Statistics 5210 Sample Survey and Causal Inference
Chapter 1: Introduction
Part 1: Survey Sampling

Zhengyuan Zhu [1]

Iowa State University

January 21, 2025

[1] Based on notes by Emily Berg and Jae-Kwang Kim

1. Introduction
2. Sampling Terminology
3. Example
4. Probability sampling
5. Survey Process
6. Survey Errors

What is sample survey?

- **Definition**: The gathering of information about a population by studying a sample of that population in such a way that the results for the sample can be extrapolated to that population.
- Population: the entire group of individuals, objects, or events that a researcher wants to draw conclusions about
  - Finite Population: all voters in an election, students in a school, etc.
  - Infinite Population: all possible location to take a soil sample, etc.
- Sample : Subset of a population
- Sampling : the process of selecting a sample to make statistical inference about the (finite) population without measuring the whole population
- Sampling error : the error that results from taking a sample instead of examining the whole population

Examples of Surveys

Surveys are everywhere in the modern society and can vary greatly along several dimensions:

- the **scope of objectives**,
- the **target population of interest**,
- **data collection methods**

The Scope of Survey Objectives

Objectives: What characteristics of the population are of interest?

- Some surveys are multi-purpose: analyst is interested in many characteristics of the population
    - National Health and Nutrition Examination Survey (NHANES):
        - Health and nutrition of adults and children in the U.S.
        - Demographic, socioeconomic, dietary, and health-related questions.
- Other surveys focus on a relatively small number of objectives
    - Annual Faculty Activity Survey (FAS): estimates the average hours per week that ISU faculty work. (see handout for questionnaire)

Target Populations

Target population, The extent of the population of interest, needs to be defined precisely.

- Some surveys study relatively broad domains
    - Current Population Survey: “National and state estimates of labor force characteristics of the civilian noninstitutionalized population 16 years of age and older” (census.gov/programs-surveys/cps/about.html)
- Other surveys are more narrow
    - National Agricultural Workers Survey restricts to the agricultural sector of the labor force.

Data Collection Methods

- Interviewer-mediated (in-person, telephone)
- Self-administered (mail, web)
- Field observation (data observed at sample sites – eg., land cover, soil sample, water quality, traffic)
- A single survey may use multiple data collection modes
  - In-person with computer-assisted self-administered component for sensitive questions
  - NHANES – interview + physical examination
  - Fisheries – aerial photography for effort (i.e., time spent fishing) + dockside monitors for catch (i.e., weight of fish caught)

Surveys vs. Other Forms of Data Collection

How do surveys compare to other types of statistical studies?

- Experimental studies: full control over factors and population (Analysis of variance, experimental design)
- Observational studies: observe part of the population without any control
- Survey sampling: observe a carefully selected part of the population without controlling any other factors

Why sampling ?

1. To reduce the cost
2. To save the time
3. Sometimes, to get a more accurate information about the population.
4. Sometimes, it is the only way of getting information about the target population.

How to do the sampling ?

- Two types of sampling
  - Probability sampling
  - Non-probability sampling
- Roughly speaking, a probability rule is assigned to obtain a sample in probability sampling.

Discussion

- Give me an example of a survey that you have encountered recently. What were the objectives, population of interest, and data collection methods?
- If you could conduct any survey of your choosing, what would be the topic? What would be the objectives, population of interest, and data collection methods?

1 Introduction

2 Sampling Terminology

3 Example

4 Probability sampling

5 Survey Process

6 Survey Errors

Defining Sampling Terms

- Element: object on which measurement is taken
  - Person, household, area of land
- Population: Collection of elements about which we wish to make inference
  - ISU faculty, all households in Ames, all farms in Iowa
- Sampling Unit: non-overlapping collection of elements from the population that cover the entire population
  - Person, city block (containing households), a rectangular segment of land
- Frame: list of sampling units from which the sample is selected
  - Email addresses of ISU facutly; list of city blocks; map divided into rectangular land area segments
  - Other relatively common examples in household surveys: telephone numbers, addresses
- Sample: collection of sampling units drawn from a frame

Defining Sampling Terms: Further Discussion

Sampling unit is not always equal to the population element

- Example:Elements = elementary school children
    - Frame at level 1: list of schools
        - Select a sample of schools; sampling unit is school
    - Frame at level 2: list of classes within each school
        - Select a sample of classes from each school; sampling unit is a class
    - Frame at level 3: list of children within each class
        - Sampling unit = element = child
- More feasible to obtain a list of schools than a list of children
- This is an example of a multi-stage design (will cover this later).

