USE OF MINI-PROJECTS IN THE TEACHING OF SURVEY SAMPLING

Lynne Stokes
Southern Methodist University
USA

Much of the material in the graduate survey sampling course is tedious to teach and learn. The classroom is enlivened and students are better able to use the concepts taught in the course when they have some experience applying it in real populations. This paper discusses some activities or mini-projects that can be used in the classroom to give students this experience without introducing come of the complexities associated with a full-scale project.

INTRODUCTION

The presentation of material in many Survey Sampling courses and books has a repetitive "design-estimator-variance" form. The same few methods are used repeatedly in derivations of standard errors for each design. This material can be dull both to present and to learn in a lecture format. This observation is not new, and others who teach the material have suggested that experience in actually implementing the methods on real populations has both increased the students' interest and improved their skill in knowing how to use the tools of the course. Gitlow (1982), Chang, Lohr, and McLaren, (1992), and Schwarz (1997) have all provided virtual "cities" from which students can sample households using a variety of designs. Realistic responses on topics such as demand for daycare or usage of cable tv are then available for the sampled households. Dietz (1993), Bishop (1998), Single (2000), and Paranjpe and Shah (2000) all describe exercises in which sampling is done from physical populations.

The purpose of this article is to make available additional activites, which we call miini-projects, with a similar purpose to those cited. Most of the exercises previously published were designed to be carried out by students outside the classroom, but we have found that some in-class time spent on these activities is also useful. In-class sampling activities are used to allow learners more variety in the methods available for them absorb material. The mini-projects we describe are easy to implement in any campus environment. In this paper, we give an overview of several of the projects and describe how they can be woven into the semester-long sampling methods. We also describe the learning objectives we had for their use.

PROJECT SELECTION

Finite populations are everywhere, and the concepts of probability sampling require virtually no assumptions about the populations to which they are applied. For this reason, it is easier to find real applications for the tools students learn in a sampling methods course than many other statistical methods courses. The mini-projects generally consist of the application of specified design and estimation procedures on a population that is easily accessible in the classroom. These applications are artificial in the sense that the parameters being estimated are not necessarily of intrinsic interest. The populations sampled consist of objects for which accurate frames can be acquired. Further, we try to set up projects so that the collection of data requires that the students make a measurement or attribute assessment on the sampled objects. This provides them with some experience on practical issues, such as the need for accurate definitions. However our projects do not require gaining the cooperation of respondents. This helps keep the skills needed for the project close to those typically covered in our courses, which do not include questionnaire design and include little on the topic of adjustment for non-sampling errors.

EXAMPLES OF MINI-PROJECTS

These activities have been used in a graduate level course sampling course and an undergraduate introductory statistics lab. We will focus mainly on the graduate course, which covers the traditional topics from a text such as Lohr (1999) or Cochran (1977). The students typically come from a variety of departments, including statistics, engineering, biology, and business. Initially the projects were introduced to give the students with less technical backgrounds another method of absorbing the material besides the equation-dense texts used in the course at the time (Cochran, 1977; Swenson, Sarndal, & Wretman, 1992). But it became apparent that even students who are accustomed to reading this type of material are better able to put the methods in practice when they have been exposed to real data collection activities.

My goal is to use an activity in class about every third 75-minute class period. Class time required is generally around 20 to 25 minutes. Most activities are done in small groups. Sometimes the groups performed identical procedures, giving an empirical view of the sampling distribution of the estimator being computed. Other times, groups use different designs or perform different parts of a larger activity, and combine information to produce a result.

The first class meeting of the course includes an activity designed to convince the students of the need for probability sampling. The rectangle sampling activity described in Gnanadesikan *et. al.* (1997) is excellent for this purpose. In it, students are asked to select two samples of 10 rectangles from a population of 100, with the goal of estimating mean area. For sample 1; they are instructed to select 10 "representative" rectangles, and in the second they are asked to use a random number table to select a simple random sample. A sample mean is computed from each, and a frequency table of the class results prepared on the board. The sampling distribution of the "representative" sample means is nearly always centered below the true mean area due to the fact that students tend to underrepresent the small rectangles in their judgment samples. The advantage of a probability design is usually dramatically illustrated with this exercise, even with small classes (e.g., less than 10). A by-product of this exercise is that students gain experience using a random number table and they also are able to use the one provided with the exercise throughout the course.

