Module 5: Sampling

**Lesson 4: Advanced Topics**

**Estimated Time**: 5 hours

**Concepts:** snowball sampling; hidden populations; sample size estimation; multi-phase sampling; multiple imputation; random number generators; reproducibility.

**Lesson Description**: This lesson explores more advanced topics in sampling and data analysis. We explore different variations of multi-stage sampling, ways to make our

**Instructor Preparation**: Read over slides and Salganik and Heckathorn (2004)

|  |  |
| --- | --- |
| **Materials and Resources** | **Learning Goals** |
| 1. Slides    1. [11-Respondent-driven sampling-slides](https://github.com/UofT-DSI/sampling/blob/main/lectures/11-Respondent-driven%20sampling-slides.pptx)    2. [12-Advanced data collection-slides](https://github.com/UofT-DSI/sampling/blob/main/lectures/12-Advanced%20data%20collection-slides.pptx)    3. [13-Reproducibility-slides](https://github.com/UofT-DSI/sampling/blob/main/lectures/13-Reproducibility-slides.pptx) 2. Key texts    1. Salganik, M. J., & Heckathorn, D. D. (2004). 5. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. Sociological Methodology, 34(1), 193–240. https://doi.org/10.1111/j.0081-1750.2004.00152.x    2. Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. DOI 10.18637/jss.v045.i03. | * Calculate asymptotically unbiased estimates of populations proportions through respondent-driven sampling * Write reproducible sampling simulations in Python and R |

|  |  |  |
| --- | --- | --- |
| **Time** | **Lesson Content** | **Instructor Notes** |
| 20 minutes  60 minutes | **11-Respondent-driven sampling**  **Introduction**   1. Review previous day’s concepts 2. Make sure students have access to Salganik and Heckathorn (2004). Give them 10 minutes to skim over the article.    1. <https://journals.sagepub.com/doi/10.1111/j.0081-1750.2004.00152.x>   **Lesson**   1. [11-Respondent-driven sampling-slide](https://github.com/UofT-DSI/sampling/blob/main/lectures/11-Respondent-driven%20sampling-slides.pptx) | These slides involve walking through long calculations which could be improved by following along on a whiteboard or iPad. |
| 60 minutes  30 minutes | **12-Advanced data collection**  **Lesson**   1. [12-Advanced data collection-slides](https://github.com/UofT-DSI/sampling/blob/main/lectures/12-Advanced%20data%20collection-slides.pptx) 2. *DSI-Sampling-Multiple Imputation Exercises*   **Exercises**   1. Have students work through their own multiple imputation workflow with another data set. Additional data can be found here under “Datasets”: <https://amices.org/mice/reference/index.html>. For Python many exist in sklearn.datasets, seaborn, statsmodels.api, etc. | Prompt students to pull up Python or R documentation for functions and follow along with slides, identifying any additional arguments of interest.    Go through slides and then demo Python or R code, or have students follow along |
| 60 minutes  30 minutes  15 minutes | **13-Reproducibility**  **Assessment**   1. ASSESSMENT-Sampling and Reproducibility.md    1. Blogpost: https://andrewwhitby.com/2020/11/24/contact-tracing-biased/   **Lesson**   1. [13-Reproducibility-slides](https://github.com/UofT-DSI/sampling/blob/main/lectures/13-Reproducibility-slides.pptx)   **Discussion**   1. Reflect upon the data documentation assignment from Lesson 2. Where do you think the articles considered are lacking in terms of their documentation and reproducibility? What would need to change or improve for the data collection process to be reproducible? | Instructions are in the md file. Have students read blogpost and work through whitby\_covid\_tracing (Python or R file). Assignment may take more or less time depending on familiarity with Python and R.      Might be good to demo set.seed in R in addition to static output provided on slides (or the Python Equivalent) |