Sampling: Estimation and Survey Quality

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

Learning Outcomes

- *How can we tell if our survey is high quality? What are some potential inaccuracies in data resulting from surveys, and what causes them?*
    i. Define and identify sources of measurement, coverage, and processing error
    ii. Assess survey quality using the concept of total survey error

Measures of Quality

- Depending on the purpose of the survey, there are many different ways to define and assess quality:
  - Relevance – results must meet an identified need
  - Accuracy – Estimates must be close to the true population quantities
  - Timeliness – Results must be available and distributed quickly
  - Accessibility – Data and results must be accessible to users in an interpretable fashion

Measures of Quality

- Cont.
  - **Comparability** – Surveys that are intended for comparison with other surveys (i.e. over time or between regions) must be designed and conducted such that estimate comparisons are meaningful
  - **Coherence** – Common definitions and standards should be used when data comes from different sources
  - **Completeness** – Statistics should be available for all areas of identified need

⚠️*Improving quality means maximizing measures of quality while minimizing errors*⚠️

Coverage Error

Coverage Error

- **Coverage error** occurs when the sampling frame does not match the target population
    - Undercoverage is most common, i.e. some of the population is missing from the sampling frame
- **Coverage bias** occurs when coverage error causes sample estimates to differ from the population value
    - Bias is low when coverage error is low or when the means of the covered and uncovered populations are similar

Measuring Coverage

- It is difficult to measure coverage, since if missing populations were easily identifiable and accessible they would already be included in the frame.
- Ways to assess coverage:
  - Compare estimates of demographic characteristics to known values from the population
    - For example, if your frame contains 75% men and 25% women, it is likely that women are undercovered
  - Compare coverage rate or estimates with an external study or data source
    - For example, coverage of households with infants could be assessed by comparing the sampling frame with recent birth records in the area of interest

Coverage and Survey Mode

- Survey mode refers to the medium in which the survey is conducted – online, telephone, mail, etc.
- Geography-based frames tend to have the highest coverage, but are also the most expensive (may require in-person interview)
- Different modes have different coverage issues

Coverage and Survey Mode

- **Mail or email**
  - Sampling frame is generally a list of addresses – coverage depends on accuracy and completeness of the list
    - People may have moved or changed email addresses
    - Excludes people who don’t use email
- **Telephone**
  - Sampling frame is either a list of phone numbers from a directory or random digit dialing
    - Coverage is difficult to track for cellphone-only households, which tend to differ from those with landlines

Coverage and Survey Mode

- Internet
  - Difficult to specify a frame and measure coverage of specific populations
  - Generally don’t use probability sampling – coverage is unknown since participants tend to be volunteers

Exercise: What kind of survey would you use to sample students from the 3 main universities in the Greater Toronto Area?

Improving Coverage

- Remove duplicates from the sampling frame
- Compare sampling frame with external data sources to check for missing units or subpopulations
- Choose a survey mode or modes that have high coverage for your target population
- Combine multiple frames to form one sample
    - For example, combining an SRS from a frame of landline numbers with an SRS from a frame of cellphone numbers

Measurement Error

Measurement Error

- **Measurement error** occurs when a value reported by a respondent differs from the true value.
- Potential sources: confusion about the question meaning, calculation error, inaccurate estimation, omission of information
- Some metrics have a true underlying value (i.e. height, weight, income) while others do not (i.e. opinion, confidence level, mood)
    - It is usually impossible to address measurement error in the context of underlying states of being

Reducing Measurement Error

- First, need to identify sources and prevalence of measurement error
  - Can be identified through randomized experiments addressing different components of survey design (question, interviewers, etc.)
- Some strategies for reducing measurement error (depending on type of survey):
  - Write clear questions
  - Test questions prior to releasing the survey
  - Write clear procedures for administering the survey
  - Hire good surveyors or interviewers
  - Provide consistent training and supervision for interviewers
  - Give surveyors a reasonable workload

Sensitive Questions and Measurement Error

- Respondents may not respond accurately to questions about sensitive topics
  - For example, income, drug or alcohol use, crime or victimization, voting
- Survey mode can influence measurement error for sensitive questions
  - Respondents are more likely to respond honestly for self-administered surveys
- Potential improvement for interview-based surveys: **randomized response**
  - Works best for “Yes” and “No” questions
  - Respondents are randomly asked either a neutral question or a sensitive question (without the interviewer’s knowledge)
  - If the probability of being asked each question and the probability of responding “Yes” or “No” to the neutral question is known, we can estimate the proportion of “Yes” or “No” responses to the sensitive question

Processing Error

Processing Error

- **Processing error** is any error that occurs due to data entry or editing.
- Potential sources:
  - Data entry clerk transcribing the wrong response into a database
  - Open ended questions – a respondent may give multiple responses which must be classified/coded as a single response
  - Flawed imputation
- Some processing error can be reduced through editing and cleaning once data is collected.

Total Survey Quality

Total Survey Quality

> Total Survey Error = Sampling Error + Nonresponse Error + Coverage Error + Measurement Error + Processing Error

- Higher total survey quality means **minimizing total survey error** – most easily improved at the design stage
- Once released, the quality of results should be communicated to users. This may include:
  - Type of sample
  - Sources of sampling and nonsampling error
  - Total sample size
  - Nonresponse rates

Next

Reproducibility