SAS SAMPLE SELECTION MACROS
Rick L. Williams and James R. Chromy
Research Triangle Institute

1. Introduction

In the survey research environment, the need exists to routinely select probability samples from a wide variety of computer accessible frames. To this end, the authors have developed several SAS macros that allow relatively unsophisticated SAS users to draw complex samples. The current macros implement an algorithm recently developed by J. R. Chromy (1979) for sequentially selecting zoned samples with unequal probabilities proportional to size. The basic selection methodology and its properties are discussed in the second section of this paper, while the specific macros are described in the third section.

2. Methodology

2.1 Sequential Probability Proportional to Size Selection

The selection of a sample can be viewed as the process of determining the values of \( n(i) \), for \( i = 1, \ldots, N \), where \( n(i) \) is the number of times sampling unit-\( i \) is included in the sample and \( N \) is the total number of sampling units on the frame. If a without replacement sample of size \( n \) is required, then \( n(i) \) will equal one for \( n \) values of \( i \) and zero for all other values of \( i \). For with replacement sampling, each sampling unit can be included in the sample more than once; that is, each \( n(i) \) can assume integer values zero through \( n \). However, in both cases, \( \sum_{i=1}^{N} n(i) = n \), the total sample size.

By definition, a sample design is said to be probability proportional to size (PPS) when the expected number of times a sampling unit is included in the sample is proportional to its size measure. More precisely, the following holds:

\[
\pi(i) = \frac{nS(i)}{S(+)} \tag{1}
\]

where,

\[
S(i) = \text{size measure associated with sampling unit-}i, \quad \text{and}
\]

\[
S(+) = \sum_{j=1}^{N} S(i). \tag{2}
\]

Furthermore, a PPS design will be termed probability minimum replacement (PMR) if \( n(i) \) can assume at most two values:

1. the integer portion of \( nS(i)/S(+) \); or
2. the next larger integer.

The algorithm under consideration is PMR with

\[
\Pr(n(i) = x) = \begin{cases} 
1 - \text{Frac}[\pi(i)], & x = \text{Int}[\pi(i)]; \\
\text{Frac}[\pi(i)], & x = \text{Int}[\pi(i)] + 1; \\
0, & \text{otherwise};
\end{cases}
\]

where \( \text{Frac}[\pi(i)] \) and \( \text{Int}[\pi(i)] \) denote the fractional and integer parts of \( \pi(i) \), respectively.

The implemented selection procedure sequentially considers each unit on the ordered sampling frame and probabilistically determines the number of times each unit will be included in the sample. In its most rudimentary form, the selection of only one unit from an ordered list, the procedure successively tests units for membership in the sample by drawing uniform \((0,1)\) random numbers, \( r(i) \) say, until

\[
r(i) \leq \theta(i) = \pi(i)/[1 - \sum_{j=1}^{i-1} \pi(j)].
\]

The first unit to satisfy this relationship is drawn into the sample. Notice that when \( r(i) > \theta(i) \) for all \( i < N \), a selection is guaranteed for the \( N \)-th unit since, with \( n = 1 \), \( \pi(N) = 1 - \sum_{i=1}^{N-1} \pi(i) \), and, therefore, \( \theta(N) = 1 \). Since the probability that unit-\( i \) is selected by this process is

\[
\theta(i)/[1 - \theta(i)] = \theta(i)[1 - \sum_{j=1}^{i-1} \pi(j)] = \pi(i),
\]

the appropriate sample selection property is easily demonstrated.

To generalize the technique to \( n \) selections, consider the \( \pi(i) \ (i = 1, \ldots, N) \) quantities arrayed as ordered segments on a line of total length \( N \) (notice that \( \sum_{i} \pi(i) = n \)) and partitioned into \( n \) zones of length one (see Figure 1). The procedure makes exactly \( n \) selections, one from each of the \( n \) zones, although this may be associated with fewer than \( n \) distinct sampling units. The selections are made so that units for which \( n(i) > 1 \), often called self-representing units, will automatically represent \( \text{Int}[\pi(i)] \) zones with the possibility of representing one additional zone. For units totally contained within a zone, the sequential test for membership is equivalent to the \( n = 1 \) case demonstrated above. A unit split by a single zone boundary might be selected in either of the zones, but not both (unless
In order for the variances of sample survey statistics to be estimable, it is necessary that all possible pairs of units have a positive chance of appearing in the same sample. To insure this property, the just discussed linear listing of the sampling units can be treated as circular by linking the lead and terminal ends of the listing. If zones are then marked off around the circular listing using a randomly selected unit as the starting point, it is clear that unit-1 could represent zone-1 and unit-(i-1) could represent zone-n in the same sample. Since all N units have a chance of becoming the lead unit for zone-1, all adjacent pairs have a chance of appearing together. By a similar argument, nonadjacent pairs also have a chance of appearing in separate zones and, therefore, have an opportunity to appear together in the same sample.

2.2 Implicit Stratification

While it is possible for large units to represent more than one zone, it is important to note that a unit will be selected from each zone. Thus, the realized sample is implicitly stratified by the frame ordering. This property can be exploited to reduce the variances of survey statistics or to control the realized sample sizes for key domains or reporting groups. To understand how a domain sample size can be controlled, consider an ordering that places all domain members consecutively on the frame. The domain can then be conceptualized as one aggregate unit, with a size measure, $S_D$, say, equal to the sum of sizes of its constituent units. The actual number of selections from the domain will differ by less than two from the expected number of times the hypothetical domain unit would be included in the sample (i.e., $nS_D/\bar{g}(\cdot)$). This difference is due to the nature of the sampling process and to the fact that the domain is not actually treated as one unit by the algorithm. If control is desired on multiple domains, it will not always be possible to place the members of every domain consecutively on the frame; however, the frame may be ordered so that each domain is listed in a few disjoint subgroups. Appealing to the logic of the preceding argument, the representation of a domain listed on the frame in $k$ subgroups will differ by less than $2k$ from its expected representation.

