Teaching Sample Design: Virtual Population Project

Fajar Budiyanto

Fakultas Ekonomi, Universitas Mayjen Sungkono Mojokerto

fajarquantum@gmail.com

Abstract

Examples of survey design are subjects that are usually taught to students pursuing minors or statistical majors at the end of their undergraduate degree. This article describes evaluation initiatives that encourage active learning and help develop important skill sets for statistical practice. The project is completed in pairs and submitted in two parts. This allows input from the first section to be forwarded to the second section. Ideally, students should have experience taking samples from the actual population. But the time it takes to receive approval from the university's ethics committee may not be appropriate for a short course. The alternative is to use an online virtual population such as Island software, which provides students with experience in setting up sampling systems, getting approval from potential participants, and collecting data. Written communication skills and teamwork are highly valued by statistical graduate employers. The project promotes collective learning in sample survey design, statistical analysis of the data obtained, and preparation of final written reports. It can be easily adapted for first year students and can also be expanded to accommodate students at Master or Honors level.

Keywords: Education; Sampling; Computational Statistics; Statistical Assessment

Article History

Received : August 7th 2020

Revision : October 11th 2020

Publication : Dec 30th 2020

INTRODUCTION

Higher education institutions that offer minor or applied statistics majors are responsible for training graduates to become trainee statisticians. To qualify with a minor or statistical major, an undergraduate will study many maths and statistics (unit) courses at different levels, each contributing to the overall curriculum and desired graduate attributes. Employers prefer multi-skilled statistical graduates, as discussed in the American Statistical Association Curriculum Guidelines for Programs in Statistical Sciences (Association, 2014). It states:

Efficient statisticians at all levels need to learn an integrated mix of skills focused on statistical theory, statistical implementation, data processing and computing, mathematics and communication (Association, 2014).

In the top-level statistics unit, there is the potential to provide a more sophisticated evaluation system, including the scaffolding skills and principles acquired in the previous unit. The goal is to integrate the synthesis of critical, low, and high skills into learning objectives. A well-designed evaluation task will be appropriate for the level of study; will involve students; be on time for active learning to take place; and will include data

manipulation and mathematical calculations to apply statistical theory. Other pedagogical considerations outlined in CGUPSS (Association, 2014) include addressing issues with meaningful context, offering students the opportunity to develop communication skills and work in teams, and providing authentic feedback.

Sample survey design is a subject that is usually offered as an optional top tier after the first year statistics unit. Typically, the goal is to combine theories and practices of sampling methods such as simple random sampling, systematic sampling, multilevel sampling, and possible cluster sampling. In addition, general estimation procedures for data analysis are also included

In this article, a survey sampling project is identified that involves sampling the island's virtual online population (Bulmer & Haladyn, 2011). The project is part of an evaluation section of a six-week to 13-week sample survey project for top-level undergraduate students; the rest of the unit focuses on experimental design. The project was conceived to be implemented in small groups, allowing students to practice a combination of statistical and critical communication skills, resulting in written studies. Each part of the assessment task is defined based on the recommendations of the Evaluation Guidelines and Instructions in the College Statistics Education Report (Association, 2016) and CGUPSS (Association, 2014). A review of these recommendations is provided in Section 2, which discusses related literature and outlines of the island's climate. Each part of the project is defined in Section 3 with Teacher's Notes and appropriate reasons; The evaluation framework is illustrated in Section 4 with a description in Section 5.

The GAISE report contains six recommendations (Association, 2014) listed here for convenience. The first two discuss what is taught: (1) to teach statistical thinking; and (2) concentrate on logical understanding. The following four discuss how to teach: (3) Combine real data with meaning and purpose; (4) Promote active learning; (5) Use technology to explore ideas and interpret data; and (6) Use tests to improve and measure student learning. While the GAISE report focuses on statistical introductory courses, six recommendations were expanded beyond the introductory stage (Association, 2014).

