Targeting Minorities using Address Based Sampling – A Simulation Study

Pedro Saavedra[1], Francine Barrington[2]
[1]ICF International, 11785 Beltsville Drive, Beltsville, MD 20705
[2] ICF International, 11785 Beltsville Drive, Beltsville, MD 20705

Abstract

Address Based Sampling (ABS) has the advantage of allowing the use of census files to oversample minorities. This study uses a file listing zip code areas and their population for three ethnic groups to identify the optimal strategy for oversampling minorities to maximize the minimum effective sample size for each minority group. Several different sampling strategies were simulated, assuming the ability to randomly sample from zip code areas or combinations of zip code areas. One of the strategies involved sampling zip code areas with replacement using various measures of size that took into account the number of residents in each of the three groups (African-American, Hispanic and other) and then randomly sampling a resident for each time a zip code was sampled. A second set of approaches created strata based on the prevailing minority and the proportion of that minority in the zip code area, and then sampled strata with high minority residence at a higher rate. A third set of approaches used a self-weighting sample and combined or cumulated it with a second sample that used one of the strategies mentioned above. The strategies were evaluated not just in terms of the proportion or number of minorities in each sample, but of the effective sample size due to weighting for each minority.

Key Words: oversampling minorities, effective sample size due to weighting, zip code areas, cumulating samples, PPS sampling without replacement, stratification

1. Introduction

Address Based Sampling (ABS) has the advantage of allowing the use of census files to oversample minorities, allowing the linkage of addresses to prevalence of race and other census information. The different approaches to selecting minorities for Address Based Sampling have also implications for certain forms of ADD samples, as landline telephones can often be linked to an address or zip code area. While cell phone surveys are less directly associated with addresses, it is possible to link exchanges to counties. In either case, the coverage varies significantly by race and other factors. Given a count of the number of persons of each of the three major ethnic groups (Non-Hispanic African-American, Hispanic and all others) and assuming one were able to sample randomly from a census zip code tabulation area or a combination of zip code areas, this study aims to identify the best strategy to maximize the effective sample size for each group. This particular question does not have an obvious answer. In real life, any study involving the sampling of minorities has to contend with the differential response rate for the different ethnic and social class groups, and with the tendency of certain groups to live in larger households than others. In addition, minority groups may often have larger households than others; access to minority groups may be more difficult than other, and lastly,

strategies for sampling different minorities can be at odds with each other. However, in order to examine this question, this simulation study ignores these factors.

The study is motivated by the fact that the objective of many sampling designs is often to obtain national estimates while oversampling minorities. The sample is said to have met its objective if the number of sample respondents exceeds a certain target. However, this is often accomplished at the expense of effective sample size, so that any estimates for the minority populations yield confidence intervals equal to those that would have been obtained from a much smaller random sample.

For this study we limit the sampling methods to those for which a design effect can be estimated prior to drawing the sample, and for which the design effect due to weighting constitutes the totality of the design effect. Therefore we assume that the design does not involve clustering that might contribute to the design effect so that the effective sample size need only consider weighting.

2. Data

The simulations use the 2010 Census Zip Code Tabulation Area File. For each zip code area the number of residents of each minority group was calculated, and race classifications were split into three groups: African-American, Hispanic, and White (which encompassed all other groups). In addition, we assumed that it is possible to sample directly from each zip code area, without having to contend with households or whether every resident has a means of being contacted.

3. Methodology & Simulations

In order to standardize the procedure, the goal was set at selecting 6,000 persons, and obtaining the maximum minimum effective sample size among the three groups. It is not difficult to obtain large minority actual sample sizes (say at least 1,000 African American and 1,000 Hispanics) but it is more difficult to obtain effective sample sizes that large.

