

CHAPTER - 13

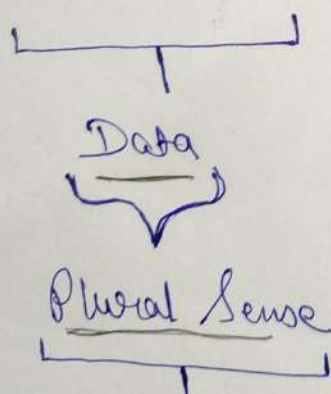
STATISTICAL DESCRIPTION OF DATA

HISTORY OF STATISTICS

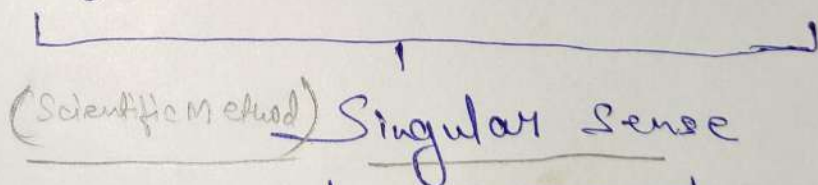
- * Latin - Status, Italian - statista,
Germany - statistik, French - statique.
- * Kautilya → Arthashastra → record of Birth and Deaths → fourth century B.C
Chandragupta's region → fourth century B.C
4CE
- * Abu Fazl → Ain-i-Akbari → Agriculture
→ Akbar → 1600 AD century, 16CE
- * Pharaoh → Egypt → the first census → 300 B.C to 2000 B.C.

DEFINITION OF STATISTICS

Statistics : [Collection, analysis & Presentation of data]

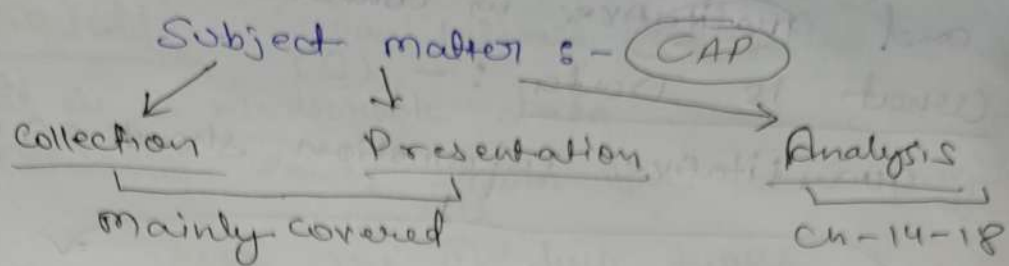


Data qualitative
or quantitative
collected to do
Statistical
analysis



used to draw statistical
inferences

Inferences means
— conclusion reached
based on evidence
and reasoning.



Application of Statistics :-

- > Economics
- > Business management
- > Industry & Commerce, etc.

• Limitations :-

- Aggregate data -
2, 7, 6, 4] Group

Individual Data -

2] single.

> Relevant for aggregate data and not individual data.

- Stats deal with only quantities
- But concerned with both quantity & quality.

quantitative → Expressed in numbers only.

qualitative → Not expressed in numbers.

So mainly quantitative data is used

and qualitative is converted into numbers.
What is Data :- (2)

