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 DATA ANALYSIS

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

“The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.”

THE DATA ANALYSIS

The topics that will be examined:

1. The types of variables.
2. The coding of data.
3. The recording of data on a computer support.
4. The revision of the collected data:
   - 4.1. Data cleaning.
   - 4.2. Missing data.
5. The data processing:
   - 5.1. Tables of data.
   - 5.2. Survey weights.

1. The types of variables

The **variables** are characteristics, usually elementary, referring to the statistical units, therefore, characteristics of the statistical population.

The **set of responses** represents, instead, the possible outcomes of a measurement, i.e. the values in which a variable can vary.

The **classification of variables** is the set of values that a variable takes on (at the level of maximum detail).

Variables can be:

- **Quantitative** (measured with a numerical result).
- **Qualitative** (not measured with a numerical result).

Quantitative variables can be:

- **Discrete** (the set of values they can assume is finite or numerable; the set of natural numbers 1,2,3,4, ...).
- **Continuous** (the set of values they can assume is the set of real numbers or a range of real numbers).

The qualitative variables can be:

- **Ordinal** (the set of responses naturally has an order, i.e. it can be arranged along a scale).
- **Nominal** (the set of responses has no natural order).

2. The coding of data

It is a matter of preparing the data collected so that they can then be transferred to a computer support system.

Through the encoding of the data, a numerical code is associated with each response value (the encoding is done when preparing the questionnaire - see questionnaire).

- When facing answers such as "I don't know", "don't answer", "I don't remember" it is preferable to use a particular code that applies to all questions (you can use digits such as 99 for "I don't know", 98 for "don't answer", 97 for "I don't remember").

- For non-applicable questions, easily distinguishable codes such as 777, 888, 999 and similar should be used.

The use of these codes is useful because blanks can cause confusion (it is not clear what they indicate) and can give computational problems in subsequent processing.

In the case of open questions, we operate as follows:

- examine the answers,
- decide how to group the answers into substantially homogeneous categories,
- associate a code.

It is then a matter of transforming open questions into closed questions.
An example of coding is the transposition of information concerning the sector of economic activity of enterprises, described in colloquial form by the respondent, into the respective codes according to the ATECO classification.

3. The recording of data on a computer support

It is a matter of transferring the coded data to a computer support suitable for subsequent statistical analysis.
The main purpose is to build the so-called data matrix that:

- contains as many rows (records) as the analysis units (e.g. interviewed companies),
- contains as many columns (fields) as the considered variables (basically the questions reported in the questionnaire).

Some survey techniques foresee the simultaneous input of data on computer support (CATI, CAPI, CAWI).

4. The revision of the collected data

Situations to check:

- **Out-of-domain values**: the values of a variable do not belong to a predefined set of allowable values (out-of-domain values give rise to errors).
- **Abnormal values**: a unit is abnormal (outlier) when it has characteristics significantly different from those of most units (abnormal values give rise to errors).
- **Incompatibility between responses**: the values of one or more variables contradict predefined logical rules and/or mathematical relationships (incompatibilities lead to error situations, but often it is not known on which variable).

Types of controls:

- **Validity or range checks**: checks that the values assumed by a given variable are within the variable's definition range.
- **Statistical checks**: used to isolate those units that have values for some of the variables contained in them that deviate significantly from the values assumed by the same variables in the rest of the units. These values are with high probability errors, but further checks are necessary.
- **Consistency checks**: check that predefined conditions of values assumed by variables taken in the same unit meet certain requirements (incompatibility rules).

Incompatibility plan

Consistency checks are used for the construction of incompatibility plans.

An incompatibility plan is a set of non-redundant and non-contradictory constraints which must be met simultaneously by each statistical unit in order for the corresponding information to be considered correct.

The rules making up an incompatibility plan can be distinguished into:

- **Formal rules**, which derive from the structure of the questionnaire, i.e. directly from the compilation rules and the internal paths of the questionnaire.
- **Substantive rules**, which derive from statistical/mathematical considerations, or from a priori specific knowledge of the phenomenon under study.

Once the records whose values violate one or more constraints of the plan of incompatibility have been identified, the problem becomes the location of the variables responsible for this violation: it is only these, in fact, the variables whose values must be considered incorrect (missing) and therefore correct.

4.1. Data cleaning

If we have found the records containing incorrect values and we know which variables are responsible for this incorrectness, we can make the following changes:

- Return to the source.
- Deterministic or probabilistic imputation.
- Treatment of the data as a non-response item.

Return to the source

If possible, request the data directly from the respondent or who completed the questionnaire (it is expensive in terms of time and work, but guarantees that you have the most correct data).

Deterministic imputation

It means assigning a value to the variable instead of the wrong value, based on other information (e.g.: if age < 14 then n° children = 0).

Typically, the information must be available within the data set and the unit itself.

Generally, to proceed with the imputation with certainty it is necessary that the information comes from more than one variable, otherwise we would not be able to decide (between two variables with incompatible values) which is the right one.

Probabilistic imputation

It means assigning a value to the variable instead of the wrong value, based on what has been observed in units with the same characteristics.

Treatment of the data as a non-response item

If there are no tools to recover or attribute the correct data, all that remains is to delete the wrong data and treat that field as a missing data.

4.2. Missing data

A very frequent problem in sample surveys is that of non-response.

Non-response can be:

- **Total** (one or more units do not answer the whole questionnaire).
- **Partial** (one or more units do not answer one or more questions).

Total non-response
For voluntary reasons (the "respondent" does not want to answer) or accidental (he is out of town for work, he is absent if the survey is done at school, etc.), there is always a quota, more or less large, of people who, despite being part of the sample identified, do not respond.

Generally, in these cases a replacement is made by selecting new people with the same characteristics as those who refused (a total non-response rate of up to 20% can be considered acceptable).

It is however appropriate to:

- carry out a rigorous control of the sample obtained (to make sure that the sample is really "similar" to the population),
- check its homogeneity with the population for some relevant variables not included in the sampling plan.

Partial non-response
There is also always a quota of people who for various reasons (they do not know or do not want to give an answer) do not answer all the questions.

In these cases, with due attention and only in the most obvious/clear cases, it is possible to assign a value to the variable, instead of the missing value, through a deterministic imputation. Partial non-response can be included in the consistency checks (incompatibility plan).

5. The data processing

Data processing enables statistical information to be provided.
Providing statistical information means:

- To set up a system of presentation and consultation of the data which is appropriate for the stakeholders and which is methodologically impeccable (whether on the Internet or on paper).
- To define and structure statistical information by means of an appropriate system of indicators.
- Prepare statistical data (tables and graphs) in a correct and readable manner.

Among the various elaborations, graphic and tabular representations are particularly important (which are those that actually interest us, descriptive analysis).

5.1. Tables of data

A data table numerically describes a variable, or the relationship between two or more variables.
The table must contain all the information needed to understand the data, regardless of the text in which it is inserted (in turn, the table supports the text, which must be clear even without a table).
It consists of:

- Title.
- Body.

The data in a table can be made more eloquent by a graph. The graph can only replace the table if it contains the same information.

5.1.1. Table title

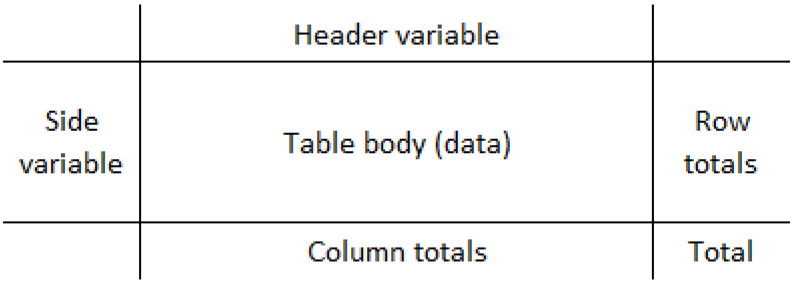
The title of the tables must be perfectly explanatory:

- What we are talking about (companies, population, users, people, ...), therefore, clearly define the unit represented in the table.
- Specify what type of data is present in the table (absolute values, percentages, rates, ...).
- Specify side and head variables.
- Specify location, reference period (if the table is constructed from administrative data or official sources, the source must be indicated).

The tables can show numbers (absolute values or ratios) or percentages.

5.1.2. Table of numbers

Title: What, described with what type of data, according to which variables. Place, year.



	Header variable	
Side variable	Table body (data)	Row totals
	Column totals	Total

Tables with absolute values must always have row and column totals.

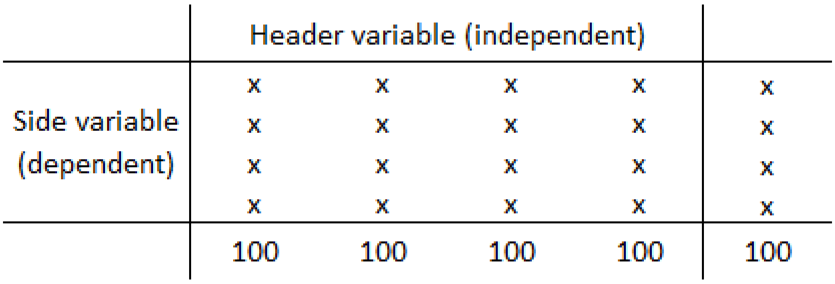
5.1.3. Table of percentages

The table may contain:

- Column percentages (which are the ones we are interested in).
- Row percentages.
- Cell percentages.

Tables with percentage distributions must be marked 100. If the distribution is per row or column, there must also be the marginal percentage distribution, which is not the sum of the percentages.

Example of a table with column percentages.
Title: What, described in percentages, according to which variables. Place, year.



	Header variable (independent)				
Side variable (dependent)	x x x x	x x x x	x x x x	x x x x	x x x x
	100	100	100	100	100

Special cases

In the case of multiple-choice questions, attention must be paid to the presentation of the percentages: their sum exceeds 100, so an explanatory note must be inserted and the sum of the percentages must not be added.

5.1.4. Missing data in the tables

In absolute number tables: the missing data is treated as a value and must be entered in the last row and last column, before the totals. Excluding missing data from the absolute number table means to present tables with different totals.

In tables with percentages:

1. Insert the percentages of missing data inside the table, treating them as a value (it might make little sense to insert the % of missing data in the table, because in this way we do not know the true distribution of the phenomenon).
2. Calculate the percentage distributions after excluding the missing data, which should be indicated at the bottom of the table.
3. Redistribute the missing data within the table, respecting the distribution of the case history (acceptable if the missing data are equidistributed).

5.2. Survey weights

- Each unit in the sample represents a certain number of units in the population, so that the whole sample represents the whole population.
- When we provide the results of a survey, each unit must be associated with its weight, so that the total number of units corresponds to N.
- The weight associated with each unit is called the survey weights and is equal to the inverse of the selection probability.

Use of survey weights

- In **self-weighing samples** the weights are constant for all the sample units. The estimates remain the same with or without the weights (but the estimate variance or sample error changes).
- In **non-self-weighting samples**, the use of weights is necessary to obtain correct estimates.