As stated previously, the frame ordering can also be utilized to reduce the variances of the survey statistics. This is accomplished in a manner akin to stratified sampling. The variance of a statistic is reduced to the extent that units situated near each other in the ordering are more homogeneous than those in the population at large. This results from the fact that differences between units sufficiently far apart in the ordering do not enter into the variance. Thus, ordering the frame so that neighboring units are similar in the survey variables should yield a reduction in variance similar to that achieved by stratified sampling. The fact that the sample is implicitly stratified by the frame ordering does not preclude the possibility of also explicitly stratifying the sample. This is accomplished by first separating the population into disjoint subgroups and then applying the selection procedure independently within each subgroup. In this way, the benefits of optimal allocation and over-representation of small domains can be realized.
2.3 Hierarchic Serpentine Ordering

Recognizing the importance of how the frame is ordered prior to sample selection leads naturally to consideration of how to best order the frame. The ordering offered here will be termed a hierarchic serpentine ordering. This ordering is best described by way of an example. Assume that a sample of schools is desired and that it has been determined that percent minority enrollment is correlated with the survey variables of interest. Also, assume that urbanicity (urban, rural) and grade range (primary, junior high, and senior high) will form the key reporting domains. Consequently, it is desirable to order the frame so that schools similar on all three of these control factors (i.e., urbanicity, grade range and percent minority) are proximate before drawing the sample. In general, a hierarchic serpentine ordering lists the sampling units from "lowest" level to "highest" level and then from "highest" level to "lowest" level on a particular control factor as the factors above it in the hierarchy of factors change levels. Thus, in the example, schools are first ordered so that those located in urban areas are listed first on the frame, while those in rural areas are listed second. Then, the urban schools are reordered so that primary schools are listed first followed by junior high schools and then senior high schools; conversely, rural schools are reordered from senior high schools down to primary schools. Finally, schools are further reordered alternately from low percent minority to high percent minority, and then from high percent to low percent within the successive combined levels of urbanicity and grade range. This process is graphically displayed in Figure 2. The three horizontal graphs illustrate the distribution of school characteristics across the ordered population.

The advantage of a serpentine ordering arises at the boundaries between control factor levels (e.g., urban to rural). To see this, note that sampling units on the boundary between two levels of a particular control factor are similar on all other factors; whereas, under a conventional nested ordering, units on either side of a level boundary are dissimilar on all factors lower in the hierarchy. The difference between the two ordering schemes becomes readily apparent by comparing Figure 3 and Figure 2. Figure 3 illustrates the distribution of school characteristics across the population under a conventional nested ordering. Note the many discontinuities in percent minority enrollment displayed in Figure 3 as opposed to the smooth transitions in Figure 2.

A final point should be noted concerning the frame ordering. When the total number of categories defined by a complete cross-classification of the control factors exceeds the number of selections made, control is not obtained on the factors near the bottom of the hierarchy. This problem is analogous to trying to select only ten units from 100 strata. Thus, it can be seen that only the last factor in the control hierarchy may be continuous since no practical benefit will be derived by including additional factors after a continuous one.

3. Implementation

The sample selection methodology outlined in Section 2 is implemented via several SAS macros. The macros accept a SAS data set containing the sampling frame, with each observation corresponding to a sampling unit, process it and return a new SAS data set containing the selected sample members, while leaving the original sampling frame data set unchanged. The three main selection macros and their functions are:

1. _SERP—sorts the frame into a hierarchic serpentine ordering;
2. _HEAD—selects a random unit to be the head of the ordered frame; and
3. _PMR—performs the actual probability minimum replacement selection.

In addition, several other stored macros are also provided to drive the three main selection macros for standard applications.

Before the stored macros can be executed, they must be provided with certain information concerning the desired sample (e.g., the sample size, where the frame is, etc.). The user communicates this information to the stored selection macros by defining several additional macros which are then used by the stored macros to control the selection. The user is required to define the following macros:

1. _FRAME—name of the SAS data set containing the frame;
2. _OUTSAMP—name of the SAS data set where the sample is to be stored;
3. _ID—variable list that uniquely identifies each sampling unit; and
4. _SAMPN—name of the variable containing the sample size.

Additionally, several options may be specified via user defined macros. A list of these macros, their functions and default values follows:

1. _STRATA—variable list whose complete cross-classification specifies the explicit strata within which the selection procedure is independently applied. If not defined, the entire frame is considered to be one stratum.
Figure 2
Hierarchic Serpentine Ordering

Figure 3
Conventional Nested Ordering
2. **CONTROL**—list of variables by which the frame is to be ordered in a serpentine fashion within each stratum. No reordering is done if this macro is not defined.

3. **SIZE**—variable specifying the size measure associated with each sampling unit. Defaults to equally sized units.

4. **OTHER**—list of any additional variables to be carried over to the data set of sample members. Only those variables referenced in other user defined macros are carried over unless this macro is defined.

5. **SEEDHD**—a constant which serves as the random number seed for selecting the lead unit in a stratum. Defaults to zero.

6. **SEEDPMR**—another constant which serves as the random number seed for the actual sample selection. Also defaults to zero.

To select a sample, the user defines the necessary macros and then executes the appropriate driving macro.

4. **Conclusion**

These macros have been successfully used at the Research Triangle Institute for several months. They have proved very effective for selecting complex samples. At the same time, experience has demonstrated the utility of the recently developed probability minimum replacement selection scheme.

**REFERENCE**