QUANTITATIVE ANALYSIS

MODULE - 1 : Introduction to Statistics

* **Statistics :-**

→ Science which deals with :

a) Collection of data
b) Tabulation of data
c) Presentation of data
d) Analysis of data
e) Interpretation of data

* **Types of data :-**

Qualitative	Quantitative
→ Data type : Descriptive statements	→ Data Type can be measured and expressed numerically
→ Text Based	→ Number Based
→ Statistical Analysis is harder	→ Statistical Analysis is simpler
→ Collected using interviews, written documents, observations	→ Collected using surveys, observations, experiments and interviews

* **Quantitative data :-**

→ **Discrete** Variable : finite Countable values. eg:- 1, 2, 3, ...

→ **Continuous** variable : infinite non-countable Values.
eg:- 1·01, 1·001, 1·0001, 1·00001, ...
2·01, 2·001, 2·0001, ...

* **Sources of data :-**

1) **Primary** source : Direct Investigation
2) **Secondary** Source : Recorded Data (data collected by some other person)

* **Tabulation of Data :-**

→ Table should be precise, easy to understand and self explanatory.

* Objectives of Tabulation of Data :-
a) Simplifies Complex data.
b) Compares related facts.
c) Facts are presented in minimum possible space
d) Information needed can be easily located.
e) Good for references and also to present the information.

* Components of Table :-
1) Table Number
2) Title of Table
3) Captions or Column Headings
4) Stubs or Row Headings
5) Body of the Table.
6) Footnotes (written at bottom of the Table)
7) Sources of data.

* Types of table :-
1) One way Table
2) Two way Table
3) Three way Table (Complex Table)

* One way Table :-
→ No. of students in CMPN for the year 2021-22.

S.E	T.E	B.E	Total
216	216	216	648

* Two way Table :-
→ No. of students in CMPN for the year 2021-22 according to sex.

Class / Sex	S.E	T.E	B.E	Total
Male	126	116	120	362
Female	90	100	96	286
Total	216	216	216	648

* Three way Table :

Sex / category	Class	S.E	T.E	B.E	Total
Male	Open				
	Reserved				
Female	Open				
	Reserved				
Total	Open				
	Reserved				
Total					

* Functions of Statistics :-

a) Simplifies the facts in a definite form.

b) Condensation : Reduce or lessen

c) Facilitates Comparison

d) Formation of Policies

e) Forecasting (Predict or estimate)

f) Estimation.

* Scope of Statistics :-

a) Statistics and Industry
b) Statistics and Commerce
c) Statistics and Agriculture
d) Statistics and Economics
e) Statistics and Education
f) Statistics and Planning
g) Statistics and Medicine

* Limitations of Statistics :-

→ Statistics deals with aggregate of facts and not individuals
→ Statistics doesnot study Qualitative Phenomenon.
→ Statistics Laws are only on an average.
→ Statistical Table may be misused
→ Statistics is only one of the methods of studying problem.

* Collection of Data :-

a) To consider the status involved in carrying out a survey.
b) To analyze the process involved in observation & interpretation
c) To describe the methods of collecting primary information.
d) To define and describe Sampling.
e) To Analyze the basis of Sampling.
f) To describe a variety of Sampling methods.

* Nature of Data :-

→ Three Types of Data :-

1) Time Series Data :-

→ Collection of set of Numerical values over a period of time.

→ eg : Consider 3 types of expenditure in Rs. for a family for 4 years.

Eg:

Year	Food	Education	Others	Total
2018	2000	3000	3000	8000
2019	2500	3500	3500	9500
2020	3000	2500	4000	9500
2021	3000	4000	5000	1200

2) Spatial Data :-

→ If the data collect is connected with that of a place, then it is termed as spatial data.

→ eg : Population of Souther States in India in 2011.

State	Population
Andhra Pradesh	8,46,65,533
Karnataka	6,11,30,704
Kerala	33387677
Pondichery	12444

* Spacio Temporal Data :-

Data collected is connected to time as well as place.