In subsequent class meeting, we try to associate one in-class activity with each design covered. The activities often require that the students go beyond the lecture material in some way. That is, they may require that the students develop a simple extension to a method presented in the lecture, that they speculate on how the method being implemented might compare to one covered earlier, or that they determine how to operationalize some concept covered only technically in the lecture. This is one justification for the use of class time for these activities. If the same exercises were assigned as homework, the students might find them frustrating, or have a greater chance of missing their points.

To facilitate comparison across designs, we often use the same population throughout the semester, in much the same way as Paranjpe and Shah (2000) describe. They use a dictionary as the finite population to be studied during the term, with the goal of estimating the number of words it contains. (In their case, they actually know the true parameter value, while in most of the examples we use, we generally do not.) We have used the following populations:

(i) The course textbook.

The table of contents serves as a frame. The parameter the students are asked to estimate is the number of equations in the book. (This always leads to a discussion of what is meant by an equation, and the importance of specifying an operational definition.) Pages can be easily sampled individually, or sections can be used as PSU's and sampled with equal probability or proportional to size (by selecting the section into which a randomly selected page falls).

(ii) A student directory

Again, pages serve as a frame. The students are asked to estimate the number of listings, the proportion of students who are foreign, and/or the proportion who list email addresses. Whatever parameter is chosen, it should be one that can be observed (either accurately or nearly so) from the listing itself.

(iii) A bookshelf in the classroom

The conference room where the class met has a wall of seldom used built-in bookshelves. Journals filled about half of the shelves, with the remainder containing hard-backed books. The students were asked to estimate the number of pages on the shelves. Because of the difficulty in determining this in some cases, other characteristics might have been better, such as mean age of the publications, or the proportion published outside the U.S.

(iv) The rectangle population

The original rectangle population was rearranged (by a former student, Elizabeth Murff) into two other configurations. One groups them by geography (to facilitate cluster sampling) and one groups them into strata defined by the size of their width. This population was used with the lower level students (business undergraduates) rather than the graduate students just to give them experience with implementing alternative sampling designs (The spreadsheet containing these files can be found at faculty.smu.edu/ICOTS/Appendix1.xls).

Now we provide a brief description of the activities used in association with each of a number of topics covered in the course.

a. Simple random sampling without replacement (srswor)

The topic of work sampling can be introduced in association with srswor. This exercise was used during a semester in which a number of business and engineering students were enrolled in the sampling class. Following a brief description of work sampling, the students were asked to suggest an activity that occurs in class whose duration could be estimated. Alternatively, they can be asked to select an event whose frequency (or total occurences) can be be estimated. For example, they might choose the amount of time the instructor has his or her back to the class (duration) or the number of steps the instructor takes during the class (total occurrences). We made a data collection assignment for each student during the next class period. Each student is assigned a randomly selected "moment" (one or more depending on the size of the class) during which they observe whether or not the activity is taking place. When estimating total occurrences, it is best to assign students a longer unit of time, such as a minute, during which they count the number of occurrences of the event. The data is collected during the class period, which proceeds as usual. A clock in the classroom is the official sampling clock. The data is recorded on a handout, and the following class period the students work in groups to complete the activity. They are asked to suggest disadvantages of the design, and how it might be improved. The goal of this question is to have them recognize that a random sample may not be as "representative" as they would like, leading them to suggest that a design that spreads out the sampling units would be better (leading up to stratified or systematic designs).

