LEGACY RDD SAMPLING METHODOLOGIES

The basic generation methods supported within GENESYS have been designed to provide the first and only set of tractable commercial RDD sampling procedures. The default RDD methodology provides single stage *epsem* samples of telephone numbers regardless of the defined sample frame. By their very nature, these samples are self-weighting in terms of residential number assignments within the set NPA/NXXs comprising the sample frame.

MOD1relates to modified RDD processes and is an extension of the *epsem* RDD process. These modified processes utilize user-defined Measure of Size (MOS) variables for each NPA/NXX.

The MOS variable provides the basis for directly varying sampling rates across NPA/NXXs. Although these processes are non-*epsem*, the defined MOS for each NPA/NXX is retained in the sample records for use in post-survey weighting adjustments.

Single Stage EPSEM RDD.

This option represents the default methodology supported by the GENESYS Sampling System. To the best of our knowledge it was the first such procedure available either commercially or even among the few proprietary sampling systems remaining in use.

By its very nature, this methodology provides a known and equal probability of selection to every possible telephone number in the defined NPA/NXX sample frame; the process produces a single stage sample, in that every possible number is implicitly included in the selection array for every sample draw. This implicit inclusion is, however, just a practical convenience which makes the generation process more efficient.

It is not necessary to explicitly show or list every possible telephone number within the defined frame to provide a single stage selection process. Rather, one needs to be able to uniquely relate the relative position within a theoretical sample frame, or selection interval, to one and only one telephone number in the actual sample frame. Any conceptual difficulties should be overcome as the procedure is detailed below.

Conceptualizing the Sample Frame. Prior to initiation of sample generation, the user has defined a sample frame comprised of a set of NPA/NXX combinations, K; has determined the total number of sample pieces required, n'; and, specified the number of replicates, R.

The operative sample frame is the variable number of residential working banks associated with each of the NPA/NXX combinations. The maximum number of telephone numbers, N^K, is equal to the number of two-digit residential banks, in the defined set of NPA/NXXs times 100 - since there are by definition 100 unique two-digit combinations, (i.e., XX00,XX01, ... XX99) in each working bank, or hundred series:

$$N^K = WB^K \times 100 = \sum_k (wb_k) \times 100 \text{ where,}$$

N^K = total possible 10-digit telephone numbers in the defined sample frame, k;

WB^K = total working banks in the defined sample frame, k;

wb_k = the number of residential working banks in the k th NPA/NXX; and,

k = the number of NPA/NXXs which comprise the defined sample frame, k.

The NPA/NXX combinations are in a specified geo-metro hierarchy, and this sequence is maintained in the file of NPA/NXXs:

$$K = K_1, K_2 \ldots K_k$$

Within each of these k NPA/NXXs, we have a specified variable number of working residential banks, wb_k:

$$WB^K = wb_{11}, wb_{12} \ldots wb_{1k}, wb_{21} \ldots wb_{km}$$

And finally, within each of the wb_{km} banks there are exactly 100 two digit suffixes:

$$wb_{11}00, wb_{11}01, wb_{11}02 \ldots wb_{11}99, wb_{12}00 \ldots wb_{km}99$$

By specifying the order of the NPA/NXXs, one has literally specified the location of every potential telephone number relative to all others. By extension, one can envision a 50 State National NPA/NXX sample frame with an exhaustive listing of all possible telephone numbers: from the first available number in the first NPA/NXX in Boston, MA, to the last available number in the last non-metro county in Alaska; a frame containing approximately 207,527,300 distinct elements - 2,075,273 working banks, or hundred series, with 100 two-digits suffixes for each.

The notational expansion clearly shows that the NPA/NXXs comprising the sample frame can be conceptualized as one long string of 10-digit telephone numbers - each unique, and each in a known and replicable position relative to all others. In other words, if it was necessary to identify the ith element, or telephone number in this string, the process would be as follows:

1) Unique Position $= 1 + (i / 100.0)$

2) The whole number portion of the result will always designate the sequential working bank, and by association the NPA/NXX - in other words, the first eight digits of the telephone number;

3) While the two-digit fraction directly indicates the suffix $(00 - 99)$.

This concept is important for understanding various operational aspects of the GENESYS sampling process as well as the single stage *epsem* RDD process.

