SAMPLE ALLOCATION FOR STRATIFIED TELEPHONE SAMPLE DESIGNS

Clyde Tucker and Robert Casady, Bureau of Labor Statistics
James Lepkowski, University of Michigan

1. INTRODUCTION

One of the most difficult tasks in conducting telephone surveys is locating households using a frame of telephone numbers. Only about twenty percent of the telephone numbers in the United States are assigned to residences, and the search for these residential numbers increases the costs of the survey and the length of the required interviewing period. The most popular method for reducing the problem of locating households was first proposed by Mitofsky (1970) and more fully developed by Waksberg (1978). The Mitofsky-Waksberg technique capitalizes on a feature of the distribution of working residential numbers (hereafter referred to as WRNs) in the United States: they tend to be highly clustered within banks of consecutive numbers.

Instead of simply dialing numbers at random, Mitofsky and Waksberg outlined a two-stage design in which banks of 100 consecutive numbers are randomly selected from a frame constructed by appending all 10,000 four-digit suffixes to the list of area code-prefix combinations obtained from BellCore Research (BCR), and a single number from each bank is called. If the number is residential, the rest of the numbers in the hundred bank are retained for use in second-stage sampling. Otherwise, the bank is discarded. By restricting calling to these screened banks, the likelihood of contacting a residence increases threefold to about sixty percent. This procedure produces, in principle, an unbiased sample of telephone households, and one only need know the universe of area code-prefix combinations.

Unfortunately, there are several disadvantages which become apparent when the Mitofsky-Waksberg technique is applied to standard, time-limited cross-sectional surveys. They include the following:

1. The concentration of the sample within certain banks results in an intraclass correlation which, depending on the characteristic being measured, could substantially increase the variance of the estimate.

2. The k residential numbers must be contacted in each retained bank at the second stage. This is not usually a serious problem for hundred banks, but it would be for smaller bank sizes. It does mean, however, that only a portion of the numbers in a bank can be used before the bank is discarded.

3. Practical limits on the length of the surveying period will prevent finding the requisite number of households in some banks even though they exist.

4. Numbers generated as replacements for non-residential numbers in the original second-stage sample will receive less varied opportunities for calling, especially near the end of the surveying period. A small residual of numbers typically accumulates at the end of the study period for which a final resolution of residential status is impossible within the time constraints.

Brick and Waksberg (1991) described a modification of the Mitofsky-Waksberg procedure suggested earlier by Waksberg (1984) which eliminates the need to contact the same number of households in every cluster. Instead, a constant number of telephone numbers are contacted in a bank, and weights are assigned to the households found in each of these clusters. The weight for a household is proportional to the reciprocal of the number of households in its cluster. Although the methodology proposed by Brick and Waksberg does simplify the Mitofsky-Waksberg procedure, it has several problems. Only a slight bias is introduced, but the variances can be affected more substantially. Not only will the variable weights increase the variances (unless they are trimmed), but the cluster sizes (10 or more) necessary to stabilize these weights may limit the number of times the banks can be reused and exacerbate the effects of intraclass correlation.

Another way to avoid the complexity of the Mitofsky-Waksberg procedure is to select banks of numbers based on external information. Sudman (1973) and Lepkowski and Groves (1986) proposed sampling blocks of numbers using probabilities developed from data on listed residential numbers. This method, however, either restricts the sample to banks with listed numbers or requires that it be supplemented with a sample drawn using the Mitofsky-Waksberg procedure. Furthermore, as Brick and Waksberg observed, the listing rate in the United States is declining to the point that the number of residential listings in a bank may not accurately reflect the total number of households.

Casady and Lepkowski (1991) offered an attractive alternative to the above designs which also uses information on listed residential numbers. They proposed using the counts of listed numbers in banks with one or more listed numbers to stratify the universe of telephone numbers available from BCR into a "high- density" stratum of numbers in banks with at least one listed number and a "low-density" stratum of all other numbers. Estimates of the probability of contacting a residence in the high-density stratum range from 52% when using hundred banks to 58% when using ten banks, rates comparable to that in the second stage of the Mitofsky-Waksberg procedure. Only about 2% of the numbers in the low-density stratum will be assigned to residences. The low-density stratum may be discarded, sampled using

an RDD procedure, or further stratified using additional information available from BCR.

