Some Basic concepts of Statistics

Population: A set of all values or elements defined on some common characteristics is called a population.

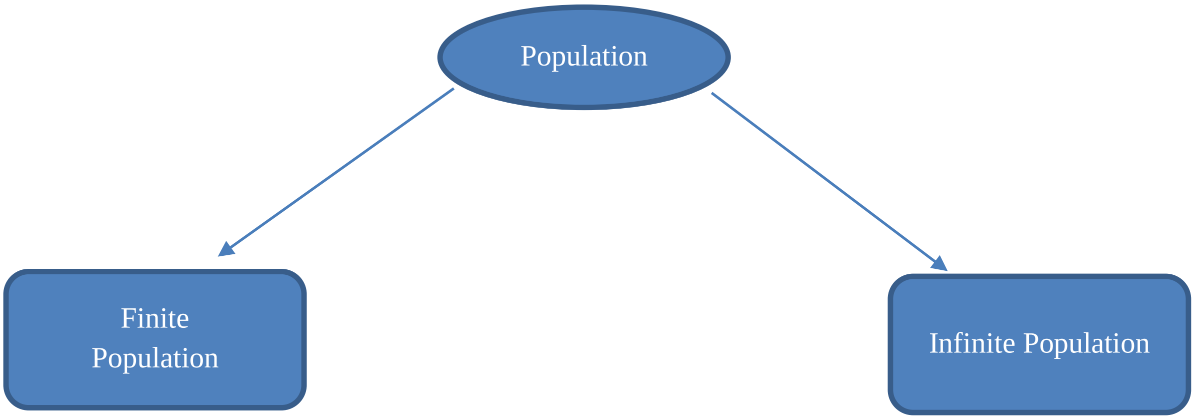
Totality or collection of all objects, items, or individuals on which observations are taken on the basis of some characteristics of the objects in any field of inquiry is called population and each object or items are called experimental units.

Thus population means an aggregate of elements possessing certain characteristics of interest in any particular investigation or enquiry.

A population consists of all the items or individual about which researcher want to draw a conclusion is known as **target population**.

'N' denotes the size of population.
Each object or individual of a population is called an experimental unit. Observations are collected on **experimental units**.

Example: If we want to study the average weight of the student of 1st semester BBA then the set that consists of all the weights of the student of 1st semester BBA will be the population in this case.



There are two types of population:

1) Finite population & 2) Infinite Population

Finite Population: A population having a finite number of units (or, individuals or, items) is called a finite population.
Example: Number of Students in BRAC University, Number of vehicles in Dhaka city.
Infinite Population: If the numbers of elements of a population are uncountable, then it is called infinite population.
Example: Number of fishes/ insect/ in the river/canal/pond. The bacteria grown in a rotten object, etc.
Parameter: A parameter is a numerical measure that describes a characteristic of a population.
Census: Any investigation based on every element of a population is called census.
Sample: A small but representative (desirably) part of population is known as sample.
In many particular situations it is impossible or even impractical to study the whole population, in such case only a small and representative part of population is taken under consideration to

draw inferences about the population by analyzing that part of population. Such a part of population is known as sample.

Sample size is denoted by ‘n’.

Statistic: A statistic is a numerical measure that describes a characteristic of sample.

Survey: The technique of collecting information from a portion of the population or sample is called survey.

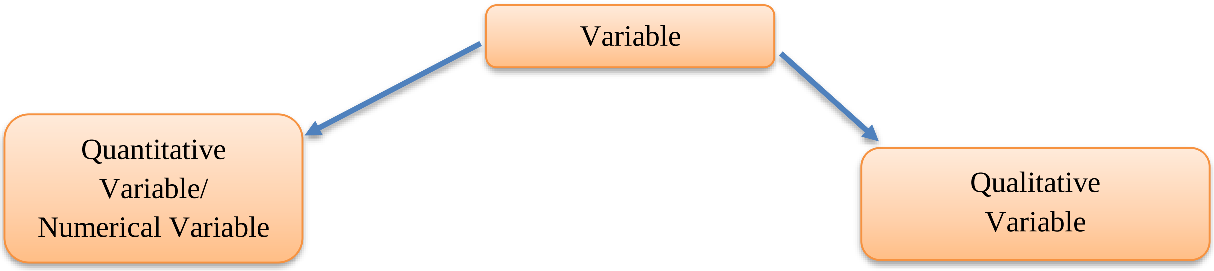
Variable: A measurable items or quantity that can varies from unit to unit, individual to individual, items to items, person to person is called variable. We denote the variables by capital letters and their values by small letters.

