TELEPHONE SURVEY DESIGNS

Robert J. Casady, U.S. Bureau of Labor Statistics
Room 2126, 441 G Street N.W., Washington D.C. 20212

KEY WORDS: Dual frame designs; Mitofsky–Waksberg design; Random digit dialing; Telephone surveys; Sample design for telephone surveys.

1. INTRODUCTION

Since the early 1970's, the use of the telephone mode for interviewing and data collection in household surveys has increased rapidly in the United States. During this same period of time there has been a parallel expansion in the development of survey designs for scientific studies targeted specifically at telephone populations. However, it appears that all of these telephone survey designs can be placed into one of three general categories based on the type of sampling frame utilized by the design. More specifically, in one category are survey designs built around convenient to use but defective frames, in another category are those survey designs built around frames which completely cover the telephone population but also include a high proportion of spurious entries, and in a third category are survey designs which use both types of frames simultaneously in an attempt to take maximum advantage of the relative strengths of each type of frame.

The first of these general types of telephone survey designs uses a directory or other commercial listing of telephone numbers as the sampling frame. The principal disadvantage is that the frame does not provide complete coverage of the telephone household population, so, in a strict sense, inferences regarding the telephone population are impossible regardless of the sample design. This deficiency is often over looked as there are other characteristics of directory/commercial list frames which make them convenient to use and provide opportunities for efficient survey design. Among these characteristics are the following:

- a very high proportion of the numbers on the frame are linked to telephone households,

- the frame tends to be convenient to use especially if it is in machine readable format,

- any auxiliary information (such as name and address) available on the frame can be used for the purposes of efficient sample design, and

- name and address information can also be used to send advance letters for the purpose of improving the response rate.

Even with all of these advantageous attributes, survey designs of this general type are little used by government agencies or research oriented survey organizations because the coverage problem would severely undermine the credibility of most scientific studies.

The second general category of survey design is characterized by the use of frames consisting of all possible telephone numbers generated from valid area code and central office code combinations. From a theoretical standpoint, proper probability based sampling designs applied to these frames will not be vulnerable to coverage deficiencies. Of course a price is paid for this benefit; sample designs in this category tend to be relatively inefficient because there is no auxiliary information available on the frame for the purposes of sample design and a large proportion of the numbers on the frame are not linked to telephone households. The latter problem is usually considered to be of greater importance and it has been rather successfully addressed by the general two stage sampling approach suggested by Waksberg (1978).

In the third general category are telephone survey designs which utilize simultaneously both of the types of frames discussed above. The basic motivation for the development of this type of survey design is the desire to capitalize on the differential strengths of the two frames. Two distinct lines of thought, one stressing sampling methodology and the other estimation methodology, have emerged in the development of these designs:

Sampling methodology – This approach emphasizes the development of innovative (and usually rather complex) sample selection methodology to incorporate the auxiliary information from a directory/list frame into a sample

design for an area/central office code frame. Examples of survey designs based on this approach include Sudman (1973) and the two phase design suggested by Lepkowski and Groves (1986). It should be noted that this approach utilizes standard estimation procedures which are included in most computer software packages.

Estimation methodology – This approach is based on the well known dual frame methodology in which standard sample designs are used to select an independent sample from each type of frame. The incorporation of the strengths of the two frames is accomplished through the optimal allocation of sampling resources to the two frames and the use of a complex estimator of the type suggested by Lund (1968).

The first objective of this paper is to unify the concepts, the definitions, and the notation that have evolved with the development of telephone survey design methodology in the statistical literature. The second objective is to utilize this unified structure to:

(a) Develop some of the more well known telephone survey designs,

(b) Develop alternative survey designs within the third category, and

(c) Discuss the relative merits of the different survey designs with respect to both statistical properties and cost considerations.

2. BASIC CONCEPTS AND DEFINITIONS

2.1 The Target Population

The target population is defined to be the collection of all households, located within some well specified geographic domain, which can be linked to one or more telephone numbers for the purpose of conducting a telephone interview. The symbol $\mathcal{T}$ will be used to denote the target population. In order to avoid a multiplicity problem, which is interesting but not of major importance in the context of this paper, it will be assumed that each household in the target population is linked to one and only one telephone number. This assumption has an added benefit as $\mathcal{T}$ can now be used to represent the target population or the set of telephone numbers linked to the target population or a set of labels assigned to the elements of the target population; the specific usage of $\mathcal{T}$ being clear from the context.

