Stratification and Allocation to Reduce Screening Costs in Telephone Surveys

Katherine Ballard LeFauve, Abt Associates
K.P. Srinath, Abt Associates
Susan Arday, Centers for Medicare and Medicaid Services

1. Introduction. In many telephone surveys, the sample from an eligible population is selected by initially screening a larger sample from the general population. The size of the sample needed for screening depends on the desired sample size, and on the proportion of the eligible population in the general population. If this proportion is small, then a very large screening sample is required to obtain the desired sample size. Screening the general population—especially if this screening sample is large—requires considerable time and effort. Furthermore, nonresponse to the screening effort may be correlated with the eligibility criteria; i.e., the target population may have a lower screening response rate than the general population, which will drive screening costs even higher. Therefore, it is of interest to implement designs that reduce screening costs.

One method of reducing screening costs is to stratify the general population into strata with varying proportions of the eligible population, and then to oversample the strata with relatively higher proportions of eligible population. Kalton and Anderson (1986) discuss screening large samples to identify members of a rare population; they suggest oversampling strata in which the rare population is concentrated, and using this data to estimate the characteristics of the rare population.

However, compared to proportional allocation, oversampling strata with high proportions of eligible populations results in disproportional allocation and therefore an increase in the variance of the estimates. Therefore, Srinath (2002) proposed an allocation to strata with varying proportions of eligible population that minimizes the loss in precision due to disproportional allocation, even as it reduces the screening sample size required to achieve the targeted sample size from the eligible population . This method was applied on the three rounds of Survey of the Racial and Ethnic Adult Disparities in Immunization Initiative (READII). The survey's objective was to estimate and compare influenza and pneumonia vaccination coverage rates among the Medicare population for three race and ethnicity groups after a community intervention in selected sites. The groups of interest were non-Hispanic whites, non-Hispanic African Americans, and Hispanics. Because the sampling frame, which consisted of lists of Medicare beneficiaries in each site, could not be used to reliably identify Hispanics, it was necessary to screen the population to identify Hispanic respondents in two sites of the study.

The method was effective in reducing the size of the sample required to achieve the targeted number of completes. However, due to other factors in the survey, including locating problems and nonresponse, it was not possible to scientifically measure the method's overall effectiveness using the survey data. Therefore, for this paper, we have taken the survey data and used it to evaluate the effectiveness of this sampling method, in a model where locating and non-response are *not* issues.

Section 2 of this paper describes the allocation to strata with varying proportions of the eligible population. Section 3 describes the survey and the method of stratification used. Section 4 compares the screening sample size required under the proposed allocation and proportional allocation. In that section we also compare the standard errors under the two allocations; and we look at the decrease in screening costs relative to the increase in the standard error of the estimates. Section 5 presents our conclusions.

2. Allocation to Strata

First, we introduce some notation. Let the general population be divided into L strata in each site. Let M_h denote the size of the general population in stratum h $(h = 1, 2 \ldots L)$. Let $M = \sum_{h=1}^{L} M_h$ denote the overall general population in the site. Let N_h denote the size of the eligible population in stratum h. The proportion of the eligible population in the general population in stratum h is denoted by

$$e_h = \frac{N_h}{M_h}.$$

For example, in the survey described earlier, the general population of size M_h is all the Medicare beneficiaries on the list in a site, whereas N_h is the size of the Hispanic population of Medicare beneficiaries in stratum h. The proportion of Hispanics out of the general population in stratum h is e_h. Generally, the exact eligible population sizes N_h are unknown, though the proportions e_h are known approximately.

Let m_h denote the screening sample size in stratum h. The allocation that minimizes the screening sample size in the survey for sampling Hispanics is

$$m_h = m \frac{M_h \sqrt{e_h}}{\sum_{h=1}^{L} M_h \sqrt{e_h}} \qquad (1)$$

where m is the overall screening sample in the site. We want the screening sample size to be large enough to yield the desired sample size from the eligible population. Let the desired sample size from the eligible population be n. We determine the required screening sample size m as follows. We have

$$n = \sum_{h=1}^{L} n_h$$

where n_h is the sample size from the eligible population in stratum h.

The expected value of n_h is $e_h m_h$. Therefore we have

$$n = \sum_{h=1}^{L} e_h m_h .$$

Now using equation (1) substitute for m_h. We get

$$n = \sum_{h=1}^{L} e_h m \frac{M_h \sqrt{e_h}}{\sum_{h=1}^{L} M_h \sqrt{e_h}} .$$

This can be written as

$$n = m \frac{\sum_{h=1}^{L} e_h M_h \sqrt{e_h}}{\sum_{h=1}^{L} M_h \sqrt{e_h}}$$

Therefore, the required screening sample size m , which will yield an expected sample of size n from the eligible population, is given by

$$m = n \frac{\sum_{h=1}^{L} M_h \sqrt{e_h}}{\sum_{h=1}^{L} M_h e_h \sqrt{e_h}} . \qquad (2)$$

The screening sample size m determined from (2) can then be used in (1) to get m_h.

