A COMPARATIVE STUDY OF TWO TELEPHONE SAMPLING DESIGNS

Manas Chattopadhyay, Rajesh Srinivasan, Elaine Christiansen, Greg Gaertner, The Gallup Org.

Manas Chattopadhyay, The Gallup Organization, One Church St, 9th Flr., Rockville, MD 20850

Key Words: RDD, Optimal allocation, cost, hit rate

1. Introduction

As the use of telephone interviewing for household surveys has grown, telephone sampling methods have increased in diversity. There are basically two major types of sampling frames used for telephone sampling. One is list or directory based where lists of published telephone numbers are used as sampling frames. An unrestricted random selection of telephone numbers from such frames yields a significantly higher rate (at least 85 percent) of working residential numbers. However, samples drawn from such lists do not include unpublished telephone numbers and studies (Bruner and Bruner, 1971) of telephone households with and without published numbers suggest that estimates based on such samples may be biased. The other type of sampling frame is the Bell Core Research (BCR) frame of all possible telephone numbers generated by appending all 10,000 four-digit suffixes to the area code prefix combinations. Random sampling from this frame consisting of both listed and unlisted numbers is referred to as Random Digit Dialing (RDD). This method, though simple, is not operationally efficient since about 80 percent of all numbers in an exchange either are not assigned or are assigned to business and other non-household telephones.

The two stage RDD telephone sampling design proposed by Mitofsky (1970) and later fully developed by Waksberg (1978) has been widely used in telephone sample of U.S. household population. This method is based on the fact that working residential numbers tend to cluster within banks of consecutive telephone numbers. While the percentage of working residential numbers is only 20 percent in RDD sampling, the corresponding percentage in a bank of 100 consecutive telephone numbers containing at least one working residential number is around 50 percent. The Mitofsky-Waksberg method uses two stages. The first stage consists of identifying clusters of 100 consecutive numbers each having at least one working residential number. In the second stage, for the sample to be self-weighting a fixed number of residential numbers must be contacted. This design greatly reduces the amount of unproductive calls thereby reducing the cost of the survey. However, in practice the sequential nature of the design under fixed field period and the increase in variance due to clustering of these numbers outweigh the benefits of fewer unproductive calls.

Casady and Lepkowski (1993) reviewed several designs which use listed number information to increase the efficiency of random digit dialing. The telephone number frame is stratified using information from published lists of telephone numbers. The high-density stratum is composed of all clusters of 100 consecutive numbers that contain x or more (x, the threshold point, can be changed) listed numbers and a low-density stratum or residual stratum composed of all remaining clusters of 100 consecutive numbers. Due to significant differences in the percentage of working residential numbers between the two strata (about 50 percent for high-density and 2 percent for the low-density with x=1), there will be a large cost difference in obtaining samples from the two strata. In the Casady-Lepkowski procedure, this cost difference is exploited by optimally allocating the sample between the strata.

There are two main objectives to this paper. First to derive estimates of frame related (z_i and h_i) and cost related parameters (c_o, c_1, γ) based on the results of our study and compare those values with estimates presented in Casady and Lepkowski (1993). Second, to examing the effectiveness of implementating the Mitofsky-Waksberg and the Casady-Lepkowski designs in terms of cost, coverage and ease of implementation in a practical setting.

In order to facilitate the comparison, The Gallup Organization commissioned a nationwide study to measure the prevalence of smoking among the US adult population using both the designs. Before describing the study design, we first provide a general description of the two sampling procedures as they were implemented.

2. Alternate Sample Designs

Mitofsky-Waksberg:

The first stage consists of obtaining clusters of telephone numbers defined as sets of numbers with the same first 8 digits of a 10 digit number for the entire study area. Next, a randomly-generated 100-bank is selected and a single complete telephone number is generated by appending a two digit randomly selected number to the first 8 digits. This number is then dialed and if it turns out to be a residence, the 100-bank is retained and an interview with the selected number is attempted. This process continues until we have the desired number of 100-banks (or primary sampling units) that would yield our planned sample size. The second stage involves sampling of 10 digit telephone numbers at random within the retained 100-banks and these numbers are dialed until a certain number (say k) of residences are contacted for the design to be self-weighting. In our study, a slightly different version of the procedure was used. Instead of continuing the second stage sampling until a specified number of residences is obtained, the sampling was actually continued until a fixed number (equal to 3) of completed interviews were obtained. Under this scheme, the probability of selection of households becomes unequal. However, if the results are analyzed without compensating weights for unequal probabilities, an implicit two step adjustment is made simultaneously for unequal probability of selection and non-response separately within each prefix area (James M. Lepkowski, 1988).