2.2 The Sampling Frames

(A) Directory/Commercial List Frames

The classical directory frame consists of the printed name, address, and telephone numbers of subscribers who live in the geographic area covered by the telephone directory. Lepkowski (1988) discusses two types of commercial list frame: city directories and master address lists. City directories are very similar to telephone directories in that they are printed address lists with a telephone number for all addresses with a telephone. A city directory may include lists ordered both geographically and by telephone number.

Master address lists are nationwide in coverage and are machine readable. Thus, they have the potential to be utilized efficiently for surveys of either widespread or compact geographic populations. Master address lists are constructed primarily from telephone directories but other sources, such as automobile registration lists, are also used to augment the coverage of the list. Some of these lists are continually updated so that the master address list stays within about six to eight weeks of current telephone directory publication (Donnelley Marketing, 1986). Master address lists often contain additional detailed geographic information such as the zip code and Census block data. In some cases they may even include household characteristics such as income and household size imputed by an algorithm based on Census block, tract, and other information.

Viewed from a survey design prospective, directories and commercial lists share many common attributes. Their principal disadvantage is their failure to include some of the telephone households in $\mathcal{T}$. The omissions are primarily due to fact that some telephone service subscribers opt for "unlisted" numbers, however, other numbers are not included because service subscription occurred after the most current publication of the directory or some type of administrative error. Another, less serious, problem is the inadvertent inclusion of some numbers that are not in $\mathcal{T}$. The primary source of these ineligible numbers is when the listed subscriber has canceled service or when the number is linked to a business or other unit that is not a household. Just as they

share the same survey design disadvantages, directories and commercial lists share many of the same advantages. These advantages stem from the fact that they both provide auxiliary information for purposes of survey design. Master address lists have the added advantage of being in machine readable format.

As directories and commercial lists are so similar when considered as frames for telephone surveys, we let the symbol $\mathscr{F}_d$ represent a generic directory/commercial list frame. In the following it is assumed that the frame $\mathscr{F}_d$ has all of the general properties ascribed to such frames in the preceding discussion; specific properties, as required, will be included on a case by case basis.

(B) Area–Prefix Frames

A telephone number consists of ten digits; the first three digits are referred to as the area code and the second three digits are referred to as the prefix. Currently there are nearly 38,000 area–prefix code combinations in use in the United States. The last four digits of the telephone number (referred to as the suffix) identify the individual telephone service subscriber. An area–prefix frame is any list of telephone numbers which is generated by appending all 10,000 suffixes to each six digit number in a specified set of viable area–prefix codes. The complete set of viable area–prefix codes can be obtained from Bell Communications Research (BCR) for a fee ($ 320 in 1988). The BCR list includes new area–prefix codes approximately three months before they are added to the telephone system.

In what follows we will let $\mathscr{F}_a$ denote a general area–prefix frame and we will assume that $\mathscr{F}_a$ was generated by a set of area–prefix numbers from the BCR list. Thus we can assume that every number in $\mathscr{T}$ is included in $\mathscr{F}_a$. We must also assume that $\mathscr{F}_a$ is essentially devoid of any auxiliary information that would be useful for sample design. In fact, area–prefix code boundaries usually do not even correspond to geopolitical boundaries except perhaps at the state level, so extensive use of screening questions may be required for a survey targeted at a specific geographic area. One final note; for the purposes of sample design the telephone numbers in an area–prefix frame are sometimes clustered into sets of 100 consecutive numbers or, less frequently, 1000 consecutive numbers. These sets of numbers are referred to as hundred banks and thousand banks respectively.

(C) Hybrid Frames

A hybrid frame, which we will denote by $\mathscr{F}_h$, is constructed from an existing directory/list frames as follows; using the directory/list frame $\mathscr{F}_d$, all hundred banks containing at least one number in $\mathscr{F}_d$ are identified, then $\mathscr{F}_h$ is formed by including all numbers from the identified hundred banks. Instead of hundred banks, thousand banks or even area–prefix codes can be used to construct hybrid frames. As the name would imply, hybrid frames have some of the characteristics of both directory/list and area–Prefix frames. For example, the hybrid frame includes all of the numbers from the original directory/list frame so auxiliary information is available for some of the numbers in $\mathscr{F}_h$. Furthermore, the hybrid frame will include many of the "unlisted" numbers in $\mathscr{T}$ so, like an area–prefix frame, its coverage may be considered nearly complete. As shall be seen, hybrid frames are utilized extensively for the purposes of telephone survey design.

