STATISTICS - I

Class: CS/DS/BCS

Two Marks

1. **Define statistics.**
   Statistics is defined as collection, organisation, presentation, analysis and interpretation of numerical data.

2. **What is meant by nominal data?**
   Nominal data refers to named or labelled variables that do not include numerical values. It classifies *items* and people by name, color, nation, and gender.

3. **Define complete enumeration survey.**
   A census is a study of every unit, everyone or everything, in a population. It is known as a complete enumeration, which means a complete count.

4. **Define diagram.**
   A diagram is a drawing or plan that explains something by showing how the parts relate to each other.

5. **What is a pie diagram?**
   A pie chart, sometimes called a circle chart, is a way of summarizing a set of nominal data or displaying the different values of a given variable (e.g., percentage distribution). This type of chart is a circle divided into a series of segments. Each segment represents a particular category.

6. **Define histogram.**
   A histogram is a diagram involving rectangles whose area is proportional to the frequency of a variable and width is equal to the class interval.

7. **Define Arithmetic mean.**
   It is defined as the sum of the values of all individual observation of a series divided by the number of observations of a series.

8. **Define Mode.**
   Mode is that value which occurs most often in the data, (i.e.,) with the highest frequency.

9. **Solve Mode, If mean = 30, median = 32**

   Mode = 3Median – 2 Mean
   = 3(32) – 2(30)
   Mode = 36

10. **What is the measures of variation?**
    Measures of variation is one such with the help of which we can study the dispersion in the values of the distribution.

11. Define range

Range is defined as the difference between the largest value and smallest value of the distribution.

12. Write the uses of Lorenz curve.

A Lorenz curve is a graphical representation of income inequality or wealth inequality. The graph plots percentiles of the population on the horizontal axis according to income or wealth.

13. Define Correlation.

The statistical tool with the help of which their relationships between two or more than two variables are studied is called correlation.

14. What is scatter diagram?

The simplest device for determining relationship between two variables is a special type of dot chart called scatter diagram.

15. What is meant by association of attributes?

Association of Attributes refers to the relationship between two or more attributes in a data set. It is a statistical technique used to analyse the relationship between attributes.

Five Marks

1. **Explain the One - dimensional Diagrammatic presentation of Statistical data**
One-Dimensional Diagrams or Bar Diagrams

One-dimensional diagrams are those that have only one-dimensional measurements, such as height or length.

- In these diagrams, the magnitude of the characteristics is shown by the length or height of the bar.
- The width of a bar is arbitrarily set to make the constructed diagram more elegant and attractive.
- The length and height of the bars vary depending on the variable value. However, the width of the bar remains constant.
- Width is also determined by the number of bars that must be accommodated in the diagrams.
- If there are a large number of items, lines can be used instead of bars.
- The bars must be equidistant from one another.
- Bars can be drawn horizontally or vertically. They are, however, usually in vertical form.

The different types of One-Dimensional Diagrams are as follows:

1. **Simple Bar Diagram**
2. **Multiple Bar Diagram**
3. **Sub-Divided Bar Diagram or Component Bar Diagram**
4. **Percentage Bar Diagram**
5. **Broken-Scale Bar Diagram**
6. **Deviation Bar Diagram**

1. Simple Bar Diagram:

A diagram in which each class or category of data is represented by a group of rectangular bars of equal width is known as a **Simple Bar Diagram**. It is the simplest type of bar diagram. In this diagram, each bar represents one figure only. The number of bars will be equal to the number of figures. These diagrams show only one characteristic of the data, such as sales, production, or population figures for various years.
The magnitude of data is determined by the bar's height (or length). The lower end of the bar touches the base line; therefore, the height of a bar starts from the zero unit. These diagrams can be vertical or horizontal in layout:

1. **Vertical Bar Diagram:** The diagram in which the magnitude of the data is presented vertically, i.e., along the Y-axis, is a Vertical Bar Diagram.
2. **Horizontal Bar Diagram:** The diagram in which the magnitude of the data is presented horizontally; i.e., along the X-axis is a Horizontal Bar Diagram.

The bars of a bar diagram can be visually compared by their relative height, and data can be easily comprehended accordingly.

