THE DESIGN AND SELECTION OF A SAMPLE FOR THE BUREAU OF THE CENSUS RANDOM DIGIT DIALING EXPERIMENT

Mary Mulry-Liggan and David W. Chapman, Bureau of the Census

Introduction

The Random Digit Dialing (RDD) Research Project has been initiated by the Bureau of the Census to investigate the feasibility and cost-effectiveness of conducting a nationwide household survey by telephone. The advantages of collecting data by telephone instead of by personal interview, which include reduced transportation costs, quicker completion of many survey operations, and the ability to monitor interviewers more closely, have prompted the Bureau to explore using RDD sampling. The project has the following goals include the establishment of a centralized RDD telephone sampling and interviewing facility and the estimation of national household and person response rates for a "typical" Bureau survey using RDD methods.

The core of the RDD Research Project is a telephone survey, the RDD Employment and Health Survey. The questionnaire for the RDD Employment and Health Survey has been designed to include questions asked in current Bureau surveys. Included are questions on health and employment similar to those asked in the Health Interview Survey and the Current Population Survey.

The sampling design for the RDD Employment and Health Survey is based on a two-stage method initially developed by Warren Mitofsky of CBS News (Mitofsky, 1970) and further analyzed by Joseph Waksberg of Westat, Inc. (Waksberg, 1978).

With this method the primary sampling units (PSUs) are selected with probability proportional to size (the measure being the number of residential telephone numbers in the PSU), even though advance knowledge or estimates of the measure of size are not generally available. Furthermore, a self-weighting sample is obtained by selecting an equal number of secondary units, i.e., k residential numbers, from each PSU. Sampling PSUs with probability proportional to size (PPS) generally provides sample PSUs that contain relatively high proportions of residential numbers. This should reduce costs and improve sampling efficiency as compared to a completely random, unclustered RDD method.

The probability of response may vary by the region of the U.S. in which the sample unit is located and by whether it is in an urban, suburban or rural environment. Therefore, the PSUs will be sorted into categories based on geography, population density and proximity to urbanized areas.

Frame

The population units for the main portion of the RDD Employment and Health Survey are the households with a telephone in the living quarters. Individuals in households with a telephone in the living quarters comprise the population for the supplement for a designated respondent. The frame is the set of all possible four-digit suffixes combined with each of the existing six-digit prefixes formed by a combination of a three-digit area code (AC) and a three-digit central office code (CO). The source used to construct the frame is a magnetic tape from AT&T Long Lines which contains all the current working area code and central office code combinations (AC-CO) for North America.

Each record on the tape contains one AC-CO code and information concerning it. The information which is useful in a telephone survey includes the following:

1. Geographical vertical and horizontal coordinates of the area which the AC-CO code services
2. Area code
3. Central office code
4. Time Zone Indicator
5. Daylight Savings Time Not Observed Indicator
6. Place Name
7. State Name
8. Point Identification variable which indicates the country a AC-CO code services and whether it is designated for DIAL-IT services or for internal AT&T purposes.

The frame will be created by transferring from the AT&T tape to a file the records for all the AC-CO codes servicing points in the 50 states and the District of Columbia. To gain some efficiency in the first stage of sampling, some obviously nonresidential AC-CO codes will be eliminated.

The file used for sampling will have approximately 33,500 AC-CO codes. The AT&T tape is updated every month, and the most recent tape available will be used at the time of sample selection.

Sample Selection

The sample for the RDD Employment and Health Survey will be chosen in seven replicates which will be interviewed in seven separate two-week periods between April and August. Based primarily on budget and time limitations, the replicate sample sizes chosen were 500 interviews for five of the replicates and 750 interviews for the remaining two replicates. It was determined that these sample sizes would provide acceptable precision for the estimates of the response rate.

Each replicate will be selected in two stages: (1) groups of 100 telephone numbers defined by the first eight digits (PSUs) and (2) residential telephones. The sample size for a replicate is of the form n = mk where m is the size of the first-stage sample, and k is the size of the second-stage sample. Waksberg (1978) derived the optimum value of k, given by

$$k = \left[\frac{t}{\pi\,(C_p/C_u) + (1 - \pi - t)}\;\frac{(1 - \rho)}{\rho}\right]^{1/2}$$

where t = proportion of PSUs with no residen tial numbers,

π = proportion of residential numbers in the frame,

ρ = intraclass correlation within PSUs for some survey characteristic,

C_u = cost of an unproductive call, i.e., to a non-residential number,

C_p = cost of a productive call.