2.3 Relationships Between the Frames and the Target Population

The general relationships between the telephone numbers linked to the target population and the various types of telephone number frames are summarized below:

1. $\mathscr{T} \subseteq \mathscr{F}_a$
2. $\mathscr{F}_d \subseteq \mathscr{F}_h \subseteq \mathscr{F}_a$
3. $\mathscr{F}_d^c \cap \mathscr{T} \neq \emptyset$ and $\mathscr{F}_d \cap \mathscr{T}^c \neq \emptyset$
4. $\mathscr{F}_h^c \cap \mathscr{T} \neq \emptyset$ and $\mathscr{F}_h \cap \mathscr{T}^c \neq \emptyset$

where all "complements" are relative to $\mathscr{F}_a$.

3. BASIC NOTATION

At this point we assume that we have a specific area–prefix frame $\mathscr{F}_a$, a specific directory/list frame, $\mathscr{F}_d$, which is in machine readable format, and a hybrid frame $\mathscr{F}_h$ generated from $\mathscr{F}_d$. Further, we assume that we have a well defined target population $\mathscr{T}$ and the relationships given in Section 2.3 hold, then we let

M_a = number of hundred banks in $\mathscr{F}_a$,

M_d = the number of hundred banks containing at least one telephone number from $\mathscr{F}_d$,

N_a = number of telephone numbers in $\mathcal{F}_a = 100\ M_a$

N_h = number of telephone numbers in $\mathcal{F}_h \cap \mathcal{T}$, and

N_d = number of telephone numbers in $\mathcal{F}_d$.

The frame $\mathcal{F}_a$ is partitioned into hundred banks by letting $\mathcal{B}_i$ represent the telephone numbers in the i^{th} hundred bank ($i = 1, 2, ..., M_a$) . For convenience we assume that the first M_d hundred banks generate $\mathcal{F}_h$ (i.e. they each contain at least one number from $\mathcal{F}_d$) . For each hundred bank three subsets are of particular interest:

$\mathcal{C}_{i1} = \mathcal{B}_i \cap \mathcal{F}_d \cap \mathcal{T}$
= set of listed target population telephone numbers in the i^{th} hundred bank,

$\mathcal{C}_{i2} = \mathcal{B}_i \cap \mathcal{F}_d^c \cap \mathcal{T}$
= set of unlisted target population telephone numbers in the i^{th} hundred bank, and

$\mathcal{C}_{i3} = \mathcal{B}_i \cap \mathcal{F}_d \cap \mathcal{T}^c$
= set of listed spurious telephone numbers in the i^{th} hundred bank.

The set of listed target population households, say $\mathcal{L}$, can be written as $\mathcal{L} = \mathcal{F}_d \cap \mathcal{T} = \bigcup_{i=1}^{M_d} \mathcal{C}_{i1}$ and the set of unlisted target population households, say $\mathcal{U}$, can be written as $\mathcal{U} = \mathcal{F}_d^c \cap \mathcal{T} = \bigcup_{i=1}^{M_a} \mathcal{C}_{i2}$. We let N_{i1} , N_{i2} , N_{i3} , N_L and N_U be the number of elements in $\mathcal{C}_{i1}$, $\mathcal{C}_{i2}$, $\mathcal{C}_{i3}$, $\mathcal{L}$, and $\mathcal{U}$ respectively. As with $\mathcal{T}$, the elements of these sets can be households, telephone numbers or labels depending on the situation. Finally, we let N_T be the size of the target population and note that

$$N_T = N_L + N_U = \sum_{i=1}^{M_d} N_{i1} + \sum_{i=1}^{M_a} N_{i2}$$

$$= \sum_{i=1}^{M_a} \sum_{j=1}^{2} N_{ij} = \sum_{i=1}^{M_a} N_{i\cdot} \ ,$$

where $N_{i\cdot}$ is the number of target population households in the i^{th} hundred bank.

Now let X be the characteristic of interest, x_{i1k} the observed value of X for the k^{th} listed household in the i^{th} hundred bank, and x_{i2k} the observed value of X for the k^{th} unlisted household in the i^{th} hundred bank, then we can define the following population means:

$$\bar{X}_T = \frac{\sum_{\mathcal{T}} x_{ijk}}{N_T}$$
= mean value of X for all households in $\mathcal{T}$,

$$\bar{X}_L = \frac{\sum_{\mathcal{L}} x_{i1k}}{N_L} =$$ mean for households in $\mathcal{L}$,

$$\bar{X}_U = \frac{\sum_{\mathcal{T}} x_{i2k}}{N_U} =$$ mean for households in $\mathcal{U}$,

and population variances :

$$\sigma_T^2 = \frac{\sum_{\mathcal{T}} (x_{ijk} - \bar{X}_T)^2}{N_T}$$
= variation of X among households in $\mathcal{T}$,

$$\sigma_L^2 = \frac{\sum_{\mathcal{L}} (x_{i1k} - \bar{X}_L)^2}{N_L}$$
= variation of X among households in $\mathcal{L}$,

$$\sigma_U^2 = \frac{\sum_{\mathcal{U}} (x_{i2k} - \bar{X}_U)^2}{N_U}$$
= variation of X among households in $\mathcal{U}$.

