OPTIMAL ALLOCATION FOR STRATIFIED TELEPHONE SURVEY DESIGNS

Robert J. Casady, U. S. Bureau of Labor Statistics
James M. Lepkowski, University of Michigan

Key Words: Random digit dialing, Telephone surveys

1. THE CURRENT STATUS OF TELEPHONE SURVEY DESIGNS

The two stage random digit dialing design for sampling telephone households, first proposed by Mitofsky (1970) and more fully developed by Waksberg (1978), has been widely employed in telephone surveys. The Mitofsky-Waksberg technique capitalizes on the fact that working residential numbers (hereafter referred to as WRNs) tend to be highly clustered within banks of consecutive telephone numbers. Currently, only about twenty percent of the possible telephone numbers within the known area code, three digit prefix combinations are WRNs. However, if a bank of 100 consecutive telephone numbers can be identified that has at least one known WRN then, on average, over 50 percent of the numbers in the bank will be WRNs. The Mitofsky-Waksberg technique, which identifies 100-banks containing WRNs in the first stage of sampling, greatly reduces the amount of screening necessary to identify telephone numbers assigned to households.

Alternatively, lists of published telephone numbers have been employed as sampling frames. These lists of published numbers are available for the entire country from commercial firms such as Donnelley Marketing Information Systems. A straightforward selection of telephone numbers from such lists provides a very high rate of WRNs (typically at least 85%) but unfortunately does not cover households with unpublished numbers. Comparisons of telephone households with and without published numbers (see, for example, Brunner and Brunner, 1971) indicates that substantial bias may result.

The purpose of this paper is to examine stratified designs based on the BellCore Research (BCR) frame as an alternative to list frames and Mitofsky-Waksberg design. As an example of frame stratification, the BCR frame could be partitioned into two strata: a "high density" stratum consisting of residential numbers in 100-banks with one or more listed numbers and a "low density" stratum consisting of all the remaining numbers in the BCR frame. stratum. The obvious cost difference of samp from the two strata can be exploited thro differential sample allocation.

The next section examines the question of appropriate allocation of sample between the st when simple random sampling is utilized within e stratum. A key feature of the stratified teleph sample approach is that it permits alterna approaches to sample selection within the diffe strata. Several alternatives are presented discussed in Section 3. The paper concludes wit general discussion contrasting the Mitofsky-Waksb procedure and stratified designs.

2. THE ALLOCATION PROBLEM FOR STRATIFIED TELEPHONE DESIGNS

2.1 Background

For the purposes of this paper we will assume the basic sampling frame is the collection of telephone numbers generated by appending four d suffixes to the BCR list of area-prefix codes. We assume that each household in the target populatio "linked" to one and only one telephone number in basic sampling frame.

We will also assume that we have access (poss only indirect) to a directory based, machine reada list of telephone numbers such as that availa through Donnelley Marketing. It should be noted because many households choose not to list t telephone numbers in a directory, any such direct based frame will not contain all of the WRNs. A directory based lists are, by nature, out of date so t will omit some numbers that are currently WI while including others that are no longer WRNs.

From a survey design point of view these frames tend to be radically different. The BCR fra includes all WRNs so it provides complete "covera of the households in the target population, but c about 20 percent of the telephone numbers include the BCR frame are actually WRNs. Thus, the rate" (and hence sampling efficiency) will be q low for a simple RDD sample design utilizing BCR frame. In contrast, a typical directory/list fra covers only about 70 percent of the target househo but the hit rate is 80 to 90 percent. In general sampling efficiency for a simple RDD design usir

stratum with a very low hit rate. The sample is allocated to the strata so as to minimize cost (variance) for a specified variance (cost). Hereafter the low hit rate stratum will be referred to as the residual stratum.

2.2 Basic Notation

Assume that the BCR frame of telephone numbers has been partitioned into *H* strata based on a 100-bank attribute which can be determined from the directory based frame. The choice of 100-banks is arbitrary, 10-banks or any other sized banks would work just as well. For the i^{th} stratum let

P_i = proportion of the frame included in the stratum,

h_i = proportion of the telephone numbers in the stratum that are WRNs (i.e., the hit rate),

w_i = average proportion of WRNs in the non-empty 100-banks (i.e., the average hit rate for non-empty banks),

z_i = proportion of the target population included in the stratum, and

t_i = proportion of 100-banks in the stratum that contain no WRNs.

