SURV 625
Applied Sampling
Spring 2023

SURVMETH 625
Applied Sampling
Winter 2023

Team Sampling Project

The State of Michigan Department of Education has been charged with responsibility for monitoring teenage smoking and drug use. State officials anticipate needing data to be able to determine whether settlements with the tobacco industry are being satisfied at present. The Department of Education officials request the design of a statewide sample of teenagers that can be used to monitor rates and extent of cigarette smoking and drug use across the state as well as by regions within the state.

As sample design experts, you have been tasked with designing a sample of teenagers that satisfies state precision requirements. You'll need to specify sample sizes and determine sampling rate(s), the number of (and selection methods for) first stage units, and optimum subsample sizes. You'll also need to stratify first stage units, allocate the sample across strata, and select first stage units in one or more strata. The client needs a written report of no more than 10 pages in length (not including appendices) describing the features of the sample design. (A description of the required elements of your report is given at the end of this document.)

The sample design is the basis for a survey to assess present teen (13-19 years of age) smoking and drug use habits. State officials have carefully reviewed alternative survey designs, including a statewide area household probability sample with in-person screening for teenagers and face-to-face or telephone interviews, a statewide screening of telephone households and telephone interviews, and a sample of schools and teenagers in school with self-administered questionnaires. For cost reasons, state officials believe that the latter school-based sample is best, despite coverage problems associated with dropping out. They are considering a separate study of the coverage properties of the school-based survey.

The population for this study thus consists of all school students grades 7 through 12 in in the state as of the time of sample selection. The survey is to be administered this fall. For the purposes of this exercise, we will accept mode, nonresponse, and noncoverage errors that could affect estimated rates and totals because of the decisions already made by the State officials.

The materials in this document will not be sufficient to develop the complete sample design. Additional information and data will be provided by Dr. Si, who will serve as a consultant. The student role is that of a sampling expert, and you will work with a team of other sampling experts to design the final sampling plan.

Population and Frame Characteristics

The study population includes all enrolled school students in grades 7 through 12 in public and other schools throughout all 83 Michigan counties. Eligibility for the study will be determined at the time of the initial contact for interview.

One part of the sampling frame is the Department of Education's 2022 head counts for public and nonpublic schools. We will assume that this list is an exhaustive compilation of schools at the time of sample selection. The PSU frame (where a school is a PSU) will be complemented by a list of currently enrolled students and currently occupied classrooms in each sampled PSU. The list of school buildings, with head counts, will be available on the Canvas site in the Sampling Project folder (**MI_school_frame_head_counts.xls**). An example of a within school list for one school that may or may not be in the sample is also stored in the same folder (**sample_school_student_list.xls**).

To the extent that there is a one-to-one correspondence between teenagers and public schools, and between public schools and the frame, a two-stage sample of school students drawn from this PSU frame will, in theory, provide total coverage of the teenage population in Michigan. In practice, sampling coverage will not be totally complete because students in home schooling and those who have dropped out will not have a chance of selection. We will accept these deficiencies for our exercise.

The sample should be designed to yield standard errors for key study variables specified by the State Department of Education. Sample sizes should be specified in terms of the number of ***completed*** questionnaires from teenagers in grades 7 through 12. To determine the sampling rate for the study, assume that 100 percent of the students within sample schools will be eligible for study, that the response rate among schools will be 30 percent, and that the response rate among teenagers within schools will be 70 percent.

State officials are interested in providing, if at all possible, separate estimates for each of nine education regions in the state, where the regions are defined by groups of counties:

Region	Counties (COUNTY_ID on the sampling frame)
1	7, 31, 42, 66
2	22, 27, 36, 55
3	2, 21, 52
4	17, 48, 49, 77
5	1, 4, 6, 16, 20, 26, 35, 60, 65, 68, 69, 71, 72
6	5, 10, 15, 18, 24, 28, 40, 43, 45, 51, 53, 57, 67, 83
7	3, 8, 11, 12, 13, 14, 34, 39, 41, 54, 59, 61, 62, 64, 70, 75, 80
8	9, 19, 23, 25, 29, 30, 33, 37, 38, 46, 47, 56, 73, 78, 81
9	32, 44, 50, 58, 63, 74, 76, 79, 82

Stratification of the sample by education region is expected to produce increased precision for survey estimates relative to that for an un-stratified sample of equivalent size. A second purpose for explicitly stratifying the sample is that it facilitates separate estimation for each region.

