SAMPLING FROM EXISTING SAMPLES

Thomas Krenzke, Keith Rust, Leyla Mohadjer, Westat
Thomas Krenzke, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Survey sampling, measures of size, ultimate sampling units, domains of interest

1. Introduction

Sampling from existing samples may occur for several reasons. A survey is restricted by costs, and if its design objectives are similar to the objectives of a larger sample survey, then a sample probably can be selected with much lower cost by using the original survey's sample instead of having to build a sampling frame and select an entirely new sample. Another reason is that the cost or burden of data collection may be very high, so that one would want to gather inexpensive data from a large group of units, before determining the most efficient design for a smaller sample of units.

The objective of this paper is to present recent experiences that involved sampling from existing samples, including a detailed description of one case of subsampling. Several issues arose from the experiences. We will explain how we used information about the original survey's design, the planned analysis on the new sample, and the types of sampling units, so that the subsample design could maintain or remove features of the original survey's sample design, and incorporate any changes in the Ultimate Sampling Unit (USU).

The main example for this paper is related to the classroom samples for the Classroom Based Writing Study (CBWS), which were selected from samples of Primary Sampling Units (PSUs) and schools from the 1998 National Assessment for Educational Progress (NAEP). The CBWS is sponsored by the National Center for Education Statistics (NCES). The objectives of the CBWS and NAEP were similar; that is, to select a sample of students and administer tests. However, the ultimate sampling units were different. Issues surrounding oversampling domains, changing measures of size, and keeping the variation of the sampling weights to a minimum are discussed.

2. Some Reasons for Using Existing Samples

Using existing samples can lower costs of fieldwork. For example, the 1992 National Adult Literacy Survey (NALS), conducted for NCES, included a sample of PSUs comprised of counties or groups of counties. Within PSUs, secondary units, or area segments that are individual Census blocks or a group of blocks, were randomly selected. An extensive listing of households was constructed within the area segments and households were sampled from the listings. The NALS measure of size was designed to give segments with a high concentration of minority groups a higher chance of selection than nonminority segments. As explained in CSFII (1997), the Continuing Survey of Food Intakes by Individuals (CSFII), sponsored by the Agricultural Research Service of the U.S. Department of Agriculture, had similar objectives to NALS, and to reduce sampling costs, advantage was taken of the NALS area sample listings of households. There was a 56% overlap between the CSFII and NALS area segments, as maximized through a sampling procedure. The use of the master segment sample reduced the cost of sampling, mapping, and listing by about 40%.

CSFII did not require any oversampling of the highly concentrated minority segments. A procedure was used to adjust for the higher-than-desired selection probabilities of these segments. The procedure was similar in motivation as the procedure explained in the example that will be discussed in Section 4, that is, to remove or maintain features of the original sample design, so that desired features are attained in the subsample or 'other survey'.

Sampling from existing samples can eliminate costs of sampling frame preparation. For example, subsamples of PSUs and schools for CBWS were drawn from NAEP samples of PSUs and schools. This example is discussed in detail in Section 4. The target populations were the same for each survey, which allowed for the subsampling to occur.

Another reason for using existing samples is that it lowers costs by drawing a large sample for inexpensive data, then drawing an efficient subsample for the high cost data, traditionally known as two-phase sampling. An extension of this approach occurs when data are gathered, analyzed, and published for two or more levels of sampling units. We briefly mention two examples:

1. The Alcohol and Drug Services Study (ADSS) is being conducted for the Substance Abuse and Mental Health Services Administration (SAMHSA). The survey design lowered costs by collecting data inexpensively through phone calls from a large sample of about 2400 facilities in Phase I (Mohadjer, Yansaneh, Krenzke, and Dohrmann (1999)). The Phase I data were

gathered and analyzed, but the Phase I data were also used to select a subsample of about 300 facilities in Phase II in order to collect the high cost data. The high cost data were gathered through an on-site facility interview, and furthermore, after a sample of client discharge episodes was selected, the clients were followed up with an interview to determine their treatment outcomes.