quantitative information shows as number

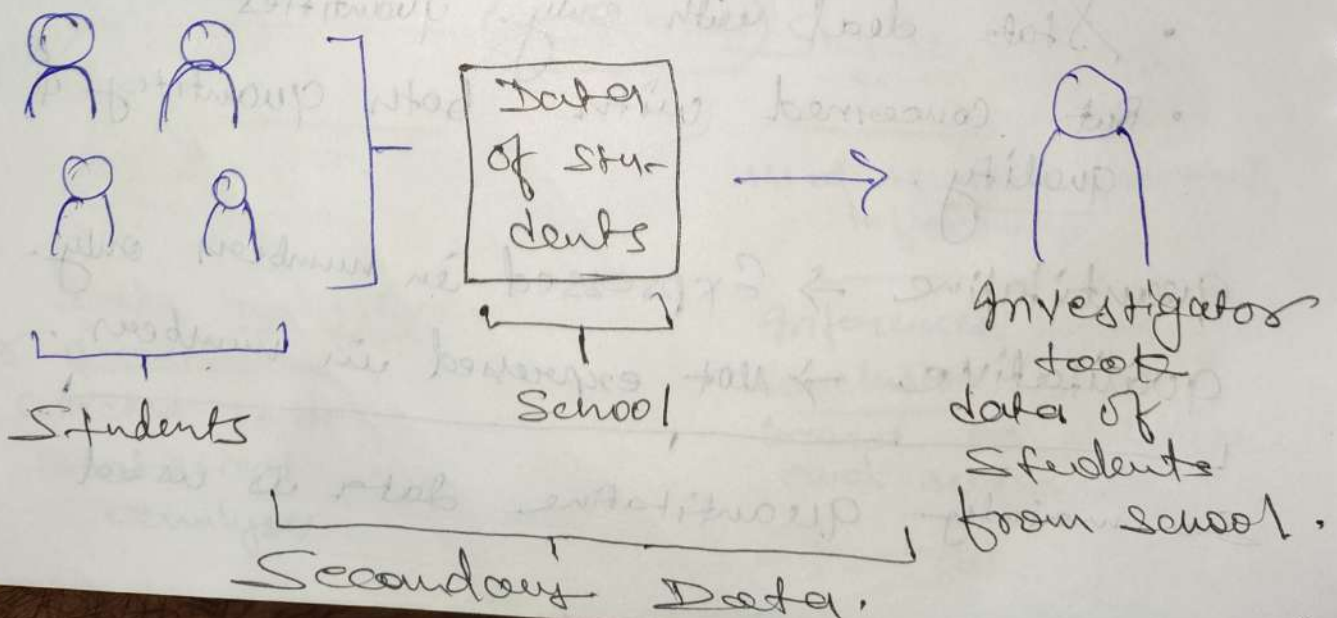
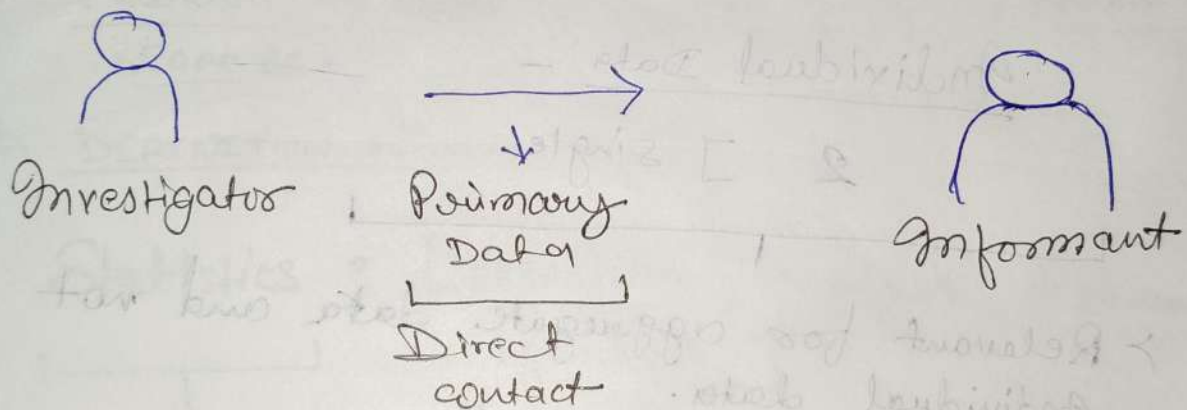
Primary Data
(unused data, fresh data, raw data, unprocessed data)

First collected
by agency,
investigator

Secondary Data

(used data, processed data)

Collected data used
by different
person / agency.



★ Attribute — qualitative (ordinal)
★ Variable (quantitative) (cardinal)

> It is a measurable data.

> Variables comes from the word vary

These are of two types

— Discrete variable

— Continuous variable



Now the Constant is :-

No of factors of number 8
 $= 4$

1, 2, 4, 8] It will not change

Variable : No of students enrolled for
CA foundation] will change.

Variable

Discrete
Variable

↓
Isolated value
no continuity

2
3
8
20
12

Gap is
not
equal

Continuous
variable

↓
continuity

42
43
44
45
46

• Gap of
1

★ Discrete Variable

When variable] or count ably infinite
Assumes a finite isolated values.

For eg : no of petals in a flower
: no of Heads throwing coin
for 2 times.

