AN INVESTIGATION OF PSU CUTOFF POINTS FOR A RANDOM DIGIT DIALING SURVEY

Carma R. Hogue and David W. Chapman, Bureau of the Census

1. Introduction

Because of rising survey costs and Federal Government budget cuts in recent years, there has been increasing interest at the Census Bureau in additional use of telephone interviews. In particular, the National Center for Health Statistics (NCHS) and the Census Bureau are investigating the possibility of using dual-frame sampling for the National Health Interview Survey (NHIS), in which the primary frame would be a telephone frame based on an AT&T tape file of the first six digits (area code-exchange prefixes) of all telephone numbers in the U.S. The secondary frame would be an area frame needed to cover the approximately 7 percent of U.S. residences that do not have telephones.

From January to May of 1984 the Census Bureau, in conjunction with NCHS, conducted a feasibility study to determine and possibly solve some of the problems that exist in adopting random digit dialing (RDD) as one of the sampling frames for the NHIS. For this study a two-stage sampling procedure--described by Waksberg (1978)--was used to select telephone numbers. On occasion a primary sampling unit (PSU) is selected at the first stage that contains relatively few residential numbers. This type of PSU is referred to as "sparse." This paper examines procedures for cutting off sampling in sparse PSUs when staff time and associated costs needed to continue sampling become intolerable.

The background for developing the sampling method and the proposed cutoff procedure is given in Section 2. The methodology that was used to determine the cutoff points is discussed in Section 3. The resulting cutoff points are presented in Section 4. In Section 5, the effect of the cutoff procedure on the weight is given. The conclusion and proposals for future research are contained in Section 6.

2. Background

The Waksberg RDD method of selection is a two-stage sample which produces an equal probability sample of phone numbers. The primary sampling units (PSUs), which are banks of 100 telephone numbers, are selected in the first stage with probability proportional to the number of residential telephone numbers within a PSU. At the second stage a fixed number of residential phone numbers is selected from each sample PSU. Each PSU (i.e., bank of 100 telephone numbers) is identified by the first eight digits of a tendigit telephone number. The 100 numbers in the PSU consist of all ten-digit numbers that can be generated by adding two digits to the specific eight-digit prefix. Telephone companies assign telephone numbers to residential and nonresidential customers in such a way that a high proportion of PSUs (100-banks) are exclusively or largely largely nonresidential. With Waksberg's procedure sampling is carried out primarily in those PSUs that contain large numbers of residential phone numbers. Specifically, PSUs are selected one at a time as follows. A PSU is initially selected by adding a random pair of digits to a working area code-exchange prefix selected at random from the six-digit prefixes on the AT&T tape.[1] Another random pair of digits is selected and added to the eight-digit prefix to form a ten-digit telephone number. This phone number is called and screened as either being residential or nonresidential. If it is residential, the PSU is referred to as residential and is retained for the sample. If the number is nonresidential, the PSU is not selected. The process of selecting an initial PSU, adding two random digits to the eight-digit prefix defining the PSU, and screening the randomly selected phone number for residential/nonresidential status is repeated until a residential PSU is selected. This process, referred to as primary screening, is repeated until the desired number of residential PSUs is selected. Within each PSU selected for the sample, telephone numbers are randomly selected and called until some fixed number, k, of residential numbers is identified. The within-PSU selection is referred to as secondary screening.

The sample size for the feasibility study was 3024 residential units. There were twelve replicates completed over a 3-month period. Each replicate was interviewed for 3 weeks with new replicates being introduced each week. Each of the twelve replicates consisted of 21 PSUs. From these 100-banks, interviewers attempted to interview twelve residential units.

The advantage of the Waksberg RDD method as compared to unrestricted RDD sampling is that sample selections are made only from residential PSUs in the Waksberg procedure. It has been determined from various RDD studies that about 63 percent of the telephone numbers in residential PSUs are residential; whereas, only about 20 percent or so of the telephone numbers in all PSUs are residential. Consequently, for a given sample size, the Waksberg method will require considerably fewer calls than will an unrestricted RDD method.

