Area Sampling for Post-secondary Institutions

Ronaldo Iachan, ORC Macro
Pedro Saavedra, ORC Macro
Stephen Kauffman, National Center for Education Statistics

KEY WORDS: Area sampling, post-secondary institutions, unduplication, matching.

This article describes the design of an area sample for the universe of non-Title-IV post-secondary institutions in the Integrated Post-Secondary Education Data System (IPEDS). The main objective of the study is to update an incomplete list of such non-Title-IV institutions, and to estimate the coverage of this list. In addition, NCES has funded the collection of a minimum data set (MDS) for participating institutions.

Currently, the IPEDS system includes a list of presumed non-Title IV institutions with approximately 2,800 listings. Because NCES has made little effort to update this list, however, the true population value could be as high as 20,000 (by some NCES estimates). Obtaining an accurate numerical estimate of the number of non-Title IV post-secondary institutions is a primary purpose of the IPEDS MDS data collection.

The primary objectives of the IPEDS MDS Project are to

1) provide a measure of coverage for the current IPEDS non-Title IV list
2) create a new list of IPEDS non-Title IV institutions based on an area sample
3) estimate total student enrollment in eligible institutions
4) estimate full-time staff in eligible institutions

This article describes the area sampling design and discusses some problems in the design and estimation process, at the time when the first-stage sample of counties has been selected, and the frame of institutions constructed within each area.

Two-stage Sampling Methods

A two-stage area sample was designed so that every candidate non-Title IV institution in the United States has the same probability of being selected. At the first stage of sampling, geographic areas (counties) were selected. At the second stage, candidate institutions will be sub-sampled from a comprehensive list of institutions at the local county level.

The design distinguishes three primary strata based on county population size: (1) certainty counties, which were included with certainty in the sample; (2) large non-certainty counties, which were selected with probabilities proportional to size (PPS); and (3) small non-certainty counties, which were sampled with equal probabilities. Candidate institutions listed in the certainty counties will be sub-sampled with the same sampling rate as was used to sample non-certainty counties. Candidate institutions in the sample large non-certainty counties will be sub-sampled at a rate that will make their probability of selection the same as that for institutions in certainty or in small counties.

It is anticipated that approximately 13,000 potential institutions will be called and screened to determine eligibility. Institutions found to be eligible will be administered the MDS.

Target Population

Post-secondary education is defined within IPEDS as "the provision of a formal instructional program whose curriculum is designed primarily for students who are beyond the compulsory age for high school. This includes programs whose purpose is academic, vocational, and continuing professional education, and excludes avocational (leisure) and adult basic education programs."

The following types of institutions are included in IPEDS: baccalaureate or higher degree-granting institutions, 2-year award institutions, and less-than-2-year institutions (i.e., institutions whose awards usually result in terminal occupational awards or are creditable toward a formal 2-year or higher award). Each of the three categories is further disaggregated by three levels of control (i.e., public, private not-for-profit, and private for-profit), resulting in nine institutional categories or sectors. Universe 2 institutions, by definition, provide post-secondary education and do not disseminate Title IV funding. Those institutions are largely unknown to the Department of Education and are the focus of the new data collection effort.

The geographical scope for the non-Title IV institution sample includeS the 50 States, the District of Columbia, and the Commonwealth of Puerto Rico.

Primary Sampling Units

Primary sampling units (PSUs) for the area sample are counties. It is important to have a PSU coincide with a jurisdiction (or groups of jurisdictions) with clear geopolitical boundaries. Among other advantages, this will facilitate the merging and unduplication of several frame sources as well as the use of sources organized by county, city or MSA. By means of a file that links Zip Codes to counties, any entry in a list of addresses can be matched to the county where its Zip Code is primarily located.

Sampling Frame

The first-stage (area) sampling frame will include 3,208 counties or county-equivalent areas defined by distinct Federal Information Processing Standards (FIPS) codes. Within each sample county, every non-Title IV institution located in that county will be identified. Operationally, target institutions will be located in a five-digit Zip Code area contained primarily in the county.

Institutions that are listed more than once in the sample county will be identified. Other procedures will be used for identifying duplicates, for handling “referrals”, and for adjusting the weights for multiplicity.

Measures of Size and Size Stratification

The measure of size was based on each county’s population. The square-root transformation was used to reduce the skewness of the distribution of county populations, and to reduce the expected number of certainty counties.

Three size strata were formed: 1) small counties, 2) large counties, and 3) very large counties constituting certainty units. Different sampling methods will be used in each of the three size strata. Small counties were defined as those with population below a pre-specified threshold (for example, counties with population of 50,000 or less). The allocation led to the selection of 472 of those small counties.

Within each sampled small county, all candidate institutions will be included in the sample of candidate institutions. A comprehensive list of potential institutions will be constructed in each small county sampled. The list will be examined (e.g., for name, address, phone number) to eliminate duplicates

and obvious out-of-scopes (e.g., elementary and secondary schools) before contacting any of the institutions.

