Sampling methods for social surveys

Gaia Bertarelli, Sant'Anna School of Advanced Studies (Italy)

January 2023

The presentation at a glance

Introduction

Survey Sampling

- Survey is conducted to measure the characteristic of a Population.
- Sampling is the process of selecting a subset of observations from an entire population of interest so that the characteristics from the subset (sample) can be used to draw conclusion or making inference about the entire population.
- Survey design + Sampling strategy = Survey sampling

Survey Sampling

- Survey Design:
  1. What survey design is appropriate for my study?
  2. How survey will be conducted/implemented?
- Sampling procedure:
  1. What sample size is needed for my study?
  2. How the design will affect the sample size?
- *Appropriate survey design provides the best estimation with high reliability at the lowest cost within the available resources*

Why do we conduct survey?

- Uniqueness: data not available from other sources.
- Standardized measurement: systematic collection of data in a structured format.
- Unbiased representativiness: selection of sample of probability distribution.
- Time: data can be collected more quickly, so estimates can be published in a timely fashion.

Types of survey design: cross-sectional and longitudinal

- Cross-Sectional survey: data are collected at a point of time.
- Longitudinal surveys:
  1. Trends: surveys of sample population at different points in time.
  2. Cohort: survey of the same population each time over a period.
  3. Panel: study of same sample of respondents at various time-points.

A sample controversy

A sample controversy

- Shere Hite's book *Women and Love: A Cultural Revolution in Progress (1987)*
  1. 84% of women are not satisfied emotionally with their relationships (p. 804).
  2. 70% of all women married five or more years are having sex outside of their marriages (p. 856).
  3. 95% of women report forms of emotional and psychological harassment from men with whom they are in love relationships (p. 810).
  4. 84% of women report forms of condescension from the men in their love relationships (p. 809).
- Hite erred in generalizing these results to all women, whether they participated in the survey or not, and in claiming that the percentages above applied to all women.

- The following characteristics of the survey make it unsuitable for generalizing the results to all women:
  1. The sample was self-selected: only 4.5% of questionnaires returned.
  2. The questionnaires were mailed to all-women groups.
  3. The survey has 127 essay questions, and most of the questions have several parts: Only very interested women fill out the questionnaire.
  4. Many of the questions are vague, e.g. using word such as love.
  5. Many of the questions are leading: the suggest to the respondent which response she should male.
- The final sample is not representative of women in the United States, and the statistics can only be used to describe woman who would have respondent to the survey.

- Hite claims that results from the sample could be generalized because characteristics such as the age, educational, and occupational profiles of women in the sample matched those for the population of women in the United States. But the women in the sample differed on one important aspect, they were willing to take the time to fill out a long questionnaire dealing with harassment by men, and to provide intensely personal information to a researcher. We would expect that in every age group and socioeconomic class, women who choose to report such information would in general have had different experiences than women who choose not to participate in the survey.

Glossary

Glossary

- **Representative Sample**: a sample is representative if the characteristics of interest in the population can be estimated from the sample with a known degree of accuracy.
- **Observation Unit**: an object on which a measurement is taken.
- **Target population**: the complete collection of observations we want to study.
- **Sample**: a subset of a population.
- **Sampled population**: The collection of all possible observation units that might have been chosen in a sample.

Glossary (ii)

- **Sampling Unit**: a unit that can be selected froma sample *(e.g. we may want to study individuals, but do not have a list of all individuals in the target population. Instead, households serve as the sampling units, and the observation units are the individuals living in the households.)*
- **Sampling frame**: a list of sampling unit in the population from which a sample may be selected *(e.g. Telephone survey and list of all residential telephone numbers; agricultural surveys and list of all farms or map of areas containing farms.)*

In an ideal survey, the sampled population will be identical to the target population, but this ideal is rarely met. In surveys of people the sampled population is usually smaller than the target population.