The average hit rate for the frame is $\bar{h} = \sum_{i=1}^{H} h_i P_i$ and the proportion of empty 100-banks in the frame is $\bar{t} = \sum_{i=1}^{H} t_i P_i$. In general only the P_i's will be known with certainty. Data from a joint research project involving the Bureau of Labor Statistics and the University of Michigan were used to provide approximate values for the parameters h_i and w_i for the two strata in the example. Values for the remaining parameters were calculated using the algebraic relationships $t_i = 1-(h_i/w_i)$ and $z_i = \frac{h_i P_i}{\bar{h}}$. The approximations for all of the frame parameters for the two stratum design are given in Table 1. The values in the table imply that for the BCR frame $\bar{h} \cong .211$ and $\bar{t} \cong .605$, which are in close agreement with those given in Waksberg (1978).

2.3 The Basic Estimation Problem, Sample Designs and Estimators

We assume the telephone numbers in the i^{th} stratum are labeled 1 through M_i and we let

$$d_{ij} = \begin{cases} 1 & \text{if the } j^{th} \text{ nb. in the } i^{th} \text{ stratum is a WRN,} \\ 0 & \text{otherwise} \end{cases}.$$

N_i denotes the number of WRNs in the i^{th} stratum

$N.$ denotes the number of WRNs in the population

We denote simple random sampling with replacement (i.e., simple RDD) from the telephone numbers in the BCR frame as design D_0 and stratified simple random sampling from the BCR frame (independent simple RDD samples are selected from each stratum) as design D_1. Under design D_0 the standard ratio estimator for μ is given by $\bar{Y}_0 = \hat{Y}_0 / \hat{N}_0$ where $\hat{Y}_0$ and $\hat{N}_0$ are the usual inflation estimators of $Y.$ and $N.$ respectively. The estimator $\bar{Y}_0$ is asymptotically unbiased for μ and its variance is given by

$$\text{var}(\bar{Y}_0) \cong \frac{\sigma^2}{m\bar{h}}$$

where m is the sample size and σ^2 is the population variance of the y's. For the design D_1 the standard ratio estimator of μ is given by $\bar{Y}_1 = \hat{Y}_1 / \hat{N}_1$ where $\hat{Y}_1$ and $\hat{N}_1$ are the standard inflation estimators of $Y.$ and $N.$ under stratified sampling. The estimator is also asymptotically unbiased for μ and

$$\text{var}(\bar{Y}_1) \cong \sum_{i=1}^{H} \frac{z_i^2 \sigma_i^2 \left(1+(1-h_i)\lambda_i\right)}{m_i h_i} \qquad (3$$

where $\lambda_i = (\mu_i - \mu)^2 / \sigma_i^2$ and m_i, μ_i, and σ_i^2 are stratum sample sizes, means, and variances respectively.

2.4 The Cost Model

There are costs associated both with determining the value of the indicator variable *d* and the value of the characteristic of interest *Y*. The cost function for determining the indicator variable is denoted by C_1 with

$$C_1(d) = \begin{cases} c_1 & \text{if } d = 1 \\ c_0 & \text{if } d = 0 \end{cases}$$

This model allows for the possibility that the cost of determining that a telephone number is not a WRN may be different than determining that a telephone number is a WRN. In fact, the cost of determining the status of telephone numbers that are WRNs is usually less. The cost of determining the value of the characteristic *Y* includes only the additional cost of determining the value of *y* after the value of *d* has been determined. Accordingly, let $C_2(\cdot,\cdot)$ represent this additional cost, then

$$\begin{cases} 0 & \text{if } d = 0 \end{cases}$$

for both designs . Letting $C(D_0)$ and $C(D_1)$ represent the total cost of conducting a survey under the two respective designs it is straightforward to show that

$$E[C(D_0)] = mc_0[1+(g-1)\bar{h}] \quad (2.2)$$

and

$$E[C(D_1)] = c_0 \sum_{i=1}^{H} m_i[1+(g-1)h_i]. \quad (2.3)$$

2.5 Optimal Allocation for $\bar{Y}_1$

The stratum sample allocation that minimizes $\mathrm{var}(\bar{y}_1)$ for a fixed expected total cost C^* (or that minimizes $E[C(D_1)]$ for a fixed variance V^*) is specified up to a proportionality constant by

