SAMPLING A RARE POPULATION: A GENERAL STRATEGY ILLUSTRATED BY A CASE STUDY

Eugene P. Ericksen, Institute for Survey Research, Temple University

1. Introduction

Studies concentrating on or restricted to subgroups of a total population comprise an increasing share of sample survey activity. At the Institute for Survey Research, we are seldom asked to carry out a straightforward cross-sectional study of the national household population where each household would have equal probability of selection. Instead, our national studies have focused on, for example, demographic subgroups defined by age, race, and sex, homeowners, or physicians concentrated in certain fields.

In such restricted studies, it is usually necessary to screen certain households based on information obtained from a door-answerer. This screening process is time-consuming, and if an extensive amount of screening is needed, the process can significantly increase the average cost of obtaining one interview. However, if the target population is at least partially segregated, the areas where it is concentrated can be sampled at a higher rate, reducing the amount of necessary screening and reducing the increase in the average cost per interview. The strategy is to form separate strata depending on variations in concentration of the target population and then to apply the strategy of optimal allocation (Hansen, Hurwitz, and Madow, 1953, section 6E; Cochran, 1963, pp. 95-97; Kish, 1964, sections 3.5, 3.6). Optimal allocation will bring about gains if the areas of concentration can be identified, and if costs vary directly with the amounts of concentration. However, it is important that a large proportion of the target population live in such segregated areas. Kish (1964, pp. 409-410) provides an illustration where the target population was high-income households. It was easy to identify the most exclusive neighborhoods, but most of the rich were scattered elsewhere and the gains from oversampling the exclusive neighborhoods were small.

Cochran (1963, p. 95) shows that in stratified sampling the variance of the overall mean of a given study variable is minimized when the specified sampling rate in a given stratum h is:

$f_h = n_h/N_h$ is proportional to $S_h/\sqrt{c_h}$, (1.1)

where

n_h = the expected number of selected observations in stratum h,

N_h = the total population in stratum h,

S_h^2 = the variance of the variable in stratum h,

c_h = the average cost of one interview in stratum h.

In most practical sampling situations, the between-stratum variations in S_h are minimal, so the sampling rates are usually functions of the cost variations. In computing these costs, it is important to include the costs of all components which vary with the numbers of interviews collected including the listing of households, travel, and the costs of coding.

2. The Specific Sampling Problem

The specific sampling problem providing our illustration was the construction of a sample of 15 through 19-year-old females living in households. The specified ratio of white to black respondents was two to one. The subject matter of the study, fertility practices and expectations, has been reported in part by Zelnik and Kantner (1972).

In order to obtain white and black interviews at the specified ratio, we estimated that it would be necessary to oversample blacks at 3.86 times the rate at which white respondents were to be selected. We also estimated that an eligible respondent would be found in one household in seven, and that completion rates would be 75 per cent for respondents of both races. Based on these computations, it was estimated that it would be necessary to visit $7/(.885 \times .75) = 10.55$ households to obtain one white interview and $7/(.115 \times .75) = 81.16$ households to obtain one black interview.

The increase in the number of required households to be contacted for one black interview was due to the necessity of screening out households with otherwise eligible white respondents. This screening would have been concentrated in predominantly white areas with few potential black respondents, and would have increased costs by an inordinate amount. It was therefore decided to subdivide the household population into two strata, with Stratum 1 to include predominantly white areas where blacks and whites were to be sampled at the same rate, and with Stratum 2 to include those areas where blacks lived in sufficient concentrations where they could be oversampled without undue increases in the amount of necessary screening. Whites in Stratum 2 were to be selected at the same rate as in Stratum 1, but blacks were to be selected at a rate 3.86 times greater.