Example: Height, weight, age, SSC and HSC marks, family size, gender, etc. are some variables of 1st semester BBA students of BRAC University.

Constant: The measurable items or quantity or characteristics of a population which does not vary from one person to another/ element to element, either in magnitude or in in quantity is called constant.

Example: Number of legs, hands, fingers, eyes of a person. The ratio of circumference and diameter of a circular (π) and the value of π=3.1416 is same for all circles.

Types of Variables: There are two basic types of variables -



1. Qualitative variable (also known as categorical variable or attribute) A qualitative variable is one for which numerical measurement is not possible. In other words, when the characteristic being studied is non-numeric, it is called a qualitative variable or an attribute.

For example: Hair color (brown, black, white etc.), religion (Muslim, Hindu, etc.), Gender (male, female), home district (Dhaka, Rajshahi, Bogra, etc.), occupational status (employed, unemployed, self-employed, others), Efficiency of worker, Quality of a finished product, etc.

An individual is simply assigned to any one of the several mutually exclusive categories on the basis of observation on the individual. The qualitative observations can neither meaningfully ordered nor physically measured, these can only be classified and then enumerated.

In dealing with the qualitative data, researchers are usually interested in how many or what proportion fall in each category.

For Example: - What percent of students of BRAC Universities of English medium background? - What proportion of people opted in favor of construction of the new Airport? - How many Muslims and how many Hindus are there in Bangladesh?

2. Quantitative variable (also known as numerical variable): A quantitative variable is one for which the resulting observations have a numeric value and thus possess a natural ordering.

Quantitative Variables are measured on a numeric or **quantitative scale.**

Example: districts population size (family size), customer's shoe size, vehicle's speed on a highway, amount of daily sell of a shop, parent income, Price of a product, Daily rainfall in a coastal area, Wage of workers, Systolic blood pressure, etc. are all **quantitative variables**

A quantitative (numeric) variable is further subdivided into discrete and continuous variables.

• **Discrete variable:** A variable, which can take, only an isolated or countable finite or infinite number of values is called a discrete variable.

Example: Number of children in a family, number of road accidents in a year, number of phone calls received in a phone booth, Number of printing mistakes per page of a book, etc.

• **Continuous variable:** A variable that can take infinitely many values over a certain interval is called a continuous variable.

A variable is said to be a continuous variable if it can theoretically assume any value within a given range or range.

Example: height of a person –since it can take any value between 5.6 feet and 5.8 feet.

Exercise:

a. Classify each variable as qualitative or quantitative:

i. Marital status of nurses in a hospital
ii. Time it takes to run a marathon
iii. Weights of lobsters in a tank in a restaurant
iv. Colors of automobiles in a shopping center parking lot
v. Ages of people living in a personal care home

b. Classify each variable as discrete or continuous:

i. Number of pizzas sold by Pizza Express each day
ii. Lifetimes (in hours) of 15 iPod batteries
iii. Weights of the backpacks of first graders on a school bus
iv. Number of students each day who make appointments with a mathematics tutor at a local college
v. Blood pressures of runners in a marathon

Another 2 types of variables called random and non-random variables will be discussed in probability.

Data: A set of observations obtained from a particular enquiry is called data or a data set.

Numerical facts gathered from a statistical investigation are called a data. In a statistical analysis the first work is to collect data the raw materials of statistics after identifying a specific problem and field of enquiry.

Data is in fact the *plural form* of *'datum'*. Single information of a phenomenon on any subject of interest is called a datum. So data is called the collection of datum.

Example: If we are interested about the height of the students of 1st semester in BBA of BU, then a single value (that is the height of a student) is called a datum, and the set of all values of height will be data.