This design has several advantages over those previously discussed. Although the information on counts of listed numbers must be purchased, first-stage screening costs are avoided, at least for the high-density stratum. Actually, only a list of the banks with one or more listed numbers is needed, and this probably overcomes much of the problem associated with the declining listing rate. The counts of listed and total residential numbers do not have to be highly correlated. Simple random sampling can be used in the high-density stratum and, possibly, throughout. Thus, variances are not affected by intraclass correlation, and implementation of the design is relatively straightforward. Finally, stratifying the frame in this way leads to a number of design options.

Casady and Lepkowski discussed some of these options, and Conner and Heeringa (1992) recently tested two designs. However, too little information on the low-density stratum has been available until now to specify all of the alternative designs or fully evaluate the ones that have been considered. This paper reports the results of a study undertaken to gather the necessary information.

2. THE STUDY DESIGN

In order to develop optimal designs using the Casady-Lepkowski methodology, information about the distribution of residential numbers in the low-density stratum is needed. The first step was to obtain a file of the counts of listed residential numbers in all of the ten banks on a frame of listed numbers kept by Donnelley Marketing. This information, purchased in April, 1990, was merged with a file containing the universe of ten banks developed from the BCR frame. The ten banks without listed numbers were assigned to sampling substrata using variables previously identified as being related to residential hit rate (Groves, Lepkowski, & Tucker, 1990). These variables, which were obtained from the BCR file, were (1) whether the area code-exchange of the ten bank was only on the Donnelley frame, only on the BellCore frame, or on both; (2) whether or not the ten banks in area code-exchanges appearing on the Donnelley frame were from thousand banks with listed numbers; (3) whether the wire center in which the area code-exchange was located contained one or more than one exchange (a surrogate for rural-urban).

At the time of the study, late 1991, financial support could not be obtained for collecting information on a sample of numbers from the entire frame of low density banks, but calling could be done to augment a sampling operation in six primary sampling units (PSUs) that had been dropped from the Consumer Price Program due to budget reductions. These PSUs were Columbus, Ohio; Salt Lake City, Utah; Phoenix, Arizona; Sacremento, California; a group of ten contiguous rural counties in central Kentucky; and two neighboring counties in North Carolina which are on the border with South Carolina. These PSUs cannot be considered representative, but they do include both rural and urban areas. Unfortunately, the urban areas are concentrated in the West, and the rural areas are in the South.

The proportion of empty banks in the six PSUs is substantially smaller than that in the national distribution. This is to be expected given that these PSUs do not contain the vast areas in the U.S., particularly in the West, that are sparsely populated. Thus, the low-density substratum most poorly represented in the study area is the one with banks in rural areas (a single area code-exchange in the wire center) that have no residential listings in the thousand bank. As a result, the proportion of numbers in the low-density stratum that are assigned to residences is probably larger than in the nation as a whole.

A simple random sample of six hundred numbers were drawn from each of the five low-density substrata represented in the six PSUs. The residential status of each number was determined by attempting as many as twelve calls. A total of 168 numbers (5.6%) were identified as residential. All remaining numbers in the ten banks containing these residential numbers were called to estimate residential densities.

3. RESULTS

The hit rates (h_i) and densities (w_i) obtained from the study are given in Table 1. As previously noted, no information was obtained in this experiment for telephone numbers in "listed 10-banks". Hence, the information provided in Table 1 for this set of telephone numbers is derived from general knowledge and experience obtained in previous telephone surveys. It should also be noted that:

(1) The estimated value of $\bar{h}$ is based on a sample of 34,000 primary numbers in a telephone survey conducted by Westat.

(2) The values of the P_i are known exactly.

(3) The z_i (except for the first stratum) were determined by the equation $z_i = h_i P_i / \bar{h}$.

(4) The value of z_1 is derived by subtraction and h_1 by $h_1 = \bar{h} z_1 / P_1$.

(5) The w_i (except for the first stratum) were estimated from an experiment conducted by BLS and the corresponding t_i were determined by the equation

$t_i = 1 - h_i/w_i$; the estimated values for w_i and t_i are based on general knowledge.

(6) The "Residual" category contains all telephone numbers from unlisted 10-banks that are found in one, and only one, of the two frames. For the six PSUs this category consisted entirely of "BellCore only" telephone numbers.