***EPSEM* Sample Selection Process.** The defined sample frame, K, is in actuality an extended string of elements, each element being a 10-digit telephone number occupying a unique and identifiable position. The length of the string is N^K. At this point the user has already determined the sample size, n', to be selected from the defined sample frame. The sample selection process first determines an initial selection interval size, H^K, by dividing the total number of elements, N^K by the desired sample size, n',

$$H^K = N^K / n' \text{ where,} \qquad H^K = \text{length of the selection interval;}$$

$$N^K = \text{number of elements in the defined sample frame; and,}$$

$$n' = \text{the desired sample size.}$$

This operation creates n' equal size selection intervals, or implicit strata, of size H^K. However, since the interval H^K will not necessarily be integral, the implied boundaries will cause telephone numbers to be divided between neighboring intervals. Consequently, the GENESYS system provides three selection options:

1) The interval H^K is modified by truncating any fractional portion,, with the new interval, $H^{k'}$, being integral. However, this will increase the expected sample size by a factor r (where r = H^k /H^k).The desired sample size is then attained by use of an external double sampling routine.

2) If the interval remains as originally defined, telephone numbers straddling interval boundaries effectively have their probabilities parsed between two neighboring strata, meaning that they have a probability of being selected twice. Consequently, the resultant sample size, will be slightly smaller than desired, due to these dual selections.

3) Alternatively, a telephone number selected a second time, from a neighboring interval can be replaced with another random selection from the same interval.

Although last procedure is not strictly epsem, as are the prior two options, it does always achieve the objective of obtaining the exact desired sample size, n'. Since RDD sampling rates are typically very small, any potential bias resulting from ignoring resultant variations in probabilities of selection will also be small. And, will be offset somewhat, by insuring a selection within each of the n' sampling intervals.

The actual RDD selection process is identical for all the above options:

1) A random number greater than 0, and less than or equal to 1.0 is then generated, this number is multiplied by the interval size, H^K or $H^{K'}$, providing a pointer to a designated element, n'_1, in the first interval. Dividing this result by 100.0 provides the sequential wb$_i$ bank in which the number is located, while the fractional portion, truncated to two digits, provides the suffix.

2) For the second interval, a new random number is generated, and again multiplied by the interval size. Adding H^K or $H^{K'}$ to the result, and dividing by 100.0, the number now points to a unique element in the second interval, n'_2.

3) The process in step three is repeated until the string is exhausted and exactly n' ten-digit suffixes are generated.

The process effectively segments the string of N^K elements into n' equal selection strata. From each stratum a single element is selected.

[Again, please note that selection interval boundaries will often "split" an element, since the stratum size H^K, is typically not an integral value. Consequently, there is a probability that a "split" element may be selected twice from neighboring strata. Following selection of a number in stratum h_i, the number is checked against the telephone number selected in h_{i-1}; if it is duplicated, the number is discarded and the selection in h_i is repeated until a unique element is selected.]

Non-*EPSEM* RDD Methods.

The use of non-*epsem* rdd methods is commonplace in the commercial research industry. In fact, the most widely used, most well-known sampling method marketed by the largest commercial sample supplier is a non-*epsem* rdd sample. The non-*epsem* procedures supported by GENESYS can be employed to replicate many of the "cost-effective" commercial rdd methods. The difference being that the GENESYS procedures are tractable, ***employing known but unequal probabilities of selection.***

Although a primary motivation for supporting such methods is the continuing need to replicate other commercial sampling procedures, these methods are eminently useful in their own right. In combination with the NPA/NXX-level demographic estimates, these procedures provide an efficient and tractable means of oversampling NXXs serving households with selected geo-demographic characteristics.

GENESYS provides the user with the ability to explicitly redefine the MOS variable associated with an NPA/NXX. The GENESYS Database contains an estimate of total households and total population, the exact count of directory listed telephone households, as well as a number of demographic variables for each NPA/NXX. These variables can be utilized individually, or in combination, to explicitly define the MOS variable associated with the defined set of NPA/NXXs.

The most popular non-*epsem* commercial RDD methods increase the probability that a particular number will result in a household contact by use of non-*epsem* sampling procedures. This is accomplished by varying sampling rates based on various implicit MOS variables that are correlated with the density of residential number household assignment.

Typically, such methods use the number of directory listed telephone households. In practice, this may not be the most highly correlated variable with density, but it is the most easily obtained. Sampling in proportion to the number of directory listed telephone households, will result in over sampling NPA/NXXs with higher densities and undersampling those with lower densities. What is usually not recognized however, is that the resultant implicit sampling rates are impacted by actual variations in residential unlisted rates:

1) Listed rates vary from 40 to over 80% depending upon geographic location - central cities as well as high income suburban areas typically have higher than average unpublished rates.