Even though the percentage of residential numbers in successfully screened (residential) PSUs averages about 63 percent, some PSUs with relatively few residential numbers pass primary screening. A "sparse" PSU is defined as one for which the proportion, P, of residential numbers is less than or equal to some threshold value, p^*. Though it need not be the case, a reasonable choice of the threshold value of P used to define a sparse PSU is the number, k, of telephone numbers to be selected from a PSU, divided by 100. This choice is reasonable because if the proportion of residential numbers in a PSU is less than k/100, there will not be enough residential numbers in the PSU to provide the target sample size k, even if all 100 numbers are called. Taking $p^* = .12$ for the feasibility study, more sparse PSUs than anticipated turned up in several of the PSUs. Telephone calls to sparse PSUs are time-

consuming and are not cost-effective. Therefore, a study of cutoff points was initiated in order to identify sparse PSU's and terminate calling the PSU before all 100 numbers are called.

There are two cutoff points of interest: 1) a cutoff point for calling the primary screening number again to find out if it was correctly identified as a residential unit and 2) a cutoff point for terminating calling in the PSU. Both of these cutoff points were investigated in this study and proposed rules have been developed for them.

3. Methodology for Determining Cutoff Points

Cutoff points were determined by calculating the probability of having a sparse PSU (i.e., one for which the proportion of residential numbers is p* or less) given the number of telephone numbers (n) in the PSU that have already been resolved, and the number of residential units, x, that were found in those n cases[2]. In this study, cutoff points were determined for p*=.04, .06, .08, .10, .12, .16, and .20, or in other words, 4, 6, 8, 10, 12, 16, or 20 residential units out of 100 units. The probability that the proportion of residential telephone numbers is less than or equal to p, given the number of residentials found in the randomly selected telephone numbers already called and resolved is obtained using Bayes's Theorem for conditional probability[3], as follows:

$$\Pr(P \le p \mid x,n)$$

$$= \Pr(P \le p \text{ and } x \text{ residences are observed in } n \text{ phone numbers}) / \Pr(\text{observing } x \text{ residences in } n \text{ phone numbers})$$

$$= \frac{\left[\sum_{M=x}^{100p} \Pr(P=M/100 \text{ and } x \text{ residences are observed in } n \text{ phone numbers})\right]}{\left[\sum_{M=x}^{100} \Pr(P=M/100 \text{ and } x \text{ residences are observed in } n \text{ phone numbers})\right]}$$

$$= \frac{\left[\sum_{M=x}^{100p} \Pr(x \text{ residences are observed in } n \text{ phone numbers} \mid P=M/100)\ \Pr(P=M/100)\right]}{\left[\sum_{M=x}^{100} \Pr(x \text{ residences are found in } n \text{ phone numbers} \mid P=M/100)\ \Pr(P=M/100)\right]}, \quad (1)$$

where M = number of residences in a PSU.

The probability of selecting x residences from n phone numbers in a PSU, given the proportion of residential numbers in a PSU, has a hypergeometric distribution and may be written as follows:

$$\Pr(x \text{ residences are selected in } n \text{ phone numbers} \mid P=M/100) = {}_{M}C_{x}\ (100-M)C(n-x) / {}_{100}C_{n}, \quad (2)$$

where ${}_{M}C_{x} = M!/(M-x)!x!$

Substituting the result from equation (2) into equation (1) the final expression is obtained for computing the probability that a PSU is sparse:

$$\Pr(P \le p \mid x,n) = \frac{\left[\sum_{M=x}^{100p} {}_{M}C_{x}\ (100-M)C(n-x)\ \Pr(P=M/100) / {}_{100}C_{n}\right]}{\left[\sum_{M=x}^{100-n+x} {}_{M}C_{x}\ (100-M)C(n-x)\ \Pr(P=M/100) / {}_{100}C_{n}\right]}$$

$$= \frac{\left[\sum_{M=x}^{100p} {}_{M}C_{x}\ (100-M)C(n-x)\ \Pr(P=M/100)\right]}{\sum_{M=x}^{100-n+x} {}_{M}C_{x}\ (100-M)C(n-x)\ \Pr(P=M/100)}. \quad (3)$$

[The upper limit of 100-n+x, rather than 100, is required in the summation in the denominator of equation (3), because if M were allowed to exceed 100-n+x, the number of nonresidential numbers in the sample, n-x, would exceed the total number of nonresidential numbers in the PSU, 100-M.]