2. Operations of the Federal Work Study (FWS) Institutional and Student Surveys were conducted for the Department of Education. The work study program offers students jobs as a financial aid tool. Similar to ADSS, the survey design lowered cost by collecting data through phone calls to college financial aid offices. Those data were analyzed, but also used to build an efficient subsample of schools, and a sample of students within schools for the student survey. Details of the sample selection process for the FWS can be found in Westat (1999).

In each of the above examples, it was important to know the features that existed in the original survey sample. Unless care is taken, what exists as features in the original survey sample, such as oversampling, may exist in the resulting sample without the sampler knowing. The example of the CBWS will show how features of the original survey sample were removed, and how features were added to the new sample. The goal of this paper is to not lay the groundwork for subsampling as was done in Cochran (1977) and Sarndal, Swensson, and Wretman (1992), but to summarize some technical issues that we have encountered in a recent application of these techniques.

3. Issues to Consider When Subsampling

There are several questions to be addressed before determining how or if a subsample is to be selected. Therefore, before discussing the CBWS example, we first offer some things that we have had to think about when designing subsamples.

1. Analysis. How are the data going to be analyzed? What are the analysis units, measures, and subgroups? What comparisons will be made? What types of analyses will be done (chi-square, analysis of variance, ratio estimation)? The survey sponsor or analyst should provide analysis plans before a sample is designed. How accurate should the analysis measures be? What is the target population for the analysis?

2. Sampling units. What should the sampling units be? An efficient design may call for a multi-stage sample of PSUs, Secondary Sampling Units (SSUs), and USUs. Based on the answers to the analysis questions, one may be able to identify the sampling units. The sampling units may not be the data collection units (e.g., a sampled campus may not be able to provide information, and it may be necessary to collect data from the university headquarters).

3. Auxiliary variables. What analysis variables are available that can be used in the design stage and in estimation in order to reduce sampling error? It is a good idea to list the variables that you know before designing the sample. Any variable that is expected to be moderately correlated with the analysis variables should be included in the design. The auxiliary information also should be used in estimation (e.g., weight adjustments for nonresponse, poststratification). A lack of auxiliary information may lead to a multi-phase design as auxiliary information may need to be collected from a large sample at an early phase or stage.

4. Cost of data collection per unit. Cost is usually the driving force on sample size, perhaps just as much or more so than accuracy. High per unit costs often leads to multi-stages and/or multi-phases of sampling.

5. Response rates. Is the data collection unit hard to locate? Are the survey questions sensitive? What response rate is to be expected? Is the time of the year a factor in obtaining response? Note that response rates will be lower if you select the subsample from respondents since the final response rate is equal to the original survey's rate multiplied by the subsample rate. In addition, if the subsample is taken from respondents, not from the entire original survey's sample, then bias due to nonresponse is inherited from the original survey.

6. Variance estimation. Is the sample design variance computation friendly? Variance estimation computations should be kept in mind when the sample is designed, so that stable variance estimates can be computed. For instance, variance estimation for complex estimation under a complex two-phase sample is an on-going research item (Kott and Stukel, 1997), and it is unclear how to provide an appropriate replication scheme under certain situations.

All the above items have an effect on the design of a survey. A discussion of one of our recent experiences will now be presented in detail to illustrate some of the main issues involved with subsampling.

4. NAEP Classroom-Based Writing Study Example

4.1 Overview of Original Survey's Samples (NAEP)

For this example, the original survey's samples come from the 1998 NAEP. The purpose for NAEP is to assess the educational achievement of students in the 4th, 8th, and 12th grades in the United States. The NAEP sample design was a multistage probability sample. The first stage featured a sample of 94 PSUs, which were counties or groups of counties. There were 22 PSUs that were selected with certainty, and 72 noncertainty PSUs, where one PSU was selected within each noncertainty stratum. The second stage consisted of a Probability Proportionate to Size (PPS) systematic sample of 4th, 8th, and 12th grade schools (about 850 schools in each grade). Important to this example is that private schools were oversampled (i.e., sampled at a rate greater than their proportion observed in the sampling frame), and public schools with larger numbers of black and Hispanic students were also oversampled. After schools were selected, sessions were randomly assigned to the sampled schools. The session types for 1998 were writing, civics, and reading. Within the sampled schools, students were randomly selected and randomly assigned to sessions. About 40,000 students were selected per grade. Also important for this example is that the USU is the student.

