A Dual Frame Sampling Design for an RDD Survey that Screens for a Rare Population

K.P. Srinath[1], Michael P. Battaglia[1], Meena Khare[2]
[1]Abt Associates Inc. [2]National Center for Health Statistics, CDC

1. Introduction

Several large surveys for government agencies are conducted by telephone using random-digit dialing (RDD) samples. Examples include the Behavior Risk Factors Surveillance System, the National Immunization Survey (NIS), and the State and Local Area Integrated Telephone Surveys. Some of these surveys focus on specific target populations because the subject matter of the survey relates to specific segments of the population. Examples might include persons age 65 years and older, children aged 0-17 years, persons with asthma, persons with special health care needs, etc. For some of these potential target populations, the percentage of households that are eligible may be fairly low. In such surveys, screening a large RDD sample of households to identify a sample of eligible households is required.

The calling of a large RDD sample of telephone numbers involves considerable effort. Screening of households can take a large pool of interviewers and a large amount of time adding a substantial cost for conducting the survey. The screening cost is driven by how rare the target population is and how large a sample of eligible households is needed in the final sample.

There are several different sampling designs to reduce data collection costs when sampling a rare population. Kalton and Anderson (1986) suggest oversampling strata in which the rare population is concentrated for estimating the characteristics of the rare population. Srinath (2002) considered allocations to strata that minimize the screening sample size and also the loss in precision due to oversampling. Other design options include oversampling residential directory listed telephone numbers to yield a sample that contains a higher proportion of residential telephone numbers, increasing the working bank threshold in a list-assisted RDD sample design to reduce the proportion of nonworking numbers in the sample, network (adaptive) sampling, and dual frame sampling. For example, Brick et al. (2002) suggest stratification of telephone numbers into two strata with different working residential number rates using a two-phase design. The dual frame approach is starting to receive more attention for use in RDD surveys. In the simplest application of dual frame sampling, one has a complete frame of households but the eligible households are not identified in the frame. The second frame only includes households that belong to the target population but that frame does not cover the entire target population. The basic idea behind dual frame sampling is to draw a sample from each of the two frames. The survey is conducted with each sample and weights are developed for each sample. Information related to the overlap of the complete frame with the partial frame is then used to develop weights that allow the two samples to be used together to provide unbiased estimates. This is an important aspect of dual frame sampling, because some eligible households are present only in the complete frame while others are present in both the complete frame and the partial frame and therefore have a multiple chance of selection. In Section 2 of the paper, we describe the National Immunization Survey and examine the possibility of using the dual frame design. In Section 3, we give the sample allocation to the two frames that minimizes the screening costs for a given variance of the estimate. In the last section, we provide results of the application of the design to the NIS.

2. Dual Frame Design in the NIS

The NIS is a large ongoing RDD survey conducted quarterly by the CDC. It measures the vaccination coverage rates among children aged 19 to 35 months. The percentage of known households with a child between 19 and 35 months is below 4% of the households in the U.S. This means that we require 25 or more sample households to identify one eligible household. The effects of nonresponse at the screening stage and at the interviewing stage increase the sample size of households that needs to be screened even further. Also, an RDD sample will contain out-of-scope telephone numbers such as nonworking and nonresidential (business) numbers. Also, the presence of unresolved telephone numbers which are defined as numbers where it is never determined whether the numbers are residential, business, or nonworking further increase the number of telephone numbers that need to be called.

In the NIS, for sample selection, the U.S. is stratified into 78 strata consisting of entire states, urban areas (cities or counties) and rest-of-state areas (e.g., Boston and Massachusetts-Rest of State are two strata). The strata are known as Immunization Action Plan (IAP) Areas. Each quarter, a list-assisted RDD sample is drawn from each stratum. The list-assisted sample design excludes banks of 100 consecutive telephone numbers that contain zero residential directory-listed telephone numbers (referred to as the zero banks). The RDD sample for each quarter is divided into subsamples (replicates) for sample release and management purposes. The sample is released on a controlled basis because each IAP area has a target number of interview per quarter. This target averages around 110 household interviews per IAP area. The first step in the dialing process is to determine if the number is a known household. Telephone numbers that are residential are then screened to determine if the household contains one or more children aged 19-35 months. If the household is eligible, an interview is attempted for each child with the person in the household that is most knowledgeable of the child's vaccinations. As noted above, nonresponse occurs at three steps in this process. First, some telephone numbers are never resolved. Some of the unresolved telephone numbers belong to households that contain age eligible children. Second, some of the known households do not complete the screening interview to determine age eligibility. Some of these households contain age eligible children. Third, some of the identified eligible households do not complete the immunization interview. It is also possible that some age eligible households will indicate