An example: telephone survey of likely voters in the USA

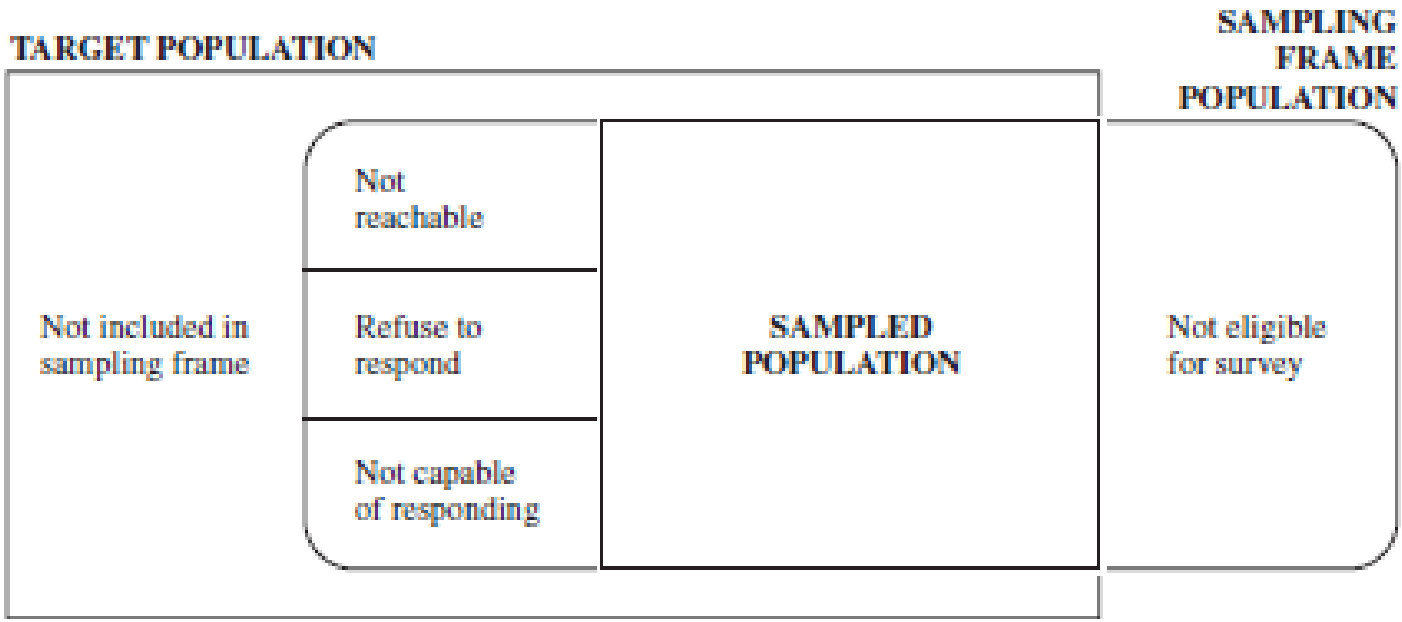


Figure 1: Not all voters have telephone. Some residents with telephone are not registered to vote. Some eligible resident do not respond.

Sampling and Nonsampling errors

Selection Bias

- Selection Bias occurs when some part of the target population is not in the sampled population: the sample is an **unrepresentative sample**.
- A **sample of convenience** is often biased: the units that are easiest to select or that are most likely to respond are usually non representative of the harder to select or nonresponding units.

Examples of selection bias

- Undercoverage is a bias that occurs when some members of your population aren't represented in a sample (e.g. telephone survey).
- Including population units in the sampling frame that are not in the target population is called Overcoverage bias. This primarily results from two situations:
  1. There are records in the sample frame that do not contain respondents or members of the target population.
  2. The same respondent is targeted by duplicate or multiple records in the sample frame.
- Overcoverage and undercoverage, essentially, depend on the mode(s) of contact and the access of a sample unit to the modes of administration.

Examples of selection bias (ii)

- Having multiplicity of listings in the sampling frame, without adjusting for the multiplicity in the analysis (sampling with unequal probability).
- Allowing the sample to consist only of volunteers. Some individuals may respond multiple times to a voluntary surveys, and a determined organization may skew the results.
- Nonresponse bias. Usually nonrespondents differ critically from the respondents but the reasons of that difference is unknow unless you can later obtain information about nonrespondents.

