**UNIT - 1**

**STATISTICS**

**INTRODUCTION**

 For a layman, ‘Statistics’ means numerical information expressed in quantitative terms.

This information may relate to objects, subjects, activities, phenomena, or regions of space. As a matter of fact, data have no limits as to their reference, coverage, and scope. At the macro level, these are data on gross national product and shares of agriculture, manufacturing, and services in GDP (Gross Domestic Product). At the micro level, individual firms, howsoever small or large, produce extensive statistics on their operations. The annual reports of companies contain variety of data on sales, production, expenditure, inventories, capital employed, and other activities. These data are often field data, collected by employing scientific survey techniques. Unless regularly updated, such data are the product of a one-time effort and have limited use beyond the situation that may have called for their collection. A student knows statistics more intimately as a subject of study like economics, mathematics, chemistry, physics, and others. It is a discipline, which scientifically deals with data, and is often described as the science of data. In dealing with statistics as data, statistics has developed appropriate methods of collecting, presenting, summarizing, and analysing data, and thus consists of a body of these methods.

 **MEANING AND DEFINITIONS OF STATISTICS**

In the beginning, it may be noted that the word ‘statistics’ is used rather curiously in two senses plural and singular. In the plural sense, it refers to a set of figures or data. In the singular sense, statistics refers to the whole body of tools that are used to collect data, organise and interpret them and, finally, to draw conclusions from them. It should be noted that both the aspects of statistics are important if the quantitative data are to serve their purpose. If statistics, as a subject, is inadequate and consists of poor methodology, we could not know the right procedure to extract from the data the information they contain. Similarly, if our data are defective or that they are inadequate or inaccurate, we could not reach the right conclusions even though our subject is well developed.

 A.L. Bowley has defined statistics as: (i) statistics is the science of counting, (ii) Statistics may rightly be called the science of averages, and (iii) statistics is the science of measurement of social organism regarded as a whole in all its mani-festations.

 Boddington defined as: “ Statistics is the science of estimates and probabilities”.

 Further, W.I. King has defined Statistics in a wider context, the science of Statistics is the method of judging collective, natural or social phenomena from the results obtained by the analysis or enumeration or collection of estimates.

According to Prof. Horace Secrist, Statistics is the aggregate of facts, affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose, and placed in relation to each other.

 From the above definitions, we can highlight the major characteristics of statistics as follows:

 (i) Statistics are the aggregates of facts. It means a single figure is not statistics. For example, national income of a country for a single year is not statistics but the same for two or more years is statistics.

 (ii) Statistics are affected by a number of factors.

 For example, sale of a product depends on a number of factors such as its price, quality, competition, the income of the consumers, and so on.

(iii) Statistics must be reasonably accurate. Wrong figures, if analysed, will lead to erroneous conclusions. Hence, it is necessary that conclusions must be based on accurate figures.

 (iv) Statistics must be collected in a systematic manner. If data are collected in a haphazard manner, they will not be reliable and will lead to misleading conclusions.

(v) Collected in a systematic manner for a pre-determined purpose .

(vi) Lastly, Statistics should be placed in relation to each other. If one collects data unrelated to each other, then such data will be confusing and will not lead to any logical conclusions. Data should be comparable over time and over space.

**TYPES OF DATA AND DATA SOURCES**

Statistical data are the basic raw material of statistics. Data may relate to an activity of our interest, a phenomenon, or a problem situation under study. They derive as a result of the process of measuring, counting and/or observing. Statistical data, therefore, refer to those aspects of a problem situation that can be measured, quantified, counted, or classified. Any object subject phenomenon, or activity that generates data through this process is termed as a variable.

In other words, a variable is one that shows a degree of variability when successive measurements are recorded. In statistics, data are classified into two broad categories: quantitative data and qualitative data. This classification is based on the kind of characteristics that are measured.

**Quantitative data** are those that can be quantified in definite units of measurement. These refer to characteristics whose successive measurements yield quantifiable observations. Depending on the nature of the variable observed for measurement, quantitative data can be further categorized as continuous and discrete data. Obviously, a variable may be a continuous variable or a discrete variable.

(i) Continuous data represent the numerical values of a continuous variable. A continuous variable is the one that can assume any value between any two points on a line segment, thus representing an interval of values. The values are quite precise and close to each other, yet distinguishably different. All characteristics such as weight, length, height, thickness, velocity, temperature, tensile strength, etc., represent continuous variables. Thus, the data recorded on these and similar other characteristics are called continuous data. It may be noted that a continuous variable assumes the finest unit of measurement. Finest in the sense that it enables measurements to the maximum degree of precision.

