LAB-MOVIE
LABOUR MARKET OBSERVATORY IN VIETNAM UNIVERSITIES

Co-funded by the Erasmus+ Programme of the European Union

LAB-MOVIE
Labour Market Observatory in Vietnam universities

WP 2
Transfer of knowledge and methodology to analyse the labour market

Outcome 2.1
Development of educational materials

THE SAMPLING

KA2 - COOPERATION FOR INNOVATION AND THE EXCHANGE OF GOOD PRACTICES
PROJECT NUMBER: 609653-EPP-1-2019-1-IT-EPPKA2-CBHE-JP Lab Movie

THE SAMPLING

The topics that will be examined:

1. Introduction.
2. Probability sampling:
   - 2.1. Selection criteria.
   - 2.2. Probability sampling designs.
   - 2.3. Sample size.
   - 2.4. Sampling error.
   - 2.5. List of units.
3. Non-probability sample:
   - 3.1. Non-probability sampling designs.

1. Introduction

Population (target): set N of statistical units that are the subject of the study (Set N = population size).

Sample: set of the n sample units selected from the N units making up the population (Set n= sample size).

Sampling: procedure by which the n sample units (sample) are selected from the population.

Sampling design
A sampling procedure consists in defining the methodology for selecting the population units to be included in the sample. The sampling design (or technical/method/strategy/sampling plan) defines the criteria/methods of sample extraction.

A first distinction can be made between probability and non-probability sampling:

- **Probability sampling**: all units have a known non zero probability of being selected (you can make inference, then extend the results obtained to the whole population. The sample is representative).
- **Non-probability sampling**: some units of the population have a zero probability of becoming part of the sample (it is not possible to make inference, therefore it is not possible to extend the results obtained to the whole population. The sample is not representative).

2. Probability sampling

Probability sampling may be:

- At **constant probabilities**: all the units of the population have the same probability of being selected (self-weighting).
- At **variable probabilities**: the units of the population have a different probability of being selected (not self-weighting).

Variable probability sampling
Probability Proportional to Size (PPS) sample:

- It represents the most efficient variable probability design.

- It is the most logical design: the most "important" units are more likely to be extracted.
- The probabilities (the weights) are determined through an auxiliary variable X, correlated to the phenomenon Y in analysis (or that however determines a > or < representativeness of the units with respect to Y).

E.g.: Y=turnover enterprises or hirings made; X=number of employees (size of the enterprise)

Variable probability sampling VS Constant probability sampling
Advantages

- Control in the selection with respect to the size of the units.
- Efficiency gain commensurate with the relationship between unit size and the variable aim of the study (if there is proportionality the sample may be smaller, with the same expected efficiency).

Disadvantages

- It is necessary to know the size of the population units.
- The resulting sample is not self-weighting.
- The estimators to be adopted are complex.

Regardless of the probability of selection, the prerequisites for obtaining a probability sample are:

- The existence of a list.
- The selection of the units must be made using random criteria.

2.1. Selection criteria

There are two criteria for the selection of units which, regardless of the probability of selection, guarantee randomness in the identification of the sample.

1. Random selection.
2. Systematic selection.

With constant probabilities
The **random selection** is made by:

- Selecting units from an urn.
- Using the random number tables.
- Generating the random numbers with the computer.

The **systematic selection** is done by ordering (randomly) the units to be sampled and selecting one of every many, starting from a randomly determined unit.
Procedure:

- Ordering (random) the statistical units.
- Definition of the "sampling step" (k=N/n).
- Identification of the n units:
  - Selection of a random number r such that 1=<r>=k.
  - Identification of the sample units r; r+k, r+2k, ..., r+(n-1)k.

With variable probabilities
The **random selection** is made:

- By matching the unit i-th Mi random numbers (where M is the total amount of random numbers).

- Cumulating the values of Mi for the units in the list, until the total M is at unit N.
- Assigning the first M1 random numbers to the first unit, from M1+1 to M1+M2 to the second, from M1+M2+1 to M1+M2+M3 to the third, and so on.
- Extracting, using the random number tables or generating the random numbers with the computer, n random numbers. The n units to which the extracted random numbers are matched will become part of the sample.

The **systematic selection** is made:

- Matching the unit i-th Mi random numbers (where M is the total amount of random numbers).
- Cumulating the values of Mi for the units in the list, until the total M is at unit N.
- Assigning the first M1 random numbers to the first unit, M1+1 to M1+M2 to the second, M1+M2+1 to M1+M2+M3 to the third, and so on.
- Definition of the "sampling step" (k=M/n).
- Identification of n units:
  - Selection of a random number r such that 1=<r>=k.
  - Identification of the sample units r; r+k, r+2k, ..., r+(n-1)k.

NOTE: Units can be selected with or without replacement. We will not go into this in depth, but consider that for our purposes the extraction must take place without replacement so, with random selection it is necessary to exclude the units (the random numbers associated with the units) that are extracted as they are extracted.

Systematic sampling VS Random sampling

Advantages

- Simplicity of the procedure.

Disadvantages

- It is effectively a pseudo-random criterion since only one random number is extracted and all the others are automatically determined.