* Diagrammatic and Graphical Representation of Data :-

Significance of diagrams and graphs :

→ They are attractive and impressive

→ They make data simple and intelligible

→ Comparison is possible

→ They save time and labour.

→ They have universal utility

→ Gives us more information.

* Rules to construct the diagrams :-

a) Diagram should be neatly drawn and attractive.

b) Measurement figures of geometrical figures should be accurate and proportional.

c) Size of diagram should match size of paper.

d) Every diagram should have a suitable and short heading.

e) Scale should be mentioned in the diagram.

f) Diagrams should have a foot note at the bottom.

g) Index must be given for identification so that the reader can easily make out the meaning of the diagram.

* Types of Diagrams :-

1) One Dimensional
2) Two Dimensional
3) Three Dimensional
4) Pictograms and Cartograms

* One Dimensional Diagrams :-

→ One Dimensional measurement

→ Height is used and width is not considered.

→ Diagrams can be in form of Bar Charts or Line Charts

→ Line Diagram, Simple diagram, Multiple Bar Diagram, Sub-divided Bar Diagram, Percentage Bar Diagram.

* Two Dimensional Diagrams :-

→ Length and Breadth both needs to be taken into account.

→ These are known as Area Diagrams or Surface Diagrams.

→ Rectangles, square, Pie-diagrams.

* Three Dimensional Diagrams :-

→ Volume Diagrams

→ Length, width, height should be considered.

→ Cubes, Cylinders, Spheres, etc.

* Pictograms and Cartograms :-

→ Pictograms : Data is represented through a pictorial symbol.

→ Cartograms : It is known as a statistical map, gives quantitative information. They will represent spatial distributions.

* Graphical Representation :-

→ It is used for representation of the data of a frequency distribution and time series.

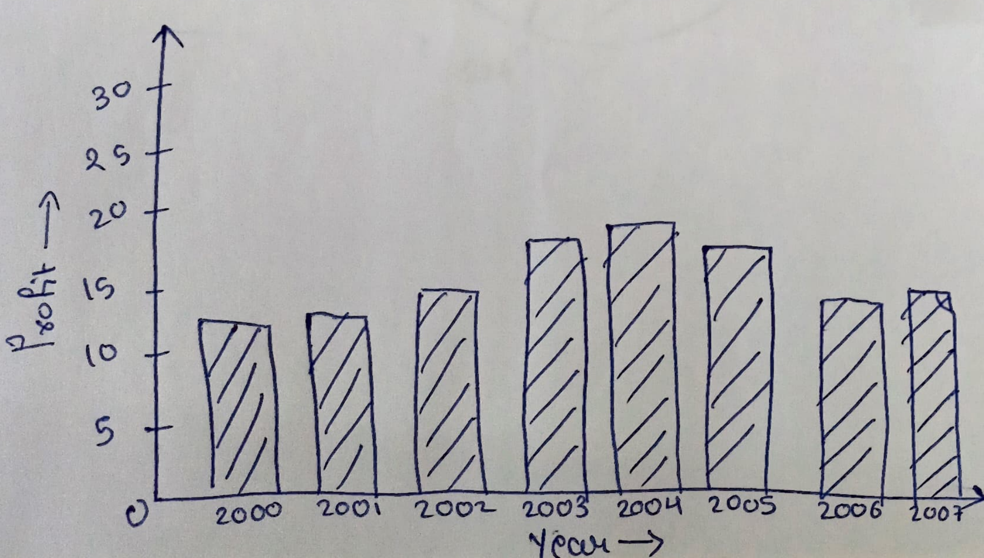
→ Graph : a) Visual form of presentation of statistical data.
b) Represents mathematical relationship between two variables.