that they do not have any children (soft refusals) and will screen themselves out. All of these factors lead to a large initial RDD sample size for the NIS. Therefore, there is considerable interest in reducing the number of RDD screening calls made in the NIS. In 2002, 30,974 household interviews were completed. The initial RDD sample size required was 3,361,396 telephone numbers. A total of 2,055,371 numbers were dialed by interviewers, yielding an eligibility rate or take rate of 1.5%. The eligibility rate is defined as the number of completed interviews divided by the number of telephone numbers dialed.

During 2003, an investigation of potential lists for a dual frame design was made. The basic idea was to maintain the RDD sample component and to add a list sample to the design. The RDD frame covers the entire target population except nontelephone households and households in the zero banks. The list frame (with telephone numbers) would offer partial coverage of eligible telephone households and also offer coverage of a portion of the eligible telephone households in the zero banks. List frames can however be very incomplete, that is, they only cover a small portion of the target population. Some list frames may offer better coverage but the quality of the information of the list may be out-of-date. This is a special concern when one is going to use the telephone numbers contained in the list frame. Working with the Marketing Systems Group, we examined the various potential commercial lists covering households with young children. We identified the Experian New Babies list, which covers households with young children as the most promising list. The Experian New Babies list is targeted to companies that want to market to households with young children. It contains name and address information, which we could use to mail advance letters. The list is constructed from birth records and other sources on a state-by-state basis. A portion of the list also includes telephone numbers. The list is kept current by using information from various sources. Though a list like the New Babies list can vary in coverage and quality from state to state, it is the best list source for a dual frame design that we were able to identify. Based on some initial expectations of list quality we expected that the eligibility rate for the list would be much higher than the RDD eligibility rate. Also, some initial state level counts of households with telephone numbers on the list indicated that coverage of the target population could be as high as 30 to 40 percent.

There are considerable risks in implementing a sample design modification to an ongoing survey. It was decided to initially test the dual frame design in five urban IAP areas. The urban IAP areas in the NIS tend to have the lowest eligibility rates. This test was implemented in Q4/2003. The RDD sample was selected for the five urban IAP areas in the usual fashion. We worked with GENESYS Sampling Systems to select a simple random sample of telephone numbers in each of the five strata from the list of telephone numbers on the Experian New Babies list. We did a telephone number match of the RDD sample and the Experian New Babies sample. Any duplicate telephone numbers were retained in the RDD sample. The final disposition of each of the duplicate cases in the sample was then transferred over the to Experian sample (this was done to avoid calling a telephone number as if it were two separate pieces of sample). To develop weights for a dual frame design, one also needs to determine the overlap between the RDD sample (assumed to come from the completed frame) and the Experian New Babies list frame (assumed to cover a portion of the target population). This was accomplished by taking the RDD sample for each stratum and matching that sample of telephone numbers against the entire Experian sampling frame (of age eligible households with telephone numbers) for that stratum.

3. Sample Allocation to the Two Frames

We have two sampling frames A and B listing telephone numbers. Let M_a telephone numbers belong to frame A only, M_b telephone numbers belong to frame B only and M_{ab} telephone numbers to both frames A and B. Because the NIS is an RDD survey, frame A has 100 per cent coverage as it consists of all the telephone numbers available for selection in an IAP area or a state and therefore $M_{ab} = M_b$. Frame B is a subset of frame A and consists of telephone numbers on the Experian list of households with a child aged 19 to 35 months.

