ESTIMATION OF SAMPLING PROBABILITIES FOR SAMPLE AGGREGATION SITES

Charles D. Cowan, National Center for Education Statistics
Sara M. Debanne, Case Western Reserve University School of Medicine
Charles D. Cowan, NCES, 555 New Jersey Ave., Rm400, Washington DC 20208

KEYWORDS: Cluster Sampling, Education Statistics

1. Introduction

There are many situations in statistics where sampling is conducted at multiple levels, and estimates are desired at each level. In education, an example would be the selection of schools and then students to give a picture of education in the country. Estimates might be made at both the school level and the student level in separate analyses, or in one analysis of school effects on students. In biostatistics, similar situations might be found with hatcheries, fish, and rivers, or with patients at hospitals who after a time transfer to other care facilities.

But mobile populations in a longitudinal study may move. The sites where the sample is found may change in second and later contacts with the sample, but still be of interest for the purposes of estimation. An example would be found in the National Educational Longitudinal Survey (NELS:88), where students were originally sampled in the eighth grade in 1988 for interviews and will be reinterviewed in the tenth and twelth grades. The problem is not knowing the probabilities associated with the students in tenth grade, because those probabilities carry over from the eighth grade. The problem is determining the probability that a tenth grade school will fall into sample because an eighth grade student has gone to that school two years later. This paper deals with determining probabilities of observation (i.e. selection) for these sites.

2. The Basic but Simplified Approach

For this problem, a model based approach was taken. Consider the process of a student going from an eighth grade to a tenth grade as a transition, similar to the type seen in transportation problems or in applications of Markov chains. There are two levels to the school sampling problem: modeling population movements for probabilities and accounting for specific sample member movements. Ultimately, the probability of a tenth grade school falling into the followup sample for this study is a combination of the model based probabilities for movements of students, and the design based probabilities for the students who were sampled in the eighth grade. The two sets of probabilities will be dealt with separately in succeeding sections.

2.1 Model Based Probabilities for the Population of 8th-10th Grade Students - A National Level Aggregated Model

In elementary-secondary education a natural hierarchical structure exists. In public education, States control education within their boundaries through school districts, and within states schools report to the school districts. School districts in this structural arrangement are referred to as Local Education Agencies (LEA's). Private schools do not have this type of structure, but geographically can be placed within State/LEA boundaries, and so for the purposes of exposition will be treated as within LEA's.

Consider then a transit by a student from an eighth grade school to a tenth grade school, either public or private, or out of the school system completely. This latter case will be terned an "out" status which, because its transitions are very infrequent among eighth to tenth grade transitions, will not be considered in the initial part of the development, but will be covered in the next section.

Ignoring the "outs", we consider a very large matrix, Z, whose elements denote numbers of students transfering from schools corresponding to row indices (i.e., eighth grade schools i, j, k, etc. within LEA's within States). to schools corresponding to column indices (i.e., tenth grade schools r, s, t, etc. within LEA's within States. Eighth and tenth grade schools are assumed to be different schools which is reasonable since rarely will the elementary and middle school be the same structures as the high school, where they will be nominally treated as different. For LEA's, sometimes the LEA's are the same controlling entities and sometimes not, with elementary LEA's feeding specific secondary LEA's. The States are always the 50 States and the District of Columbia (hereafter referred to as the 51 states).

A simple transformation could be used to transform the elements of Z which are counts to the corresponding probabilities of transitions between schools. Note that the counts and probabilites

discussed here are at the student level, not the school level.

Unfortunately, these counts cannot be observed directly (they can be, but only through a massive data collection that would cost upwards of a million dollars). Some counts are available, however, and those are the marginals of the table. These are available from the Common Core of Data (CCD), collected by the National Center for Education Statistics, as the enrollment or membership figures for each eighth or tenth grade school. Notationally, we may think of the matrix $\underline{Z}$, with entries:

Z_{ij} being the counts of students in eighth grade school i who transferred to tenth grade school j

We know the marginal vectors with entries Z_{i+} and Z_{+j}, where

Z_{i+} = the number of eighth grade students in school i in 1988

Z_{+j} = the number of tenth grade students in school j in 1990

from the Common Core of Data.

