UNIT-2

classification is the process of arranging things in groups or classes according to their resemblances and affinities, and gives expression to the unity of attributes that may exist amongst a diversity of individuals

OR

The grouping of data into different categories or classes with similar / homogeneous characteristics is known as CLASSIFICATION OF DATA

characteristics of classification

- performs homogeneous grouping of data.
- brings out points of similarity and dissimilating.
- may be either real or imaginary
- is flexible to accomodate adjustments

characteristics of good classification

- comprehensiveness
- homogeneity
- stability
- Clarity
- Suitability
- Elastic

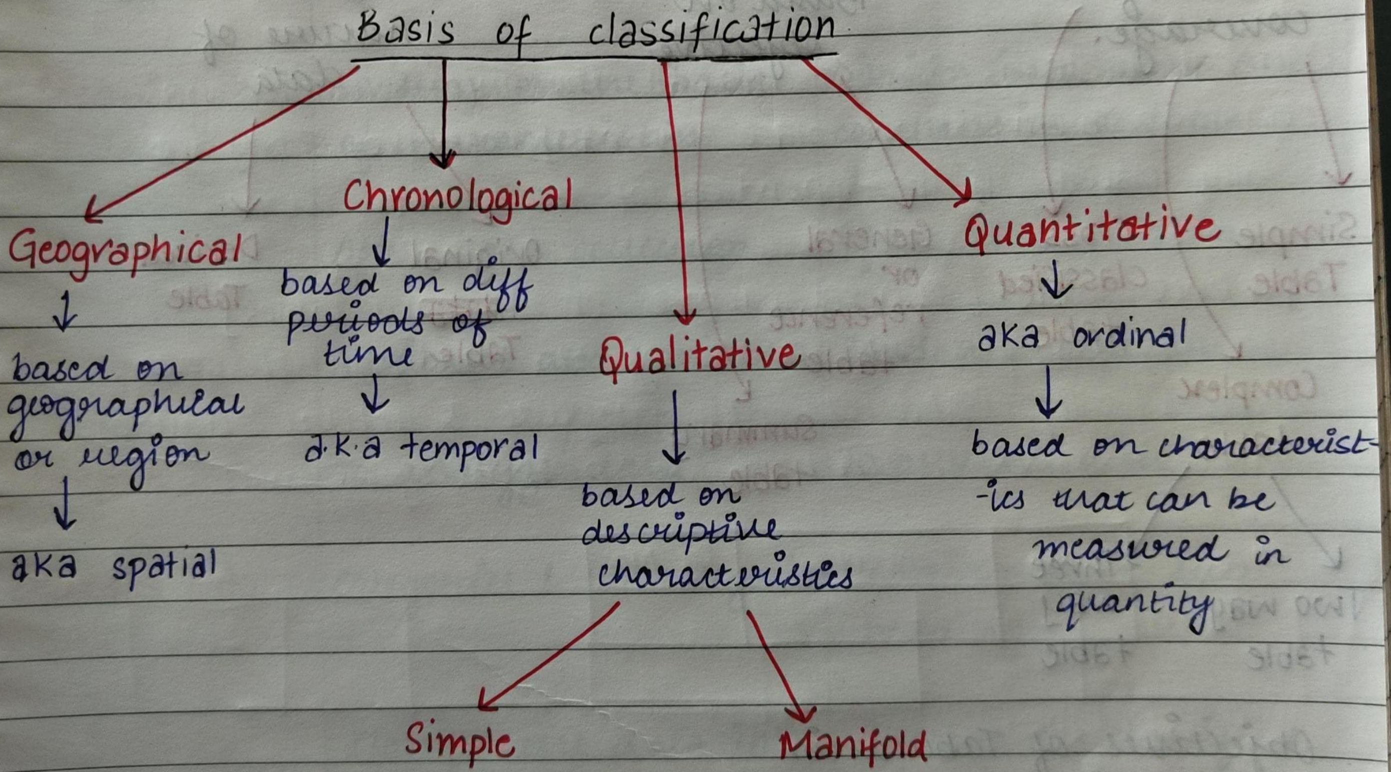
objectives of classification

- Brief and simple
- utility
- Distinctiveness
- Comparibility
- Scientific arrangement
- Attractive and effective.

Objectives:

- explain similarities and differences of data
- simplify and condense data's mass
- facilitate comparisons
- study the relationship
- prepare data for tabular presentation
- ~~presentation~~ present a mental picture of the data

Types.

