ESTIMATION ISSUES RELATED TO THE STUDENT COMPONENT OF THE SASS

Karen Ellen King, Bureau of the Census and Steve Kaufman, National Center of Education Statistics[1]
Karen Ellen King, DSMD, U.S. Bureau of the Census, Washington, DC 20233

Key Words: Sampling, Probability of Selection, Hypergeometric Distribution

I. INTRODUCTION

The Schools and Staffing Survey (SASS) is a periodic integrated system of surveys of schools, school districts, school administrators, and teachers. For the 1993-94 SASS, a student component was added.

SASS is sponsored by the National Center for Education Statistics (NCES) of the U.S. Department of Education. Users of the survey data are educators, researchers, policy makers, and others interested in educational issues.

The survey data is collected by mail, with telephone followup of nonrespondents. All levels of the SASS are interrelated. Selection of sample schools, both public and private, is the starting point. For each sample school, a sample of its teachers is selected and data is also collected from its principal. The school district of each selected public school is also in the sample. For the current SASS, a sample of students was selected from sample teachers; continuing the relationship of one component with the other components of the survey.

The NCES planned to add a student component to the SASS for several years. The goals of this component is to examine the quality of teachers through their students and analyze student characteristics. This is accomplished by selecting a few sample students from a class taught by each sample teacher.

A student component in SASS was tested initially as part of a 1991 SASS Research Study. In this study, student sampling and the collection of administrative data on selected student was attempted for the first time. Several problems were encountered during the sampling and the collection phases which discouraged any attempt at estimation.

A second feasibility study was conducted during the 1992-93 school year to solve the operational problems encountered in the first study. It is also where we began to deal with the issue of estimation, in particular, to develop an estimator for the student's probability of selection using only the amount of information that an already over burdened school could easily provide.

This paper gives an overview of the second feasibility study and a summary of the components that make up our estimator of the probability of selection of students.

II. OVERVIEW OF SAMPLING

A. School Selection

As with all SASS surveys, the selection of samples of public and private schools was the starting point for the feasibility study. Three hundred public and 200 private schools were selected and mailed forms for listing teachers. A teacher listing form asks schools to provide the names and some demographic information for every eligible teacher at that school. Eligible teachers consist of regular full-time and part-time teachers whose main assignment was teaching in kindergarten or any of grades 1 to 12 during the school year.

Completed listing forms were returned to the Census Processing Center in Jeffersonville Indiana. Two hundred thirteen public and 133 private schools returned completed teacher listing forms.

Interviewers specially trained for this operation did the teacher selection, class period selection, and the student selection through a series of telephone conversations with participating schools.

B. Teacher Selection

Three teachers (if available) were systematically selected from each of the returned teacher listing form.

Each school was called to confirmed that each sampled teacher was eligible, i.e., did they teach at least one regularly scheduled class of K-12 grade students in a week. Once the ineligible teachers were screened out, the call continued by asking questions to classify the eligible teachers as either self-contained or

---

[1] This paper reports the general results of research undertaken by Census Bureau staff in collaboration with staff from the National Center for Education Statistics. The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau or National Center for Education Statistics.

departmental. Sampling instructions for class period selection varied by this teacher classification.

* Self-contained is defined as teaches several different subjects to the same group of students all day.

* Departmental is defined as teaches only a limited number of subjects to more than one group of students per day.

C. Class Period Selection

For departmental teachers, a double sampling procedure was used to select the sample class period. We started by asking the school how many periods they had per week, and then, using this value, selected a set of five class periods as the initial sample. If all the teachers were departmental at the school then all three teachers had the same set of class periods.

For example, suppose the school told us that there were 25 class periods in a week (not counting homeroom). For this number of periods per week, the selected set of class periods were the fifth on Monday, the fourth on Tuesday, the third on Wednesday, the second on Thursday, and the first on Friday.