Sample Design Specification

Your team is to design a two-stage sample of students that meets sample size requirements (which will be provided in class). The sample specification should include sample sizes in terms of the number of schools and students to be selected, the number of schools expected to participate, and the expected number of students completing a questionnaire. This design specification will require examination of expected *cv(.)* for each survey characteristic of interest, and the value of key design parameters for each characteristic of interest, such as *roh*, the desired cluster size *b** of students completing interviews, *deff*, and the number of clusters of the desired size *b** to be selected. Relevant quantities from similar prior surveys for the variables of interest in this survey will be provided in class.

The population of schools is to be stratified by the regions above. The strata may not be of equal size, but an integer number of schools must be specified. The subsample size however does not need to be an integer, and the cluster size may vary across strata. The sample of schools and students will need to be allocated across strata. The allocation must allow for a **paired difference method of variance estimation** across clusters, which needs to be clearly described in the report.

For a subset of **two strata (only!)**, select the sample of schools, including the calculation of a sampling interval for each stratum to achieve *epsem* within the stratum, choice of a random start for each stratum, and specification of a minimum measure of size for a school that can be tolerated before linking must occur. Create a table similar to the following, summarizing the first stage stratification and selection:

Stratum	f_h	Mos_h	a_h	b_h	k_h	Random Start
1: Description						
2: Description						

The selection of schools will require an unbiased technique for handling undersized units *after* selection occurs. In addition, once schools have been selected, compute for each the within school sampling rate that must be applied in order to achieve the stratum sampling rates above. Summarize the results of the selections in a table similar to the following one for **each stratum**:

Selection Number	Stratum	School	Cumulative Mos	Within school interval
1				
2				
3				
…				

When, through the linking procedure, more than one school is selected, list each school in the selected linked unit on a **separate line**. That is, each selected school within a linked unit will have the same Selection Number.

Each team must specify the procedure that will be used for selection of students within selected schools. This includes an example of how the selection method would be applied to one school for which we already have the list of students (see the aforementioned example list on Canvas).

The client is interested in a summary of the expected levels of precision for the survey. For selected key variables, including **ever smoked one cigarette** (expected proportion = 0.25), **ever smoked marijuana** (expected proportion = 0.15), and **age when first approach to smoke cigarettes or marijuana** (expected mean = 12, expected SD = 1), include in the report a table of expected 95% confidence intervals for the total state, and for a hypothetical subclass that consists of approximately 20% of the overall state population of students (low income households), report the same expected confidence intervals.

Estimation Plan

Finally, to complete the design specification, describe estimation methods for means and proportions (rates) for the total state and for the 20% subclass. The methods must include the specification of any weights to compensate for unequal probabilities of selection and variance estimation procedures that account for weighting, stratification (explicit and implicit), and clustered selection. A plan for computing 95% confidence intervals for means and proportions must also describe how many degrees of freedom are available for interval formation based on the paired difference sampling error calculation model used.

Report Preparation and Contents

The report will be written by four- to five-person teams organized by Dr. Si. Each team will submit a single report, and all team members earn the same technical score on the project (maximum 100 points). The required technical elements of the design are as follows:

- Overall Design. Overall sampling rate, *cv(p)* achieved for each variable, *roh* and *deff* for each variable, b_{opt} for each variable and b^* overall, n_{SRS} for each variable, and n and a for the overall sample. **(25 points)**

- Allocations. For each of the nine strata, the sample size *n* and PSUs *a*, average cluster size b_h^*, sampling interval k_h, random start for PSU selection, and minimum MOS. **(25 points)**
- Selections. Selection of first stage PSUs in two of the strata using a systematic PPeS method, including the linking of undersized units, the identification of oversized units, the calculation of within-PSU sampling rates and intervals, and a detailed description of the within-PSU sampling procedure, including an example for one school. **(25 points)**
- Estimation Procedures. Description of the creation of sampling error codes (stratum and cluster codes) for variance estimation, the formulas and descriptions of all symbols for estimation of a mean or proportion for the total sample and for the 20% subclass, the formulas and descriptions of all symbols for estimation of the standard error of a mean or proportion for the total sample and for the 20% subclass, and the calculation of 95% confidence intervals for the means and proportions, for both the total sample and the 20% subclass. **(25 points)**