It was then necessary to estimate the amounts of screening and increases in costs which would be obtained in areas varying in the proportions of households which were black. Because white respondents were to be selected at a constant, lower rate throughout, the additional costs of screening by race were all applied to the black interviews. The following assumptions were made:

a. Each cluster of households, or "listing area," would include an expected 100 households yielding an expected 9 or 10 interviews if there were no screening on

the basis of race.

b. Three hours would be required to list all households in a given listing area, 15 minutes would be required to determine for each household whether or not an eligible respondent lived there who was willing to be interviewed, and a total of 2 hours would be required to complete and code the interview.

c. The costs per hour would be the same for the various components of the total process.

The time necessary to complete each black interview was computed for listing areas varying in racial composition. These computations are shown in Table 1. There we can see that the expected time necessary to obtain a black interview rose sharply when the proportion of households which were black fell below 25 per cent. Applying equation (1.1), when 3 per cent of all households were black, the optimal sampling rate was 1/3.83 times the rate at which blacks would be sampled in a totally black area. If the proportion of black households was 10 per cent or less, the optimal sampling rate for blacks was closer to the sampling rate for whites than it was to the specified oversampling rate for blacks which was 3.86 times greater. It was therefore decided that if it were reasonably clear that a given listing area included fewer than 10 per cent black respondents, it would be placed in Stratum 1, and would otherwise be placed in Stratum 2.

The actual stratification procedure was composed of two stages. Estimates of the racial composition for the selected primary sampling units (psus) included in the national sampling frame were computed on the basis of (1) the 1960 Census, which was unfortunately 10 years out-of-date at the time, and (2) estimates of the racial composition of other listing areas in the same psus which had been used in recent surveys. If we confidently estimated, given the limitations of the data, that a psu was less than 3 per cent black, that psu was placed in Stratum 1, and listing areas were selected at 1/3.86 times the Stratum 2 rate. Household listings were then obtained from all selected listing areas in the two strata, and estimates of the racial composition were obtained. A certain amount of error in these estimates was expected, but if it was estimated that a given listing area in Stratum 2 included fewer than 10 per cent black households, and if this estimate did not contradict estimates computed from census data and past surveys, the listing area was transferred to Stratum 1. Listing areas thus transferred to Stratum 1 were then subselected at the rate of 1 in 3.86. Within Stratum 2, white households were subselected at this rate at the time of interviewing. Therefore, the difference in the sampling procedure for whites in the two strata was that in Stratum 1, all subselection was done in advance and no screening of households was necessary, whereas in Stratum 2, the sampling rate for households was 3.86 times greater, and households including potential white respondents were subselected in the field at the rate of 1/3.86.

3. Results: Numbers of Interviews

Two thousand nine hundred and fifteen (2,915) interviews were obtained with white respondents and 1,438 interviews with black respondents. Two thousand five hundred and thirty-five (2,535) interviews were obtained in Stratum 1, of which 58, or 2.29 per cent, were with black respondents. Multiplying the 438 white respondents in Stratum 2 and the 58 black respondents in Stratum 1 by 3.86, we estimate that 45 per cent of potentially eligible respondents in Stratum 2 were black, and that 86 per cent of all potentially eligible black respondents lived in Stratum 2.

The classification of areas into the two strata thus appears to have been accurate. Multiplying the 1,438 black interviews by the estimated 81.16 households necessary to obtain one interview, it would have been necessary to include 116,708 households in the sample had we not followed the stratification procedure. Had all blacks been selected at the higher rate, $1,380 + 58 \times 3.86 = 1,603.88$ interviews with blacks would have been obtained, but the sample would have had to include 130,170 households. Ninety-two thousand five hundred and ninety-nine (92,599) housing units were in fact listed. Of these, 11,594 were eliminated from the sample on the basis of the enumerator's estimate of the racial composition of listing areas, leaving a final sample of 81,005 housing units, 56.33 per black interview.

