Sampling: Introduction

```
$ echo "Data Science Institute"
```

Learning Outcomes

- Ability to implement simple probability samples.
- Ability to understand more complicated sampling procedures and the tradeoffs involved.
- Ability to identify and understand sources of error or inaccuracies in data as a result of sampling strategies.
- Development of intuition around survey quality.

Foundations of Probability (level: Beginner)

- *How do we calculate and interpret probabilities? What is a statistical distribution?*
  - Foundations
  - Distributions
  - Random variables
- Reference: Pitman, 1993, *Probability* , Springer, Chapters 1-3

Populations, Censuses, Surveys, and Observational data (level: Beginner)

- *Who are you intending to study? Who is receiving your survey or being observed? How will this impact the resulting data and analysis?*
    - Defining a target population and what are the (statistical) units
    - Introduce representative and non-representative sampling and illustrate with examples
    - Differences between censuses, surveys and observational studies with Canadian applications
- Reference: Wu and Thompson, 2020, *Sampling Theory and Practice* , Springer., Chapter 1

Essentials of Sampling, Asking, and Observing (level: Beginner)

- *What makes a good sample or study? How does sampling in theory differ from sampling in practice?*
  - Requirements of a good sample
  - Observational studies and sampling
  - Probability sampling: theory vs. practice
  - Questionnaire design
- Reference: Lohr, 2019, *Sampling Design and Analysis* , 2nd Edition, CRC Press., Chapter 1; ** Salganik, 2018, *Bit by Bit: Social research in the Digital Age* . Princeton University Press., **Chapter 3

Errors (level: Intermediate)

- *How might your sampling and surveying approach cause inaccuracies in your data?*
  - Sampling and nonsampling errors
  - Selection bias
  - Total survey error
- Reference: Lohr, 2019, *Sampling Design and Analysis* , 2nd Edition, CRC Press., Chapter 1; ** Salganik, 2018, *Bit by Bit: Social research in the Digital Age* . Princeton University Press., **Chapter 3

Simple Probability Samples (level: Intermediate)

- *How might we select and study random individuals from a population? How do we effectively study a sample selected in this manner?*
  - Simple random sampling
  - Weights
  - Systematic sampling
- Reference: Lohr, 2019, *Sampling Design and Analysis* , 2nd Edition, CRC Press., Chapter 2

Stratified Sampling (level: Intermediate)

- *How might our study be impacted if we divide our population into groups by shared characteristics before sampling? How do we effectively study a sample selected in this manner?*
    - Introductory concepts
    - Weights, again
    - Defining strata
    - Using quotas
- Reference: Lohr, 2019, *Sampling Design and Analysis* , 2nd Edition, CRC Press., Chapter 3

Cluster Sampling (level: Intermediate)

- *How might our study be impacted if we sample entire groups of individuals from our population based on shared characteristics? How do we effectively study a sample selected in this manner?*
    - Introductory concepts
    - One-stage clusters
    - Two-stage clusters
- Reference: Lohr, 2019, *Sampling Design and Analysis* , 2nd Edition, CRC Press., Chapter 5

Non-Response (level: Intermediate)

- *Why do some individuals not respond to surveys? How can we encourage people to respond consistently to surveys when sampled? What can be done when non-response is unavoidable?*
    - Introductory concepts
    - Designing to reduce non-response.
    - Dealing with non-response.
- Reference: Lohr, 2019, *Sampling Design and Analysis* , 2nd Edition, CRC Press., Chapter 8

Estimation and Survey Quality (level: Intermediate)

- *How can we tell if our survey is high quality? What are some potential inaccuracies in data resulting from surveys, and what causes them?*
  - Measures of quality
  - Dealing with various errors including coverage, non-response, measurement, processing, etc
  - Total Survey Quality.
- Reference: Lohr, 2019, *Sampling Design and Analysis* , 2nd Edition, CRC Press., Chapter 15

Differential Privacy (level: Advanced)

- *How can probability be used to create privacy and anonymity in large data sets?*
  - Informational risk and anonymization
  - Basics of differential privacy
  - Implementation
  - Practical and ethical considerations
- Reference: Wood, Altman, Bembenek, Bun, Gaboardi, Honaker, Nissim, OBrien, Steinke & Vadhan, 2018, Differential privacy: A primer for a non-technical audience. *Vanderbilt Journal of Entertainment & Technology Law, * 21(1) 209-275.

Additional Topics (level: Advanced)

- Reproducibility
- Sampling and seeds
- Data documentation
- Ethics
- Respondent burden
- External validity
- Inequity
- Collecting and using data about race and ethnicity