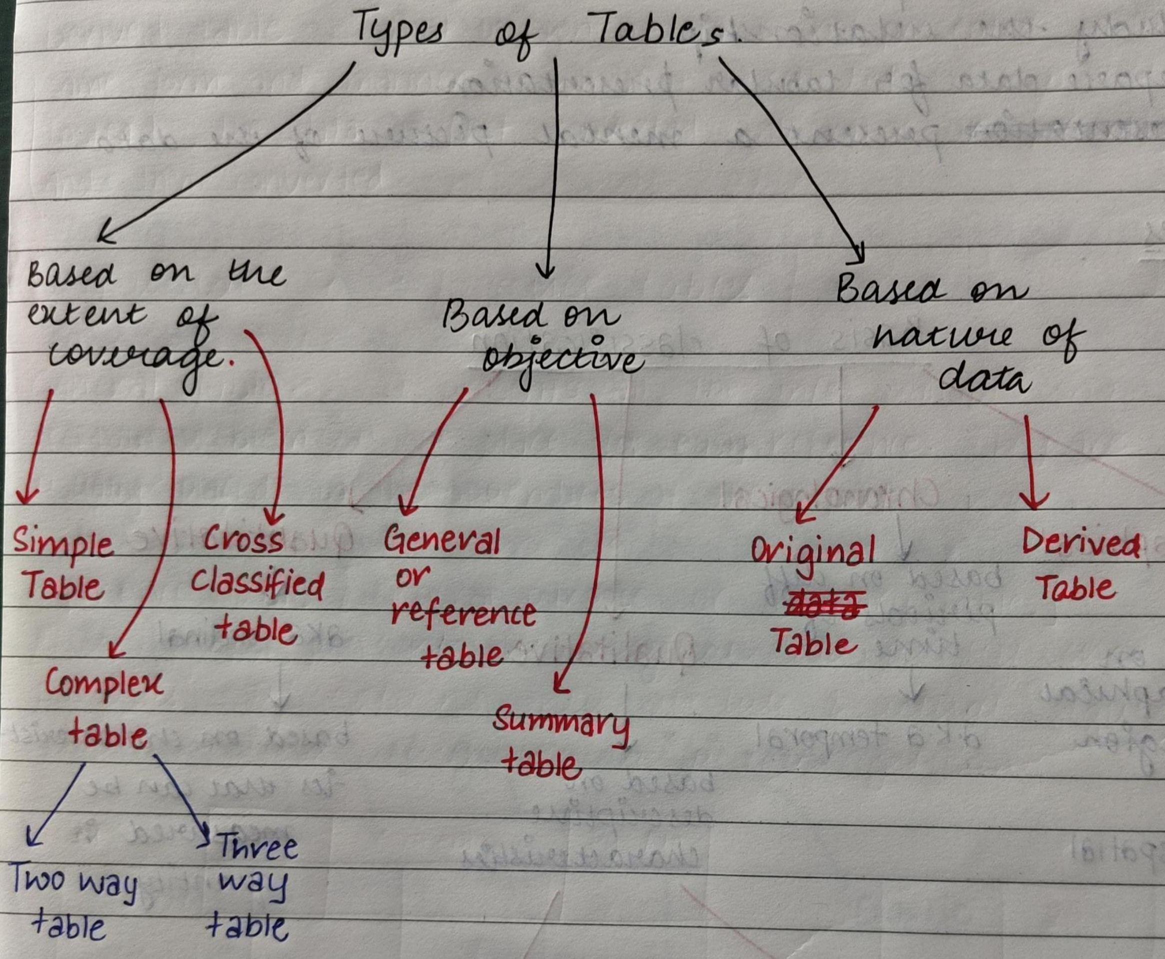


Basis of classification
Geographical
based on geographical or region
aka spatial
Chronological
based on diff periods of time
a.k.a temporal
Qualitative
based on descriptive characteristics
Simple
Manifold
Quantitative
aka ordinal
based on characteristics that can be measured in quantity

Qualitative data: attributes, labels, or non-numerical entries. This is is also called as categorical data

Quantitative data: numerical measurements or counts

TABULATION. It is a systematic arrangement of data in columns and rows, that represent data in concise and attractive way


Types of Tables.

Based on the extent of coverage.

Simple Table

Cross Classified table

Complex table

Two way table

Three way table

Based on objective

General or reference table

Summary table

Based on nature of data

Original ~~data~~ Table

Derived Table

Objectives of Tabulation

- To simplify the complex ~~to~~ data
- To facilitate comparison
- To economize the space
- To draw valid inference/ conclusions
- To help for further ~~analysis~~ analysis

Uses / Objectives / Advantages / Importance of Tabulation of data :

Simplify complex data
Highlight important information
Enable easy comparison
Help in the stastical analysis
Saves space.

Basis for comparison	Classification	Tabulation
Meaning	Division on the basis of characteristics	Division into rows and columns in a table.
Order	After data collection	After classification
Arrangement	Attributes and variables	Rows and columns
Purpose	To analyze / sorting data	To present data
Bifurcates data	Categories & sub categories	Headings and sub headings
	Classify the data into diff groups	Present the classified data in tabular form.

① Original & Derived Table [Based on Nature]

Original table is that in which data is presented in the same form and manner ~~as~~ in which they are collected.

Derived table is that in which data is not presented in the same form and manner in which it is collected.
Instead, the data is first converted into ratios or percentages and then converted.

② General or reference tables [Based on objective].

General table. : It presents all the info available on a certain problem. at one place for easy reference. They are usually placed in the appendices of the report

Reference Table [General purpose or Primary Tables]
These tables present the original data for reference purposes. It contains only absolute & actual figures and round numbers or percentages.

ex:

S. No.	Contents	Pg. No.

Summary table.
The info contained here, aims at analysis and inference. Hence, they are a.k.a. interpretative tables

③ Simple, Complex & Cross classified table. [Based on extent of coverage]

Simple Tables - based on single characteristic.

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of students	10	12	17	20	15	6

Complex Tables- A complex table summarizes the complicated info and presents them into 2 or more interrelated categories.

Rounds.