Let $M = M_a + M_{ab}$ denote the total number of telephone numbers available for selection in frame A. M_a and $M_b = M_{ab}$ are known. Let N (unknown) denote the number of telephone numbers of households with children between 19 and 35 months in frame A. Let $e = \frac{N}{M}$ denote the eligibility rate in the main frame. Similarly, let $N_b = N_{ab}$ denote the residential telephone numbers with children between 19 and 35 months in frame B. Let $e_b = \frac{N_b}{M_b}$ denote the eligibility rate in frame B. e_b will be less than one due to errors in the Experian frame. However, we expect e_b to be much bigger than e. Let $\alpha = \frac{M_b}{M}$ be the proportion of telephone numbers covered by the Experian list out of the total population of numbers in frame A. We have $(1-\alpha) = \frac{M_a}{M}$. Let N_a denote the number of households with children between 19 and 35 months out of the M_a telephone numbers in Frame A and not on the Experian list. Let $e_a = \frac{N_a}{M_a}$ be the eligibility rate for the part of the frame not on the Experian list. We have $e = (1-\alpha)e_a + \alpha e_b$.

Assume that we select a simple random sample of m telephone numbers from frame A and a simple random sample of m_b telephone numbers from frame B. The total sample selected is $m_0 = m + m_b$. Let n denote the number of households with children between 19 and 35 months out of the m selected telephone numbers. We will assume that there is one child between 19 and 35 months per household. Let n_b denote the number of households with children from frame B resulting from

calling m_b telephone numbers. The total number of children selected is $n_0 = n + n_b$.

We match the sample of telephone numbers of n eligible households with the telephone numbers contained in the entire Experian list. Let n_{ab} numbers selected from frame A match with the numbers on the Experian list. n_{ab} will also contain some numbers appearing in both the samples. Let $n_a = n - n_{ab}$ denote the households with numbers not matching with the numbers on the Experian list. We therefore have three samples. A sample of size n_a belonging to frame A and not to B, a sample of size n_{ab} selected from frame A but identified as belonging to B after matching, and a sample of size n_b selected independently from frame B. We want to use all the three samples to estimate the vaccination coverage rate. We use the Hartley (1962) type design-based ratio estimator (Cochran, 1977) given below.

$$\hat{R} = \frac{\hat{Y}_a + p\hat{Y}_{ab} + (1-p)\hat{Y}_b}{\hat{N}_a + p\hat{N}_{ab} + (1-p)\hat{N}_b} \qquad (1)$$

where $\hat{Y}_a$ is the estimated number of children who are up-to-date for a specific vaccine or vaccine series based on the sample from frame A not belonging to B, $\hat{Y}_{ab}$ is the estimated number of children who are up-to-date based on sample from frame A belonging to B, and $\hat{Y}_b$ is the estimated number of children who are up-to-date based on the sample from frame B. Both $\hat{Y}_{ab}$ and $\hat{Y}_b$ are estimates of the number of children who are up-to-date in Frame B. The estimates in the denominator are the estimated number of children between 19 and 35 months based on the three samples. Both $\hat{N}_{ab}$ and $\hat{N}_b$ are estimates of N_b, the total number of children in frame B. p and $(1-p)$ are the weighting factors to combine the estimates relating to frame B. We can find the values of $p, m,$ and m_b such that for a given m_0, the conditional variance of $\hat{R}$ is minimized.

The totals in the numerator and the denominator of $\hat{R}$ and the variance of $\hat{R}$ can be estimated using domain weights. We have $\hat{Y}_a = \frac{M}{m} y_a$ where y_a is the number of children who are up-to-date out of n_a children in frame A not belonging to B. $\hat{Y}_a$ can be written as

$$\hat{Y}_a = \frac{M}{m} n_a r_a$$

where r_a is the proportion of children who are up-to-date out of n_a children. Similarly,

$$\hat{Y}_{ab} = \frac{M}{m} n_{ab} r_{ab}$$

where r_{ab} is the estimated proportion of children who are up-to-date based on the sample of n_{ab} children and

$$\frac{M_b}{m_b} n_b r_b$$

where r_b is the estimated proportion of children who are up-to-date based on the sample n_b children. The estimated total number of children in frame A not belonging to B is $\hat{N}_a = \frac{M}{m} n_a$. The estimated total number of children in frame A belonging to B is $\hat{N}_{ab} = \frac{M}{m} n_{ab}$. The estimated total number of children in frame B is $\hat{N}_b = \frac{M_b}{m_b} n_b$.

