13A. STATISTICAL DESCRIPTION OF DATA

CONCEPT 01 : INTRODUCTION TO STATISTICS

Definition	**Singular Noun**	The **scientific method** that is employed for collecting, analysing & presenting data, leading finally to drawing statistical inferences about important characteristics. It is **Science of Counting** or **Science of Averages**.
	Plural Noun	**Data, qualitative as well as quantitative**, that are collected, usually with a view of having statistical analysis.

History of Statistics

- **Origin of the word Statistics** : This is a debatable topic but various theories are as follows :
  - **(a) Latin Word :** Status
  - **(b) Italian Word :** Statista
  - **(c) German Word :** Statistik
  - **(d) French Word :** Statistique
- The first census was conducted by Pharaoh in Egypt during 300BC to 2000BC.
- Kautilya (Chanakya) kept a record of births & deaths and some other precious information in his book 'Arthashastra' during Chandragupta Maurya's reign in 4th Century BC.
- Statistical records on Agriculture are also found in Ain-i-Akbari by Abul Fazl in 16th Century AD.

Application	**Economics**	▪ **Econometrics :** Branch of Economics that interacts with Statistics ▪ Time Series Analysis, Index Numbers, Regression Analysis, Demand Analysis etc. are some overlapping areas.
	Business Management	**Statistical Decision Theory** focuses on the analysis of complicated business strategies with a list of alternatives.
	Commerce & Industry	Data on previous sales, raw materials, wages etc. are collected, analysed & experts are consulted in order to maximise profits.

Limitations

(a) **It deals with aggregates.** An individual has no significance.
(b) **It is concerned with quantitative data.** However, qualitative data can also be converted into quantitative data by assigning a numerical value.
(c) **Future projections are possible under specific set of conditions.** If any of these conditions are violated, projections are likely to be incorrect.
(d) **The theory of statistical inferences is built upon random sampling.** If the rules of random sampling are not strictly adhered to, the conclusion drawn based on these unrepresentative samples would be erroneous.

CONCEPT 02 : COLLECTION OF DATA

What is Data?

- Data is a quantitative information about some particular characteristic(s) under consideration.
- A **Variable** is a measurable quantity i.e. quantitative information.

Discrete Variable	If it can assume only finite or countably infinite no. of isolated values. *Examples :* *(a) No. of Petals in a Flower* *(b) No. of Misprints in a Book* *(c) No. of Road Accidents in a Locality* *(d) Annual Income of a Person* *(e) Marks of a Student* *(f) Distribution of Shares* *(g) Salary of a Person*
Continuous Variable	If it can assume any value from a given interval. *Examples :* *(a) Height, Weight, Age of a Person* *(b) Sales / Turnover of a Company* *(c) Distribution of Profits of a Company*

- An **Attribute** is a qualitative characteristic.
  Examples :
  (a) Gender of a Baby,
  (b) Nationality of a Person
  (c) Colour of a Person,
  (d) Drinking Habit of a Person, etc.