It is possible to reduce some forms of selection bias by using probability sampling methods. Accurate responses can often be achieved through careful design testing the survey instrument (e.g. training of interviewers) and pre-testing the survey.

Measurement Error

- Measurement error is the difference between the observed value of a variable and the true, but unobserved, value of that variable. It occurs when the response has a tendency to differ from the true value in one direction.
- Measurement errors sometimes are unavoidable (e.g. count of birds) but can be adjusted during the analysis.

Measurement Error (ii)

- Measurement error is surveys of people:
  1. People sometimes do not tell the truth.
  2. People do not always understan the questions.
  3. People forget (e.g. time).
  4. People can give different answers to different interviewers.
  5. People may say what they think an interviewer wants to hear.
  6. Interviewer may affect the accuracy of the response.
  7. Certain words mean different things to different people.

Sampling Errors

- Sampling Error: Sampling error is the error that arises in a data collection process as a result of taking a sample from a population rather than using the whole population.
- Sampling error is one of two reasons for the difference between an estimate of a population parameter and the true, but unknown, value of the population parameter. The other reason is nonsampling error.
- Even if a sampling process has no nonsampling errors then estimates from different random samples (of the same size) will vary from sample to sample, and each estimate is likely to be different from the true value of the population parameter.

Sampling Errors (ii)

- The sampling error for a given sample is unknown but when the sampling is random, for some estimates theoretical methods may be used to measure the extent of the variation caused by sampling error.
- Most surveys that you see report a margin of error: the margin of error given in a survey or a polls is an expression of sampling error, the error that results from taking one sample intead of examining the whole population.
- Sampling errors are usually reported in probabilistic terms.
- The larger the margin of error, the less confidence one should have that a poll result would reflect the result of a survey of the entire population.

Nonsampling errors

- Nonsampling error is the error that arises in a data collection process as a result of factors other than taking a sample.
- Nonsampling errors have the potential to cause bias in polls, surveys or samples.
- There are many different types of nonsampling errors.
- Selection bias and measurement error are examples of nonsampling errors, which are any errors that cannot be attributed to the sample-to-sample variability.