Variance of $\hat{R}$

We will look at the conditional variance of $\hat{R}$ assuming that n, the number of children sampled from frame A and n_b, the number of children sampled from Frame B are fixed in repeated sampling and replace these values by their expected values (Des Raj and Promod Chandhok, 1998). These are $E(n) = m\frac{N}{M} = me$ and $E(n_b) = m_b \frac{N_B}{M_B} = m_b e_b$.

We also have

$$E(n_a) = m\frac{N_a}{M} = m\frac{N_a}{M_a}\frac{M_a}{M} = me_a(1-\alpha)$$

and

$$E(n_{ab}) = m\frac{N_{ab}}{M} = m\frac{N_{ab}}{M_{ab}}\frac{M_{ab}}{M} = me_b\alpha .$$

Since n is fixed, we can write $n_b = n - n_a$. That is, $\hat{N}_b = N - \hat{N}_a$. Therefore we can write $\hat{R}$ as

$$\hat{R} = \frac{\hat{Y}_a + p\hat{Y}_{ab} + (1-p)\hat{Y}_b}{\hat{N}_a(1-p) + p\hat{N} + (1-p)\hat{N}_b} . \qquad (2)$$

We determine the variance of $\hat{R}$ keeping $\hat{N}$ and $\hat{N}_b$ fixed.

We can show that the variance of $\hat{R}$ is

$$V(\hat{R}) = \frac{1}{N^2}[\frac{M^2}{m} R(1-R)\{e_a(1-\alpha) + p^2 e_b \alpha\} + \frac{M_b^2}{m_b}(1-p)^2 e_b \alpha R(1-R)] \quad (3)$$

We want to determine sample sizes m, m_b and p to minimize $V(\hat{R})$ for a specified screening sample size $m_0 = m + m_b$. The values of p and the allocation that minimizes the variance are given by

$$p = \frac{\sqrt{e_a}}{\sqrt{e_b}} \quad (4)$$

.

$$m_b = \frac{m_0 \alpha(\sqrt{e_b} - \sqrt{e_a})}{\alpha(\sqrt{e_b} - \sqrt{e_a}) + \sqrt{e_a}} \quad (5)$$

and

$$m = \frac{m_0 \sqrt{e_a}}{\alpha(\sqrt{e_b} - \sqrt{e_a}) + \sqrt{e_a}} \quad (6)$$

If we want the expected sample number of children in the sample to be n, then we want

$$m_0 = n \frac{\alpha(\sqrt{e_b} - \sqrt{e_a}) + \sqrt{e_a}}{e_b \alpha(\sqrt{e_b} - \sqrt{e_a}) + e\sqrt{e_a}} \quad (7)$$

Note that the value of p depends only on e_a and e_b and the allocation depends only on

α, e_a and e_b.

4. Application to the NIS

In this section we describe the dual frame approach adopted on a trial basis in the NIS using the results derived in Section 3. As indicated in Section 2, we selected 5 urban areas for implementing this design. We give some results of the implementation of the dual frame design in these areas. Frame A described in Section 3 is the RDD sampling frame in each IAP area. Frame B is the list of telephone numbers contained in the Experian new Babies list in each IAP area. Frame B is a subset of frame A. For implementing the dual frame design, we need to know the eligibility rates both for the RDD frame and the Experian frame. We estimated the eligibility rate for the RDD frame for each of the 5 IAP areas using the NIS data from the previous quarters. A single eligibility rate of 32% was assumed for the Experian frame for all the 5 IAP areas based on a previous national test sample of 500 telephone numbers from the Experian list. Based on the frame data, we were also able to get the value of α which is the proportion of the Experian frame in the RDD frame for each IAP area. Table 1 shows the sizes of the two frames and the actual value of α after the implementation of the design.