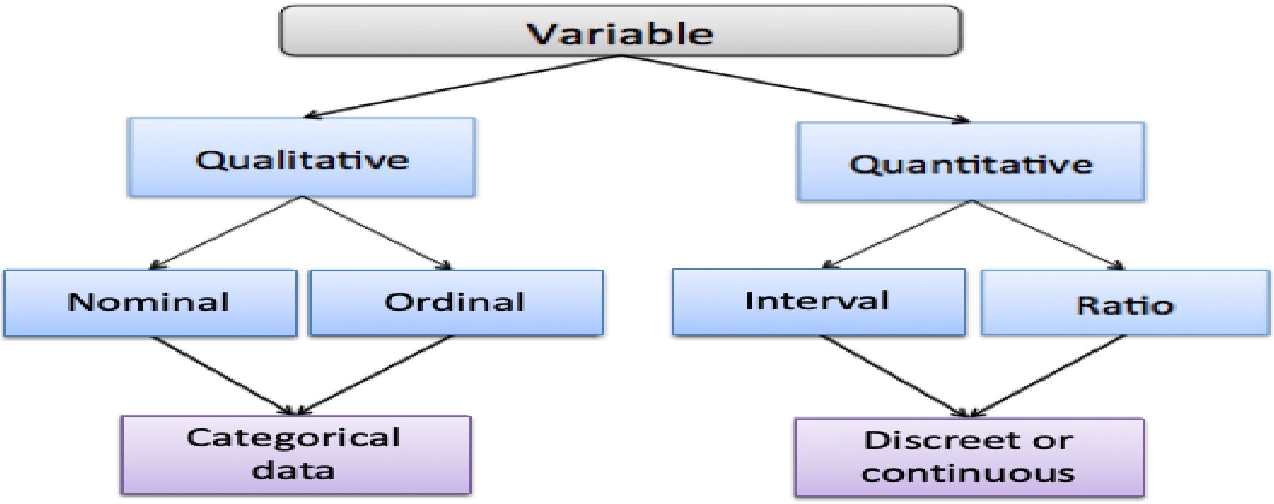
Sources and Types of data: Based on the sources data can be of two types.

Primary data: A data is said to be primary data if it is obtained from an investigation conducted for the first time. Thus, the data collected for the first time by the investigator as original data are known as primary data.

Secondary data: When a statistical analysis is conducted on a data set available from a prior investigation is called a secondary data.
Example: National income data collected by the government are primary data but they become secondary data for those who use them.
Raw data: In any statistical investigation, when data first collected usually appear in raw form where, information has been recorded merely in arbitrary order in which they happened to occur. This is known as the raw data set.

Raw data, collected for any statistical investigation, is unable to represent the summary information, which are although preliminary but necessary for analyses with advanced statistical method. So it is necessary to represent the raw data in such a way, which will enable us to extract the preliminary ideas about the variable(s) under study, to get some summary measures, and also to perform further statistical analysis.
Measurement: It is a process of assigning numbers to some characteristics or variables or events according to scientific rules. Data according to a scale of measurement is 4 types:

1. Nominal
2. Ordinal
3. Interval &
4. Ratio



Nominal: The scale of measurement by which we can classify and identify a qualitative variable according to different categories is called nominal data.
Example:

1) Gender of a worker in a factory (Male, Female)
2) Religion of a student (Muslim, Hindu, Buddhist, Christian)
3) Marital status of a person (Single, married, widowed, divorced, separate)

Ordinal: The scale of measurement by which we can classify, identify, and rank a qualitative variable according to different categories is called ordinal scale.
Example:

1) Grading of a student (A, B, C, D)
2) Economic status of a citizen (Higher class, Middle Class, Poor)
3) Health status of a person (Excellent, Good, Poor)

Interval: The scale of measurement by which we can measure a quantitative variable numerically on the experimental unit with arbitrary zero as origin is called interval scale.