$$m_i \propto \frac{z_i s_i}{\sqrt{h_i}}\left[\frac{c_0(1-h_i)}{c_1}\right]^{1/2} \Big/ [1+(g-1)h_i] \quad (2.4)$$

where the proportionality constant is determined by substitution into the expected cost equation (or the variance equation, as appropriate). Relative to RDD sampling, the proportional reduction in variance (cost) under optimal allocation for fixed cost (variance), denoted by $R(\bar{y}_1, \bar{Y}_0)$, is approximately

$$-\bar{h}\,\frac{\left[\sum_{i=1}^{H} \frac{z_i s_i}{\sqrt{h_i}}\left[\frac{c_0(1-h_i)}{c_1}\right]^{1/2}\right]^2}{s^2[1+(g-1)\bar{h}]}. \quad (2.5)$$

2.6 Practical Problems Associated With Optimal Allocation

The problem of specifying the values for the parameters in the allocation equations is generic to optimal allocation schemes. For our particular case there are three basic types of parameters: frame related (z_i and h_i), cost related (g and c_0) and those specific to the variable of interest ($\bar{Y}_i$ and s_i^2). Currently, we have a fairly good working knowledge of the frame related parameters for the two stratum example and certain other specific stratification schemes. In Section 4, we will discuss several active research projects that should further expand our knowledge in this area.

It is clear that $g \geq 1$ but the actual value can vary widely. Waksberg (1978) considers values of g between 2 and 20. Potentially the variable specific parameters pose the most serious problem. Usually our knowledge regarding the values of these parameters is limited and, in the case of multipurpose

Therefore, with caution, we assume that $\bar{Y}_i = \bar{Y}$ $s_i^2 = s^2$ for $i = 1,2,\ldots,H$. Optimal allocation achieved by

$$m_i \propto \frac{z_i}{\sqrt{h_i}}[1+(g-1)h_i]^{1/2} \quad (2$$

and the proportional reduction in variance is

$$R(\bar{y}_1, \bar{y}_0) \approx (1-\bar{h})\,\frac{\left[\sum_{i=1}^{H} z_i\sqrt{h_i}[1+(g-1)h_i]^{1/2}\right]^2}{H\sum_{i=1}^{H} z_i^2 h_i[1+(g-1)h_i]}. \quad (2$$

In the case of the two stratum example, the allocat specified by (2.6) implies that the allocation relat to the residual stratum (i.e., m_1/m_2) is 2.54 w $g = 2$ and 1.42 when $g = 10$. In the first case projected proportional reduction in variance $R = .283$ and in the second $R = .077$. In fact follows from (2.7) that as the relative cost determining the value of the variable of inte increases, the relative benefit of optimal allocat decreases.

The Mitofsky-Waksberg sample design, deno by D_3, employs two stages of sample selection (i non-empty 100-banks are selected in the first st and WRNs are selected in the second stage). T Mitofsky-Waksberg estimator, denoted by $\bar{Y}_3$, unbiased for $\bar{Y}$. Under "optimal" within 100-b sample allocation, the reduction in variance relative simple RDD for the estimator $\bar{Y}_3$, denoted by $R(\bar{y}_3$, is approximately

$$-\frac{[1+(g-1)\bar{h}-\bar{t}]\rho^2 + (1-\rho)^2 + \bar{t}]^2}{1+(g-1)\bar{h}} \quad (2$$

where ρ is intra-bank correlation. At the natio level Groves (1977) reports that $\rho \approx .05$ for econom or social statistics. Using this value of ρ, toget with the values of $\bar{h}$ and $\bar{t}$ from the two strat example, the projected proportional reduction variance for the Mitofsky-Waksberg procedure R=.281 when $g = 2$ and R=.060 when $g = 10$.

The two methodologies appear to prod essentially identical variance reduction for both val of the cost ratio. However, too much should not read into this simple comparison as the projec reduction for each of the procedures is based simplifying assumptions that will not be strictly t for any application. The only inference intended that the two procedures appear to highly competit

efficiency of the estimators, not their expectations. Unfortunately an extremely high price is paid for the assurance of unbiasedness because sampling from the residual stratum provides information on only a small proportion of the population and at a relatively high cost. If we are willing to settle for an estimate of the population mean exclusive of those households linked to telephone numbers in the residual stratum, we can "truncate" the original frame by eliminating the residual stratum and selecting a stratified RDD sample from the remaining telephone numbers. For the two stratum example the "truncated frame" would consist only of those telephone numbers in the first stratum. The hit rate for the sample from the truncated frame would be .521, in contrast to a hit rate of .211 for the entire frame. However, only about 94% of the target population would remain in scope.