★ Continuous Variable

When there is continuity] A variable assumes
any value from
the given interval

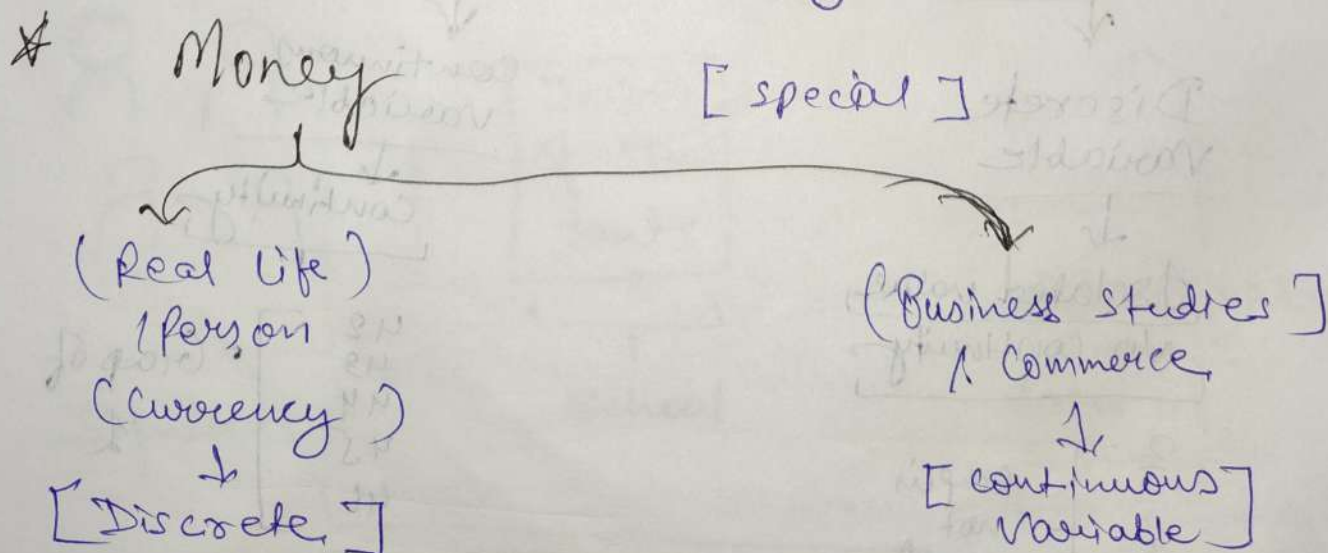
For eg : Can also be in decimals.

Age of Person in a Camp

8.25 , 9.37 , 10.35 , 10.85 , 11.25

But Continuous -

ex : Height , Sale
Weight , money

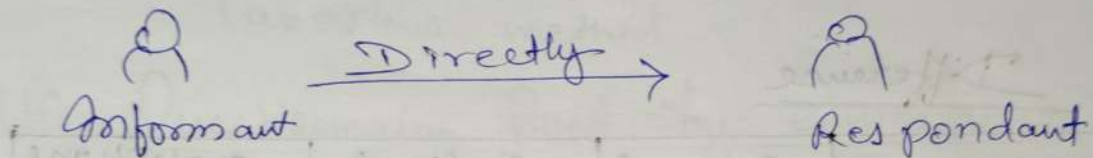


COLLECTION OF DATA

will be of $\rightarrow$ Primary data
[Bcz. secondary data is already collected]

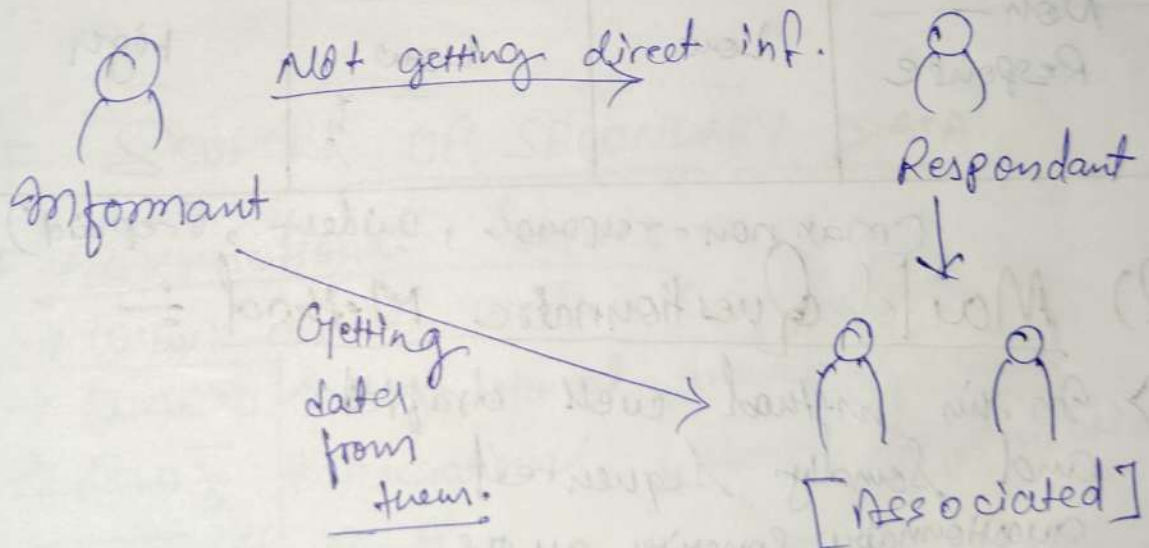
(A) INTERVIEW METHODS: - (4)

(a) Personal / Direct Interview - (NC)



(b) INDIRECT INTERVIEW:- (HB)

When we didn't get direct information from respondant we will get this from his associated persons.



(c) Telephonic Interview:-

quick, non-expensive, wide coverage.