<table>
<tr><td>Classification of Data</td><td colspan="2">PRIMARY DATA</td><td>SECONDARY DATA</td></tr>
<tr><td></td><td colspan="2">The data which is collected for the first time by an investigator or an agency.</td><td>Data, already collected, used by a different person or agency.</td></tr>
<tr><td rowspan="6">Methods of Collecting Primary Data</td><td colspan="3">[A] INTERVIEW METHOD</td></tr>
<tr><td>Personal Interview</td><td colspan="2">▪ The investigator meets the respondents directly & collects the required information then & there, from them.
▪ In case of a Natural Calamity, this method can be quick & accurate.</td></tr>
<tr><td>Indirect Interview</td><td colspan="2">▪ The investigator collects necessary information from the persons associated with the problems.
▪ If there are some practical problems in reaching the respondents directly (like in a rail accident), then this method can be used.</td></tr>
<tr><td>Telephonic Interview</td><td colspan="2">▪ The information can be gathered by researcher by contacting the interviewee over the phone.
▪ It is a quick & non-expensive way to collect data.
▪ It is less consistent method, but has a wide coverage.
▪ Non-Responses are maximum in this method.</td></tr>
<tr><td colspan="3">[B] QUESTIONNAIRE METHOD</td></tr>
<tr><td>Mailed Questionnaire</td><td colspan="2">▪ It involves framing of a well-drafted & soundly-sequenced questionnaire covering all the important aspects of the problem & sending them to the respondents with pre-paid stamp & necessary guidelines.
▪ It has a wide coverage but non responses are maximum in this method.</td></tr>
<tr><td>Questionnaire by Enumerator</td><td colspan="2">▪ Enumerators collect information directly by interviewing the persons having information by explaining the questions.
▪ It is used for larger enquiries from persons who are being surveyed.</td></tr>
<tr><td colspan="3">[C] OBSERVATION METHOD</td></tr>
<tr><td colspan="3">▪ In this method data is collected by direct observation or using an instrument (like height or weight of a group of students).
▪ It is time consuming & laborious method having small coverage.
▪ It is the best method for data collection.</td></tr>
<tr><td>Sources of Secondary Data</td><td colspan="3">1. International Sources : WHO, ILO, IMF, World Bank etc.
2. Government Sources : Statistical Abstract by CSO, Indian Agricultural Statistics etc.
3. Private & Quasi Govt. Sources : ISI, ICAR, NCERT etc.
4. Unpublished Sources of various research institutes, researchers etc.</td></tr>
<tr><td>Scrutiny of Data</td><td colspan="3">▪ Since the statistical analysis are made only on the basis of data, it is necessary to check whether the data under consideration are accurate as well as consistent.
▪ No hard & fast rules can be recommended for scrutiny of data. One must apply his intelligence, patience & experience while scrutinising the given information.
▪ If two or more series of related data are given, they may be checked for Internal Consistency.
E.g. If data for Population, Area & Density for some places are given, then we may verify that they are internally consistent by examining whether the relation, Density = Population / Area holds.
▪ A good enumerator can also detect whether the returns submitted by some enumerators are exactly of the same type, thereby implying the lack of seriousness on the part of enumerators.
▪ The bias of enumerators may also be reflected by the returns submitted by him.
Rectification : This type of errors can be rectified by asking the enumerator(s) to collect the data of disputed cases, once again.</td></tr>
</table>

CONCEPT 03 : CLASSIFICATION / ORGANISATION OF DATA

Definition

The process of arranging data on the basis of the characteristic under consideration into a number of groups or classes according to the similarities of the observations.

Characteristic	Type of Data	Example
	NON-FREQUENCY DATA	
Time Points / Intervals	**Chronological / Temporal / Time Series Data**	*No. of students appeared in CA Foundation in last 20 years.*
Region	**Spatial Series / Geographical Data**	*No. of students appeared in CA Foundation in 2025 in accordance with different states.*
	FREQUENCY DATA	
Variable	**Quantitative / Cardinal Data**	*Height, Weight, Profits*
Attribute	**Qualitative / Ordinal Data**	*Gender, Nationality*

Advantages

(a) It puts data in a neat, precise and condensed form so that it is **easily understood & interpreted**.
(b) It makes **comparison possible** between various characteristics.
(c) **Statistical Analysis** is possible only for the classified data.
(d) It **eliminates unnecessary details** & makes data more readily understandable.

CONCEPT 04 : PRESENTATION OF DATA

4.1 TEXTUAL PRESENTATION OF DATA

Definition

- This method comprises presenting data with the help of a paragraph(s).
- The official reports of enquiry commissions are usually made through this method.

Example

"In 2024, out of a total of five thousand workers of Roy Enamel Factory, four thousand and two hundred were members of a trade union. The total number of female workers was eight hundred & six per cent of total members were members of the Trade Union.

In 2025, the number of workers belonging to the trade union was increased by twenty per cent as compared to 2024 of which four thousand and two hundred were male. The number of workers not belonging to trade union was nine hundred and fifty of which four hundred and fifty were females."

Advantages	Disadvantages
▪ Simplicity ▪ Observations with exact magnitude can be presented ▪ First step towards other modes of presentation	▪ Dull & Monotonous ▪ Comparison is not possible ▪ Cannot be recommended for manifold classification

4.2 TABULAR PRESENTATION / TABULATION

Definition

A systematic presentation of data with the help of statistical table having a number of rows & columns & complete with reference number, title, description of rows as well as columns and foot notes, if any.