2. Multiple Bar Diagram:

The Multiple Bar Diagram is used to compare two or more variables such as revenue and expenditure, import and export for different years, marks obtained in different subjects in different grades, and so on. It is often referred to as a **Compound Bar Diagram**. The method for creating multiple bar diagrams is the same as for creating a Simple Bar Diagram. However, to distinguish the bars from each other, different bars are differentiated by different shades or colours.

3. Sub-Divided Bar Diagram:

In these diagrams, the bar corresponding to each phenomenon is divided into several components. Each part or component occupies a proportional part of the bar to its share in the total. **For example,** the bar corresponding to the number of students enrolled in a course can be further sub-divided into boys and girls.

- When preparing a sub-divided bar diagram, the various components in each bar should be kept in the same sequence.
- It is important to use different colours or shades to differentiate between different components.
- A suitable index should explain these various colours or shades.
- These diagrams are quite useful for comparing the sizes of various parts and throwing light on the relationship between these integral parts. **For instance**, such diagrams are used to present data such as sales profits from various products, a family's expenditure pattern, the budget outlay for receipts and expenditures, and so on.

4. Percentage Bar diagram:

A Percentage Bar Diagram is a sub-divided bar diagram that indicates the total percentage of each component rather than the magnitude. The absolute magnitudes of several components are presented using a subdivided diagram. These magnitudes can be converted into relative values by describing them as a percentage of the total.

Each data component is expressed as a percentage of the corresponding total. Thus, in a percentage bar diagram, all of the bars are of height 100, while the different segments of the bar representing the various components vary in height depending on their % value of the total. Just like in the sub-divided bar diagram, in the percentage bar diagram, different components can be differentiated by different shades or colours.

5. Broken-Scale Bar Diagram:

This diagram is used when the value of one variable is extremely high or extremely low in comparison to others. Larger bars may be broken to make space for the smaller bars of the series. Every bar has its value written on the top of the bar.

When the majority of the data figures are of low magnitude and one or more of the figures are of unusually large magnitude, a broken-scale diagram is used to present the data.

6. Deviation Bar Diagram:

These diagrams are used to represent net changes in data such as net profit, net loss, net exports, net imports, etc.

- In these diagrams, only changes are shown, not the original data.
- The values in these diagrams might be both positive and negative.
- Positive values are displayed above the X-axis (Base line), while negative values are displayed below it.

2. **Draw a pie chart to reflect the following expenditure for an ordinary working-class family**.

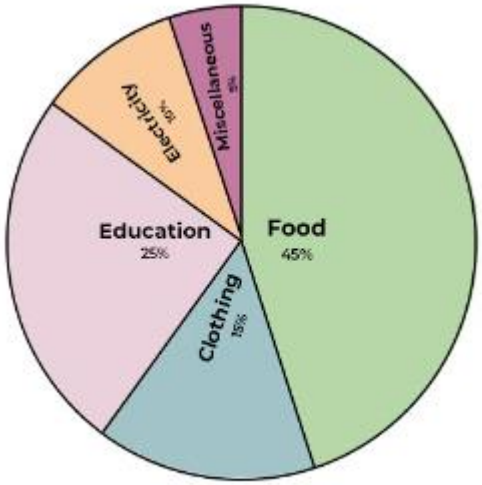
Items	% of Total Expenditure
Food	45
Clothing	15
Education	25
Electricity	10
Miscellaneous	5

Solution:

The sum of the percentages of total expenditure is 100. It means that the total angle; i.e., 360° represents 100. Now, multiply each of the percentages by or 3.6° to determine the proportionate angles.