Table 2 shows the sample size selected from the RDD frame and Experian frame using the allocation formulas given above and the estimated eligibility rates from previous surveys. It also shows the actual number of completes. From Table 2, we see that large portion of the total sample is allocated to the RDD frame, as the proportion of the Experian frame out of the total is small.

Table 3 shows the actual eligibility rates in the two frames after data collection. It also shows the value of the weighting factor p to combine the two estimates from Frame B.

As seen from Table 3, the observed Experian eligibility rates were less than the assumed eligibility rates of 32% for all IAP areas. Table 4 shows the required sample size under a single RDD frame design to get the same number of total completes assuming the observed overall eligibility rate. We compare this with total sample size under the dual frame design. There is a loss in precision due to the use of dual frame design as compared to the single RDD frame design due the use of unequal weights. Table 4 also shows the square root of the design effect (DEFT) due to the use of unequal weights because of two samples from the two frames.

Conclusions

Dual-frame designs can offer savings during data collection. Since there is some loss in precision due to the use of unequal weights, we need to balance the savings in screening costs with the possible increase in variance. That is, if the estimates from the dual frame design have a large design effect (DEFF), then it is not desirable as the savings in the screening sample size is offset by a relatively larger loss in precision of the estimates. In such cases, it may be desirable to go with a single frame design. Also, it is important that we are able to match the sample from the RDD frame with the entire Experian frame to be able to develop composite weights to combine the two estimates for the list

frame. In the NIS, the Experience New Babies List seems to provide a useful partial list frame for implementing the dual frame design. For planning the allocation, we need estimates of eligibility rates and the proportion of the Experian frame out of the total RDD frame in all IAP areas.

References

Brick , J.M., Judkins, D., Montaquila, J., and Morganstein, D. (2002) "Two-Phase List-Assisted RDD Sampling" Journal of Official Statistics, Vol. 18, No.2, pp.203-215.

Cochran, W.G. 1977. Sampling Techniques. 3rd Ed. New York: Wiley

Des Raj and Promod Chandhok 1998. Sample Survey Theory. Narosa Publishing House.

Hartley , H.O.(1962) Multiple Frame Surveys. Proceedings of the Social Statistics Section, American Statistical Association. 203-206.

Kalton , G and Anderson, D. (1986), "Sampling Rare Populations", Journal of the Royal Statistical Society, series A.149 pp. 65-82.

Srinath, K.P. (2002), "Allocation to Strata in Surveys that Screen for Rare Populations", Proceedings of the International Conference on Recent Advances in Survey Sampling, Ottawa, Canada, July 2002

Table 1: Sizes of the Two Frames in the 5 IAP Areas

IAP Area	Total Number of RDD Numbers (M)	Telephone Numbers in the Experian Frame	Proportion of Experian Frame (α) (%)
1	608,400	986	0.162
2	1,604,500	3,875	0.241
3	699,300	2,356	0.337
4	937,000	4,479	0.478
5	457,200	1,739	0.380

Table 2: Allocation of the Sample to the Two Frames

IAP Area	Sample Size RDD	Sample Size Experian	Number of Completes: RDD	Number of Completes: Experian
1	6,720	168	86	40
2	6,411	162	75	35
3	4,600	190	69	39
4	3,759	77	65	22
5	6,195	122	70	35

Table 3: Observed Eligibility Rates

IAP Area	Overall Eligibility Rate (%) (e)	Experian Eligibility Rates (%) (e_b)	$p = \frac{\sqrt{e_a}}{\sqrt{e_b}}$
1	1.28	23.8	0.232
2	1.17	21.6	0.232
3	1.50	20.6	0.270
4	1.73	28.7	0.245
5	1.13	28.8	0.198

Table 4: Comparison of Sample Sizes for the Single and Dual Frame Designs

IAP Area	Sample Size Single Frame	Sample Size Dual Frame	Difference	Expected Design Effect
1	9,844	6,888	2,956	1.14
2	9,402	6,573	2,829	1.14
3	7,200	4,790	2,600	1.13
4	5,029	3,836	1,194	1.11
5	9,292	6,317	2,975	1.14