→ Histogram, Frequency Polygon, Frequency curve, Lorenz curve

* Problems :-

1) Draw a simple bar diagram for the following data relating to profit achieved by a business firm during 2000-2007

Year	Profit (in Rs. Lakhs)
2000	10.6
2001	12.3
2002	15.6
2003	19.2
2004	20.1
2005	19.1
2006	17.7
2007	18.9



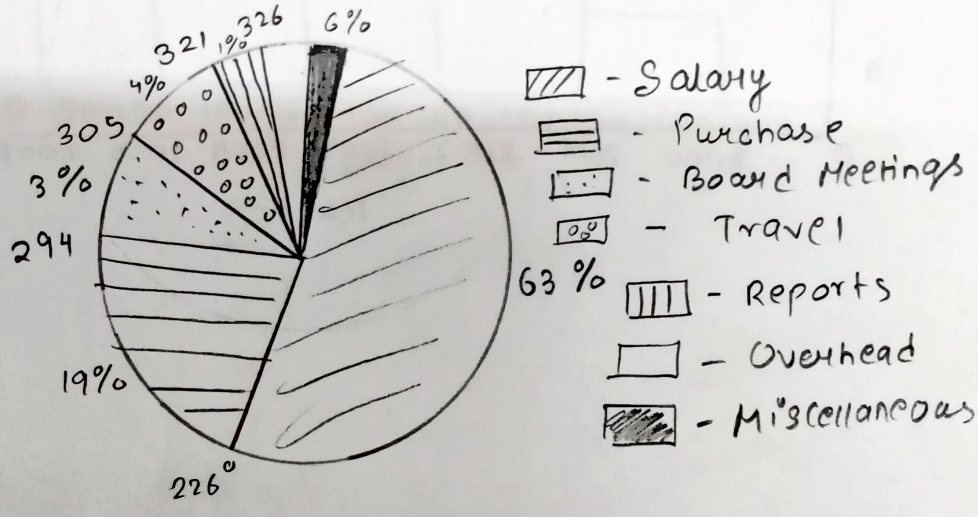
) The following table represents the details of sales & profits achieved by a business firm during 2000 - 2007. Draw a simple bar diagram to represent both series of data.

Year	Profit (in Rs. Lakhs)	Sales (in Rs. Lakhs)
2000	10.5	126.3
2001	12.3	120.9
2002	15.6	140.3
2003	19.2	162.8
2004	20.1	161.7
2005	19.1	168.3
2006	17.7	166.1
2007	16.9	168.2

3) Annual Budget allocation for a Bussiness firm under various heads of expenditure for the financial year 2020-21 is given. Draw a Pie chart.

Heads of expenditure	Budget Allocation (in Rs. Lakhs)	Degree	Percentage
Salary	100	$\frac{100 \times 360}{159} = 226$	63
Purchase	30	67	19
Board Meetings	5	11	3
Travel	7	16	4
Reports	2	5	1
Overhead	5	11	3
Miscellaneous	10	23	6
Total	159		