(ii) Discrete data are the values assumed by a discrete variable. A discrete variable is the one whose outcomes are measured in fixed numbers. Such data are essentially count data. These are derived from a process of counting, such as the number of items possessing or not possessing a certain characteristic. The number of customers visiting a departmental store everyday, the incoming flights at an airport, and the defective items in a consignment received for sale, are all examples of discrete data.

**Qualitative data** refer to qualitative characteristics of a subject or an object. A characteristic is qualitative in nature when its observations are defined and noted in terms of the presence or absence of a certain attribute in discrete numbers. These data are further classified as nominal and rank data.

 (i) Nominal data are the outcome of classification into two or more categories of items or units comprising a sample or a population according to some quality characteristic.

Classification of students according to sex (as males females), of workers according to skill (as skilled, semi-skilled, and unskilled), and of employees according to the level of education (as matriculates, undergraduates, and post-graduates), all result into nominal data. Given any such basis of classification, it is always possible to assign each item to a particular class and make a summation of items belonging to each class. The count data so obtained are called nominal data.

 (ii) Rank data, on the other hand, are the result of assigning ranks to specify order in terms of the integers 1,2,3, ..., n. Ranks may be assigned according to the level of performance in a test. a contest, a competition, an interview, or a show. The candidates appearing in an interview, for example, may be assigned ranks in integers ranging from I to n, depending on their performance in the interview. Ranks so assigned can be viewed as the continuous values of a variable involving performance as the quality characteristic.

 Data sources could be seen as of two types, viz., secondary and primary. The two can be defined as under:

 **(i) Secondary data**: They already exist in some form: published or unpublished - in an identifiable secondary source. They are, generally, available from published source(s), though not necessarily in the form actually required.

 **(ii) Primary data:** Those data which do not already exist in any form, and thus have to be collected for the first time from the primary source(s). By their very nature, these data require fresh and first-time collection covering the whole population or a sample drawn from it.

**TYPES OF STATISTICS**

There are two major divisions of statistics such as descriptive statistics and inferential statistics. The term descriptive statistics deals with collecting, summarizing, and simplifying data, which are otherwise quite unwieldy and voluminous. It seeks to achieve this in a manner that meaningful conclusions can be readily drawn from the data.

Descriptive statistics may thus be seen as comprising methods of bringing out and highlighting the latent characteristics present in a set of numerical data. It not only facilitates an understanding of the data and systematic reporting thereof in a manner; and also makes them amenable to further discussion, analysis, and interpretations.

**SCOPE OF STATISTICS**

Apart from the methods comprising the scope of descriptive and inferential branches of statistics, statistics also consists of methods of dealing with a few other issues of specific nature. Since these methods are essentially descriptive in nature, they have been discussed here as part of the descriptive statistics.

 These are mainly concerned with the following:

 (i) It often becomes necessary to examine how two paired data sets are related. For example, we may have data on the sales of a product and the expenditure incurred on its advertisement for a specified number of years. Given that sales and advertisement expenditure are related to each other, it is useful to examine the nature of relationship between the two and quantify the degree of that relationship. As this requires use of appropriate statistical methods, these falls under the purview of what we call regression and correlation analysis.

 (ii) Situations occur quite often when we require averaging (or totalling) of data on prices and/or quantities expressed in different units of measurement. For example, price of cloth may be quoted per meter of length and that of wheat per kilogram of weight. Since ordinary methods of totalling and averaging do not apply to such price/quantity data, special techniques needed for the purpose are developed under index numbers.

 (iii) Many a time, it becomes necessary to examine the past performance of an activity with a view to determining its future behaviour. For example, when engaged in the production of a commodity, monthly product sales are an important measure of evaluating performance. This requires compilation and analysis of relevant sales data over time.

 The more complex the activity, the more varied the data requirements. For profit maximising and future sales planning, forecast of likely sales growth rate is crucial. This needs careful collection and analysis of past sales data. All such concerns are taken care of under time series analysis.