After the survey was completed, the racial composition of all listing areas included in Stratum 2 was calculated from screening forms filled out for each listed household, and compared to the estimates made by enumerators when listing households. These estimates tended to be accurate, with the mean absolute error being 13.39 per cent, the mean actual error (actual minus estimated percentage black) 3.08 per cent, and the median absolute error 6.03 per cent. Of the 477 estimates, 105 had greater than 20 per cent error, 20 had greater than 50 per cent error, and 7 had larger than 80 per cent error. Many of the larger errors contradicted the estimate expected on the basis of census and past survey data, and where there was doubt, a listing area was retained in Stratum 2. In most such cases, the error turned out to be a case of the interviewer giving the percentage black where the percentage white was intended.

The relative effects of actual racial composition of listing area, race of interviewer, region of the U.S., and central city-suburban status on the accuracy of the estimates of racial composition were measured by using these as explanatory variables in a multiple regression equation used to estimate the size of the absolute errors. This exercise was limited to SMSAs, because black interviewers did not work in non-metropolitan areas. The results are presented in Table 2. There we see that only one variable,

the actual racial composition of an area, provided any substantial clue to the accuracy of the estimate. The closer the actual composition of the area was to 50 per cent, the greater was the chance of error, and the other characteristics added only negligible amounts of explanation. When the larger, presumably random errors were eliminated from consideration, the explanatory power of the variables increased noticeably.

4. Results: Costs Per Interview

A computer record was kept of the results of calls of each listed address. Records of the salaries and expenses paid to each interviewer were also kept for both the household listing and interviewing phases of the study. The results of calls and salaries and expenses were aggregated to the psu level for computation of costs per interview. Where an interviewer worked in more than one psu, the psus were combined, and the full sample of 126 psus was thus reduced to 115 units.

It was estimated that the average cost of coding one interview was $4. Adding this cost to the tabulated costs of listing and interviewing, the average cost per interview of these components was $27.27. In the 69 psus where all listing areas were included in Stratum 1, the average cost per interview was $22.49, and in the remaining areas the average cost was $31.67. Assuming an average cost per white interview of $22.49 in these remaining 46 areas, the average cost per black interview rose to $36.21.

Regression equations estimating the average costs per interview are presented in Table 3. There we see that the increases in costs due to screening and to nonresponses were comparable, and that small reductions in costs were obtained in psus where the number of interviews obtained was large. Because the variation in the amount of screening was much greater than the variation in the other two variables, this was the primary determinant of variations in cost per interview.

We are now in a position to evaluate our decision regarding the optimal cutting point for stratifying listing areas. Within Stratum 1, an average of 10.5 households were screened out on the basis of age for every interview collected and an average of .67 nonresponses, usually refusals or cases where the presence of an eligible respondent could not be determined, were obtained per interview. Applying equation 5 from Table 3, the average per interview cost in a given psu in the absence of screening on the basis of race could then be expected to be

$$\hat{Y} = 11.79 + (1.18)\ (10.5) + (1.38)\ (.67) - .03K$$
$$= 25.09 - .03K \text{ dollars, where}$$

K = the number of interviews obtained in the psu.

Within Stratum 2, the presence of white households increased the amount of screening in proportion to the number of such households. The formula expressing the expected increase is

$$I = W + (I-W)/3.86 \text{ where} \qquad (4.1)$$

W = the proportion of households which are black, and

10.5/1 = the expected number of screened households per interview.

For example, in totally white areas, I = .259 and the expected number of screened households was 10.5/.259 = 40.5. In an area where 50 per cent of households were black, I = .630, and the expected number of screened households, 16.7. In the first case, the expected cost per interview was raised to (60.49 - .03K) dollars, and in the second case to (32.41 - .03K) dollars where K equals the number of interviews.

Since the same number of white interviews could have been obtained without racial screening by placing all listing areas in Stratum 1, the costs of such screening must be added to the costs of obtaining black interviews. Therefore, the expected cost per black interview would be higher than the overall cost per interview, and the increase would be considerable in areas with few blacks. The estimated costs for areas of different racial composition are presented in Table 4. There we see that the cutoff points between the two strata were about as predicted in Table 1. By subsampling listing areas using the enumerators' estimates of racial composition, the proportion of black eligible respondents in those listing areas of a psu where household screening on the basis of race was carried out never fell below 10 per cent. The actual cost per interview was over $100 in only one psu, and the estimated cost per black interview exceeded $100 in only three psus.