Then the interviewer probed the school about each class period in the initial set of five to determine if the teacher actually taught a class of eligible students. Eligible students are those in kindergarten through the twelfth grade, that are receiving instruction and are not in study hall, recess, lunch or homeroom. If a teacher did not teach a class in one of the class periods, the period was considered ineligible to go to the next step of sampling. Once the eligibility of each class period was determined, one out of the remaining set of eligible class periods was randomly selected.

For example, suppose teacher Jane Doe taught four out of five class periods given in the above example. (She supervises study hall the third period on Wednesdays.) To select the class period we ordered the four remaining periods by days in the week (Monday through Friday) and picked one.

The third class period in our ordered set was selected. Thus, we wanted three student names from the second period on Thursday.

For the self-contained teachers, no class period sampling procedure was needed since they only taught one class of students.

Schools were asked to get selected class period rosters. Generally the first call was terminated so that the school could look up the roster. Another time was set for a call back to do the next phase of selection.

There are two reasons to justify this elaborate scheme to select a class period. The first is the double sampling guaranteed that we selected a class period where the teacher was actually teaching. During the initial study, we selected one class period randomly in the school week for each departmental teacher. Many times the school simply said that the teacher was not teaching during the selected period. Subsequently, no students were selected for these teachers and the student sample size was much smaller than expected. The second reason was to reduce the chances of bias being introduced into the student sample. If we pick only one class period, there is the possibility that a subset of the student body would be in ineligible classes (study hall, homeroom, lunch, or recess) and have no chance of selection. When we increase the number of class periods selected to five, the chances of a student being in an ineligible class for all five class periods becomes small.

D. Student Selection

When the class period roster was available, over the phone we gave the school instructions to select three sample students from the roster. A random number table was used to indicate the line numbers of the students selected.

For example: Suppose Jane Doe's second period class on Thursdays had 26 students. Using a table, interviewers would have asked for the 3rd, the 14th, and 24th name from the top of the roster.

Student names or some other unique student identifier was requested so that we could uniquely label each student's forthcoming questionnaire. Eleven schools refused to provide student names for our survey fearing parental displeasure.

Two months after telephone sampling, student questionnaires were mailed to the schools of over 1600 public students and over 1000 private students.

III. ESTIMATOR DEVELOPMENT

If we selected our sample of students from a list of students enrolled in a school, the probability of selection within the school would be straight forward since a student would only be listed once, i.e., (1/enrollment). However, the main goal is to provide data on sample students that are taught by sample teachers in an eligible class in sample schools. This involves several level of sampling to obtain our sample student.

Due to the many levels of sampling, the probability of selection of each student for a sample teacher within a sample school is actually made up of several component probabilities and some random variables.

$N \cdot P(\text{class period}) \cdot P(\text{student within class}) \cdot P(\text{teacher}) \cdot C \cdot P(\text{school})$

Several of the components are straight forward and easy to define. However, several components (those dealing with sampling within the school) turned out to be quite a challenge. The first subsection defines the easier components of the estimator and the following three subsections show the more challenging components.

A. Probability of Selecting the Teacher and the Student Within the Class Period

The probability of selecting the teacher within the school is three out of the total head count of teachers (H) or

$$P(teacher) = \left(\frac{3}{H}\right)$$

The probability of selecting the student from the selected class period (l) of teacher (j) is three out of the class size S_{lj} or

$$P(student\ within\ class) = \left(\frac{3}{S_{lj}}\right)$$

B. Multiplicity of Teachers and Class Periods (N)

The student universe within school is a combination of every list of every class period roster of every eligible class period taught by each eligible teacher in the school during a school week. In schools containing mostly self-contained teachers, such as lower elementary schools, each student's name only appears on one teacher's class period roster. However, in schools containing mostly departmental teachers, such as high schools, each student's name can appear on many class period rosters.

The word multiplicity has come to represent the total number of ways a student can end up in the student component considering all teachers that teach the student and all class periods each teacher has the student. This is equivalent to the number of time the student's name appears on the list if we combined every class roster.