Methods	Examples
★ <u>Personal Interview</u>	Natural calamity, door to door survey
★ <u>Indirect Interview</u>	Rail accident (Human Behaviour)
★ <u>Telephone Interview</u>	By calling / messaging

Difference

Parameter	Personal	Indirect	Telephone
Accuracy	High	Medium	Low
Coverage	Low	Low	High
Non-Response	Low	Low	High

(Max non-response, widest, Prepaid)

(2) Mailed Questionnaire Method :-

- > In this method well drafted and soundly sequenced questionnaire covering all the important aspects of the data requirement is sent to respondent for filling
- > Here coverage is wide but amount of non-response will be maximum.

Q _____

Q _____

Q _____

Online / form → Email

(3) Observation method :-

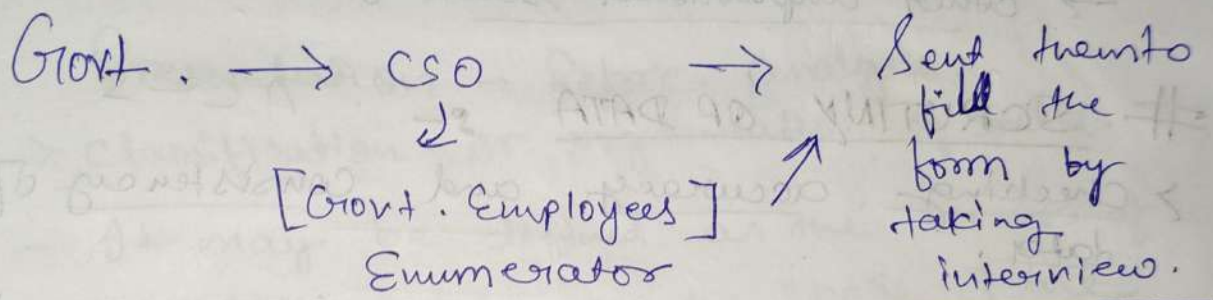
(Smallest area, Best method, Time consuming)

> In this method data is collected by direct observation or using instrument

> Data on weight or height [Example]

> Although more accurate -
Low Coverage -
Laborious method -

(4) Questionnaire filled by enumerators :-



SOURCES OF SECONDARY DATA

★ International Sources -

- (WHO) world health Organisation
- (IMF) International monetary fund
- (ILO) International Labour Organisation
- world - Bank

★ Government Sources :-

(In India)

- central statistics office (CSO)

- National Sample Survey Office (NSSO)
- RBI, SBI
- RERA
- IRDA

* Private or Quasi govt. sources -

- Indian Statistical Institute (ISI)
- Indian Council of Agriculture
- NCERT

Others

- Research Papers and
- Other unpublished sources.

SCRUTINY OF DATA :- AC-S

> checking accuracy and consistency of data.

> There is no rule for it, one must apply his intelligence, Patience, experience while scrutinizing the given information.

> Internal Consistency :- when two or more series of related data are given, we should check consistency among them

For eg : Age - 10-7-2000
DOB - 21

Related

> Types of Error

Type I - Rejection of true hypothesis

Type II - Non Rejection of false hypothesis

§ Density = $\frac{\text{Population}}{\text{Area}}$ holds.

> The elimination of false, fake or biased data is called scrutiny of data.

PRESENTATION OF DATA :-

Before analysis

↓
Raw data

After analysis

↓
Summaries / Report

Presentation - Before analysis

> Classification or organisation of Data.

- It may be defined as the process of arranging data on the basis of the characteristics under consideration into a number of groups or classes according to the similarities of the observations

> objectives - (no unnecessary details)

- Neat, Precise, condensed Data
- Facilitate comparison
- Statistical Analysis

* Classified as - (1)

- Non Frequency
- (i) Chronological or Temporal (Time Series) data. Wrt Time
Eg: Revenues year on year.
 - (ii) Geographical or Spatial Series data. Wrt Space
Arrangement based on regions
Eg: Country wise revenue of a company

On the basis of quality :-

- Frequency
- Qualitative or Ordinal data. wrt quality
Based on some attribute
Eg: Olympic medals by a company.
 - Quantitative or Cardinal Data. wrt quantity
Based on some Variable

2, 2, 3, 3, 3, 5, 4, 5, 2, 4

V	F
2	3
3	3
4	2
5	2

Classification	{	Temporal	-	Time
		Spatial	-	Space / Region
		Ordinal	-	Qualitative
		Cardinal	-	Quantitative

② MODE OF PRESENTATION :- ③

(i) • Textual Presentation -

There is a company which has 30 male and 40 female, 20 are married and all

↓
This is Textual Presentation.

→ most simplified method (Report)

(ii) • Tabular Presentation or Tabulation -

→ most common method (Best method)

• ~~even data~~ is preferred over textual form because of the below reasons.

→ Useful in easy comparison.

→ Complicated data can be presented

→ Table is must to create a diagram.

→ No analysis possible without table.

• when data is shown in the form of table.