Based on the information provided in Table 1, the frame was partitioned into the four basic strata defined below:

Very High Density Stratum:
{All telephone numbers in a listed 10-bank}

Moderate Density Stratum:
{All telephone numbers in an unlisted 10-bank}$\cap$
{All telephone numbers in a listed 1000-bank}

Low Density:
{All telephone numbers in an unlisted 10-bank} $\cap$
{All telephone numbers in an unlisted 1000-bank} $\cap$
{All telephone numbers in a 2+ prefix exchange} $\cup$
{Residual}

Very Low Density:
{All telephone numbers in an unlisted 10-bank} $\cap$
{All telephone numbers in an unlisted 1000-bank} $\cap$
{All telephone numbers in a 1 prefix exchange}

As can be seen in Table 2, each of the four strata comprises a significant portion of the frame and can be clearly distinguished from the others on the basis of hit rate. Two alternative stratification schemes were developed by collapsing the basic strata in different ways. These alternatives are given in tables 3 and 4.

Stratified designs based on the frame stratification given in tables 2-4, as well as the Mitofsky-Waksberg design, were compared to simple RDD sampling of the combined frame using the cost model described by Waksberg (1978) and used by Casady and Lepkowski (1991). Specifically, the sample designs included in the study were

Design 1. Mitofsky-Waksberg sampling applied to combined frame. It is not practical to use 10-bank second stage clusters for this design so 100-bank clusters were assumed. The proportional reduction in variance is from Casady and Lepkowski (1991).

Design 2. Frame stratified as in Table 2.\ Simple RDD sampling within each of the four strata with stratum sample sizes determined by optimal allocation.

Design 3. Frame stratified as in Table 3.\ Simple RDD sampling within each of the three strata with stratum sample sizes determined by optimal allocation.

Design 4. Frame stratified as in Table 4.\ Simple RDD sampling within each of the three strata with stratum sample sizes determined by optimal allocation.

The proportional reductions in variance or cost , for typical cost ratios, are in table 5. The cost ratios compare the cost of a productive number (obtaining a completed interview) to an unproductive number, be it residential or not. There are virtually no differences in efficiency among the designs for this range of cost ratios.

The sample designs discussed above assume that the sample will be drawn from the entire frame using optimal allocation, but, at the discretion of the researcher, part of the frame can be discarded to further improve efficiency at the risk of some bias. Designs using only part of the frame are referred to here as "truncated" designs; our attention will be limited to designs that achieve truncation through the elimination of an entire stratum. Several options are available depending on the initial stratification scheme chosen and the amount of potential bias that can be tolerated. The Mitofsky-Waksberg design is not considered in the following because the stratification schemes, and hence the truncation strategies, are based on ten banks. This dictates that the Mitofsky-Waksberg design would of necessity be applied to 10-banks, which is not practical. Had the strata been constructed from hundred banks, truncated Mitofsky-Waksberg designs could have been evaluated.

The first three truncated designs discard the "Very Low Density" stratum as defined in tables 2 and 3:

Design 1. Simple RDD sampling applied to the First Truncation Frame.

Design 2. First Truncation Frame stratified as in Basic Stratification Scheme and Simple RDD sampling within each of the three remaining strata; stratum sample sizes determined by optimal allocation.

Design 3. First Truncation Frame stratified as in First Alternative Stratification Scheme and Simple RDD sampling within each of the two remaining strata; stratum sample sizes determined by optimal allocation.

Results for these three designs are given in Table 6. Just eliminating the "Very Low Density" stratum (about a fourth of all ten banks) from a simple RDD sampling design increases efficiency about 10%, but the gain is greater when either the basic stratification or the first alternative (Table 3) is used. However, this is only a gain of about 5-6% over using these

sampling strata with the entire frame. In all of these designs, about 1% of the frame is lost. The potential bias is likely to be inconsequential, especially when surveying the general population of telephone households.

The other two truncated sample designs, which discard the "Low/Very Low Density" stratum (as given in Table 4), are

Design 1. Simple RDD sampling applied to the Second Truncation Frame.

Design 2. Second Truncation Frame stratified as in Second Alternative Stratification Scheme and Simple RDD sampling within each of the two remaining strata; stratum sample sizes determined by optimal allocation.

As can be seen in the results presented in Table 7, almost 6% of the population is not covered for these two designs. On the other hand, over half of all telephone numbers have been discarded. The resulting increase in efficiency is substantial. With this much of the frame eliminated, stratification does not offer much of an advantage over simple RDD sampling. The potential bias created from the loss of 6% of the telephone households could be serious, depending on the characteristics of interest.

