Selection of Small and Large Schools in State and County School Surveys

Pedro Saavedra, **Tonja Kyle**, James Ross
Macro International Inc., 11785 Beltsville Drive, Suite 300, Calverton, MD, 20705

Abstract

The most common method when conducting state or county school surveys is to first select schools with probabilities proportional to size using enrollment as the measure of size, and then sample the same number of students from every school. Ideally, this would result in the same probability of selection for every student in the frame. One difficulty is that in both state and county surveys there are usually a number of schools with enrollments below the number of students targeted for each participating school. A second difficulty, primarily present in county surveys, is that there are often schools where the calculated probability of selection exceeds 1.0. This paper explores different ways of handling the above problems and their impact on design effect for variables with low and high intra-class correlations.

KEY WORDS: Probabilities Proportional to Size, minimum replacement, systematic sampling

1. Introduction

In a recent thread in the Survey Research Methods Section list a question was asked about cluster sampling (Marker, 2007).

"The correct formula is $p_{ij} = m*(n_j/N)*(n/ n_j)$. You are selecting the clusters proportional to their number of second-stage (you refered to them as level 1) units, and there are m chances of the cluster being selected. Then you are selecting n second-stage units regardless of the number of units in cluster j.

Assuming your measures of size are correct the n_j cancel out leaving $p_{ij} = mn/N$ which is constant across all cluster."

That comment elicited the following remark (Chapman, 2007):

"I agree with David's response, but there are two cautions. The formula assumes that there aren't any clusters so large that $m*(n_j/N)$ would exceed 1. Second, the formula assumes that there aren't any clusters so small that n/ n_j exceeds 1." [2]

This paper discusses the two exceptions and how one may take them into account in sampling elementary and secondary school students where schools are clusters and students are sub-sampled from each cluster. There are different ways that one can modify a sample design to take into account very large or very small clusters, and different approaches can be optimal under different circumstances. In this paper we present various considerations which will lead to a preference for one approach or another, and present some simulations that illustrate the principles presented.

2. The Basic Design and Notation

This section will describe a basic design in a ideal world where one is to select m schools and sample n students from each school, for a total of mn students in the sample. The sampling would use probabilities proportional to size (PPS). In this ideal world, there is a sufficiently large number of schools so that if N is the total number of students in the population and N_j is the number of students eligible for selection from the school then $p_j = m(N_j / N)$ is never greater than one and N_j is never smaller than n. This condition is easily met when one eliminates small schools from the frame and is conducting a survey at the state or national level. Subsequently one must distinguish between the initially designated n and m and the actual numbers m' and n' that design modifications end up selecting.

The number p_j in this case is the probability of selection one would assign to each school. Schools may be selected with this probability using any of several methods, and the merits of each approach are beyond the scope of this paper. There have been good reasons to select the same using the random systematic selection approach developed by Goodman and Kish (because implicit stratification is rather effective), Chromy's approach (allowing a particular variance estimation formula) or Pareto sampling (useful when separate samples by grades have to be linked). Only the first method will be described to facilitate discussion later in the paper.

In order to sample the schools with PPS one can sort the schools in any random or systematic order.

In practice, when doing a state survey, one efficient practice is to assign a random number to each county and a random number to each school. The counties are sorted by the random number and schools are sorted within counties by the school random number. This approach guarantees that counties will be within one school of their expectation given their size.

Once the schools have been ordered, a random number p_0 between zero and one is selected as a starter. Number the schools from 1 to M (the total number of schools). Now let s_j be the selection indicator for school j and $c_j = p_0 + p_1 + \ldots + p_j$ then we make $s_j = \text{lim}(c_j) - \text{lim}(c_{j-1})$ where lim(x) denotes the largest interval that is smaller than x. It is easy to see that the s_j will be 1 or 0 and will add up to precisely m. This approach is simply the PPS systematic or Goodman-Kish method. The use of probabilities as the sampling interval facilitates understanding the procedure and using a simple function in programming the sample selection,

Unfortunately, what may work when one is to select state samples with no schools so large that their calculated probabilities exceed 1.0 or their enrollment is smaller than n, does not work when one has smaller domains or must include smaller schools. There are, however, a variety of methods that can handle these situations, and the choice of an approach depends both on the statistical properties of the variables to be estimated and the logistic considerations in selecting schools.

It should be noted that in many school surveys costs by schools and students are not easily estimated. Often surveys are conducted by school system personnel and the cost is in time which would have been used to do something else. Yet, an excess number of students or schools would lead to an increased burden, and this must be taken into account, though it cannot always be easily quantified. In some instances the number of schools to be selected or the number of students to be initially sampled or targeted (i.e. the expected n) is contractually mandated, and in others it is somewhat flexible within budget constraints.