In what follows we assume that the truncated frame is simply the original BCR frame less the residual stratum. Accordingly, for the truncated frame $\bar{h}^* = (\bar{h} - P_K h_K)/(1 - P_K)$ is the hit rate, the proportion of empty 100-banks is $\bar{t}^* = (\bar{t} - P_K t_K)/(1 - P_K)$ and $\mu^* = (\mu - z_K \mu_K)/(1 - z_K)$ is the population mean. Let design D_4 be stratified simple random sampling from the truncated frame, and $\bar{Y}_4$ the standard ratio estimator of the population mean. The estimator $\bar{Y}_4$ is asymptotically unbiased for μ^*, and, in general, it is biased for μ. The (asymptotic) bias is given by

$$B(\bar{Y}_4) = \mu^* - \mu = \frac{z_K(\mu - \mu_K)}{(1 - z_K)}. \qquad (3.1)$$

In most practical circumstances the bias tends to zero monotonically as the proportion of the target population in the residual stratum becomes small, although, as indicated by (3.1), this is not necessarily the case. In any event, since the value of $\mu - \mu_K$ is never known, an upper limit on the proportion of the population in the residual stratum is usually the key specification to be determined when considering the use of a truncated frame. For the two stratum example approximately 6% of the target population is excluded from the sampling frame and, in almost all cases, this would not be tolerable for Federal agencies.

The equations for cost, variance, allocation, and proportional reduction in variance (or cost) are essentially the same as those presented in Section 2. In fact, the only modifications required for equation

$$R(\bar{Y}_4, \bar{Y}_0) = 1 - \frac{\bar{h}(1 + \bar{h}^* g - 1)}{\bar{h}^*(1 + \bar{h} g - 1)}. \qquad (3$$

Thus for the two stratum design, the proportio reduction in variance (cost) is approximately .49 w $g = 2$ and .21 when $g = 10$. In both cases reduction is substantially greater than achieved by two methods in the previous section. However, nea 6% of the population is not covered by the frame.

In an attempt to retain the relative efficiency truncation while reducing the magnitude of coverage problem, BLS and the University Michigan are investigating several alternat stratification plans in an effort to reduce proportion of the population in the residual stratu One promising approach calls for the partitioning the residual stratum in the two stratum example form a new residual stratum which consists telephone numbers in 100-banks thought to primarily assigned to commercial establishments not yet activated for either residential or commer use. Estimated frame parameters for the result three stratum design are given in Table 2.

These data were used to compute the projec proportional reduction in variance for both the th stratum design and the truncated three stratum desi These results, together with a summary of the res for the two stratum designs and the Mitofs Waksberg design, are presented in Table 3. Tabl also includes the projected reduction in variance fo cost ratio of 20.

It appears that the proposed partitioning strat was reasonably successful as the percent of population out of scope was reduced from nearly to approximately 2%. The projected proportio reduction in variance for the truncated three strat design is approximately .41 when $g = 2$ and .16 w $g = 10$. From an efficiency point of view, it occup the middle ground between the highly effici truncated two stratum design and unbiased desig Of course the issue to be faced when considering s a design is the coverage problem. For any particu application the risk inherent in sampling from a fra that does not include all of the target population m be weighed against the potential gain in efficien As expected, the standard three stratum design slightly more efficient than the two stratum desi However, the increase in efficiency is so small tha is doubtful that the added cost of partitioning the B frame into an additional stratum is justified except

sampling in others. The motivation for this type of design is based on the following two considerations:

(a) Mitofsky-Waksberg sampling tends to be "administratively complex", and if the gain in efficiency is small, simple RDD is preferred.

(b) If the proportion of empty banks in a stratum is "small" then Mitofsky-Waksberg sampling offers little, if any, increase in efficiency.

Thus, we propose to utilize simple RDD sampling in strata with a "small" proportion of empty hundred banks and Mitofsky-Waksberg sampling in the remaining strata. A complete discussion of this topic may be found in Casady and Lepkowski (1991).

