**Chapter 5: Data Processing and Analysis**

**1) Coding, editing and cleaning the data**

Once the field data begin to flow in, the attention is turned to data processing and analysis. Data processing implies the editing, coding, classifying and tabulation of the collected data so that they are anable to analysis. Back in the research design stage decision has been made on how to analyze the expected data. Data processing is important because it helps to:

* Check for the accuracy of the data
* Its conversion from raw to classified and reduced forms more appropriate for the analysis and interpretation states.

The following activities are generally involved in the data processing exercise.

*a) Editing the first step in data processing is to edit the raw data.* Editing of data is the process of examining the collected raw data to detect errors and omissions and to correct theses when possible. Editing involves a careful scrutiny of the completed questionnaires. One edits the data to ensure that the minimum data quality standards are achieved .In general one edits to assure that the data are:

* Accurate
* Consistent with other information/facts gathered
* Uniformly entered
* As complete as possible
* Arranged to facilitate coding and tabulation.

The editing can be done at two levels; on the field and in the office

1. **Field level Editing-** during the stress of data collection, the interviewer often uses ad hoc abbreviations special sample and the like as son as possible often an interview, field workers should review their reporting fonns, complete what was abgorviated, translate personal shorthand, and rewrite illegible entries. where needed to fill gaps, a callback should be made rather than guessing what the respondent probably would have said investigator must restrain himself/herself and must not correct errors of omissions by simply guessing what the informant would have said if the questions had been asked.

*ii) Central editing*- At this stage the research form should get a through editing and takes place when all forms or schedules have been completed and returned to the office. One editor or a learn of editors may connect obvious errors such as entry in wrong place, recorded in wrong units, etc.

b) **Coding** refers to the process of assigning numerals or other symbols to answers so that responses can be put into a limited number of categories of classes. By this method several thousands of replies or answers can be reduced to a few categories, which contain the critical information needed for analyses.

Coding is necessary for efficient analysis and through it the several replies may be reduced to small number of classes, which contain the critical information required for analysis. Data are transcribed from a questionnaire to a coding sheet. However, the coding must be:

1. Appropriate which implies that the classes or categories must provide the best partitioning of data for testing hypothesis and showing relationships.
2. Exhaustive – three must be a class for every data item. If there are large number of ‘other’ responses it indicates that our classification set may be too limited. Under such cases we may not tap the full range of information in the data.
3. Mutual exclusively- category components should be mutually exclusive meaning that specific answers can be placed in one and only one cell in a given category set.

**c) Classification** – most research studies result in a large volume of raw data, which must be reduced into homogeneous groups if we are to get meaningful relationships. This necessitates classification of data, which happens to be the process of arranging data in groups or classes on the basic of common characteristics. Data having common characteristics are placed in class and in this way the entire data get divided into a number of groups or classes. Classification can be one of the following two types depending upon the nature of the phenomenon.

* 1. Classification according to attributes- data are classified according to some common characteristics which can either be descriptive (literacy, sex, honesty, etc) or numerical such as weight, height income, etc.) Descriptive characteristics refer to qualitative phenomenon, which cannot be measured quantitatively. Only their presence can be noticed. Data obtained in this way on the basis of certain attributes are known as statistics of attributes and their classification is said to be classification according to attributes and their classification is said to be classification according to attributes. Such classification can be simple classification or manifold classification. In the case of simple classification only one attribute is considered and the universes is divided into two classes-those possessing the attribute and those, which do not posses the given attribute. In the case of manifold classification two or more attributes are considered simultaneously and data are divided into a number of classes. The total number of classes would be 2.
  2. Classification according to class intervals- statistical data quantitative phenomenon like income population, age, weight, etc. are classified on the basis of class intervals. For instance, persons whose income say are within 401 to 600 can form another group and so on. The entire data may be divided into a number of groups or classes or what are usually called, ‘class interval’. We may have classes with equal class magnitudes or with unequal class magnitudes. In the classification exercise the number of classes to be used, the choice of the class limits, the frequency of each class have to be determined.

**d) Tabulation-** when a mass of data has been assembled, it becomes necessary for the researchers to arrange the same in some kind of concise and logical order. This procedure is referred to as tabulation. Tabulation is the process of summarizing raw data and displaying it in compact form (i.e in the form of statistical tables) for further analysis. It is an orderly arrangement of data in columns and rows. Tabulation provides the following advantages.

* It conserve space and reduces explanatory and descriptive statement to a minimum.
* It facilitates the process of comparison.
* It facilitates the summation of items and the detection of errors and omissions.
* It provides a basis for various statistical computations such as measures of central tendencies, dispersions, etc.

Tabulation may be done by hand or by mechanical or electronic devices such as the computer. The choice is made largely on the basis of the size and type of study, alternative costs, time pressures and the availability of computer facilities. In the case of hand tabulation one may use the direct tally method where data are directly tailed from the questionnaire, the list and tally method where data are coded from a larger worksheet or the card sort method. In the case of computer tabulation computer programs such as SPSS, Lotus, excel, STATA, etc. could be used. Tabulation may be classified as simple and complex. Simple tabulation gives information about one or more groups of independent questions. Complex tabulation shows the division of data into two or more categories.