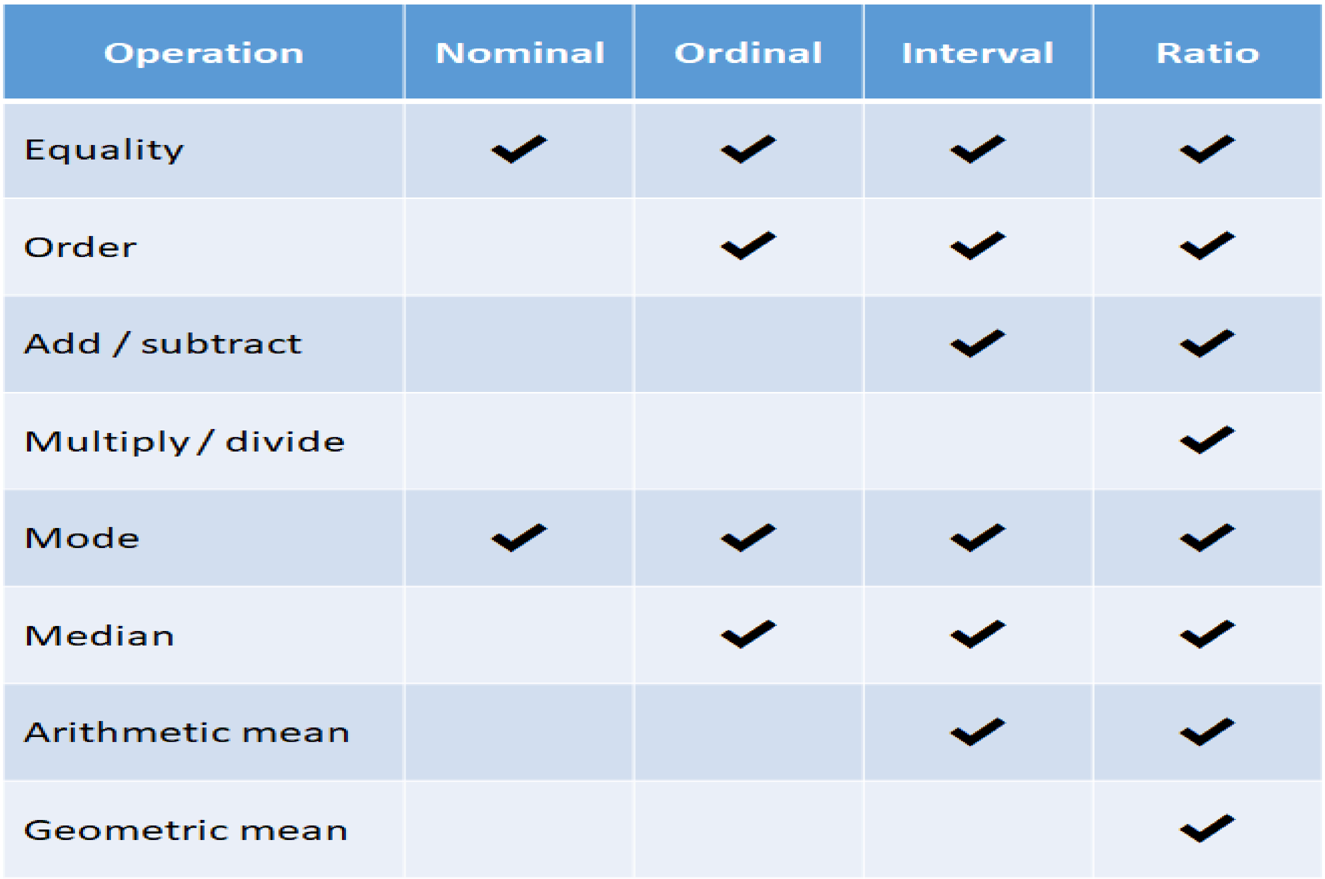
Example:

i) Body temperature of a patient
ii) Marks obtained by students in an exam.
iii) Calendar time
iv) Scholastic Aptitude test (SAT)

Ratio: The scale of measurement by which we can measure a quantitative variable numerically on the experimental unit with absolute zero as origin is called ratio scale.

Example:

i) Age, Height, Weight of a worker
ii) Number of printing mistakes per page of a book
iii) Number of children per family
iv) Number of defects of a product



Operation	Nominal	Ordinal	Interval	Ratio
Equality	✓	✓	✓	✓
Order		✓	✓	✓
Add / subtract			✓	✓
Multiply / divide				✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Arithmetic mean			✓	✓
Geometric mean				✓

Dealing with Raw Data: How to prepare data for further Statistical operation: In the next few subsequent segments, we are going to discuss on some techniques of statistics that we usually used to condense raw data, to make the data prepared for further statistical application.

The most frequently used methods for data condensation or/and representation are

i. Classification
ii. Tabulation
iii. Graphical representation

Classification: Classification is the process of arranging data values of a variable in groups or classes according to their affinities or of our interest. It is the first step towards further processing of a heterogeneous mass of data in to a number of homogeneous groups and subgroups by their respective characteristics. Broadly, the data can be classified on the following four basis:

i) Geographical, i.e., area-wise, e.g., cities, districts, etc.
ii) Chronological, i.e., on the basis of time
iii) Qualitative, i.e., according to some attributes.
iv) Quantitative, i.e., in terms of magnitudes.