Guidelines for Tabulation

Table Number
Title

Stub (Row Heading)	***Caption (Column Heading)***				***Total (Rows)***
	Sub-Head		***Sub-Head***		
	Column Head	***Column Head***	***Column Head***	***Column Head***	
⋮ ***Stub Entries (Row Entries)*** ⋮	***Body***				
Total(Columns)					

Source :
Footnotes :

1) Serial Number with a Self-Explanatory Title
2) Table should be divided into Caption, Box-Head (Entire Upper Part), Stub & Body.
3) Table should be well-balanced in length & breadth.
4) Data must be arranged in such a way that comparisons are facilitated.
5) Row Totals, Column Totals, Units of Measurement must be shown.
6) Data should be arranged intelligently and appealing to the eyes as far as possible.
7) Source of Data and Footnotes should be shown (if any).

TERMINOLOGY

(a) Box Head : The entire upper portion of the table which includes columns & sub-column numbers, unit(s) of measurement & caption.
(b) Caption : Upper Portion of the table describing columns & sub-columns.
(c) Stub : Left Portion of the table providing description of the rows.

Example

Table 13.1
Status of the workers of Roy Enamel Factory on the basis of their trade union membership for 2024 & 2025

Status	Member			Non-Member			Total		
Year	M	F	T	M	F	T	M	F	T
2024	3900	300	4200	300	500	800	4200	800	5000
2025	4200	840	5040	500	450	950	4700	1290	5990

Source :
Footnote : M stands for Male, F stands for Female & T stands for Total.

Advantages

- Facilitates comparison between rows & columns.
- Complicated data can also be represented using tabulation.
- It is a **must for diagrammatic representation**.
- It is a **must for statistical analysis**.
- It is the **most accurate & best method** of presentation of data.

4.3 DIAGRAMMATIC PRESENTATION

Basics

- This can be used for **both educated & uneducated section of society**, unlike the previous two.
- Any **hidden trend** can be noticed only in this mode of presentation.
- If there is a priority for accuracy, recommend tabulation. **It is less accurate than tabulation.**
- It is the **most attractive method** of presentation of data.

Line Diagram /Historiagram

It is used for **Time Series** Data.

(a) If wide range of Fluctuations
Logarithmic / Ratio Chart
(b) Multiple Time Series Data
- **Same Unit :** Multiple Line Chart
- **Distinct Unit :** Multiple Axis Chart

Bar Diagram

(a) Time Series Data / Quantitative Data
Vertical Bar Diagram
(b) Spatial Data / Qualitative Data
Horizontal Bar Diagram
(c) For Comparing Related Series
Multiple / Grouped Bar Diagram
(d) For Representing Data into Parts
Component / Sub-Divided Bar Diagram
(e) For Comparing Different Components or Relating the Components to the Whole
Divided / Percentage Bar Diagram

Pie Chart

- For Circular Presentation of Data & For Comparing Different Components or Relating the Components to the Whole.
- **Segment Angle**
  Segment Value / Total Value x 360°

CONCEPT 05 : FREQUENCY DISTRIBUTION

Frequency

No. of times a particular observation / class occurs.

Frequency Distribution

It is a statistical table that distributes the total frequency to a number of classes.

TYPES OF FREQUENCY DISTRIBUTION

DISCRETE / UNGROUPED / SIMPLE F.D.	GROUPED F.D.
When tabulation is done in respect of a Discrete Random Variable & frequency is assigned to each one of them.	When tabulation is done in respect of a Continuous Variable & frequency is assigned to a group of values & not individual values. **Types of Grouped Classification** (a) Non-Overlapping / Mutually Inclusive (b) Overlapping / Mutually Exclusive

Mutually Inclusive Classification

Class Interval	x
01 – 10	
11 – 20	
21 – 30	

CI includes Upper Class Limit. (10 will be counted in 01-10 CI)

Mutually Exclusive Classification

Class Interval	x
01 – 10	
10 – 20	
20 – 30	

CI excludes Upper Class Limit. (10 will be counted in 10-20 CI)

Basic Terminology

Class Limit (CL)

The Minimum Value (Lower Class Limit) & Maximum Value (Upper Class Limit) of a Class Interval (CI).