Groves (1977) has reported data from which t, ρ and π can be estimated for national surveys. The estimates of t and π are .65 and .2, respectively. The estimates of ρ are approximately .02 for attitude questions and about .05 for economic or social statistics.

The amount of time for the interviewer to conduct a productive interview, is estimated to be 35 minutes. An unproductive call should require about five minutes. Assuming that the cost of the interviews is proportional to the amount of time they require, the ratio of the cost of a productive to an unproductive call is seven, $C_p/C_u = 7$.

The design of the RDD Employment and Health Survey will be based on $\rho = .01$ because estimating response rates is of primary importance, and response to the survey is not believed to be highly correlated within a PSU. Therefore, k has been chosen to be six.

Once the desired number of interviews is determined, the appropriate sample size is derived from the anticipated response rate, R. For a replicate requiring 500 completed interviews, the sample size should be $n = 500 \div R$.

For the initial design it will be assumed that the response rate will be about 83 percent, although the literature indicates that this is conservative (i.e., probably low). For $R = .83$, the sample size will be $n = 600$. Therefore, since $k = 6$, $m = 100$ PSUs (i.e., $600 \div 6$) will yield the optimum design. When 750 interviews are required, the sample size will be $n = 900$, and $m = 150$ PSUs will give the optimum design.

The first stage of sample selection will be accomplished in two steps.[1] A systematic sample of the file of area code and central office code combinations will be the first step. The selected prefixes will be placed in random order. As the next step, a random choice of the next two digits will be added to each six-digit prefix selected to form an eight-digit number identifying a PSU of 100 telephone numbers. Another random selection of the next two digits will be made to identify a telephone number in the PSU. If, when this number is dialed, it proves to be a residence, the eight-digit PSU is retained. If the number is nonresidential, the PSU is rejected. Based on the random ordering of prefixes, telephone numbers are called until the target number of PSUs (i.e., 100 or 150) is obtained.

The placement of the six-digit prefixes in a random order permits the PSUs to then be screened one at a time in their new order until m are accepted without introducing a bias from the sort and systematic sampling procedure. Also, the procedure allows the number of PSUs retained in the sample to be fixed, instead of only fixing the expected number of PSUs.

In order to determine the appropriate number of six-digit prefixes which should be drawn to obtain m residential numbers, the appropriate probability model is the negative binomial distribution where the probability of success is .2, the proportion of residential telephone numbers in the frame. The negative binomial distribution, $p(x)$, gives the probability of requiring x independent Bernoulli trials to achieve m successes. It has been decided that the number of six-digit prefixes selected at the first step should be large enough so that there is only a 1 percent chance of failing to accept 100 PSUs in the second step. Consequently, the size of the sample drawn from the AC-CO code file will be equal to the 99th percentile of the negative binomial distribution, which is $m' = 610$ for $m = 100$ and $m' = 884$ for $m = 150$.

For the systematic sample of the file of AC-CO codes, the skip interval L will be computed as follows:

$$L = \frac{\text{No. of records}}{m'}, \quad \text{rounded to two decimal places.}$$

A random number r, where $0.01 < r \leq L$, will be generated. The sample will consist of records at positions r, r + L, r + 2L, ..., until the end of the file.

In the second stage of sample selection, two-digit numbers will be selected at random and added to each of the eight-digit PSUs to form telephone numbers. The numbers will be generated and dialed until $k = 6$ distinct residential telephones have been selected.

The first person who answers the telephone and is at least 16 years old will be the respondent for the household portion of the questionnaire. The interviewer will select a designated respondent at random from those at least 16 years of age on the household roster to answer a supplement for an individual.