To determine the probabilities of student transitions from eighth grade to tenth grade, we will consider a hierarchy of types of transitions, modeled after data from the survey in 1988 and subsequent information collected in 1990 (after the students have transferred), from geographically distant transitions to local ones:

1) the student transfers from one State to another,
2) the student stays in the same State, but
   a) transfers from one LEA to another, or
   b) dies or drops out (to be estimated from third and fourth sources)
3) the student stays in the same State and LEA.

It is necessary to use a slightly more complex notation for $\underline{Z}$ if the intuitive flavor of the hierarchy of transitions is to be utilized. Reindexing subscripts we redefine elements of $\underline{Z}$ as:

Z[ijk,i'j'm] = the count of students who were in eighth grade in eighth grade State i, LEA j, and school k who transited to tenth grade State i', LEA j', and school m.

Some characteristics of this notation:
if i = i', the students stayed in the same State,
if i ≠ i', then j = j' has no meaning for this problem,
if i = i' and j = j', the students stayed in the same State and LEA, the school subscripts k and m are different instead of accented to emphasize that the schools are always different.

Finally, some practical insights:

(A1) transfers between States are relatively rare, and so the data will be too sparse to consider modeling even finer transitions, like specific LEA to LEA across States;
(A2) transfers between States can be modeled, but smaller States may have to be grouped to do so;
(A3) transfers between LEA's within States can be modeled separately within each State, but again smaller States may have to be grouped;
(A4) there is nothing random or good natured about transfers within LEA's, but for this model the only nformation required is for LEA's in the sample. Additional data collection is required from schools/LEA's to get specific information for this part of the model.

For every cell of the matrix $\underline{Z}$ we can calculate a transition probability. For situations involving infrequent transitions, we estimate the following probabilities:

(P1) for i ≠i',
P[ijk,i'j'm] =
P[i++,i'++]*P[+++,i'j'+]*P[+++,i'j'm]
(P1) assumes equal chances of transferring from state i to state i', regardless of LEA's or schools

(P2) for i = i', j ≠j',
P[ijk,i'j'm] =
P[i++,i++]*P[ij+,ij'+]*P[+++,ij'm]
(P2) also assumes equal chances of transfers between states, but also equal chances of transfers within states between LEA's

And we would directly estimate:
(P3) for i = i', j = j', **P[ijk,i'j'm]**
(P3) is the probability for state i, LEA j of a student transferring from eighth grade school k to a tenth grade school m

In each of the products, (P1) & (P2), there are three components: the first is the probability of a student remaining in or transferring between states, the second is the probability within a state of a student transferring to a specific LEA, and the third is the probability within a state/LEA of transferring to a specific school. (P3) has a single component. All of the components above can be estimated directly from the survey or from CCD. Other sources of information can be used for better esti-

mates of some of the pieces. For example, the transit probabilities from state i to state i' might be better estimated by combining several months of data from the Current Population Survey conducted by the U.S. Bureau of the Census. But P[i++,i'++] also can be estimated directly from NELS as the ratio of the number of students in state i who transferred to state i' over the number of students in state i.

P[+++,i'j'+] is estimated from CCD as the ratio of the number of tenth grade students in State i' in LEA j' over the number of tenth grade students in State i'. Admittedly this may not be a good substitute for knowledge of transfers between States, but it is a reasonable approximation with no specific information about what happens between States. The assumption is that students transferring in from other States transfer in proportion to the size of the LEA's actual membership.

Within LEA's we make the same assumption for P[(+++,i'j'm], estimating it as the ratio of the number of tenth grade students in school m within State i' and LEA j' over the total number of tenth grade students in State i' and LEA j'.

Probabilities for transfer between States and for transfer between LEA's within States (P[i++,i'++] and P[ij+,ij'+]) are calculated directly from NELS as ratios. P[i++,i'++] is estimated as the ratio of the number of eighth grade students in State i who transfer to tenth grade in State i' over the number of eighth grade students in State i. P[ij+,ij'+] is estimated as the ratio of the number of eighth grade students in State i and LEA j who transfer to a tenth grade school in State i, LEA j' over the number of eighth grade students in State i.

For the most frequent case, transfers within the same state and same LEA, we must collect additional information from the tenth grade schools to determine the distribution of the number of students from each feeder eighth grade school in the same LEA, and this distribution is used to calculate the probabilities of transit within the LEA. This would include transit in or to private schools (located within the LEA's).