The calculation from equation (3) that a PSU is sparse required knowledge of, or an approximation to, the probability distribution of (M/100), the proportion of telephone numbers in a residential PSU that are residential. An approximation to this probability distribution was developed, based on some data given by Groves and Kahn (1979) on p. 337, along with the knowledge that $E(P) \doteq .63$. The data provided by Groves and Kahn consist of the numbers of phone numbers that had to be called in order to obtain 9 residential units in each of 104 residential PSUs. Since the estimate of E(P) is .74 for this set of data, the distribution based on the 104 PSUs had to be somewhat modified in order to provide appropriate estimates for NHIS/RDD. In addition to shifting the mean from .74 given in Groves and Kahn to .63 for NHIS/RDD, the distribution was smoothed. To simplify programming, three linear functions were used to approximate the probability distribution, as follows:

$$\Pr(P=M/100) = \begin{cases} (M+2)/4800 & \text{if } M=4,\ldots 52, \\ (25M-46)/100{,}000 & \text{if } M=53,\ldots 77, \\ (115-M)/2000 & \text{if } M=78,\ldots 100, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 1 is a graph of this probability mass function. The distribution will be updated when more data become available.

Using this distribution, the expected value of P, i.e., the expected proportion of residences in a PSU, is

$$E(P) = \sum_{i=4}^{52} (i/100)(i+2)/4800 + \sum_{i=53}^{77} (i/100)(25i-46)/100{,}000 + \sum_{i=78}^{100} (i/100)(115-i)/2000$$

$$= .627.$$

This expectation is consistent with the estimates of E(P) derived from other RDD surveys. The standard deviation of P for this distribution is .225.

4. Determination of Cutoff Points

In order to determine the two cutoff points (for verifying the primary screening number and for curtailing calling in the PSU), a FORTRAN program was written, based on equation (3), to calculate $Pr(P<p^*|x,n)$, i.e., that a PSU is sparse. Values of x, the number of residential units found in calling the PSU, were 0,1,2, ..., 10. [For the NHIS/RDD Feasibility study, 12 residences were needed from each PSU; but cutoff points for x>10 were so high that the entire PSU could easily be called to try to find the remainder of the necessary residential units.] Values of n, the number of resolved telephone cases in the PSU, ran from x+1 to $100(1-p^*)+x$. The bounds on n force the probability to be between 0 and 1. The values of p^*, the threshold proportion of residential units, were .04, .06, .08, .10, .12, .16, and .20.

The cutoff point for calling the primary screening number again was established as the value of n for which $Pr(P<p^*|x,n) \geq .50$. These cutoff points are given in Table 1. For example, if only 2 residential telephone numbers have been found out of 24 resolved (business, non-working number, etc.) cases, the primary screening number would be called again to determine if the PSU was correctly identified as a residential unit. In cases where a sparse PSU is defined as having 16 or fewer out of 100 residences, 2 residentials out of 19 resolved cases would instigate a second call to the primary screening number. If the PSU was incorrectly identified as a residential PSU, a replacement "residential" PSU would be selected immediately. Phoning in the PSU would continue if the PSU screening number was correctly identified as a residence.

Three cutoff points for terminating calling in the PSU are presented in the tables. These points are for $Pr(P<p^*|x,n)$ greater than or equal to .8, .9, and .95. The selection of a cutoff procedure will depend upon the amount of risk that one is willing to take of continuing to call numbers in a PSU that is sparse and of terminating calling in a PSU that is not "sparse." (These risks are analogous to probability of type I and II errors in hypothesis testing.)

For a sample size of 12 units per PSU, cutoff points for a probability of .9 and a threshold proportion of residential units of .12 seem reasonable. Cutoff points for a probability of .8 are displayed in Table 2; those for .9 are shown in Table 3; and those for .95 are in Table 4. As an example of how to use Table 3, assuming that a sparse PSU is one with 12 or fewer residential numbers, the fortieth resolved telephone number would result in termination of calling in the PSU if only 2 residences had been found. All previously called, unresolved cases would be called until they were determined to be business, residential, non-working, etc. but no other numbers in the PSU would be called. Therefore, the PSU will have fewer than the required number of residential units and consequently the variance of survey estimates will be higher.