(iii) • Diagrammatic Presentation

> Can be helpful for layman (without having much knowledge of numbers)

> Hidden trend can be traced

> But table is more accurate than diagrams.

→ most Attractive method.

(b) Tabulation Presentation

Serial No-24 : Title (Self Explanatory)

Symbol Test	A/c		Law		P-A		Eco	
	B	G	B	G	B	G	B	G
T-1								
T-2							*	
T-3								
T-4								

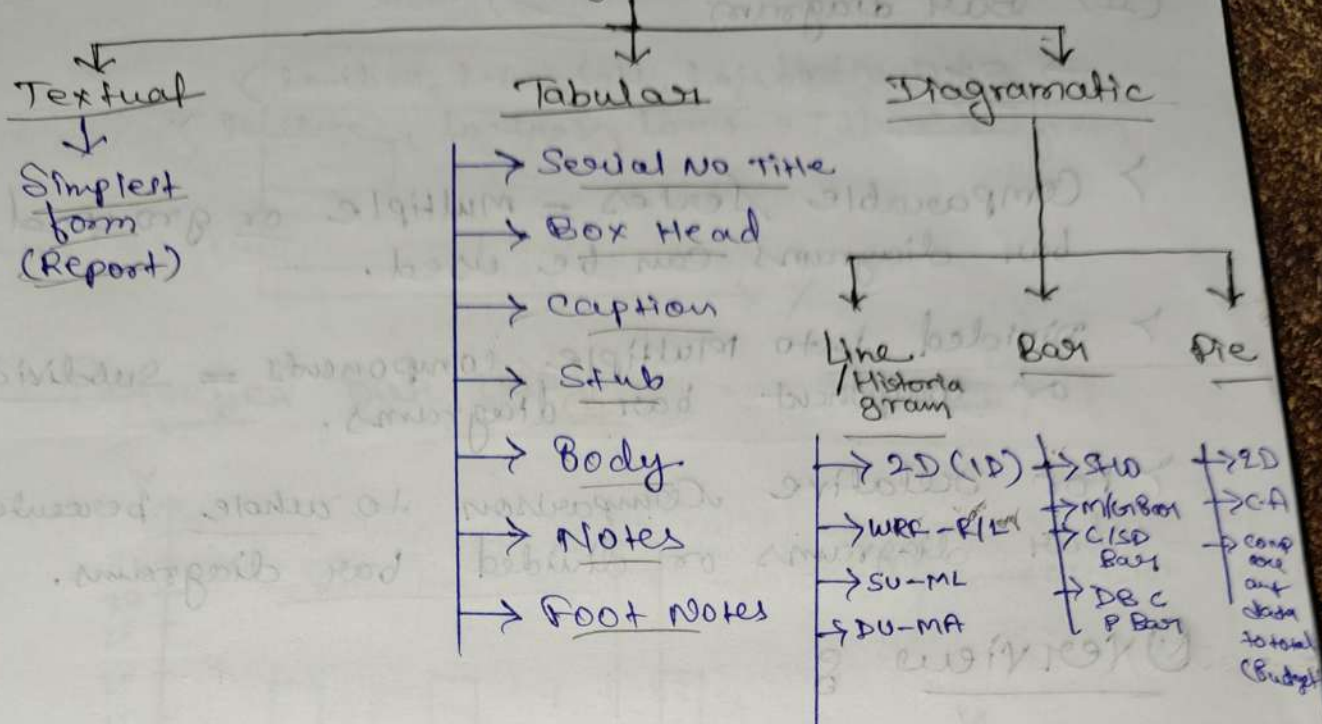
Stab

Body (Main Part)

Bottom
Part of
table

Notes : Source of data
Foot Notes : Classification about any data.

Mode Of Presentation of Data

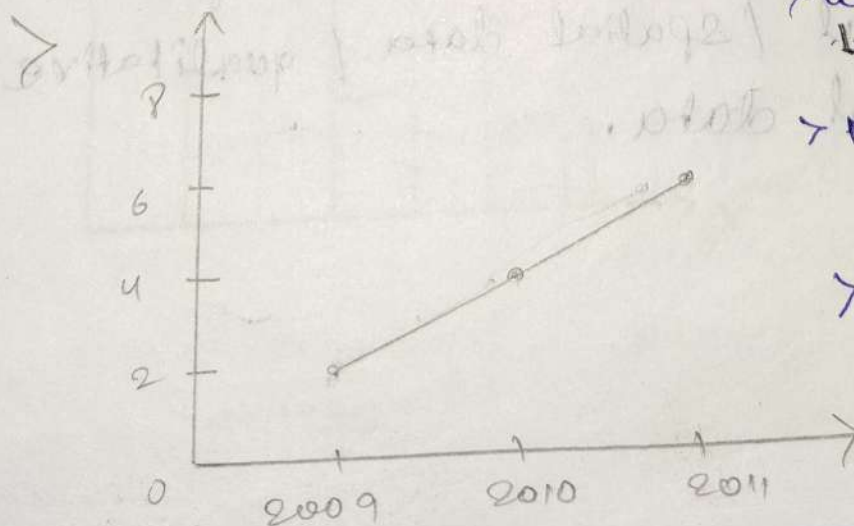


(C) Diagrammatic representation of data.