For dropouts and other students who do not transit to tenth grade, something different is required. To deal with this problem easily, we create a new pseudo-LEA for each state, which represents the set of eighth grade students who do not transit to tenth grade. We use the same procedures to estimate the probabilities of this occurring in each state.

Using these procedures, we can estimate a transition probability for any eighth grade student to any tenth grade. These probabilites are the estimates of the population probabilities for transition and will be used in the next section to determine the probabilities of having a tenth grade school fall into sample in the tenth grade administration of NELS.

2.2 Design Based Probabilities for the Selection of 8th Grade Schools and Students

Eighth grade students in NELS:88 were selected by sampling eighth grade schools initially, then a set number of students within each school. The number of students selected and interviewed in each school was determined in advance, as were the probabilities of selection. Standard methods of selection, such as stratification and oversampling for some subgroups were employed. The probabilities for the eighth grade schools and students should be taken as given, since they have already been applied to the population to obtain a sample.

With stratification within schools, students can actually be selected with different probabilities within the school, but for the purposes of this section we will act as though all students were selected with equal probability. We will return to this when we explore a more deliberate method of calculating the population probabilities of transition.

Design based probabilities of selection for eighth grade schools and students can be represented as:

(S1) **S[ijk]** is the probability of selection for the kth eighth grade school in the jth LEA in the ith State.

(n1) **n[ijk]** is the number of students selected in the kth eighth grade school in the jth LEA in the ith State.

(N1) **N[ijk]** is the total number of eligible students in the kth eighth grade school in the jth LEA in the ith State.

Probabilities for G8 schools with large clusters and oversampling within after the school has been selected with probability S[ijk] should reflect this.

2.3 Combining Design Based Probabilities for the Selection of 8th Grade Schools and Students with Model Based Probabilities for the Transition of Eighth Grade Schools

For the tenth grade school represented by the column indexed by (i',j',m), we now want to know the probability that the [i'j'm]th school will be observed

(come into sample) because a student transferred there from eighth grade. But this could happen if any student or any number of students transferred from an eighth grade school or any number of eighth grade schools. In fact, there are a very large number of ways this could happen since each eighth grade school could contribute 0, 1, 2, ..., up to n[ijk] students to the tenth grade school, and all possibilities for all schools would need to be considered.

However, there is only one way that a tenth grade school would not fall into the sample (be observed), and that is if no sample students came from any of the eighth grade schools. This relationship can be summarized as (the abbreviation BY refers to Base Year, meaning that the students come from the eighth grade base year sample, and G8 and G10 refer to grade eight and grade ten respectively):

Prob(G10 school [i'j'm] has at least one BY Student)

= 1 - Prob(G10 school [i'j'm] has no BY Students)

= 1 - Prob(G10 school [i'j'm] has no BY Students from any G8 school)

$$= 1 - \prod_{ijk} \textbf{Prob(G10 [i'j'm] has no BY Students from G8 [ijk])}$$

To describe when a G8 school did not fall into the BY sample,

$$Q[ijk,i'j'm] = \text{Prob(G10 [i'j'm] has no BY Students from G8 [ijk])}.$$

The probability that G8 school ijk contributes no students to G10 school i'j'm is the combination of the probabilities that the G8 school was selected and that no students from G8 school ijk go to G10 school i'j'm. This can be expressed as:

$$Q[ijk,i'j'm] = (1 - S[ijk]) + S[ijk]*R[ijk]*B[ijk],$$

where B[ijk] is the binomial probability

$$B[ijk] = C_{n[ijk]\ 0}\ (P[ijk,i'j'm])^0 \times (1 - P[ijk,i'j'm])^{n[ijk]}$$

$$= (1 - P[ijk,i'j'm])^{n[ijk]}$$

and $C_{n[ijk]\ 0}$ is the combinatoric n[ijk] items taken zero at a time

$$Q[ijk,i'j'm] = (1 - S[ijk]) + S[ijk]*(1 - P[ijk,i'j'm])^{n[ijk]},$$

So the probability of G10 school [i'j'm] falling into sample is:

$$I[i'j'm] = \textbf{Prob(G10 school [i'j'm] has at least one BY Student)}$$

$$= 1 - \prod_{ijk} \{ (1 - S[ijk]) + S[ijk]*R[ijk]*(1 - P[ijk,i'j'm])^{n[ijk]} \}$$

For schools not actually drawn into the Base Year sample, and so for which we do not have an exact measure of n[ijk] since no sample was drawn, the expected value of n[ijk] would be used since in every school there is some possibility of a contribution of students.