b. Stratified sampling

Graduate students in statistics generally find it easy to follow the computations justifying Neyman allocation and deriving an expression for the variance reduction available through stratification. They find it more difficult to recognize how to form strata or to allocate sample to them in a real population. The project used in this unit asks students to select first a simple random sample of a specified size and then a stratified sample from one of the populations described above, such as the textbook or student listings in the directory. They are to make estimates of a specified parameter and its standard error from each. They are then asked to select the same size sample using their stratified design, but no further details are offered on how this should be done, except that the goal is to make the design as efficient as possible. They do this activity in teams of 3 or 4, which allows the sampling and estimation to proceed more quickly. They must determine not only how many, but also how to form the strata, and then how to allocate the sample to them. In the textbook population, it is usually obvious that introductory chapters and the indexes are less dense with equations than other parts of the text, and thus are can reasonably make up a stratum, leaving the remainder for one or more additional strata. In the directory population, if the parameter of interest is the proportion of foreign students, students may make note that stratifying by letter grouping is useful (e.g., X and Z sections have higher proportions of foreign students than others). Once the students decide on a design, select the sample, and calculate their estimates and standard errors, the groups share their results. Some observations can usually be made from this comparison. First, fewer strata generally do as well as more strata, as assessed at least from estimated standard errors. Second, the extra effort of Neyman allocation is generally not useful in these population, and a discussion of why this is is enlightening for the students. (In the textbook population, the guesses of stratum variance are so poor that the allocation is not really optimal; in the directory population, the strata variances are similar since they are of the form p(1-p).)

c. Systematic sampling

I use an activity after this design is introduced to encourage students to think about ways to assess the standard error of an estimate from a systematic design. I first ask each student individually to select a systematic sample of a given size from a population, such as the textbook. Then I ask that they form groups of about 3 students each, pool their data to form a group estimate, along with its standard error. After they decide to average their individual estimates, it is usually easy for them to figure out how to estimate its variance as $\sum_{i=1}^{k}(t_i-\bar{t})^2/k(k-1)$, where k is the number of students in the ith group and t_i is the estimate of the ith one of them.

d. Use of cost information

In this activity, students are asked to compare a cluster design, where clusters are chosen by srswor, with a two-stage design with srswor at each stage and a constant number of ssu's chosen from each psu. They are asked to determine which is most efficient for estimating a specified parameter. The populations I have used for this activity are the textbook and the bookshelf populations. The same psu's are used in each design (a section and a shelf, respectively, for the two populations). The groups work in teams, and one person on each team collects the cost information by timing the data collection processes. The sample sizes for the two designs will not be identical since the psu's vary in size, so that the students must figure out a way to compare efficiency when neither cost nor variance are constant for the two designs. I have found that students need fairly explicit instructions on how to do this, so they are led through a series of questions. First they are asked to determine how many psu's would be needed to achieve the same variance with the cluster design as with the two-stage design, and then they are asked how much this would cost, based on the cost data collected. We discuss as a group what they need to compute from their sample data to make such a determination, and then the calculations are done outside of class. During the next class period, the results are compared. In the textbook population, the cluster design is usually more efficient because most of the effort (cost) goes into locating the sampled pages. In the bookshelf population, the opposite is true, since the shelves tend to be homogeneous with regard to the types of books on them.

e. Horvitz-Thompson Estimation

I have a unit on the Horvitz-Thompson estimator and weighting, in which students are shown that as long as the selection probabilities can be calculated, then they can produce an unbiased estimator of mean or total. In this exercise, students are given a student directory and asked to consider the following designs: (i) Select a srs of pages, and a systematic sample of lines on the selected pages (note: the number of lines per listing varies); (ii) Select a srs of pages and randomly select one listing from each page; (iii) Select a systematic sample of pages and observe all listings on the selected pages. For each design, the students are asked to determine how to form unbiased estimates of total (i.e., to calculate sample weights).