		Round 1	Round 2	Round 3	Round 4	Round 5
Phases	Phase 1	Mary	Stanley	Mary	Stanley	-
	Phase 2	Mary	Stanley	Mary	Stanley	-
	Phase 3	SYSTEM	SYSTEM	SYSTEM	SYSTEM	-
	Phase 4	Stanley	Mary	Stanley	Mary	-
	Phase 5	SYSTEM	SYSTEM	SYSTEM	SYSTEM	-

Two Way Table. If there are 2 coordinate factors, the table is called two way table or Bi-variate Table

	Basketball	Baseball	Football	Total
Males	15	13	20	48
Females	16	23	13	52
Total	31	36	33	100

Three Way Table If the number of coordinate groups is 3, ~~and it is based on more than 3~~ it is a case of 3 way tabulation.
It is also known as Derive Table

	Engineering		English	
	Male	Female	Male	Female
Admit	30	10	5	10
Deny	30	10	15	30
Total	60	20	20	40

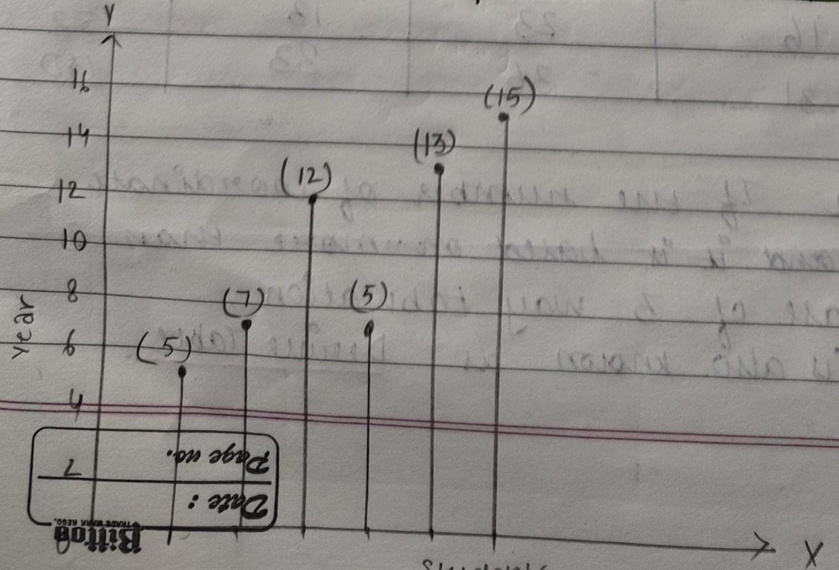
Cross Classified Table describes the relationship b/w 2 or more categorical variables. It is a.k.a contingency table.

Diagrammatic Presentation

A diagram is a visual form for presentation of stastical data. The diagram refers various types of devices such as bars, circles, maps, pictorials, cartograms etc

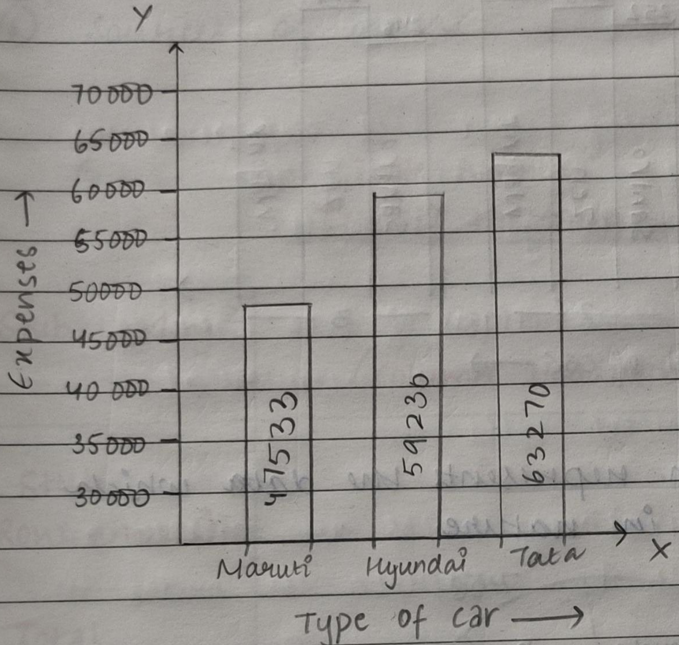
Line diagrams on the basis of the figures, heights of the bars/lines are drawn. Distance between the bars is kept uniform

Year	2001	2002	2003	2004	2005	2006
Students	5	7	12	5	13	15

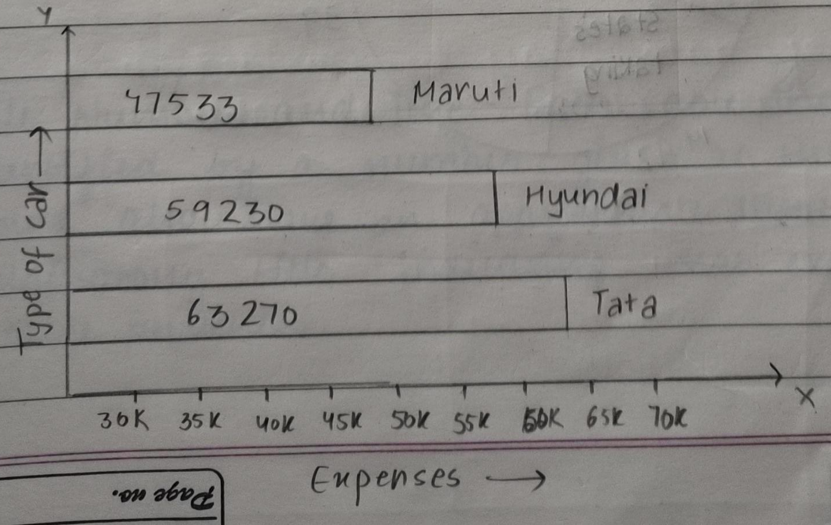


Simple Bar chart The vertical bars represent the numbers, and the horizontal axis represent the variables

