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://docs.google.com/presentation/d/1ICRU32ho6NAiQwjITGOi5yXsjg0dHOkhVTRaqc7fcXY/edit#slide=id.g10eebaa84e3_0_0)    2. [12-Advanced data collection-slides](https://docs.google.com/presentation/d/11p9UXAR5ge6Muvkl3Gt0FOVeGaTqFoLMg-jd6zjjNR8/edit#slide=id.g109f9c406b8_0_0)    3. [13-Reproducibility-slides](https://docs.google.com/presentation/d/1ahaMw3izwqz-Kbd_nCBUEgAtOBVhe0t8KPEUAqEYuwg/edit#slide=id.g10a1f0bc2cf_0_57) 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 * Conduct multiple imputation using the mice() R package * Write reproducible sampling simulations in 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://docs.google.com/presentation/d/1ICRU32ho6NAiQwjITGOi5yXsjg0dHOkhVTRaqc7fcXY/edit#slide=id.g10eebaa84e3_0_0) | 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://docs.google.com/presentation/d/11p9UXAR5ge6Muvkl3Gt0FOVeGaTqFoLMg-jd6zjjNR8/edit#slide=id.g109f9c406b8_0_0) 2. *DSI-Sampling-Multiple Imputation Exercises.R*   **Exercises**   1. Have students work through their own multiple imputation workflow with another data set from the mice() package. Additional data can be found here under “Datasets”: https://amices.org/mice/reference/index.html | Prompt students to pull up R documentation for functions and follow along with slides, identifying any additional arguments of interest.    Go through slides and then demo R code, or have students follow along |
| 60 minutes  30 minutes  15 minutes | **13-Reproducibility**  **Assessment**   1. ASSESSMENT-Sampling and Reproducibility in R.Rmd    1. Blogpost: https://andrewwhitby.com/2020/11/24/contact-tracing-biased/   **Lesson**   1. [13-Reproducibility-slides](https://docs.google.com/presentation/d/1ahaMw3izwqz-Kbd_nCBUEgAtOBVhe0t8KPEUAqEYuwg/edit#slide=id.g10a1f0bc2cf_0_57)   **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 Rmd file. Have students read blogpost and work through whitby\_covid\_tracing.R. Assignment may take more or less time depending on familiarity with R.      Might be good to demo set.seed in R in addition to static output provided on slides |