5. Results: Design Effects

The sampling frame used for this study had one unfortunate aspect. The primary sampling units were constant in size, each one being defined to include 10,000 housing units as of the 1960 Census. The geographic area covered by such psus varied greatly. Among the selected psus, the range extended from a psu covering about one square mile on the south side of Chicago to a psu covering over 20,000 square miles in ten counties in eastern Montana. The sample psus were typically small and homogeneous in comparison to psus selected in other frames such as that of the Current Population Survey or the Survey Research Center at the University of Michigan. This difficulty was anticipated before the study began, but because of time constraints, it was not possible to construct a new sampling frame with more heterogenous psus such as the two mentioned above or that now used at ISR.

These characteristics of the psus had two important effects. One was that they were homogeneous, and the values of the intraclass correlation coefficients for study variables were greater than they would have been for other studies. The other was that the black population was concentrated in only a few of the sample psus, so that over half the black interviews were obtained in just 10 of the 126 sample psus. Therefore the design effects, which measure the increase in variance over what would have been ob-

tained from a simple random sample of the same size and is the product of the intraclass correlation and the average cluster size, were increased over what would have been expected even for such homogeneous psus.

Design effects were computed for 12 variables. These are demographic and economic variables for which the design effects are usually larger than for other variables, particularly attitudes. These were computed for four groups, blacks in Stratum 2 subdivided into three groups depending on racial composition, and hence the cost per interview, of the psu, and whites. There were not enough blacks in Stratum 1 to merit separate computations. The three black groups included approximately the same numbers of interviews. These are shown in Table 5.

The intraclass correlations for the blacks were lower than they were for the whites, indicating greater heterogeneity within black areas than white areas. However, the clustering of black interviews was so much greater than the clustering of white interviews that the design effects for blacks were much greater in the two groups of interviews obtained in particularly black areas and nearly as great in the third group, where the cost per interview was much higher.

The lower costs per interview obtained in the areas of greatest black concentration appear to have been obtained at the price of greater design effects. When the costs per interview were multiplied by the design effects, giving the costs per equivalent simple random sampling interview, these latter costs were comparable among the three black groups, ranging from about $100 to $120 per equivalent interview.

The lessons to be learned from this exercise can be summarized as follows:

a. The additional costs of screening, given these costs of the various components of interviewing, become great when the subpopulation comprises about 10 per cent of the total, and rise sharply below that level.

b. When sampling blacks, gains can be made by identifying areas of greater concentration, and applying optimal allocation.

c. However, it is important that the black interviews not be clustered in a few small, homogeneous psus as the design effects will be unduly large.

d. This difficulty can be overcome by having more diverse psus with smaller numbers of black interviews in each. We saw from equation 5 in Table 3 that little reduction in cost per interview is obtained from having many interviews in one psu. However, if areas of greatest black concentration can be identified within psus, these can be sampled at a higher rate, minimizing the amount of screening, spreading the black sample over a greater number of areas, and reducing the number of black interviews in any one psu.

