Applied Statistics

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

Statistics, Statistical Learning, Computing and Data Analytics
February, 15, 2022

The presentation at a glance

Introduction

A sample controversy

Questionnaire Design

Modes of Survey Administration

Glossary

Errors in Survey Sampling: Sampling and Nonsampling Errors

- Selection Bias - Coverage Errors
- Measurement Error
- Sampling and Nonsampling Errors

Simple Probability Sample

General Information about the course

- Aim: This part of the course *Applied Statistics* concentrates on the statistical aspects of taking and analyzing a sample.
- 8 hours.
- Final exam: Assignment on theoretical or practical study.
- Course materials:
  1. Slides
  2. Sharon L. Lohr (2022). *Sampling, Design and Analysis* - Second Edition - CRC Press Taylor & Francis Group. `https://www.sharonlohr.com/`
  3. Scientific articles reported during the course.
  4. Nicolini, G., Marasini, D., Montanari, G. E., Pratesi, M., Ranalli, M. G., & Rocco, E. (2013). Metodi di stima in presenza di errori non campionari. Springer Science & Business Media.

Introduction

Introduction to sampling from finite populations

- A 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

Introduction to sampling from finite populations

- *y* target variable
- *N* finite population size
- *census*: all the units in the population are observed and, in the hypothesis that no type of error occurs in the data collection, the real value of the parameter of interest T is determined.
- *sample survey*: only some of the population units are observed. It follows that, even in the absence of detection errors, the relative value of the parameter of interest T is unknown but can be estimated through the values observed for the units in the sample.

If the real value of the parameter T is unknown, the estimate must be *the best*. This raises the problem of the reliability of the estimate and of the choices of sampling techniques.

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.

Questionnaire Design

Questionnaire Design

- The most important step in writing a questionnaire is to decide what you want to find our: **write down the goals of your survey and be precise.**
- *Always test your questions before taking the survey*: test on a small sample, try different versions, ask for interpretations.
- *Keep it simple and clear*: questions that seem clear to you may not be clear to someone listening to the whole question over the telephone, or to a person with a different native language.
- *Define your term*
- *Use specific questions instead of general ones, if possible.*

Questionnaire Design (ii)

- *Decide wether to use oper or closed questions*:
  1. an **open question** allows respondents to form their own response categories;
  2. in a **closed question (multiple choice)** the respondent chooses from a set of categories read or displayed.
  - Each has advantages: a closed question may prompt the respondent to remember responses that might otherwise be forgotten, and is in accordance with the principle that specific questions are better that general one, but if you use a closed question, you'll always have an "other" category.
- *Avoid leading questions*.
- *Avoid double negative*.

Questionnaire Design (iii)

- *Consider the social desiriability of responses to questions, and write questions that elicit honest responses*.
- *Use forced-choice, rather than agree/disagree questions* (some persons will agree with almost any statement)
- *Ask only one concept per question*: In particula, avoid what are called **double-barreled** questions, so named because if one barrel of the shotgun does not get you, the other will.
- *Pay attention to question order effects*. Book example: ask if the respondent has read a book in the last 7 days and only then ask if he is interested in reading. The answers change following the order.

Modes of Survey Administration

Modes of Survey Administration

- Personal Interview
- Telephone
- Mail
- Computer-Assisted self-interviewing (CASI)
- Computer-Assisted personal-interview (CAPI)
- Computer-Assisted telephone-interview (CATI)
- Combination of methods

CAPI

- In the CAPI (Computer assisted personal interviewing) surveys, the interview is conducted Face 2 Face by interviewers who have a PC or tablet on which to attribute the answers.
- It is a method that allows direct contact with the interviewer, which helps in completing the questionnaire, both in terms of stimulus to completion and possible explanations, if necessary or requested by the respondent.
- The number of interviews that can be conducted daily is however more limited than oder methods, as well as the geographical area of reference. Furthermore, costs can also be more significant than other data collection.

CAPI (ii)

- The CAPI methodology is indicated for complex investigations, with long and structured questionnaires where the presence of an interviewer can explain the questions or how to fill in the questionnaire, or that appropriately give videos or stimuli of various kinds.
- The target audience for this survey may be those who are not easily reachable by telephone or with little computer experience, or when they intend to refer to a specifically localized population (eg: frequenters of shopping centres, areas of interest, points of aggregation, etc).

CATI

- In the CATI (Computer Assisted Telephone Interview) surveys, the interview is conducted via telephone by interviewers supported by software that allow data collection and scheduling of the interview according to the times and the respondent's availability.
- This method ensures a high quality of data collection, because the interviewer is guided by the software in the administration of the questions thus avoiding errors of interpretation. In addition, the software automatically reprograms the callback in case the respondent is not available.

CATI (ii)

- Being conducted by telephone, it is not possible to share stimuli, images or videos with the respondents. The methodology offers good value for money and an optimization of time.
- CATI is also suitable for geographically dispersed targets or for B2B targets as it is possible to plan the interview based on the interviewee's availability.

CAWI

- In the CAWI (Computer Aided Web Interviewing) surveys, a link is sent via email that the respondent proceeds to complete independently. It is possible to share visual and audio stimuli.
- This method allows reaching a large number of respondents in a very limited time interval. People are autonomous in completing the questionnaire and can do so at a time of their choice and take all the time they need to respond.
- It offers the best efficiency in terms of costs and timing. Any limits are related to the lack of comparison with the interviewer and the risk of abandonment by the respondent without having completed the interview.
- There are no geographical limits of targets, but those who are not very familiar with computers and technology are excluded.

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

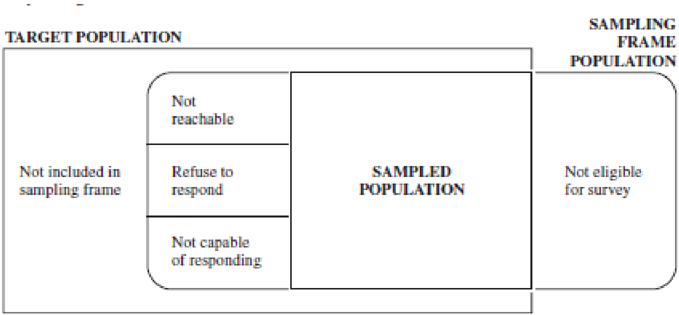


Figure 1: Not all voters have telephone. Some residents with telephone are not registered to vote. Some eligible resident do not respond.

Glossary (iii)

- **Population parameters**: Let y_j; $j = 1, \dots, N$; the value of the target variable Y on the unit j in the population. The population target parameter is a function of the N values in the population:

$$T = f(y_1, \dots, y_N)$$

Glossary (iv)- More parameters for the same variable

- Total:
$$Y = \sum_{j=1}^{N} y_j$$
- Mean:
$$\bar{Y} = \frac{Y}{N}$$
- Proportion:
$$P = \frac{A}{N}$$
- Quantiles:
$$Q_q = inf\{y = \Phi(y) \geq q\}$$
where $\Phi(y)$ is the distribution function of y, $0 < q < 1$
- Variance:
$$S_y^2 = \frac{1}{N-1} \sum_{j=1}^{N} (y_j - \bar{Y})^2$$

Glossary (v)- Parameters that highlight the relationship between variables

- Ratio estimator of total or mean

$$R = \frac{Y}{X} = \frac{\bar{X}}{\bar{Y}}$$

where $\sum_{j=1}^{N} x_j = X$

- Regression coefficient:

$$\beta_{yx} = \frac{S_{yx}}{S_x^2}$$

where

  - S_x^2 variance of x
  - Covariance of xy

$$S_{yx} = \frac{1}{N-1} \sum_{j=1}^{N} (y_j - \bar{Y})(x - \bar{X})$$

Glossary (vi)- Parameters that highlight the relationship between variables

- Linear correlation coefficient:

$$\rho_{yx} = \frac{S_{yx}}{S_y S_x}$$

linked to **Use of auxiliary variables in probabilistic sampling design**

Errors in Survey Sampling: Sampling and Nonsampling errors

Errors

- Errors of varying severity can occur during the different stages of an investigation.
- There are different types of errors but more than their type, we must pay attention to their extent.
- Statistical inference: Small errors are acceptable.

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

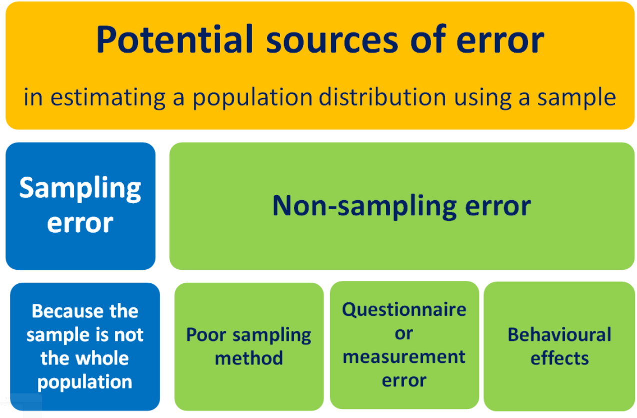

Potential sources of error
in estimating a population distribution using a sample
Sampling error
Non-sampling error
Because the sample is not the whole population
Poor sampling method
Questionnaire or measurement error
Behavioural effects

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

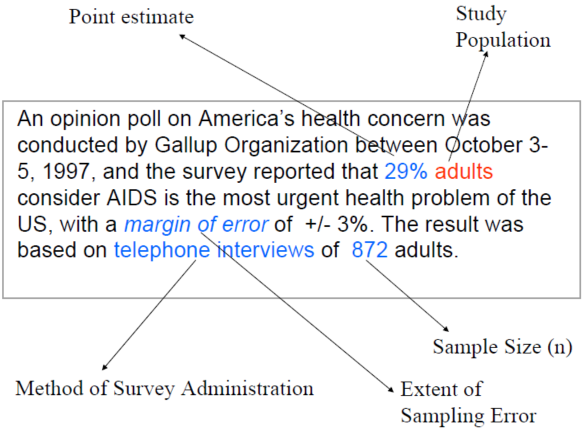
Recap

- **Census**: A survey in which the entire population is measured.
- **Coverage**: The percentage of the population of interest that is included in the sampling frame.
- **Measurement Error**: The difference between the response coded and the true value of the characteristic being studied for a respondent.
- **Nonresponse**: Failure of some units in the sample to provide responses to the survey.
- **Nonsampling error**: An error from any source other that sampling. Examples include nonresponse and measurement error.

Recap (ii)

- **Sampling error**: Error in estimation due to taking a sample instead of measuring every unit in the population.
- **Sampling frame**: A list, map, or other specification of units in the population from which a sample may be selected.
- **Selection bias**: Bias that occurs because the actual probabilities with which units are sampled differ from the selection probabilities specified by the investigator.

Recap (iii)

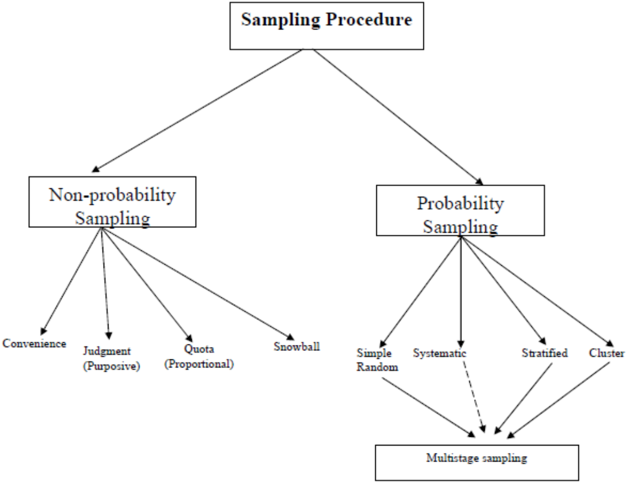


An opinion poll on America's health concern was conducted by Gallup Organization between October 3-5, 1997, and the survey reported that 29% adults consider AIDS is the most urgent health problem of the US, with a *margin of error* of +/- 3%. The result was based on telephone interviews of 872 adults.

Figure 3: source: S. Ahmed.

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 (General idea)

- 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 (General idea)

- 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 (General idea)

- 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.

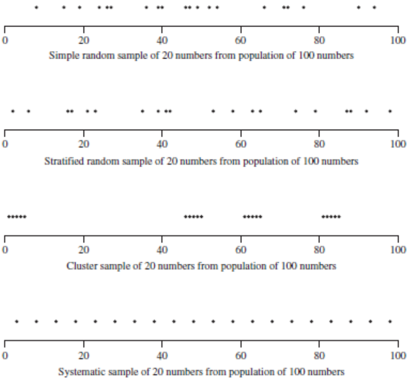
Systematic sample (General idea)

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

FIGURE 2.1

Examples of a simple random sample, stratified random sample, cluster sample, and systematic sample of 20 integers from the population {1, 2, . . . , 100}.



Simple random sample of 20 numbers from population of 100 numbers

Stratified random sample of 20 numbers from population of 100 numbers

Cluster sample of 20 numbers from population of 100 numbers

Systematic sample of 20 numbers from population of 100 numbers

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 possible probabilities sum to 1).
- The Sampling design is a function $p(s)$ defined on S_0 where:
  1. $p(s) \geq 0$
  2. $\sum_{s \in S_0} p(s) = 1$

  General advice: We are not interested in the order of extraction.

- 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

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

  is called First order inclusion probability and it is the probability of drawing a sample that includes the label i.

- The Second order inclusion probability is

  $$\pi_{i,i'} = \sum_{i,i' \in s} p(s)$$

  represents the probability of extracting a sample that contains the aforementioned pair i and $i^{'}$.

- A sampling design is said to be **probabilistic** if each label included in U has a positive and calculable probability of first order inclusion.
- A sampling design is said to be **self-weighting** if the first order inclusion probability is the same for each label.
- A sampling design is said to be **measurable** if the second order inclusion probabilities are all positive and calculable.
- If $\pi_i = 0$ for some i the sampling design is no more probabilistic because some units of the population of interest are not considered.

Gaia Bertarelli

Department of Economics and Management in the Era of Data Science

Department of Excellence 2018 - 2022

EMbeDS
Economics and Management
in the era of Data Science

ISTITUTO DI MANAGEMENT
Sant'Anna
Scuola Universitaria Superiore Pisa

management e sanità

gaia.bertarelli@santannapisa.it