Comparison of Two Weighting Schemes for Sampling with Minimal Replacement

Pedro J. Saavedra
ORC Macro, 11785 Beltsville Dr., Calverton, MD 20705

Abstract

In his 1979 paper on PPS sampling with minimal replacement (SMR) for multistage samples, Chromy developed an estimator based on the inverse of the expected number of times a PSU would be selected. This yields an unbiased Horvitz-Thompson estimator, which takes into account the sampling process at every stage. An alternative estimator also is used with SMR based on Stage 1 weights computed as the inverse of the PSU probabilities of selection. With this approach, the allocations at the second stage are taken as a given and not as the result of a probabilistic process at Stage 1. This is exactly the estimator most commonly used for sampling without replacement, and thus many statisticians may be more familiar with it than with the Chromy estimator. This study compares the design effect for the two estimators using real sampling frames with varying numbers of selected PSUs with and without perfect correspondence of the number of units per PSU and the size measure. The alternative estimator was more efficient when the intra-class correlation was high.

Keywords: two stage sample, estimator, simulations, sequential sampling

1. Introduction

In his 1979 paper, James Chromy introduced the term “probability minimum replacement” (PMR) to designate an alternative to probability replacement (PR) designs and probability nonreplacement (PNR) design. Since the terms Sampling With Replacement (WR) and Sampling Without Replacement (WOR) have become more usual, the term Sampling With Minimal Replacement (WMR) will be used in this paper. Chromy introduced the term to describe one particular sampling scheme, but the concept is applicable to any of a number of schemes where only PSUs which are selected with certainty can be selected more than once.

Consider a cluster sample design where m Primary Sampling Units (PSUs) or clusters are to be selected among M, where the number of units is a measure of size, where cluster j has N_j units, and where N is the total number of units across all clusters. The expectation of a cluster can be expressed as $E_j = mN_j / N$ (one can, of course, use a size measure that correlates with the number of units in the population, and most of what is said here will hold). In a WMR design, if all the expectations are less than one, the expectations represent probabilities of selection, and the design is identical to a WOR design. Now, let Int(E_j) be the integer portion of E_j and Frac(E_j) be the fractional part. A design is WMR if the number of times cluster j is selected is Int(E_j)+ 1 with probability Frac(E_j), and Int(E_j) times with probability 1-Frac(E_j),

2. WMR Sampling Methods

Any WOR method can be made into a WMR method in the following way. First one calculates the expectations E_j as described above. Then one selects each unit Int(E_j) times where Int(E_j) is the integer part of E_j . Finally one assigns a probability of P_j=E_j-Int(E_j) to each unit and selects the units with that probability. If a unit selected with certainty Int(E_j) times is selected, then it will be in the sample Int(E_j)+ 1 times.

However, there are at least two methods where the WMR sampling can be implemented directly. One is an extended Goodman-Kish (1950) approach. Once expectations have been calculated, the PSUs are sorted randomly (or randomly by stratum) and labeled 1 through j. Let C_j=(E_1+E_2+ …E_j) where the E_j are expectations of PSUs 1 to j. Let s_j=Int(C_j+r)-Int(C_{j-1}+r) where r is a starting random number (0<r<1) and s_j the number of times PSU j is sampled. This approach is the WMR equivalent of Procedure 2 in Brewer and Hanif (1983).

A second method is the one introduced by Chromy in his 1979 paper, and presented as Procedure 50 in Brewer and Hanif (1953). The Chromy method begins the same way as the extended Goodman-Kish approach, and hence one can begin with the same notation and terminology. The units are ordered and the terminology is the same as before. Let I_j = Int(C_j) and F_j =Frac(C_j) Let n(j)= s_1+ s_2+ …+ s_j. It can be seen that for any j, n(j) will be I_j or I_j + 1. The selection is sequential. If F_j =0 then n(j)= I_j . Now let r_j be a random number associated with PSU j. Given that the sample has been drawn up to unit j-1, n(j-1) must be I_{j-1} or I_{j-1} + 1. The following rule will be used to define n(j):

1) If n(j-1)= I_{j-1} and $F_j \geq F_{j-1} \geq 0$ and $r_j < (F_j - F_{j-1})/(1-F_{j-1})$ then n(j)= I_j +1
2) If n(j-1)= I_{j-1} and $F_j \geq F_{j-1} \geq 0$ and $r_j \geq (F_j - F_{j-1})/(1-F_{j-1})$ then n(j)= I_j
3) If n(j-1)= I_{j-1} and $F_{j-1} > F_j > 0$ then n(j)= I_j
4) If n(j-1)= I_{j-1} +1 and $F_j \geq F_{j-1} \geq 0$ then n(j)= I_j +1
5) If n(j-1)= I_{j-1} +1 and $F_j > F_{j-1} > 0$ and $r_j < (F_j / F_{j-1})$ then n(j)= I_j +1
6) If n(j-1)= I_{j-1} +1 and $F_j > F_{j-1} > 0$ and $r_j \geq (F_j / F_{j-1})$ then n(j)= I_j

Now, one simple calculates s_j = n(j)-n(j-1), where, once again s_j is the number of times PSU j is sampled.

The Chromy method has one advantage over the extended Goodman-Kish. Consider a population of five units where two are to be sampled with probabilities .1, .6, .6, .1, and .6 . There is no way that the two units with the smallest probabilities can be both sampled. However, using the Chromy approach (where one randomly sorts the units as a starting point) one can easily see that a very low r_j for the first and fourth unit will lead to their selection. This means that a variance estimate can be obtained, and Chromy presents one in his paper.

3. Weighting aproaches

The question examined in this paper is what the weights should be for a WMR sample. Suppose that indeed we decide to sample k_j n units from each PSU, where k_j is the number of times cluster j was sampled. There are two possible ways of looking at the design, leading to two different sets of weights. The first is equivalent to the Chromy's approach. One determines the probability of selection of a particular unit in the population. In order to generalize, let us suppose that the measures of size are exact (we will discuss the other situation later) and that as before we sample n units for every time the PSU is selected. It is easy to realize that the probability of selection of a unit is Frac(E_j)(Int(E_j)+1) n/ N_j + (1-Frac(E_j) Int(E_j)n/N_j= (n/N_j)(Int(E_j)+Frac(E_j)) = (n/N_j)E_j = (n/N_j) (m N_j/N) = mn/N. In other words, every unit has exactly the same probability of selection. Of course, if it turns out that the real number of units of the PSU is N_j' then the probability becomes (mn/N) (N_j'/ N_j).

However, there is another way of looking at the same probability of selection without altering the sampling method. The sample could be treated as a WOR sample. The probability of selection would be P_j = min($E_{j,}$1) and the probability of selection of a unit would be P_j(n_j/N_j) where n_j is the number of units sampled from cluster j, which in turn depends on how many times the cluster was actually selected. This approach treats the n_j as if they had been arbitrarily selected.

For either approach we can use the inverse probability as a weight. We will call the first weight the unconditional (because the weight are not dependent on the first stage results) or Chromy weights and the second weight the alternateor conditional weights. Now, in the case where we know the exact number of units in each cluster beforehand, the unconditional weights will add up to the population. The conditional weights will not, but can be adjusted so that they do add up to the population. This makes no difference in estimating means and proportions, but it will affect the estimates of totals.

Let us take an example. Suppose a design calls for 100 clusters with 20 units to be sampled per cluster, and there are 2,000,000 units in the population. Suppose a cluster has 44,000 units. The expectation is E=100(44,000/2,000,000) or 2.2. This means there is a 20% probability that the PSU would be selected three times and an 80% probability that it be selected twice. The probability of selection will be 1/1,000 for all units, whether in this cluster or not. On the other hand, the unadjusted conditional weight will depend on the number of times the PSU is selected. If the cluster is selected twice, the probability of selection of the units in that cluster will be 40/44,000 and the weight will be 1,100 whereas if the cluster is selected three times, the probability of selection will be 60/44,000 and the weight will be 733.33.

It would seem that the unconditional weights yield an additional design effect due to weighting, and that the conditional weights would be preferable. But a contrived example will show that this is not necessarily so. Suppose that in the above example the cluster in question were the only certainty cluster. Suppose furthermore that there were a variable with a uniform value of 10,000 in the cluster and 1,000 in every other cluster. It is easy to calculate that the population mean is 1,198 and that using unconditional weights the sample estimates could be 1,180 or 1,270, depending on the number of times the certainty cluster was selected. However, using the conditional weights, the estimates would be exactly 1,198 regardless of how many times the certainty cluster was selected.

The totals are a different matter. The population total is 2,396,000,000. The estimates using the unconditional weights are 2,360,000,000 80% of the time and 2,540,000,000 20% of the time, with the average estimate across samples equaling the population. The WOR sample yields 2,400,000,000 every time, showing a bias, but being closer to the population all the time. And the bias can be corrected by adjusting the weights to the known number of units in the population. Thus in this contrived instance the conditional weights yield the lower mean square error, even though they do show a bias if the totals are uncorrected.

In order to explore the two kinds of weights, we decided to do simulations using real data, sampling from a frame where the values were known for every unit in the frame. The first set of simulations used States as the PSUs and schools as the units. We used schools in an old Common Core of Data (CCD) file for which at least one of grades 6, 8, 10 and 12 were present and for which the proportion of students of each race was reported. Estimates were made for two variables, the number and percent of schools with a sixth grade and the number and percent of schools with enrollments less than 80% white.

4. The School Simulations

Several simulations were conducted. In each, States were selected as PSUs and schools were selected at the second stage. In two, the number of schools in the States was used as a size measure, and in the other two, the number of students was used. Each of these simulations was first conducted using the extended Goodman-Kish procedure and later repeated using the Chromy procedure. Estimates were obtained using the unconditional estimator developed by Chromy and the alternate conditional estimator.

The school simulations were conducted using a subset of the CCD, including schools with ethnicity information and at least one of grades 6 through 10 in 49 States and the District of Columbia (Hawaii was excluded, as its number of schools in the frame was too small). The process of drawing a sample began with a list of States indicating the number of units (schools in this case) in each State. For the first two simulations the proportion of all schools in the frame that were found in that State were multiplied by the number of States to be sampled (40 in this case) to provide the expectation of selection for the State. If this expectation was less than 1.0, it represented the probability that the State would be selected once. If it was greater than 1, the integer part was equal to the number of times the State was to be selected with certainty, and the fractional part became the probability that the State would be selected one additional time. The sum of all expectations equaled 40, or the number of States to be sampled.

Two estimates were selected for the simulation. The first was the percentage of schools nationwide, among those with grades 6 through 10 that included a sixth grade. The second was the percentage of schools nationwide with more than 20% minority enrollment. These variables have very different distributions across States. The first has a low intra-class correlation, as the percentage of schools that include a sixth grade is not that different from State to State. The second has a high intra-class correlation, as one finds that in some States, such as Vermont there are no schools with more than 20% minorities and in others, such as the District of Columbia, every school has at least 20% minority.

Several sets of samples were drawn, selecting 1000 schools, 40 States and 25 schools per State, for each time the State was selected. The frame had 47,104 schools, 44% of which had over 20% minority enrollment and 63% of which had a sixth grade.

5. The Assisted Housing Simulations

Using a count of assisted housing tenants in three programs by county (obtained as part of an evaluation of assisted housing recertification practices) the counties were clustered into 1,196 PSUs. Of these 100 clusters were selected and the selection of 40 tenants per cluster was simulated. Estimates were obtained of the percentage of tenants in each of three programs: Traditional Public Housing, Tenant-Based Section 8 and Project-Based Section 8. As with the school simulations, 1000 samples were drawn.

6. Evaluation of Results

Though the relative effectiveness of the Goodman-Kish and Chromy approaches was not a primary concern of this paper, comparisons were made for each estimate using an ordinary t-test. Two dependent variables were used. The first was the absolute value of the difference between the estimate and the parameter obtained from the frame. The other was the square of that difference. The squared deviations criterion, of course, gives greater weight to large discrepancies.

Then for each parameter being estimated within a method, the two estimators were compared, also using both absolute deviations and squared deviations. Only these comparisons were compared using matched paired t-tests. To present an example, suppose y were the parameter being estimated (e.g. the proportion of schools with sixth grades in the country). Let $\hat{y}_{i1}$ be the estimate using the unconditional weights for the sample drawn at sample i and $\hat{y}_{i2}$ be the corresponding estimate using the conditional weights. Now let $d_i = |\hat{y}_{i1} – y| - |\hat{y}_{i2} – y|$. A t-test was used to determine if one of the two estimators was significantly closer to the population parameter than the other, by determining if the d_i were significantly different from zero. A similar test was used using $d_i' = (\hat{y}_{i1} – y)^2 - (\hat{y}_{i2} – y)^2$. Coefficients of variations were used for descriptive purposes in the comparisons.

7. Results

Three designs are reported in this paper:

1) 40 States, 25 schools per State per selection, exact size measures
2) 40 States, 25 schools per State per selection, approximate size measures
3) 100 PSUs, 40 tenants per PSU, exact size measures.

Comparing the precision of the two sampling methods, no difference was found for either estimator for the high intra-class correlation variables, but for a low intra-class correlation variable (proportion of schools with a sixth grade) the Chromy method was slightly better for either estimator ($p<.05$ using squared deviations) when the measure of size was exact. All the comparisons of the two estimators yielded similar results regardless of sampling method, so only the results using the Chromy method will be presented. The absolute deviation tests and the squared deviation tests led to the same conclusions when comparing estimators.

For the first design, the conditional estimator was significantly better for estimating proportion of schools with over 20% minority students (high intra-class correlation variable). There was no difference for the low intra-class correlation variable. For the second design, where the measure of size was not exact, the unconditional estimator was better for the low intra-class correlation variable and the conditional estimator was better for the high intra-class correlation variable.

For the third design, for two out of the three estimates, the conditional estimator performed better. Indeed, these two were the estimates of the proportion of traditional public housing, more prevalent in large metropolitan areas, and tenant-based Section 8 housing (more prevalent in less densely populated counties). Both of these tend to have a higher intra-class correlation. The project-based Section 8 housing is more widespread, and is not managed by the local PHA, so the intra-class correlation can be expected to be lower.

Table 1 presents the results of the estimator comparisons:

Variable	Uncond.	Condit.	P<.01
Des.1 - low	.01901	.01888	
Des.1 - high	.05432	.05247	*
Des.2 - low	.02770	.02806	*
Des.2 - high	.07864	.07740	*
Design 3 – a	.05084	.04959	*
Design 3 – b	.05015	.04886	*
Design 3 – c	.05318	.05293	

8. Conclusions and Summary

The results are consistent with the observation that Chromy’s estimator for samples where PSUs are selected with minimal replacement works best for variables with a low intra-class correlation, whereas the conditional estimator works best for variables with a high intra-class correlation. It should be noted that it is precisely for variables with a low intra-class correlation that one would want to use sampling with minimal replacement. If there was a high degree of homogeneity within PSU, it would not be cost-effective to increase the allocations of the larger PSUs. On the other hand, when the intra-class correlation is low, it is more important to sample each unit with close to the same probability of selection.

References

Brewer, K.R.W., and Hanif M. (1983) Sampling with Unequal Probabilities, New York, Springer.

Chromy, J.R., 1979. "Sequential Sample Selection Methods". Proceedings of the American Statistical Association, Survey

Goodman R. and Kish, L. (1950) "Controlled Selection - a Technique in Probability Sampling" J. Americ. Statist. Assoc. 45, 350-372.