The burden issue can even have political considerations. A design with unequal n_j by selected school may be perceived as posing unequal burdens or as confusing to the local team that manages the survey for the school system, whereas in other cases this is not an issue at all. Some student surveys may involve individual interviewing of students, and therefore, it might be desirable that n_j be a multiple of the number of students that can be interviewed in one day. Other surveys may not have this constraint. Thus the selection of approach will depend on many considerations that may be specific to the survey and may even change from cycle to cycle for the same survey as concerns about burden, budgetary considerations, modes of data collection and information about the properties of the variables of interest inform the decisions.

One assumption that will be made throughout the discussion is that there is a need to keep the targeted n (the expected number of students) fixed, but that some modification of the actual number of distinct schools selected is permissible. The ambiguous nature of cost information in some surveys and the inclusion of the sample size in some contracts makes this a realistic assumption for purposes of this paper.

Finally, it should be noted that in many schools surveys intact classes are sampled, and thus the method of selection is not to select n students from the school, but to target the selection of n students from the school. Thus if a school has an enrollment of 500 and n is 100 one would target 20% of the classes. The actual number selected may be larger or smaller, depending on the size of the classes selected. The relevance to this discussion is that n need not be an integer if it represents the targeted sample rather than the actual sample. This in turn simplifies the discussion and avoids the issue of rounding.

3. The Problem of Certainty Schools

We will first address the situation where some schools are so large that $p_j > 1$. There are various ways of approaching this situation, and we will discuss three of them. We will refer to them as:

1) Probability Minimum Replacement (PMR)
2) Sampling Without Replacement (WOR)
3) Proportional Allocation Method (PAM)

Each has advantages and disadvantages, and there are variations for each.

3.1 Probability Minimum Replacement

The term was introduced by Chromy (1979) in the paper where he introduced his sampling algorithm. In general if $p_j = i_j + q_j$ where i_j is an integer and $q_j < 1$ then unit j would be sampled i_j times with certainty and have a probability of q_j of being sampled an additional time. For each time selected n students would be sampled from the school.

Chromy presented a direct approach to sample with PMR. In the method presented in the previous section, the procedure is followed as above understanding the s_j could be an integer other than 0 or 1 if $p_j > 1$. In fact, $s_j = \lim(p_j)$ or $s_j = \lim(p_j)+1$.

If one uses another procedure such as Pareto sampling, one can simply implement the following steps:

1) Sample the schools where $p_j > 1$ exactly i_j times.
2) Let i equal the sum if the i_j.
3) Sample m-i schools with probability q_j each.

Now, this procedure will yield fewer than m distinct schools because the certainty schools may be sampled more than once. The major advantage of this approach is that the number of students selected from each school will be a multiple of n. If one is sampling intact classes and wants to fix the number of classes to be sampled this is a good option. If one is interviewing students and it takes one day to interview n students, this is a way of avoiding an assignment of interviewees that would occupy the interviewer for a fraction of a day.

There are two ways in which one can assign weights to students under this scheme (Saavedra, 2005). One of them is to calculate the probability of selection before the first selection has taken place (unconditional weights). This should be the same for every student in the population, and hence there will be no design effect due to weighting. The alternative procedure is to assign a probability of one to the certainty schools, and treat the number to be selected as a fixed sample size for that school (disregarding its dependence on the number of times the school was selected). The second weighting approach (called conditional weights because the probability is conditional on the results of the first stage sample) does yield unequal weights, but is preferable for high intra-class correlation variables.

Thus the advantages are:

1) Equal weights (with unconditional probabilities).
2) Samples of students are always a multiple of n.
3) Variance formula based on PSUs if Chromy method is used.

The disadvantages are:

1) Two larger schools of equal sizes may have different student sample sizes.
2) If a variable has a high intra-class correction, sampling more students from a school may not add much to the precision.

3.2 Sampling Without Replacement

The second method is simply sampling without replacement. First the p_j are calculated. The certainty schools are selected, and their number is subtracted from m. A new set of p_j are then calculated. The procedure is repeated until there are no more certainty schools. Now the remaining schools are sampled with a PPS approach. Exactly n students are selected from each school.

A major advantage of this approach is the preservation of the original m as the number of distinct schools selected. A major disadvantage is that students from the larger schools will have smaller probabilities of selection, and hence larger weights. This last factor will produce a design effect due to weighting. If a variable has a high intra-class correlation (schools tend to be homogeneous with respect to the variable) then this is countered by the larger number of distinct schools. Indeed, one can see that if students in a school all tend to answer the same way, there is no point in increasing the sample size within that school. By the same token, if a variable has a very low intra-class correlation, then there is no point in having more clusters, and increasing the sample in large certainty schools will simply reduce the design effect due to weighting.