**II) Data Analysis**

One data are edited, and coded the data presentation exercise begins. Large volume of raw statistical information need to be reduced to more manageable dimensions if one is to see the meaningful relationships in it. Data analysis is the computation of certain indices or measures. It refers to the computation of certain measures along with searching for patterns of relationship that exists among data group. In the case of survey or experimental data, analysis inovvles estimating the values of the unknown parameters of the population and testing of hypotheses for drawing inferences. Analysis may therefore, be categorized as descriptive analysis or inferential analysis (statistical analysis). With respect to the number of variables involved in the analysis, it can also be divided into uni-variate analysis and multivariate analysis.

**Uni-variate Analysis**

This refers to analysis with respect to one variable. It is also called a one-dimensional analysis. Such analysis is the study of distribution of one variable. The process of data reduction is the preserve of descriptive statistics. The initial uni-variate analysis may be the presentation of descriptive analysis in the form of frequency distributions. It provides with a profile of different groups on any of multitude of characteristics such as size, composition, efficiency, or preferences of persons or other entities. The creation of frequency distribution is a common example. The data in a frequency distribution are transferred by the calculation of a number of statistical indices, which summarizes the results even further. Measures of central tendency are examples. It can be used with nominal, interval, ordinal or ratio level data. A frequency distribution is a table in which values of variables are classified according to size. The uni-variate analysis could either be presented in the form of statistical measures such as measures of central tendencies and measure of variations or dispersions or in the form of graphs. Graphical illustrations could also be used to demonstrate the frequency distribution such graphical illustrations include, histograms, ogives, polygons, bar graphs, line graphs and circular graphs or pie charts.

Correlation analysis studies the joint variation of two or more variables for determining the amount of correlation between two or more variables. Similarly casual relationships is concerned with the study of low one or more variables affect changes in another variable. It is thus a study of functional relationships existing between two or more variables.

**Multivariate Analysis**

Multivariate analysis involves **???** considerations of two or more **varieties**. If we have two variables then we have bi-variate analysis but if we have more than two variables then we have multivariate analysis. Several multivariate analyses could be undertaken such as the construction of bi-variate tables or multivariate analysis such as multiple regression, ANOVA, discriminal analysis, caronical analysis, etc.

**Association**

Another data analysis approach involves the establishment of associations between two or more variables. A causal association can be established between two variables is unlikely to exist unless there is some tendency for a change in one variable to be associated with a change in the other. For example, income causes, a change in educational attainment.

To variables can be related in a simple linear fashion or in more complex way. Linear relationship refers a one unit change in one variable would always be associated with a certain number of values of changes in the other variables. For example, while one additional year of education is associated with an average increase of Birr 200 in a person’s annual income an increase in income from 1 year to 2 years of education **ma** produce in average increase of Birr 200 in annual income and that an increase from 12 to 13 years may produce an average increase of Birr 3000 in annual income is a complex non linear relationship.

**Temporal order**

When the cause precedes the effect we have temporal order. A change in the suspected cause must precede observed changes in the suspected effect. For example, years of schooling precedes the attainment of current income. When a change in one variable produces a change in the second, which in turn produces a change in the first and so on is an example of a reciprocal causation.

**Alternative causes**

Even if two variables are fermed to be associated and if change in one always preceders changes in the other it is possible that the one that changed first does not cause the one that changed second. Instead they are both caused by a third variable.

Example, a researcher is interested in the relationship between house size and vacation behavior, if it is found that families who live in bigger houses tend to take vacation more frequently it is not possible to prove that size of house causes vacation behavior. Both may be caused by family income. Every time a family’s income increases they may move into **larger** houses and they take vacation. So the relationship between house size and vacation behaviour is not cansal, but spumous.

It is therefore, important to test alternative explanations for an observed relation between two variables. Hence, an alternative hypothesis is presented.

**Casual Mechanism**

Another way to acquire confidence in causal hypothesis is to develop an understanding of the causal mechanism involved. If the investigator cannot think a possible causal mechanism relacing the variables, then the apparent relationship might just **be** coincrence.

Summary chart concerning analysis of data

Cathonical Analysis

Multi-ANOVA

Multiple discriminant analysis

Multiple regression

Blvanate analysis (analysis of two variables or attributes in a two way classification

Simple regression and correlation

Two way ANOVA

Association of Attributes

Calculation of several measures mostly concerning one variable

i. measures of central tendency

ii. measures of disporsion

iii. measures if skewness

iv. simple correlation

v. one way ANOVA

Non Parametric test or distributa free tests

Parametric test

Interval estimation

Point estimation

Multi-varaitilu analysis (simultaneous analysis of more than two variables of attributes in a multiway classification)

Testing hypothesis

Estimation of parameter values

Uni-dimensional analysis

Tabulation

Internal analysis/statistical

Descriptive & causal analysis

Using percentages

Classification

Coding

Editing

Analysis of data (analysis proper)

Processing of data (preparing of data for analysis)

**III) Testing Hypothesis**

Stating the hypothesis – analysis of survey data can begin by stating descriptive or explanatory hypothesis.

* Descriptive hypothesis provides proportions, frequencies or means of a single variable. For example, what proportion of the households is literate.
* Explanatory hypothesis attempts to find the causes of social phenomenon. For example, what factors determine the demand of a given item is an explanatory hypothesis.

# Testing Causal hypothesis

Testing of causal hypothesis requires four steps.

* Checking to see whether the variables are associated with one another- association.
* Verifying that changes in presumed causes precedes changes in the presumed effect (temporal order).
* Eliminating alternative explanation, and
* Generting a plausible causal mechanisms.