Letting $P_L = N_L / N_T$ be the proportion of listed households in $\mathcal{T}$ it follows that

$$\bar{X}_T = P_L \bar{X}_L + (1 - P_L) \bar{X}_U \text{ and}$$

$$\sigma_T^2 = [P_L \sigma_L^2 + (1 - P_L) \sigma_L^2] + P_L (1 - P_L) (\bar{X}_L - \bar{X}_U)^2$$

Further, if the variance among the values of X is approximately the same for the listed and the unlisted households then we have

$$\sigma_T^2 = \sigma_W^2 + P_L (1 - P_L) (\bar{X}_L - \bar{X}_U)^2$$

where $\sigma_W^2 \doteq \sigma_L^2 \doteq \sigma_U^2$.

4. SURVEY DESIGNS FOR THE TELEPHONE HOUSEHOLD POPULATION

4.1 Designs Using $\mathcal{F}_d$ as the Frame

Survey designs utilizing directory/list frames depend to a great extent on specific characteristics of the actual frame and the purpose of the study. It is not the intent of this paper to study any particular directory/list frame so specific design details are not provided in this section. We will simply assume that the ultimate goal of the survey is to provide an estimate of $\bar{X}_T$, that the sample design provides a sample of n households in $\mathcal{F}$, and that some form of sample mean, denoted by $\bar{x}_d$, is the survey based estimator for $\bar{X}_T$. Under these general conditions the statistical characteristics for survey designs based on directory/list frames are as follows:

Bias – Designs based on directory/list frames provide biased estimates of $\bar{X}_T$. Specifically we assume that $E(\bar{x}_d) = \bar{X}_L \neq \bar{X}_T$.

Variance – Designs based on directory/list frames allow for the use of efficient design techniques, such as stratification, so the variance of $\bar{x}_d$ is less than simple random sampling variance. That is, for a sample of n households, we can assume that $\mathrm{Var}(\bar{x}_d) < \frac{\sigma_L^2}{n}$. An alternate statement of this property is that the design effect, denoted by $\delta(\bar{x}_d)$, for directory/list frame designs is less than 1.

Mean square error – Combining these two characteristics, the mean square error of $\bar{x}_d$ is given by

$$\begin{aligned}
\mathrm{MSE}(\bar{x}_d) &= \mathrm{Var}(\bar{x}_d) + (\mathrm{Bias}(\bar{x}_d))^2 \\
&= \delta(\bar{x}_d)\left[\frac{\sigma_L^2}{n}\right] + (\bar{X}_L - \bar{X}_T)^2 \\
&\doteq \delta(\bar{x}_d)\left[\frac{\sigma_W^2}{n}\right] + (1 - P_L)^2 (\bar{X}_L - \bar{X}_U)^2 \\
&< \frac{\sigma_W^2}{n} + (1 - P_L)^2 (\bar{X}_L - \bar{X}_U)^2
\end{aligned}$$

4.2 Designs Using $\mathcal{F}_a$ as the Frame

(A) Single Stage Designs

One simple method for selecting a "with replacement" sample of n households from $\mathcal{F}$ is to randomly select telephone numbers from $\mathcal{F}_a$ (with replacement) until exactly n "hits" in $\mathcal{F}$ have been accumulated. This type of sample design is usually referred to as a simple or single stage RDD design. A through discussion of the single stage RDD design, including several variations, can be found in Lepkowski (1988).

If we let $\bar{x}_s$ be the sample mean for single stage RDD sample of size n , it is straightforward to verify that $\bar{x}_s$ is unbiased for $\bar{X}_T$ and

$$\begin{aligned}
\mathrm{Var}(\bar{x}_s) = \frac{\sigma_T^2}{n} &= \frac{1}{n}\left[P_L \sigma_L^2 + (1 - P_L)\sigma_L^2\right] \\
&\quad + \frac{P_L (1 - P_L)}{n}(\bar{X}_L - \bar{X}_U)^2 \\
&\doteq \frac{\sigma_W^2}{n} + \frac{P_L (1 - P_L)}{n}(\bar{X}_L - \bar{X}_U)^2 .
\end{aligned}$$

(B) Two Stage Designs

The primary problem with the single stage RDD design is its inherent inefficiency in identifying households in $\mathcal{F}$. In fact, in most cases the ratio N_T/N_a is between .20 and .25 , so the sample selected from $\mathcal{F}_a$ must be approximately four to five times as large as the specified sample size n to be selected from $\mathcal{F}$. As the elimination of spurious cases from the $\mathcal{F}_a$ sample is both expensive and time consuming the single stage RDD is rarely used in practice.