Types of Errors

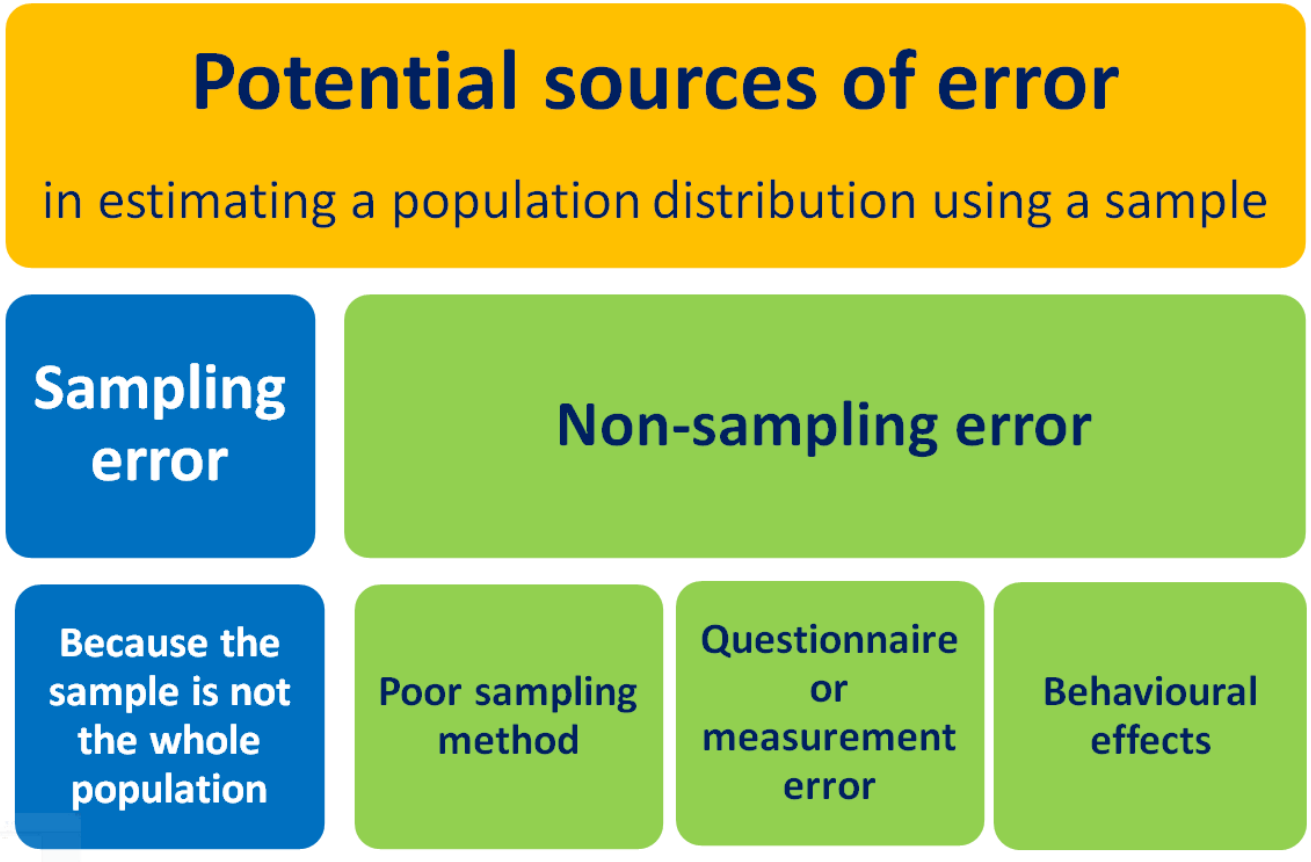