Using the square-root population size measure yields 31 certainty counties, and a sample allocation of 287 counties to the stratum of large non-certainty counties. For each of these large non-certainty counties, the (second-stage) sub-sampling rate will be determined as 0.20 divided by the county's probability of selection, so as to make the overall probabilities of selection approximately uniform across strata (about 0.20). The stratification of counties by size is summarized in Table 1.

Table 1. Size Stratification by Counties

Size Stratification	Number of Sample Counties
Small Counties	472
Large Counties	288
Very Large Counties	31
Total Counties	791

Stratification and Allocation

The first-stage sample is stratified along three dimensions:

- Region
- Urban status
- Size.

As discussed previously, the size stratification variable has three levels and urban status has two levels (high versus low) based on population density. There are five regions formed by combining pairs of the 10 Bureau of Economic Analysis (OBE) regions.

The sample was allocated to each small non-certainty stratum in proportion to the number of counties in the stratum. The allocation to each large non-certainty stratum was in proportion to the sum of the size measures. The initial sample of 791 counties, which represents 20 percent of the small non-certainty counties, assigns to each institution a probability of selection equal to 0.20 by means of sub-sampling of the institutions in the larger counties.

The allocations were initially made at the size stratum level. Within the small-county stratum, there were five regional strata, each of which was subdivided into two urban-status strata. The sampling rate will be approximately the same for each of those (10) substratum cells. The large non-certainty counties were also subdivided into ten substratum cells. The allocation for each of the cells was approximately equal to the sum of the probabilities of selection. Table 2 shows the stratum definition, and Table 3 shows the allocation of the sample counties to strata.

Table 2. Stratification

Stratum Label	Size Stratum	Region	Urban Status (Population Density)
1	Small	1	Low
2	Small	1	High
3	Small	2	Low
4	Small	2	High
5	Small	3	Low
6	Small	3	High
7	Small	4	Low
8	Small	4	High
9	Small	5	Low
10	Small	5	High
11	Large	1	Low
12	Large	1	High
13	Large	2	Low
14	Large	2	High
15	Large	3	Low
16	Large	3	High
17	Large	4	Low
18	Large	4	High
19	Large	5	Low
20	Large	5	High
21	Certainty	1	--
22	Certainty	2	--
23	Certainty	3	--
24	Certainty	4	--
25	Certainty	5	--

Table 3. Sample Allocation to Strata

Stratum Label	Frame Total Number of Counties	Sample Counties
1	50	10
2	50	10
3	358	72
4	359	72
5	379	76
6	380	76
7	318	64
8	318	64
9	71	14
10	71	14
11	60	15
12	61	33
13	145	34
14	146	64
15	129	31
16	130	55
17	32	8
18	33	15
19	43	11
20	44	22
21	7	7
22	4	4
23	9	9
24	1	1
25	10	10
Total	3,208	791

Sub-sampling Institutions in Certainty and Large Counties

In certainty counties, the list of potential institutions will be sub-sampled at the same rate as the small non-certainty counties (0.20, or a 20 percent rate). In the large counties, the sub-sampling rate will depend on the county's probability of selection, so that each institution will have an overall probability of selection of 0.20. This approach will result in an approximate self-weighted sample of institutions (i.e., a sample where the probability of selection of each potential institution is the same). Institutions will be sub-sampled with systematic random sampling from the list of institutions in the county ordered by data source and by SIC codes. This implicit stratification will ensure that the distribution of sub-sampled institutions mirrors the total institutions in the county along with data source and SIC code dimensions.

The sampling plan allows for fine-tuning the sub-sampling rate in the certainty and large counties to achieve the target sample sizes. Adjustments are expected to be small enough that accurate projections can be made for the total number of candidate institutions nationwide as well as for the screening rate. Therefore, changes in the uniform sampling rates will lead to precision losses that are likely to be small.

Identification and Unduplication of Institutions

Two national databases comprised the main information sources for the sampling frame of candidate institutions. While InfoUSA is made up of business white page listings, Dunn and Bradstreet are business listings. An analysis of these two sources indicates that the overlap between InfoUSA and Dunn and Bradstreet listings is approximately two-thirds. Listings obtained from Dunn and Bradstreet will augment the InfoUSA listings; using both sources will provide a more complete sampling frame.

Secondary listings of institutions were obtained from professional associations, accreditation agencies and state sources. After these files were obtained and duplicate listings eliminated, institutions listed in the Common Core of Data (CCD), IPEDS – Universe 1, and the Private School Survey were eliminated. These secondary listings form a separate stratum that needs not be confined to the sample counties.

Specialized matching software (Automatch) has been used to compare the IPEDS universe with information from directories and commercial files. Automatch is a highly generalized matching algorithm that uses probabilistic methods to match records from any two

files. The Automatch process will lead to a developing comprehensive master list of potential post-secondary institutions. After the preliminary set of replicate samples has been determined, the master file will be subset to the sampled areas and augmented with information collected locally. To subset the master file of telephone listings to the sample counties, it is necessary to map the sample counties to Zip Code areas included in the counties.

Diagram 1 provides a summary results of the unduplication process conducted within and across data sources.