4. CONCLUSION

Even if the Mitofsky-Waksberg procedure can be easily administered or the Brick-Waksberg modified design used, potential intraclass correlation can increase the variance in estimates. This problem is eliminated with the list-assisted designs presented here; and, furthermore, the increase in efficiency for the designs using the whole frame is otherwise comparable to the second stage of Mitofsky-Waksberg. The truncated designs, especially the second one, provide additional increases in efficiency if the potential biases can be tolerated. These conclusions hold for most reasonable cost ratios. If the cost ratio is very large, 20 or more, none of the designs, including Mitofsky-Waksberg, are much better than simple RDD sampling.

For sensitive situations in which it is important to demonstrate that the entire population has been covered, the basic stratified design and the two alternatives are comparable. In less sensitive situations where a small potential bias is acceptable, the first truncation designs using the basic stratification scheme or the first alternative do increase efficiency somewhat. It is the second truncation, however, which produces large gains compared to using the whole frame, and simple RDD sampling in the reduced frame does about as well as stratifying. The problem is that the potential bias in this case can be large, especially if the interest is in certain subpopulations. For instance, the portion of the frame truncated that comes from the "Residual" category has a disproportionately large share of college housing.

The results reported here do not take into account first-stage costs. More information about these costs is needed, and they should be incorporated in the cost model. For the Mitofsky-Waksberg design, these are screening costs. For the list-assisted designs, the Donnelley file must be purchased. In addition, all of the designs require the BellCore file, and some programming and processing costs always will be incurred. The processing costs for passing the Donnelley file may be quite large, depending on the available hardware and software. Regardless of the design chosen, costs usually can be amortized over several survey administrations.

A study which will provide the stratum hit rates at the national level will be completed next year. As mentioned earlier, these estimates are expected to be slightly smaller than those found in the 6 PSUs. Thus, the potential biases that result from discarding part of the frame may be even less than reported here.

5. REFERENCES

Brick, J.M., and Waksberg, J. (1991), "Avoiding Sequential Sampling with Random Digit Dialing," Survey Methodology, Vol. 17, No. 1, June, pp. 27-42.

Casady, R.J., and Lepkowski, J.M. (1991), "Optimal Allocation for Stratified Telephone Survey Designs," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 444-447.

Conner, J.H., and Herringa, S.G. (1992), "Evaluation of Two New Cost Efficient RDD Designs," paper presented at the annual meeting of the American Association for Public Opinion Research, St. Petersburg, FL.

Groves, R.M., Lepkowski, J.M., and Tucker, C. (1990), "Assessing Telephone Sample Designs That Use Counts of Listed Numbers to Improve Efficiency," paper presented at the annual meeting of the American Association for Public Opinion Research, Lancaster, PA.

Lepkowski, J.M., and Groves, R.M. (1986), "A Two Phase Probability Proportional to Size Design for Telephone Sampling," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 73-98.

Mitofsky, W. (1970), "Sampling of Telephone Households," unpublished CBS News memorandum.

Sudman, S. (1973), "The Uses of Telephone Directories for Survey Sampling," Journal of Marketing Research, Vol. 10, No. 2, May, pp. 204-207.

Waksberg, J. (1978), "Sampling Methods for Random Digit Dialing," Journal of the American Statistical Association, Vol. 73, No. 361, March, pp. 40-46.

Waksberg, J. (1984), "Efficiency of Alternative Methods of Establishing Cluster Sizes in RDD Sampling," unpublished Westat Inc. memorandum.

Table 1. Approximate values of frame parameters when 10-bank characteristics are used to partition the combined BCR/Donnelley frame of telephone numbers. The "Residual" class consists of those telephone numbers found in one, but not both, of the two primary frames. A "Listed Bank" is a bank containing at least one listed number and a "Non-Empty Bank" is a bank containing at least one Working Residential Number.