2) High growth areas and the NPA/NXXs serving them, typically have high effective unlisted rates because telephone directories are more out-dated. This situation is exacerbated when new NPA/NXXs are created.

In short, uncontrolled sampling processes utilizing listed households or other variables may result in significant sample biases, both on a geographic or demographic basis. Although one achieves data collection cost reductions, by increasing the average likelihood of reaching a household by 10% or more, the issue of potential sample bias cannot be resolved since neither actual nor relative probabilities of selection are known or reported.

GENESYS supports a non-*epsem*, or modified RDD sampling method. This alternative has been appropriately named MOD1.

Although these modified RDD methods are patterned after widely used methodologies, their significance is that they represent the industry's first tractable sampling applications. By assigning an explicit measure of size (MOS) to the individual NPA/NXXs comprising the sample frame, or within a particular stratum, MOD1 produces a single-stage PPS (probability proportional to size) sample of residential telephone numbers.

The explicit MOS values are retained in each sample record for use in constructing weighting factors to compensate for the disproportionate sampling.

The following sections detail the methodological procedures employed in these non-*epsem* RDD processes.

MOD1 Methodology. Operationally, the MOD1 procedure parallels the *epsem* RDD process. However, where the *epsem* RDD process assumes equal measures of size (MOS) across NPA/NXXs, this assumption is now relaxed, allowing for unequal MOS assignments. Whether one examines individual NPA/NXXs or working banks, each has an equal number of elements, 10,000 or 100, respectively. In other words, the *epsem* RDD process assumes a constant, implicit MOS variable associated with each sampling unit.

The GENESYS MOD1 procedure does not alter the operative sample frame. The frame remains as detailed above. It comprises the variable number of residential working banks associated with each of the NPA/NXX combinations. However, the sampling rate applicable to working banks comprising each NPA/NXX is a variable determined explicitly by the MOS_k assigned to each respective combination.

For any defined sample frame K, the sum of the k Measures of Size, $MOS1_T$, is defined as

$$MOS1_T = \sum_K MOS_k \quad \text{where,} \, MOS_k = \text{the user defined Measure of Size associated with}$$

the kth NPA/NXX.

The expected sample take from the kth NPA/NXX can then be expressed as

$$E[n_k] = n' * (MOS_k / MOS1_T) \quad \text{where} \quad n' = \text{the desired overall sample size.}$$

And, the sampling fraction, f_k, within the kth NPA/NXX is equal to,

$$f_k = E[n_k] / (WB_k * 100)$$

The previous equations can also be expressed in terms of the relative sampling fraction and a constant.

$$f_k = (n' / MOS1_T) * (MOS_k / (WB_k * 100)) = c * f'_k \quad \text{where,} \, c = \text{a constant based on the desired sample size and the de-}$$
fined sample frame's total MOS1.

f_k = the relative sampling fraction associated with the **kth** NPA/NXX

This is an important result as the inverse of the relative sampling fraction, $f_k'^{-1}$, is included in the output record of each sample telephone number generated, and is intended for use in developing probability-based selection weights.

As with the *epsem* RDD process, there is a determinable maximum sample file yield for the MOD1 process, which is in general less than that for the *epsem* process. The maximum yield is a function of the relative sampling fractions and the constraint that the maximum sampling rate must be less than or equal to one, $f_{\max} \leq 1.0$. First, the NPA/NXX with the largest ratio of MOS_k to available numbers is identified as this will provide the limiting sampling rate.

$$f_{\max} = \max_k [MOS_k / (wb_k * 100)]$$

The maximum sample file yield can then be determined directly as

$$N1_k = MOS1_T * f_{\max}^{-1}$$ where, $N1_k$ = the maximum sample file yield from the sample frame, K, using a MOD1 process based on the defined MOS1.

Again, the ultimate sampling fractions, f_k, are a function of the desired sample yield, constrained by the maximum sampling rate, $f_{\max} \leq 1.0$. As the sample frame's defined MOS1 approaches uniformity in distribution, the maximum sample yield approaches that of the *epsem* process as a limit.