Purpose of classification: Classification is necessary to serve the following purpose:

i. To eliminate unnecessary details.
ii. To bring out clearly point of similarity and dissimilarity.
iii. To enable one to form mental picture of the object.
iv. To enable one to make comparisons.
v. To pinpoint the most significant features of the data at glance.
vi. To enable statistical treatment of the collected data.

Principles of determining the number of classes: Usually we determine the number of classes in the light of the following conjoined considerations-

1) The number of observations of a variable.
2) The lowest and highest value of a variable.
3) Even distribution of the values with in classes.
4) A regular sequence of frequencies.
5) Avoidance of extremely large or small number of classes.

Tabulation: A statistical method of data condensation by which we can represent summary information of one or more variables, is defined as tabulation.

A statistical table is the logical listing of collected data in vertical columns and horizontal rows of numbers with sufficient explanatory and qualifying words, terms and statements in the form of titles, headings and notes which make clear the full meaning of data and their origin.

Principles of the constructions of a table:

Some of the most basic principles that one should consider in constructing table are as follows:

1. The table should be self-explanatory. The title describing the contents of the table should be clear, concise and to the point.
2. The table should be as simple as possible. Two or three tables are often preferable to a large table containing too many details and variables.
3. The specified units of measurements for the data should be given.
4. Necessary code or symbols used in table should be explained in a footnote.
5. Sources of data should be mentioned.

Frequency distribution: Frequency distribution is a tabular summary showing the number of times each value of the variable occurs in the data.

That is frequency distribution can be defined as a summary presentation of a number of observations of an attributes or values of a variable arrange according to their occurrence either individually (in case of discrete data) or in a range (in case of both discrete or continuous data).

There are two types of frequency distribution: 1) Discrete & 2) Continuous frequency distribution.

Table 01: Frequency distribution of number of children per family (Discrete)

Number of children	Number of families
0	10
1	27
2	15
3	18
4	9

Table 02: Frequency distribution of height of trees in Sundarbans (Continuous Frequency distribution)

Height of the tree (In Feet)	Number of trees
0-50	1000
50 – 100	2735
100 – 150	1589
150 – 200	1518
More than 200	719

Class limit: Class limits are the highest and the lowest values that can be included in the class. For example, if we consider the class 50 – 100, here 50 is the lower limit and 100 is the upper limit. In such case no values greater than 100 shall fall into that class. Similarly, no values less than 50 shall fall into that class either.

Class interval: The difference between the upper limit and the lower limit of a class is called the class interval. Class interval is usually denoted by c, *i*, h, or w.

For example, the class interval of the class ‘50 – 100’ is 50.

Class frequency: The number of observations falling with in a particular class is called its frequency or class frequency.

Class midpoint or class mark: The value of the variable that lies in the middle of the upper and lower limits is called mid value or midpoint of the class. It can be obtained as follows:

$$\text{Class midpoint: } \frac{l+s}{2}$$

Where, *l*= Upper limit of the class, *s*= Lower limit of the class

Relative frequency (also known as proportion): Instead of presenting the frequencies in absolute terms, it is sometimes convenient to express the frequencies in percentages. The relative frequency (also known as proportion) corresponding to a class is simply the ratio of the total number of items in that class to the total number of elements in the total set.

Multiplying relative frequency by 100 one can obtain the percentage of observation that belongs to any particular class.

$$\text{Relative frequency = Proportion = } \frac{\text{Frequency in each class}}{\text{Total number of observations}}$$

Cumulative frequency: The cumulative frequency corresponding to a class is the total of all frequency up to and including that class.

Example: Let us consider the following table showing the distribution of mark of 27 students.

Class limit	Class mid value	Frequency	Relative frequency	Cumulative frequency	Cumulative relative frequency
0 – 10	5	4	0.148	4	0.148
10 – 20	15	8	0.296	4+8	0.444
20 – 30		5		4+8+5	
30 - 40		4			
40 – 50		3			
50 – 60		2			
60 – 70	65	1			
Total					

- **Important steps for constructing a frequency distribution:**
  Step 1: Determination of range: Range is the difference between the largest value and the smallest value of the data set. Therefore, Range= Largest value – Smallest value
  Step 2: Finding the number of class: Number of classes should not be too large or too small. As a general rule, the number of classes should range from 5 to 25. Another rule of thumb is that the number of classes should be around $\sqrt{n}$ where n is the number of observations in the data. According to M.A. Struge's, the number of classes can be determined using formula

$$\mathbf{K= 1+ 3.322\ Log_{10}{}^{N}}$$

  Step 3: Decide the types and number of column.
  - Decide whether the frequency distribution will be exclusive or inclusive.
  - Decide that the class limit will be equal or unequal.

