Sampling Theory and Practice

STAT 454[1]

Winter 2022 (1221)[2]

Cameron Roopnarine[3]　　Changbao Wu[4]

7th January 2022

[1] STAT 454$\equiv$ STAT 854

[2] Online Course until January 27th, 2022

[3] LaTeXer

[4] Instructor

Contents

1 Review of Basic Concepts in Survey Sampling 2

Chapter 1

Review of Basic Concepts in Survey Sampling

Week 1
5th to 7th January

Survey sampling as a scientific discipline:

- Started from Jerzy Neyman's 1934 paper (1894–1981).
- Fast development since the 1940s and 1950s.
- Became an important area of statistics and social science.
- (Used to be) the primary tool of data collection for official statistics and researchers in social science and health studies.
- Face challenges in the big data and internet era.

Some ongoing well-known surveys:

- The Current Population Survey of the US (CPS).
- The National Health and Nutrition Examination Survey of the US (NHANES).
- The General Social Survey of Canada (GSS).
- The Canadian Community Health Survey (CCHS).
- The International Tobacco Control Policy Evaluation Surveys (The ITC Surveys, headquartered at UWaterloo).
- The Canadian Longitudinal Study of Aging (CLSA, McMaster, McGill, and Dalhousie).

Statistics Canada:

- One of the most respected survey organizations in the world.

Some Canadian survey statisticians:

- J.N.K. Rao (Carleton University, retired).
- David Bellhouse (University of Western Ontario, retired).
- Jiahua Chen (University of British Columbia).
- David Haziza (University of Ottawa).

- Carl E. Särndal (University of Montreal, retired).
- Louis-Paul Rivest (Laval University).
- David Binder (Statistics Canada, 1949–2012).
- Carl Schwarz (Simon Fraser University, retired).
- Steve Thompson (Simon Fraser University).
- Randy Sitter (Simon Fraser University, 1961–2007).
- V. P. Godambe (University of Waterloo, 1926–2016).
- Mary E. Thompson (University of Waterloo, retired).
- Matthias Schonlau (University of Waterloo).
- Changbao Wu (University of Waterloo).

Example 1.1. The Math Faculty plans to conduct a survey to study the well-being of recent graduates from the faculty.

- What is exactly the group to be studied?
  (The target population)
- What information is to be collected?
  (Variables to be measured; sample data)
- From what can we select individuals to be surveyed?
  (Sampling frame(s))
- How to select individuals to be surveyed?
  (Sampling methods; sampling procedures)
- What method to use to collect data?
  (The mode of data collection: Face-to-face? Telephone? Mailed questionnaire?)
- How to use the data to draw conclusions?
  (Statistical analysis)

Three versions of survey populations (with reference to Example 1.1):

- *The target population*: The set of all units covered by the main objective of the study.
  (All students who received a formal degree from Waterloo between 2016 and 2019)
- *The frame population*: The set of all units covered by the sampling frame(s).
  (Sampling frame: The list of personal email addresses of students who graduated between 2016 and 2019)
- *The sampled population* (*the study population*): The population represented by the sample. Under probability sampling, the sampled population is the set of all units which have a non-zero probability to be selected in the sample.
  - The sampled population is not the set of sampled units!
  - Units which cannot be reached or do not respond to surveys (non-response) are not part of the sampled population.

Population structures and sampling frames:

$$U = \{1, 2, \ldots, N\},$$

where N is the population size, and the labels $1, 2, \ldots, N$ represent the N units.

- **Unstructured population**: There exists a single complete list of all N units, which can be used as the sampling frame.

- **Stratified population**: The population U has a stratified structure if it is divided into H non-overlapping subpopulations:
$$U = U_1 \cup U_2 \cup \cdots \cup U_H,$$
where the subpopulation U_h is called stratum h, with stratum population size N_h, $h = 1, 2, \ldots, H$. It follows that
$$N = \sum_{h=1}^{H} N_h.$$
Sampling frames for stratified sampling: H separate lists, each list consists of all units in one stratum.

- **Clustered population**: If the survey population can be divided into groups, called *clusters*, such that every unit in the population belongs to one and only one group, we say the population is clustered.

  First stage sampling frame for cluster sampling: A complete list of clusters (but not all the units within each cluster).

- Stratified sampling versus cluster sampling:
  - Under stratified sampling, sample data are collected from every stratum.
  - Under cluster sampling, only a portion of the clusters has members in the final sample.

Example 1.2. Survey of the population of high school students in the Waterloo region. There are a total of 15 high schools. Take a sample of 300 students from the population.

- **Plan A**. Randomly select 20 students from each high school. (Stratified sampling)
- **Plan B**. Randomly select 5 high schools from the list of 15 schools, and then randomly select 60 students from each of the 5 selected schools. (Two-stage cluster sampling)
- **Plan C**. The Waterloo region can be divided into KW area (8 high schools) and non-KW area (7 high schools). First, randomly select 3 schools from the KW area and 2 schools from the non-KW area, then randomly select 60 students from each of the 5 selected schools. (Stratified two-stage cluster sampling)

Sampling units and observational units:

- *Sampling units*: Units used to select the survey sample.
  - Under clustering sampling, sampling units are the clusters.
  - Under non-clustering sampling, sampling units are the individual units.
- *PSU and SSU*: Under two-stage cluster sampling, the first stage sampling units are clusters, called the *primary sampling unit* (PSU); the second stage sampling units are individual units, called the *secondary sampling unit* (SSU).
- *Observational units*: Observational units are always the individual units from which measurements are taken.

Example 1.3. An educational worker wanted to find out the average number of hours each week (of a certain month and year) spent on watching television by four and five-year-old children in the Waterloo Region. She conducted a survey using the list of 123 pre-school kindergartens administered by the Waterloo Region District School Board. She first randomly selected 10 kindergartens from the list. Within each selected kindergarten, she was able to obtain a complete list of all four and five-year-old children, with contact information for their parents/guardians. She then randomly selected 50 children from the list and mailed the survey questionnaire to their parents/guardians. The planned sample size is $10 \times 50 = 500$ and the sample data were compiled from those who completed and returned the questionnaires.

- *The target population*: All four and five-year-old children in the Region of Waterloo at the time of the survey. This is defined by the overall objective of the study.
- *Sampling frames*: Two-stage cluster sampling methods were used (further details to follow). The first stage sampling frame is the list of 123 kindergartens administered by the school board. The second stage sampling frames are the complete lists of all four and five-year-old children for the 10 selected kindergartens.
- *Sampling units and observational units*: The first stage sampling units are the kindergartens; the second stage sampling units are the individual children (or equivalently, their parents); observational units are individual children.
- *The frame population*: All four and five-year-old children who attend one of the 123 kindergartens in the Region of Waterloo. It is apparent that children who are homeschooled are not covered by the frame population. Thus, as is frequently the case, the frame population is not the same as the target population.
- *The sampled population*: All four and five-year-old children who attend one of the 123 kindergartens in the Region of Waterloo and whose parents/guardians would complete and return the survey questionnaire if the child was selected for the survey.