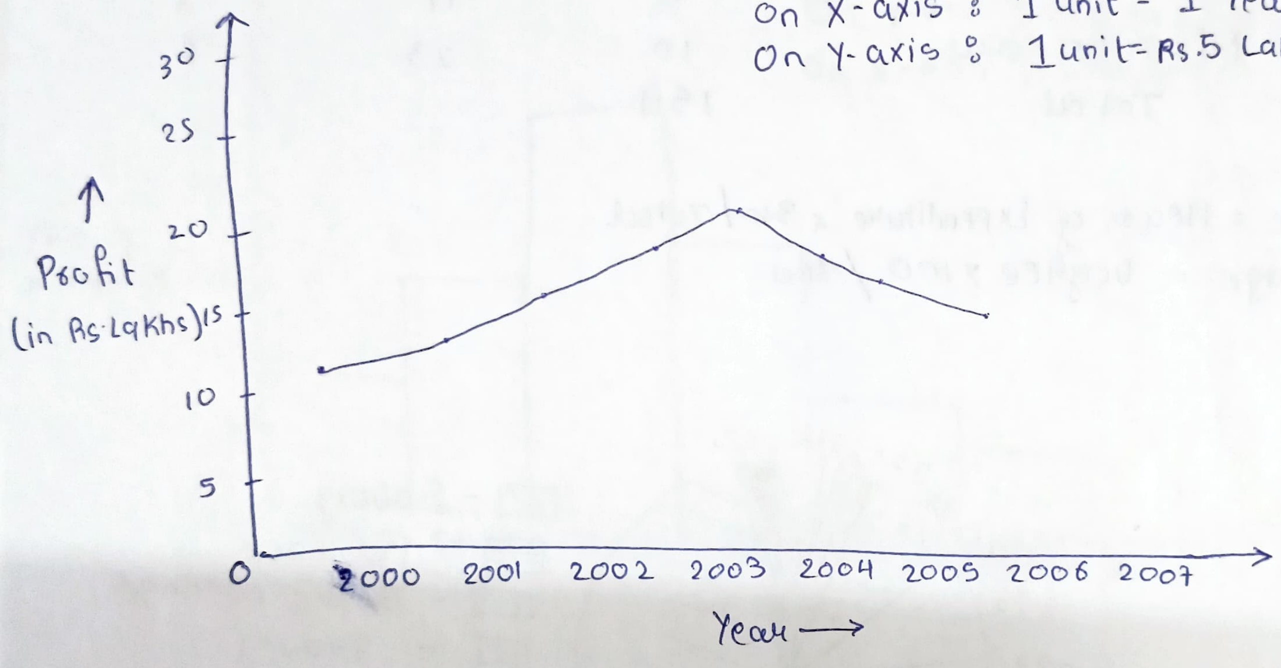
Degree = Heads of Expenditure x 360 / Total.
Percentage = Degree x 100 / 360



4) Find the time series Chart/ for the given data draw a table line Chart.

Year	Profit (in Rs. Lakhs)
2000	10.5
2001	12.3
2002	15.6
2003	19.2
2004	20.1
2005	19.1
2006	17.7
2007	16.9

Scale :-
On X-axis : 1 unit = 1 Year
On Y-axis : 1 unit = Rs. 5 Lakhs



5) Draw a histogram for the following data :

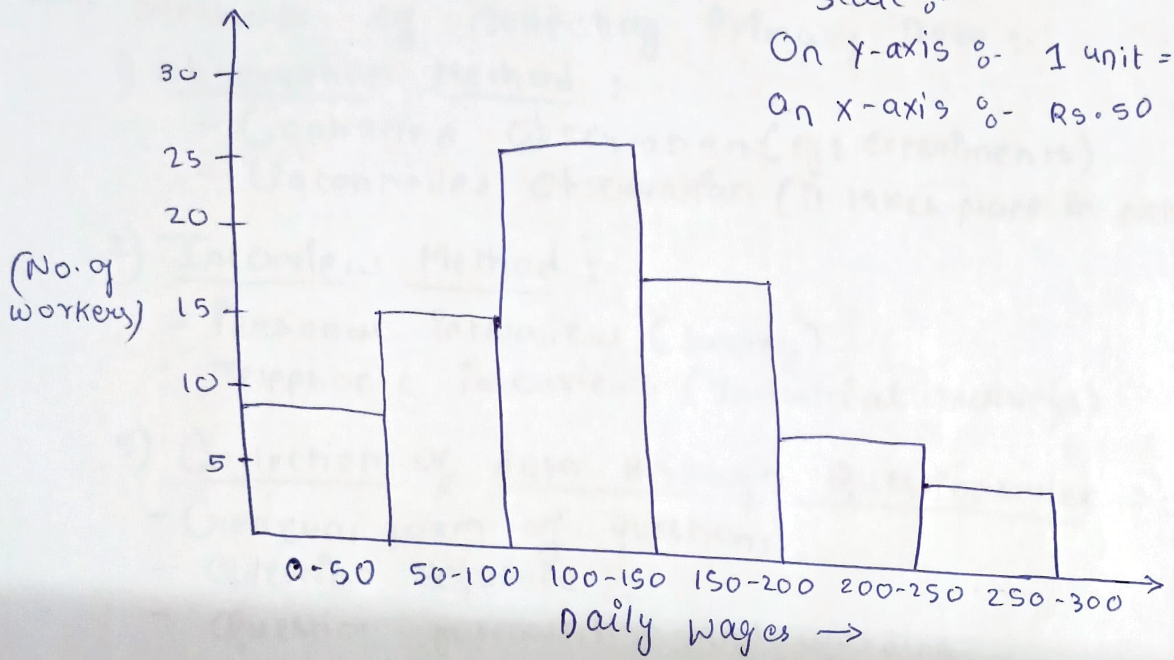
Daily wages (in Rs.)	Number of workers
0-50	8
50-100	16
100-150	27
150-200	19
200-250	10
250-300	6