Suppose Student A has four subjects with four teachers and each subject is taught once a day or five times a week. Let us assume that the second period on Thursday was the period used to select the student.

To get the true probability of selection, we would have to obtain all this information to count all the possible ways this student could have been selected.

In the first study, we tried to get an idea of the multiplicity using the following question:

"How many class periods does the student have each week that are taught by .....only 1 teacher? two or more teachers?"

This question did not work well and went unanswered by many of the school administrators. Of course, for our example, the correct answers are twenty for only 1 teacher and zero for two or more teachers.

This particular example of all possible ways of getting Student A is very simple. When we add more teachers, more periods per day, classes that don't meet everyday, and some sort of period rotation, it gets very confusing.

When planning the second study, we debated whether to ask for all the information about a sample student's school week or reduce respondent burden by collecting for each sample student only information about the three sample teachers. It was decided to reduce respondent burden, ask for less information, and concentrate on the sample teachers only. The multiplicity question was reorganized and reworded to ask specifically for the association of the student to each of the sample teachers in the school. Basically, it was broken down into three smaller questions.

1. Does this teacher have this student?
2. Is the student with the teacher all day?
3. If not all day, what subjects does the student have with the teacher and how often does the class meet?

The same set of questions is repeated for each sample student and each sample teacher in the sample school.

A term adopted for use during this study was the "certainty" teacher. The certainty teacher is defined to be the teacher we initially went through to get the sample student. At the very least, we expected to see information for the certainty teacher filled out in the multiplicity question. Any information appearing under the other two teacher names was an added bonus.

You might wonder why we are interested in the other two teachers. We had to determine if the student had a chance of being selected through the other two teacher. If the student has more than one sample teacher then the student's probability of selection is the sum of the probability of selection through the each sample teacher (j).

$$\sum_{j=1}^{3}\left[N_j \cdot P(\text{class period}_{lj}) \cdot P(\text{student within class}_{lj}) \cdot P(\text{teacher}_{lj})\right] \cdot C \cdot P(\text{school})$$

Most of the time in the feasibility study, the probability of selection through the other two teachers

was zero because they didn't have the student. Occasionally, a student did have more than one sample teacher and twice, the same student was selected for sample through two different sample teachers.

Let us look at the multiplicity for student A again. Suppose by chance, two of this student's teachers were selected for sample. The new question would have given us the following information. Ms. Jane Doe teaches this student English and the class meets five times per week. Mr. John Smith teaches this student Social Studies and the class meets five times per week. Jane Doe became the certainty teacher when we selected student A in her Thursday second period class and as expected, we picked up all five second period classes. The information about John Smith teaching of student A in the five class periods was a welcomed surprise. So the multiplicity or total number of ways student A could be selected through Ms. Doe is five and for Mr. Smith is also five. We also know that we probability of selection for student A will be the sum of the probability of selection through each sample teacher.

Using the multiplicity information as seen in the example, we could estimate a student probability of selection conditioned on selecting the three sample teachers in the school.

C. Probability of Selecting the Sample Class Period

Another component that we had to estimate was the probability of selecting the class period. For self-contained teachers, this probability is one since their one class is in with certainty. For departmental teachers, the double sampling procedure for selecting class period (described in section II) guaranteed an eligible class, but it added some complication to calculating this component. Recall that the procedure involved selecting a set of five class periods for the departmental teachers in a school. For each sample teacher, we determined which class periods contain an eligible class and select one of the eligible classes.

To do this, we had to calculate the probability of selecting at least one eligible class from a set of five class periods and then selecting one of them. From the start we knew that we had to consider all possible combinations of five class periods where T define the total number of class periods in the school week. Initially we came up with:

$$\textit{Initial } P(\textit{class period}) = \binom{T}{5} * \left(\frac{1}{5}\right)$$

Unfortunately, the resulting weights were large implying that the probability was too small. After several more dead ends, it occurred to us that we needed to consider the eligibility of the class period as a success in a series of trails, i.e., the probability of having at least one eligible class out of a possible set of five was a hypergeometric random variable. Actually it is a sum of hypergeometrics since we have to estimate the probability of all possible combinations of sets of five class periods that contained at least one eligible class.