Items	% of Total Expenditure	Proportionate Angles
Food	45%	45 x 3.6 = 162°
Clothing	15%	15 x 3.6 = 54°
Education	25%	25 x 3.6 = 90°
Electricity	10%	10 x 3.6 = 36°
Miscellaneous	5%	5 x 3.6 = 18°
	100%	360°

According to the degrees of angles at the centre, the circle is divided into five parts. Hence, the pie diagram for the given data will be represented as:



3. **Calculate mode from the following data.**

C.I	0-20	20-40	40-60	60-80	80-100
f	5	20	35	20	12

Solution

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$= 40 + \frac{35-20}{2(70)-20-20} \times 20$$

Mode = 43

4. **Write the merits and demerits of Standard Deviation.**

Merits

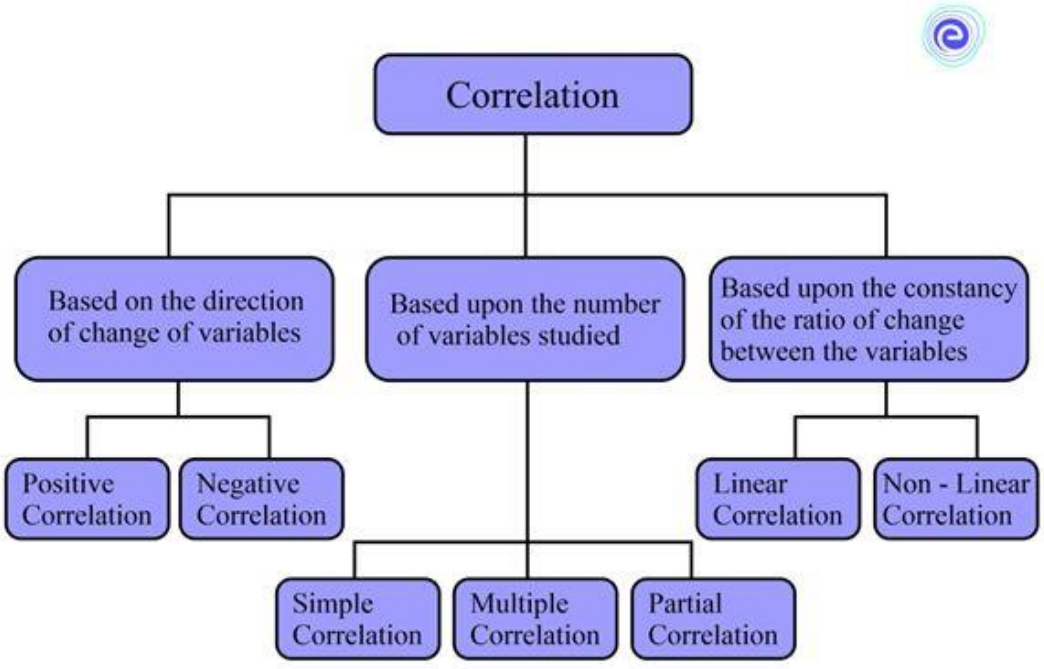
- It is rigidly defined and free from any ambiguity.
- Its calculation is based on all the observations of a series and it cannot be correctly calculated ignoring any item of a series.
- It strictly follows the algebraic principles, and it never ignores the + and – signs like the mean deviation.
- It is capable of further algebraic treatment as it has a lot of algebraic properties.
- It is used as a formidable instrument in making higher statistical analysis viz.: correlation, skewness, regression and sample studies, etc.
- It is not much affected by the fluctuations in sampling for which is widely used in testing the hypotheses

Demerits

- It is not understood by a common man.
- Its calculation is difficult as it involves many mathematical models and processes.
- It is affected very much by the extreme values of a series in as much as the squares of deviations of big items proportionately bigger than the squares of the smaller items.
- It cannot be used for comparing the dispersion of two, or more series given in different units.

5. **Explain the types of correlation.**

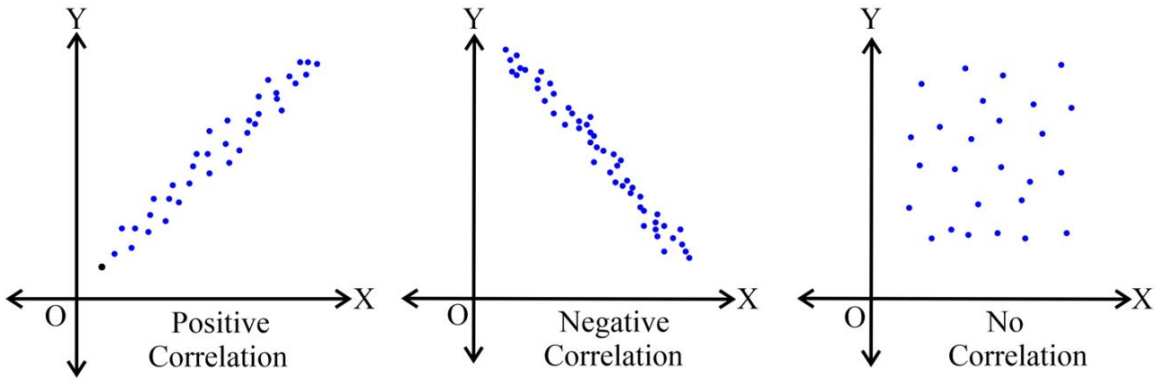
We can classify the correlations based on the direction of change of variables, the number of variables studied and the constancy of the ratio change between the variables.



TYPE I: BASED ON THE DIRECTION OF CHANGE OF VARIABLES

Correlation is classified into two types based on the direction of change of the variables: positive correlation and negative correlation.

- **PositiveCorrelation:**
  When the variables change in the same direction, the correlation is said to be positive. The sign of the positive correlation is +1.
  Example: When income rises, so does consumption and when income falls, consumption does too.
- **NegativeCorrelation:**
  When the variables move in opposite directions, the correlation is negative. The sign of negative correlation is –1.
  Example: Height above sea level and temperature are an example of a negative association. It gets colder as you climb the mountain (ascend in elevation) (decrease in temperature).



TYPE II: BASED UPON THE NUMBER OF VARIABLES STUDIED

There are three types of correlation, based on the number of variables.

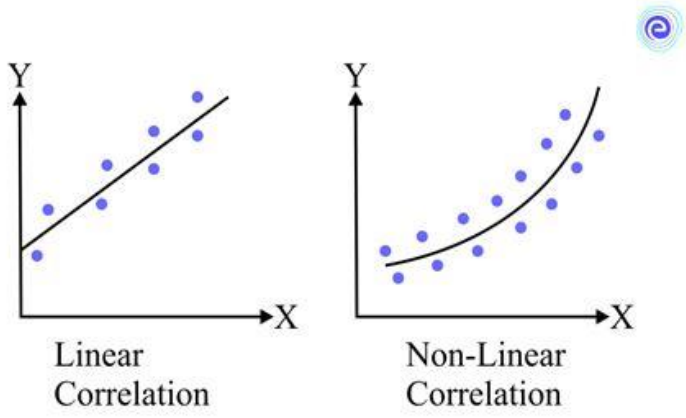
1) **Simple Correlation:**
   The study of only two variables is referred to as simple correlation. The usage of fertilisers and paddy yield is an example of a simple connection, as paddy yield is dependent on fertiliser use.
2) **Multiple Correlation:**
   Multiple correlations is defined as the study of three or more variables at the same time. Crimes in a city, for example, maybe influenced by illiteracy, growing population, and unemployment, among other factors.
3) **Partial Correlation:**
   If there are three or more variables, but only two are considered while keeping the other variables constant, the correlation is said to be partial. For example, while controlling for weight and exercise, you would wish to investigate if there is a link between the amount of food consumed and blood pressure.

TYPE III: BASED UPON THE CONSTANCY OF THE RATIO OF CHANGE BETWEEN THE VARIABLES

Correlation is classified into two types based on the consistency of the ratio of change between the variables: linear correlation and non-linear correlation.

1) **Linear Correlation:**
   The correlation is said to be linear when the change in one variable bears a constant ratio to the change in the other.

2) **Non-Linear Correlation:**
   If the change in one variable does not have a constant ratio to the change in the other variables, the correlation is non-linear.



