Data Analysis

Data can be analysed in different ways-

- Total comparison of different runs of data acquisition.
- Total comparison of acquired data with standard data (for discrete data, one may compute precision-recall, etc.)
- Simple summary statistics.
- Bivariate analysis (linear or higher order).
- Multivariate analysis (can be complex).

The following is a list of the main types of analysis:

− Simple derivations (e.g. means, medians, quantiles, histograms, two and three way tables, scatter plots, etc.).
− Complex derivations (e.g. Gini coefficients, expectation of life, etc.).
− Data models (e.g. population estimates, population projections, etc.).
− Social indicators (e.g. fertility rate, dependence ratio, labor participation rate, etc).
− Mathematical modeling (e.g. linear regression, factor analysis, etc.).

Descriptive versus explanatory purpose

If the survey has a descriptive purpose, then focus is to estimate the parameters (values such as total number, percentage, average, etc) in a population e.g. the number employed, unemployed, etc. This can also be labelled the 'What' analysis that focuses on significant facts surrounding an issue.

A survey may also have an explanatory purpose. Apart from knowing the facts around a particular issue, one may also want to address the 'Why' questions i.e. not merely showing a trend moving in a certain direction (e.g. by a graph or indicator) but also to explain why it has so moved. In that case the purpose is to establish the causal relationship between two or more variables. This may be connected to theory, a model, or earlier findings.

In a statistical report a typical descriptive comment to the table could be:
Agricultural work is the most frequent occupation among both men and women. 90 per cent of the women and 75 per cent of the men are employed in agriculture. The table could, however, be interpreted in another way – as showing correlation between the variables sex and occupation. In this case we look at the table as two- dimensional, one dimension being occupational distributions, the other being sex. One could surmise from the table that sex affects occupation.

Relationships between variables

The variable that one may wish to explain is generally labelled the **dependent variable**. The variable expected to explain the change in the dependent variable is referred to as the **independent variable**. Most of the phenomena that one may wish to analyse also often call for more than one independent variable to explain variation in the dependent variable. This happens because of the complexity of phenomena. One independent variable usually explains only a certain amount of the variation in the dependent variable, and more independent variables have to be introduced in order to explain more of the variation.

Example: In a network traffic study, the number of communicating applications can be an independent variable and the number of data packets arriving at router may be a dependent variable.

The object (counting unit or unit of analysis)

Decisions about the issue to be analysed, the data to be used, and the statistical technique to be applied, are fundamental to good analysis. However, an even more basic level question is the object or unit of analysis.

An object is defined as any concrete or abstract entity (physical object, living creature, organization, event, etc.) that the users may want to have information about. Objects for the particular survey are items (elements or units) that the users would like to have information on (e.g. computing node, data packet, etc).

Different types of variables

For every object, there will be several variables or properties of interest, e.g. for the object PERSON we can have age, income, occupation, marital status, etc., as variables. Variables may be grouped according to classifications (e.g. 5-year age groups, countries of birthplace, etc.). Apart from the choice of the object of analysis, it is the variables and their classifications that determine the limits and richness of the analysis.

Variables may be qualitative or quantitative.

The first division whether a variable is numerical (i.e. a variable for which the value is indicated using numbers) and is called quantitative (e.g. the variable weight – 8 kg, etc.). If instead we use words then we speak of a qualitative variable e.g. the variable place of residence is described in words (urban-rural).

Among the quantitative variables we distinguish between those that can only take certain values and those that can take all the values in a range. Variables that can only take certain values (usually whole numbers) are called **discrete** – the number of children is, for example discrete. Variables that can take all numbers in a range (e.g. age) are called **continuous.**

In some cases the type of variable determines the type of measurement whilst in some cases there is a choice. The variable ‘age’, for example, can be given in years but can also have measurement values such as youth- middle age-old.

Different scales of measurement-

Level of measurement	**Characteristics, the measurement values can be ...**
Nominal scale	Distinguished
Ordinal scale	Distinguished and ranked
Interval scale	Distinguished, ranked and measured with constant units o measurement

Different types of tables or charts suit different types of variable. A typical example is the histogram, which is used for continuous variables. It would not be appropriate to use a histogram to illustrate a discrete variable.

Table construction

(a) Frequency distributions – one dimensional tables

(b) Percentage distributions – Often absolute frequency values are not informative enough; not suitable for many comparisons. Proportions and percentages permit the comparison of two or more frequency distributions.

Bivariate distributions (two dimensional tables, two way tables) – Cross Tabulation

The presence of 'significant' relationships between variables. In terms of analysis a relation implies a relation between two or more variables. When we say X and Y are related, we mean that there is something in common to both variables; the two 'go together', i.e. that they covary. Typically, relationship between variables are identified through cross tabulation.

In a bivariate table, there are two cross-classified variables. Such a table consists of rows and columns. The categories of one variable are labels for the rows while the categories of the second variable are labels of the columns. Usually the independent variable is the column variable (listed across the top) and the dependent variable is the row variable (listed on the left side of the table).

Percentages for bivariate distributions

To summarise a bivariate table it is useful to compare its univariate distributions. Presenting its frequencies in percentages can achieve this. It is also the predominant method of analysis. The scale or level of measurement has no importance in this case. The method can be used in the analysis of variables at nominal level, ordinal level, or interval level.

Control techniques

An association between two variables is not a sufficient basis for inferring that the two variables are causally related. Other variables must be ruled out as alternative explanations. For example, a relationship between height and income can probably be accounted for by the variable age. Age is related to both income and height, and this joint relationship produces a statistical relationship that has no causal significance. The original relation between height and income is said to be a *spurious relation*.

Cross tabulation is a method of control that can be compared to the mechanism of matching, used in experiments. It involves the division of the sample into subgroups of the controlled variable. The original bivariate relation is then tested within each subgroup. Only variables that are associated with both the independent and dependent variable can potentially bias the results and are to be selected as control variables.

The reporting, presentation and dissemination of survey data

There are three major basic outputs from a statistical survey:
− *Macrodata* – 'statistics' representing estimates for certain statistical characteristics; these data are the primary purpose of the survey being carried.
− *Microdata* – 'observations of individual objects', underlying the macrodata produced by the survey; these data are essential for future use and interpretation of the survey results.
− *Metadata* – 'data describing the meaning, accuracy, availability and other important properties of the underlying micro and macro data'; these are essential for correctly identifying and retrieving relevant statistical data for a specific problem as well as for correctly interpreting and (re)-using the statistical data.

These basic outputs need to be packaged and should be made available in a user-acceptable form through appropriate distribution channels.

Survey reports

The presentation of the results of a survey or data collection is expected to meet the following demands:

− The main objectives of the survey
− The main issues/questions
− The methods used
− The definitions
− Quality of data
− Interpretation of data
− References to the technical report (if any).
− Main results illustrated by tables, diagrams/charts on an aggregated level.
− Detailed tabulations.