Again, let T be the total periods in the school week. Let L_j define the total number of class periods that teacher (j) taught an eligible class in the school week. Finally let l be the number of eligible periods in the set of five.

The probability of selecting at least one eligible class and choosing one is:

$$\textit{Final } P(\textit{class period}) = \sum_{l=1}^{\min(5,L_j)} \frac{\binom{L_j}{l}\binom{T-L_j}{5-l}}{\binom{T}{5}} * \frac{1}{l}$$

In words this is saying the probability of selection of the class period is equal to the sum of

(the probability of getting one eligible class out of five)

PLUS (the probability of getting two eligible classes out of five and selecting one)

PLUS (the probability of selecting three eligible classes out of five and selecting one)

PLUS (the probability of selecting four eligible classes out of five and selecting one)

PLUS (the probability of selecting five eligible classes out of five and selecting one).

D. Multiplicity of Students (C)

How often can a student's name appear in the set of distinct students taught by a set of three sample teachers over all possible sets of three sample teachers? It depends on how many distinct teachers the student has during the week. This was a second multiplicity problem that we encountered and our final obstacle in a pursuit of an estimator. We didn't have any way calculating this because we didn't ask for the number of teachers the student had in the school week during student sampling. Again, due to the decision to lighten the respondent burden on school administrators, we would have to approximate this component. We felt we could estimate it as an

average across all students by using the following adjustment:

$$C = \left(\frac{X_s}{S_s}\right)$$

where S_s is the number of students in scope for the survey in the school and X_s is one over the sum of all student probability of selection within the school

$$X_s = \frac{1}{\sum_{h=1}^{3}\left[\sum_{j=1}^{3} N_{ij} * \left(\sum_{l=1}^{\min(5,L_j)} \frac{\binom{L_j}{l}\binom{T-L_j}{5-l}}{\binom{T}{5}} * \frac{1}{l}\right) * \frac{3}{S_{ij}} * \frac{3}{H}\right]}$$

One benefit of this ratio adjustment was the joint probability of selection of the three sample teachers cancels out and does not appear in the final weight.

IV. SUMMARY OF RESULTS

We have an approximation of probability of selection for each student which provides an unconditional estimator of student basic weight. This estimator depends heavily on collected data which is open to item nonresponse or response error. The basic weight for sample student i is given by:

$$BW_i = \left[\frac{1}{\sum_{j=1}^{3} N_{ij}\left(\sum_{l=1}^{\min(5,L_j)} \frac{\binom{L_j}{l}\binom{T-L_j}{5-l}}{\binom{T}{5}} * \frac{1}{l}\right)\frac{3}{S_{ij}} * \frac{3}{H}}\right] * \frac{S_s}{X_s} * B$$

Where j is a teacher.

L_j is the total number of class periods taught by teacher j.

l is a class period.

i is the student.

N_{ij} is the number of class periods student i has with teacher j.

T is the total number of class periods in the school.

S_{lj} is the number of students in teacher j's selected class period l.

S_s is the school enrollment.

X_s is one over the sum of student probabilities within school before adjustment.

H is head count of eligible teachers at the school.

V. FUTURE PLANS

Sampling and data collection has been completed for the 1993-94 SASS student component. We used the sampling methodology developed in the research studies to implement the student sampling successfully. The weighing methodology includes the estimator given earlier to generate the basic weights with one additional component as of the publishing of this paper. The component probability of $\left(\frac{l}{L_j}\right)$ has been added to the probability of selecting a class period. This probability covers the chances of selecting the particular set of eligible periods in the initial set of five sample class periods.

Tinkering with the estimator will probably continue until the weighting is run. After the estimation checks currently planned have been completed, more research may be desired.