Statistical Analysis

Analysis means the categorizing, ordering, manipulating, and summarizing of data to obtain answers to research questions. The purpose of analysis is to reduce data to intelligible and interpretable form so that the relations of research problems can be studied and tested. A primary purpose of statistics, for example, is to manipulate and summarize numerical data and to compare the obtained results with chance expectations. A researcher hypothesizes that styles of leadership affect group-member participation in certain ways. He plans an experiment, executes the plan, and gathers data from his subjects. Then he must so order, break down, and manipulate the data that he can answer the question: How do styles of leadership affect group-member participation? It should be apparent that this view of analysis means that the categorizing, ordering, and summarizing of data should be planned early in the research. The researcher should lay out analysis paradigms or models even when working on problem and hypotheses.

Data Editing

Editing is the process of examining errors and omissions in the collected data and to make necessary corrections. Data should be edited after getting the filled up questionnaire or schedule and before entering in to the step of data processing. It is done to assure the data are accurate, consistent, uniformly entered, complete and well arranged. Editing is carried out in two stages:

  (i)   field editing
  (ii)   central editing

Field editing is the review of reporting forms by enumerator or investigator for completing what the signs and symbols have written in abbreviated form in the time of recording respondents' response. Central editing is editing obvious errors such as entry in wrong place, missing replies etc. by editor when all schedules or forms have been completed and returned to office.

Data Coding

It is process of assigning numerals or other symbols to answers so that response can be put into a limited number of class or categories. The quantitative data collected using questionnaire or schedule is numeric so that no need of coding. For the data which is qualitative in nature the numeric codes are to be used before the analysis. For the statistical treatments qualitative responses are to be converted into numerical figures which satisfy, all the rules of arithmetic operation. Different social scales are used on assigning numerical figures to the qualitative response. For example for male and female code 1 and 0 are used.

Classification of Data

The data contained in questionnaire or schedule will not enable us to see quickly all possible characteristics. In order to make data easily understandable the classification is adopted. Classification is the process of arranging the related facts or data into different groups or classes according to their similarities. Facts differ from class to class with re characteristics which is the basis of classification.

The classification should be

(i)   according to research problem
(ii)   exhaustive
(iii)   mutually exclusive
(iv)   independent

The main objectives of classification are

i)   to condense mass of data
ii)   to facilitate comparison
iii)   to pinpoint feature of data at a glance
iv)   to enable statistical treatment

Types of classification: Statistical data are classified in respect of their characteristics. Broadly there are four basic types of classification namely

a) Chronological classification
b) Geographical classification
c) Qualitative classification
d) Quantitative classification

a) **Chronological Classification**
   In chronological classification the collected data are arranged according to the order of time expressed in years, months, weeks, etc.. The data is generally classified in ascending order of time. For example, the data related with population, sales of a firm, imports and exports of a country are always subjected to chronological classification. For example, The estimates of birth rates in Nepal during 1970 - 76 are

| Year	1970	1971	1972	1973	1974	1975	1976
Birth rate	36.8	36.9	36.6	34.6	34.5	35.2	34.2

b) **Geographical Classification**

In this type of classification the data are classified according to geographical region or place. For instance, the production of wheat different countries etc.

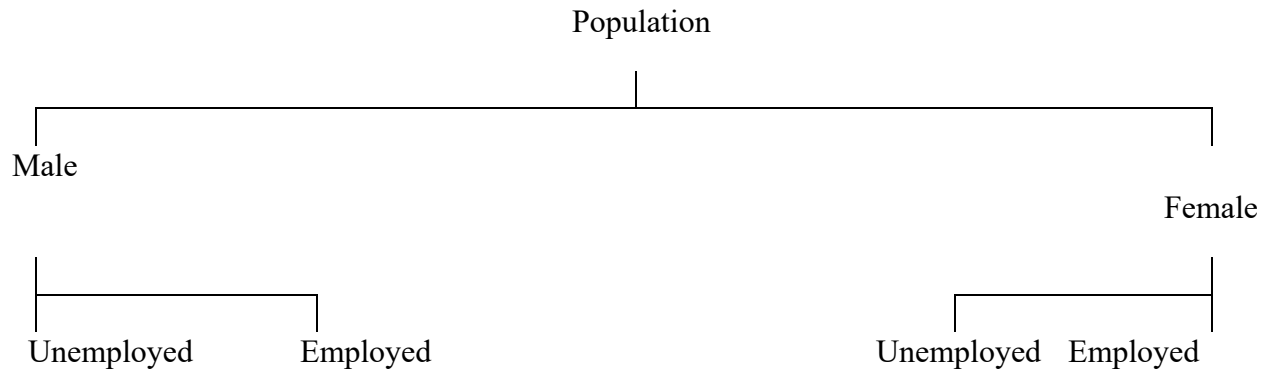| Country	America	China	Denmark	France	Nepal
Yield of wheat kg/acre	1925	893	225	439	862

c) **Qualitative Classification:**

In this type of classification data are classified on the basis of some attributes or quality like: sex, literacy, religion, employment etc. Such attributes cannot be measured along with a scale.

Population

Male   Female
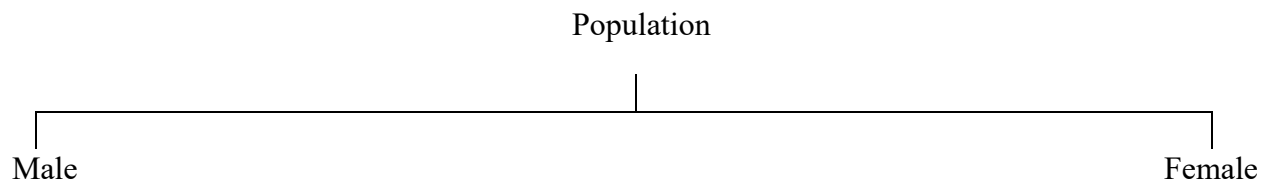
(i)   The classification, where two or more attributes are considered and several classes are formed, is called a manifold classification. For example, if we classify population simultaneously with respect to two attributes e.g., sex and employment then population are first classified with respect to 'sex' into 'males' and 'females'. Each of these classes may then be further classified into 'employment' and 'unemployment' on the basis of attribute 'employment' and as such population are classified into four classes namely. (i) Male employed (ii) Male unemployed (iii) Female employed ((iv) Female unemployed

Still the classification may be further extended by considering other attributes like marital status etc. This can be explained by the following chart

```
                                    Population
                                        |
      ┌─────────────────────────────────┴─────────────────────────────────┐
Male                                                                   Female
      |                                                                    |
  ┌───┴───┐                                                          ┌─────┴─────┐
Unemployed   Employed                                          Unemployed   Employed
```

(d) Quantitative classification

Quantitative classification refers to the classification of data according to some characteristics that can be measured such as height, weight, etc.

Data Entering into Spreadsheet

A spreadsheet is an interactive computer application program for organization and analysis of data in tabular form. Spreadsheets developed as computerized simulations of paper accounting worksheets. The program operates on data represented as cells of an array organized in rows and columns. Each cell of the array is a model-view-controller element that can contain either numeric or text data or the results of formulas that automatically calculate and display a value based on the contents of other cells. The user of the spreadsheet can make changes in any stored value and observe the effects on calculated values. This makes the spreadsheet useful for "what-if" analysis since many cases can be rapidly investigated without tedious manual recalculation. Modern spreadsheet software can have multiple interacting sheets and can display data either as text and numerals or in graphical form.

Management of Missing and Inconsistent Information

Generally, in data management activity of research work a researcher faces the threats of

i)   Missing data
ii)   Impossible values
iii)   Inconsistencies and
iv)   Transcription errors.

       Missing and inconsistent data (information) are a part of almost all research and all the researcher have to deal with it from time to time. There are various alternative ways of dealing with missing data. To get data with less missing and inconsistent information attention should be given from the designing questionnaire to data entry. Most of the

missing are arisen in the survey field because of the imperfection of the field workers and the less skill of the person who involved in data entry work.

There are a number of strategies for handling missing and inconsistent data and common will be described here. These methods can be accomplished with standard statistical software packages (SAS, SPSS).

List wise deletion

In this method, cases with any missing values are deleted from an analysis. This method is sometimes called *complete case analysis* because only cases with complete data are retained. This is the default procedure for many statistical programs but it is generally not an advisable method.

Pairwise deletion

In this method, the maximum amount of available data is retained and so this method is sometimes referred to as *available case analysi*s. Cases are excluded from only operations in missing which data are missing on a variable that is required. In a correlation matrix, for example a case that was data on one variable would not be used to calculate the correlation coefficient between that variable and another but would be included in all other correlations.

Ways of managing Missing and Inconsistent data

The quality of the data can be kept up by careful and systematic method of data cleaning. The following are the steps by using which we can reduce the inconsistent observation and problems of missing data.

- o Develop a plan for data management
- o Make a an intensive training before data entry
- o Make strategy of getting quality data
- o Maintain the question that can cross check the responses
- o Adopt checking system of impossible values
- o Record the variables and if possible create composite variables
- o Preferably do not make a change in raw data set, if changed logically it should be documented
- o Coding system should be preferably used to reduce such errors
- o Use standard methods of data cleaning using software
- o Use the method of labeling the values and merge cells if possible
- o Once the data set is cleaned, the next step is to format it for analysis
- o Data formatting should be done using code
- o Maintain a master dataset that is distributed to everyone conducting analyses

Descriptive Statistical measure

Types of average

   I.   Arithmetic Mean
  II.   Geometric mean
 III.   Harmonic mean
 IV.   Median
  V.   Mode

Measure of Dispersion

Absolute and relative measure of dispersion

Types of dispersion

Range

Quartile deviation
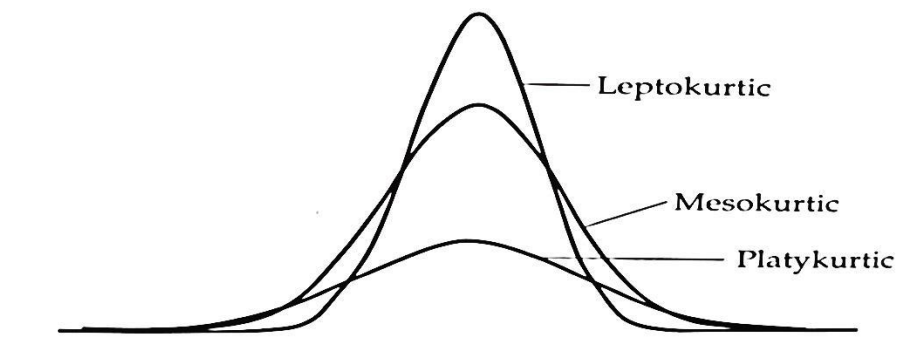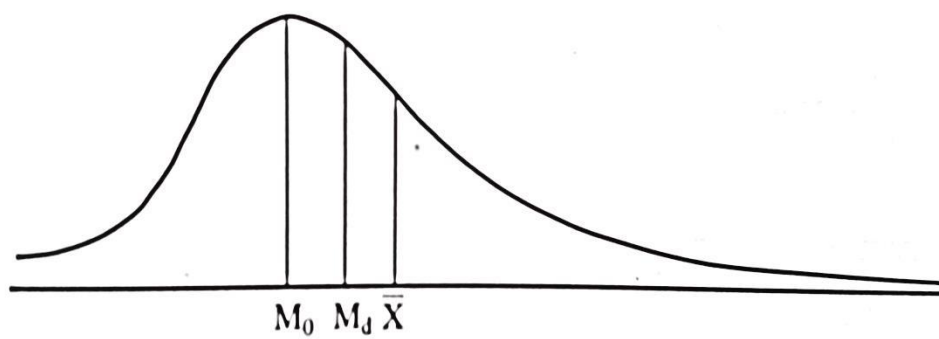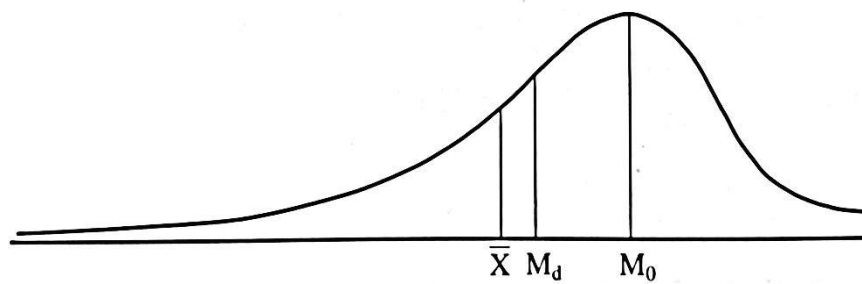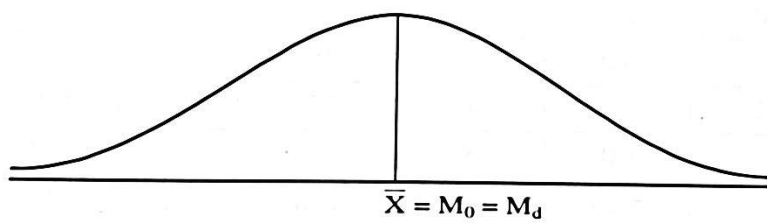
Mean deviation

Standard deviation

(Coefficient of Variation)

Skewness   **kurtosis K**$=\dfrac{Q_3-Q_1}{2(P_{90}-P_{10}}$ **K=0.263**

$$\overline{X} = M_0 = M_d$$

$$\overline{X} \quad M_d \quad M_0$$

$$M_0 \quad M_d \quad \overline{X}$$

Leptokurtic

Mesokurtic

Platykurtic

Correlation and Regression