3. Methodology and Objectives of the Racial and Ethnic Adult Disparities in Immunization Initiative Survey

The National Immunization Program (NIP) in the Centers for Disease Control and Prevention (CDC) partnered with the Centers for Medicare and Medicaid Services (CMS) to assess the disparities between elderly minorities and elderly whites in immunization rates for the influenza and penumococcal vaccines. A survey of Medicaid beneficiaries of three race/ethnic groups was conducted in five geographic areas. These sites were

1. Bexar County, Texas (San Antonio)
2. Chicago, Illinois
3. Milwaukee, Wisconsin
4. Monroe County, New York (Rochester)
5. Selected counties in rural Mississippi

NIP had earlier provided community groups in these areas with funds and support to increase immunization rates among African-Americans and/or Hispanics. Not all groups were targeted in all sites. The following table shows the race/ethnic groups targeted in various sites for both the intervention and the subsequent telephone survey.

Table 1: Population of Interest in Each READII Site

Site/Community	Whites	African-Americans	Hispanics
San Antonio	X		X
Chicago	X	X	X
Milwaukee	X	X	
Rochester	X	X	
Mississippi	X	X	

The objectives of the READII survey were to estimate the vaccination coverage rates among whites, Hispanics, and African Americans in the five communities listed above, and to determine whether the community efforts are helping to bridge the health gap between whites and minority groups. Three rounds of data were collected over three years in order to evaluate the progress toward the goal in each area.

The goal was to complete 400 interviews in each group. The sampling frame for the selection of the sample was the Medicare Enrollment Database (EDB); a sample was selected for each site from the EDB. As indicated earlier, the EDB does not contain reliable information on Hispanic ethnicity. Therefore, to find Hispanics in San Antonio and Chicago, two tools were used to stratify the sample. Census data were used to determine the proportion of population of each zipcode that is Hispanic; and the Passel-Word surname list was used to identify respondents who were likely to be Hispanic based on their last names.

The EDB does not include telephone numbers, so telephone numbers had to be located for selected respondents. Because of the combined effect of nonresponse and lack of telephone numbers, and nonresponse, only around 45% of the sample was able to be screened.

4. Application of the Allocation Procedure to the READII Survey

As stated in the introduction, this method was applied to the READII survey. In the first round of the survey, we had only estimates of the proportion of Hispanics in each stratum. These estimates turned out to be too high, and additional sample had to be added. Later rounds used data from previous rounds to refine these estimates, but more sample still had to be added in each stratum to compensate for missing telephone numbers and nonresponse, which were higher in the high-Hispanic strata.

For the purposes of this paper we will analyze the effectiveness of the method by using the data from the third year of the study. To simplify the analysis, we will assume a100% interview completion rate, since the purpose of this exercise is to compare this particular stratification and allocation procedure to procedures that do not use stratification, or that use stratification with proportional allocation.

The city of San Antonio was divided into 11 strata, based on zipcode and surname. The following table gives the strata, the proportion of Hispanics in each stratum, and the general population of Medicaid beneficiaries. As stated above, the objective was to complete 400 interviews with each group.

Table 2: Percentage of Hispanics by Strata

Stratum	Population of Medicare Beneficiaries (M_h)	Percentage of Hispanics in the population (%) (e_h)
s 1	15,750	90.6
s 2a	12,120	90.8
s 2b	2,960	34.6
s 3a	10,840	88.0
s 3b	8,740	4.7
s 4a	6,120	84.9
s 4b	14,340	5.1
s 5a	6,550	77.9
s 5b	35,020	2.9
s 6a	1,400	53.8
s 6b	19,350	2.1
Total	133,190	37.1%

From Table 2, we see that though the overall percentage of Hispanics in San Antonio is 37.1, one stratum has a percentage of 90.8 while another stratum has a percentage of 2.1.

If we were selecting a simple random sample of beneficiaries in San Antonio, then we would need to screen 400/0.371 = 1,078 persons. The same screening sample size is required if we stratify the population as given in Table 1 and select a sample of 400 that is allocated in proportion to the number of Hispanics in each stratum.

The screener sample size required under the proposed allocation is given below. The numerator and the denominator of equation (2) are shown in Table 3. The required sample size is n, which is equal to 400 in this case.

Table 3: Screening Sample Size under the Proposed Allocation

Stratum	Proportion of Hispanics in the Population (e_h)	$M_h\sqrt{e_h}$	$M_h e_h\sqrt{e_h}$
s 1	0.906	14,992	13,584
s 2a	0.908	11,547	10,481
s 2b	0.346	1,742	603
s 3a	0.880	10,170	8,953
s 3b	0.047	1,885	88
s 4a	0.849	5,639	4,788
s 4b	0.051	3,227	163
s 5a	0.779	5,783	4,507
s 5b	0.029	5,954	172
s 6a	0.538	1,027	553
s 6b	0.021	2,778	57
Total	0.371	64,744	43,949

To compute the screening sample size required under this allocation, we divide the aggregated number in column (3) of the table by the number in column (4) and multiply this ratio by 400. The required sample size is

$$m = 400\frac{64744}{43949} = 400x1.4725 = 589.$$

Under stratification and the proposed allocation, we require only a sample of 589 screeners to identify 400 Hispanics, compared to 1,078 under simple random sampling.