 (iv) Obtaining the most likely future estimates on any aspect(s) relating to a business or economic activity has indeed been engaging the minds of all concerned. This is particularly important when it relates to product sales and demand, which serve the necessary basis of production scheduling and planning. The regression, correlation, and time series analyses together help develop the basic methodology to do the needful. Thus, the study of methods and techniques of obtaining the likely estimates on business/economic variables comprises the scope of what we do under business forecasting. Keeping in view the importance of inferential statistics, the scope of statistics may finally be restated as consisting of statistical methods which facilitate decision-- making under conditions of uncertainty. While the term statistical methods is often used to cover the subject of statistics as a whole, in particular it refers to methods by which statistical data are analysed, interpreted, and the inferences drawn for decisionmaking.

**IMPORTANCE OF STATISTICS IN BUSINESS**

There are three major functions in any business enterprise in which the statistical methods are useful. These are as follows:

(i) The planning of operations: This may relate to either special projects or to the recurring activities of a firm over a specified period.

(ii) The setting up of standards: This may relate to the size of employment, volume of sales, fixation of quality norms for the manufactured product, norms for the daily output, and so forth.

 (iii) The function of control: This involves comparison of actual production achieved against the norm or target set earlier. In case the production has fallen short of the target, it gives remedial measures so that such a deficiency does not occur again.

**LIMITATIONS OF STATISTICS**

Statistics has a number of limitations, pertinent among them are as follows:

 (i) There are certain phenomena or concepts where statistics cannot be used. This is because these phenomena or concepts are not amenable to measurement. For example, beauty, intelligence, courage cannot be quantified. Statistics has no place in all such cases where quantification is not possible.

 (ii) Statistics reveal the average behaviour, the normal or the general trend. An application of the 'average' concept if applied to an individual or a particular situation may lead to a wrong conclusion and sometimes may be disastrous. For example, one may be misguided when told that the average depth of a river from one bank to the other is four feet, when there may be some points in between where its depth is far more than four feet. On this understanding, one may enter those points having greater depth, which may be hazardous.

 (iii) Since statistics are collected for a particular purpose, such data may not be relevant or useful in other situations or cases. For example, secondary data (i.e., data originally collected by someone else) may not be useful for the other person.

 (iv) Statistics are not 100 per cent precise as is Mathematics or Accountancy. Those who use statistics should be aware of this limitation.

(v) In statistical surveys, sampling is generally used as it is not physically possible to cover all the units or elements comprising the universe. The results may not be appropriate as far as the universe is concerned. Moreover, different surveys based on the same size of sample but different sample units may yield different results.

(vi) At times, association or relationship between two or more variables is studied in statistics, but such a relationship does not indicate cause and effect' relationship. It simply shows the similarity or dissimilarity in the movement of the two variables. In such cases, it is the user who has to interpret the results carefully, pointing out the type of relationship obtained.

(vii) A major limitation of statistics is that it does not reveal all pertaining to a certain phenomenon. There is some background information that statistics does not cover. Similarly, there are some other aspects related to the problem on hand, which are also not covered. The user of Statistics has to be well informed and should interpret Statistics keeping in mind all other aspects having relevance on the given problem.

**Types of Data**

Data can be classified broadly on the basis of the following 4 criteria:

1. Geographical i.e. area-wise e.g. cities, districts e.t.c

2. Chronological i.e. on the basis of time

3. Qualitative i.e. according to some attributes

4. Quantitative i.e. in terms of magnitude

**Graphical data**

Data are classified on the basis of geographical or vocational differences between the various items like states, cities, regions, zones, areas e.t.c.

**Chronological data**

When data are observed over a period of time the type of classification is known as chronological classification e.g. we may present the figures of population (or production, sales …) as follows:

Population of Kenya from 1969 to 1988(hypothetical figures) .

Time series are usually listed in chronological order; normally starting with earliest period..

**Qualitative data**

Data are classified on the basis of some attribute or quality such as sex, colour of hair, literacy, religion e.t.c.

The attribute under study cannot be measured; one can only find out whether it is present or absent in the units of the population under study.

E.g. if the attribute under study is population, one can find out how many persons are living in urban areas and how many in rural areas.

**Quantitative Classification**

Under this type of classification, the collected data are classified on the basis of certain variable viz. mark, income, expenditure, profit, loss, height, weight, age, price, production etc. which is capable of quantitative is also otherwise known as ‘classification by variables’.

**Characteristics of classification**

 a) Classification performs homogeneous grouping of data

b) It brings out points of similarity and dissimilarities.

c) The classification may be either real or imaginary.

d) Classification is flexible to accommodate adjustments.