Table 1. Expected Interviewing Costs in Listing Areas Varying in Racial Composition

Percent Black in Listing Area	Estimated Number of Interviews[1] White	Black	Total	Expected Total Hours Spent in Listing Area[2]	Expected Total Hours Required for all White Interviews in Listing Area[3]	Expected Total Hours Required for all Black Interviews in Listing Area	Expected Hours Per Black Interview	Increase in Time Per Black Interview Because Screening Necessary (Square Root)
0	2.78	0	2.78	33.56	12.82	20.74	---	---
1	2.75	.11	2.86	33.72	12.68	21.04	191.27	41.49 (6.44)
2	2.72	.21	2.93	33.86	12.54	21.32	101.52	22.02 (4.69)
2.5	2.71	.27	2.98	33.96	12.49	21.47	79.52	17.25 (4.15)
3	2.69	.32	3.01	34.02	12.40	21.62	67.56	14.66 (3.83)
5	2.64	.54	3.18	34.34	12.17	22.17	41.06	8.91 (2.98)
10	2.50	1.07	3.57	35.14	11.53	23.61	22.07	4.79 (2.19)
25	2.08	2.68	4.76	37.52	9.59	27.93	10.42	2.26 (1.50)
50	1.39	5.36	6.75	41.48	6.41	35.07	6.54	1.42 (1.19)
75	.69	8.04	8.73	45.54	3.18	42.36	5.27	1.14 (1.07)
90	.28	9.64	9.92	47.84	1.28	46.56	4.83	1.05 (1.02)
100	0	10.71	10.71	49.42	--	49.42	4.61	1.00 (1.00)

[1] Assuming 100 households per listing area, 1 household in 7 to include an eligible respondent, and a 75 percent completion rate and blacks selected at 3.86 times the rate of whites.

[2] 3 hours required to list households, 15 minutes to process each household before interviewing, 2 hours to carry out and code each interview.

[3] This figure equals 2 hours times the number of expected white interviews plus 28 hours times (% white/3.86), where 28 hours is estimated time needed to list and contact 100 households.

Table 2. Relative Effects of Four Variables on Accuracy of Estimates of Racial Composition in SMSAs

Equation	Coefficient of Determination	Cases Included
$\hat{Y} = 34.686 - .299X_1 + .282X_2 + 1.357X_3 - .571X_4$	$R^2 = .195$ $r^2_{4y} = .189$	All Estimates n = 253
$\hat{Y} = 35.194 - .109X_1 - .495X_2 + 3.018X_3 - .604X_4$	$R^2 = .257$ $r^2_{4y} = .249$	Estimates Where Error Less than 80% n = 251
$\hat{Y} = 33.665 - .208X_1 - 1.948X_2 + 2.561X_3 - .568X_4$	$R^2 = .386$ $r^2_{4y} = .373$	Estimates Where Error Less Than 50% n = 243

Y = absolute error of estimate,

X_1 = 0 if area located in South, 1 otherwise,

X_2 = 0 if area located in suburb, 1 in central city,

X_3 = 0 if enumerator's race was black, 1 otherwise,

X_4 = absolute difference of actual percentage black from 50 per cent.

Table 3. Regression Equations Estimating Variations in Costs Per Interview Over Primary Sampling Units

1. $\hat{Y}_1 = 2.72 + .72X_1 + .75X_2 + .01X_3$ $R^2 = .485$

2. $\hat{Y}_2 = 3.64 + .21X_1 + .30X_2 - .02X_3$ $R^2 = .145$

3. $\hat{Y}_3 = 6.37 + .93X_1 + 1.03X_2 - .02X_3$ $R^2 = .416$

4. $\hat{Y}_4 = 1.67 + .24X_1 + .33X_2 - .02X_3$ $R^2 = .253$

5. $\hat{Y}_5 = 11.79 + 1.18X_1 + 1.39X_2 - .03X_3$ $R^2 = .443$

X_1 = number of households screened/number of interviews collected,

X_2 = number of nonresponses/number of interviews collected,

X_3 = number of interviews collected.

Y_1 = total salaries paid to interviewers for interviewing/number of interviews collected,

Y_2 = total expenses paid to interviewers for interviewing/number of interviews collected,

Y_3 = total salaries and expenses paid to interviewers for interviewing/number of interviews collected,

Y_4 = total salaries and expenses paid to interviewers for listing/number of interviews collected,

Y_5 = total costs paid for interviewing, listing, and coding/number of interviews collected.