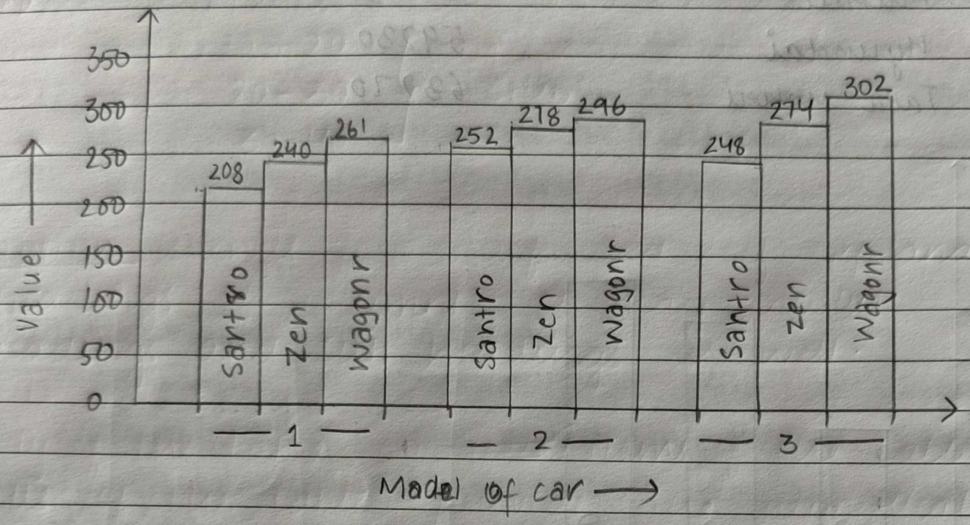
Type of Car	Expenses Rs/Year
Maruti	47533
Hyundai	59230
Tata Motors	63270



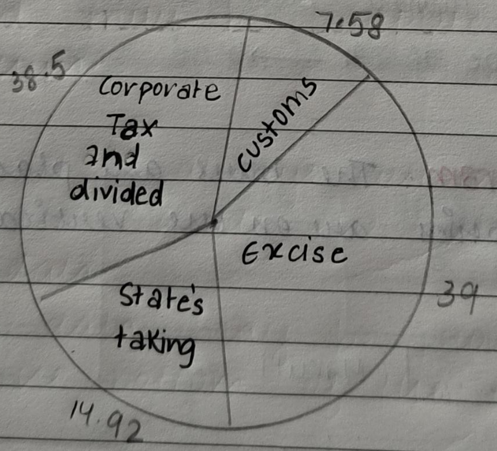
Horizontal bar diagram. The bars are placed horizontally and the categories are on the vertical axis.



Compound Bar Diagram The information represented is complex in nature with multiple categories in multiple years.



Pie Diagram The diagram represents the data which can be comparitive in nature



Components of a Table

1. Table number and Title
2. Stub (Heading of rows)
3. Caption (Heading of columns)
4. Body of the table
5. Foot notes
6. Sources of data

Table Number

Title of the table

Stub Heading	Caption (Column Headings)	Total
Stub (Row entries)	Body	
Total		

Foot note (if any)

Source of data (if any)

Table Number and Title - Each table should be identified by a number given at the top. It should also have an appropriate short and self explainatory title indicating what exactly the table presents.

Stub. stands for brief and self explantory headings of rows.

ex: Marks } — stub head
10 - 20 }
20 - 30 } stub.
30-40 }

Caption stands for brief and self explanatory headings of columns. It may involve headings and sub headings as well.

Body of the table or ~~item~~ items of the table or cells of the table provide the numerical information in different cells.

Foot Notes. The explanatory notes are given as foot notes and must be complete in order to understand them

Sources of data It is always customary to provide source of data to enable the user to refer the original data.
The source of data may be provided in a foot note at the bottom of the table.

Characteristics of a Good Table. / General Precautions

- Table should suit the size of the paper.
- No. of rows and columns should neither be too large nor too small.
- Possible figures should be estimated before tabulation to reduce unecessary details.
- Table must be as precise as possible and easy to understand.
- Items should be arranged alphabetical / chronological / geographical order or acc to their size.
- Sub-total and total of the items should be mentioned
- Mass of data should be presented in more than one table.
- Percentages, totals and averages should be kept close to each other.

Frequency Distribution

A tabular arrangement of raw data by a certain number of classes and the number of items (called frequency) belonging to each class is known as frequency distribution.

Types of frequency distribution

- Discrete freq distribution
  ↓
  values that cannot be further opened up
- Continous frequency distribution

Formation of freq distribution

- Inclusive Method
- Exclusive Method

Discrete frequency distribution   Here tally chart is constructed

Age	Tally Marks	Frequency
14	||||	4
15	𝍸 |||	8
16	𝍸 𝍸	10
17	|||	3
Total		25

continous frequency distribution A large mass of data that is summarized in such a way that the data values are distributed into groups, or classes, or categories along with the frequencies is known as continous or grouped frequency distribution.