Types of diagrams :- (3)

- Time
- (I) Line diagram or Histogram 2D (simplest form-1D)
 - (II) Bar Diagram $\rightarrow$ 1D
 - (III) Pie chart $\rightarrow$ 2D
 - (IV) Cuboidal $\rightarrow$ 3D

(I) Line Diagram :- (Time Series) (2D)



- > wide Range fluctuation
log or ratio chart
- > Multiple line chart
 $\rightarrow$ 2 or more related time series (same unit)
- > Multiple axis chart
(distinct unit)
 $\rightarrow$ 2 or more series

(II) Bar diagram (1D)

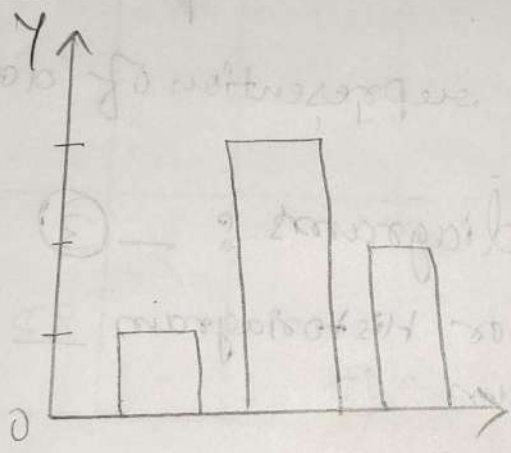
→ SHO (Spatial, Horizontal, ordinal)
→ TVC (Temporal, Vertical, Cardinal)

> Comparable series = multiple or grouped bar diagrams can be used.

> Divided into multiple components = subdivided or component bar diagrams.

> For relative comparison to whole percentage bar diagrams or divided bar diagrams.

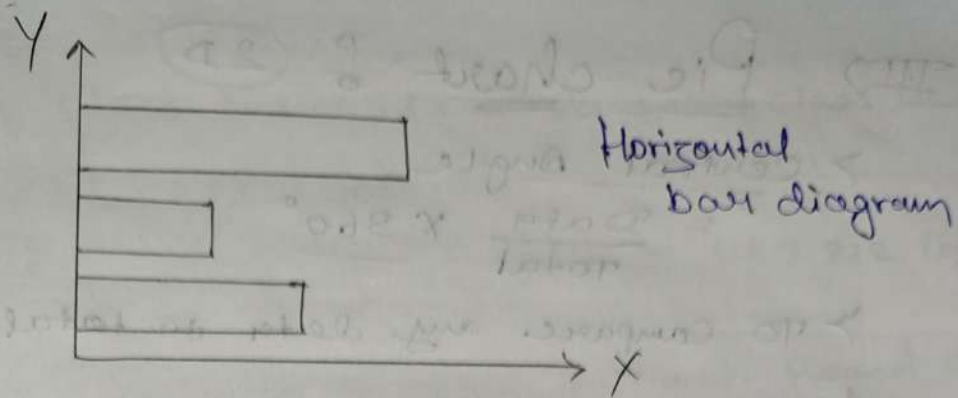
Overview :



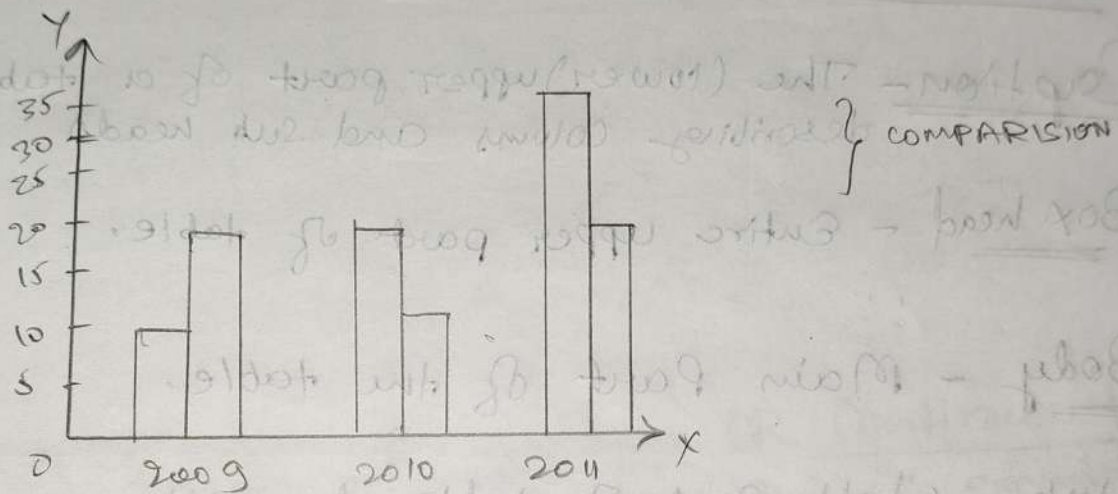
(Vertical)

* Vertical bar diagram is used for time series data or quantitative data / cardinal data.