Sol^n :-

Scale :-

On Y-axis :- 1 unit = 5 workers

On X-axis :- Rs. 50



MODULE -2 :- Data Collection & Sampling Methods.

* **Categories of Data :-**
  1) Primary Data
  2) Secondary Data

* **Primary Data :-**

→ Collected by the investigator himself for the purpose of a specific enquiry or study.

→ It is termed as original data and it is generated by conducting a survey.

→ Methods of Collecting Primary Data :-

1) observation Method :-
   - Controlled observation (eg: experiments)
   - Uncontrolled observation (it takes place in natural settings)

2) Interview Method :-
   - Personal interview (survey)
   - Telephonic interviews (Industrial surveys)

3) Collection of data through Questionaries :-
   - General form of questions
   - Question sequence.
   - Question formulation and wording

4) Collection of data through Schedules :-
   - Schedule : Proforma containing a set of questions required to be filled for a specific purpose.
   - Difference between questionaire and Schedule :

     Questionaire : Sent through mail and any person related to the questionaire can fill the data

     Schedule : Generally filled out by a research Scholar

5) Other Methods of Primary Data Collection :-

- Warranty Cards
- Audits
- Use of Mechnical Devices
  (eg: Eye camera, Motion Picture camera etc.)

* **Secondary Data** :-

→ Data has been already collected and analyzed by some agency for its own use and later the data is used by a different agency.

→ Sources of Secondary data :

a) Published Sourcess

b) UnPublished Sources.

- **Published Sources** :-

→ Government Publications
(Indian Trade Journals, Reports on Currency and Finance, Reserve Bank of India Bulletin, Agricultural statistics of India, Indian Foreign statistics, Economic surveys)

→ International Bodies.
(UNO, WHO, International labour Organization, Food & Agriculture Organization, World Meterological Organization)

→ Semi-Government Publications
(Publish reports related to birth, death, education, sanitation etc.)

→ Repoots of Committee and Commissions.

→ Private Publications
(Institute of Economic Growth, Stock Exchanses, NCERT, NCAER)

→ Newspapers and Magazines

→ Research Scholars

- **Unpublished Sources :-**
  Data here is not published and records are maintained by government ; agencies ; private offices and firms

*** Frequency Distribution :-**

→ A table in which the data is grouped into classes and the number of cases which fall in each class are recorded.

→ Example :

10	50	55	35	39	75	81
09	12	13	32	21	25	90
08	47	62	26	40	42	19
20	51	15	17	45	85	80

→ The above data is raw or ungrouped data.
→ Such data is recorded without any pre-Consideration.
→ Thus, we consider Discrete Frequency Distribution.
→ The distribution shows no. of times each value has occured.
→ Discrete Frequency Distribution is termed as ungrouped frequency distribution.

Q. The problem is as follows

3	1	3	2	1	5	6	6
2	0	0	3	4	2	1	2
1	3	1	5	3	3	2	4
2	2	3	0	2	1	4	5
4	2	3	4	1	2	5	4

1) In a survey of 40 families in a village, the number of children per family was recorded and the following data was obtained : Represent the data in the form of discrete frequency distribution.

