive Random Sampling.</td>
</tr>
<tr>
<td>SPRS</td>
<td>Stratified Progressive Random Sampling.</td>
</tr>
<tr>
<td>SRS</td>
<td>Stratified Random Sampling.</td>
</tr>
<tr>
<td>( s_{i} )</td>
<td>Standard error for the total claim amount in period ( i ).</td>
</tr>
<tr>
<td>( s_{i,h} )</td>
<td>Standard error estimate for the claim amount per individual in period ( i ) and stratum ( h ).</td>
</tr>
<tr>
<td>( t_{0.2} )</td>
<td>( t ) Student distribution (1-( \alpha/2 ))100% percentile.</td>
</tr>
<tr>
<td>( T_{i} )</td>
<td>Total claim amount in period ( i ).</td>
</tr>
<tr>
<td>( V_{i,j,h} )</td>
<td>Estimated variance per individual in period ( i ) and stratum ( h ).</td>
</tr>
<tr>
<td>( w_{i,h} )</td>
<td>Sum of the claim amounts for the population members in period ( i ) belonging to the set ( Y_{i,h} ).</td>
</tr>
<tr>
<td>( X_{i,j,h} )</td>
<td>Claim amount in period ( i ) for population member ( j ) in stratum ( h ).</td>
</tr>
<tr>
<td>( \bar{x}_{i,h} )</td>
<td>Average claim amount per individual in period ( i ) and in stratum ( h ).</td>
</tr>
<tr>
<td>( Y_{i,h} )</td>
<td>Set containing only those population members in period ( i ) and stratum ( h ) whose claim amount are known from earlier sampling periods.</td>
</tr>
</tbody>
</table>

Note: In the case of RS and PRS techniques, the stratum sub-index \( h \) is not necessary because under those techniques the total number of strata equals 1. As an example, the sampling fraction under PRS is \( f_{i} \) instead of \( f_{i,h} \).

Specific period \( i \) and stratum \( h \) has some claim amount \( X_{i,j,h} \), with variance \( \sigma^{2}_{i,j,h} \) and mean \( \bar{x}_{i,j,h} \). The purpose of drawing a sample of size \( n_{i,h} \) for claiming period \( i \) and for each stratum \( h \) is to estimate the total of the \( x_{i,j,h} \) values over all population members \( j \) contained in all strata \( h \). Member \( j \) may also be present in subsequent periods, and will have a unique label for identification purposes.

Since the sampling is done retrospectively, the period from which the sample is drawn (i.e., the claiming period) is always a number of periods in the past relative to the period when
the sample is physically gathered (i.e., the sampling period). As a consequence, there is a collection of periods for which information may be obtained while sampling the claiming period. This aggregation of periods is called the sample window, and it begins in the claiming period and ends at the sampling period. To illustrate these concepts, consider the following example: a sample containing information from period $i$ is physically gathered in period $i + k$; under this example, the claiming period is $i$, the sampling period is $i + k$ and the length of the sampling window is $k + 1$.

Suppose that the claim amounts of member $j$ pertaining to subsequent periods within the sample window (if member $j$ is present in the populations of those periods) can be obtained without significant additional expenditure of effort. Then, for any specific claiming period $i$, it is possible to identify some members from a population $N_{i,h}$ whose $x_{i,j,h}$ values were obtained from previous samples of earlier sampling periods. The members present in period $i$ that are known from earlier sampling periods are members of the set $Y_{i,h}$. This set is excluded from the population that is sampled in period $i$ and stratum $h$. The sum of the claim amounts from these “known” members in period $i$ and stratum $h$ is represented by

$$w_{i,h} = \sum_{j \in Y_{i,h}} x_{i,j,h}$$

The typical length of a claiming period is a fiscal quarter, i.e., three months, but the period could be any convenient time interval. Given a specific significance level $\alpha$, a claim for the period $i$ is usually based upon the lower $(1 - \alpha)100\%$ confidence bound $LT_i$ of the total claim amount $T_i$ for that claiming period.

The fundamental concept behind PRS methods is the use of information obtained from previous samples to complement sample data obtained for the current claiming period. This concept requires that members of a population for which information is already on hand from previous samples are to be removed from the population prior to sampling for the current claiming period, since such information is known and is not subject to sampling error in the current claiming period. The statistical estimates derived for the current claiming period are therefore obtained from a current period sampling component that is subject to sampling variation, together with previous periods’ components that are not subject to sampling variation in the current period.

Consider the following example. Suppose that it is required to estimate the total claim amount for period $i$ after already drawing samples in periods $i - 1, i - 2, \ldots, i - k$. In this case and for each stratum $h$ in $i$, the set of elements $Y_{i,h}$ is going to be composed of $m_{i,h}$ members of the population $N_{i,h}$ whose claim amounts were obtained in the samples drawn in periods $i - 1$ up to $i - k$. Then, under progressive random sampling (PRS), the estimate of the total amount for period $i$, $\hat{T}_i$, is composed of two terms:

1) the expected value of the total claim amount obtained from the random sample taken from the adjusted populations in period $i$ and each stratum $h$;
2) the sum of claim amounts of the population elements contained in all $Y_{i,h}$ sets.

More specifically, under stratified progressive random sampling (SPRS) the sampling estimate for the total $\hat{T}_i$ can be represented by the following:

$$\hat{T}_i = \sum_{h=1}^{L_i} \{ (N_{i,h} - m_{i,h}x_{i,h}) \} + \sum_{h=1}^{L_i} w_{i,h}$$

where the first term of (2) is the estimate based on the sample from the reduced population in claiming period $i$, and the second term is the previously known value for the population members belonging to the $Y_{i,h}$ sets. Also, under SPRS, the sampling estimate of the standard error of the total claim amount is the standard error of the first term of (2), i.e.,

$$S(T_i) = \sum_{h=1}^{L_i} \sqrt{V[x_{i,h}]} \left( \frac{(N_{i,h} - m_{i,h} - n_{i,h})(N_{i,h} - m_{i,h})}{n_{i,h}} \right)$$

The following remarks should be made concerning the above two equations.

1) The length of the sample window need not be fixed, and may be varied by the decision maker over time. As an example, samples may be made for financial audits on a monthly, quarterly, or annual basis.
2) All of the samples are taken without replacement.
3) Confidence bounds on the total amount are based upon the standard error derived from the random sample drawn for the current claiming period alone.
4) PRS is a special case of SPRS that occurs when the population is not stratified or when $L_i$ equals 1 for all $i$.
5) Simple RS is a special case of PRS that occurs when $m_{i,h}$ equals zero (i.e., when no information from population members is carried over from previous sampling periods).
6) Similarly, SRS is a special case of SPRS that occurs when $m_{i,h}$ equals zero for all $h$ in period $i$.

The PRS techniques can improve RS or SRS results in either of two ways: by demanding smaller sample sizes for confidence intervals of a given width, or by narrowing confidence intervals for equal sample sizes. In the next section, each of these areas of improvement is empirically illustrated.

IV. EMPIRICAL RESULTS

In this section, the empirical results from comparing the following multiperiod sampling methods are summarized for:

1) Sample random sampling (RS);
2) Stratified sample random sampling (SRS);
3) Progressive sample random sampling (PRS);
4) Stratified progressive sample random sampling (SPRS).

To assess the relative effectiveness of these four sampling methods, at different sampling fraction levels, a cumulative claim (CC) performance metric is employed. Such a metric assumes that the total claim is the summation of the lower $(1 - \alpha)100\%$ confidence bound for the total claim amount in each claiming period truncated at zero. For each empirical setting, this metric is calculated as follows:

$$CC = \sum_{i=1}^{q} \text{Max} [\hat{T}_i - t_{\alpha/2}S(T_i), 0]$$

(4)
When employing stratified sampling schemas (i.e., SRS or SPRS), the samples are allocated to each stratum using Newman’s optimal allocation of samples as follows:

$$n_{i,h} = n_i \left( \frac{N_{i,h}S_{i,h}}{\sum_{h=1}^{L_i} N_{i,h}S_{i,h}} \right)$$  \hspace{1cm} (5)

When considering (5), this assumes that the unit sampling cost in each stratum in the period \(i\) is the same.

Some important information associated with the empirical experiments is:
1) the payment time aggregation unit employed is quarters (i.e., three payment months);
2) each experimental setting was replicated 100 times;
3) experiments are conducted using four levels of target sampling fraction (i.e., \(f^t = 0.05, 0.10, 0.15, \text{and } 0.20\)). Such information is necessary to compute a target sample size to draw from in any claiming period \(i\), as follows:

$$n_i^t = \begin{cases} \text{Int}(f^t N_i), & \text{if } \{f^t N_i - \text{Int}(f^t N_i)\} = 0 \\ \text{Int}(f^t N_i) + 1, & \text{otherwise} \end{cases}$$  \hspace{1cm} (6)

However, the actual sample size \(n_i\) is bounded by the size of the population subject to random variation. This implies that

$$n_i = \min \left[ n_i^t, N_i - \sum_{h=1}^{L_i} m_{i,h} \right]$$  \hspace{1cm} (7)

As a consequence, under progressive sampling schemas, the actual sample size is smaller than the target in quarters when the following condition occurs

$$f^t + \sum_{h=1}^{L_i} m_{i,h} > 1$$  \hspace{1cm} (8)