Sorting the File for the First Stage

Prior to the first stage of sample selection, the file of AC-CO codes will be sorted to decrease the possibility of underrepresentation in the sample of units with particular characteristics correlated with response. The sort of the file will be based on geography, population density, and proximity to urbanized areas. The major sort of the file will be census region in the following order:

1. Northeast
2. South
3. North Central
4. West

Within each region, the AC-CO codes will be placed in size-urbanization categories which are based on two criteria. One factor determining the classification of an AC-CO code will be the number, N_{VH}, of AC-CO codes in its exchange area, that is, the number of AC-CO codes with the same vertical coordinate V and horizontal coordinate H. The other criterion will be the distance the V and H coordinates for the AC-CO code is from an exchange area with $N_{VH} \geq 30$. Each unit of a V or H coordinate is approximately one-third of a mile. The first size-urbanization category will contain large cities and the second, their surburban areas. The remaining AC-CO codes will be grouped into four categories based on their N_{VH} numbers. The scheme for identifying the surburban areas is similar to the one used by CBS News (Yusef, 1977). The size-urbanization categories are shown in Table 2.

The number of AC-CO codes, N_{VH}, in an exchange area is used only as a sort (implicit stratification) variable; it is not used as a measure of size for setting the first-stage selection probabilities. The use of N_{VH} as the measure of size in a PPS sampling procedure would be questionnable since determining the exact relationship between the number of AC-CO codes in an exchange area and its resident population is difficult. Although the exchanges are geographically defined areas with place names, approximately 20 percent of the area serviced by the AC-CO codes may lie outside the incorporated boundaries of the place name (Groves and Scott, 1976). Also an exchange area may have more than one place name listed for its AC-CO codes. The use of the size-urbanization categories is based on the conjecture that whether a person lives in an urban, surburban, or rural area is correlated with responding to a telephone survey and to many "typical" survey questions.

To assist with the development of the specific categories, the Statistical Abstract of the United States, 1980 and 1980 Census of Population and Housing, Advance Reports were consulted. According to the Statistical Abstract of the United States, 1980, Category 1 generally contains large cities which are the cores of Standard Metropolitan Statistical Areas with a population greater than 500,000. Category 2 contains the surburban areas, AC-CO codes within 20 miles of a city in Category 1. The operational criterion to identify the AC-CO codes in a suburban area of an exchange area in Category 1 with coordinates (V,H) is those in exchange areas with coordinates (V',H') lying inside a square, 40 miles on a side, centered at (V,H).

Using 1980 Census of Population and Housing, Advance Reports, Category 3 contains other large cities with a population approximately between 500,000 and 100,000. Category 4 contains medium-sized cities with a population roughly between 100,000 and 50,000. The exchange areas with only one AC-CO codes have been isolated in Category 6 because they are probably rural farm areas with very little urbanization.

Within each size-urbanization category in a region, the AC-CO codes will be ordered by area code. The sequence of area codes is a geographical ordering which respects state lines by grouping all area codes servicing a state together.

When using systematic sampling, abrupt changes in characteristics in members of a file or list tend to increase the variance of an estimator. To minimize the number of abrupt changes in the characteristics of AC-CO codes when the file is sorted, the order of the size-urbanization categories will be reversed across the region boundaries and the order of the area codes will be reversed across size-urbanization categories.

Within area codes, the V and H coordinates will be used to order the AC-CO codes geographically, alternating between sweeping northwest to southeast and southeast to northwest. If the V and H coordinates were used without adjustment, they would provide a sweeping band that is too narrow. A wider band is obtained by truncating the V coordinate as follows:

$$V_T = V/30, \text{ rounded to the nearest integer.}$$

The use of the V_T and H coordinates to define a geographic sweep will provide bands that are 10 miles wide. The sweeping effect will be accomplished by ordering numerically in ascending order the code C_{VH}, where the C_{VH} code for an AC-CD code with coordinates V and H is defined by

$$C_{VH} = \begin{cases} V_T{*}100000 + H, & \text{if the last digit in } V_T \text{ is even,} \\ V_T{*}100000 + (99999-H), & \text{if the last digit in } V_T \text{ is odd.} \end{cases}$$

The AC-CO codes with the same C_{VH} will be ordered numerically in ascending order. In some areas, telephone companies assigned CO codes by zones so a numerical ordering gives some geographic continuity. In areas where telephone numbers with the same CO code are spread over the entire exchange area, a numerical ordering is essentially a random ordering.

The final sort will be in the following sequence of codes:

1. Census Region
2. Size-urbanization category
3. Area code
4. C_{VH} code
5. Central office code in ascending order.

Refusals

During the calls which screen the PSUs in the first stage of sampling, the interviewer should make every effort to determine whether the telephone number is residential. If a second call also does not determine whether the number is residential, another telephone number within the PSU will be generated at random and dialed. Whether or not the replacement number is residential will determine if the PSU is retained or discarded.