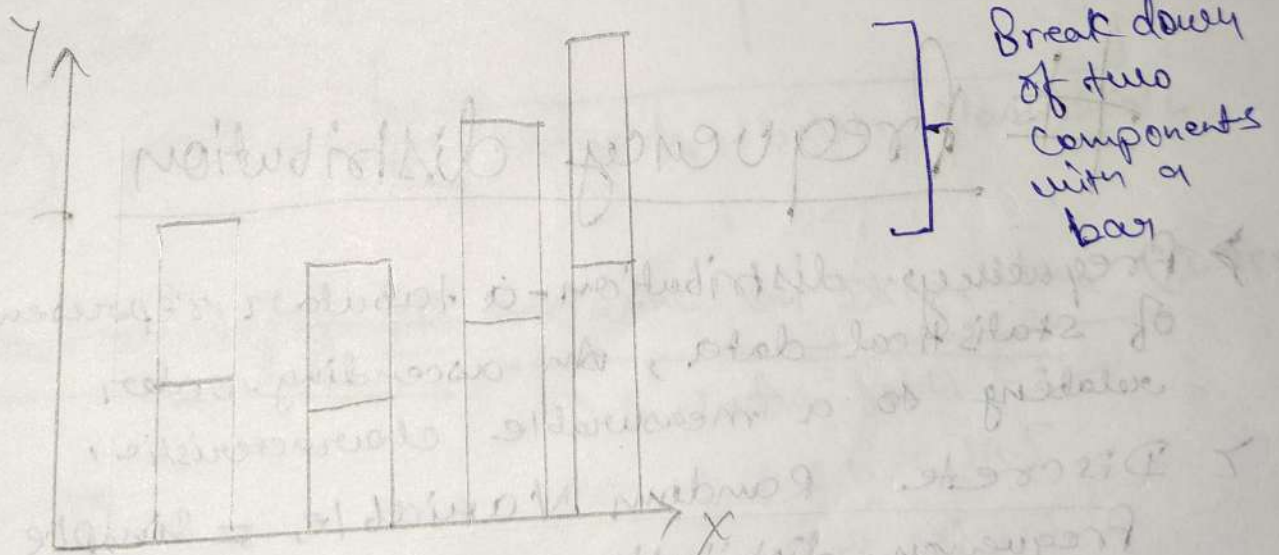
* Horizontal bar diagram is used for geographical / spatial data / qualitative or ordinal data.



Grouped Bar diagram



COMPONENT BAR DIAGRAM



(III) Pie chart : (2D)

> Central Angle

$$= \frac{\text{Data}}{\text{Total}} \times 360^\circ$$

> TO compare any data to total (Budget)

About Table :

1. Caption - The (lower) upper part of a table describing columns and sub heads
2. Box head - Entire upper part of table.
3. Body - Main part of the table.
4. Stub - Left part of table describing rows.
5. Footnote : A note mentioning source of data.

Frequency distribution (Imp)

- > Frequency distribution - a tabular representation of statistical data, in ascending order relating to a measurable characteristic.
- > Discrete Random Variable = Simple frequency distribution.
- > Continuous Variable = Grouped frequency distribution.

NOTE :

1. No of class intervals = Range / class lengths

2. STURGES' FORMULA

No of class intervals (K) = $1 + 3.22 \log N$
when N is very large

Ⓢ For case no of class interval round off always high side.

Ex max = 73 min = 47, CL = 5

$$\text{No of CI} = \frac{73 - 47}{5} = 5.2$$

(a) 4 (b) 5 (c) 6 (d) None.

* Overlapping OR Exclusive OR Continuous (variable)
↳ upper limit is excluded

0 → 10
10 → 20
20 → 30

$$LCL = 10$$

$$UCL = 20$$

$$LCB = 10$$

$$UCB = 20$$

C.B coincide with class limit.

