DEFINITION OF STATISTICS, IMPORTANCE & LIMITATIONS

- The word 'Statistics' has been derived from the **Latin word 'statisticum', Italian word 'statistia' and German word 'statistik',** each of which **means a group of numbers or figures** that represent some **information of human interest**.
- Data mean information, which can be of two types –

I. Qualitative
II. Quantitative.

Statistics means quantitative or numerical data, which can be used for further calculations.

4 DISTINCT PHASES OF STATISTICAL ANALYSIS OF DATA

I. **COLLECTION OF DATA:** In this first stage of investigation, numerical data is collected from different sources.
II. **CLASSIFICATION AND TABULATION OF DATA:** The raw data is divided into different groups or classes and represented in a form of a table.
III. **ANALYSIS OF DATA:** Classified and Tabulated data is analysed using different formulas and methods
IV. **INTERPRETATION OF DATA:** At the final stage, relevant conclusions are drawn after the data is thoroughly analysed

IMPORTANCE OF STATISTICS

- Business and economics
- Medical
- Weather forecast
- Stock market
- Bank
- Sports

FUNCTION OF STATISTICS

1. Statistics present the **facts in definite form.**
2. Statistics **simplify** complex data.
3. It provides a **technique of comparison**.
4. Statistics study the **relationship between two or more variables**.
5. It helps in formulating **policies**.
6. It helps in **forecasting outcomes**.

LIMITATIONS OR DEMERITS

- **STATISTICS DO NOT DEAL WITH INDIVIDUALS**: Statistics is the study of mass data or a group of observations and deals with aggregates of facts.
- **STATISTICS DOES NOT STUDY QUALITATIVE DATA**: Statistics is the study of only those facts which are capable of being stated in number or quantity.
- **STATISTICS GIVE RESULT ONLY ON AN AVERAGE**: Sample is collected for study and draw conclusion from, as a representative for the whole.
- **THE RESULTS CAN BE BIASED:** The data collection may sometime be biased which will make the whole investigation useless.

POPULATION: It is the entire collection of observations from which we may collect data.

Example: If we are studying the weight of adult men in India, the population is the set of weights of all men in India.

SAMPLE: a part of population is selected for study.

Example: The population for a study of infant health might be all children born in India in one particular year. The sample might be all babies born on one particular day in that year.

Data can be classified into two types, based on their characteristics:

1. Variates
2. Attributes

VARIATE: A characteristic that varies from one individual to another and can be expressed in numerical terms.

Example: Prices of a given commodity, wages of workers, heights and weights of students in a class, marks of students, etc.

ATTRIBUTE: A characteristic that varies from one individual to another but can't be expressed in numerical terms.

Example: Colour of the ball (black, blue, green, etc.), religion of human, etc.

QUANTITATIVE/ NUMERICAL VARIABLES CAN BE CLASSIFIED

- **DISCRETE VARIABLE:** A variate which takes discrete or distinct value or in other words can take only a countable and usually finite number of values.

Example: Number of members in a family, Number of accidents, Age in years.

- **CONTINUOUS VARIABLE**: A variate that can take any value within a range (integral/fractional).

Example: Percentage of marks, Height, Weight.

IMPORTANT POINTS

- Samples and populations use measures such as mean, median, mode and standard deviation.
- For a sample, they are called **statistic**
- For a population, they are called **parameters**.
- A statistic is a characteristic of a sample; a parameter is a characteristic of a population,
- statisticians use lower case Roman letters to denote sample statistics and Greek or Capital letters to denote population parameters.

	POPULATION	SAMPLE
Definition	Collection of all items	Part of the population
Characteristics	Parameters	Statistics
Symbols	Size - N	Size – n
	Mean - μ	Mean - $\overline{x}$
	Standard Deviation - σ	Standard Deviation - s

COLLECTION OF DATA

PRIMARY DATA is the data which is collected directly or first time by the investigator or researcher from the respondents.

Primary data is collected by using the following methods:

- Direct Interview Method
- Questionnaires
- Census and sample survey

SECONDARY DATA

Secondary data are the **Second-hand information**. The data which have already been **collected and processed by some agency or persons** and is collected for the second time are termed as secondary data.

DISTINCTION BETWEEN PRIMARY AND SECONDARY DATA

1. The **data collected for the first time is called Primary data** and data collected through some published or unpublished sources is called Secondary data.
2. The **primary data in the hands of one person can become secondary for all others**.
3. **Primary data are original** as they are collected first time from the respondents directly or by preparing questionnaires. So, they are more accurate than the secondary data.
4. the collection of primary data requires more money, time and energy than the secondary data.

CLASSIFICATION AND TABULATION

The method of arranging data into homogeneous group or classes according to some common characteristics present in the data is called Classification.