It is possible that resolution of the unresolved cases would produce more residential units which would raise the cutoff limit. Since calling the primary screening number again would not interrupt calling in the PSU, this cutoff rule could be implemented without considering the status of the unresolved cases in the PSU. On the other hand, the status of the unresolved numbers at the time that the cutoff point is reached for curtailing sampling in a PSU is considerably more important. Typically, nonworking telephone numbers and businesses would be easier to resolve than residential phone numbers. Therefore, it would be possible to accumulate several out-of-scope cases prior to resolving some residential cases. Consequently, if a cutoff point for truncating calling in a PSU is reached, based on the resolved cases in the PSU, the decision to discontinue sampling in the PSU should be considered tentative. Once the unresolved cases have been classified, the decision should be reevaluated. In particular, the probability of the PSU being sparse should be computed, if feasible, from equation (3), using the additional resolved telephone numbers. Based on this probability and on other factors, such as the time remaining in the interview period, an updated decision should be made regarding the continuation of sampling in the PSU.

5. Effect of the Cutoff Procedure on the Weights

For survey estimation purposes a weight is generally assigned to each sample unit in a survey. Though often adjusted to account for nonresponse and to incorporate ratio estimation, this weight is basically the inverse of the selection probability. The Waksberg RDD sampling method has been designed so that all residential telephone numbers in the country have a uniform probability of selection. This uniform selection probability, p, which is derived in the appendix, is the following:

$$p = m''k \,/\, 10000\ M, \qquad (4)$$

where m'' = the number of six-digit prefixes on the AT&T tape that were selected for the sample and used to obtain the desired number of PSUs,

k = the target cluster size (12 for the feasibility study),

M = the total number of in-scope six-digit prefixes on the AT&T tape.

The uniform weight, w, for each case, excluding any ratio and nonresponse adjustments, is the inverse of equation (4):

$$w = 10000\ M / m''k. \quad (5)$$

For the case in which a sampling cutoff is applied to a given PSU, a slight modification of the selection probability and assigned weight is needed for all residences selected in that PSU. Specifically, if only k_i (less than k) residences are selected from PSU i because of the cutoff procedure, the selection probability, p_i, for each residence selected in PSU i is

$$p_i = m''k_i / 10000\ M. \quad (6)$$

Consequently, the appropriate weight, w_i, to assign each sample residence in PSU i is the inverse of equation (6):

$$w_i = 10000\ M / m''k_i. \quad (7)$$

Upon comparing the weights given in equations (5) and (7), it is evident that the basic uniform selection weight of each residence selected from PSU i has to be multiplied by the factor k/k_i if sampling is curtailed in PSU i after only k_i of the desired k residences are reached.

6. Conclusion

The cutoff point procedures were not implemented in the 1984 NHIS/RDD feasibility study because the computer program for the call-scheduling procedures could not be altered as needed before the last replicate of the survey was finished. However, cutoff procedures will be used for other Census Bureau RDD surveys.

One aspect of future research in this area will focus on obtaining better estimates of the probability distribution of P, the proportion of residential units in residential PSUs. Data from the NHIS/RDD feasibility study will be examined to see the effect on the cut-off points of providing a better approximation to the probability distribution of P. Another aspect of future research involves the investigation of alternate approaches to determining cutoff points. Specifically, an attempt will be made to associate a cost saving and variance increase with each cutoff rule. This would allow the development of an "optimum" cutoff procedure. Also, sequential testing methods will be examined to determine if they could be easily applied and if they would reduce the number of telephone calls that are needed to classify the PSU as acceptable or as sparse.

In summary, we have derived two cutoff point procedures: 1) a cutoff point for recalling the primary screening number and 2) a cutoff point for terminating calling in the PSU. These points were determined by calculating the probability of a given PSU being a sparse PSU after observing a certain number of residences, x, in n resolved cases.

