ORGANISATION OF DATA

MEANING OF CLASSIFICATION

The process of classifying data involves dividing it into distinct but related sections or organizing it into sequences and groups based on shared traits.

Purpose of classification of data

1. It simplifies the unprocessed data into a format appropriate for statistical examination simplifies and draws attention to the data's characteristics.
2. It makes comparing and deriving conclusions from the data easier.
3. It offers details on how different components of a data set are related to one another.
4. By grouping the data set's components into homogeneous groups,
5. it facilitates statistical analysis by highlighting similarities and differences.

Objective of classification

1. **To simplify and condense the mass of data:** the primary goal of classification is to remove extraneous details and reduce the enormous amount of complex data to a form that is simple, condensed, logical, and understandable. It aids in emphasizing the important aspects of the data.
2. **To explain similarity and dissimilarity of Data**: Classification makes it easier to organize the data into groups based on specific affinities and diversities. This makes them easy for the investigators to understand. Classes are used to store information such as married and single, employed and unemployed, and educated and ignorant.
3. **To facilitate comparisons:** We can identify facts, form conclusions, and make insightful comparisons thanks to classification.
4. **To study the relationships**: Using a set of criteria, classification assists in determining the cause and effect links among the data. After categorizing the vast majority of data, for instance, the attributes of education and wealth can be connected.
5. **To get the data ready for tabulation**: Tabular presentation is only possible for classified data. Consequently, classification offers a foundation for tabulation and additional statistical processing.

6. **To present a mental image**: The classification process helps people visualize objects, and summarized information is simple to recall.

Prequisites of a Good Classification

1. **Suitability:** The classification must be appropriate for the goal of the investigation. *For example*, it will be useless to categorize employees based just on their religion if research is done to find out about their financial circumstances.
2. **Unambiguous**: There should be no room for doubt or confusion as a result of the classification. Sorting units into distinct groups based on shared attributes shouldn't be too tough.
3. **Exhaustiveness**: The classification system ought to be so comprehensive that each unit in the series is assigned to a specific category.
4. **Flexibility**: A successful classification should be able to be modified in response to evolving circumstances.
5. **Mutually Exclusive**: In order for an observed value to belong to only one class, there must be no overlap between the classes. Nothing must exist that can be used in more than one class.
6. **Stability:** If the categorization principle is chosen, it must be adhered to throughout the analysis in order to produce results that are relevant.
7. **Homogeneity:** When related things are grouped together in a class, it is referred to as homogeneous classification. All units that are part of a group should have similar traits.

METHODS OF CLASSIFICATION

1. **Geographical classification.** The term "geographical classification" refers to the classification of data based on region or place. When the population of various states is shown, we refer to it as being classified geographically.

2. **Chronological Classification (Temporal Classification)**

   Chronological classification is a sort of classification that occurs when data is categorized according to distinct time periods. For instance, the population of Delhi can be categorized chronologically for various years.

3. Classification of Qualitative Information

When data is classed qualitatively, it is done so using descriptive features or non-quantifiable variables such as caste, region, sex, literacy, and education. There are two varieties of this classification type:

I. **Simple Classification:** A classification is considered simple if it divides facts into two groups based solely on a single attribute. It is easy to classify a city's population, for instance, if we split it up into two groups: males and females.

II. **Manifold Classification:** A classification is referred to as manifold if facts are categorized based on more than one attribute and if each class has more than two subclasses. For instance, there are several classes established if we divide a city's population into males and females, then into literate and illiterate people, and finally into religious groups.

4. Quantitative Classification (Or Numerical Classification)

Data is categorized using this method according to certain measurable attributes, such height, weight, revenue, expenses, production, or sales.

CONCEPT OF VARIABLE

The term variable is derived from the word 'vary' which means to differ or change. Hence variable means the characteristic which differs or changes. A variable refers to quantity or characteristic whose value varies from one investigation to another.

There are two kind of variable discrete and continuous variable

1. **Discrete variable:** Discrete variables are those that have the ability to accept just exact values and not any fractional values. Stated differently, discrete variables are represented using whole numbers.

2. **Continuous Variable:** Continuous variables are those that have the ability to take any value within a specific range, including fractional and integral values.
   Individuals' weights and heights can therefore be any figure within the bounds. In this scenario, measurements are used to collect data.

FREQUENCY

The number of times a specific value occurs in a distribution is referred to as its frequency.

- Let's take an example where there are 20 pupils in the class and among them:
- Nine pupils received 70 scores,
- six received 85 marks, and
- five received 92 marks.

  Frequencies will now be 5, 6, and 9, in that order.

STATISTICAL SERIES

Statistical Series is the logical arrangement of classified data according to some feature, either measurable or not. Examples of such qualities include size, time of occurrence, and other attributes.

KINDS OF STATISTICAL SERIES

On the basis of characteristics

1. **Time Series:** A series that is created when the various values that a variable has taken over time are grouped chronologically is referred to as a time series.The data is displayed as a statistical series according to a time unit (day, week, month, or year).
2. **Spatial Series:** A spatial series is made up of data organized geographically or by location. In this series, locations vary while the time factor stays constant.
3. **Condition Series**: Information is categorized in this series based on changes that take place under specific circumstances. Pupils in a particular class, for instance, grouped by age, height, weight, grades, etc.