4. CONCLUDING REMARKS

The strengths of the Mitofsky-Waksberg technique for generating telephone samples are clear: high hit rates in the second stage of selection, an efficient method for screening out empty banks of telephone numbers, and a conceptually ingenious approach to sample generation. It is a remarkable testimony to the strength of the technique that after many years it is still considered to be the standard method of random digit dialing with few serious competitors . The weaknesses of the technique (first stage screening and replacement of non-residential numbers during the data collection) do not, on the surface, seem to be important relative to the general strength of the technique. However, these features can cause substantial difficulty, especially in short time period telephone survey operations.

Several alternatives are considered. In the two stratum methods, telephone numbers are divided into two groups, the high density stratum consisting of all telephone numbers in 100-banks containing listed numbers and the low density stratum consisting of all remaining telephone numbers. In the three stratum methods, the "all other numbers" or original low density stratum is further subdivided, on the basis of auxiliary data, into two strata. One of the new strata is expected to contain nearly all residential numbers in the original low density stratum and the other is now the (new) low density stratum. For both the two and three stratum design, two general alternatives are considered: (1) selecting simple random samples from all strata except the low density stratum frame where the Mitofsky-Waksberg method is used and (2) selecting simple random samples from all strata except the low density stratum which is not sampled at all.

residential numbers at the second stage, since the o telephone numbers that must be dialed in the h density stratum are those that are generated at beginning of the study.

Even for low cost ratios, the two and three strat designs are as efficient as the Mitofsky-Waksb approach. When numbers can be dropped from low density stratum, these alternatives are much m efficient, at the price of unknown bias due excluding part of the target population. When c ratios are high, the two and three stratum approac are clearly superior.

A critical issue is the magnitude of the b introduced by dropping the low density stratum. E though the proportion of the population in the strat is small, the magnitude of the bias may be relativ large for some characteristics and for some subgro of the population. Further empirical investigati are necessary.

The cost of the auxiliary list frame is addressed in this investigation because the fra information used in stratification was derived from specialized research file. Further investigation needed into this cost as it must be considered in practical application.

In order to improve the hit rates in the h density stratum, smaller banks of numbers can used. In another investigation we have found that banks will have hit rates in the neighborhood of compared to the .52 reported here for 100-ban Working with 10-banks substantially increases size of files and processing operations used generate samples. And, the cost of a 10-bank fram likely to be much higher than the 100-bank frame.

Clearly the results presented here are insuffici to draw final conclusions about the overall value these alternative designs. Further cost data empirical evidence on the size of the bias caused eliminating the numbers from the low density strat is needed before a final conclusion can be reached.

5. References

Brunner,J. A., and Brunner, G. A., "Are Voluntaril Unlisted Telephone Subscribers Really Different? *Journal of Marketing Research*, Vol. 8, February 1971, pp.121-124.

Casady, R. J. and Lepkowski, J. M., "Stratified Telephone Designs," unpublished report of the U Bureau of Labor Statistics, Washington D.C.

Groves, R. M. ,"An Empirical Comparison of Two

Table 1. Approximate values of the frame parameters for a two stratum design based on the BCR frame and the Donnelley list frame.

	Proportion of Frame (P_i)	Proportion of Population (z_i)	Hit Rate (h_i)	Proportion of Empty 100-Banks (t_i)	Hit Rate Within Non-empty Banks (w_i)
1	.3804	.9402	.5210	.0300	.5371
2	.6196	.0598	.0204	.9584	.4900

Table 2. Estimated frame parameters for a proposed three stratum design based on the BCR frame and the Donnelley list frame.

Stratum	Proportion of Frame (P_i)	Proportion of Population (z_i)	Hit Rate (h_i)	Proportion of Empty 100-Banks (t_i)	Hit Rate Within Non-empty Banks (w_i)
1	.3804	.9402	.5210	.0300	.5371
2	.2000	.0399	.0420	.9143	.4900
3	.4196	.0199	.0100	.9796	.4900

Table 3. Projected proportional reduction in variance (or cost) relative to simple RDD sampling for five alternative telephone sample designs.

Sample Design	Proportional Reduction in Variance or Cost			Proportion of Frame Not in Scope
	g = 2	g = 10	g = 20	
Two Stratum	.2829	.0766	.0320	.0000
Two Stratum (Truncated)	.4917	.2055	.1189	.0598
Mitofsky-Waksberg	.2811	.0597	.0135	.0000
Three Stratum	.3001	.0866	.0389	.0000
Three Stratum (Truncated)	.4095	.1574	.0879	.0199