Casady-Lepkowski:

This design stratifies the Bellcore(BCR)[1] frame into two strata: a "high density" stratum consisting of 100-banks with one or more listed residential numbers and a "low density" stratum consisting of all the remaining numbers in the BCR frame. About 52 percent of the numbers in the high-density stratum are estimated to be working residential numbers whereas in the low-density stratum, the corresponding percentage is only about 2 percent. The Casady-Lepkowski procedure, exploits the significant difference in the cost of sampling between the two strata by optimally determining the sample size in each stratum. In the truncated version of the procedure, sampling is done only from the high-density stratum.

In this study, a stratified simple random sampling design was used i.e., simple RDD samples were drawn from each stratum. Following Casady and Lepkowski (1993), the optimal sample allocation was calculated using the relation:

$$m_i \propto (z_i/\sqrt{h_i})\,(1 + (\gamma - 1)h_i)^{-1\backslash 2} \qquad \text{............(1)}$$

where

m_i: number of households contacted/dialed in stratum i = 1,2

z_i: proportion of the target population included in stratum i = 1,2

h_i: proportion of telephone numbers that are WRNs (i.e., hit rate) for stratum i = 1,2

γ: cost-ratio = $(c_1 + c_2)/c_0$

c_0: cost of determining that a telephone number is not a WRN

c_1: cost of determining that a telephone number is a WRN

c_2: cost of completing an interview of a WRN.

So c_1+c_2 is the total cost of a productive call and c_0 is the cost of an unproductive call and γ is the ratio of these two costs. Let n_1 and n_2 be the effective sample sizes i.e. the number of interviews to be completed in stratum 1 (high stratum) and stratum 2 (low stratum) respectively. Hence $n_1=m_1h_1$ and $n_2=m_2h_2$. We need to know h_i, z_i (i=1,2) and γ to be able to determine n_1 and n_2 using (1). Based on our preliminary estimates of cost, γ was estimated to be equal to 10. For h_i and z_i, the values as reported in Casady and Lepkowski (1993) were used to determine the sample allocation. A key ingredient of this process is the stratification of 100-banks into high density and low density strata. Information required to create the strata were obtained from Survey Sampling Inc (SSI). SSI utilizes the Donnelley Listed Household Database to determine the number of listed numbers in each 100-bank. The frame was purchased in March of 1994.

3. Background

As part of Gallup's mission to conduct research and disseminate data on important public health problems, this study was commissioned to examine the prevalence of smoking, attitudes towards smoking and smoking behavior in general and particularly among the adult US population. Gallup completed a total of 3,019 interviews using the two different telephone sampling methodologies. Two teams of interviewers at two different locations were employed for this project, and each team administered both methodologies to complete approximately half the number of interviews. As an additional protection to maintain complete independence of sampling and administration, while one team worked towards completing one "half" of the Mitofsky-Waksberg sample the other team worked on the first "half" of the Casady-Lepkowski interviews. When one team had completed their interviews, they were required to wait

[1] The Bellcore frame used in this study contains only categories of exchanges that have been set aside for residential usage and excludes those such as 800 numbers, exclusively cellular phones etc. The exclusion of these numbers is done for both designs.

until the other team had completed their portion before switching to the other methodology. Although this contributed to the rather long field period, it was deemed important more as a safeguard due to the experimental nature of the design. Each team was comprised of Gallup's most experienced interviewers and once formed, no additions to the team were permitted. In addition, no interviewer left the team during the course of the survey.