ACKNOWLEDGEMENT

The authors wish to thank Dr. Paul Biemer of the Bureau of the Census for his initial suggestions regarding the use of cutoff procedures and for his helpful comments on earlier drafts.

FOOTNOTES

[1]Actually, some six-digit prefixes on the AT&T tape that are obviously nonresidential (e.g., long-distance information numbers) are removed from the frame prior to sampling.

[2]The methodology developed in this section is based on the assumption that all 100 numbers in a PSU are available for calling. In most surveys, however, including the feasibility study, the primary screening number is not allowed to be called again for interview. The assumption that all 100 numbers are available simplifies the presentation substantially and has only a trivial impact on the cutoff points derived. Specifically, each is the same or one number higher than the corresponding cutoff point based on the availability of 99 numbers. Consequently, the 100-number assumption provides somewhat conservative cutoff points.

[3]Bayes's Theorem is given in many texts on probability theory and methods. See, for example, Parzen (1960), p. 119.

REFERENCES

Groves, Robert M. and Kahn, Robert L. (1979), Surveys by Telephone, New York: Academic Press.

Parzen, Emanuel (1960), Modern Probability Theory and Its Applications, New York: John Wiley and Sons, Inc.

Waksberg, Joseph (1978), "Sampling Methods for Random Digit Dialing," Journal of the American Statistical Association, 73, 40-46.

Table 1: Cutoff Points for Calling the Primary Screening Number Again (Probability of .5)

No. of Residentials (x)	Values of the Threshold Probability (p*)						
	.04	.06	.08	.10	.12	.16	.20
0	54	26	18	16	11	8	6
1	63	36	26	23	17	13	11
2	71	46	35	32	24	19	16
3	80	56	44	40	31	24	20
4	***	67	54	49	38	30	25
5	***	80	64	58	46	35	30
6	***	***	74	67	53	41	35
7	***	***	84	76	60	47	40
8	***	***	***	86	67	52	45
9	***	***	***	95	74	58	49
10	***	***	***	***	82	63	54

Table 2: Cutoff Points for Terminating Calling in the PSU for a Probability of .8

No. of Residentials (x)	Values of the Threshold Probability (p*)						
	.04	.06	.08	.10	.12	.16	.20
0	80	45	32	28	20	15	12
1	84	54	40	36	27	20	17
2	88	62	31	44	34	27	23
3	92	71	57	52	42	33	28
4	***	80	66	61	49	38	33
5	***	90	75	70	56	44	38
6	***	***	84	78	63	50	43
7	***	***	92	86	70	56	48
8	***	***	***	93	76	61	53
9	***	***	***	99	83	66	58
10	***	***	***	***	89	72	62

Table 3: Cutoff Points for Terminating Calling in the PSU for a Probability of .9

No. of Residentials (x)	Values of the Threshold Probability (p*)						
	.04	.06	.08	.10	.12	.16	.20
0	88	56	40	35	26	19	16
1	91	63	48	43	33	25	21
2	93	70	56	51	40	31	27
3	96	77	64	59	47	37	32
4	***	85	72	67	54	43	37
5	***	93	80	75	61	49	42
6	***	***	88	82	68	55	47
7	***	***	95	89	75	60	52
8	***	***	***	95	81	66	57
9	***	***	***	100	86	71	62
10	***	***	***	***	92	76	66

Table 4: Cutoff Points for Terminating Calling in the PSU for a Probability of .95

No. of Residentials (x)	Values of the Threshold Probability (p*)						
	.04	.06	.08	.10	.12	.16	.20
0	92	64	48	42	31	23	19
1	94	70	55	49	38	29	25
2	96	76	62	56	45	35	30
3	98	82	69	64	52	41	36
4	***	88	76	71	59	47	41
5	***	95	84	79	66	53	46
6	***	***	91	86	72	59	51
7	***	***	96	92	78	64	56
8	***	***	***	97	84	69	61
9	***	***	***	100	89	74	65
10	***	***	***	***	94	79	70

FIGURE 1: PROBABILITY DISTRIBUTION OF THE NUMBER OF RESIDENTIAL PHONE NUMBERS IN A PSU