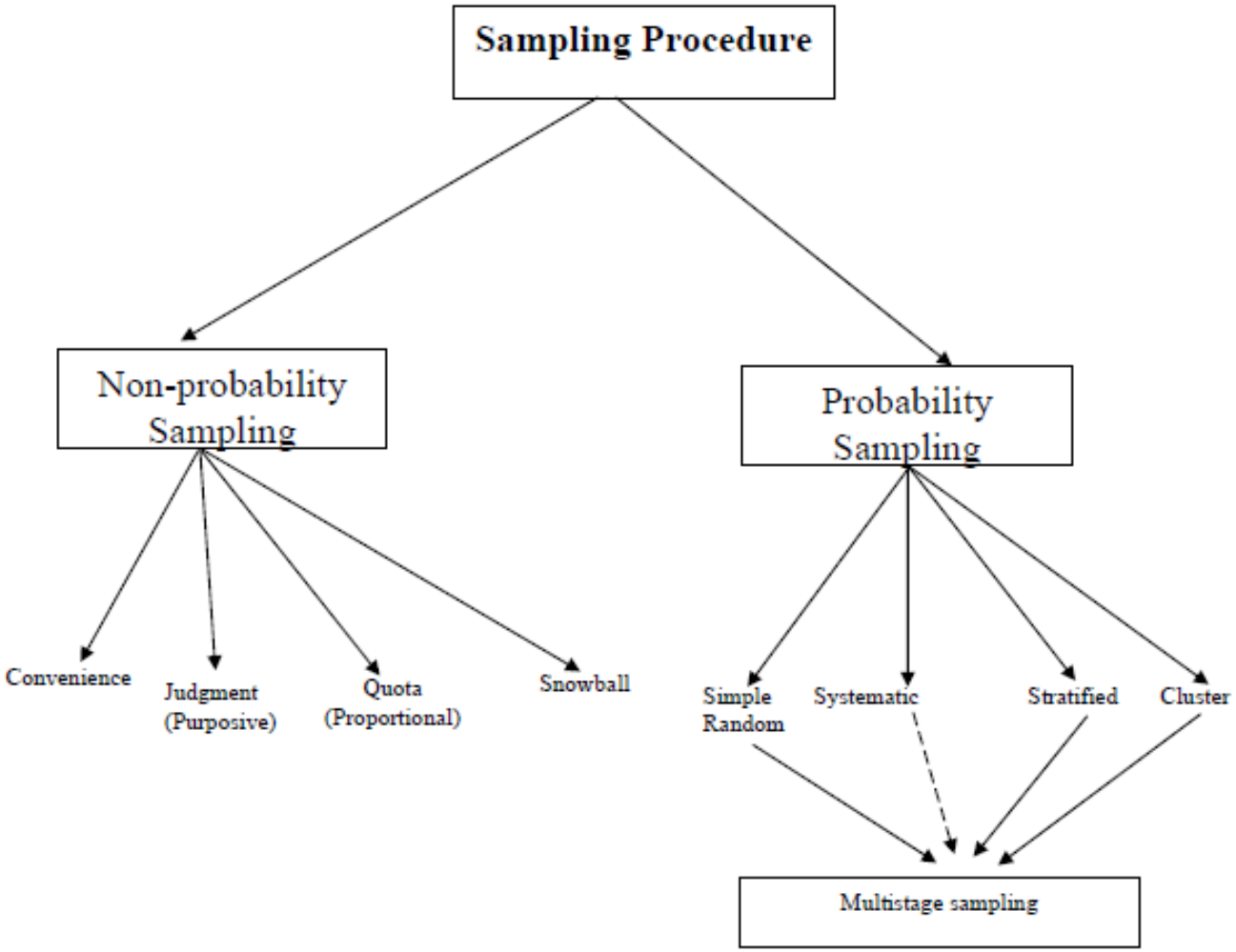