The WOR approach sacrifices the weighting design effect for the clustering design effect. The fact that m and n are adhered to allows the survey to also adhere to a precise budget and to determine what logistic difficulties will need to be met prior to drawing the sample. This may or may not be a significant advantage.

3.3 Proportional Allocation Method

This approach first identifies the certainty schools and makes each of them a stratum. The non-certainty schools form a separate stratum. Now, let N_j be the population of each stratum. If m is the designated number of schools (the number that would be selected if there were no certainty school) and n the number that would be selected per school if there had been no certainty schools, then $mnN_j /$

N would be the number of students to be selected from each stratum.

Let N_0 be the non-certainty number of students enrolled, and let k be the number of certainty schools. Now, one has two choices. One can retain the number of schools by subtracting the number of certainty schools from m and defining $n' = mnN_j / N(m-k)$ or one can redefine m as $m' = mN_j / N$ (rounded) and preserve the original n to be sampled. The first approach preserves the number of schools to be selected and the second preserves the sample size to be selected per non-certainty school.

4. The Problem of Small Schools

The second problem which often arises in school surveys is that of schools whose enrollment (of number of enrolled students in the population in scope for the survey) is smaller than n. There are three distinct approaches that are commonly used:

1) Do nothing at the first stage and sample all students in any such school that is selected (the Unequal Weights option).
2) Pair up schools so the sum of the enrollments exceed n and treat each pair as if it were a single school at the first stage of selection (the Pairing Option).
3) Assign the small schools a probability of mn/N, but assign the other schools the probability $m(N_j / N)$ where m is the originally intended number of schools (the Minimum Probability Option).

Each of the approaches has minor variations, each of which in turn has advantages and disadvantages. We will mention that an additional approach is to simply consider students attending very small schools to not be in the population of interest, and one frequently finds surveys where students in very small schools are automatically out-of-scope.

4.1 The Unequal Weight Option

This option is most often used because the problem was not considered at the first stage of sampling, or because the enrollment in the small schools is not that much smaller than the number to be sampled for each school. The probability of selection for a student in a school that is large enough is mn/N. But for a small school, that probability will be $m(N_j / N)$ since all students will have to be selected, and if $N_j < n$ then the probability of selection of a student in the school will be smaller that that of students in larger schools. This means that the weights for those students, if calculated as the inverse of the probability of selection, will be larger.

The design effect due to weighting will obviously be a factor in using this approach, as will be the slightly smaller sample size. In practice the effect is usually not that large if few small schools are selected. One situation where the number of schools becomes an issue is where multiple surveys are administered, and hence n is rather large. At this point the number of schools where $N_j < n$ (where n is now the number needed for all the surveys) can be considerable.

One effort to avoid the loss of sample size has been to spread the shortfall among all the schools that are sufficiently large. This may retain the sample size, but the effective sample size due to weighting will still be smaller.

4.2 The Pairing Option

This option involves pairing schools and treating each pair as a school. There are several ways of doing this. One way is to pair small schools. One orders the small schools and pairs the largest small school with the smallest, the second largest with the second smallest and so forth. If need be some schools with enrollments slightly larger than n can be added to the schools that are to be paired, so that the enrollment of each pair exceed n. This approach has the advantage that the sample will definitely include students from both members of a pair. The sample can be spread proportionately between the two schools.

A second approach involves pairing a small school with a nearby school, even if the other school is much larger. Then a second step is taken. Let N_1 be the enrollment (less than n) of the small school and N_2 be the enrollment of the larger school, where $N_1 < n < N_2$.
One can always design a sampling scheme so that there will be a substantial probability that the entire sample will be drawn from the larger school. If $N_1 + N_2 > 2n$ one can, for example, divide the pair into two equal partitions, so that one is entirely contained in the larger school. One of the partitions is selected and if both schools are represented the sample is selected proportionately. There are more efficient methods for achieving this objective, but they are beyond the scope of this paper.

This second approach is particularly helpful when the data collection method requires visiting the schools, and sampling two schools in proximity to each other becomes more cost efficient. In practice, however, when there is a large frame of schools, it may be difficult to pair up every small school with a neighboring one.

4.3 The Minimum Probability Option

In a survey where every school is sufficiently large the probability of selection of every student will be mn/N, where m is the number of schools to be selected, n is the number of students to be selected from each school and N is the number of students in the population. Thus, if every student is to have the same probability of selection, students in a school where every student is to be selected, the school itself must have probability of selection mn/N. Therefore, a simple way of assuring equal weights is to simply assign every small school the probability of selection mn/N and retaining for the other schools their probability proportional to enrollment.