- Classification condenses the data by **removing unimportant details.**
- It enables us to accommodate large number of observations into few classes and **study the relationship between several characteristics**.
- Classified data is presented in a **more organised way** so it is easier to interpret and compare them, which is known as Tabulation.

THERE ARE FOUR IMPORTANT BASES OF CLASSIFICATIONS:

1. **QUALITATIVE BASE:** Here the data is classified according to some quality or attribute such as age, religion, literacy, intelligence, etc.
2. **QUANTITATIVE BASE:** Here the data is classified according to some quantitative characteristic like height, weight, age, marks, etc.
3. **GEOGRAPHICAL BASE:** Here the data is classified by geographical regions or location, like states, cities, countries, etc. like population in different states of India.
4. **CHRONOLOGICAL OR TEMPORAL BASE:** The data is classified or arranged by their time of occurrence, such as years, months, weeks, days, etc. This classification is also called Time Series data. Example: Sales of a company for different years.

TYPES OF CLASSIFICATION

- If we classify observed data for a **single characteristic**, it is known as **One-way Classification**.

Example: Population can be classified by Religion - Hindu, Muslim, Christians, etc.

- If we consider **two characteristics at a time** to classify the observed data, it is known as a **Two-way classification**.

Example: Population can be classified according to Religion and age.

- If we consider **more than two characteristics** at a time in order to classify the observed data, it is known as **Multi-way Classification**.

Example: Population can be classified by Religion, age and literacy.

FREQUENCY DISTRIBUTION

Frequency: If the **value of a variable** (discrete or continuous) e.g., height, weight, income, etc. **occurs twice or more in a given series of observations,** then the **number of occurrences of the value** is termed as the **"frequency" of that value.**

Intervals are normally of equal size covering the sample observations range.

CLASS-LIMITS OR CLASS INTERVALS:

A class is formed within the two values, class-limits or class-intervals. The **lower value is called lower class limit** or lower-class interval and the **upper value is called upper class limit** or upper-class interval.

CLASS LENGTH OR CLASS WIDTH

Class Length → Class Width → Upper Class Interval - Lower Class Interval

MID-VALUE OR CLASS MARK

The mid-point of the class is called mid-value or class mark.

$$\boldsymbol{Class\ Mark = \frac{Lower\ class\ limit + upper\ class\ limit}{2}}$$

TYPES OF CLASS INTERVALS

There are two types of class-intervals

a. Exclusive type

b. Inclusive type

EXCLUSIVE TYPE: Class intervals like 0-10, 10-20, 500-1000, 1000-1500. Here the **upper limits of the classes are excluded from the respective classes** and put in the next class while considering the frequency of the respective class.

INCLUSIVE TYPE: Class intervals like 60-69, 70-79, 80-89, etc. are inclusive type. **Here both the lower- and upper-class limits are included in the class-intervals** while considering the frequency of the respective class, e.g., 60 and 69 are both included in the class 60-69.

CLASS BOUNDARIES

Inclusive classes can be converted to exclusive classes and the new class intervals are called class boundaries.

Example: The classes 5-9, 10-14 can be converted to exclusive type of classes using the formula:

New UCI = Old UCI + (10-9)/2 = 9 + 0.5 = 9.5.

New LCI = Old LCI - (10-9)/2 = 5 - 0.5 = 4.5.

So the class-boundaries are 4.5-9.5, 9.5-14.5, etc.

OPEN-END CLASS INTERVAL

In open-end class interval either the lower limit of the first class or upper limit of the last class or both are missing.

Example: Below 10, Above 40

RELATIVE FREQUENCY

$$\frac{Frequency}{Total\ Frequency}$$

PERCENTAGE FREQUENCY

$$\frac{Frequency}{Total\ Frequency}*100$$

Percentage frequency of the class interval = (12/32) x 100 = 37.5.

FREQUENCY DENSITY

Frequency density of a class interval = $\frac{Class\ Frequency}{Width\ of\ Class}$

Frequency Distribution is of two types.

- Discrete Frequency Distribution
- Continuous Frequency Distribution

1. Discrete Frequency Distribution: Variable takes distinct values.

Problem 1: 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

Solution: This data can be classified into a discrete frequency distribution.

Number of Post-graduates (x)	***Frequency***
0	2
1	2
2	4
3	1
4	1

2. Continuous Frequency Distribution:

Variable takes values which are expressed in class intervals **within certain limits.**

Marks obtained by 20 students in an exam for 50 marks are given below- convert the data into continuous frequency distribution form.

18, 23, 28, 29, 44, 28, 48, 33, 32, 43, 24, 29, 32, 39, 49, 42, 27, 33, 28, 29.

Marks	Frequency
15-20	**1**
20-25	**2**
25-30	**7**
30-35	4
35-40	1
40-45	3
45-50	2