Note that it is not true that:

$$\sum_{i'j'm} I[i'j'm] = 1.0.$$

since the only constraint on the probabilities is that they sum to one row-wise, and only as representing a zero for a column and the whole of the sample distributed in all other columns, for each column. There is no reason or expectation that the probabilities would sum down a column, and in fact one would expect the sum in many cases to be greater than one. These sums are meant to be a measure of the relative representations in the population.

3. The Expanded Model - Accounting for Differences Between Students by Characteristics

The basic model is useful for describing the process of estimating the probabilities of transition, but not very realistic. An expanded version of the model would find probabilities of transition by demographic subgroup.

By adding demographic subclassification to the model structure so that probabilities of transition may be estimated using combinatorial techniques. Indeed, the underlying assumption of combinatorics is that there are homogeneous subgroups. Notwithstanding the inherent individuality of specific students, it is useful for modeling purposes to consider demographic subgroups among eighth graders since within them the a priori likelihood that any two students will advance academically, transfer among specific LEA's, or move to individual tenth grade schools within LEA's is clearly the same.

In this section the notation involving counts of student subgroups and transition probabilities is expanded to depict tion of student demographic types. I, a vector of indices of those demographic

characteristics which are significantly related to housing patterns, educational preferences and academic success (viz., race, sex, socio-economic status, etc.) is specified along the lines of thought advanced in [1].

As before, let the notation be such that i refers to state, j to LEA, k to eighth grade school and m to tenth grade school. Thus, define

$Z_I[ijk,i'j'm]$ = number of students of demographic type I who were ijk as eighth graders and i'j'm as tenth graders.

Taking the joint specifications ijk and i'j'm to denote, respectively, row and column designations, it is possible to consider the matrix **Z(I)** of transitions of students of demographic type I

$$Z(I) = \left[\!\left[Z_I[ijk,i'j'm] \right]\!\right]$$

In order for the matrices to serve as a complete description of transitions of all type I students in either the eighth grade population, the tenth grade population, or both, it is necessary that the "out" status (defined above) be further specified. We must add in addition to academic ijk's rows with indices representing those who drop in (immigrate from home study, from private school, from illness), those who fail once (ninth graders in base year), those who fail twice (tenth graders in base year), those who skip once (seventh graders in base year), and those who skip twice (sixth graders in base year). Similarly, columns must include, in addition to academic i'j'm's, indices representing those who drop out (of school, to private school, to home, to illness), those who fail once those who skip once (11th graders in follow-up year). In this way, the row and column totals will each sum to the grand total, $Z_I[+++,+++]$.

Next, define transition probabilities

$P_I[ijk,i'j'm]$ = P{transition of type I student to i'j'm ¦ ijk at eighth grade}

A "+" in the place of a subscript may be read as "any", to give the same sense as the mathematical totals, x_{i+}, and x_{+j}.

The following equalities are clear from their definitions.

$$P_I[i++,i'++] = \frac{\sum_{jkj'm} Z_I[ijk,i'j'm]}{\sum_{jkuj'm} Z_I[ijk,uj'm]}$$

$$= \frac{\sum_{jkj'm} P_I[ijk,i'j'm]\, Z_I[ijk,+++]}{\sum_{jkuj'm} P_I[ijk,uj'm]\, Z_I[ijk,+++]}$$

Summing numerators and denominators over I gives a similar formula for $P_+[i++,i'++]$.

The probability that a student of type I goes from LEA j to LEA j' is expressed by:

$$P_I[ij+,i'j'+] = \frac{\sum_{km} Z_I[ijk,i'j'm]}{\sum_{kuvm} Z_I[ijk,uvm]}$$

$$= \frac{\sum_{km} P_I[ijk,i'j'm]\, Z_I[ijk,+++]}{\sum_{kuvm} P_I[ijk,uvm]\, Z_I[ijk,+++]}$$

Summing numerators and denominators over I gives a similar formula for $P_+[ij+,i'j'+]$.

