Stat472/572 Sampling: Theory and Practice

Slides are intended for a course based on the book: *Sampling: Design and Analysis, Third Edition* by Sharon L. Lohr

Instructor: Yan Lu

University of New Mexico

December 27, 2024

Chapter 1: Introduction

Three Elements of Statistical Study

- Collecting Data: observational data, experimental data, survey data
- Describing and Presenting Data: graphical and numerical description, point estimation and interval estimation
- Drawing Conclusions from Data

Statistical Inference

- We have limited visibility into the population, but we are interested in understanding the typical characteristics of the population.
  - Population parameter: A numerical value that represents a characteristic of the entire population, such as the population mean $\bar{y}_U$ or population total t.
- The only information we currently possess is based on the sample.
  - Statistics derived from samples are used to estimate population parameters.

Goal: Our objective is to utilize the information obtained from the sample to draw inferences about the population and its parameters.

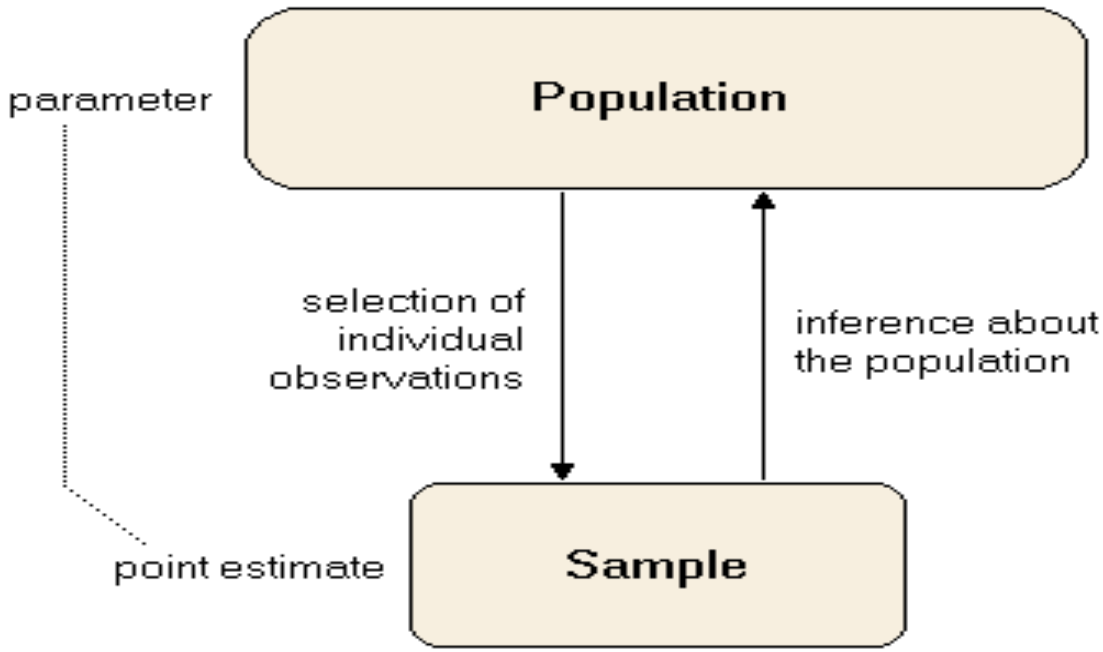

Figure 1: Population, sample and statistical inference

Survey Sampling

We aim to utilize sample information to make inferences about a finite population.

- The remaining statistics, $Y_1, Y_2, \cdots, Y_n$, are random variables following a certain distribution, such as the normal distribution $N(u, \sigma^2)$. The observed values of these random variables are $y_1, y_2, \cdots, y_n$.
- In general probability sampling (design-based analysis), $Y_1, Y_2, \cdots, Y_N$ represent the population. We select a sample of n units, denoted as $y_1, y_2, \cdots, y_n$, based on a predetermined design that assigns a probability of selection to each possible subset of the population with size n. Neither $Y_1, Y_2, \cdots, Y_N$ nor $y_1, y_2, \cdots, y_n$ are random variables. The random variables are Z_i's, defined as:

$$Z_i = \begin{cases} 1 & \text{if unit} \quad i \in S \\ 0 & \text{otherwise} \end{cases}$$

Some definitions:

- Observation Units: Objects on which measurements are taken, sometimes referred to as elements.
- Target Population: The complete collection of observations that we aim to study.
  — Defining the target population is an important and often challenging aspect of the study.
  — For instance, in a political poll, should the target population encompass all eligible adults to vote, all registered voters, or all individuals who participated in the last election? The choice of target population significantly influences the resulting statistics.
- Sampled Population: The population from which the sample was drawn.
  Note: In an ideal survey, the sampled population would be identical to the target population, although this ideal scenario is rarely achieved precisely.

- Sample: A subset of a population.
- Sampling unit: The unit that is actually sampled.
  Example: When studying individuals and lacking a comprehensive list of all individuals in the target population, households may serve as the sampling units, with individuals living in those households as the observation units.
- Sampling frame: The list of sampling units.
  Example: In telephone surveys, the sampling frame could be a list of all residential telephone numbers in the city. In personal interviews, the sampling frame might consist of a list of all street addresses.
- Census: When data is collected for every unit of the population, it is referred to as a census.