Table 4: Estimated Costs Per Interview and Per Black Interview in Areas Varying in Racial Composition

Percent of all Households which are Black	Percent of Potentially Eligible Respondents not Screened on Basis of Race	Estimated Number of Screened Households per Interview	Cost per Interview[1]	Percent of all Interviews which are with Black Respondents	Cost per Black Interview[2]	Square Root, Cost per Black Interview Divided by Expected Cost per Interview with no Racial Screening
0	25.91	40.5	59.35	0	---	---
1	26.65	39.4	58.05	3.75	933.28	6.24
2	27.39	38.3	56.75	7.30	473.26	4.45
3	28.13	37.3	55.57	10.66	320.54	3.66
4	28.87	36.4	54.51	13.86	244.09	3.19
5	29.61	35.5	53.45	16.89	198.62	2.88
10	33.32	31.5	48.73	30.01	106.52	2.11
20	40.73	25.8	42.00	49.10	60.71	1.59
30	48.13	21.8	37.28	62.33	45.34	1.38
40	55.54	18.9	33.86	72.02	37.71	1.25
50	62.95	16.7	31.27	79.43	33.17	1.18
60	70.36	14.9	29.14	85.28	30.04	1.12
70	77.77	13.5	27.49	90.01	27.88	1.08
80	85.18	12.3	26.07	93.92	26.23	1.05
90	92.59	11.3	24.89	97.20	24.94	1.02
100	100.00	10.5	23.95	100.00	23.95	1.00

[1] Estimated using equation 5, Table 3, setting the number of interviews equal to 38, the mean obtained over the 115 areas.

[2] This assumes that the expected cost per white interview in Stratum 2 equals the expected cost per interview in Stratum 1 equals 25.09-(.03)(38) = $23.95.

Table 5: Design Effects and Relative Costs for Black and White Interviews

Group[1]	Mean, Intraclass Correlation	Average Cluster Size	Mean Design Effect[2]	Cost per Interview	Cost Per Equivalent Simple Random Sampling Interview[3]
Blacks, Set 1	.044	76.2	4.29	23.36	100.21
Blacks, Set 2	.059	52.6	4.02	26.61	106.97
Blacks, Set 3	.043	27.2	2.13	55.90	119.07
Whites	.070	23.7	2.59	23.95	62.05

[1] Blacks, Set 1 includes all black interviews in psus where blacks comprised 90 to 100 percent of population.

Blacks, Set 2 includes all black interviews in psus where blacks comprised 40 to 90 percent of population.

Blacks, Set 3 includes all remaining black interviews in Stratum 2.

Whites includes all white interviews

[2] The design effect was equal to $1 + \text{roh}\,(B-1)$, where roh = intraclass correlation, B = average cluster size

12 variables were used. They are:

- % 8th grade education or less
- % never married
- % yearly household income above $15,000
- % never had intercourse
- % does not pay rent
- % living in owner-occupied house valued under $10,000
- % lived at present address less than 5 years
- % respondents unemployed
- % household heads unemployed
- % never been pregnant
- % intend to have 5 or more children
- % living with parents

[3] This is the product of the cost per interview and the design effect.

BIBLIOGRAPHY

Cochran, William G. Sampling Techniques, Wiley, New York, 1963.

Hansen, M.H., W.N. Hurwitz, and W.G. Madow. Sample Survey Methods and Theory, Vol. I, Wiley, New York, 1953.

Kish, Leslie. Survey Sampling, Wiley, New York, 1964.

Zelnik, Melvin, and John F. Kantner. "Sexuality, Contraception, and Pregnancy in the United States," in U.S. Commission on Population Growth and the American Future, Vol. I, Charles F. Westoff and Robert Parke, Jr., eds., Washington, D.C., Government Printing Office, 1972.

ACKNOWLEDGEMENT

I would like to acknowledge the very capable assistance of Christine Amoroso, who carried out many of the computations presented in this paper.