This approach makes the sum of the probabilities not an integer. The result is that before the sampling takes place, the number of schools to be selected (a number now greater than m) will not be known. Neither will the exact number of students selected. The variation in the number of small schools selected if one uses the Goodman-Kish (PPS systematic) approach can be controlled by grouping all the small schools in an implicit stratum. Ordering the school by enrollment within the stratum should also reduce the variation in the total number of students selected. Given that one can accept the slight variation in the number of schools selected, this approach seems the more practical of the three approaches, and it is easiest to implement.

5. Integration of the Two Issues in County Samples

If the number of schools in the frame is fairly large, one seldom has the first difficulty, but one may still have small schools. If the number of students to be sampled is small, the second difficulty seldom arises, but one may still have certainty schools. Small counties typically require that all schools be sampled and hence proportional allocations assigned to each school. However, both problems can arise in large counties, where the sampling fraction for schools is relatively large and one has both large and small schools.

Essentially this situation calls for three types of schools. Those that are certainties will be assigned an allocation equal to mnN_j/N. Those with enrollments under n will be selected with probability mn/N and all the students will be sampled. And the remaining schools will be selected with probability mN_j/N and n students will be sampled.

An example will clarify the issue. A county has 43 high schools and one wishes to select 6,000 students, ideally 200 from each of 30 high schools. Table 1 presents the data to illustrate the design. As can be seen twelve schools turned out to be certainties and six schools had enrollments under 200. The total number of different schools selected in the example was 29, though 28 was also possible (the expectation for the number of distinct schools was 28.43). A total of 6,097 students were sampled, though 6,000 was the expectation.

References

Chapman, David, Message in the Survey Research Methods Section (SRMSNET) listserv, 2007.

Chromy JR. Sequential sample selection methods. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp 401-406, 1979.

Goodman R. and Kish, L. (1950) Controlled Selection - A Technique in Probability Sampling, *Journal of the American Statistical Association*, 45, 350-372.

Marker, David, Message in the Survey Research Methods Section (SRMSNET) listserv, 2007.

Saavedra, P.J. (2205) Comparison of Two Weighting Schemes for Sampling with Minimal Replacement. Presented at the Joint Statistical Meetings, Minneapolis, MN

Table 1: Example of County Sample Selection

School	Enrollment	Expectation	Probability	Cumulative Probability	Sample Flag	Sampled Students
1	4,291	1.305	1.000	1.000	1	261
2	4,171	1.268	1.000	2.000	1	254
3	4,065	1.236	1.000	3.000	1	247
4	3,771	1.147	1.000	4.000	1	229
5	3,765	1.145	1.000	5.000	1	229
6	3,720	1.131	1.000	6.000	1	226
7	3,679	1.119	1.000	7.000	1	224
8	3,629	1.103	1.000	8.000	1	221
9	3,594	1.093	1.000	9.000	1	219
10	3,529	1.073	1.000	10.000	1	215
11	3,461	1.052	1.000	11.000	1	210
12	3,409	1.036	1.000	12.000	1	207
13	3,287	0.999	0.999	12.999	1	200
14	3,146	0.957	0.957	13.956	1	200
15	3,112	0.946	0.946	14.902	1	200
16	3,068	0.933	0.933	15.835	1	200
17	2,913	0.886	0.886	16.721	1	200
18	2,841	0.864	0.864	17.584	1	200
19	2,773	0.843	0.843	18.427	1	200
20	2,772	0.843	0.843	19.270	1	200
21	2,768	0.842	0.842	20.112	0	0
22	2,701	0.821	0.821	20.933	1	200
23	2,666	0.811	0.811	21.744	1	200
24	2,557	0.777	0.777	22.521	1	200
25	2,502	0.761	0.761	23.282	1	200
26	2,258	0.687	0.687	23.968	0	0
27	2,223	0.676	0.676	24.644	1	200
28	2,097	0.638	0.638	25.282	1	200
29	1,559	0.474	0.474	25.756	0	0
30	1,553	0.472	0.472	26.228	1	200
31	1,540	0.468	0.468	26.696	0	0
32	1,354	0.412	0.412	27.108	0	0
33	1,039	0.316	0.316	27.424	1	200
34	925	0.281	0.281	27.705	0	0
35	479	0.146	0.146	27.851	0	0
36	469	0.143	0.143	27.993	0	0
37	251	0.076	0.076	28.069	0	0
38	193	0.059	0.061	28.130	0	0
39	174	0.053	0.061	28.191	0	0
40	156	0.047	0.061	28.252	1	156
41	94	0.029	0.061	28.313	0	0
42	73	0.022	0.061	28.374	0	0
43	44	0.013	0.061	28.434	0	0
Total	98,671	30	28.43	---------	29	6,097