A total of 1,494 interviews were completed using the Mitofsky-Waksberg method. The sample of the primary stage units was drawn and then screened by Survey Sampling Inc. for non working banks. The total sample (including non working banks) were randomly "scrambled" and then sequentially numbered. The sample was then divided into two sub samples, one for each interviewing team. Known non working banks (as determined by SSI) were marked and did not actually go through a primary stage telephone call. A total of 498 working banks were identified and 3 interviews were completed in each one of these banks.

A total of 1,525 interviews were completed using the Casady-Lepkowski method. Based on our preliminary estimate of $\gamma = 10$ and using the values of z_i and h_i as reported in Casady and Lepkowski (1993), the optimal sample allocation for a total sample of size 1,525 is 1,484 for the high-density and 41 for the low-density stratum. The actual number of completes in the two strata in our study was 1,473 and 52 respectively. The allocation however, tends to be "flat" in the neighborhood of optimal allocation (Cassady & Lepkowski, 1993). After all interviewing was completed, the Mitofsky-Waksberg portion of the sample was matched with the Bell Core tape to determine strata assignment. The final dataset of 3019 interviews contained 2901 from the high-density stratum and 118 from the low-density stratum.

Selection of a respondent within the household was accomplished using a random selection grid. The interviewer obtained a list of adult household members and the CATI software selected one member to be interviewed with equal probability. The sample data were weighted to compose estimates for population parameters. The basic weight attached to every sampled unit was the inverse of the selection probability. The number of eligible adults and the number of residential phone lines in a household were taken into consideration in the calculation of selection probabilities. In the Casady-Lepkowski procedure, necessary weighting was done to correct for disproportional allocation. Gallup also used post-stratification weighting to make the final sample reflect the population it is intended to represent. For the purpose of post-stratification weighting, current estimates of the 1990 census data were used. Data was collected between June 1994 and March 1995.

4. Results

Estimation of frame and cost parameters:

Based on our study, the frame related parameters P_i (proportion of frame included in the stratum), z_i (proportion of the target population included in the stratum) and h_i (proportion of the telephone numbers in the stratum that are working residential number i.e., the hit rate) were estimated. Table 1 below provides that information along with the corresponding values reported in Casady and Lepkowski (1993).

Table 1: Estimated values of frame parameters for a two-stratum design

Stratum	Proportion of frame (P_i)	Proportion of population (z_i)	Hit rate (h_i)
High-density	.3617 (.3804)*	.9311 (.9402)	.5298 (.5210)
Low-density	.6383 (.6196)	.0689 (.0598)	.0222 (.0204)

*figures in parentheses are corresponding estimates reported in Casady & Lepkowski (1993).

No significant difference is observed between the two sets of estimates. In this study, the total number of telephone numbers sampled in the high and low stratum were 6730 and 6050 respectively under the Casady-Lepkowski methodology. The working rates were obtained by multiplying the SSI screening rate which = numbers identified as working/total number sampled, and the field working rate = (working -busy-no answer)/(used-no answer-busy). In the Mitofsky-Waksberg sample, the hit rate in the second stage of sampling was 49 percent. The hit rate in the first stage of sampling, however, was much lower (18.6 percent).

Following Casady & Lepkowski (1993), we used a cost function of the following type:

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

where d=1 if the number is a working residential number (WRN) and d=0 otherwise. Hence c_1 is the cost of determining that a number is a WRN and c_0 is the cost of an unproductive call. This allows for the possibility of differential cost of determining a WRN as opposed to a number that is not working. Although in most cases the cost of determining the status of a WRN is usually less, due to the unique location and set up of our phone facility, we found that in this study our estimate of c_1 was slightly higher than c_0. Since this was a multipurpose survey, c_0 and c_1 was further amortized over the cost of collecting responses to the substantive variables. Our estimates were as follows: c_0 =1.6 cents, c_1 = 5 cents and c_2 = 17 cents, where c_2 was the additional cost of determining the value of a characteristic after identifying whether it was WRN or not. Hence $\gamma = (c_0 + c_1)/c_0 = 13.75$. Using our initial estimate of γ =10 before the start of the study and the values of h_i and z_i as reported in Casady & Lepkowski (1993), the optimal allocation required a sample size of 1,484 to the high-density and 41 to the low-density stratum. Using estimates of γ = 13.75, h_i and z_i based on the results of our study, the corresponding optimal allocation would be 1,470 and 55 in the high and low density stratum respectively.