f. Complex designs

Toward the end of the course, the students are given an activity that takes about 60-70 minutes of class time spread over 3 class periods. It provides the students with experience in combining the various methods learned in the course into a single design, a topic which is often not covered in a standard course. In the first of the three class meetings used for this activity, the students are asked to plan a design to estimate the proportion of empty parking spaces available in university lots at a given time (during the next class meeting time), as well as the number of vehicles illegally parked at that time. The students are assigned to teams. All teams are given a university parking map which serves as the frame. They are free to try to obtain any additional auxiliary data they can to improve their design or estimation process. (University police do have available records of number of parking spaces and number of stickers owned for each lot, which several groups did obtain.) In the second class meeting, the studens are given about 30 minutes of class time (15 minutes before and after the target time) to collect the data. They had to work between the two class meetings to finalize a sample design with the objective that it should be as efficient as possible. Since the campus is large, they also must consider the cost of data collection, since some lots require much greater time to reach than others. In the third class meeting, students share their designs and estimates, which they have computed between class meetings. Often it happens that estimates differ by larger margins that they should, based on standard errors. This provides an appreciation for non-sampling errors, such as measurement error and non-response bias. Both of these can occur in this exercise, since students may use different definitions of *illegally parked*, for example. Non-response can occur if students are not able to execute the design as planned in the allotted time.

CONCLUSIONS

We believe that activities such as those presented above are valuable in helping the students learn to use the tools that are covered so abstractly in most sampling courses. Though I have no defensible experimental data, I have for several years used the same series of questions on the final exam. These questions require that the student plan and execute designs on physical populations that are made available to them during the exam period. For example, one question asks students to estimate the amount of money (monopoly money) present in a set of closed envelopes. On the outside of each envelope is written the number of bills in the envelope (ranging from 1 to 6). They are allowed to look at (but not in) the envelopes before they plan their designs, and they may not sample more than a given number of envelopes. They must determine a design and make an estimate. In semesters when less class time is taken with in-class exercises of the type described here, students score lower on these questions.

On a less positive note, however, I have not noticed that course evaluations are higher in semesters when these exercises are used extensively. I believe that one reason for this is that statistics students are not accustomed to this type of classroom experience. Some students who are skilled in mathematics have had difficulty with these exercises, perhaps because of less interest in applications. On the other hand, I have had students to become extremely enthusiastic about sampling after taking the course. My impression is that while the average evaluation is about the same, the variance of the evaluation may be larger when this approach is used in teaching. One of the most important benefits of this approach to teaching this material is that it is so much more fun for the instructor. I find that I look forward to the class meetings when certain activities are planned in a way that I did not in the lecture format. The classroom atmosphere is livelier, and there is more opportunity for interaction with students.

REFERENCES

Bishop, G. (1998). A series of tutorials for teaching statistical concepts in an course: I. Sampling from an aerial photograph. *Journal of Statistics Education*, *6*(2).

Chang, T.C., Lohr, S.L., & McLaren, C.G. (1992). Teaching survey sampling using simulation. *The American Statistician*, *46*, 232-237.

Cochran, W.G. (1977). *Sampling techniques*. New York: John Wiley and Sons.

Dietz, E.J. (1993). A cooperative learning activity on methods of selecting a sample. *The American Statistician*, *47,* 104-108.

Gitlow, H.S. (1982). *Stat city: Understanding statistics through realistic applications*. Homewood, IL: Irwin.

Gnanadesikan, M., Schaeffer, R.L., Watkins, A.E., & Witmer, J.A. (1997). An activity-based statistics course. *Journal of Statistical Education, 5*(2).

Lohr, S. (1999). *Sampling: Design and Analysis*, Pacific Grove, CA: Duxbury Press.

Paranjpe, S.A., & Shah, A. (2000). How many words in a dictionary? Innovative laboratory teaching of sampling techniques. *Journal of Statistics Education, 8*(2).

Sarndal, C.E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Single, R.M. (2000). Using the National Health Interview Survey and the 2000 Census to introduce statistical sampling and weights. *Journal of Statistics Education*, *8*(1).

Schwartz, C.J. (1997). StatVillage: An on-line, www-accessible hypothetical city based on real data for use in introductory class in survey sampling. *Journal of Statistical Education*, *5*(2).