In the second stage, anyone who refuses to respond to the survey will be called again by another interviewer and requested a second time to participate. Weight adjustment will be made to account for final refusals and other nonrespondents.

Ring-No-Answer and Other Unresolved Numbers

If the first time a telephone number is dialed, there is a ring but no answer or the call is not completed, the number will be dialed three more times at varied times of the day and evening. If, after the three additional calls, there is no resolution of the status of the number, the business office will be called and asked whether the number is working or nonworking. If a number is identified by the business office as nonworking or nonresidential during the first stage of sampling, the PSU will be rejected. If this occurs during the second stage of sampling, the search for the k residential numbers in the PSU will continue.

Sparse PSUs

In some cases a PSU will be selected that contains a relatively small proportion of residential numbers--perhaps as low as 10 percent or less. It is anticipated that the number of such "sparse" PSUs will be very small. Consequently, if necessary, all 100 telephone numbers will be called in a PSU in attempting to reach six residential numbers. The anticipated low frequency of occurrence of sparse PSUs (Groves and Kahn, 1979) does not merit setting up a stopping rule or other special sampling procedures which may create some bias and complications in developing appropriate weights.

This conclusion was based on an examination of the hypergeometric waiting-time distribution of the random variable X, the number of telephone numbers which must be drawn from the remaining 99 in a PSU to obtain k=6 residential numbers. The expected number of successive draws, μ, required to obtain k=6 residential numbers when there are N_p residential numbers out of the remaining 99 in the PSU, is shown below for various values of N_p.

N_p	μ
6	85.7
9	60
14	40
19	30
39	15
62	9.5

If the proportion of residential numbers in an accepted PSU equals .63, which is the estimated proportion of residential numbers in a PSU with at least one residential number, the expected number of telephone numbers which must be drawn to obtain k=6 residential numbers is 9.5. If the proportion of residential numbers in a PSU is 0.2, which is the estimated proportion of residential numbers in all PSUs combined, the expected number of telephone numbers which must be drawn to obtain k=6 residential numbers is 30.

The only special operational procedure which will be implemented to address sparse PSUs is that if very few residential numbers have been found after a substantial number of calls in a PSU have been contacted, the supervisor will call the number used to accept the PSU in the first stage of sampling to confirm that it is residential. If the supervisor finds that the number is not residential, the PSU will be replaced.

Suggested criteria for when the number used to accept the PSU should be called again are shown below. If r or less residential numbers have been found after N_r contacts, the status of the number used to accept the PSU should be checked.

N_r	r
10	0
15	1
20	2
30	3

The check points were based again on the hypergeometric waiting-time distribution. If the PSU has the same proportion of residential numbers as all the PSUs combined, 0.2, then the probability of contacting r or less residential numbers in N_r contacts is approximately 0.10.

If there is a case where all 100 telephone numbers in a PSU are called, only to discover that the PSU contains less than k=6 additional residential telephone numbers, then the data for the households contacted will be weighted accordingly.

Estimation and Variance Estimation

Although the sample for each replicate has been designed to be self-weighting, there will be differential weights due to the adjustments for households with multiple telephones, nonresponse, and PSUs with less than k=6 residential telephone numbers. Also, a ratio adjustment will be applied to the weights on a region basis so that their sum will approximate the total number of telephone households, based on data from the 1980 census. Therefore, an estimator for a total of a survey variable x is given by

$$x' = \sum_{i=1}^{m} \sum_{j=1}^{k} w_{ij}\, x_{ij}$$

where x_{ij} = the value of the x variable for the j-th unit in the i-th PSU,

w_{ij} = the weight for the j-th unit in the i-th PSU, which is basically the inverse of the selection probability, adjusted for nonresponse and ratio estimation.

An estimator for the mean of the variable x is given by

$$\bar{x} = \frac{1}{w} \sum_{i=1}^{m} \sum_{j=1}^{k} w_{ij}\, x_{ij},$$

where $w = \sum_{i=1}^{m} \sum_{j=1}^{k} w_{ij}.$

The variance estimators will be developed by using the PSUs as ultimate clusters (Hansen, Hurwitz, Madow, 1952). Since there are differential weights, the estimator of the variance of x' will have the following form:

$$v = \frac{m}{m-1} \sum_{i=1}^{m} \left(x'_i - \frac{x'}{m}\right)^2 ,$$

where

$$x'_i = \sum_{j=1}^{k} w_{ij}\, x_{ij} \, .$$

The variance of the estimator of the mean $\bar{x}$ analogously will be based on the ultimate cluster approach although it requires a Taylor series approximation of the variance of a ratio estimator.

The data collected in all the seven replicates will be combined to form estimates of the total and mean of survey variables. Since the estimates from each replicate are independent estimates from essentially the same population, an estimator for the total of a survey variable is given by

$$x_T = \frac{1}{7} \sum_{j=1}^{7} x'_j ,$$

where x'_j = the estimator of the total of the survey variable from the j-th replicate.

The estimator of the variance of x_T will have a form similar to the estimator of the variance of x'.

Etimators for the mean of a survey variable $\bar{x}_T$ and the variance of $\bar{x}_T$ which use the entire sample will be formed analogously.

Appendix

An analysis of the geographical distribution of the PSUs retained for the RDD Employment and Health Survey indicates that they are distributed approximately proportionally as expected.

The relative frequency distribution of the PSUs chosen in the first step of the first state using the AT&T tape is shown by region and size-urbanization category in Table 1.

After the PSUs were screened, the distribution of the 800 PSUs retained for all the replicates combined is shown in Table 2.

Although no data are available to check the correctness of the distribution of the PSUs among the size-urbanization categories, the distribution is informative. For example, having only two retained PSUs in Category 2 in the North Central and none retained in Category 2 in the West indicates that the central city exchanges in these regions are large enough to include most of the suburban areas.

For evaluating the geographical distribution of the retained PSUs, comparing the distribution of the retained PSUs by census region or division to the distribution of residential telephone numbers would be optimal. However, the distribution of residential telephone numbers is not available. A distribution which is available is the distribution of housing units from the 1980 Census of Population and Housing, Advance Reports. Although the states probably do not have either the same percentage of housing units without telephones or the same percentage of multiple telephone households, the distribution of housing units should approximate the distribution of residential telephone numbers closely enough to provide some insight as to the adequacy of the sample design and procedures. The significance probabilities of the results of the Chi-square tests of the goodness-of-fit of the distribution of the PSUs from individual replicates to the distribution of housing units, where the cells are the four census regions, range between .16 and .93. When the distribution of the PSUs from all the replicates combined is tested, the significance probability is .75.

Also of interest is whether the number of PSUs from each state is reasonable, and which states, if any, have more or less than they should. By considering the probability that a retained PSU is in a state to be equal to the proportion of the housing units in the U.S. that are in the state, probability intervals for the number of PSUs in a state can be developed from the binominal distribution.

Approximate .95 probability intervals and the number of PSUs in each state for all seven replicates combined are given in Table 3. For example, assuming the probability that a retained PSU in Alabama is .017, then 95 percent of the time when 800 are selected, the number of PSUs which are in Alabama would be greater than or equal to six and less than or equal to 20.

Alaska is the only state not represented among the PSUs, but that is not unreasonable. For all the replicates combined, the number of retained PSUs falls outside the .95 probability interval for seven states. These minor discrepanices could be attributed to the difference between the distribution of the residential telephone numbers and the distribution of the housing units.

An examination of each individual replicate shows that there are between one and four states in each for which the number of retained PSUs is outside the .95 probability interval. However, none of the states have the number of selected PSUs falling outside the .95 probability interval consistently.

Therefore, these results indicate that the sampling design and procedures provide an acceptable geographical distribution of PSUs.