On the basis of construction

1. **Individual series:** Individual series refers to that series in which items are listed singly, i.e. each item is given a separate value of measurement.
   (i) ***Unorganized Individual Series:*** An unstructured mass of data is known as an unorganized series (raw data). Data in its unaltered state is referred to as raw data. Raw

data, or disorganized data, is data that has been gathered by the investigator but has not been arranged in a systematic way.

(ii) ***Organized Individual Series:*** Raw data is arranged in an orderly fashion in an organized series. Two methods are available for presenting an ordered individual series:

(a) According to Serial Number

(b) According to order of Magnitude

2. **Discrete series:** Discrete series data expression is a far superior method of data presentation. A discrete series is one in which there is a defined amount of difference between each value in the series. In a discrete frequency distribution, various values of the variable are shown along with their corresponding frequencies. 'Frequency Array' is the term used to describe the data classification of a discrete variable. Discrete variables do not accept fractional values; instead, they require a fixed integral value. As a result, we have frequencies for each of its integral values.
3. **Continuous series:** A continuous variable, as contrast to a discrete variable, can have any value within an interval. A continuous series is one that displays the range of values for each item in the series and reflects continuous variables.

   Frequency distribution, grouped frequency distribution, series with class intervals, and series of grouped data are other names for continuous series.

   (i) ***Class:*** Class hereby means a group of numbers in which items are placed such as 0-10, 10-20, 20-30, etc.

   The classes should be clearly defined and should not lead to any confusion.

   Classes should be exhaustive and mutually exclusive, so that any value of the variable corresponds to one and only one of the classes.

   (ii) ***Number of Classes:*** The judgment of each individual investigator is a major factor in determining the number of classes.

   Although there isn't a hard-and-fast guideline for how many classes should be created, there shouldn't be an extreme amount of them.

   There ought to be five to fifteen classes. If there are fewer than five classes, each term will not be included correctly. Likewise, calculations and computations become challenging as it exceeds 15.

(iii) ***Class limits***'Class limit' refers to the lowest and maximum values of the variables within a class.

Every class in a continuous series lies between two numbers. The class limit is these two integers.

A class's "upper limit," or "L_2" is its maximum value, while its "lower limit," or "L_1" is its lowest value.

The bottom and upper bounds, if class is 10–20, will be 10 and 20, respectively.

It is practical to set a class's lower limit to a multiple of five or to zero. Class bounds ought to encompass all of the given data and, to the greatest extent feasible, be whole numbers.

(iv) ***Class-Interval Class-Interval:*** Class-interval is the difference between the upper limit (L2) and lower limit (L1).

Generally, the symbols 'e' and 'i' represent the class-interval.

The class-interval is sometimes referred to as the class's "magnitude," "size," or "length."

(v) ***Class-Interval Width:*** When creating the frequency distribution, it is ideal for each class-interval's width to have the same size. The following formula can be used to calculate the width (or size) of each class-interval:

Width of the class interval= largest observation – smallest observation / no. of class desired.

(vi) ***The range:*** The range of a frequency distribution can be defined as the difference between the lower limit of first class-interval and the upper limit of the last class-interval. For example, if classes ate 0-10, 10-20 till 70-80, then range is 80-0=80.

(vii) ***Mid-Value or Mid-Point***: The middle point of a class interval is called the midpoint. The computation involves dividing the sum of the magnitudes of the lower and upper bounds by two.

(viii) ***Frequency:*** The quantity of things (observations) falling into a specific class is referred to as frequency. For instance, frequency is 10 if there are 10 students in classes 0–20. This indicates that ten pupils received grades ranging from zero to twenty.

(ix) ***Distribution of Frequencies*****:** A frequency distribution is a table that illustrates the distribution of a variable's various values among its many classes and the related class

frequencies. A thorough method for categorizing unprocessed data for a quantitative variable is the frequency distribution.

(x) ***Class frequency:*** Class frequency, also referred to as the frequency of that class, is the quantity of observations that belong to that particular class.

Generally, it is indicated by f.

The symbol for N represents the total frequency.

4. Types of continuous series :

(i) ***Exclusive Series:*** The classes of the type 10-20, 20-30, etc., wherein the upper limit of one class interval becomes the lower limit of the next class, are known as exclusive classes.

(ii) ***Inclusive Series:*** The classes of the type 10-19, 20-29, etc., wherein all observations with magnitude greater than or equal to the lower limit and less than or equal to the upper limit of a class are included in it, are known as inclusive classes.

(iii) ***Open-End Distribution:*** In a frequency distribution, if the lower limit of the first class and the upper limit of last class is not given, it is known as open-end distribution

(iv) ***Cumulative Frequency Series*** ('Less than' and 'More than')

- 'Less than' cumulative frequency distribution: In a 'less than' cumulative frequency distribution, the frequencies of each class-interval are added successively from top to bottom.
- 'More than' cumulative frequency distribution: In a 'more than' cumulative frequency distribution, the cumulative frequencies of each class-interval is obtained by finding the cumulative totals of frequencies starting from the highest value of the variable (class) to the lowest value (class).

(v) Equal and Unequal class-interval Series:

- Equal Class-interval Series: When the classes in a series are of the same interval, it is called the equal class-interval series.
- Unequal Class-Interval Series: When class-intervals are not equal, it is called unequal class- interval series.

(vi) ***Mid-value Series:*** When middle value of a class-interval are given, it is called mid-value series.

5. **Bivariate Frequency Distribution:** When the data is classified on the basis of two variables such as height and weight, marks in statistics and economics etc., the distribution is known as Bivariate frequency distribution or Two-way frequency distribution.