One such design includes sampling zip code areas with probabilities proportional to size (PPS) with replacement, sampling a single resident for each time the zip code area is selected. Any measure of size will result in known probabilities, and the challenge is to optimize the measure of size to obtain both the desired number of minority respondents and the desired effective sample size.
A second approach is stratification, where the zip code areas are assigned to strata based on the proportions of each minority in the zip code area. A random sample is then assumed for each stratum, with the sampling fraction being different for each stratum.

A third approach is preferred by some sampling statisticians in order to guarantee the national population estimates while adding a minority oversample. First one draws a national sample, with no oversampling of minorities. Then one draws a second sample from a more restricted domain where minorities prevail. Finally the two sample samples are combined. The non-minority domain respondents retain their same weights, but the weights for respondents from the minority domains are adjusted to account for the two probabilities of selection.

Aside from clustering, there is one other approach which is not considered here. Clearly, if one wishes to select a large sample of minorities with a large effective sample size, this can be accomplished by first selecting a large self-weighting sample and then screening out some or all the non-minority respondents. This method is omitted, not only because it is costly, but because one can always optimize effective sample size by simply having a large sample and screening out a proportion of the non-minorities.

2.1 PPS Sampling

The first design was a PPS sample of zip code areas. In order to avoid clustering the PPS sample, we "drew" one person for each time the zip code area was selected. The effective sample size can be calculated strictly from the weights. This allows a comparison of the different methods, both in terms of the actual sample size for each group and of the effective sample size. Let c_j be the number of non-minorities in zip code area j, a_j the number of African-Americans and h_j the number of Hispanics. Let C_T, A_T and H_T be the national total for each ethnic group. Let p_j be the probability of selection for zip code area j.

Approach 1: $p_j = 6{,}000 * \frac{(c_j + a_j + h_j)}{(C_T + A_T + H_T)}$ The result from one simulation yielded:

Group	*Actual*	*Effective*
Total	6,000	6,000
White	4,421	4,421
African-American	654	654
Hispanic	925	925

This approach yields a self-weighting sample. The expected results are exactly the same as if one were to select 6,000 cases using simple random sample. The effective sample size due to weighting is exactly the same as the actual sample size. It is presented to illustrate the different results which different size measures yield.

Approach 2: $p_j = 2{,}000 * \left[\frac{c_j}{C_T} + \frac{a_j}{A_T} + \frac{h_j}{H_T}\right]$ The result from one simulation yields:

Group	*Actual*	*Effective*
Total	6,000	4,862
White	3,707	3,205
African-American	1,058	872
Hispanic	1,235	1,071

This approach talks the average of the proportion of each ethnic group found in each zip code area. A similar design has been used the HUD Assisted Housing Quality Control Studies, and in other improper payment studies where cluster samples were required to sample participants in multiple programs. The ratios can be used as a base for a number of size measures, consisting of giving a different weight to each ratio. The general form of the size measure of probability is:

Approach 3: $p_j = \left[\frac{6{,}000}{(w_c + w_a + w_h)}\right] * \left[w_c \frac{c_j}{C_T} + w_a \frac{a_j}{A_T} + w_h \frac{h_j}{H_T}\right]$ where w_c, w_a, and w_h are coefficients applied to each ratio to increase or decrease the proportion of each group in the sample.

Several examples follow. Using $w_c = 1$, $w_a = 8$ and $w_h = 4$ we obtained:

Group	*Actual*	*Effective*
Total	6,000	2,649
White	3,197	1,599
African-American	1,594	906
Hispanic	1,209	908

If we use as our criterion the size of the smallest of the eight sizes (actual or effective) this is better than either of the two previous ones.

Another, more extreme, set of coefficients is: $w_c = 1$, $w_a = 12$ and $w_h = 12$, yielding:

Group	*Actual*	*Effective*
Total	6,000	2,187
White	3,199	1,323
African-American	1,286	731
Hispanic	1,515	1,042

While the actual sample sizes are large for the minorities, the African-American effective sample size is smaller than the second approach.