Example: Telephone survey of likely voters
Target population: All likely voters
Sampling frame: A list of telephone numbers
— (1) Not all households have telephones
— (2) Some individuals with phones are not registered to vote and are therefore ineligible for the survey
— (3) Certain eligible individuals with phones may be unreachable, refuse to respond, or incapable of responding
Sampling unit: A phone number
Observation unit: An individual associated with the phone number

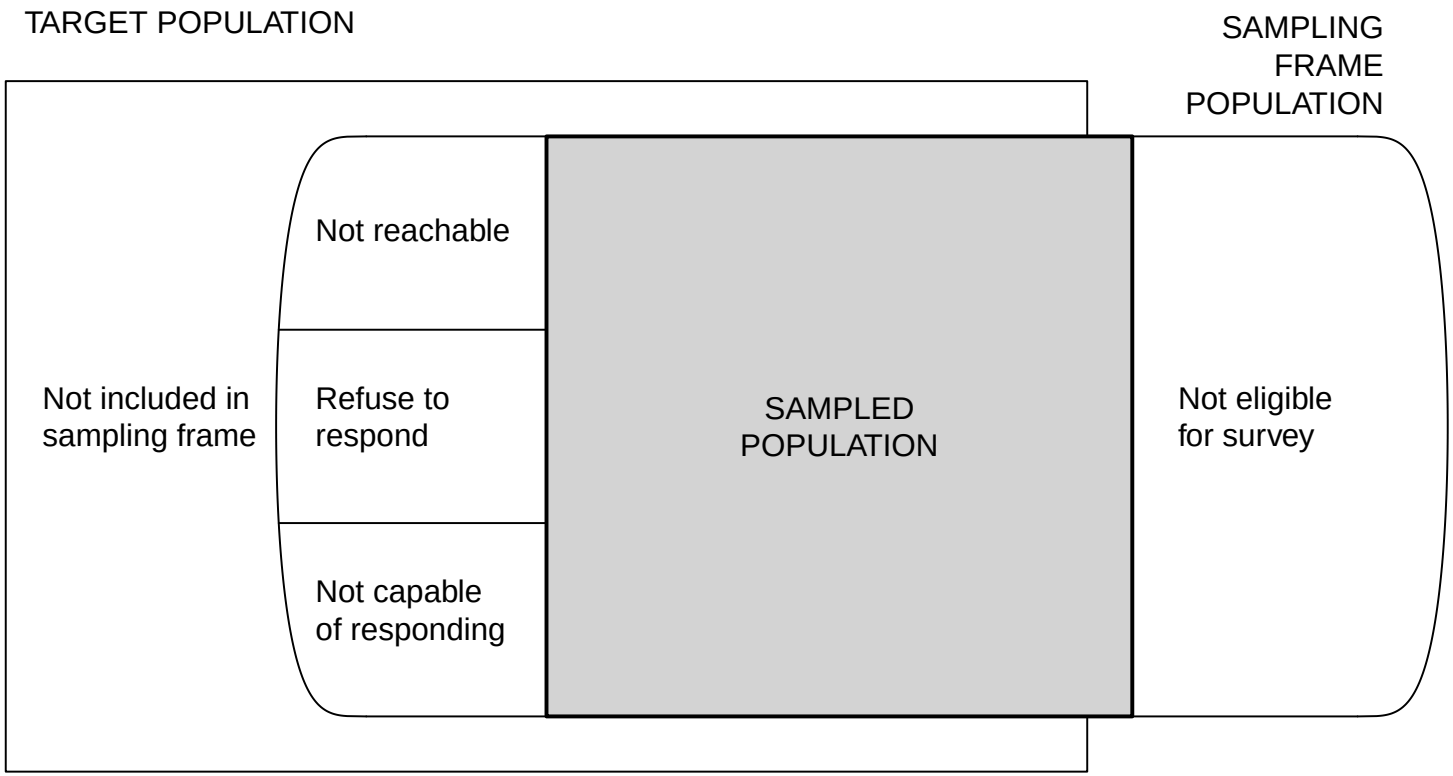

Figure 2: Telephone survey of likely voters

Source: Figure 1.1 of Sampling: Design and Analysis, 3rd edition, by Sharon L. Lohr

Why Sampling?

- Cost: A census is expensive.
- Time: Conducting a census is very time-consuming.
- Impractical: In some applications, a census can be impractical. Example: The government requires automakers who want to sell cars in the U.S. to demonstrate that their cars can survive certain crash tests. Obviously, the company can't be expected to crash every car to see if it survives! So the company crashes only a sample of cars.

Types of Samples:

1. Non-probability (non-random) samples: These samples focus on volunteers, easily available units, or those that happen to be present when the research is conducted. Non-probability samples are useful for quick and cost-effective studies, case studies, qualitative research, pilot studies, and developing hypotheses for future research. However, non-probability samples often introduce bias.