Similar cost figures for the Waksberg design though not very precise suggest the advantages to using a list assisted design. Screening costs were fairly high in the first stage of the Waksberg design. The complexity of the two stage design also necessitated the use of incentives for interviewers to be productive which increased interviewing costs. In effect the total cost of conducting the same study using the Waksberg design was roughly 2.5 times what it cost using the list-assisted design. Assuming the cost of purchasing the frame can be amortized over several such studies, this is a substantial cost savings for any organization. Further, if the estimates of substantive variables are statistically insignificant, then the choice is quite clear.

Estimators for the telephone survey on smoking:

In order to examine the effect of particular telephone sampling methodology on survey variables in our study, we compared the estimates of smoking prevalence under both methods. The question asked to obtain data on smoking prevalence was: "Have you, yourself, smoked any cigarettes in the past week?" Table 2 below provides estimates of smoking prevalence (i.e. proportion of smokers) for both Casady-Lepkowski and Mitofsky-Waksberg samples.

Table 2: Estimates of Smoking Prevalence

Strata (Design)	Sample size unweighted	Prevalence Estimates Overall	Male	Female
High density	1,473	24.46% (1.27)*	25.92% (1.93)	23.14% (1.68)
Low density	52	28.63% (7.23)	26.71% (10.07)	30.51% (10.40)
Overall (Casady-Lepkowski)	1,525	24.84% (1.33)	26.00% (2.00)	23.79% (1.79)
Overall (Mitofsky-Waksberg)	1,494	24.82 (1.24)	23.47% (1.78)	26.03% (1.68)

* figures in parentheses are standar errors of the corresponding estimates. The standard errors were calculated using the software SUDAAN.

Between estimates obtained from the two different strata of the Casady-Lepkowski design, we notice, as expected, that the estimate based on the low-density stratum has a much larger standard error. However, it has relatively less influence on the overall estimate which largely depends on the estimate based on the high-density stratum. Similar pattern is observed for males and females separately with the exception of the Waksberg design where qualitatively the smoking prevalence is higher among women. The estimates and standard error of the estimates under the two sampling methods are found to be quite close. Restricting attention to the high-density stratum only, we also notice significant agreement between the results of the truncated Casady-Lepkowski design and the Mitofsky-Waksberg design.

Table 3: Estimates of Smoking Prevalence by Subgroup

Subgroup	Prevalence Estimates Overall	C-L	M-W
Age			
18-34	28.07%	27.85%	28.20%
35-54	28.01%	27.89%	28.26%
55+	16.92%	17.34%	16.39%
Gender			
Male	24.89%	26.00%	23.47%
Female	24.79%	23.79%	26.03%
Race			
White	25.36%	25.33%	25.39%
Non-White	21.62%	21.80%	21.26%
Ethnicitiy			
Hispanic	23.04%	26.12%	19.17%
Non-hispanic	25.01%	24.72%	25.36%

Table 3 presents smoking prevalence estimates for few other important demographic subgroups under both designs. The findings are more or less consistent with what we observed in Table 2. In general smoking prevalence among young adults in general, men, white and non-hispanic in particular is higher than for their counterparts.

Effect of ignoring the low-density stratum:

In the Casady-Lepkowski design, sampling from the lower stratum is very expensive because of an extremely low hit rate (about 2 percent). A high price is paid to ensure the unbiasedness of the estimator. In the truncated Casady-Lepkowski design consisting of just two strata (high and low), sampling is done only from the high-density stratum. By doing so, about 6 percent of the population remain out of scope. However, there is significant gains in terms of cost and efficiency. The decision, therefore, regarding the choice of a full or truncated design in a specific problem will depend on the relative importance of efficiency and coverage. In order to examine the effect of ignoring the lower stratum on estimates of smoking prevalence, we tested the null hypothesis of eqality of strata means (or proportions). The observed value of the t-statistic was -0.568 with an approximate significance level of 0.2843 indicating no statistically significant difference. We also conducted similar tests based on the combined sample of size 3019 consisting of 2091 cases in the high-density stratum and 118 cases in the low-density stratum. The observed value of the t-statistic in that test was -0.3471 with a significance level of 0.3632 implying no statistically significant difference. In this context, we may note that the optimal allocation calculations in the Casady-Lepkowski design were carried out under simplifying assumptions about the variable specific parameters. The assumptions are equivalent to equality of strata means and strata variances (i.e. $\mu_i = \mu$ and $\sigma_i = \sigma$ for all strata i). In case of proportions (smoking prevalence), equality of means will imply equality of variances and we noticed no significant differences in strata means.

In a multipurpose survey, however, there will be several variables of interest and the effect of ignoring the low-density stratum may be quite different for different variables. Table 4 provides estimates of some important variables for both high and low density stratum from the Casady-Lepkowski sample. The purpose of this table is to illustrate the lack of significant differences among selected subgroups observed in this study. This is not surprising given the small sample sizes in the lower stratum as we break the sample into its sub-domains.

Table 4. Distribution of some key Demographics by Strata

Subgroup	High	Low
Age		
18-34	33.41%	42.82%
35-54	37.11%	37.94%
55+	29.48%	19.24%
Gender		
Male	47.2%	51.46%
Female	52.8%	48.54%
Race		
White	86.43%	81.63%
Non-White	13.57%	18.37%
Ethnicitiy		
Hispanic	8.75%	8.13%
Non-hispanic	91.25%	91.87%

The low-density stratum appears to be quite similar to the high-density stratum in terms of most of the variables. Any characterization of the low-density stratum in terms of these variables for the purpose of distinguishing it from the high-density stratum is not feasible with such small sample sizes.

5. Concluding remarks:

Both the telephone sampling methods proved effective in improving the overall hit-rate as compared to a simple RDD. A comparison between the two methods suggests that they seem to produce, at least in this study, comparable estimates and standard errors of estimates. In terms of our cost estimates, the Casady-Lepkowski design was found to be more economical. This was particularly true, as expected, in case of the truncated Casady-Lepkowski design. However the decision regarding ignoring the low-density stratum for the purpose of sampling will depend, in a specific study, on several factors including survey variables of interest, target subgroups of population etc. No general conclusions on this issue could be drawn in light of a small sample size in the lower stratum. In terms of operational efficiency, the Casady-Lepkowski design was easier to implement. Interviewers found this design easier to work with. In the Mitofsky-Waksberg design, it was necessary in some cases to make significant number of calls to complete three interviews in a primary sampling unit. Although this study was commissioned by Gallup the sequential nature of this design posed some problems particularly in cases where the wait was much longer than expected. This resulted in an increased field period. In this study, the Casady-Lepkowski design, on the whole, appeared to be more cost-efficient and easier to work with as compared to Mitofsky-Waksberg design. It must be mentioned, however, that we examined only those versions of the methods that were implemented in this study. Other possible versions of both methods, discussed in detail in Casady and Lepkowski (1993) were not evaluated. A choice between the methods (and possible variations of them) in other settings will certainly be influenced by other considerations and hence no general conclusions can be drawn based on our results.

Acknowledgements

This study was supported by The Gallup Organization. However,the findings and opinions expressed in this paper are those of the authors and do not necessarily represent those of The Gallup Organization.

References:

BRUNER, J.A., and BRUNER,G.A.(1971). Are voluntarily unlisted telephone subscribers really different? Journal of Marketing Research, 8, 121-124.

MITOFSKY,W.(1970). Sampling of telephone households. Unpublished CBS Newsmemorandum, 1970.

ROBERT,J.CASADY and JAMES,M.LEPKOWSKI (1993). Stratified Telephone Survey Designs. Survey Methodology, 19, 103-113.

WAKSBERG,J. (1978). Sampling methods for random digit dialing. Journal of the American Statistical Association,73,40-46.