Location of Telephone Number	Prop. of Frame (P_i)	Prop. of Pop. (z_i)	Hit Rate (h_i)	Prop. of Empty Banks (t_i)	Hit Rate Within Non-Empty Banks (w_i)
Listed 10-Bank	.3390	.8865	.5736	.0440	.6000
Unlisted 10-Bank, Unlisted 1000-bank, 1 Prefix Exchange	.2397	.0109	.0100	.9714	.3500
Unlisted 10-Bank, Listed 1000-bank, 1 Prefix Exchange	.0450	.0191	.0930	.8371	.5710
Unlisted 10-Bank, Unlisted 1000-bank, 2+ Prefix Exchange	.0982	.0170	.0380	.9283	.5300
Unlisted 10-Bank, Listed 1000-bank, 2+ Prefix Exchange	.0761	.0371	.1070	.7714	.4680
Residual	.2020	.0294	.0320	.8933	.3000

$\bar{h} = .2194$ and $\bar{t} = .6157$

Table 2. Approximate values of the frame parameters for the Basic Stratification Scheme. Stratum definitions are given below.

Stratum	Prop. of Frame (P_i)	Prop. of Population (z_i)	Hit Rate (h_i)	Prop. of Empty 10-Banks (t_i)	Hit Rate Within Non-Empty Banks (w_i)
Very High Density	.3390	.8865	.5736	.0440	.6000
Moderate Density	.1211	.0562	.1018	.7958	.4985
Low Density	.3002	.0464	.0339	.9047	.3557
Very Low Density	.2397	.0109	.0100	.9714	.3500

Table 3. Approximate values of the frame parameters for the First Alternative Stratification Scheme; the Moderate Density Stratum and Low Density Stratum have been collapsed into a single Moderate\Low Density Stratum.

Stratum	Prop. of Frame (P_i)	Prop. of Population (z_i)	Hit Rate (h_i)	Prop. of Empty 10-Banks (t_i)	Hit Rate Within Non-empty Banks (w_i)
Very High Density	.3390	.8865	.5736	.0440	.6000
Moderate\Low Density	.4213	.1026	.0534	.8733	.4214
Very Low Density	.2397	.0109	.0100	.9714	.3500

Table 4. Approximate values of the frame parameters for the Second Alternative Stratification Scheme; the Low Density Stratum and the Very Low Density Stratum have been collapsed into a single Low\Very Low Density Stratum.

Stratum	Prop. of Frame (P_i)	Prop. of Population (z_i)	Hit Rate (h_i)	Prop. of Empty 10-Banks (t_i)	Hit Rate Within Non Empty Banks (w_i)
Very High Density	.3390	.8865	.5736	.0440	.6000
Moderate Density	.1211	.0562	.1018	.7958	.4985
Low\Very Low Density	.5399	.0573	.0233	.9343	.3546

Table 5. Projected proportional reduction in variance\cost (relative to simple RDD sampling of the combined frame) for four alternative sample designs. All four of the alternative designs sample from the entire combined frame and hence cover all of the target population. Cost ratios are typical of research situations.

Sample Design	Proportional Reduction in Variance or Cost			Prop. of Population Not in Scope
	$\gamma = 4$	$\gamma = 6$	$\gamma = 8$	
1. Mitofsky-Waksberg	.1719	.1161	.0783	.0000
2. Basic Strat. \ OA	.1683	.1185	.0890	.0000
3. First Alter. \ OA	.1601	.1119	.0836	.0000
4. Second Alter. \ OA	.1595	.1109	.0823	.0000

Table 6. Projected proportional reduction in variance\cost (relative to simple RDD sampling of the combined frame) for three alternative sample designs based on sampling from the combined frame less the "Very Low Density" stratum. Cost ratios are typical of research situations.

Sample Design	Proportional Reduction in Variance or Cost			Prop. of Population Not in Scope
	$\gamma = 4$	$\gamma = 6$	$\gamma = 8$	
1. Trun1\RDD	.1359	.1102	.0912	.0109
2. Trun1(Basic Strat.) \ OA	.2234	.1670	.1325	.0109
3. Trun1(First Alter.) \ OA	.2153	.1606	.1272	.0109

Table 7. Projected proportional reduction in variance\cost (relative to simple RDD sampling of the combined frame) for two alternative sample designs based on sampling from the combined frame less the "Low Density" stratum. Cost ratios are typical of research situations.

Sample Design	Proportional Reduction in Variance or Cost			Prop. of Population Not in Scope
	$\gamma = 4$	$\gamma = 6$	$\gamma = 8$	
1. Trun2\RDD	.3087	.2442	.2019	.0573
2. Trun2(Second Alter.) \ OA	.3233	.2534	.2083	.0573