Exclusive freq distribution

class interval	frequency
150 - 153	7
154 - 157	7
158 - 161	15
162 - 165	10

20 or more but less than 30 →

20 – 30	12
30 – 40	15
40 – 50	13
50 – 60	25

← 30 will be included here.

[20, 30)
[30, 40)

Terminology

Class If the observations of data set are divided into groups and the groups are bound by limits, ~~tho~~ then each group is called a class

Class limit The end values of a class is called the class limit. The smaller value of the class limit is called the lower limit (L) and the bigger value is called the upper limit (U).

Class interval The difference between the lower limit and the upper limit is called the class interval (I)

$\boxed{I = U - L.}$

Class boundaries are the midpoints b/w the upper limit of the class and the lower limit of the succeeding class.

Class width of a class is the difference b/w the upper class boundary and the lower class boundary

Mid point Half of the diff b/w upper class boundary and the lower class boundary.

Inclusive Method of freq distribution

- Both the lower & the upper class limits are included in the classes.
- may be used for grouped frequency distribution for discrete variables.
- cannot be used in case of continous variables like height, weight etc as integral/ fractional values are not permissable.

Exclusive Method

- The values which are equal to upper limit of class are not included in that class, instead they are included into next class.
- fractional/ decimal values are ~~pore~~ used here

Converting inclusive to exclusive.

Inclusive →	Exclusive
5 - 15	4.5 - 15.5
16 - 24	15.5 - 24.5
25 - 35	24.5 - 35.5
36 - 44	35.5 - 44.5

$(x-0.5) - (y+0.5)$.

Cumulative Frequency Distribution. to a class interval is defined as the total frequency of all values less than upper boundaries of the class. A tabular arrangement of all cumulative frequencies together with the corresponding classes is called cumulative freq distribution

Type	Freq	Cum Freq.
upto 1000	22	22
1001-2000	45	67
2001-3000	57	124
3001-4000	97	221
4001-5000	152	373
5001-6000	241	614
6001-7000	153	767

2 types of cumulative Frequency

more than type

less than type.

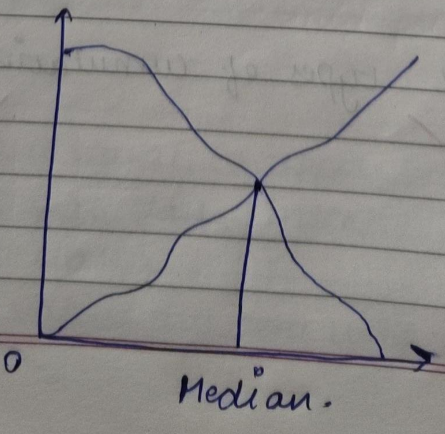
Frequency Distribution	Cumulative frequency distribution
A particular class interval according to how many items lie within it is described	The number of items that have values either above or below a particular level is described.

Less than type. The cumulative freq of each class shows the number of elements in the data whose magnitudes are less than the upper limits of the respective class.

More than type. The cumulative freq. for each chass shows the number of elements in the data whose magnitudes are more than the lower limits of the respective class.

class	freq.	less than		more than	
0 - 10	18	10	18	0	101
10 - 20	32	20	50	10	83
20 - 30	15	30	65	20	51
30 - 40	17	40	82	30	36
40 - 50	19	50	101	40	19
	101				

ogives are used.



Relative Frequency Distribution

- defined as the ratio of cumulative freq to the total freq.
- usually expressed in terms of a %.
- arrangement of relative cumulative frequencies against the respective class boundaries is termed as relative cumulative freq distribution

class	freq.	cumm freq	relative freq.
50-59	5	5	5/45 = 0.11
60-69	8	13	8/45 = 0.18
70-79	12	25	12/45 = 0.27
80-89	13	38	13/45 = 0.29
90-99	7	45	7/45 = 0.16

Bivariate Freq distribution

When a data set consists of a large mass of observations, they may be summarized using a two-way table.

Diagrams
↓
Qualitative data

Graphs
↓
Quantitative data.

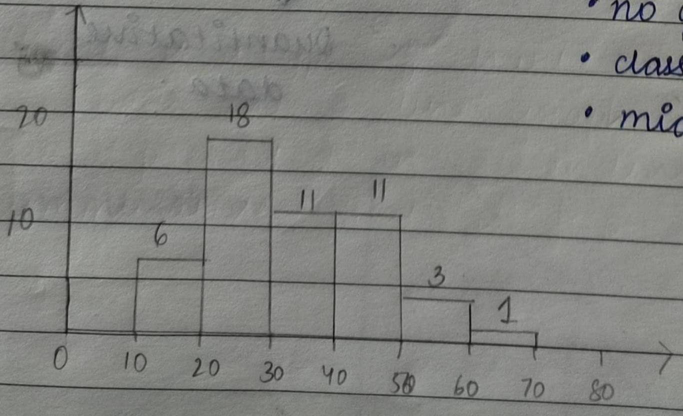
~~Con~~ Common statistical Graphs.