Ten Marks

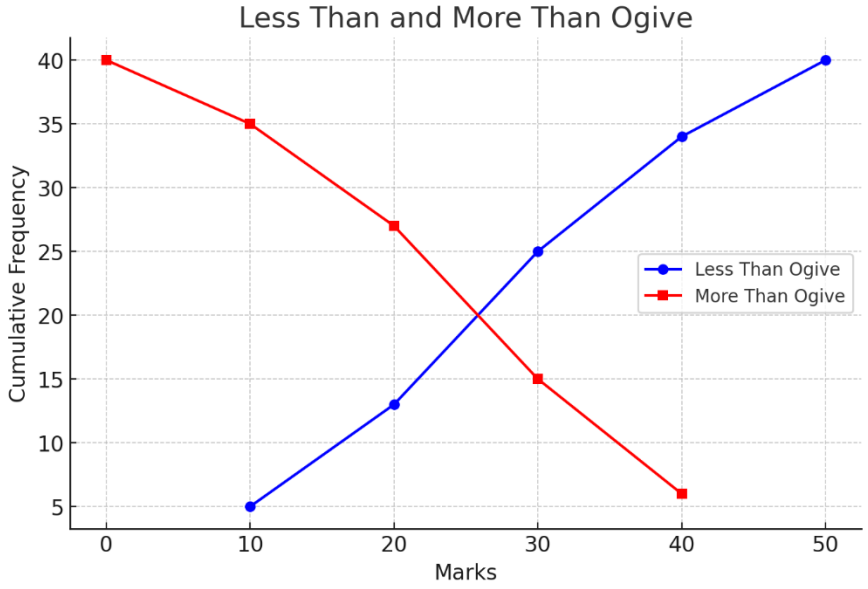
1. **Draw the Ogives for the following data.**

CI	0-10	10-20	20-30	30-40	40-50
f	5	8	12	9	6

Solution:

Calculation of Cummulative frequencies:

CI	f	Upper limit	Less than cf	Lower limit	More than cf
0-10	5	10	5	0	40
10-20	8	20	13	10	35
20-30	12	30	25	20	27
30-40	9	40	34	30	15
40-50	6	50	40	40	6



2. Compute Mean, Median and Mode for the following data given below

Expenditure (Rs.)	160	200	250	320	410	500	570
No. of families	5	12	23	18	13	8	3

Solution:

X	f	fX	cf
160	5	800	5
200	12	2400	17
250	23	5750	40
320	18	5760	58
410	13	5330	71
500	8	4000	79
570	3	1710	82
	$\sum f = 82$	$\sum fX =$ 25750	

Mean, $\bar{X} = \frac{\sum fX}{N} = \frac{25750}{82} = 314.02$

Median = Size of $(\frac{N+1}{2})^{th}$ item

= Size of $(\frac{82+1}{2})^{th}$ item

= Size of $(41.5)^{th}$ item

Median = 320

Mode = 250

3. What is Quartile Deviation? Write the merits and Demerits.

Quartile Deviation is defined as the half of the difference between upper quartile and lower quartile.

Merits:

- It is rigidly defined.
- It is superior to range in as much as its calculation is based on middle 50% of the items of a series.
- It is easy to calculate, and especially in case of an open-end series, no 50% of the items of a series.
- It is not very much affected by the extreme values of a series.

Demerits

- It is not based on all the observations of a series.
- It is not capable of further algebraic treatment.
- It is affected by fluctuations in sampling.
- It is not understood by a common man.
- It does not exhibit any scatter around an average for which is remarked as a measure of partition rather than a measure of dispersion.
- It takes more time to calculate as its calculation involves finding out of the values of the lower quartile, and upper quartile.

4. Find the correlation coefficient for the following data.

X	12	25	21	20	22	35	30	28	26
Y	25	32	42	45	39	72	68	47	55

Solution:

X	**Y**	**d_x = X- A**	**d_y = Y-A**	**d_x^2**	**d_y^2**	**$d_x d_y$**
12	25	-10	-14	100	196	140
25	32	3	-7	9	49	-21
21	42	-1	3	1	9	-3
20	45	-2	6	4	36	-12
22	39	0	0	0	0	0
35	72	13	33	169	1089	429
30	68	8	29	64	841	232
28	47	6	8	36	64	48
26	55	4	16	16	256	64
		21	**74**	**399**	**2540**	**877**

$$r = \frac{N\sum d_x d_y - \sum d_x \sum d_y}{\sqrt{N\sum d_x^2 - \sum (d_x)^2}\ \sqrt{N\sum d_y^2 - \sum (d_y)^2}}$$

$$= \frac{9(877) - (21)(74)}{\sqrt{9(399) - (21)^2}\ \sqrt{9(2540) - (74)^2}}$$

$$r = 0.958$$