**Objectives / purposes of classifications**

 i) To simplify and condense the large data

 ii) To present the facts to easily in understandable form

 iii) To allow comparisons

iv) To help to draw valid inferences

v) To relate the variables among the data

vi) To help further analysis

vii) To eliminate unwanted data

viii) To prepare tabulation

**Tabulation of data**

**Major Objectives of Tabulation**

1. To simplify the complex data

2. To facilitate comparison

3. To economise the space

4. To draw valid inference / conclusions

5. To help for further analysis

**Differences between Classification and Tabulation**

1. First data are classified and presented in tables; classification is the basis for tabulation.

 2. Tabulation is a mechanical function of classification because is tabulation classified data are placed in row and columns.

3. Classification is a process of statistical analysis while tabulation is a process of presenting data is suitable structure.

**Frequency Distribution**

Frequency distribution is a table used to organize the data. The left column (called classes or groups) includes numerical intervals on a variable under study. The right column contains the list of frequencies, or number of occurrences of each class/group. Intervals are normally of equal size covering the sample observations range. It is simply a table in which the gathered data are grouped into classes and the number of occurrences, which fall in each class, is recorded.

 **Definition**♦ A frequency distribution is a statistical table which shows the set of all distinct values of the variable arranged in order of magnitude, either individually or in groups with their corresponding frequencies. - Croxton and Cowden

 A frequency distribution can be classified as

 a) Series of individual observation

b) Discrete frequency distribution

c) Continuous frequency distribution

a) **Series of individual observation**

 Series of individual observation is a series where the items are listed one after the each observation. For statistical calculations, these observation could be arranged is either ascending or descending order. This is called as array.

|  |  |
| --- | --- |
| **Roll number** | **Marks obtained in english** |
| 1 | **23** |
| **2** | **25** |
| **3** | **14** |
| **4** | **29** |

The above data list is a raw data. The presentation of data in above form doesn‟t reveal any information. If the data is arranged in ascending / descending in the order of their magnitude, which gives better presentation then, it is called arraying of data.

**Discrete (ungrouped) Frequency Distribution**

 If the data series are presented in such away that indicating its exact measurement of units, then it is called as discrete frequency distribution. Discrete variable is one where the variants differ from each other by definite amounts.

 Ex: Assume that a survey has been made to know number of post-graduates in 10 families at random; the resulted raw data could be as follows.

 0, 1, 3, 1, 0, 2, 2, 2, 2, 4

|  |  |
| --- | --- |
| No. of post graduates (x) | Frequency(f) |
| 0 | 12 |
| 1 | 1 |
| 2 | 13 |
| 3 | 15 |
| 4 | 9 |

 This data can be classified into an ungrouped frequency distribution.

The number of post-graduates becomes variable (x) for which we can list the frequency of occurrence (f) in a tabular from .

**Continuous frequency distribution (grouped frequency distribution)**

Continuous data series is one where the measurements are only approximations and are expressed in class intervals within certain limits. In continuous frequency distribution the class interval theoretically continuous from the starting of the frequency distribution till the end without break.

 According to Boddington „the variable which can take very intermediate value between the smallest and largest value in the distribution is a continuous frequency distribution.

 Ex: Marks obtained by 20 students in students‟ exam for 50 marks are as given below – convert the data into continuous frequency distribution form

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 18 | 23 | 28 | 29 | 44 | 15 | 32 | 24 | 17 | 19 |
| 24 | 29 | 19 | 39 | 42 | 2 | 33 | 28 | 29 | 20 |

. By grouping the marks into class interval of 10 following frequency distribution tables can be formed.

|  |  |
| --- | --- |
| Marks  | No. of students |
| 18 | 24 |
| 23 | 29 |
| 28 | 1 |
| 29 | 39 |
| 44 | 42 |
| 15 | 2 |
| 32 | 33 |
| 24 | 28 |
| 17 | 29 |
| 19 | 20 |

### Significance of Diagrams and Graphs:

Diagrams and graphs are extremely useful because of the following reasons.

1. They are attractive and impressive.
2. They make data simple and intelligible.
3. They make comparison possible
4. They save time and labour.
5. They have universal utility.
6. They give more information.
7. They have a great memorizing effect.

### ****Types of Diagrams and Graphs****

One of the most effective and interesting alternative ways to present statistical data is through diagrams and graphs. There are several ways in which statistical data may be displayed pictorially, such as different types of graphs and diagrams. The most commonly used diagrams and graphs shall be discussed in subsequent posts, as listed below.