4) Experimental results are compared in terms of:
   a) The expected claim net growth rate (\(G\)). This metric is the proportion of growth in the expected cumulative claim amount obtained from the initial to the last sampling fraction under a particular sampling approach (SA). \(G\) is calculated as follows:

$$G = \frac{E[\Delta \text{CC}|\text{SA}, f] - E[\text{CC}|\text{SA}, f]}{E[f|\Delta \text{CC}|\text{SA}, f] - E[f|\text{SA}, f]}$$  \hspace{1cm} (9)

   b) The expected claim surplus \((E[\Delta \text{CC}|\text{SA}, f])\) under a sampling approach SA and expected sampling fraction \(f\). This represents the expected additional claim amount that can be attributed to the use of a particular sampling approach for a given expected sampling fraction. It can be computed as follows:

$$E[\Delta \text{CC}|\text{SA}, f] = E[\text{CC}|\text{SA}, f] - \min_{\text{All } \text{SA}, f} E[\text{CC}]$$  \hspace{1cm} (10)

5) The size of the sampling window length is set to 8 quarters. In this sense, the first 8 quarters are called “carryover growth quarters” because during this period the number of quarters carrying-over payment information increases. The quarters beginning at the quarter number 9 and ending in the last claiming quarter are called “carryover maturity quarters” because during this period the number of quarters carrying-over information stays approximately constant.

The empirical experiments employ quarterly payment information from a real-life research dataset of social welfare claims. It should be noted that the research dataset has been partially transformed to safeguard its confidentiality and anonymity. This analysis requires the summation, on a quarterly basis, of all the payments associated with a recipient considering the actual payment transaction date.

Given that recipient payment information came from various sources, the stratification variable is chosen to be the quarterly primary payment source. This variable contains the source of payment that provides the greatest amount to the recipient in a particular quarter. For example, as illustrated in Table II, the recipient \(j\) shows different primary payment sources in each of the observed quarters: in quarter \(i\), the primary payment source is #1; in quarter \(i + 1\), the primary payment source is #3; and in quarter \(i + 2\), the primary payment source is #2.

As presented in Table III, in spite of the fact that each stratum shows some similarity in mean and median values, the data also contain a large number of outliers. This situation causes high variability in each stratum and degrades the potential improvement in the reliability of the sampling estimates that could be obtained using stratification techniques. Consequently, the degree of improvement from using stratified sampling techniques is not expected to be large.

The summarized experimental findings are as follows.

1) As presented in Table IV, progressive sampling approaches (PRS and SPRS) yield significantly different expected cumulative claim amounts as compared to simple and stratified random sampling approaches (RS and SRS).
2) The net growth rate associated with the expected cumulative claim is higher under progressive sampling conditions. That is, as presented in Table V, PRS yields a higher $G$ than RS, and SPRS yields a higher $G$ than SRS. This implies that progressive sampling approaches use random samples more efficiently than the nonprogressive sampling approaches.

3) As observed in Fig. 1, progressive sampling approaches improve the reliability of multiperiod sampling estimates as compared to simple and stratified random sampling approaches (RS and SRS). Also, due to the poor performance...
of the employed stratification variable, even PRS outperforms SRS. However, SPRS yields more reliable sampling estimates than PRS at any of the employed sampling fractions. This improvement in the reliability of the sampling estimates is evident by the existence of larger claim amounts under the progressive sampling schemes. As an example, with expected sampling fraction of a 0.10:

1) PRS is expected to collect $4.4 million (or 6%) more than RS and $3.1 million (4%) more than SRS.
2) SPRS is expected to collect $5.2 million (or 7%) more than RS and $3.9 million more (5%) than SRS.

As observed in Fig. 1, progressive sampling approaches reduce the sampling effort at each claim collection level. Some examples are:

1) To obtain the same expected collection amount as PRS with an expected sampling fraction of a 0.05, RS requires an expected sampling fraction of about 0.10 and SRS requires an expected sampling fraction of about 0.07.
2) Similarly, in order to obtain the same expected collection amount as SPRS with an expected sampling fraction of a 0.05, RS requires an expected sampling fraction of about 0.12 and SRS requires an expected sampling fraction of about 0.15.

By comparing the aforementioned examples, it is possible to observe that SPRS makes a more efficient use of the sampling effort as compared to PRS.

1) As presented in Fig. 2, the carryover proportion resulting from the application of progressive random sampling approaches increases as the target-sampling fraction increases. Furthermore, as observed in Fig. 3, the expected carryover proportion reaches its limit in the last quarter of the carryover growth quarters, and it stays relatively constant during the “carryover maturity quarters.” This is explained by the following facts: a) during the carryover growth quarters, all the population members carrying-over information enter the progressive estimation process and stay active during the sampling window length; and b) during the carryover maturity quarters, the carry-over component of the population is progressively updated.