The reduction in screening sample size is 1,078-589 = 489, which is 45% of the sample under simple random sampling. Table 4 shows the allocation of the screening to strata and the expected number of Hispanics in each stratum.

Table 4: Screening Sample Size by Strata

Stratum	Proportion of Hispanics in the Population (e_h)	Screening Sample Size under Proportional Allocation	Expected Number of Hispanics in the Sample	Screening Sample (m_h)	Expected Number of Hispanics (n_h)
s 1	0.906	127	116	136	123
s 2a	0.908	98	90	105	95
s 2b	0.346	24	8	16	6
s 3a	0.880	88	77	93	82
s 3b	0.047	71	3	17	1
s 4a	0.849	50	42	51	43
s 4b	0.051	116	6	29	2
s 5a	0.779	53	41	53	41
s 5b	0.029	283	8	54	1
s 6a	0.538	11	6	10	5
s 6b	0.021	157	3	25	1
Total	0.371	1,078	400	589	400

We had estimated the standard error, under proportional allocation, of the estimate of the percentage of Hispanics receiving a flu shot as 2.455%. In doing so we assumed that the expected sample of Hispanics is constant under repeated sampling. The standard error of the estimate under the proposed allocation is 2.6249%. (The square root of the design effect is 1.0692.) That is, there is a slight increase in the standard error from the proportional allocation method.

We compare the two allocations by taking into account both the reduction in sample size and the increase in standard error as follows. First, we compute the ratio of the reduced sample size under the proposed allocation to the sample size required under simple random sampling. In this case, the ratio is 589/1078 = 0.546. This indicates a reduction of 45.4% in the required sample size to get the desired number of completes.

But the reduced sample size results in an increase in variance. This increase is measured by taking the ratio of the standard error under the proposed allocation to the standard error under simple random sampling. This ratio is **0.026249/0.024550** = 1.0692. We multiply the ratio of the sample size by this ratio and this gives 0.546 x 1.0692 = 0.584. Therefore, the reduction in sample size after accounting for increase in the **standard error** is 41.6% instead of 45.4%.

One can use other methods to decrease the sample size. For example, one could use an even larger sample size than proposed above from strata with a larger percentage of Hispanics, and thereby achieve larger savings. But this would lead to a larger increase in the standard error than what is achieved under the in the proposed allocation. If we adjust the decrease in sample size to offset corresponding increase in the standard error, the resulting adjusted decrease will be smaller than what is proposed under this allocation. In other words, this allocation is optimum when we consider both the decrease in the sample size and the increase in the standard error.

Sampling Whites

We also needed a sample of 400 whites in San Antonio. If we used a sample of 1,078 under simple random sampling, then the usual procedure would be to sample both Hispanics and whites, and to include all screened Hispanics in the sample and take a subsample of whites. Subsampling adds to the complexity of sampling, as we have to keep track of the number of whites selected. Nonresponse to the survey makes this tracking difficult.

The sampling method we are discussing here also provides a means of avoiding subsampling. First, include all the whites in the sample obtained by screening the sample for Hispanics. For example, we expect to get a sample of 189 whites in the sample when we screen 589 selected persons in various ZIP codes. We need a total sample of 400 whites, which means that we need to select another 211 whites in the sample. To select this number, we identify strata in which the number that we would need to select under proportional allocation is larger than the number we have already selected. We select an additional 211 cases from strata in which we were unable to select enough whites, while retaining all the selected whites in the sample from strata in which the number selected is more than or equal to the number required.   This method reduces the loss in precision due to disproportional allocation. The total screening sample size is 800: smaller than the 1,078 required under simple random sampling or under stratified sampling with proportional allocation.

Conclusions:

The proposed method provides the optimal means for reducing the size of the screening sample with minimal increase in the standard error. However, like all methods, it is subject to the realities of data collection.   Because only 45% of the sample could be screened into the READII survey (due both to lack of telephone numbers and nonresponse), the actual screening sample sizes were much larger than the target sample predicted by this method.

We did find that the proportion that could be screened varied greatly between the strata, with respondents in high-Hispanic strata being more difficult to screen. This reduced the proportional savings somewhat, as we had to select more sample in the high-Hispanic strata. However, as the overall size of the screening sample increased, even those smaller proportional savings provided a significant actual savings.

References:

Kalton, G. and Anderson (1986), “Sampling Rare Populations”, *Journal of the Royal Statistical Society,* ser. A, 149: pp.65-82.

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

Srinath, K.P. (2002) “Allocation to Strata in Surveys That Screen for Eligible Populations”, Proceedings of the International Conference on Recent Advances in Survey Sampling , Ottawa, 2002.