Applied Statistics

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

Statistics, Statistical learning, Computing and Data Analytics

March, 1, 2022

The presentation at a glance

Non-probability sampling

- Focus on Quota sampling
- Focus on Snowball sampling

Non-response or Missing data

Non-probability sampling

- Non-probability techniques, relying on the judgment of the researcher or on accident, cannot generally be used to make generalizations about the whole population.
- Non-probability sampling represents a group of sampling techniques that help researchers to select units from a population that they are interested in studying.
- A core characteristic of non-probability sampling techniques is that samples are selected based on the **subjective judgement** of the researcher, rather than random selection (i.e., probabilistic methods)

Principles of non-probability sampling

- Non-probability sampling represents a valuable group of sampling techniques that can be used in research that follows
  - qualitative,
  - mixed methods,
  - and even quantitative research designs.
- Researchers following a quantitative research design often feel that they are forced to use non-probability sampling techniques because of some inability to use probability sampling (e.g., the lack of access to a list of the population being studied). However, this is not the case for researchers following a **qualitative research design**.

Theoretical reasons

- Unlike probability sampling, **the goal is not** to achieve objectivity in the selection of samples, or necessarily attempt **to make generalisations** (i.e., statistical inferences) from the sample being studied to the wider population of interest.
- Instead, researchers following a qualitative research design tend to be interested in the **intricacies** of the sample being studied. Whilst making generalisations from the sample to the population under study may be desirable, it is more often a secondary consideration.
- Even whether this is desired, there are additional problems of bias and transferability (or validity).

Practical reasons

- Non-probability sampling is often used because the procedures used to select units for inclusion in a sample are much easier, quicker and cheaper when compared with probability sampling.
- To sample hidden or hard-to-reach population where a list of the population simply does not exist.
- Non-probability sampling can also be particularly useful in exploratory research where the aim is to find out if a problem or issue even exists in a quick and inexpensive way.

Main types of non-probability sampling

- Quota sampling
- Convenience sampling
- Purposive sampling
- Self-selection sampling
- Snowball sampling

Basic ideas: Quota sampling

- Quota sampling is a sampling methodology wherein data is collected from a homogeneous group.
- It involves a two-step process where two variables can be used to filter information from the population.
- It can easily be administered and helps in quick comparison.
- Many persons confuse quota sampling with stratified sampling. In quota sampling, quota classes are formed that serve the role of strata, but **the survey taker uses a nonprobability sampling method such as convenience sampling to reach the desired sample size** in each quota class.

Basic ideas: convenience sampling

- A convenience sample is simply one where the units that are selected for inclusion in the sample are the **easiest to access**.

Basic ideas: purposive sampling

- Purposive sampling, also known as judgmental, selective, or subjective sampling, is a form of non-probability sampling in which researchers **rely on their own judgment** when choosing members of the population to participate in their surveys.
- This survey sampling method requires researchers to have **prior knowledge** about the purpose of their studies so that they can properly choose and approach eligible participants for surveys

Purposive vs Convenience sampling

- The terms purposive sampling and convenience sampling are often used interchangeably, but they do not mean the same thing.
- Convenience sampling is when researchers leverage individuals that can be identified and approached with as little effort as possible. These are often individuals that are geographically close to the researchers or those who have previously completed an online survey.
- Purposive sampling is when researchers thoroughly think through how they will establish a sample population, even if it is not statistically representative of the greater population at hand. As the name suggests, researchers went to this community on purpose because they think that these individuals fit the profile of the people that they need to reach.

Purposive vs Convenience sampling

- While the findings from purposive sampling do not always have to be statistically representative of the greater population of interest, they are qualitatively generalizable.
- The more prior information that researchers have about their particular communities of interest, the better the sample that they're going to select.

Basic ideas: Self-selection sampling

- Self-selection sampling is appropriate when we want to allow units or cases, whether individuals or organisations, to choose to take part in research on their own accord.
- The key component is that research subjects (or organisations) volunteer to take part in the research rather than being approached by the researcher directly.
- A sample is self-selected when the inclusion or exclusion of sampling units is determined by whether the units themselves agree or decline to participate in the sample, either explicitly or implicitly.

Basic ideas: Snowball sampling

- Snowball sampling is particularly appropriate when the population you are interested in is **hidden and/or hard-to-reach** (drug addicts, homeless people, individuals with AIDS/HIV, prostitutes . . .)
- This is a sampling technique, in which **existing subjects provide referrals to recruit samples** required for a research study.
- This sampling method involves a **primary data source nominating other potential data sources** that will be able to participate in the research studies.
- Snowball sampling consists of two steps:
  1. Identify potential subjects in the population. Often, only one or two subjects can be found initially.
  2. Ask those subjects to recruit other people (and then ask those people to recruit. Participants should be made aware that they do not have to provide any other names.