General undergraduate statistics programs have been widely distributed in literature; list of main articles published in (Horton & Hardin, 2015). They emphasized the need for student participation in the overall data analysis process. Programs can be innovative with their curriculum by providing a combination of theory, engineering, coding, and application (Horton & Hardin, 2015). Panel discussion paper by Fecso et al. (1996). is one of the papers mentioned under the "Second Course" on the teaching of sampling surveys. It offers insight into the approach to survey teaching sampling and the materials that should be included in the curriculum. Depending on the diversity of student groups in the classroom and course objectives, the main challenge is to reconcile theory with realistic. Each of the five panel members emphasized the value of practical experience (Fecso et al., 1996).

Ideally, a realistic experience would include data collection according to a specific sampling design; data analysis using statistical software that integrates sampling design; and communication of the results produced in the report; thus combining theory with practice. In cases involving human participants, approval from the university's ethics committee is required. Obtaining this approval may take time and may not be possible within a limited timeframe for the undergraduate unit. The solution is to enter population data, putting the focus of evaluation on the post data collection process. The drawback of providing population data is that students do not experience: the difficulties and chaos often associated with the creation of sampling frameworks; sampling of individuals who may choose not to respond; data collection process; and subsequent data manipulation is often necessary. See Hulsizer & Woolf (2009) for more discussion on the advantages and disadvantages of using a reality-based and ar dataset. The alternative is to use online virtual populations, such as islands, that do not require ethical approval from university committees.

Virtual island application

The island is a virtual online community created by Bulmer & Haladyn (2011) consisting of twenty-seven villages spread over three islands (<https://islands.smp.uq.edu.au/>). Islanders live in houses in communities of different sizes; the villages have schools and other city facilities, such as desks, classrooms, clinics, museums, and three major cities also have universities. There are several ways to find islanders: individuals can choose a house that lists the inhabitants of the island; local schools register teachers and students who are enrolled; the office has a population pool, a list of residents working in several professions or members of recreational clubs; the hall lists births, deaths and marriages; the university enrolls staff members and students. Islanders can be contacted, asked to perform certain tasks, or answer questions.

The prospects for using Island for statistical projects are endless. At undergraduate level, the island's virtual environment became popular with educators of introductory statistics courses for teaching students general statistics and experimental design (see, for example, Baglin, Bedford, et al., 2013; Baglin et al., 2015; Baglin, Reece, et al., 2013; Baglin & Huynh, 2015; Linden et al., 2011)

One of the main benefits of the Icelandic curriculum is that it allows students to observe the entire statistical problem solving process within an attractive framework.

The use of the island is conducive to active learning (Prince, 2004; Tirlea et al., 2016), because it offers students the ability to participate in the learning process. A small pilot study found that common myths about sampling can be corrected by including active learning — either with virtual online simulation environments (such as islands) or interactive classroom exercises (Tirlea et al., 2016). Focusing on investigating exploration data, the islands have been used for problem solving exercises in Australian selective high schools. Pilot studies show that students' attitudes to statistics improved significantly after surgery (Baglin & Huynh, 2015; Huynh et al., 2014). While undergraduate statistics

students are a different target group to high school students, the use of the island can also increase university-age student participation. This article contributes to the growing literature on how the island can be used for statistical projects (new instructors on the island can email the island@maths.uq.edu.au to create a new account). In particular, the project mentioned in the next section is planned for students enrolled in the top-level survey sampling unit.

METHOD

Learning outcomes from this project include:

1. the construction of sampling systems;
2. critical assessment of methods used in surveys and reliability of results;
3. the use of statistical tools to perform effective analysis and interpretation of outputs;
4. direct communication about findings and assumptions in official written statements;
5. work collaboratively with other students.

A brief description of the project objective and the predetermined context is listed below. To gain a deeper understanding of the physiology of islanders, large-scale health surveys are planned to gather information about specific physiological actions, including basic actions such as age, height, weight and gender, as well as other clinical measures. The volume of a person's lungs can be measured by the volume of forced expiration (FEV) measured in liters by a spirometer. The purpose of this project is to conduct a pilot study using a variety of sampling designs to recommend one of them to the survey designer. The goal is to estimate the average lung volume of the population of individuals who are at least 15 years old.