Figure 2: source: https://creativemaths.net/blog/sampling-and-non-sampling-error/

Simple Probability Samples

Types of Samples



Simple Probability Sample

- In a probability sample each unit in the population has a known probability of selection, and a random number table or a randomization mechanism is used to choose the specific units to be included in the sample.
- If a probability sampling design is implemented well, an investigator can use a relatively small sample to make inferences about an arbitrarily large population.

Basic types of Probability Samples - Simple Random Sampling

- Simple Random Sampling (SRS) is the simplest form of probability sample.
  - An SRS of size n is taken when every possible subset of n units in the population has the same probability of being part of the sample.
  - SRSs are at the basis of more complex designs.
  - In taking a random sample the investigator is mixing every member of the population before selecting n units.
  - The investigator does not need to examine every member of the population.

Stratified random sample

- In the Stratified random sample the population is divided into subgroups called strata.
- Then a SRS is selected from each stratum.
- The SRSs in the strata are selected independently.
  - The strata are often subgroups of interest to the investigator (to the survey).
  - Elements in the same stratum often tend to be more similar than randomly selected elements from the whole population.
  - As a consequence stratification increases precision.

Cluster sample

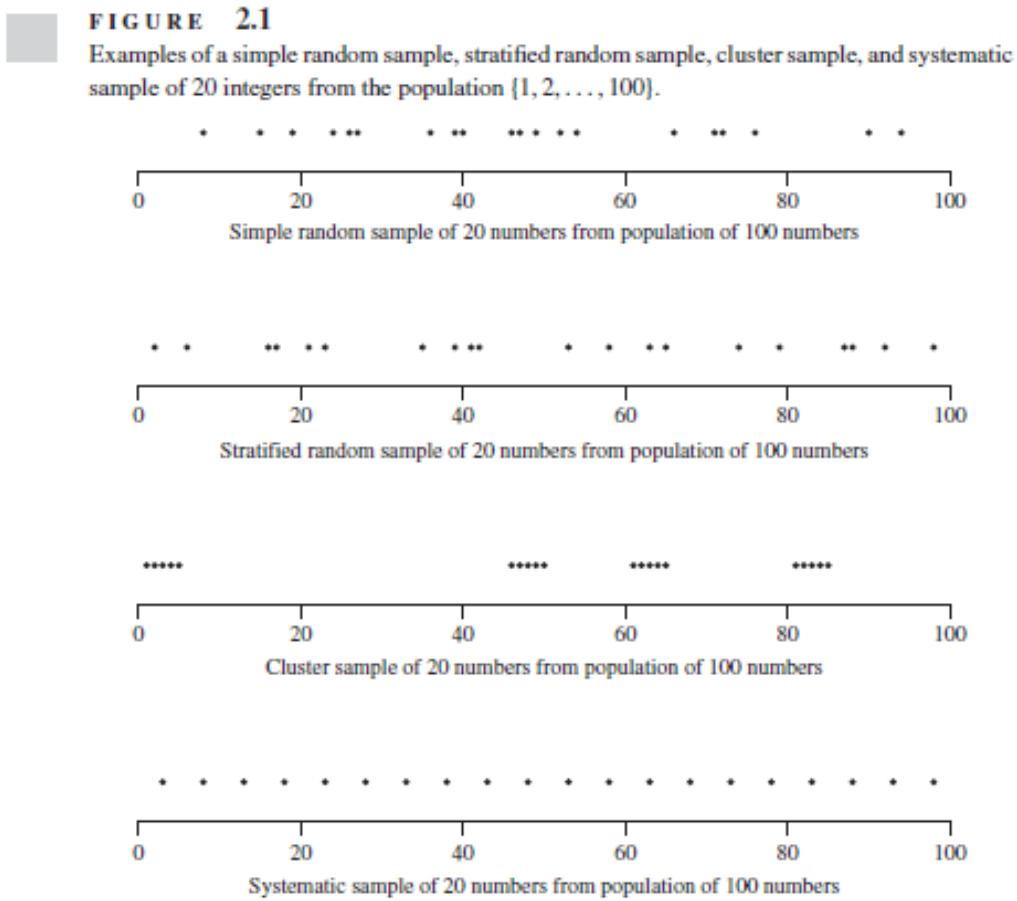
- In a Cluster sample observation units in the population are aggregated into larger sampling units called clusters.
- Cluster sampling is used when natural groups are present in a population.
- The investigator takes a SRS of the clusters (e.g. schools) and then subsample all or some members of the clusters. They then randomly select among these clusters to form a sample.
- Pay attention: In stratified sampling, individuals are randomly selected from all strata to make up the sample. On the other hand cluster sampling, the sample is formed when all individuals are taken from randomly selected clusters.

- In stratified sampling, there is homogeneity within the group, while in the case of cluster sampling the homogeneity is found between groups.
- Heterogeneity occurs between groups in stratified sampling. In contrast, group members are heterogeneous in cluster sampling.
- When the sampling method adopted by the researcher is stratified, then the categories are imposed by him. Otherwise, categories are groups that already exist in cluster sampling.
- Stratified sampling aims to improve accuracy and representation. Unlike cluster sampling, the goal of which is to improve cost effectiveness and operational efficiency.

Systematic sample

- In a Systematic sample a starting point is chosen from alist of population members using a random numeer.
- That unit, and every k-th unit thereafter, is chosen in the sample.
- A systematic sample thus consists of units that are equally spaced in the list.

Figure 3: Source: S. Lohr (2019). Sampling Design and Analysis. CRC press. (pg. 27)



Framework

- We need to be able to list the N units in the finite population.
- The finite population(Universe) of N units is denoted by the index set $U = \{1, 2, \ldots, N\}$
- Out of this population we can choose different samples, which are subsets of U
- Suppose $U = \{1, 2, 3, 4\}$ we can select 6 samples from this finite population $S_1 = \{1, 2\}$, $S_2 = \{1, 3\}$, $S_3 = \{1, 4\}$, $S_4 = \{2, 3\}$, $S_5 = \{2, 4\}$, $S_6 = \{3, 4\}$.

- Each possible sample S from the population has a know probability $P(S)$ of been selected (and the posisble probabilities sum to 1).
- In a probability sample, since each possible sample has a known probability of being the chosen sample, each unit in the population has a known probability of appearing in our selected sample.

$$\pi_i = P(\text{ unit } i \text{ in sample})$$

Gaia Bertarelli

Department of Economics and Management in the Era of Data Science

gaia.bertarelli@santannapisa.it