* Non-overlapping OR Inclusive OR Discontinuous (Discrete)

↳ includes both C.L

10 → 19

20 → 29

30 → 39

$$LCL = 20$$

$$UCL = 29$$

$$LCB = 19.5$$

$$UCB = 29.5$$

Ex

CI	f	Less than C.F.	more than cf
0-20	2	2	23
10-20	5	7	21
20-30	3	10	16
30-40	7	17	13
40-50	6	23	6

$$\Sigma f = 23$$

- (i) find less than cf for 20-30 = 10
 (ii) find more than cf for 10-20 = 21
 (iii) find cf for 30-40 = 17 or 13

Ex

CI	0-9	10-19	20-29	30-39	40-49	50-59	60-69
freq.	2	3	7	8	4	6	10

- (i) find FDC for 30-39 = $\boxed{FDC = \frac{f}{\Sigma f}} = \frac{8}{10} = 0.8$
 (ii) find RFC for 40-49 = $\frac{4}{40} = 0.1$

$$\boxed{RFC = \frac{f}{\Sigma f}}$$

- (iii) And PFC for 20-29 = $\boxed{PFC = \frac{f}{\Sigma f} \times 100}$
 $PFC = \frac{7}{40} \times 100 = 17.5\%$

CLASS LIMIT

For a class interval CL is the minimum and maximum value the class interval may contain.

> Minimum : LCL (Lower)

> Maximum : UCL (Upper)

Eg: From discrete Series :-

Class Interval	Frequency	LCL	UCL
10-19	10	10	19
20-29	5	20	29
30-39	8	30	39

Non overlapping
or
mutually inclusive.

In case of continuous Series

Class Interval	Frequency
10-20	
20-30	
30-40	
40-50	

where to take 20?

⇒ where 20 is UCL

[overlapping or mutually exclusive]

Previously classification of grouped frequency distribution.

- > mutually exclusive.
- > mutually inclusive.

Mutually exclusive :

- > Also known as overlapping classification
- > Here, UCL of an interval as LCL of next interval are same.
- > This is usually applicable for continuous variable.
- > An observation which is equivalent to common class limit is excluded from the class interval where it is UCL and taken in the class where it is LCL

Class	LCL	UCL
10 - 20	10	20
20 - 30	20	30

Mutually Inclusive :

- > Also known as non overlapping classification
- > There is no common class limits between two intervals.
- > This is usually applicable to discrete variable
- > All observation including UCL and LCL will be taken in the same class interval

as there is no confusion.

<u>Class</u>	<u>LCL</u>	<u>UCL</u>
10-19	10	19
20-29	20	29
30-39	30	39

CLASS BOUNDARY (actual class limit)

> For mutually exclusive / overlapping.

- Class boundary = class limit.
- $LCL = LCB$
- $UCL = UCB$

> For mutually inclusive / non-overlapping :

- mid of the two class limits

$$- LCB = LCL - \frac{D}{2}$$

$$- LCB = UCL + \frac{D}{2} \quad \text{--- class interval different}$$

mid point / mid value / class mark

$$1) \frac{LCL + UCL}{2}$$

$$2) \frac{LCB + UCB}{2}$$

Answer will be same.

Height, width or size of a class interval or class unit class length.

$$= UCB - LCB$$

For eg : $UCB = 29.5$

$LCB = 19.5$

$$\therefore \text{Class length} = 29.5 - 19.5 = 10$$

* CUMULATIVE FREQUENCY

(Imp)

less than more than

* Frequency density of a class interval

> It may be defined as the Ratio of the frequency of that class interval to the corresponding class length.

$$\text{Frequency density class} = \frac{\text{frequency}}{\text{Class length}}$$

* Relative frequency

> For a class interval, relative frequency

$$= \frac{\text{class frequency}}{\text{total frequency}}$$

$$RFC = \frac{f}{\Sigma f}$$

Ratio, Class frequency; total frequency

* Percentage frequency class interval

> Ratio, Class frequency, total frequency

$$PFC = \frac{f}{\Sigma f} \times 100$$

Graphical Representation of a Frequency distribution. — (3)

- (i) Histogram or Area diagram or Rectangle diagram
- (ii) Frequency Polygon.
- (iii) Ogives or Cumulative frequency graphs.

(I) Histogram or Area Diagram.

For equal
Class Interval

For unequal
Class Interval.
(FDC)

- > A comparison between frequency of two different classes possible.
- > It is useful to calculate mode also. Lies on x axis.

(II) Frequency Polygon.

- > class mark and frequency is required to draw frequency polygon.
- > Points are joined by Scale (Straight line)
- > Frequency Polygon can be drawn from Histogram.

x axis = Class value
y axis = Frequency

(iii) Ogives or Cumulative Frequency Graph

less than
ogive

more than
ogive

> Ogives further can be put into use for making short term projection.

> Median can be find out from ogive either by intersection or by any one of these and lies on x-axis.

> The shape of less than ogive is J-shaped.

Frequency Curve -

> A frequency curve is a smooth curve for which the total area is taken to be unity.

> It is a ~~same~~ limiting form of a histogram or frequency polygon.

> These are of below four types.

(a) Bell Shaped curve
most commonly used (Blue chip company)

(b) U-shaped curve

(c) J-shaped curve

Startup or New edge company

(d) Mixed curve

Combination of curves