The project is divided into two parts and completed in pairs. Part A consists of three parts: Part 1 — Variables and Context; Part 2 — Population framework and sampling; and Part 3A — Simple random sampling. Part B consists of three parts: Part 3B — Simple random sampling analysis; Part 4 — Random sampling of strata; and Section 5 — Summary and conclusion. Project evaluation is 15% of the total value in the unit.

In any analysis, statisticians are faced with three initial problems: understanding problems, understanding context, and understanding how interest variables are evaluated. In Part 1 of Part A, students are expected to review general knowledge of pulmonary volume or pulmonary capacity (FEV) spirometer measurements such as how FEV is measured and the estimated range of values for adult humans. Two specific studies were needed that reported an estimate of the average lung volume in the general healthy population; for example, for different age groups, genders, occupations or ethnic groups. For each study, it is important to summarize goals, including the target population; sampling design, including sample size; identified estimates and their standard errors or relative standard errors; and helper variables used in the analysis. Teacher's Note: In

practice, because students operate in pairs, each student is expected to find one specific analysis. Proper reference is also planned.

Rational: This section encourages students to read literature on previous relevant research, to think statistically in a particular field and to be an objective consumer of relevant papers (Goal 1) (Association, 2014). CGUPSS pedagogical considerations include "presenting problems with substantive context" (Association, 2014).

Population and sampling

In Part 2 of Part A, each student's spouse must select and determine the target population, indicating that the population is at least 200. For example, students currently enrolled at a particular university can be a target demographic. An administrative list is required to conduct any sampling of the identified target population.

The sampling framework is a list of sample units from which samples can be taken (Lohr, 1999). Students are expected to use island administration tools to build their sampling systems. For example, Colmar University has 498 students (at the time of writing) in different faculties.

The next step is to set up a sampling framework using spreadsheet tools (such as Microsoft Excel or Google Sheets). With a list of names, id variables can be easily created, and then each person's gender, age, place, and island where they live can be collected by selecting each name. The About page includes a basic profile from which you can get additional information (Figure 1).

For example, if a group decides to research a student, this page will contain information similar to what will appear in the student administrative list. No one in the sampling framework has been contacted at this time

The last sampling framework was created by combining separate spreadsheet files from each student in the school. Some data cleanup may be required after merging, as each student may have used a different cell format or category code to record the data. Examples involve using different numeric codes or combinations of uppercase and lowercase letters when recording observer functions for nominal variables such as gender. Students are then expected to import data from comma-separated files into sas.

Population

Populations can be represented by creating acceptable summary tables and plots for people included in the sampling framework. For example, age distribution by gender can be seen in a side-by-side box plot. Teacher's Note: Here, students are not given clear instructions — data manipulation and the appropriate selection of tables and plots are part of the assignment. Creating a sampling framework is not an easy task as there is no immediate way to retrieve the selected name; they should be copied from the Islands web page. When selecting students, this can be done in a variety of ways: first, all names with faculty titles can be copied and pasted into a text editor and then into comma-separated files; or any set of names under each faculty can be copied and pasted. Any lines that make

up the headers on the website can then be deleted. Because it takes time to record data, students in the group are advised to share these tasks and merge separate files.

Some students find errors when importing data from a spreadsheet file into sas if the cells are empty or if a specific cell format has been set. Another problem that the community is reporting is that some university graduate students are in the midst of data collection on their sampling frameworks. The group addressed the issue by taking footage from students on a specific date and explaining what they had done in their learning. It is also possible to change the age of certain participants (as they may be birthdays) during the data collection process.

This unexpected issue prompted a discussion of issues that could occur during the lengthy study-gathering process and the need to identify the reach of the target population. Rational: In Recommendation 1, recommendations include using technology to effectively manage data, explore and visualize data (Association, 2014). CGUPSS pedagogical considerations include" experience with statistical computing and data-related skills (Association, 2014)

Simple random sample

The project noted that the budget requires a sample of $n=80$ people to be included in the pilot analysis. In Part 3A of the project, students were asked to select a simple random sample (SRS) using SAS and report as well as interpret the resulting sampling weight ($w_i=N/n$) where N is the population size, n is the sample size and $I=1, \dots, N$.

The next step requires the approval of each islander in their sample to participate in the analysis. For each person sampled, select the Tasks page and then the Spirometer under Physiology will seek approval from the FEV to be measured. If permission is granted, FEV calculations must be performed. It takes time to complete: the white heart is filled to show improvement, as seen in Figure 2. The results of each task are reported on the Tasks page of the individual.

Students get experience with a non-response unit if someone disagrees. Adjustments to sampling weights for any non-response can be made in such a way that the number of sampling weights for respondent sample participants is equal to population size N . Students are expected to report response rates and identify their individual samples by creating summary tables and plots suitable for any additional variables.

Teacher's Note: Choosing to set the same interest variable for all students allows multiple evaluation comparisons to be submitted for assessment. FEV measurements are chosen for a number of reasons: these are continuous variables, there are important studies in the literature; and there is variability in measurements between individuals. The definition of measurement can also be a topic of discussion in the classroom. For example, are there different ways to calculate FEV? In the case of repeated observations,

why is the FEV's highest observation to be that paling relevant in this context? Students are given examples in the classroom on how to change the weight of the sample.

Rational: The use of physiological measures such as FEV allows students to think statistically about acceptable statistical measures and brightly recommend 1 (ASA). 2016, p. 12. The willingness of the islanders to approve the implementation of tasks such as FEV calculations raises the discussion of ethical issues as stipulated in Goal 9 (Association, 2014). Issues related to lost data can also be the focus of class discussions..

RESULTS AND DISCUSSION

Report part A

Part A of the project consists of stuon three pages for Part 1. 2 and 3A to be sent to each pair in pdf format at the end of the third week of the semester. Until submitting the paper, each student in pairs is expected to read, review and give input on their respective contributions. In addition to the assessment scheme, various suggestions for teamwork and some tips for writing reports are also given. Feedback is provided in the form of written comments on PDF documents and tagging files (see Appendix B). For reporting styles, only formative comments are made in Part A, offering students the opportunity to develop their written communication skills in Part B.

Teacher's Note: At this stage, students may have little knowledge of writing statistical studies. The author's experience with students in this unit in previous years prompted the decision to split the report into two parts for three main reasons: (1) to provide feedback on their population choices which was helpful in correcting misconceptions; (2) provide feedback on their report writing style; and (3) to provide a deadline for Section A for students to participate in the analysis. The consultation period is well spent by students who ask very important questions before the Part A deadline, showing a timely dedication to the content of the subject.

Rational: Recommendation 6 involves offering "useful and timely feedback" in formative assessments of "monitoring and improving student learning." Suggestions include well-coordinated tests with subjects as taught; written assignments that allow students to improve strong communication skills; and encourage students to work in groups to facilitate learning.

Easy random sampling studies

Part B of the project consists of analysis of data obtained from various sampling schemes. Section 3B focuses on the study of data from simple randomized samples collected in Section 3A.

Review of exploration data

Section 3B of the project includes the use of SAS to measure summary statistics and create plots, e.g. histogram and box map, to illustrate a simple random sample for

FEV. Students are expected to comment on the average population estimate and other summary statistics, such as Mini Mom and Maximum grades, with reference to the results of previous studies mentioned in Section A.

Variance Sampling dan Normal Error

Students measured variance sampling calculations and standard sampling errors generated for population estimation of the FEV mean. This involves applying the appropriate formula using (calculated sample variants obtained from SAS.

Teacher's Note: Students soon learn that each time they take a sample, the average sample result is different. This encourages discussion about realistic use of random seeds when samples are taken using software. Further discussions may include the definition of parameters; assessors and their properties; and parameter calculation (including notation). In other words, for each parameter, different samples can produce different estimates when samples are taken using different random seeds. Although each estimated particle average cannot match the actual average value, the overall estimator is not biased, as shown in sampling theory. Applying a variance sampling formula allows students to associate sampling theory (applying notation) with SAS PROC SURVEYMEANS performance.

Rational: The idea that different samples can produce different estimates of sample averages and standard errors is the result of active learning related to Target 2: 'understanding and applying representative sampling concepts for observational analysis' (Association, 2014). Goal 3 refers to the interpretation of statistical summaries and their graphical appearance (Association, 2014). The central function of variability is clarified in Goal 4 (Association, 2014). Students measure and interpret various measures of variability, such as: sample variance; standard deviation; distance; and the inter-quartile range. Interpretation includes recognition that these numbers are projections of a limited population. Students need to understand the principles of sampling variability calculated by variance calculation of sampling and standard errors, and the differences between limited and infinite populations.

Goal 5 involves the introduction of the idea of "random", the difference between probabilistic and non-probabilistic samples, and the generalization of observational analysis findings on populations. Students are required to determine the probability of selection and the person in the sample, known and not zero for the probability of the sample (Lohr, 1999). In the case of SRS, the selection opportunities are the same for all individuals sampled. The sample weight (w_i , $i = 1, \dots, n$), the probability of reciprocal selection, is expressed in the production of the SAS Survey Tool. It is important to interpret weights and incorporate the concept of generalization. This part of the project confirms the idea that the number of sampling weights is equal to the size of the N population; and that each sample unit represents a certain number of units in the population (Lohr, 1999).

Cascade random storied

In Section 4A of the project, students plan to perform random, multilevel samples using their original sampling framework. Students need to propose and identify two or more strata for the purpose of estimating the average FEV in their target population and justifying their preferences. For example, if the target population is currently enrolled at Colmar University, gender can be described as a strata variable. Its job is to assess the population size (N_h) for each strata (h) and evaluate their SRS sample from Part 3 according to stratum. The study involved comparisons of results between strata involving simple summary statistical tabulations, the development of histograms and side-by-side box plots, and comments about results.

Proportional assignment

Proportional allocation sets the sample size (n_h) of strata h to be proportional to the size of the population (N_h). Students are expected to determine strata sample sizes using a proportion allocation given the total sample count ($n = 80$) and the unique population size for their analysis. The resulting sampling weight for each strata can then be determined and interpreted.

Optimal assignment

Optimal allocation minimizes variance of population estimates with restrictions, typically fixed sample sizes or fixed costs. In this segment, students are required to calculate strata sample size using optimal allocation considering fixed sample size (n). The resulting sampling weight for each strata can then be determined and interpreted.

Teacher's Note: This part of the project opens up a debate about strata concepts; search for sample mean and variance variability between strata in sample summary statistics; and variations between proportional and optimal allocations. The discussion may also refer to previous research findings that have been questioned in Part A.

Rational: Goals 2 through 5 are applied here in the same way as in the previous SRS section. But in this section to propose and identify two or more appropriate strata, students are asked to explain and interpret the distribution of FEV variables according to one or more other variables. This concerns the statistical thinking mentioned in Goal 6 (Association, 2014).

Cascade Randomized Samples — Statistical Studies

Section 4B of the project includes the use of SAS to take two random, multilevel samples based on the specified strata and strata sample sizes, assuming both: (a) proportional allocation; and (b) optimal allocation. For each allocation, the average FEV estimate and associated sampling standard errors can be calculated using SAS.

Teacher's Note: Discussions in the classroom should include: why random storied samples can be taken from SRS; and when different allocation methods can be used. Students need to create new variables (using numeric codes) depending on the strata they choose. SAS demands that data be sorted according to strata variables before multilevel sampling takes place. SAS tips can be given to remind students of these criteria. Some

students first chose to measure the average estimate of sampling and variance for each strata. They then demonstrated how they were implemented to achieve total estimation, thus demonstrating the conceptual understanding of the theory.

Rational: Writing SAS code to evaluate their data using multilevel samples; draw conclusions and interpret output; understand the theory behind the SAS process; are all concepts and skills recommended in Goal 8 (Association, 2014)

CONCLUSION

The last part of the project (Part 5) includes a review of the findings, as well as conclusions and suggestions in the sense of the problem. This includes developing a summary table to report results from three pilot study sampling methods: SRS; multilevel sampling with proportional allocation; and multilevel sampling with optimal allocation. The design effect (deaf) is a useful summary measure used to equate the effectiveness of sample estimation (sample mean in this case) measured by a complex sample design (in this case multilevel) with that specified by SRS with the same number of observations. Calculation of ratio variants under each place must be calculated in such a way

Students are expected to measure two design effects, one for each tiered allocation method, and then discuss which sampling method they would suggest to the survey designer. Students are also asked to propose alternative definitions for usable strata, to provide reasons for their proposals, and to comment on predictable design consequences.

Teacher's Note: Some students have difficulty summarizing their grades, which contributes to a good conversation about writing papers in general. Topics include analysis of measured design results. For alternative strata concepts, some students go beyond what is requested and do more research

REFERENCES

- Association, A. S. (2014). Curriculum guidelines for undergraduate programs in statistical science. *Retrieved December, 15, 2017.*
- Baglin, J., Bedford, A., & Bulmer, M. (2013). Students' experiences and perceptions of using a virtual environment for project-based assessment in an online introductory statistics course. *Technology Innovations in Statistics Education, 7*(2).
- Baglin, J., & Huynh, M. (2015). Using the island to teach statistics through data investigations: A pilot project in Australian secondary schools. *Proceedings of The Australian Conference on Science and Mathematics Education (Formerly UniServe Science Conference).*
- Baglin, J., Reece, J., & Baker, J. (2015). Virtualising the quantitative research methods course: An island-based approach. *Statistics Education Research Journal, 14*(2).
- Baglin, J., Reece, J., Bulmer, M., & Di Benedetto, M. (2013). Simulating the data investigative cycle in less than two hours: Using a virtual human population, cloud collaboration

- and a statistical package to engage students in a quantitative research methods course. *Proceedings of the Joint IASE/IAOS Satellite Conference, August*.
- Bulmer, M., & Haladyn, J. K. (2011). Life on an Island: A simulated population to support student projects in statistics. *Technology Innovations in Statistics Education*, 5(1).
- Fecso, R. S., Kalsbeek, W. D., Lohr, S. L., Scheaffer, R. L., Scheuren, F. J., & Stasny, E. A. (1996). Teaching survey sampling. *The American Statistician*, 50(4), 328–340.
- Horton, N. J., & Hardin, J. S. (2015). Teaching the next generation of statistics students to “think with data”: Special issue on statistics and the undergraduate curriculum. *The American Statistician*, 69(4), 259–265.
- Hulsizer, M. R., & Woolf, L. M. (2009). *A guide to teaching statistics: Innovations and best practices* (Vol. 10). John Wiley & Sons.
- Huynh, M., Baglin, J., & Bedford, A. (2014). Improving the attitudes of high school students towards statistics: An island-based approach. *Sustainability in Statistics Education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9), Flagstaff, Arizona, USA. Voorburg: International Association of Statistics Education*.
- Linden, M., Baglin, J., & Bedford, A. (2011). Teaching clinical trial design and management using an online virtual environment. *Proceedings of The Australian Conference on Science and Mathematics Education (Formerly UniServe Science Conference)*, 17.
- Lohr, S. L. (1999). *Sampling: Design and analysis*. Brooks. Cole Publishing Company. Pacific Grove, CA.
- Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education*, 93(3), 223–231.
- Tirlea, L., Baglin, J., Huynh, M., & Elphinstone, B. (2016). A pilot experimental study comparing different active learning strategies for understanding sampling. *Proceedings of the 9th Australian Conference on Teaching Statistics*, 27–31.