- Convenience sample: Also known as an "accidental" sample or "man-in-the-street" sample. The researcher selects units that are convenient, easily accessible, or nearby.
- Purposive sample: The researcher selects units with a specific purpose in mind, such as students who live in dorms on campus or females.
- Quota sample: The researcher constructs quotas for different types of units. For example, interviewing a fixed number of shoppers at a mall, with half being male and half being female.

2. Probability-based (random) samples: These samples are based on probability theory. Every unit of the population of interest must be identified, and all units must have a known, non-zero chance of being selected into the sample.

- Simple random sample (SRS): Randomly select a sample of size n from a population of size N.
  a) The sampling unit and observation unit are the same.
  b) Each subset of size n has an equal probability of being selected as the sample.
  c) Each unit has an equal chance of being selected in the sample.

  — Random number generators

  — Lottery method

- Systematic random sampling: Randomly select the first item or subject from the population and then select every *n*th subject from the list.
  - The results are representative of the population unless certain characteristics of the population are repeated for every *n*th individual, which is highly unlikely.
  - Systematic sampling is useful for selecting large samples, such as 100 or more. It is less cumbersome than a simple random sample using either a table of random numbers or the lottery method.
  - If the selection interval matches some pattern in the list (e.g., the list is male, female, male, female, $\cdots$), and you select the No.1, No.3, No.5, $\cdots$ observations to form a systematic sample, you will introduce systematic bias into your sample.

- Stratified random sampling: Divide population into H strata, take an SRS of size n_h from stratum h, $h = 1, \cdots, H$, select the sample independently.
  Example: You want to find out the attitudes of students on your campus about immigration.
  —— 27,000 students: 22,000 West; 3,000 East; 1000 Midwest; 600 South; 400 Foreign.
  ——Select a simple random sample of 1500 students, you might not get any from the Midwest, South, or Foreign.
  ——Divide the students into these five groups (Stratum), and then select the same percentage of students from each group using a simple random sampling method. This is proportional stratified random sampling.
  ——Divide students into the five groups and then select the same number of students from each group using a simple random sampling method. This is disproportionate stratified random sampling.

- Cluster sampling: A cluster is a naturally-occurring grouping of the members of the population. For example, city residents are also residents of neighborhoods, blocks, and housing structures.
  Randomly select n clusters, then observe all the elements in the selected clusters or partial of the elements in the selected clusters.

Example: To obtain information about the drug habits of all high school students in New Mexico.
—-Obtain a list of all the high schools in NM
—-Select an SRS of high schools
—-Within each selected high school, list all classes, and select an SRS of classes
—-The students in the selected classes are the observations in your sample

Biases

- Selection Bias: If some part of the target population is not in the sampled population, a bias called Selection Bias occurs.
  Examples:
  —- In a survey to estimate per capita income, if transient people are ignored.
  —Mis-specification of the target population
  —-Failure to include all the target population in the sampling frame, also called undercoverage
  —-Substituting a convenient member of a population for a designated member not readily available

—-Non Response: failure to obtain responses from those chosen in the sample

—-Allowing a sample to consist entirely of volunteers (Radio, TV, or call-in polls)

Note that large samples are generally considered good but if the sample is unrepresentative, it can be quite bad. The design of the survey is far more important than the absolute size of the sample.

- Measurement Bias: Measurement bias occurs when the measuring instrument has a tendency to record in one direction more often than the other. Measurement biases are more common when dealing with people.
  —-People may not tell the truth
  —-Lack of understanding of questions
  —-Lack of proper account of events in memory
  —-Variations in responses due to interviewer
  —-Misreading questions, or miss recording responses
  —-Desire to impress the interviewer
  —-Ordering and wording of questions have effects on responses
  Many of these problems can be avoided by proper questionnaire design

Questionnaire Design:

- Decide what you want to find out; this is the most important step in writing a questionnaire
- Pilot study: Test questions before sending out the questionnaire.
- Keep the questions *Simple* and *Clear*: Questions should be neither too lengthy nor too technical. They should be easily understood by non experts
- Questions should be specific and not general
- Decide whether to use open or closed questions.
  —-Open Question: The respondent is not prompted with categories for responses. It allows responses to form their own response categories.
  ——Closed Question: A question is closed when specific response categories are provided.
  —— Closed questions with well thought and researched categories elicit more accurate responses.

- Avoid questions that prompt or motivate the respondent to say what investigator wants to hear
- Use choices rather than Agree/Disagree type questions
- Ask only one concept in one question
- Pay attention to question-order effect. Ask general questions first then follow with specific questions.

Sampling and Non-Sampling Errors:

- Sampling Errors: Sampling errors are results of inherent variability in the sampling process. These arise because the results vary from sample to sample. Margin of errors reported are a result of sampling error. These can only be reduced by increasing the sample size but not be eliminated.
- Non-Sampling Errors: These result from selection bias, measurement error and inaccuracies of responses. These can not be attributed to sample-to-sample variability. Such errors can be eliminated by proper precautions. Selection bias can be reduced by using probability sample. Accurate responses can be achieved through proper and careful design of survey instrument and training of interviewers.