#### ****Types of Diagrams/Charts****

1. Simple Bar Chart
2. Multiple Bar Chart or Cluster Chart
3. Stacked Bar Chart or Sub-Divided Bar Chart or Component Bar Chart
	* Simple Component Bar Chart
	* Percentage Component Bar Chart
	* Sub-Divided Rectangular Bar Chart
	* Pie Chart

#### ****Types of Diagrams/Charts****

1. Histogram
2. Frequency Curve and Polygon
3. Lorenz Curve

**Bar charts (bars horizontal or vertical)**

Most appropriate use: to compare categories (qualitative data, the independent variable is non-numerical) and grouped discrete quantitative data (scores on a test, amount spend by customers in a shop)

How to draw: Rectangles with equal width are used. The height/length represents the frequency of the category. Do not draw the bar adjacent. Label the diagram as a whole (title), the bars and the frequency axis. Indicate scale on the frequency axis.



Examples of qualitative data display.

**PIE CHART**

A pie chart is a circular graph that shows the relative contribution that different categories contribute to an overall total. A wedge of the circle represents each category’s contribution, such that the graph resembles a pie that has been cut into different sized slices. Every 1% contribution that a category contributes to the total corresponds to a slice with an angle of 3.6 degrees.



**Histograms**

 Most appropriate use: to represent grouped continuous variables. Always depicts frequency (or count) versus a continuous or nearly continuous variable.

How to draw: Rectangles whose areas are proportional to the frequencies. The rectangles are adjacent (that is, the rectangles touch each other.) The axes are labelled, the graph has a title.

****

### Frequency Polygons

Another type of graph that can be drawn to represent the same set of data as a histogram represents is a **frequency polygon**. A **frequency polygon** is a graph constructed by using lines to join the midpoints of each interval, or bin. The heights of the points represent the frequencies. A frequency polygon can be created from the histogram or by calculating the midpoints of the bins from the frequency distribution table. The **midpoint** of a bin is calculated by adding the upper and lower boundary values of the bin and dividing the sum by 2.



# Ogive Graph / Cumulative Frequency Polygon



An ogive (oh-jive), sometimes called a cumulative frequency polygon, is a type of [frequency polygon](https://www.statisticshowto.datasciencecentral.com/frequency-polygon/) that shows [cumulative frequencies](https://www.statisticshowto.datasciencecentral.com/cumulative-frequency-distribution/). In other words, the cumulative percents are added on the graph from left to right.

An ogive graph plots **cumulative frequency** on the y-axis and **class boundaries**along the x-axis. It’s very similar to a [histogram](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/descriptive-statistics/histogram-make-chart/), only instead of rectangles, an ogive has a single point marking where the top right of the rectangle would be. It is usually easier to create this kind of graph from a frequency table.

## How to Draw an Ogive Graph

**Example question:** Draw an Ogive graph for the following set of data:
02, 07, 16, 21, 31, 03, 08, 17, 21, 55
03, 13, 18, 22, 55, 04, 14, 19, 25, 57
06, 15, 20, 29, 58.

Step 1: Make a [relative frequency table](https://www.statisticshowto.datasciencecentral.com/relative-frequency-distribution/) from the data. The first column has the class limits, the second column has the frequency (the count) and the third column has the relative frequency (class frequency / total number of items):


Step 2: Add a fourth column and cumulate (add up) the frequencies in column 2, going down from top to bottom. For example, the second entry is the sum of the first row and the second row in the frequency column (5 + 5 = 10), and the third entry is the sum of the first, second, and third rows in the frequency column (5 + 5 + 6 = 16):


Step 3: Add a fifth column and cumulate the **relative frequencies** from column 3. If you do this step correctly, your values should add up to 100% (or 1 as a decimal):


Step 4: Draw an [x-y graph](https://www.statisticshowto.datasciencecentral.com/what-is-a-cartesian-plane/) with percent cumulative relative frequency on the y-axis (from 0 to 100%, or as a decimal, 0 to 1). Mark the x-axis with the class boundaries.

Step 5: Plot your points.

 **Note**: Each point should be plotted on the upper limit of the class boundary. For example, if your first class boundary is 0 to 10, the point should be plotted at 10.

Step 6: Connect the dots with straight lines. the ogive is one continuous line, made up of several smaller lines that connect pairs of dots, moving from left to right.

The finished graph for this sample data:
