Module 5: Sampling

**Lesson 2: How to Sample**

**Estimated Time**: ~7.5 hours

**Concepts:** Questionnaire design; bias; error; probability samples; data documentation

**Lesson Description**: This lesson introduces students to the main sampling strategies, as well as questionnaire design and sources of error in sampling and survey design.

**Instructor Preparation**: Look over the slides; familiarize yourself with the survey data used in data documentation exercise, complete assignment/go through solutions

|  |  |
| --- | --- |
| **Materials and Resources** | **Learning Goals** |
| 1. The slides 2. Core texts    1. Lohr, 2019, *Sampling Design and Analysis*, 2nd Edition, CRC Press.    2. Salganik, 2018, Bit by Bit: Social research in the Digital Age. 3. Optional: white board or iPad | * Design an effective questionnaire * Identify sources of error in sampling and survey design * Distinguish between different sampling methodologies and identify their strengths and weaknesses |

|  |  |  |
| --- | --- | --- |
| **Time** | **Lesson Content** | **Instructor Notes** |
| 10 minutes  30 minutes | **04-Errors**  **Introduction**   1. Review previous day’s concepts   **Lesson**   1. [04-Errors-slides](https://github.com/UofT-DSI/sampling/blob/main/lectures/04-Errors-slides.pptx)    1. Section on selection bias (Slides 9 to 12) will require reference to Lohr’s *Sampling Design and Analysis,* Section 1.5 |  |
| 10 minutes  40 minutes  40 minutes | **05-Simple probability samples**  **Introduction**   1. Watch this video from YouTube channel Dr. Nic’s Math and Stats: <https://www.youtube.com/watch?v=be9e-Q-jC-0>   **Lesson**   1. [05-Simple probability samples-slides](https://github.com/UofT-DSI/sampling/blob/main/lectures/05-Simple%20probability%20samples-slides.pptx)   **Exercises**   1. **Lohr Chapter 2, Exercise 33** Have students use amazon.com to conduct the following exercise. This can be done in groups or individually. This will also require students to self-teach confidence intervals for SRS (Lohr section 2.5) 2. In the books search window, type in a genre you like, such as mystery or sports; you may want to narrow your search by selecting a subcategory since an upper bound is placed on the number of books that can be displayed. Choose a genre with at least 20 pages of listings. The list of books forms your population. 3. What is your target population? What is the population size, *N*? 4. Take an SRS of 50 books from your population. Describe how you selected the SRS, and record the amount of time you spent taking the sample and collecting the data. 5. Record the following information for each book in your SRS: price, number of pages, and whether the book is paperback or hardback. 6. Give a point estimate and a 95% CI for the mean price of books in the genre you selected. 7. Give a point estimate and a 95% CI for the mean number of pages for books in the genre you selected. 8. Interpret your confidence intervals. | It may be useful to copy and paste these instructions into a document to display for students, or direct them to reference the exercise in the textbook.    Ensure students record their work/results, as they will be referenced again for the following two exercises. |
| 40 minutes  30 minutes | **06-Stratified sampling**  **Lesson**   1. [06-Stratified sampling-slides](https://github.com/UofT-DSI/sampling/blob/main/lectures/06-Stratified%20sampling-slides.pptx)   **Exercises**   1. **Lohr Chapter 3, Exercise 36** This is a continuation/extension of the previous SRS exercise. Instruct students to use the same book genre for this problem. 2. Stratify the population into two categories: hardcover and paperback. You can obtain the population counts in the paperback category by refining your search to include the word paperback. 3. Take a stratified random sample of 40 books from your population using proportional allocation. Record the price and number of pages for each book. 4. Give a point estimate and a 95% CI for the mean price of books and the mean number of pages for books in the population. 5. Compare your CI’s to those from the previous exercise where you used an SRS. Does stratification appear to increase the precision of your estimate? 6. BONUS: Use your SRS from Chapter 2 to estimate the within-stratum variance of book price for each stratum. In this case, you are using the SRS as a pilot sample to help design a subsequent sample. Find the optimal allocation for a stratified random sample of 40 books. How does the optimal allocation differ from the proportional allocation? | It may be useful to copy and paste these instructions into a document to display for students, or direct them to reference the exercise in the textbook. |
| 60 minutes  30 minutes | **07-Cluster sampling**  **Lesson**   1. [07-Cluster sampling-slides](https://github.com/UofT-DSI/sampling/blob/main/lectures/07-Cluster%20sampling-slides.pptx)   **Exercises**   1. **Lohr Chapter 5, Exercise 35** You may have noticed in the previous two exercises that it took quite a bit of time to locate the records chosen for the SRS. It may be faster to take an SRS of pages from the website, then look at some or all of the books listed on that page. Use the following procedure to take a cluster sample of books from the genre you studied in the section about SRS, recording the amount of time you spend selecting the sample and collecting the data. Take an SRS of 10 pages, then sample 5 books per page. For each sampled book, record the price, number of pages, and whether the book is paperback or hardback. Estimate the mean of each variable, and give a 95% CI. Do you think clustering decreased precision relative to an SRS? Compare the precision per unit time for the SRS and the cluster sample by calculating 1/[(estimated variance) × time] for each method. |  |
| 60 minutes | **Assessment**   1. [ASSIGNMENT-Data Documentation Comparison Worksheet](https://github.com/UofT-DSI/sampling/blob/main/assessment/ASSIGNMENT%20-%20Data%20Documentation%20Comparison%20Worksheet.md)    1. Students are intended to find the documentation on their own, however the links for instructor use are as follows:       1. 2018 GSS GVP: <https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&amp;SDDS=4430>       2. 2019 CES Online: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DUS88V>       3. “Tropic niche flexibility…”: https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2435.12192 | This assignment is intended to introduce students to finding and working with documentation for real-life survey data sets and identify features of sampling design that have been taught in lessons 1 and 2. It will also be considered in lesson 13-Reproducibility when data documentation is taught explicitly. |