Focus on Quota sampling

- **Example**: We are interested in comparing the differences in career goals between male and female students at the University of Pisa. The university has $N = 10000$ students.
- Define a sample size wich is reasonable.
- To create a quota sample, there are three steps:

Focus on quota sampling

1. *Choosing the relevant stratification and dividing the population accordingly*: If we wanted to look at the differences in male and female students: gender as the stratification.
2. *Calculating a quota for each stratum*: The number of cases that should be included in each stratum will vary depending on the make-up of each stratum within the population.
3. *Continuing to invite cases until the quota for each stratum is met*: Once you have selected the number of cases you need in each stratum, you simply need to keep inviting participants to take part in your research until each of these quotas are filled.

Focus on Quota sampling

- Advantages:
  1. As representative as possible of the population being studied.
  2. Quicker and easier to carry out because it does not require a sampling frame and the strict use of random sampling techniques.
  3. The quota sample improves the representation of particular strata (groups) within the population, as well as ensuring that these strata are not over-represented.
  4. The use of a quota sample, which leads to the stratification of a sample (e.g., male and female students), allows us to more easily compare these groups (strata).
- Disadvantages:
  1. It's impossible to determine the possible sampling error.
  2. Since the strata must be mutually exclusive, this means that we would need to sample four strata from the population: undergraduate males, undergraduate females, graduate males, and graduate females. This will increase overall sample size required

Focus on Quota sampling

- Disadvantages:
  1. It's impossible to determine the possible sampling error.
  2. Since the strata must be mutually exclusive, this means that we would need to sample four strata from the population: undergraduate males, undergraduate females, graduate males, and graduate females. This will increase overall sample size required for the research, which can increase costs and time to carry out the research.

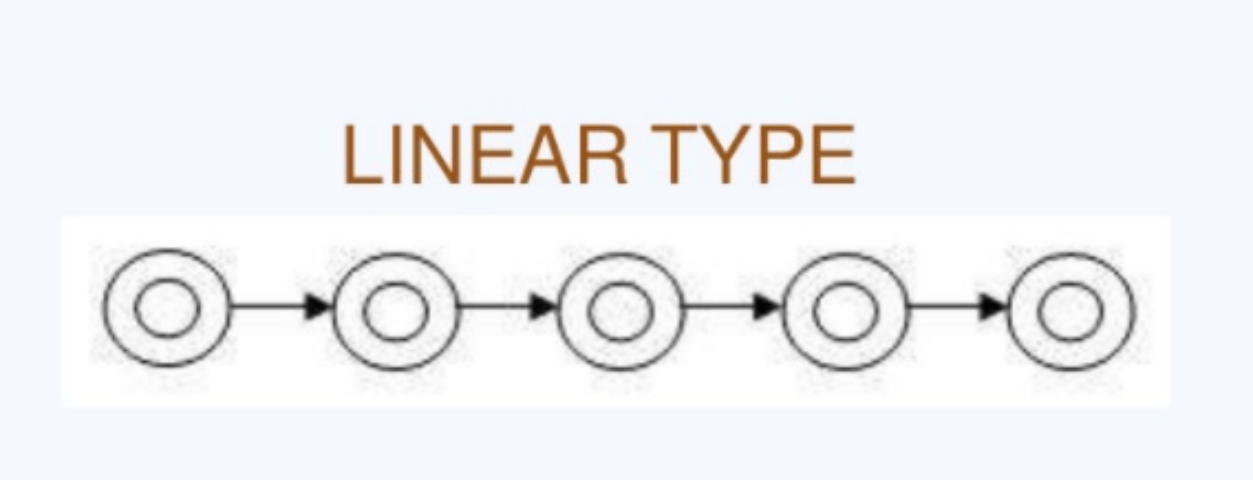
Focus on Snowball sampling

- Some populations that we are interested in studying can be hard-to-reach and/or hidden.
- These include populations such as drug addicts, homeless people, individuals with AIDS/HIV, prostitutes, and so forth.
- Such populations can be hard-to-reach and/or hidden because they exhibit some kind of social stigma, illicit or illegal behaviours, or other trait that makes them atypical and/or socially marginalized.
- Snowball sampling is a non-probability based sampling technique that can be used to gain access to such populations.

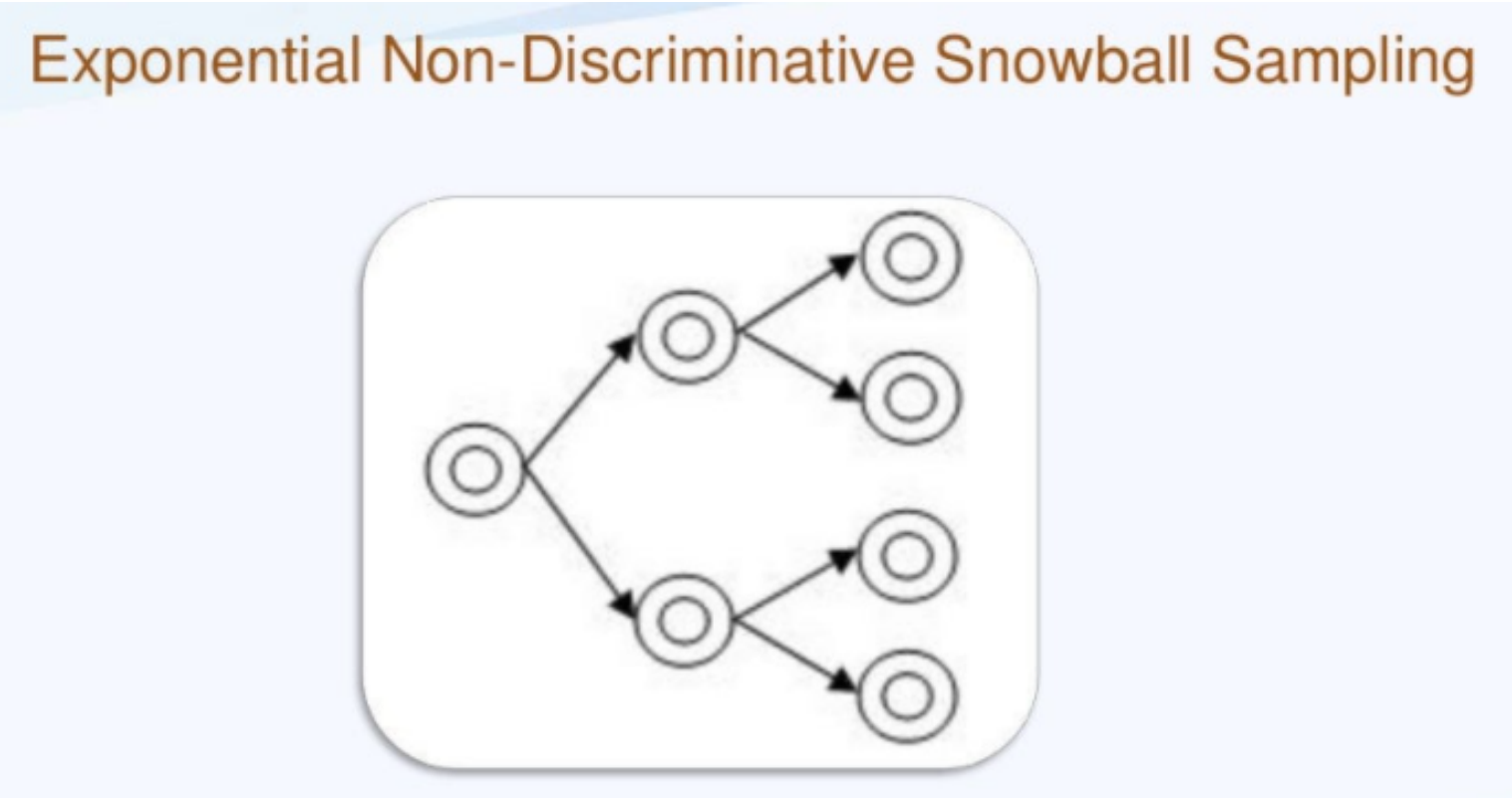
Focus on Snowball sampling

- To create a snowball sample, there are two steps:
  1. trying to identify one or more units in the desired population;
  2. using these units to find further units and so on until the sample size is met.
- It works as a chaining system: participants refer to othe participants.

Types of Snowball sampling

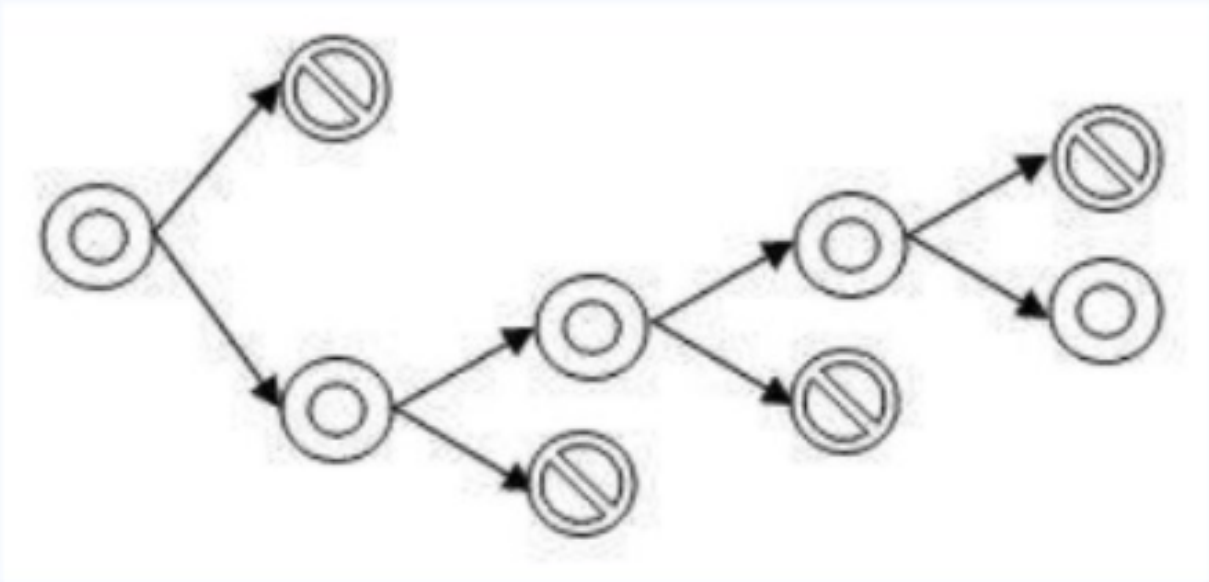


Types of Snowball sampling



Types of Snowball sampling

Exponential Discriminative Snowball Sampling



Advantages of Snowball Sampling

1. **Locate Hidden Populations**
2. **Locating People of a Specific Population**: there is no listsot other obvious sources for locating members of the population of specific interest.
3. **Cheap**: the process is cheap, simple, cost-efficient. This sampling technique needs little planning and fewer workforce compared to other sampling techniques.

Disadvantages of Snowball Sampling

Community Bias

- The first participants will have strong impact on the sample. Snowball sampling is inexact and can produce inaccurate results. The method is reliant on the skill of the individual conducting the actual sampling.
- To be successful requires previous contacts within the target areas, and the ability to keep the information flow going throughout the target group.

Wrong Archoring

- A disadvantage in snowball sampling is the lack of definite knowledge as to wheter or not the sample is an accurate reading of the target population.
- By targeting only a few selected people, it is not always indicative of the actual trends within the result of the group.
- Identifying the right person to conduct the sampling, as well as locating the correct targets is a time consuming process which renders the benefits only slightly outweighting the costs.

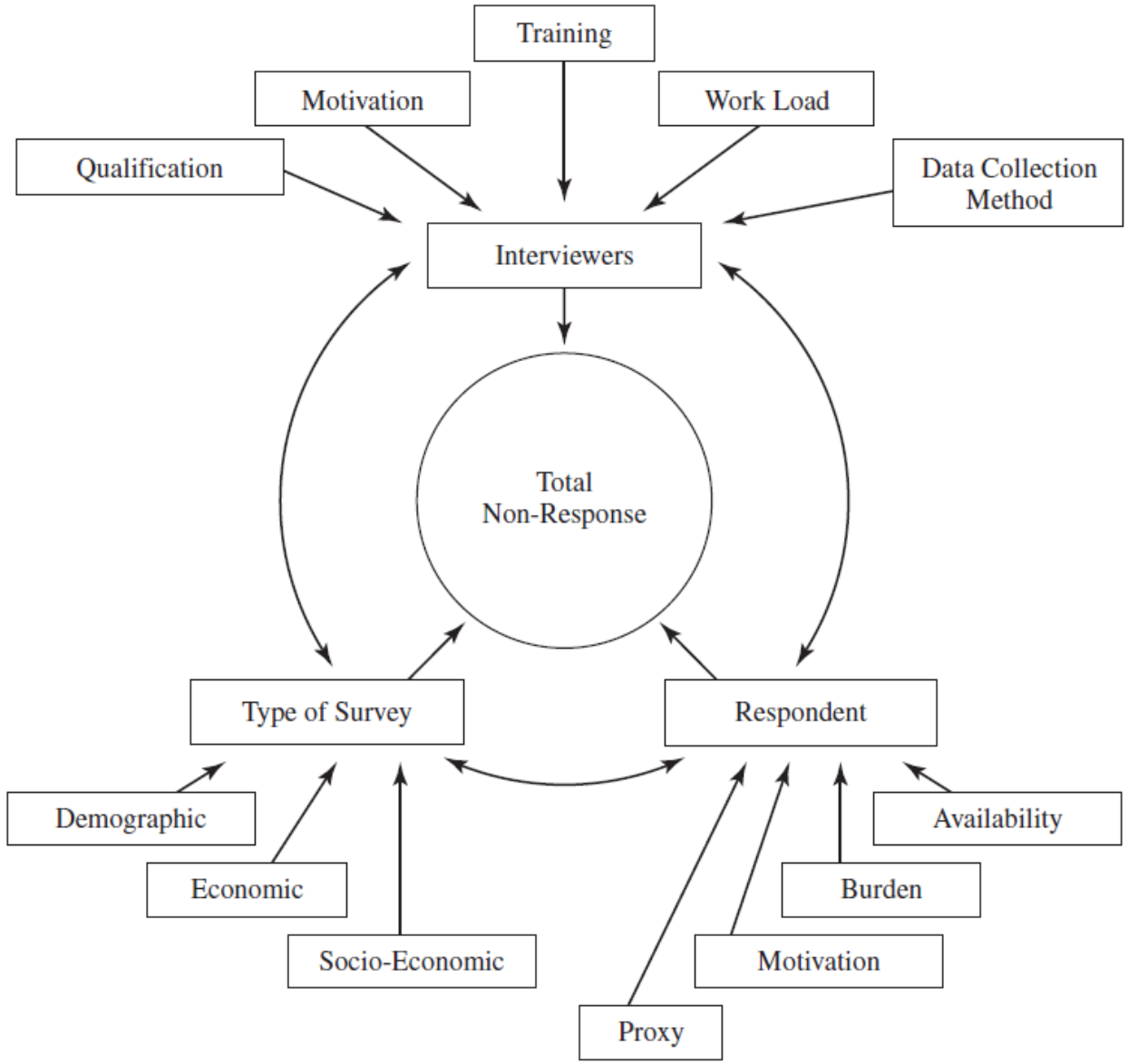
Introduction to non-response

- The best wayto deal with non-response (missing data) is to prevent it.
- Two type of non-responses:
  1. **unit non-response**: the person provides no information for the survey.
  2. **item non-response**: some information are present, the person does not respond to a particolar item.

Introduction to non-response

- Four approaches to dealing with nonresponse:
  1. Prevent it: design the survey so that the non-response is low.
  2. Take a representative subsample of the non-respondents: use the subsample to make inference about the other nonrespondents.
  3. Use a model to predict the values of the nonrespondents. Imputation often adjusts for item non-response, weighting class adjustment methods use a model to adjust for unit non-response. Parametric models may be used for unit and item non-response.
  4. Ignore the non-response (not correct even if common): non-response and undercoverage present serious problem for survey inference. The failure to obtain information from some units in the selected sample (nonresponse), or the failure to include parts of the population in the sampling frame (indercoverage) can lead to bias estimates of the population quantities.

SOURCE: "Some Factors Affecting Non-Response," by R. Platek, 1977, *Survey Methodology, 3,* 191–214.



Mechanisms for Non-response

- Even if the survey is well defined non-responses can occur.
- The methods for try to fix the non-responses are model-based.
- If we want to make inference on non-respondents we must assume that they are related to respondents ins ome way.

$$R_i = \begin{cases} 1 & \text{if unit } i \text{ responds} \\ 0 & \text{if unit } i \text{ does not respond.} \end{cases}$$

- $\phi_i = P(R_i = 1)$ (unknow and assumed positive) is the probability that a unit selected for the sample will respond (propensity score of unit i.

Mechanisms for Non-response

- **Missing completely at random**
- **Missing at random given covariates**
- **Not missing at random**

Weighting methods for non-response

- Weights can be used to adjust for non-response.
- Z_i indicator variable for presence in the selected sample $P(Z_i = 1)\pi_i$
- If R_i indep. of Z_i

$$P(\text{ unit i selected in sample and responds}) = \pi_i\phi_i$$

- The probability of responding ϕ_i is estimated for each unit in the sample using auxiliary information that is known for all units in the selected sample.
- The final weight for a respondent is $\frac{1}{\pi_i\hat{\phi}_i}$
- **Weighting methods assume that the response probabilities can e estimated from variables nown for all units.**
- **Weighting methods assume MAR data.**

- **Regression imputation** predicts the missing value using a regression of the item of interest on variables observed for all cases.

Gaia Bertarelli

Department of Economics and Management in the Era of Data Science

gaia.bertarelli@santannapisa.it