In response to this problem several ingenious two stage RDD sample designs have been proposed for improving efficiency. Important papers on this subject include Sudman (1973), Waksberg (1978), and Potthoff (1987). The Sudman procedure uses thousand banks as first stage units and actually provides a sample from a hybrid type frame and not $\mathcal{F}_a$. The Waksberg paper elaborates on a procedure originally proposed by Mitofsky (1970). This particular two stage design, which has come to be known as the Mitofsky–Waksberg design, utilizes hundred banks from an area–prefix frame as the first stage sampling units. Finally, Potthoff developed a general class of two stage RDD designs which includes both the Sudman design and the Mitofsky–Waksberg design as special cases.

All of the designs in Potthoff's general class derive their improved efficiency from the manner in which the telephone company assigns numbers to household subscribers; specifically, numbers are not assigned at random within an area–prefix code, rather they tend to be assigned in blocks or groups of nearly consecutive numbers. Thus a given hundred bank within a working area–prefix code will tend to have either no numbers linked to households in $\mathcal{T}$ or it will have a considerable number (typically 50 or more). So, if only hundred banks with assigned numbers are selected in the first stage, the within hundred bank (or second stage) "hit" rate will tend to be quite high relative to single stage RDD.

We will consider a Mitofsky–Waksberg type design in which a pps random sample (with replacement) of m hundred banks is selected in the first stage and, in the second stage, a simple random sample (with replacement) of b households is selected from the target population of households in the selected hundred bank. More specifically we assume:

(a) on each of the m first stage selections, the probability of selecting the i^{th} hundred bank is $N_i ./ N_T$,

(b) conditional on the i^{th} hundred bank being selected, a simple random sample (with replacement) of size b is selected from $\mathcal{B}_{i1} \cup \mathcal{B}_{i2}$,

(c) the total sample size is given by n = m b .

Letting $\bar{x}_t$ be the sample mean for a household sample selected via the two stage design described above, it is straightforward to verify that $\bar{x}_t$ is unbiased for $\bar{X}_T$ and

$$\mathrm{Var}(\bar{x}_t) = \frac{\sigma_T^2}{n}\left[1 + (b-1)\frac{B_T^2}{\sigma_T^2}\right]$$

$$= \frac{\sigma_T^2}{n}\left[1 + (b-1)\rho_I\right]$$

$$= \frac{\sigma_T^2}{n}\,\delta(\bar{x}_t)$$

where $B_T^2 = \frac{1}{N_T}\sum_{i=1}^{M_a} N_i .\,(\bar{X}_i . - \bar{X}_T)^2$ and ρ_I represents intra–hundred bank correlation .

4.3 Designs Using Both $\mathcal{F}_d$ and $\mathcal{F}_a$

(A) Methods Using the Hybrid Frame

There are many possible ways to incorporate the hybrid frame into sample designs for the telephone population. The particular design presented in this section was motivated by the two phase design proposed by Lepkowski and Groves (1986) and the stratified design proposed by Mohadjer (1988). First we will consider a method for sampling from the sub population $\mathcal{T} \cap \mathcal{F}_h$:

First stage – Select a sample of m hundred banks from the M_d hundred banks in $\mathcal{F}_h$. The sample is selected pps (with replacement) and the measure of size is the total number of listed telephone numbers for the hundred bank. For example, the measure of size for the i^{th} hundred bank is $N_{i1} + N_{i3}$.

Second stage – For each sample hundred bank , select (with replacement) numbers within the bank at the rate of $\frac{b}{N_{i1} + N_{i3}}$ and retain all "hits" from $\mathcal{B}_{i1} \cup \mathcal{B}_{i2}$ in the sample.

This sample design does not produce a fixed sample size but the expected sample size is given by $\frac{m\ b\ N_h}{N_d}$. As the simple sample mean, denoted by $\bar{x}_h$, is a ratio type estimator, it can be shown that $E(\bar{x}_h) \doteq \bar{X}_H \neq \bar{X}_T$ where $\bar{X}_H$ is the mean value of X for households in $\mathcal{F}_h \cap \mathcal{T}$. In general it is safe to assume that the bias of $\bar{x}_h$ is considerably less than the bias of $\bar{x}_d$ because the frame $\mathcal{F}_h$ covers a much larger proportion of $\mathcal{T}$ than does $\mathcal{F}_d$.