- histogram
  vertical bar chart of frequencies
- Frequency polygon
  line graph of frequencies
- Ogive
  line graph of cumm frequencies
- Bar chart
- Pie chart
  proportional presentation for categories of a whole
- Pareto chart
- Scatter plot

Graphs for Quantitative Data

- histograms
- Relative freq and percent freq distributions
- Dot plot
- Frequency distribution
- Ogives.

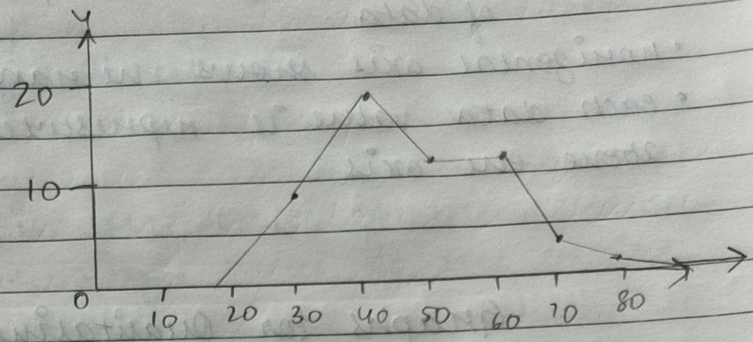
Histogram.



features:
- no gaps b/w bars
- class boundaries
- mid points.

Frequency polygon :

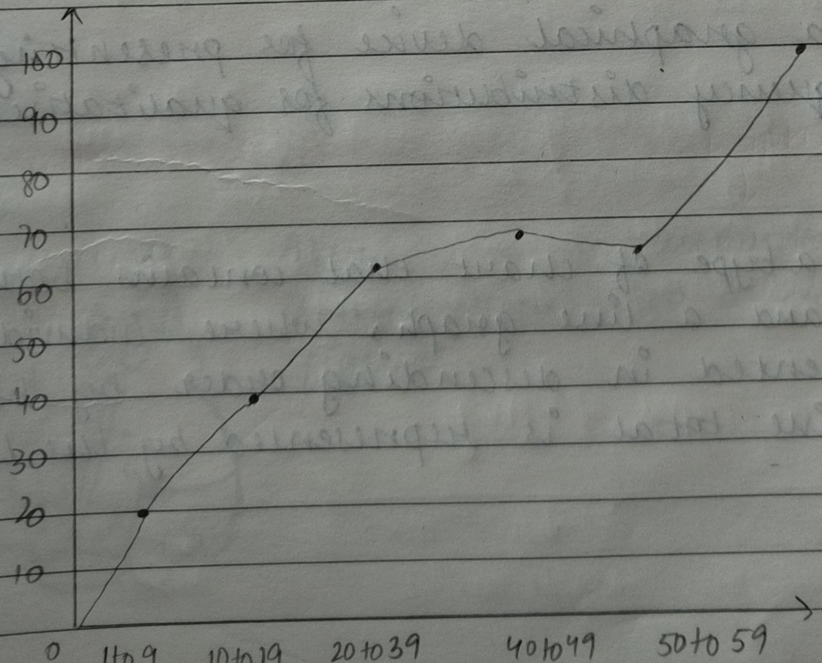
20 under 30	6
30 under 40	18
40 under 50	11
50 under 60	11
60 under 70	3
70 under 80	1



Ogives :

- An ogive is a graph of cummulative distribution
- Data values are shown on horizontal axis.
- on vertical axis
  ↳ cummulative frequencies OR
  ↳ cummulative relative frequencies OR
  ↳ cummulative percent frequencies.

- The frequency of each class is plotted as a point.
- The plotted points are connected by straight lines



Dot Plot. It is one of the simplest graphical summaries of data.

- horizontal axis shows the range of data values
- each data value is represented by a dot placed above the axis.

Graphs for Qualitative data

Bar Graph

Pie chart

Pareto chart

Bar chart.

- on the horizontal axis, we specify the labels that are used for each of the classes.
- A frequency, relative frequency, or percent frequency scale can be used for the vertical axis.
- The bars are seperated to emphasize the fact that each class is a seperate category.

Pie Chart. is a graphical device for presenting relative frequency distributions for qualitative data.

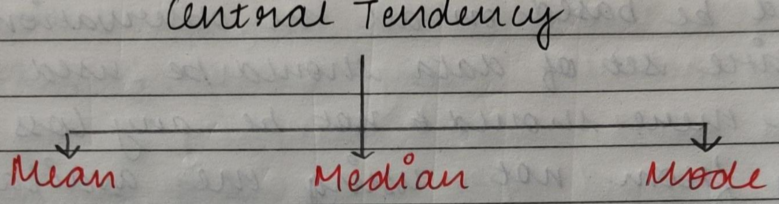
Pareto Chart. is a type of chart that contains both bars and a line graph, where individual values are represented in descending order by bars and cummulative total is represented by the line.

Standar deviation = $\sqrt{\text{variance}}$.

$$= \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}}$$

CENTRAL TENDENCY is the stastical measure that represents the single value of the entire distribution or a dataset.

It aims to provide an accurate description of the entire data in the distribution.



Purpose of central Tendency

The purpose is to provide an exact representation of the entire collected data. It is often defined as the single value that is representative of the data.

Purpose and Functions of Average

Average is an attempt to find one single figure to describe whole of figures.