Table 1
PSUs Retained for the Sample

Region	Size-Urbanization Category						Total
	1	2	3	4	5	6	
Northeast	43	24	8	25	39	37	176
North Central	41	2	10	27	52	61	193
South	72	33	17	24	48	80	274
West	42	0	33	43	25	14	157
Total	198	59	68	119	164	192	800

Table 2
Size-Urbanization Categories

Size-Urbanization Category	Heuristic Description	Operational Criteria
1	very large cities	AC-CO codes in exchange areas with 30 or more AC-CO codes.
2	suburban areas	AC-CO codes not in category 1 with coordinates (V,H) such that there is an exchange area in category 1 with coordinates (V',H') such that $\lvert V-V'\rvert \leq Q$ and $\lvert H-H'\rvert \leq Q$, where $Q = 60$. (The choice of $Q = 60$ provides for the inclusion of AC-CO codes within approximately 20 miles of any exchange area in Category 1.)
3	large cities	AC-CO codes not in category 2 which are in exchange areas with an N_{VH} such that $13 \leq N_{VH} \leq 29$.
4	medium cities	AC-CO codes not in category 2 which are in exchange areas with an N_{VH} such that $6 \leq N_{VH} \leq 12$.
5	small towns and rural areas	AC-CO codes not in category 2 which are in exchange areas with an N_{VH} such that $2 \leq N_{VH} \leq 5$.
6	rural farm and other rural areas	AC-CO codes not in category 2 which are in exchange areas with $N_{VH} = 1$.

Table 3: .95 Probability Intervals for the Number of Retained PSUs from Each State for All Seven Replicates Combined

State	Proportion of housing units	Lower Bound	Upper Bound	No. of PSUs
AK	.002	0	4	0
AL	.017	6	20	12
AR	.010	3	13	13
AZ	.012	3	15	7
CA	.105	67	100	97
CO	.013	4	16	9
CT	.013	4	16	5
DC	.003	0	5	4
DE	.003	0	5	6*
FL	.049	27	51	52*
GA	.023	10	26	21
HI	.004	0	6	2
IA	.013	5	16	15
ID	.004	0	6	3
IL	.049	27	51	45
IN	.024	11	27	10*
KS	.011	3	14	15*
KY	.015	5	18	11
LA	.018	7	21	13
MA	.025	11	28	21
MD	.018	7	21	20
ME	.006	1	9	3
MI	.041	22	43	32
MN	.018	7	21	15
MO	.022	10	25	17
MS	.010	3	13	3
MT	.004	0	6	3
NC	.026	12	29	15
ND	.003	0	5	2
NE	.007	2	10	3
NH	.004	0	6	6
NJ	.031	16	34	24
NM	.006	1	9	2
NV	.004	0	6	6
NY	.078	48	77	65
OH	.046	26	48	33
OK	.014	5	17	7
OR	.012	4	15	6
PA	.052	30	53	42
RI	.004	0	6	6
SC	.014	5	17	8
SD	.003	0	5	1
TN	.020	9	24	19
TX	.063	37	63	34*
UT	.007	1	10	4
VA	.023	10	26	23
VT	.002	0	4	3
WA	.019	8	22	19
WI	.021	9	24	5*
WV	.008	2	11	13*
WY	.002	0	4	1

*Number of PSUs Retained Does Not Lie in the .95 Probability Interval

Footnotes

1/This election procedure is based on a method described by Joseph Waksberg of Westat during a telepehon conversation with David Chapman in October 1981.

References

Casady, Robert J., and Monroe G. Sirken (1980), "A Multiplicity Estimator for Multiple Frame Sampling," Proceedings of the American Statistical Association,Survey Research Section.

Groves, Robert M. (1977), "An Empirical Comparison of Two Telephone Sample Designs," unpublished report of the Survey Research Center of the University of Michigan, Ann Arbor, Michigan.

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

Groves, Robert M. and John C. Scott (1976), "An Attempt to Measure the Relative Efficiency of Telephone Surveys for Social Science Data Collection," paper prepared for the American Association of Public Opinion Research Annual Conference 1976.

Hansen, M. H., W. N. Hurwitz, W. G. Madow (1953), Sample Survey Methods and Theory, John Wiley & Sons, Inc., New York.

Mitofsky, Warren (1970), "Sampling of Telephone Households," unpublished CBS memorandum.

Waksberg, Joseph (March, 1978), "Sampling Methods for Random Digit Dialing," Journal of the American Statistical Association, Volume 73, No. 361, pp. 40-46.

Wilks, Samuel S. (1962), Mathematical Statistics, John Wiley & Sons, Inc., New York.

Wolter, Kirk M., Introduction to Variance Estimation, Springer-Verlag, New York (to be published in 1983).

Yusuf, M. (1977), "Telephone Survey Specifications," unpublished CBS memorandum.