Letting σ_H^2 and B_H^2 be defined analogously to σ_T^2 and B_T^2 , except restricted to $\mathcal{F}_h \cap \mathcal{T}$, and assuming that

$$\frac{B_H^2}{\sigma_H^2} \doteq \frac{B_T^2}{\sigma_T^2} = \rho_I \quad , \text{ and}$$

$$\frac{N_{i1} + N_{i3}}{N_d} \doteq \frac{N_{i1} + N_{i2}}{N_T} \equiv \frac{N_i .}{N_T}$$

for i = 1, 2, ..., M_d , it can be shown that

$$\mathrm{Var}(\bar{x}_h) \doteq \frac{\sigma_H^2}{n}\left(1 + (b^* - 1)\rho_I\right)$$

where $b^* = b\,(N_h / N_d)$ and $n = m\,b^*$. Note that n represents expected sample size.

If the bias is considered to be a problem it can be dealt with by selecting an independent sample from $\mathcal{A}_a \cap \mathcal{A}_h^c$ via the Mitofsky–Waksberg method and utilizing a stratified type estimator, say

$$\bar{x}_{st} = \frac{N_h}{N_T}\bar{x}_h + \left(1 - \frac{N_h}{N_T}\right)\bar{x}_{a\text{-}h}$$

where $\bar{x}_{a\text{-}h}$ is the mean for the sample from $\mathcal{A}_h^c \cap \mathcal{S}$. We assume that N_h / N_T is known as the more general case is tedious and not particularly interesting. If we let n_h represent the expected sample size from $\mathcal{A}_h \cap \mathcal{S}$ and $n_{a\text{-}h}$ the sample size from $\mathcal{A}_h^c \cap \mathcal{S}$ then it can be shown that

$$\mathrm{Var}(\bar{x}_{st}) \doteq \left[\frac{N_h}{N_T}\right]^2 \frac{\sigma_H^2}{n_h}\left[1 + (b^* - 1)\rho_I\right] + \left[1 - \frac{N_h}{N_T}\right]^2 \frac{\sigma_T^2}{n_{a\text{-}h}}\left[1 + (b - 1)\rho_I\right].$$

If we assume that $\sigma_H^2 \doteq \sigma_T^2$ and use proportional allocation of sample to the two frames then it follows that

$$\mathrm{Var}(\bar{x}_{st}) \doteq \frac{\sigma_T^2}{n}\left[1 + \frac{N_h}{N_T}(b^* - 1)\rho_I + \left(1 - \frac{N_h}{N_T}\right)(b - 1)\rho_I\right]$$

where n is total (expected) sample size.

(B) Dual Frame Methods

We will assume that we have a sample of size n_d target households selected via a directory/list frame and a sample of n_a target households selected via an area–prefix frame. For the latter sample, we will let $n_{a,l}$ and $n_{a,u}$ be the number of sample households from $\mathcal{L}$ and $\mathcal{U}$ respectively. In addition, we let $\bar{x}_{a,l}$ and $\bar{x}_{a,u}$ be the corresponding sample means. A dual frame estimator for $\bar{X}_T$ is

$$\bar{x}_{dual} = (1 - P_L)\,\bar{x}_{a,u} + P_L\left[\lambda\,\bar{x}_{a,l} + (1 - \lambda)\,\bar{x}_d\right]$$

where, following Lund (1968), we set

$$\lambda = \frac{\mathrm{Var}(\bar{x}_d)}{\mathrm{Var}(\bar{x}_{a,l} \mid n_{a,l}) + \mathrm{Var}(\bar{x}_d)}.$$

It is straightforward to show that

$$\mathrm{Var}(\bar{x}_{dual}) \doteq \sigma_W^2\left((1 - P_L)^2\frac{\delta(\bar{x}_{a,u})}{n_{a,u}} + P_L^2\left[\frac{n_{a,l}}{\delta(\bar{x}_{a,l})} + \frac{n_d}{\delta(\bar{x}_d)}\right]^{-1}\right)$$

but unfortunately no further simplification is possible without some rather strong assumptions.

5. RELATIVE MERITS OF THE SURVEY DESIGN

5.1 General Considerations

Before preceding further it is necessary to consider the relative sizes of the various frames and some of the key sub sets defined in the earlier sections. The values given below are based on consideration of information from several papers including Waksberg (1978); Landenberger, Groves, and Lepkowski (1984); Traugott, Groves, and Lepkowski (1987); Lepkowski (1988); and Tucker (1989). Clearly we can only consider the stated values as approximations; based on further information we may have to modify the approximations , which may in turn lead to modification of our conclusions.