- Direct element sampling – sample selection from a frame that directly identifies the individual elements of the population of interest (We will talk about this first).

Defining Sampling Terms: Further Discussion

Frame does not always match the target population

- Undercoverage – eligible elements are excluded from the frame
    - Can lead to bias if elements excluded differ systematically from population with respect to characteristics of interest
    - Example: Want to estimate the total number of Iowa residents who read the Des Moines register last week
        - Frame of telephone numbers with Iowa area code
        - Frame excludes Iowa residents with cell phones number from other states
- Overcoverage – frame contains ineligible units
    - Inefficient – need to screen ineligible units out of the sample
    - Example: Target population – farmers in Michigan who grow specialty crops (fruits, vegetables)
        - Frame: list of all farmers from the Agricultural Census and administrative data
        - Screening procedures intent to exclude farmers who do not grow specialty crops
        - Some interviewers do not properly implement the screening procedures, causing some farmers who do not grow specialty crops to be included

Defining Sampling Terms: Further Discussion

Area frames vs. List frames

- Area Frame – A geographic frame consisting of area units; every population element belongs to an area unit, and it can be identified after inspection of the area unit. The area units may vary in size and in the number of elements that they contain.
    - Examples: city map, forest map, aerial photograph
- List Frame – List of sampling units in the population
    - Examples: telephone numbers, postal addresses, email addresses, social security numbers, business identification numbers

Defining Sampling Terms: Further Discussion

Auxiliary inforamtion

- Frames can contain useful **auxiliary information** about population elements that can be used in sample design and estimation
  - Examples of auxiliary variables for area frames: location, elevation, average temperature, average precipitation
  - Examples of auxiliary variables for list frames: number of employees in a business, age of an individual, number of children enrolled in a school

Defining Sampling Terms: Further Discussion

Population is **finite**

- Finite population – In survey sampling, the population is a finite list of N elements
  - Universe: $U = \{1, \ldots, N\}$
  - Examples: finite number of ISU faculty in 2024; finite number of farms in Iowa; finite number of businesses in Canada in 2006.
- This differs fundamentally from other courses involving model based inference, where the population is typically infinite
  - Ex: $X_i \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2) : i = 1, 2, \ldots$

Discussion

- Consider the faculty activity survey. What is the...
  - Element?
  - Population?
  - Sampling unit?
  - Frame?

1. Introduction
2. Sampling Terminology
3. Example
4. Probability sampling
5. Survey Process
6. Survey Errors

Township Example

- Consider an artificial example of a township with a population of four farms (N = 4)

ID	Size of farms (Acres)	Corn Acreage (y)
1	4	1
2	6	3
3	6	5
4	20	15

- Farm acreage is known in advance for all 4 farms in the population (i.e., available on the frame)
- Corn acreage is not known
- Parameter of interest: average acres of corn per farm in the town

$$\theta = (y_1 + y_2 + y_3 + y_4)/4$$

- We can only afford to sample 2 farms, and the value of y is obtained only from the sampled units.

- 6 possible samples

case	sample ID	sample mean	sampling error
1	1, 2	2	
2	1, 3	3	
3	1, 4	8	
4	2, 3	4	
5	2, 4	9	
6	3, 4	10	

- Each sample has a sampling error.
- Two ways of selecting one of the six possible samples.
  - Nonprobability sampling : (using size of farms or etc.) select a sample subjectively.
  - Probability sampling : select a sample by a probability rule.

Probability sampling

- Design A: Simple Random Sampling : Assign the same selection probability to all possible samples

case	sample ID	sample mean ($\bar{y}$)	sampling error	selection probability
1	1, 2	2	-4	$1/6$
2	1, 3	3	-3	$1/6$
3	1, 4	8	2	$1/6$
4	2, 3	4	-2	$1/6$
5	2, 4	9	3	$1/6$
6	3, 4	10	4	$1/6$

- In this case, the sample mean($\bar{y}$) has a discrete probability distribution.