2) In any claiming period, where the summation of the expected carry-over proportion and the target sampling fraction is equal to or greater than 1, the expected results are equivalent to conducting a census. In these cases, the sampling variation is zero and all the expenditures in payments to social welfare recipients are claimed. This is observed under PRS and SPRS with target sampling proportion greater than 0.10.

V. CONCLUSION

As demonstrated in Section IV and using quarterly payment information from a transformation of a real-world dataset of social welfare claims, the progressive sampling approaches can improve the reliability of the sampling estimates when compared to traditional approaches. As a consequence, the proposed sampling methods can be employed either for increasing the precision of multiperiod sampling estimates or for reducing sample size requirements in retrospective sampling conditions.

On the other hand, as outlined by the environmental protection agency [22], stratified random sampling methods require reliable a priori knowledge of the population in order to define the strata and allocate the sample sizes, including a prior knowledge of: 1) the strength of the relationship between the stratification variable(s) and the variable of interest and 2) knowledge of the variability within the strata.

For the case of the stratified progressive random sampling (SPRS) approach, the nature (or strength) of the relationship between the stratification variable(s) and the variable of interest could change over time. In addition, the expected gain from using this approach is directly associated with the amount of information that the population members carry over in subsequent sampling periods. Although the potential impact of these two factors should be considered in future research of the SPRS approach, it is expected that the proposed methodology can yield significant improvements in the reliability of the sampling estimates (as demonstrated in Section IV with a real-life application).
Future work for extending the SPRS method could focus on:
1) making it adaptive and (potentially) robust to changes in the stratification and information carryover patterns over time;
2) integrating the progressive sampling concepts into more complex sampling designs, including cluster and multiple stage sampling; and
3) combining progressive sampling techniques with sampling-related resource allocation decisions, including the development of methods for sequencing multiple sampling projects and for the efficient management of sampling resources.

REFERENCES

[21] United States Code, False Claims, Title no. 31, Subtitle 3, Ch. 37, Subch. 3, Section 3729.

Plinio A. De los Santos, Jr., received the B.S. degree in industrial engineering from the Instituto Tecnológico de Santo Domingo, Santo Domingo, Dominican Republic, in 1986, the Master in Business Administration (MBA) degree from the University of Puerto Rico, Rio Piedras, in 1992, and the M.S.I.M.E. degree in industrial management engineering, and the Ph.D. degree in decision sciences and engineering systems from the Rensselaer Polytechnic Institute, Troy, NY, in 1996 and 1998, respectively.

He has experience in the development and application of statistical methods in pharmaceutical, social services, and software development environments. Dr. De los Santos is a member of the IEEE Computer Society and the American Statistical Association.

Richard J. Burke received the B.Eng. degree in naval architecture from Maritime College of the State University of New York, in 1972, the S.M. degree in naval architecture and marine engineering from the Massachusetts Institute of Technology, Cambridge, in 1974, and the Ph.D. degree in industrial engineering and operations research from the University of Massachusetts, Amherst, in 1990.

He was a Statistician with the Knolls Atomic Power Laboratory, Lockheed-Martin, New York. He was also with the University of Massachusetts, Amherst and the Rensselaer Polytechnic Institute, Troy, NY, and is currently a Professor and Chairman of Engineering at Maritime College. His current research interests include applying statistical techniques in industrial settings, especially the designs of effective studies and experiments.

James M. Tien (M’78–SM’82–F’91) received the B.E.E. degree in electrical engineering from Rensselaer Polytechnic Institute, Troy, NY, in 1966, and the S.M.E.E. degree in electrical and systems engineering and the Ph.D. degree in systems engineering and operations research from the Massachusetts Institute of Technology, Cambridge, in 1967 and 1972, respectively.

He was with Bell Telephone Laboratories, the Rand Corporation, and the Structured Decisions Corporation. In 1977, he joined the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, where he has been the Acting Chair, the Founding Chair in the Department of Decision Sciences and Engineering Systems, and the Acting Dean of Engineering. His current research interests include the development and application of computer and systems analysis techniques to information and decision systems. He is the author or coauthor of several publications. Dr. Tien is a Fellow of the Institute for Operations Research and the Management Sciences (INFORMS) and the American Association for the Advancement of Science. He is also an elected member of the U.S. National Academy of Engineering. He is the recipient of the IEEE Joseph G. Wohl Outstanding Career Award, the IEEE Major Educational Innovation Award, the IEEE Norbert Wiener Award, and the IBM Faculty Award.