The proportion of the telephone population covered by a directory or commercial list frame can vary considerably depending on such factors as geographic location, proportion of the target population in urban areas, etc. . For the United States as a whole the proportion appears to be between .60 and .70 , therefore, we will set $P_L = .65$. Although there may be exceptions, the omission of approximately 35% of the target population from the frame would, in general, preclude the exclusive use of a directory/list based estimator $\bar{x}_d$.

For a hybrid frame the situation is less clear. It seems that in most cases it is safe to assume that $\frac{N_h}{N_T}$ is very close to 1. In fact we will assume that $\frac{N_h}{N_T} = .95$ but in most cases it is probably even larger. This level of coverage seems to be tolerable in most cases. For

example nearly all Federal population based sample surveys utilize either Census based and/or area frames which fail to cover approximately 10% of the target population. Furthermore, as the non–response rate is typically 20% to 30% in telephone based surveys, possible bias due to non–response must be considered a much more serious problem. On balance the coverage problem alone does not preclude the use of the estimator $\bar{x}_h$.

The following frame and population parameters will determine the relative cost and efficiency of the proposed design/estimation strategies. First we will assume that $\pi = \frac{N_T}{N_a} = .20$ so obviously the single stage strategy for sampling for $\mathcal{A}_a$ will not be very efficient. Next we assume that 65% of the hundred banks in $\mathcal{A}_a$ do not contain any telephone numbers in $\mathcal{A}$ (or in Waksberg's notation t = .65). These two assumptions combine to imply that (on average) there are approximately 57 target population households in each hundred bank containing target households. It is this concentration of target households in a relatively small proportion of the hundred banks that makes the two stage design more efficient than the single stage design.

Finally we let θ represent the proportion of listings in $\mathcal{A}_d$ that are linked to target population households so that

$$\theta = \frac{\sum_{i=1}^{M_d} N_{i\,1}}{N_d}$$

and we assume that $\theta = .85$. This, together with the fact that we are assuming that $P_L = .65$ and $\frac{N_h}{N_T} = .95$, then implies that $\frac{N_h}{N_d} = \frac{.95 \times .85}{.65} = 1.24$.

5.2 Cost Considerations

Sample designs utilizing frames of the type described above result in one of two kinds of telephone call; either a productive call to a target population household or an unproductive call. Following Waksberg (1978) we let C_p be the cost of a productive call and C_u the cost of an unproductive call. Waksberg shows that the expected cost of a two stage estimator based on a sample of size n = m b is given by

$$EC(\bar{x}_t) = m\,b\left[C_p + \frac{1-\pi-t}{\pi}C_u\right] + \frac{m\,t}{\pi}C_u$$

and it is straightforward to show that the expected cost for a single stage estimator is

$$EC(\bar{x}_s) = n\left[C_p + \frac{1-\pi}{\pi}C_u\right] .$$

For a directory/list frame it can be shown that

$$EC(\bar{x}_d) = n\left[C_p + \frac{1-\theta}{\theta}C_u\right] + C_D$$

where C_D is the (fixed) cost of purchasing the directory/list and writing the programs to select the sample. Actually there are certain fixed costs associated with the single stage and two stage designs which are not shown in the expected cost functions; thus C_D should be considered as <u>additional</u> fixed cost.

For the hybrid frame estimator the situation is somewhat different because the sample of hundred banks is not selected via a telephone call. Rather the sample of hundred banks is selected via a programed algorithm based on the relative size of the hundred banks in terms of listed numbers. We assume there is an (additional) fixed cost, say C_H , to purchase the size information but it is much less than the cost of purchasing an entire commercial list. The cost of writing the programs to select the first stage sample is also included in C_H and we assume the cost of actually selecting the first stage sample is negligible. The expected cost of the hybrid frame estimator is

$$EC(\bar{x}_h) = m\,b\,\frac{N_h}{N_d}\left[C_p + \frac{1-\gamma}{\gamma}C_u\right] + C_H$$

where γ is the proportion of numbers in $\mathcal{A}_h$ that are linked to target households (note that $\gamma = \frac{N_h}{100\ M_d} \doteq .57$) and the expected sample size is $n = (m\ b\ N_h)/N_d$. It follows that in the case of the stratified type estimator the expected cost is given by

$$EC(\bar{x}_{st}) = n_h\left[C_p + \frac{1-\gamma}{\gamma}C_u\right] + C_H$$

$$+ m_{a\text{–}h}\,b_{a\text{–}h}\left[C_p + \frac{1-\pi^*-t^*}{\pi^*}C_u\right] + \frac{m_{a\text{–}h}\,t^*}{\pi^*}C_u .$$

where t^* and π^* are defined on $\mathcal{A}_a \cap \mathcal{A}_h^c$ so that $t^* \doteq .974$ and $\pi^* \doteq .015$.