Class Boundary (CB)

- ***Mutually Exclusive Classification :*** *Class Boundary = Class Limits.*
- ***Mutually Inclusive Classification :***
  Class Boundary will be calculated as follows :

LCB = LCL – D/2	*UCB = UCL + D/2*
LCB = 10 – ½ = 9.5	*UCB = 19 + ½ = 19.5*
9.5, 19.5, 29.5, 39.5 etc.	*19.5, 29.5, 39.5 etc.*

**D = Difference between UCL of given CI & LCL of next CI.*

Class Length

Class Length / Width / Size = UCB – LCB

Class Mark

Mid-Point / Mid-Value = $\frac{\text{LCL + UCL}}{2} = \frac{\text{LCB + UCB}}{2}$

Frequency Density

= Frequency of a CI / Class Length

Relative Frequency

= Class Frequency / Total Frequency [They add up to 1 (Unity)]

% Frequency

= Class Frequency / Total Frequency x 100 [They add up to 100%]

Note : *If question asks to calculate No. of Class Intervals, then follow these steps :*

- ***Step 01 :*** *Calculate Range (Largest Observation – Smallest Observation).*
- ***Step 02 :*** *No. of CI = Divide Range by Class Length required (Rounded Up).*
  (Question might give 1+3.322logN as an option. Learn this if asked again.)

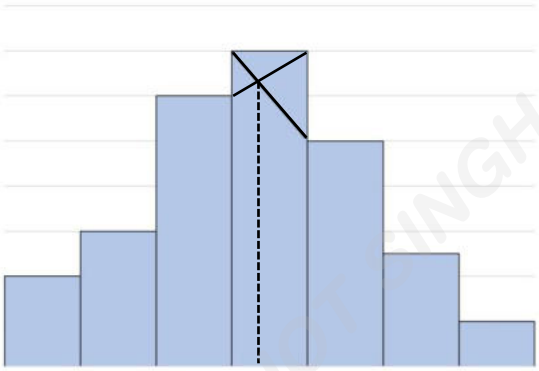
Cumulative Frequency

TYPES OF CUMULATIVE FREQUENCY

Class Interval	Frequency (f)	Less than	CF	More than	CF
10 – 15	20	Less than 15	20	More than 10	100 (80 + 20)
15 – 20	32	Less than 20	52 (20 + 32)	More than 15	80 (48 + 32)
20 – 25	18	Less than 25	70 (52 + 18)	More than 20	48 (30 + 18)
25 – 30	30	Less than 30	100 (70 + 30)	More than 25	30
	$\sum f = 100$				

For any Class Boundary, Less than CF + More than CF = $\sum f$

CONCEPT 06 : GRAPHICAL REPRESENTATION OF FREQUENCY DISTRIBUTION

Histogram (Area Diagram)	▪ It is a very **convenient way** to represent frequency distribution. ▪ Comparison among different CI's is possible. ▪ It is used to calculate **Mode**.  **MODE**
Frequency Polygon	▪ It is usually meant for Simple / Ungrouped Frequency Distribution. ▪ But it can also be used for Grouped Frequency Distribution, provided the width of the Class Intervals remains the same (Use Mid-Points). ▪ We can also obtain a Frequency Polygon starting with a Histogram, by adding the mid-points of the upper sides of the rectangles successively & then completing the figure by joining the two ends.
Ogives	▪ By plotting cumulative frequency against the respective class boundary, we get ogives. ▪ By plotting less than cumulative frequency, we get **Less than type Ogive**. ▪ By plotting more than cumulative frequency, we get **More than type Ogive**. ▪ They can be considered for obtaining **Quartiles** graphically. ▪ If a perpendicular is drawn from the point of intersection of the two ogives on the horizontal axis, then the x-value of this point gives us the value of **Median**. ▪ They can also be used for making **short term projections**.
Frequency Curve	▪ It is a limiting form of a Histogram or Frequency Polygon. ▪ It can be obtained by drawing a smooth & free hand curve through the mid points of the upper sides of the rectangles forming Histogram. ▪ Total Area is taken to be 1 (Unity). ▪ **X – Axis :** Class Boundary & **Y – Axis =** Frequency Density

TYPES OF FREQUENCY CURVE

Bell Shaped Curve	U-Shaped Curve	J-Shaped Curve	Mixed Curve
Frequency is maximum near central part & minimum near extremities. It is the most commonly used curve.	Frequency is minimum near the central part & maximum near extremities.	Starts with minimum frequency & reaches maximum at other extremity.	Combination of different curves.