4.2 Overview of "the Other Survey's" (CBWS) Subsampling Design

The other survey, the CBWS, had a target population of 4th and 8th grade students in the United States, which is consistent with NAEP. The reason for subsampling from NAEP's PSUs and schools was threefold:

1. To obtain a substantially smaller number of schools (129 4th grade schools and 134 8th grade schools were selected for the CBWS from the initial NAEP samples of 4th and 8th grade, respectively);

2. To reduce field costs by subsampling NAEP PSUs; and

3. To take a subsample of NAEP schools that were assigned a writing session, since the NAEP writing assessment was used in the analysis of CBWS data.

The subsampling procedure for the CBWS included subsampling NAEP PSUs. All 22 NAEP certainty PSUs were selected for the CBWS. The 72 noncertainty PSUs were paired within pseudostrata, and one of the two PSUs from each pair was randomly selected with equal probability.

For grades 4 and 8, within the selected CBWS PSUs, a sample was selected from 1998 NAEP schools that were assigned a writing session. To arrive at 100 responding schools for each grade, a sample of 129 grade 4 schools and 134 grade 8 schools was selected, accounting for anticipated loss due to nonresponse and ineligibility of schools and teacher nonresponse for the teacher survey.

Within the subsample of schools, one language arts classroom was selected from a list of all language arts classrooms for the grade at which the school was sampled. Within the selected classroom, all students were asked to participate in the study, thus the USU was the classroom, not the student as in NAEP.

4.3 Objectives of the CBWS Design

Several objectives were incorporated into the sample design:

- To oversample public schools in high minority areas, in the same way as NAEP samples;

- To remove the rate of oversampling private schools, present in the NAEP design, so that the number of private school students in the sample was proportionate to the number of private school students in the sampling universe;

- To arrive at approximately equal selection probabilities for classrooms within public schools with larger numbers of black and Hispanic students, with all other classrooms having selection probabilities half as great; and

- To deal with the USU for NAEP (student) being different from the USU for the CBWS (classroom). Details of the 1998 NAEP and CBWS sample design and weighting procedures are contained in Krenzke, Rust, and Wallace (1999).

The measure of size assigned to schools for the CBWS sample incorporated factors that attempted to meet the design objectives. The design also incorporated the change in the ultimate sampling unit from the student in NAEP to the classroom for the CBWS.

To meet the objectives discussed above, several factors were constructed for the CBWS subsample of schools in order to cancel out the various factors that comprise the overall selection probability of a NAEP school assigned at least one writing/civics session. To begin, the overall selection probability of CBWS school *i* is presented as:

$$\pi_i = \pi_g \times \pi_{i|g \in G1} \times \pi_h \times \pi_{g|g \in G1} \times \pi_{i|g \in G2, i \in I}; \qquad (1)$$

where

I = NAEP sample of schools;

$G1$ = set of PSUs selected for NAEP;

$G2$ = set of PSUs selected for CBWS as a subset of the NAEP PSUs;

π_g = the NAEP PSU selection probability for the PSU *g* containing the school *i*;

$\pi_{i|g \in G1}$ = the conditional NAEP selection probability for school *i* given the sample of NAEP PSUs;

π_h = the probability that the school was assigned at least one writing session;

$\pi_{g|g \in G1}$ = the conditional selection probability of a CBWS PSU *g* given the sample of NAEP PSUs *G1*; and

$\pi_{i|g \in G2, i \in I}$ = the conditional selection probability of CBWS school *i* for the CBWS given the sample of NAEP schools *I* and the sample of CBWS PSUs *G2*.

For the CBWS school probability, we want equation (1) to reduce to:

$$\pi_i = K_2 \times c_i \times \hat{m}_i;$$

where

K_2 = the inverse of the sampling interval for the subselection of NAEP schools;

c_i = 2, for public school *i* with larger numbers of black and Hispanic students; = 1 otherwise;

$\hat{m}_i$ = a function of the estimated number of classrooms within school *i*; and

$$= \begin{cases} round\left(\dfrac{x_i}{25}\right) & \text{if } x_i > 12 \\ 0.25 & otherwise \end{cases} .$$

The function of the estimated number of classrooms, $\hat{m}_i$, is included in anticipation of selecting one classroom within the school so that we optimize our chance that we arrive with approximately equal selection probabilities across classrooms within oversampling domains. The exception is that small schools, with just a single classroom, have a selection probability of one quarter of other comparable classrooms. In order to do this, we need to know more about how the original survey's sample of schools was drawn in order to make the design as efficient as possible.

4.4 The NAEP School Selection Probabilities

Let the conditional NAEP selection probability for school *i* within the set of NAEP PSUs be defined as:

$$\pi_{i|g \in G1, pss} = K_1 \times \left(1/\pi_g\right) \times f_i \times z_i;$$

where

K_1 = inverse of the sampling interval for the selection of NAEP schools; and

z_i = measure of size;

$$= \begin{cases} 0.25 & \text{if } x_i < 6 \\ x_i/20 & \text{if } 6 <= x_i <= 19 \\ 1 & \text{if } 20 <= x_i <= n_{max} \\ x_i/n_{max} & \text{if } x_i > n_{max} \end{cases} ;$$

where

x_i = estimated grade enrollment for school *i*; and

n_{max} = maximum within school sample size of students.

For public schools with larger numbers of black and Hispanic students, the oversampling factor is $f_i = 2$. For nonpublic schools, $f_i = 3$. By incorporating the formula for $\pi_{i|g \in G1, pss}$ into equation (1), the first two factors reduce to $K_1 \times f_i \times z_i$, therefore equation (1) can be written as:

$$\pi_i = \left[\left(K_1 \times f_i \times z_i\right) \times \pi_h\right] \times \pi_{g|g \in G1} \times \pi_{i|g \in G2, i \in I}. \qquad (2)$$

The terms in the brackets in equation (2) show what we know about the NAEP sample of schools that were assigned a writing session. That is, we know that its selection probability is a function of:

- An oversampling factor, f_i;
- Estimated grade student enrollment; and

- The probability of being assigned a writing session.

4.5 The CBWS School Selection Probabilities

For subsampling schools and sampling classrooms for the CBWS, we needed to realize that the NAEP oversampling features were not quite what we wanted, which was only to oversample public schools with larger numbers of black and Hispanic students. It was also helpful to know that it was a function of the estimated grade enrollment, which was included in NAEP since students were the USU. However, for the CBWS, we were sampling classrooms, so we needed a function of the number of classrooms.

To arrive at approximately equal probabilities of selection among classrooms (within oversampling domains), the last factor in equation (2) must be formulated to do the following:

- Cancel out the first five terms of equation (2); and

- Anticipate the sampling of one language arts classroom for each school from which all students within the classroom participate in the study, so that the student weights do not vary due to a varying number of language arts classrooms across the sampled schools.

Therefore, the conditional selection probability of school i for the CBWS given the sample of NAEP schools i in the sample of CBWS PSUs $G2$ was formulated as:

$$\pi_{i|g\in G2, i\in I} = K_2 \times c_i \times \hat{m}_i \times \left(1/(K_1 \times f_i \times z_i)\right) \times \left(1/\pi_h\right) \times \left(1/\pi_{g|g\in G1}\right).$$

We set up the conditional selection probability of the CBWS schools so that we

- Take out the oversampling factors that existed for NAEP;

- Take out the function of the estimated grade enrollment (z_i);

- Take out the probability of being assigned a writing session; and

- Take out the conditional probability of selecting the CBWS PSUs.

We added back in the oversampling factor, c_i, for public schools with larger numbers of black and Hispanic students. We also added the function of the estimated number of classrooms.

Incorporating the conditional selection probability of the CBWS school i into equation (2), we get:

$$\begin{aligned}\pi_i &= (K_1 \times f_i \times z_i) \times \pi_h \times \pi_{g|g\in G1} \times \pi_{i|g\in G2, i\in I}\\ &= (K_1 \times f_i \times z_i) \times \pi_h \times \pi_{g|g\in G1} \times K_2 \times c \times \hat{m}_i \times\\ &\quad \left(1/(K_1 \times f_i \times z_i)\right) \times \left(1/\pi_h\right) \times \left(1/\pi_{g|g\in G1}\right)\\ &= K_2 \times c_i \times \hat{m}_i\end{aligned} \tag{3}$$

Now we see that any complication of the USU being the student for NAEP, and the classroom for the CBWS has been eliminated since the z_i constructed for sampling NAEP students has been removed and the $\hat{m}_i$ constructed for sampling classrooms has been added.

One classroom is selected at random from each sampled school. Therefore, the conditional probability of selection for classroom j in school i is $1/m_i$; where m_i is the number of language arts classes in the selected grade. Extending equation (3), the overall selection probability of a classroom is:

$$\pi_{ij} = K_2 \times c_i \times \hat{m}_i \times (1/m_i).$$

4.6 Variation in Classroom Probabilities

The overall selection probabilities for classrooms, and consequently the final classroom sampling weights, will vary across classrooms for four reasons:

1. Undersampling of small schools, since $\hat{m}_i / m_i \neq 1$ (or does not cancel out), because $\hat{m}_i = 0.25$ for small schools;

2. Oversampling of public schools with larger numbers of black and Hispanic students;
3. To the extent that the actual number of language arts classrooms differ from the estimated number of language arts classrooms; and

4. Nonresponse adjustments to the weights.

Relating to reason #3, if there is any difference between the estimated number of classrooms, $\hat{m}_i$, and the actual number of classrooms, m_i, it can cause quite a bit of variation in sampling weights, since they are small integers. For instance, if $\hat{m}_i = 1$, and $m_i = 2$, then the classroom will have half the selection probability of others in the same oversampling domain. This example shows the importance of accurate auxiliary information.

Further investigation revealed that much of the variation among the final classroom weights was due to estimating the number of classrooms. The second largest component was due to oversampling. A weighting procedure that adjusts the weights to account for nonresponse is done to reduce the bias due to nonresponse. This adjustment procedure also caused some variation in the resulting sampling weights.

5. Summary

In summary, we provided a short discussion of some of our recent experiences with regards to subsampling in the survey methodology context. Issues that were discussed relate to: meeting analytical objectives, choosing measures of size, cost implications on sample size and design, subsampling from existing 'samples' not just respondents, desiring to arrive at a equal probability design and how resulting weights may vary, and subsampling from an original survey while switching ultimate sampling units between surveys.

Furthermore, we provided a case of sampling from an existing sample, that is, an original survey's sample. We showed how the sampler can benefit from knowing information about the original survey's sample, so that he/she can maintain or remove characteristics that existed in the original survey's sample.

6. References

Cochran, W.G. (1977). Sampling Techniques. John Wiley & Sons.

CSFII (1997). Continuing Survey of Food Intakes by Individuals/Diet and health Knowledge Survey 1994-96, 3-Year Survey Operations Report. Conducted by Westat for Agricultural Research Service of the United States Department of Agriculture.

Krenzke, T., Rust, K., and Wallace, L. (1999), forthcoming. 1998 Main National Assessment of Educational Progress Sampling and Weighting Report. Prepared for the National Center for Education Statistics.

Kott, P. and Stukel, D. (1997). Can the Jackknife Be Used With a Two-Phase Sample? Survey Methodology, Vol. 23, No. 2, pp 81-89.

Mohadjer, L., Yansaneh, I., Krenzke, T., and Dohrmann, S. (1999), forthcoming. Sample Design, Selection and Estimation for Phase I of ADSS Final Report. Prepared for the Substance Abuse and Mental Health Services Administration.

Sarndal, C., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. Springer-Verlag.

Westat (1999), forthcoming. The National Study of the Operation of the Federal Work-Study Program. Volume Two: Technical Appendices for the Institutional Survey. Prepared for the U.S. Department of Education.

Acknowledgments

We would like to thank Fran Cohen for her computer programming efforts related to this paper.