Characteristics of a Good Average:

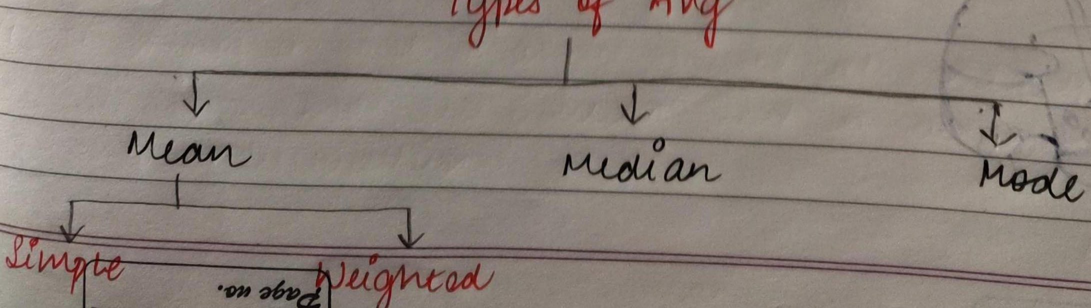
① It should be rigidly defined. It means that the def^n should be clear so that it leads to one and only one interpretation.

② It should be easy to understand and simple to calculate. It should be so easy that even a non-mathematical person can calculate it.

③ It should be based on all the observations. It means that entire set of data should be used in computing avg. and there should not be any loss of info resulting from not using the available data.

④ It should be capable of further algebraic treatment. Avg should be capable of further mathematical and statistical computations to expand or enhance its utility.

⑤ It should not be unduly affected by extreme observations. Avg should be such that it should not be affected by the presence of one/two very small or very large observations



Arithmetic Mean The most popularly used measure of central tendency is arithmetic mean or simply mean.

Ⓐ Simple Arithmetic Mean.

I In case of ungrouped data

1. Individual observations

$$x_1, x_2, x_3 \ldots x_n$$

$$\bar{X} = \frac{\Sigma x}{n}$$

2. Discrete frequency distribution.

~~a~~ Direct Method

$$\bar{x} = \frac{\Sigma f x}{\Sigma f}$$ f = frequency

II In case of ~~ung~~ grouped freq distr.

OR

continous series

$$\bar{x} = \frac{\Sigma f m}{\Sigma f}$$ m → mid value of each class interval

Step-derivation Method,

$$\bar{x} = A + \frac{\Sigma f\mu}{\Sigma f} \times i$$

$$\mu = \frac{m-A}{i}$$

A = assumed mean
m = mid value
i = step factor

Properties of Arithmetic Mean

Merits

1. Simple to calculate and easy to understand
2. Based on each and every observ^n of the series
3. Does not fluctuate with sampling.
4. Does not depend upon the position in the series.
5. Is capable of further algebraic treatment
6. It is rigidly defined. Everyone will get the same answer when apply the formula of avg.

De Merits of Arithmetic Mean

1. It is unduly affected by extreme values, ie, the presence of very large/small numbers
2. It cannot be determined by inspection like mode and it cannot be located graphically.
3. Mean is not a suitable avg in class of qualitative data.
4. It is not a good measure of central tendency in case of normal distribution and in case of U shaped distribution.

Median. The median of a set of data is the middlemost number of centre value in the set.

Median is also the number that is halfway into the set.

(A) Odd Number of Observations

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{term}$$

(B) Even number of observations

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} \text{term} + \left(\frac{n}{2}+1\right)^{th} \text{term}}{2}$$

Median is that value of the data which divides the group into two equal parts.

Median is the middle value of the series, when items are arranged in either ascending or descending order.

Merits of Median:

- It is easy to calculate and understand
- It is not affected by extreme values because it is a positional average and not dependent on the magnitude.

- It has a definite and a certain value because it is rigidly defined.

- It is the best measure of central tendency while dealing with qualitative data where ranking is preferred instead of measurement or counting.

- It can be calculated even if the values of the extremes are not known. However, the number of items should be known.

- Its values can be determined or represented graphically with the help of ogive curves.

Demerits of Median:

- Arranging the data in ascending/descending order of magnitude is time consuming in case of a large number of observations.

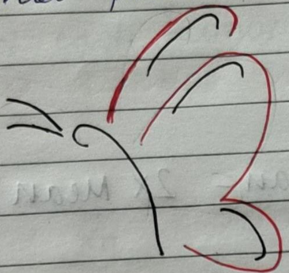
- It is a positional avg and does not consider the magnitude of the items. It neglects the extreme values.

- It is not dependent on all the observ"s, so it cannot be considered as their good representative

- In case there is a big variation b/w the data, it will not be able to represent the data

- It is affected by the fluctuations in sampling and this effect is more than that in the case of arithmetic mean

- It is a positional average, so further algacbric treatment is not possible.



Purpose of central Tendency

- To condense the data in a single value that is representative of the whole
- To facilitate comparison.
- It makes it easy for the researcher and the reader to comprehend the data.
- With the help of a sample, it provides us the idea about the mean of the whole population.

