2. DESCRIPTIVE STATISTICS

2.1 Some terms in use

Population - The term "population" is used in statistics to represent all possible measurements or outcomes that are of interest to us in a particular study.

Sample - The term "sample" refers to a portion of the population that is representative of the population from which it was selected.

Parameter - A parameter is a numerical characteristic of a given population. i.e. the parameter tells us something about the whole population.

Statistic - A statistic is a characteristic of a sample. A statistic can be used to estimate the value of a population parameter.

Element – A unit in the population or a sample which provides data.

Example:
The population for a study of infant health is all the children born in the Sri Lanka in the 1990s. The sample might be all babies born on 7th May in any of the years. Average height at birth of a child born in the Sri Lanka in the 1990s is a *parameter*. Average weight at birth of all babies born on 7th May in any of the years is a *statistic*.

2.2 Collection of Data

The first step in any enquiry (investigation) is collection of data. The data may be collected for the whole population or for a sample only. It is mostly collected on sample basis. The enumerator or investigator is the well trained person who collects the statistical data. The respondents (information) are the people whom the information is collected from.

- **Census**. A census is a study that obtains data from every member of a population. In most studies, a census is not practical, because of the cost and/or time required.

- **Sample survey**. A sample survey is a study that obtains data from a subset of a population, in order to estimate population attributes.

- **Experimental study**. An experiment is a controlled study in which the researcher attempts to understand cause-and-effect relationships. The study is "controlled" in the sense that the researcher controls (1) how subjects are assigned to groups and (2) which treatments each group receives.

In the analysis phase, the researcher compares group scores on some dependent variable. Based on the analysis, the researcher draws a conclusion about whether the treatment (independent variable) had a causal effect on the dependent variable.

- **Observational study**. Like experiments, observational studies attempt to understand cause-and-effect relationships. However, unlike experiments, the researcher is not able to control (1) how subjects are assigned to groups and/or (2) which treatments each group receives.

2.3. Types of Data:

There are two types (sources) for the collection of data.

- Primary Data
- Secondary Data

2.3.1 Primary Data

The primary data are the first hand information collected, compiled and published by organization for some purpose. They are most original data in character and have not undergone any sort of statistical treatment.

Example**:** Population census reports are primary data because these are collected, complied and published by the population census organization.

2.3.1.1 Methods of Collecting Primary Data

Primary data are collected by the following methods:

- **Collection through Questionnaire:** The researchers get the data from local representation or agents that are based upon their own experience. This method is quick but gives only rough estimate. It can be posted or e-mailed.

Advantages

- Can be used as a method in its own right or as a basis for interviewing or a telephone survey.
- Can cover a large number of people or organizations.
- Relatively cheap.
- No interviewer bias (in general).
- Possible anonymity of respondent.

Disadvantages

- Design problems.
- Historically low response rate and high time consumption.
- No control over who completes it or total completion.
- Not possible to give assistance if required

Designing

- Theme and covering letter
- Instructions for completion
- Appearance, Length and order
- Question types and proper designs
- Thanking and giving complements

A pilot survey is needed before distribute the finalized questionnaire.

- **Interviewing**
  - **Personal**: The researcher conducts the survey him/herself and collects data from it. The data collected in this way is usually accurate and reliable. This method of collecting data is only applicable in case of small projects.

Advantages

- Good response rate.
- Completed and immediate.
- Possible in-depth questions and investigate motives and feelings.
- If one interviewer used, uniformity of approach.
- Used to pilot other methods.

Disadvantages

- Need to set up interviews.
- Time consuming and expensive
- Respondent bias – tendency to please or impress, create false personal image
- Embarrassment possible if personal questions are sensitive in nature.
- If many interviewers, training required.

- **Through Telephone:** The researchers get information through telephone this method is quick and give accurate information.
- **Through investigators:** Trained investigators are employed to collect the data. These investigators contact the individuals and fill in questionnaire after asking the required information. Most of the organizing implied this method.

Interviews can be get the form of **Structured, Semi-structured or Unstructured** according to the requirement and the available sources.

Planning an interview

- List the areas in which you require information.
- Decide on type of interview.
- Transform areas into actual questions.
- Try them out on a friend or relative.
- Make an appointment with respondent(s)

- **Focus groups interviews:** Here, a small group of individuals join to talk about the problem.

- **Observation method**: This method lets one to assess the dynamics of a situation. This is a systematic way of data collection. Researchers make use of all their senses to evaluate people in naturally occurring situations.

  Types of Observation method
  - Structured or unstructured – observe either specified area or all the area
  - Disguised or Undisguised – respondents are unaware or aware of being observed
  - Natural or Contrived – observe at a natural or an artificial environment
  - Participant or Non-participant – researcher is part of observers or not

- **Case-studies:** Intensive examination of a single unit such as a person, a small group of people, or a single company. Case studies involve measuring what is there and how it got there.

2.3.2 Secondary Data

The secondary data are the second hand information which are already collected by someone (organization) for some purpose and are available for the present study. The secondary data are not pure in character and have undergone some treatment at least once.
Example**:** Doing a study based on the data from the central bank report.

Methods of Collecting Secondary Data

The secondary data are collected by the following sources:

- **Official:** e.g. The publications of the Statistical Division, Ministry of Finance, the Federal Bureaus of Statistics, Ministries of Food, Agriculture, Industry, Labor etc…

- Publication of Trade Associations, Chambers of Commerce etc…

- Technical and Trade Journals and Newspapers.

- Research Organizations such as Universities and other institutions.

2.4 Editing of Data

After collecting the data either from primary or secondary source, the next step is its editing. Editing means the examination of collected data to discover any error and mistake before presenting it. It has to be decided before hand what degree of accuracy is wanted and what extent of errors can be tolerated in the inquiry. The editing of secondary data is simpler than that of primary data.

To derive conclusions from data, we need to know how the data were collected; that is, we need to know the method(s) of data collection.

2.5 Presentation of Data

The collected data in any statistical investigation are known as raw data. Usually when data is collected there are a lot of numbers, results, responses, etc. They cannot be easily understood by people and are not fit for further analysis and interpretation. In fact, there is usually so much data that it needs to be summarized before it can be analyzed. Hence after having collected and edited data, the next important step is present the data in a systematic manner. This lesson describes some commonly used tools for presenting categorical data.

2.5.1 Tabulation of Data

By tabulation we mean, a systematic presentation of numerical data in columns and rows in accordance with some salient feature of characteristics.

The main objectives of tabulation are:

1. To clarify the object of investigation
2. To simplify complex data
3. To clarify the characteristics of data
4. To present facts in the minimum of space
5. To facilitate comparison
6. To detect errors and omission in the data
7. To depict trend and tendencies of the problem under consideration
8. To facilitate statistical processing
9. To help reference

Example 1: Present the following information in a suitable tabular form. In 2010, out of 4000 workers in a factory, 2500 were members of a trade union X. The number of women workers employed was 250, out of that 200 did not belong to the trade union. In 2011, the number of union workers was 2650 of which 2600 were men. The number of non-union workers was 380, among whom 155 were women.

Solution: Comparative study of the membership of Trade Union X in 2010 and 2011

Year	**2010**			**2011**		
	Males	**Females**	**Total**	**Males**	**Females**	**Total**
Members						
Non Members						
Total						

Source: Annual Report of Union X

2.5.2 Diagrammatic Presentation

A picture is indeed worth a thousand words, we hope to be able to detect patterns or be able to draw conclusions once we see data represented graphically. A diagram is a visual form for presentation of statistical data. Diagram refers to the various types of devices such as bars, circles, maps, pictorials, cartograms, etc.

General Rules

While graphing statistical data, the following guidelines may be kept in mind.

1. Every graph must have a title, indicating the facts presented by the graph.
2. It is necessary to plot the independent variables on the horizontal axis and dependent variables on the vertical axis.
3. Problem arises regarding the choice of a suitable scale. The choice must accommodate the whole data.
4. The principle of drawing graphs is that the vertical scale must start from zero. If the fluctuations are quite small compared to the size of variables, there is no need of showing the entire vertical scale from the origin. The scale just sufficient for the purpose need to be shown and for this purpose a false line may be used. The portion of the scale which lies between zero and the smallest variable is omitted, by drawing two horizontal lines.
5. The graph must not be overcrowded with curves.
6. If more than one variable is plotted on the same graph, it its necessary to distinguish them by different lines, dotted lines, broken lines, dots, dot-cum-dash, thick, thin, dashed-lines etc.
7. Index should be given to show the scales and the meaning of different curves.
8. All lettering must be horizontal.
9. It should be remembered that for every value of independent variable, there is corresponding value of the dependent variable. It is these matched values (pair of value) that are to be plotted. Each pair of value is represented on the graph by a point.
10. Source of information should be mentioned as footnote.

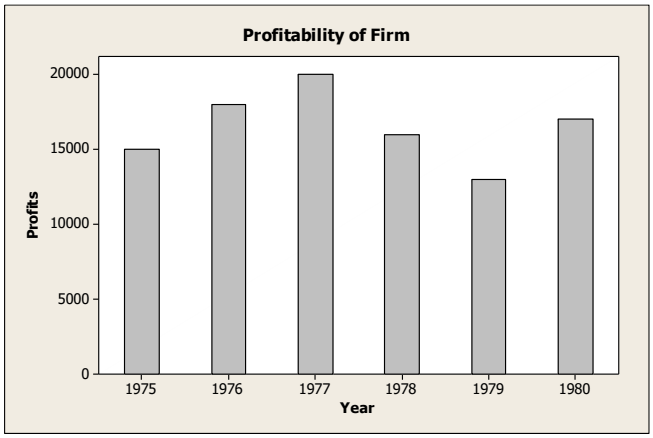
Types of Diagrams

2.5.2.1 Simple Bar Diagram

A simple bar diagram can be drawn either on horizontal or vertical base. Bars on horizontal base are more common. A bar diagram is simple to draw and easy to understand. In business and economics it is commonly used.

Example 2:Draw a suitable bar diagram showing the following data.

Year	1975	1976	1977	1978	1979	1980
Profits Rs('000)	15000	18000	20000	16000	13000	17000



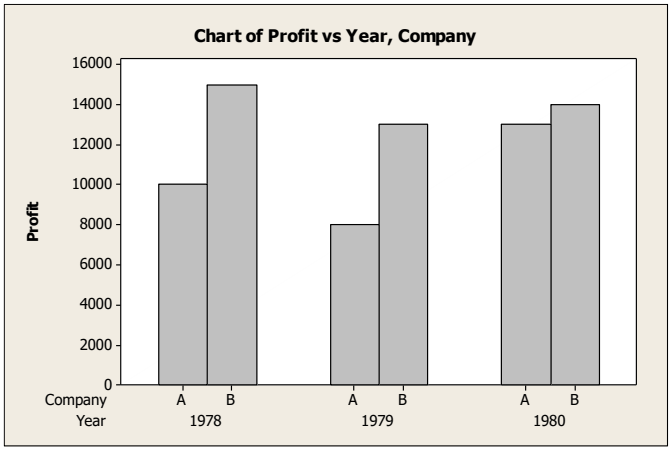
2.5.2.2 Multiple Bar Diagram

Multiple bar diagrams are used to denote more than one phenomenon, e.g., for import and export trend. Multiple bars are useful for direct comparison between two values. The bars are drawn side by side. In order to distinguish the bars, different colors, shades etc.,, may be used and a key index to this effect be given to understand the different bars.

Example 3

Draw a suitable bar diagram showing the following, data.

Year	**Profit (Rs)**	
	Company A	**Company B**
1978	10000	15000
1979	8000	13000
1980	13000	14000



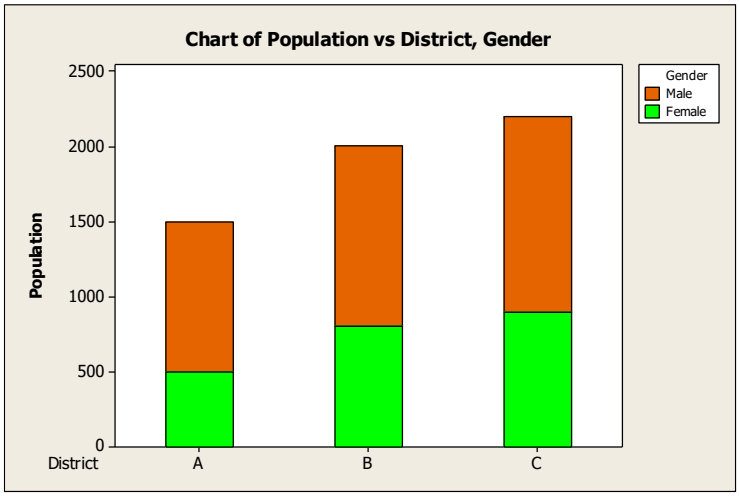
2.5.2.3 Sub-divided Bar Diagram (Component Bar Diagram)

The bar is subdivided into various parts in proportion to the values given in the data and may be drawn on absolute figures or percentages. Each component occupies a part of the bar proportional to its share in the total.

Example 4

Represent the following data in a suitable diagram.

District	A	B	C
Male	1000	1200	1300
Female	500	800	900
Total	1500	2000	2200



2.5.2.4 Percentage Sub-divided Bar Diagram

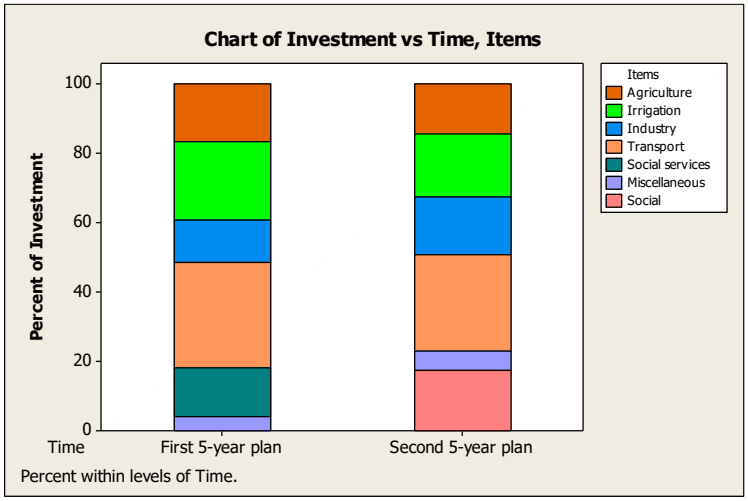
The above mentioned diagrams have been used to represent absolute value. But comparison is made on a relative basis. The various components are expressed as percentage to the total. For dividing the bars these percentages are cumulated. In this case, the bars are all of equal height. Each segment shows the percentage to the total.

Example 5

Represent by a percentage bar diagram the following data on investment for the First and Second Five-year plans.

Investment in the Public Sector

Items	First 5-year plan	Second 5-year plan
Agriculture	357	768
Irrigation	492	990
Industry	261	909
Transport	654	1495
Social services	306	945
Miscellaneous	90	300