- Probability mass function of $\bar{y}$:

$$P_{\bar{y}}(y) = \begin{cases} 1/6 & \text{if } y \in \{2, 3, 4, 8, 9, 10\} \\ 0 & \text{otherwise.} \end{cases}$$

- Is it unbiased?

- Variance

Remark

- No model assumption about y_i in the example: totally different framework !
- Design-based approach: the reference distribution is the sampling distribution generated by the repeated application of the given sampling mechanism.

1 Introduction

2 Sampling Terminology

3 Example

4 Probability sampling

5 Survey Process

6 Survey Errors

Probability Sampling: Definition & Notation

- $U = \{1, 2, \cdots, N\}$: index set of finite population
- A : subset of U, index set of the sample.
- $\mathcal{A}$: set of samples under consideration, sample support.
- $\hat{\theta} = \hat{\theta}(y_i; i \in A)$: statistic

Probability Sampling: Definition & Notation

- Probability distribution of samples, or sample distribution: probability mass function $P\left(\cdot\right)$ defined on $\mathcal{A}$. That is, $P\left(\cdot\right)$ satisfies
  1. $P\left(A\right) \in \left[0,1\right], \qquad \forall A \in \mathcal{A}$
  2. $\sum_{A\in\mathcal{A}} P\left(A\right) = 1.$
- (Induced) probability distribution of a statistic
  - Expectation : $E(\hat{\theta}) = \sum_{A\in\mathcal{A}} P(A)\hat{\theta}\left(A\right)$
  - Variance : $Var(\hat{\theta}) = \sum_{A\in\mathcal{A}} P(A)\left[\hat{\theta}\left(A\right) - E(\hat{\theta})\right]^2$
  - Mean squared error :

$$\begin{aligned} MSE(\hat{\theta}) &= \sum_{A\in\mathcal{A}} P(A)\left[\hat{\theta}\left(A\right) - \theta\right]^2 \\ &= Var\left(\hat{\theta}\right) + \left[E(\hat{\theta}) - \theta\right]^2 \end{aligned}$$

  - $MSE = Variance + (Bias)^2$

Probability Sampling

- Definition : For each element in the population, the probability that the element is included in the sample is known and greater than 0.
- Advantages
  1. Exclude subjectivity of selecting samples.
  2. Remove sampling bias (or selection bias)
- What is sampling bias ? (θ : true value, $\hat{\theta}$: estimated value of θ)

$$
\begin{aligned}
\text{(sampling) error of } \hat{\theta} &= \hat{\theta} - \theta \\
&= \left\{\hat{\theta} - E\left(\hat{\theta}\right)\right\} + \left\{E\left(\hat{\theta}\right) - \theta\right\} \\
&= \text{variation} + \text{bias}
\end{aligned}
$$

- In nonprobability sampling, variation is 0 but there is a bias. In probability sampling, there exists variation but bias is 0.

Probability Sampling

- Main theory
  1. Law of Large Numbers : $\hat{\theta}$ converges to $E(\hat{\theta})$ for sufficiently large sample size.
  2. Central Limit Theorem : $\hat{\theta}$ follows a normal distribution for sufficiently large sample size.
- Additional advantages of probability sampling with large sample :
  1. Improve the precision of an estimator
  2. Can compute confidence intervals or test statistical hypotheses.
- With the same sample size, we may have different precision.

Example - Continued : Probability sampling 2

- Design B: use auxiliary info on farm size in design: every sample has to contain farm 4, with the remaining farms selected with equal probability.

Sample ID	y value	Mean Estimator	Selection probability
1, 4	1, 15	4.5	$1/3$
2, 4	3, 15	6	$1/3$
3, 4	5, 15	7.5	$1/3$

- What is the probability distribution of the mean estimator ?
- What is the expected value of the mean estimator ?

$$E(\bar{y}) = \frac{1}{3}(4.5 + 6 + 7.5) = 6.0$$

- Compute the variance. Compare it with that of SRS.

1. Introduction
2. Sampling Terminology
3. Example
4. Probability sampling
5. Survey Process
6. Survey Errors

Outline of a survey process

1. Define Target population: this is the population to which the conclusions apply.
2. Determine population characteristics of interest (e.g. acres of corn, daily intake of calcium)
3. Find sampling frame: device that associates elements by sampling units (e.g. phone book)
4. Obtain a sample by a probability sampling design.
5. Measure the study variables.
6. Use measured values to compute point estimates and standard errors.