  Step 4: Calculate the class interval If equal class limit: The width or size of the class interval is h = $\frac{\text{Range}}{\text{Number of class}} = \frac{\text{Largest value}-\text{Lowest value}}{1+3.322\ \text{Log}_{10}^{\text{N}}}$
  Step 5: Tally Mark and Frequency: Take each item from the data, one at a time and put a tally mark (|). Counted tally marks will be the required frequency. Lastly, we have to give the suitable title to the frequency distribution table.

Assignment W1D2_001

Question 1: The following data relate to the audit-time of 20 clients.

10, 15, 20, 28, 13, 18, 24, 29, 12, 16, 23, 34, 14, 17, 22, 17, 21, 16, 18, 19.

For the given data construct a suitable frequency distribution table.

Question 2: The following information, extracted from a survey of a Microfinance institution (MFI) represents the amount of loan request of 50 potential borrowers from any particular branch of that MFI.

1850	9250	6100	4500	5100	1800	6100	6500	6999	6780
3100	7475	6400	4950	8789	6100	6480	7050	9900	4790
4400	7900	6900	3865	5556	4859	6999	6780	8050	9900
5600	6600	9980	4800	8855	5550	1200	4790	6500	8050
3858	7300	8050	6200	7155	4980	8050	6480	7050	1500

For the given data construct a suitable frequency distribution table featuring the following components

i. Class mid value
ii. Tally Bars
iii. Frequency
iv. Relative Frequency
v. Cumulative frequency & vi. Cumulativerelative frequency

Using the aforementioned information also answer the following

a. Determine the number of loan request between tk 4000-6000 12

b. Determine the proportion of loan request between 4000 – 6000.

c. Determine the number of loan request below tk. 7000. 34

d. Determine the proportion of loan request below tk. 7000.

Question-3. Fill in the blanks:

(*i*) Quantitative Variables are of two kinds … … and … …
(*ii*) … … is the process of arranging data into groups according to their common characteristics.
(*iii*) In chronological classification, the data are classified on the basis of … …
(*iv*) … … classification means the classification of data according to location.
(*v*) Class-mark (mid-point) is the value lying half-way between … …
(*vi*) According to Sturges' rule, the number of classes (k) is given by: k = … …
(*vii*) The magnitude of the class (i) is given by: i = … …
(*viii*) … … of data is a function very similar to that of sorting letters in a post office.
(*ix*) Different bases of classification of data are … …
(*x*) The data can be classified into … … and … … type classes.
(*xi*) While forming a grouped frequency distribution, the number of classes should usually be between … …
(*xii*) In exclusive type classes, the upper limit of the class is … …
(*xiii*) In the continuous classes 0—5, 5—10, 10—15, 15—20 and so on, the class 15—20 means that the variable X takes the values … …
(*xiv*) Two examples of discrete variable are … … and … … and continuous variable are … … and … …
(*xv*) The classes in which the lower limit or the upper limit are not specified, are known as … …
(*xvi*) The difference between the upper and the lower limits of a class gives … … of the class.
(*xvii*) The number of observations in a particular class is called the … … of the class.
(*xviii*) If the data values are classified into the classes 0—9, 10—19, 20—29, and so on and the frequency of the class 20—29 is 12, it means that … … .

Ans. (*i*) discrete, continuous, (*ii*) classification. (*iii*) time. (*iv*) geographical. (*v*) the upper and the lower limits of the class. (*vi*) $k = 1 + 3\cdot322 \log_{10}N$; N is total frequency. (*vii*) i = (upper limit – lower limit) of the class. (*viii*) classification. (*ix*) geographical, chronological, qualitative and quantitative. (*x*) inclusive, exclusive. (*xi*) 5 and 25. (*xii*) is not included in the class. (*xiii*) 15 and more but less than 20 *i.e.*, $15 \leq X < 20$. (*xiv*) marks in a test, number of accidents; height in inches, weight in kgs. (*xv*) open end classes. (*xvi*) the width or the magnitude. (*xvii*) frequency.
(*xviii*) there are 12 observations taking values between 20 and 29, both inclusive *i.e.*, $20 \leq X \leq 29$.