2.2. Probability sampling designs

Probability sampling designs can be distinguished between:

- Simple sample designs.
- Complex sample designs.

2.2.1 Simple sample designs

<u>Random sampling</u>

It consists in the selection of the units adopting a selection criterion that guarantees randomness.

A distinction is made between:

- Simple random sampling.
- Variable probability (random) sampling.

Simple random sampling
Simple Random Sampling consists of selecting units by extracting them from the population with the same probability (at each step of the extraction), adopting random selection or systematic selection (**Systematic sampling or Sampling with systematic selection**).

When using Simple random sampling:
- The population is homogeneous.
- You have good lists of the entire population.
- The cost to reach each unit is homogeneous and does not vary if more complex designs are used.
- You want to use simple estimators.
- You want to estimate complex relationships and other designs have comparable costs.

When you can do better than Simple random sampling:
- You have auxiliary information about the population.
- The population is divided into homogeneous groups.
- The lists are present for groups of units and not for the entire population (hierarchical structure of the lists).
- The costs to reach the units can vary considerably and different designs result in much lower costs.

Variable probability (random) sampling
Variable probability sampling consists of selecting units from the population with different probabilities, using either random selection or systematic selection (**Systematic variable probability sampling or Variable probability sampling with systematic selection**).

2.2.2 Complex sample designs

Why complex sample designs:
- Random sampling is often a theoretical reference that is not very applicable in reality (e.g.: you need a single list of the whole population, it is not feasible with face-to-face interviews on a very large population because of time and costs).
- We often want to have more control over the selection while keeping it random.

The complex sample designs are:
- Stratified sampling.
- Multi-stage sampling (we will limit ourselves to a simple description).
- Cluster sampling (we will limit ourselves to a simple description).
- Area sampling (we will not go into detail).
- Double sampling (we will not go into detail).

<u>Stratified sampling</u>

It consists of partitioning the population into sub-populations, called strata, and selecting a probability sample within each stratum (a sample can be formed with a different criterion in each stratum).

A distinction is made between:

- Proportional stratified sampling.
- Non-proportional (or variable probability) stratified sampling.

Proportional stratified sampling

Proportional stratified sampling consists of extracting an equal proportion of units from each strata, using either random selection or systematic selection (**Systematic proportional stratified sampling or Proportional stratified sampling with systematic selection**).

Implicit stratification

It is a particular type of systematic selection that allows to select a stratified sample by taking control over the selection deeper than is made explicit with stratification.

- **With a stratification variable (quantitative).**
  It consists in sorting the units from the one with the highest value to the one with the lowest value and selecting the sample using systematic selection (**Sampling with implicit stratification**).

- **With one or more qualitative stratification variable and one quantitative variable (serpentine process).**
  An interesting case of implicit stratification applicable to various strata is the serpentine procedure which consists in dividing the population into strata (defined by one or more qualitative variables) and ordering the units within the first stratum by increasing values of the implicit (quantitative) variable, then by decreasing values those of the second stratum. In the following strata (if present), the units are divided by increasing and then decreasing values and so on (**Sampling with serpentine stratification**).

Proportional stratified sampling VS Simple random sampling

The proportional stratified sample is more efficient than the simple random sample with the same sample size. The efficiency is directly proportional to the variance between strata.

Non-proportional (or variable probability) stratified sampling

Non-proportional stratified sampling consists of extracting different proportions of units from each strata, using either random selection or systematic selection (**Systematic non-proportional stratified sampling or Non-proportional stratified sampling with systematic selection**).

Also in this case it is possible to adopt the implicit stratification (**Sampling with non-proportional implicit stratification, Sampling with non-proportional serpentine stratification**).

It is recommended to sample with high fractions the strata for which:

- The variance within the stratum is high.
- It is not expensive to sample in the stratum.
- The stratum is very large.

Why stratification should be used:

- Because the population is naturally organized into subpopulations.
- To highlight sets of units that are significant for research.
- To separate subpopulations with specific characteristics.

- To identify units to be detected with particular techniques.
- To introduce maximum control into the selection, while keeping it random.
- To make the sub-populations homogeneous with respect to the variables to be collected, so that the estimates are more efficient than those obtained by random sampling.

The stratification variables

- There must be a relationship between stratification variables and the variable of interest.
- The stratified sample is efficient if the strata are homogeneous within them and very different from each other.
- It is more efficient to use several stratification variables rather than several options of answer of the same variable.
- Stratification variables are qualitative variables (categorical) or quantitative variables reduced into classes.

Multi-stage sampling

It is based on the concept of hierarchical populations: the final population of units is contained in an aggregate of units of a higher level (or stage), which can be contained in units that are increasingly small in number and large in size.
It consists of dividing the population into hierarchical levels or subgroups (clusters). At a first stage we will proceed by selecting a probabilistic sample of the subgroups (clusters), at a second stage we will proceed by selecting a probabilistic sample of detection units from each previously selected subgroup.

The extraction of the sample can be carried out with different criteria at each stage. It can be done with constant or variable probabilities, from variously stratified lists at each stage, adopting random selection or systematic selection.

E.g.: If you want to conduct a survey on a sample of customers of a certain chain first sample the stores then the customers (of the selected store), in this case we have adopted a two-stage sampling.

NOTE: multi-stage sampling can be at two or more stages depending on how many hierarchical levels you decide to adopt (e.g. Regions, provinces, ..., shops, customers).

Why stage sampling should be used:

- You do not have a single list of the statistical population and creating it from the available lists is too expensive.
- The population is naturally organized in clusters.
- The population is distributed over a large territory, and organisational and economic constraints prevent random or stratified sampling.

Cluster sampling

This is a special case of multi-stage sampling. It consists, in fact, in selecting all the detection units belonging to the subgroups (clusters) selected at the first stage.

2.3. Sample size

The optimal number of a sample is that which allows the objectives of the survey to be achieved at minimum cost and it will be the smallest number by which the estimates reach the level of reliability expected by the researcher.
The factors influencing the determination of the optimal sample size may be:

- The variance of the observed phenomenon.
- The sample error that is considered admissible.
- The cost (fixed cost + cost per observation).
- The presence of many variables.
- The need for significant estimates for sub-populations.

2.4. Sampling error

Difference between what results from the sample analysis (estimation) and the true value of the population that is not known (distortion).

The sample error depends on the case and on the goodness of the selected sample (on the sampling design adopted and on the stringency with which it is applied).

The sampling error is inversely proportional to the sample size (zero if the survey involves the entire target population).

2.5. List of units

List of units making up the statistical population.
The list must be:

- Complete (must coincide with the target population).
- Without duplication.
- Stable over time.
- Computerized.

Problems:

- Often no single list is available, but it is fragmented (e.g.: municipal registries).
- The list may not be available (e.g. supermarket customers).
- To prepare the complete list can be extremely expensive, if not impossible.

Actually it is not necessary to have the complete list (simple list): it must be defined in the contents, but then, depending on the sampling design, it is sufficient to have the parts of the list from which the sample is extracted (complex list).

Simple list: list of labels that correspond one to one to the units of the population (Simple random sample).
Complex list (made up of several lists): distinguished by subpopulations, hierarchical, dynamic (Complex Samples).

NOTE: Probability sampling assumes that the statistical population (the units that make up the list at our disposal) coincides with the target population (the theoretical population to which we want to extend the results).
The non coincidence between these populations generates the (non-sampling) error of non coverage.
In the case in which the list used does not coincide with the target population for evident limits it is always possible to realize a probability sample whose results can be extended to the statistical population, but not to the target one.
E.g.: Sample of companies selected from the lists of the chamber of commerce (it is possible to make inference on the target population because all companies are obliged by law to register to the chamber of commerce, the coverage error is linked to the fact that, for example, the lists may not be very up-to-date, therefore, do not coincide perfectly with the real companies currently active).
Sample of companies selected from the Yellow Pages or from those that are registered in trade associations (it is possible to make inferences on the statistical population - on companies on the Yellow Pages or registered in a trade association - but not on the target population because there are many companies that are not on the Yellow Pages and are not registered in a trade association).

3. Non-probability sample

The prerequisites for obtaining a non-probabilistic sample are:

- The lack of a list (but it can also exist).
- The selection of the units must take place through non-random criteria.

3.1. Non-probability sampling designs

The non-probability sampling designs are:

- Convenience or accidental sampling.
- Voluntary sampling.
- Snowball sampling.
- Judgment sampling.
- Quota sampling.
- Privileged witness sampling.

Convenience or accidental sampling

The choice of units is based on those that are most easily available.

Voluntary sampling

The units decide whether to be part of the sample or not (e.g.: samples formed by newspaper readers who respond spontaneously to certain surveys).

Snowball sampling

Units are selected using the relational networks of a group of units initially identified (this is used in the case of populations made up of individuals who tend to conceal their identity or are difficult to find.
E.g: Survey on immigrants without residence permit: an immigrant is contacted, interviewed and then asked to indicate another immigrant of his or her knowledge willing to answer the interview).

Judgment sampling

The choice of the units is based on the judgment of the researcher who knows the phenomenon and with criteria, more or less personal, tries to extract a sample of the population.

Quota sampling

The reference is proportional stratified sampling:

- It subdivides the population according to some variables of which the distribution in the population is known (size of the holdings from the census).
- The sample is constructed respecting the proportions (quotas) of the stratification variables in the population.
- Units are selected until the quotas are reached (the surveyor is free to choose at his discretion the units to be interviewed as long as he keeps to the quotas).

Privileged witness sampling

You do not select units that are part of the population, but "privileged witnesses".

Privileged witnesses are people who are recognised as having a particular knowledge and/or competence in the subject matter and a particular ability to interpret facts.

This type of non-probabilistic sampling is mainly used in Focus Groups (which we will explore separately) and with the Delphi Method.

Why is probabilistic sampling not used?

- Too long time compared to information needs.
- Costs are too high.
- Absence of the population list.
- The non-sampling error does not guarantee the complete reliability of the results.

Typically, non-probabilistic sampling is used in market research, electoral and opinion polls, where little information needs to be found very quickly.