Basic procedures for survey sampling

1. Planning
    1. Statement of objectives
    2. Selection of a sampling frame
2. Design and development
    1. Sample design
    2. Questionnaire design
3. Implementation
    1. Data collection
    2. Data capture and coding
    3. Editing and Imputation
    4. Estimation
    5. Data analysis
    6. Data dissemination
4. Evaluation - Documentation

Two aspects of Survey Design

1. Sampling design: how to collect a sample ?
    - What is your target population ?
    - What is your sampling frame ?
    - What information is available from the sampling frame ?
    - What is the cost for observing each unit in the sample ?
2. Questionnaire design: how to obtain measurement from the selected sample ?
    - What is your research questions ?
    - What are the variables that you want to measure from the sample ?
    - How to ask the questions properly ?
    - Is there any other auxiliary variables that you want to measure to use in the weighting or editing stage ?

1. Introduction
2. Sampling Terminology
3. Example
4. Probability sampling
5. Survey Process
6. Survey Errors

Source of Errors

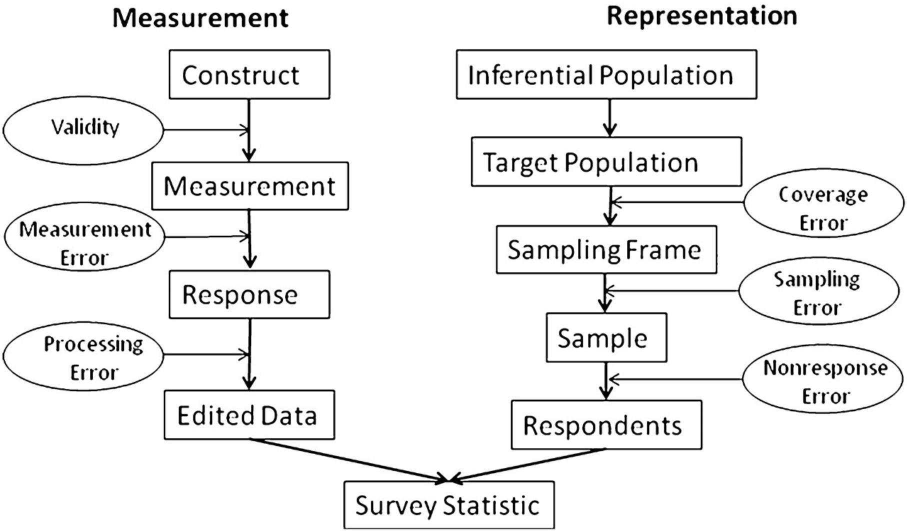
1. Errors of nonobservation
    - Coverage error
        - (Target) population $\neq$ Frame
        - Some elements are not listed.
    - Sampling error
        - Frame $\neq$ sample
        - Some listed elements not sampled.
    - Non-response error
        - sample $\neq$ respondents
        - Some sampled elements don't respond.
2. Errors of observation
    - Measurement error: Interviewer, respondent, instrument, mode
    - Processing error

Source of Errors

Consider the faculty activity survey

- What are two sources of potential measurement error?
- What are two sources of potential selection error?

Total survey errors



- Sampling error:
  - If $n = N$, no sampling error
  - In fact, if $n \uparrow$ then sampling error $\downarrow$.
- Nonsampling error: Everything else
  - Even if $n = N$, you have nonsampling error.
  - In practice, we can decrease nonsampling error by decreasing n.
- Because of nonsampling error, a sample is often more accurate than a Census.

Survey Methodology

- Psychology (Cognitive science), social science
- More interested in non-sampling errors
- By properly asking questions, we can reduce survey errors.
- Questionnaire design, survey management

Survey Statistics

- Statistics
- More interested in sampling errors
- Want to measure the uncertainty of survey errors and incorporate them into estimation.
- Sampling design, estimation, editing etc.

Summary

- The survey sample part of the course is mostly about the sampling error of $\hat{\theta}$
  1. What is it ?
  2. How to reduce it ?
  3. How to measure it ?