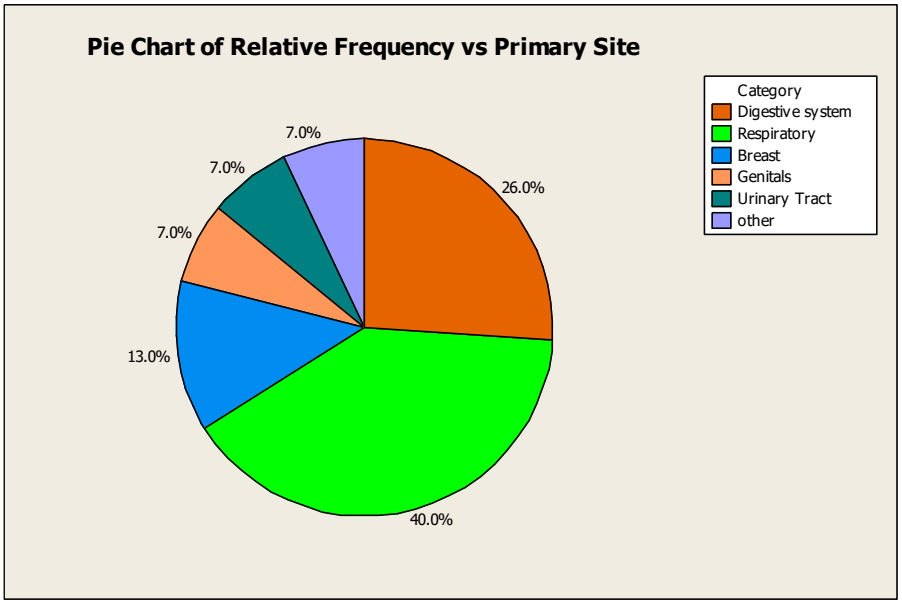


2.5.2.5 Pie Diagram

A pie chart is a simple descriptive display of data that sum to a given total. A pie chart is probably the most illustrative way of displaying quantities as percentages of given total. The pie diagram ranks high in understanding and be careful that you do not have too many slices in that pie, or they will become meaningless. Also, note that a pie chart is usually used as a snapshot of ONE moment in time.

Example 6

Primary Site	Relative Frequency	Angle size
Digestive system	0.26	360 x 0.26 = 93.6°
Respiratory	0.40	360 x 0.40 = 144°
Breast	0.13	360 x 0.13 = 46.8°
Genitals	0.07	360 x 0.07 = 25.2°
Urinary Tract	0.07	360 x 0.07 = 25.2°
Other	0.07	360 x 0.07 = 25.2°



2.5.2.6 Line Graph

When only one variable is to be represented graphically, we use the line graph. Line charts may also be used to show how the value of a variable changes over time. Unlike bar and column charts, line charts imply continuous change rather than a number of discrete points. The

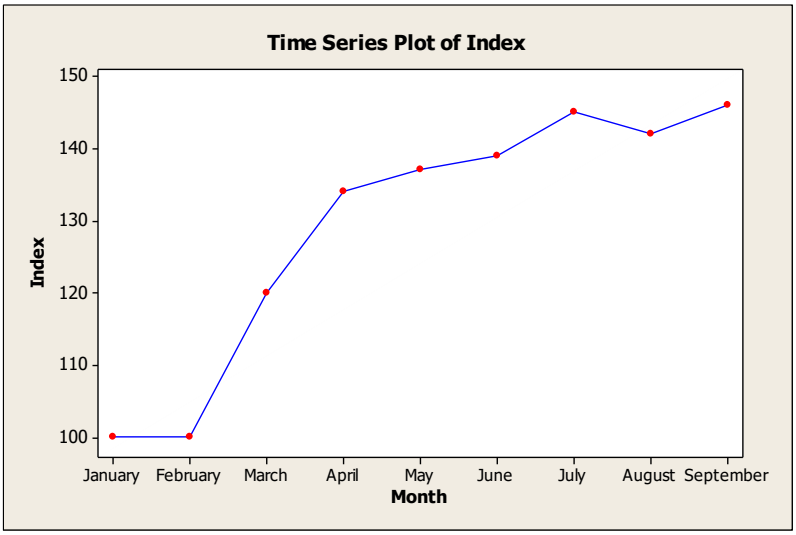
plotted peaks and dips on the grid allow you to monitor and compare improvement and decline. For this reason, line charts are better at implying a trend.

Be careful when interpreting such charts that you do not automatically assume intermediate values by the line placement.

Example 7

Represent the following data by a line graph.

Month	**January**	**February**	**March**	**April**	**May**	**June**	**July**	**August**	**September**
Index	100	100	120	134	137	139	145	142	146



When independent variable relating to two or more related variables, all the variables are shown on the same graph. When two or more variables are shown on the same graph, they are distinguished from one another by drawing dotted-dashes or different types of lines-thin, thick, dotted lines, broken etc.