Also, the probability that a student of type I goes from school k to school m is:

$$P_I[ijk,i'j'm] = \frac{Z_I[ijk,i'j'm]}{\sum_{uvm} Z_I[ijk,uvm]}$$

$$= \frac{P_I[ijk,i'j'm]\, Z_I[ijk,+++]}{\sum_{uvm} P_I[ijk,uvm]\, Z_I[ijk,+++]}$$

Summing numerators and denominators over I gives a similar formula for $P_+[ij+,i'j'+]$.

The probability that a student of type I chosen at random is from school k is obtained from the expression:

$$Z_I[ijk,+++] \;/\; \sum_{uvm} Z_I[uvm,+++]$$

Also, the probability that a student chosen at random is of type I and school k is:

$$\sum_I Z_I[ijk,+++] \;/\; \sum_I \sum_{uvm} Z_I[uvm,+++]$$

Suppose that the $Y_I[ijk,i'j'm] = Y$ (subscripts suppressed where the reference is clear) denote realizations of the random variables $Z_I[ijk,i'j'm]$. If, for given (ijk), $Z_I(k)$ denotes the size

of sample at school k of type I(ijk) among $Z_{I}[ijk,+++]$ (i.e., total students of type I from sample in school k), then the probability that there are $Y = Y_{I}[ijk,i'j'm]$ at school m from k of type I(ijk,i'j'm) is:

$$\left[\begin{matrix} Z_{I}(k) \\ Y \end{matrix}\right]\left[\begin{matrix} Z_{I}[ijk,+++] - Z_{I}(k) \\ P_{I}[ijk,i'j'm] \times Z_{I}[ijk,+++] - Y \end{matrix}\right]$$

$$\left[\begin{matrix} Z_{I}[ijk,+++] \\ P_{I}[ijk,i'j'm] \times Z_{I}[ijk,+++] \end{matrix}\right]$$

where Y is, then, the number of students among the $Z_{I}(k)$ in the sample of type I, who went from school k to school m; the second factor in the numerator is the combinatorial "those in school k who are not in the sample" choose "those not in the sample who go to school m" and the denominator is the combinatorial "total number of students in school k of type I" choose "those who go to school m".

Finally, setting Y=0 in the last equation, one obtains the probability that school m is in the sample as:

I[i'j'm] = P{school m has at least one student from base year sample} =

1 - P{school m has none of the students from base year sample} =

$1 - \pi_{I}$ P{school m has none of the type I students from base year} =

$$1 - \pi_{\text{all possible}} \Sigma_{I} Z_{I}(k) =$$

$$n[ijk] = \frac{\left[\begin{matrix} Z_{I}[ijk,+++] - Z_{I}(k) \\ P_{I}[ijk,i'j'm] \times Z_{I}[ijk,+++] - Y \end{matrix}\right]}{\left[\begin{matrix} Z_{I}[ijk,+++] \\ P_{I}[ijk,i'j'm] \times Z_{I}[ijk,+++] \end{matrix}\right]^{2}}$$

which simplifies to

$$1 - \pi_{\text{all possible}} \Sigma_{I} Z_{I}(k) = n[ijk]\, F_{I} / G_{I}$$

where

$F_{I} = \{Z_{I}[ijk,+++]\}\{Z_{I}[ijk,+++]-1\} \ldots \{Z_{I}[ijk,+++]-Z_{I}(k)+1\}$

and

$G_{I} = \{Z_{I}[ijk,+++]x(1-P_{I}[ijk,i'j'm])\}\{Z_{I}[ijk,+++]x(1-P_{I}[ijk,i'j'm])-1\} \ldots \{Z_{I}[ijk,+++]x(1-P_{I}[ijk,i'j'm])-Z_{I}(k)+1\}$

4. Conclusions

A number of assumptions are made in the solution presented above. These include the assumptions about the availability of data and the applicability of the model. A test of these assumptions has been proposed and will probably be undertaken in Fall of 1989.

1 The authors want to thank Bruce Spencer at NORC for assistance in developing this final connection of students to the probabilities for the G10 schools.
2 The authors wish to thank Doug Rowland at Case Western Reserve University for suggesting the combinatorial solution to the expanded model.