The expected cost of the dual frame estimator is given by

$$EC(\bar{x}_{dual}) = n_d \left[C_p + \frac{1-\theta}{\theta} C_u \right] + C_D$$

$$+ m_a b_a \left[C_p + \frac{1-\pi-t}{\pi} C_u \right] + \frac{m_a t}{\pi} C_u \ ,$$

where $n = n_d + n_a$ and $n_a = m_a b_a$.

5.3 Fixed Cost Efficiency

For the purposes of an illustrative example we will let C^* represent the total resources available for the survey and assume that $\rho_I = .05$ and $\frac{C_p}{C_u} = 2$. It can then be shown that

$$Var(\bar{x}_s) \doteq 6 \frac{C_u}{C^*} \sigma_T^2$$

$$Var(\bar{x}_t) \doteq 4.08 \frac{C_u}{C^*} \sigma_T^2 \quad (\text{where } b = 5)$$

$$Var(\bar{x}_h) \doteq 2.75 \left[1 - \frac{C_H}{C^*} \right]^{-1} \frac{C_u}{C^*} \sigma_T^2 \quad (\text{using } b^* = 1)$$

$$Var(\bar{x}_{st}) \doteq 3.05 \left[1 - \frac{C_H}{C^*} \right]^{-1} \frac{C_u}{C^*} \sigma_T^2$$

$$(\text{where } n_h = .95\, n \text{ and } b_{a\text{-}h} = 21)$$

$$Var(\bar{x}_{dual}) \doteq 3.30 \left[1 - \frac{C_D}{C^*} \right]^{-1} \frac{C_u}{C^*} \sigma_T^2$$

$$(\text{where } n_d = .3\, n \ , \ \delta(\bar{x}_d) = 1 \ ,$$
$$\delta(\bar{x}_{a,u}) = \delta(\bar{x}_{a,l}) = (1 + (b_a/2 - 1)\rho_I) \ , \ b_a = 7)$$

Of course we do not want to generalize too much from this one example but it appears that if the coverage of $\mathcal{A}_h$ is adequate and the cost C_H is small relative to C^* then the estimator $\bar{x}_h$ will be difficult to beat. Assuming that the strategy of selecting several samples from one sample of hundred banks reduces the cost of $\bar{x}_t$ by 20 to 40% (see Waksberg, 1978) then it will compete strongly with $\bar{x}_h$.

NOTE

Any opinions expressed are those of the author and do not reflect policy of the Bureau of Labor Statistics.

REFERENCES

Donnelley Marketing, *Donnelley Marketing Advantages*, Stamford, CT, R.H. Donnelley Corp., 1986.

Landenberger, B.D., Groves R.M., and Lepkowski, J.M., "A Comparison of Listed and Randomly Dialed Telephone Numbers," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1984, pp. 280–284.

Lepkowski, J.M., "Telephone Sampling Methods in the United States," in R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, and J. Waksberg (eds.), *Telephone Survey Methodology*, John Wiley & Sons, 1988, pp. 73–98.

Lepkowski, J.M., and Groves R.M., "A Two Phase Probability Proportional to Size Design for Telephone Sampling," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1986, pp. 357–362.

Lund, R.E., "Estimators in Multiple Frame Surveys," *Proceedings of the Social Statistics Section, American Statistical Association*, 1968, pp. 282–288.

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

Mohadjer, L., "Telephone Sampling Methods in the United States," in R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, and J. Waksberg (eds.), *Telephone Survey Methodology*, John Wiley & Sons, 1988, pp. 161–173.

Potthoff, R.F., "Generalizations of the Mitofsky–Waksberg Technique for Random Digit Dialing," *Journal of the American Statistical Association*, Vol.82, No.398, June 1987, pp. 409–418.

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

Traugott, M.W., Groves, R.M., and Lepkowski J.M., "Using Dual Frame Designs to Reduce Nonresponse in Telephone Surveys," *Public Opinion Quarterly*, Vol. 51, No. 4, Winter 1987, pp. 522–539.

Tucker, C., "Characteristics of Commercial Residential Telephone Lists and Dual Frame Designs," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1989, to appear.

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