If we define c as the desired sample size, n', f_k reflects the actual sampling rate within each NPA/NXX. And, as indicated previously, the expected yield from each NPA/NXX is

$$E[n_k] = f_k * wb_k * 100$$

The underlying order of the NPA/NXX combinations comprising the defined sample frame is in the identical geo-metro hierarchy, described in the *epsem* RDD process.

$$K = K_1, K_2 \ldots K_k$$

And again, within each of these k NPA/NXXs, we have a specified variable number of working residential banks, wb_k:

$$WB_k = wb_{11}, wb_{12} \ldots wb_{1k}, wb_{21} \ldots wb_{km}$$

And finally, within each of the wb_{km} banks there are exactly 100 two-digit suffixes:

$$wb_{11^{00}}, wb_{11^{01}}, wb_{11^{02}} \ldots wb_{11^{99}}, wb_{12^{00}} \ldots wb_{km^{99}}$$

The specified order of the NPA/NXXs, still uniquely identifies the location of every potential telephone number relative to all others.

However, the MOD1 process is non-*epsem*, and where the implicit selection intervals were conceptualized in terms of equal segments of 10-digit telephone numbers, the MOD1 selection interval is now defined as equal $MOS1$ segments. In other words, the selection interval, $H1^K$ is measured in $MOS1$ units rather than individual telephone numbers.

$$H1^K = MOS1_T / n'$$ where, $H1^K$ = length of the selection interval;

$MOS1_T$ = sum of MOS_k in the defined sample frame;

n' = desired sample size, $(\leq N1_K)$

The result is n' equal size $MOS1$-based intervals of size $H1^K$, comprised of a variable number of ten-digit telephone numbers. The actual selection interval definition and selection process is sequential, beginning with the first NPA/NXX in the ordered sample frame.

1) A random number RN greater than 0, and less than or equal to 1.0 is generated. This number is multiplied by the interval size, $H1^K$, providing a pointer that is then mapped to an individual element, n'_1, in the first interval. This sample element is identified by accumulating, sequentially, the element-based MOS measures until the indicated total is reached. In other words,

$$n'_1 = n_{kj}' \{RN * H1^k = \sum_k \sum_j MOS_k / (WB_k * 100)\} \text{ where, } n'_1 = \text{sample element selected from the first interval;}$$

$$j = \text{jth element, or four-digit suffix in the kth NPA/NXX.}$$

2) For the second interval, a new random number is generated, and again multiplied by the interval size. Adding $H1^K$ to the result, the accumulation is continued until n'_2 is identified.

$$n'_2 = n_{kj}' \{H1^K + (RN * H1^K) = \sum_k \sum_j MOS_k / (wb_k * 100)\}$$

3) The process is repeated until the string is exhausted and exactly n' ten-digit suffixes are generated.

$$n'_i = n_{kj}' \{H1^K(i - 1) + (RN * H1^K) = \sum_k \sum_j MOS_k / wb_k * 100)\}$$

The process segments the string of N^K elements into n' equal selection strata. From each stratum a single element is randomly selected. Please note, that selection interval boundaries will often "split" an element, since the stratum size $H1^K$, is typically not an integral value. Consequently, there is a probability that a "split" element may be selected twice from neighboring strata.

Following selection of a number in stratum h_i, the number is checked against the telephone number selected in h_{i-1}, if it is duplicated, the number is discarded and the selection in h_i is repeated until a unique element is selected (i.e, the sampling is accomplished without replacement).

Measure of Size (MOS) Manipulation. As detailed in prior sections, the Measure of Size (MOS) is an explicit, user specified, combination of household or population estimates derived for each telephone exchange. And, the MOS explicitly controls the allocation of sample and the sampling rates across the exchanges comprising each sample frame.

In most cases the user will find that the specification of the desired MOS is closely approximated by the demographic categories included in the GENESYS database. However, the MOS definition field does allow for both arithmetic and logical operators. Providing the following types of manipulations:

- Combinations of age and/or income categories;

- Creating MOS values for categories which are not explicitly defined in the database; one example would be estimating the number of households with $60,000+ by apportioning, say 45% of the of the $50-75,000, and adding it to the $75,000+ category, to approximate a $60,000+ MOS;

- Inserting an MOS of zero, for exchanges which meet a specified criteria, or increasing the MOS for certain exchanges; although operations such as this may be better handled through specific stratification and sample allocation procedures, the capabilities are there.

Finally, it is also possible to alter the MOS field to include a specific MOS. This can be accomplished through an arithmetic equation or identity, with or without logical operators.