Characteristics of Central Tendency:

- It should be well defined so that a unique answer can be obtained
- It should be used to understand, calculate and interpret.
- It should be based on all the observations of the data.
- It should be based on

Characteristics of a Good Measure:

- It should be amenable for further mathematical calculations.
- It should be least affected by the fluctuations of the sampling.
- It should not be unduly affected by the extreme values

MODE The observation which occurs the most frequently in the series is Mode

for grouped data

$$\text{Mode} = 3\times \text{Median} - 2\times \text{Mean}.$$

Or

$$\text{Mode} = L + \frac{(fm - f_1) \times h}{(fm - f_1) + (fm - f_2)}$$

l= lower boundary of modal class-
h= size of modal class.
fm = frequency correspondent to modal class-
f_1= frequency preceding the modal class.
f_2 = frequency ~~pro~~ exceeding the modal class -

Merits of Mode:

- It can be easily located by mere inspection
- It eliminates extreme variations
- It is commonly understood
- Mode can be determined graphically.

De Merits of Mode:

- It is a measure having limited practical value
- It is not capable for further mathematical treatment.
- It is ill-defined and indefinite and so trustworthy

NOTE

The list having2 modes is Bimodal List and the one having 3 modes is Trimodal List

HARMONIC MEAN is a type of numerical average calculated by dividing the number of observations by the reciprocal of each number in the series -

$$HM = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$$

for grouped continous data

$$HM = \frac{n}{\Sigma\left(\frac{f}{x}\right)}$$

f = frequency of the class
x = mid-point of the class
n = total frequency

Special Applications:

- It is useful for computing the average rate of increase of profits of a concern or avg speed at which a journey has been performed.

- The avg price at which an article has been sold. The rate usually indicates the relation b/w 2 different types of measuring units that can be expressed reciprocally.

Demerits When the number of items is large, the computation of Harmonic Mean becomes tedious.

Standard Error of a stastic is the approximate standard deviation of a statistical sample population

~~Standard error is the approximate standar~~

The STANDARD ERROR is a ~~statical~~ statistical term that measures the accuracy with which a sample distribution represents a population by using standard deviation.

In ~~stasti~~ statistics, a sample mean derivates from the actual mean of a population. This is known as standard error.

$$SE = \frac{\sigma}{\sqrt{n}}$$

σ → population standard deviation.
n → sample size.

DATA	INFORMATION
Data is unorganized and unrefined facts	Info comprises processed, organized data presented in a meaningful context.
Data is an individual unit that contains raw materials which do not carry any specific meaning	Info is a group of data that collectively carries a logical meaning.
Data does not depend upon Inf	Info depends upon data

Raw data alone is insufficient for decision making	Info is sufficient for decision making.
Ex: student's test score	Ex: average score of class derived from given data

<u>SAMPLE</u> is a group of individuals who will actually participate in the research

SAMPLING METHOD

Probability Sampling	Non probability Sampling
involves random selection, allowing you to ~~base~~ make strong stastical inferences about the whole group	involves non-random selection based on convenience ~~of~~ or other criteria allowing you to easily collect data

Judgmental Sampling (ie. nonstastical) includes gathering a selection of items for testing based on examiner's professional judgement, expertise and ~~knowd~~ knowledge to target known or probable areas of risk.

Stratified Sampling Here, researchers divide subjects into sub groups called strata based on characteristics that they share. Once divided, each subgroup is randomly sampled using another probability sampling method.

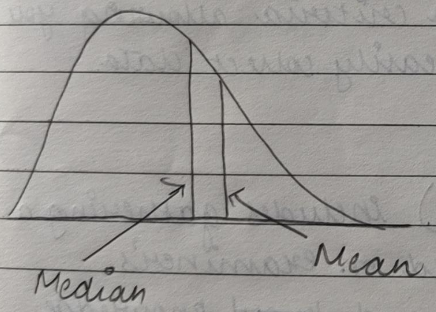
Comparison of mean, median and mode.

- In symmetric distributions, median and mean are equal.

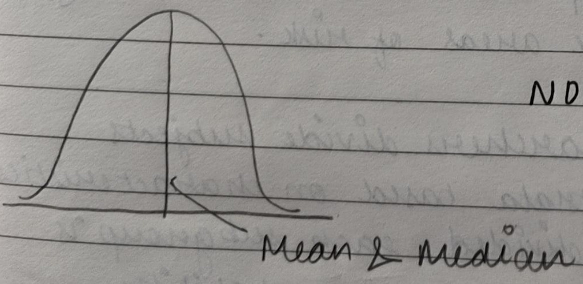
for normal distribution,
mean = median = mode

- In positively skewed distributions,
mean > ~~mo~~ median

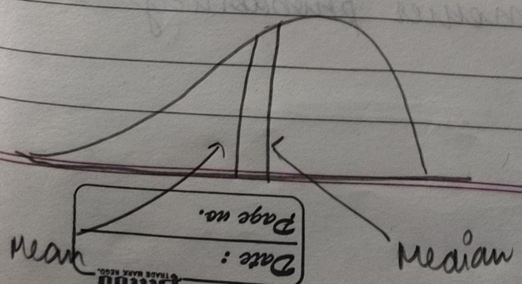
- In negatively skewed distribution.
mean < median



POSITIVELY SKEWED



NORMAL DISTRIBUTION



NEGATIVELY SKEWED

PROS	CONS
Sensitive as it takes all data values into account (reliable)	Biased output if outliers / extreme values exist in the data set
Not affected by extreme values	• Less sensitive than Mean as it only focusses on giving out the middle data point irrespective of how far the other values are from the middle. • Needs the data to be arranged in ascending order before computing
Not affected by extreme values and can be used with non-numerical data	There may be more than one mode or no mode at all and it may not reflect data summary accuratel