Using $w_c = 1$, $w_a = 25$ and $w_h = 14$ we obtain:

Group	*Actual*	*Effective*
Total	6,000	2,059
White	3,149	1,236
African-American	1,606	870
Hispanic	1,245	860

This results in the largest African-American actual sample size, but the smallest total effective sample size.

Lastly, using the coefficients: $w_c = 1$, $w_a = 4$ and $w_h = 2$ yields:

Group	*Actual*	*Effective*
Total	6,000	3,414
White	3,327	2,113
African-American	1,543	937
Hispanic	1,130	900

This again is promising, as both minority effective sample sizes are at least 900, and both actual minority sample sizes are over 1,100.

However, not every possible equation will yield reasonable outcomes. A size measure was used where $p_j = \left[\frac{6{,}000*(c_j + a_j^2 + h_j^2)}{\sum(c_j + a_j^2 + h_j^2)}\right]$ yielding the following results:

Group	*Actual*	*Effective*
Total	6,000	15
White	1,893	10
African-American	1,321	198
Hispanic	2,786	359

While PPS sampling is elegant in theory, its implementation can be somewhat complex. From the perspective of simulation, one cannot calculate the expected distribution of the ethnic groups in the sample without actually drawing several samples to get an approximation. The sampling is also a two-stage sample, and the presence of a large number of PSUs with one or two cases to be sampled from each is a barrier. For these reasons a stratified random sample may be more practical.

2.2 Stratification

One advantage that a stratified random sample (with equi-probable selection from each stratum) has over PPS is that one can calculate (rather than simulate) the expected number of minorities in each stratum. Indeed after defining the strata and conducting some preliminary calculations as the composition of each stratum, one can use a spreadsheet to calculate the expected actual and effective sample sizes associated with each sample design.

Naturally, the way to control the number of minorities to be sampled from a stratified random sample is to assign the proper allocation to each stratum. If each stratum receives an allocation proportional to its representation in the population, the expected proportion of every ethnic group in the sample will be the same as in the population. Thus the point of interest is examining the effects of different allocations and different stratifications.

There have been a number of studies that have defined the strata using the following criteria:

- Strata 1: At least 60% white
- Strata 2: At least 40% African-American and more African-Americans than Hispanics
- Strata 3: At least 40% Hispanics and more Hispanics than African Americans
- Strata 4: Less than 60% white, but no other group reaches 40%

The units to which these criteria applied have varied from survey to survey. In some cases, the criteria have been applied at more than one level (e.g. first district and then school). In this study they will be applied to zip code areas.

In order to obtain a sample with equal probabilities of selection, one requires allocations of strata where:

- S1 = 4,559,
- S2 = 417,
- S3 = 689
- S4 = 336

This leads to the same expectations as in the first table presented. Suppose, however we redistributed the allocations to:

- S1 = 3,200,
- S2 = 1,200,
- S3 = 1,200,
- S4 = 400

This leads to the following expected actual and effective sample sizes:

Group	*Actual*	*Effective*
Total	6,000	4,920
White	3,787	3,758
African-American	1,080	725
Hispanic	1,183	969

As mentioned above, the strata described in this section have been used in other studies. Efforts to sample using strata with more extreme distributions led to smaller expected effective sample sizes. Repeated simulations using different allocations adding to 6,000 for the four basic strata failed to produce a single design where the minimum effective sample size among the three groups was above 1,000.

4. Conclusions

The study seems to demonstrate the difficulty of sampling for ethnic minorities in the United States if the objective is not merely obtaining a given actual sample size for the two major minority groups, but also obtaining a sufficiently large effective sample size. The approaches presented can be useful in practice, however, if feasible or cost-effective, screening is recommended. When oversampling minorities using ABS, it is important to keep in mind the intended analysis. The relative importance of the *actual sample size* and the *effective sample size* will depend on the analysis. Lastly, when simulating various approaches, it is recommended that one start with a barebones approach and to slowly increase the scope and depth of the simulations, building in differential coverage and response rates.