Solⁿ:- Frequency distribution of the no. of children is as follows:

No. of Children	Tally Marks	Frequency
0	III	3
1	~~IIII~~ II	7
2	~~IIII~~ ~~IIII~~	10
3	~~IIII~~ III	8
4	~~IIII~~ I	6
5	IIII	4
6	II	2
	Total =	40

* **Grouped Frequency Distribution**:-

→ Continuous Frequency Distribution

→ Distribution refers to group of values

→ Example :

Weekly Wages (in Rs.)	No. of Employees
1500 - 2000	4
2000 - 2500	12
2500 - 3000	22
3000 - 3500	33

→ The terms used in the distribution :

a) Class Interval :-

- Various groups into which the values of the variable are classified.
- Eg : 25-35 represents a group or class which includes values from 25-35.

b) Class Limits :-

- Minimum and Maximum values (lower limit & upper limit)

c) Class Mark / Midpoint :-

- $\text{Midpoint} = \dfrac{\text{Lower Limit} + \text{Upper Limit}}{2}$

- Eg :-

Class :- 25 - 50

Lower Limit = 25

Upper Limit = 50

$\text{Midpoint} = \dfrac{25 + 50}{2} = 37.5$

Length = Upper Limit - Lower limit

= 50 - 25

= 25 //

d) Length of the class :-

- Length = Upper Limit - Lower Limit

* Mesaures of Central Tendency :-

→ These are also known as Statistical Averges or Averages.

→ It is a values around which all the observations have a tendency to cluster.

→ Such a value can be termed as most representative figure of the entire data-set.

→ Mean, Median and Mode.

* Mean :-

→ Sum of observed values of ~~a set~~ a set divided by the no. of observations in the set.

→ Mean value = Average value.

→ If variable 'X' assumes 'N' values i.e $X_1, X_2, X_3, \cdots, X_N$ then the mean is given by,

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N}$$

i.e, $\boxed{\bar{X} = \frac{\Sigma X}{N}}$

→ This is for ungrouped or raw data

→ Eg: A student's marks in 6 subjects are 15,25,35,45, 45, 65. Find the average marks scored by student.

Soln: Let $X_1 = 15$, $X_2 = 25$, $X_3 = 35$, $X_4 = 45$, $X_5 = 45$, $X_6 = 65$

$N = 6$

$$\therefore \bar{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6}{N}$$

$$\therefore \bar{x} = \frac{15 + 25 + 35 + 45 + 45 + 65}{6}$$

$$\boxed{\therefore \bar{X} = 38.33}$$

→ Shortcut Method :

- This method is based on assumed mean.
- Here, mean is given by,

$$\boxed{\bar{X} = A \pm \frac{\Sigma d}{N}}$$

where, $\bar{X}$ = Mean / arithematic Mean

A = assumed Mean

d = deviations of the items from assumed mean. (X - A)

N = No. of observations.

* PROBLEMS :

1) Calculate arithmetic mean of height of 10 students in a class

Roll No.	1	2	3	4	5	6	7	8	9	10
Height cms.	160	157	162	155	158	159	161	154	152	156

Solⁿ :- Here, $N = 10$

Let, Height of Students $= x$

$\therefore \Sigma x = 160 + 157 + 162 + 155 + 158 + 159 + 161 + 154 + 152 + 156$

$\therefore \Sigma x = 1574$

$\therefore$ Arithmetic mean is,

$$\bar{x} = \frac{\Sigma x}{N}$$

$$\therefore \bar{x} = \frac{1574}{10}$$

$$\boxed{\therefore \bar{x} = 157.4 \text{ cms}}$$

2) Marks obtained by 10 students in Statistics are,

Roll no.	1	2	3	4	5	6	7	8	9	10
Marks	46	64	55	47	74	78	88	90	59	63

Calculate arithmetic mean by Shortcut methods.

Solⁿ :- $N = 10$; Let, the assumed mean $(A) = 74$

Let, 'x' represent marks obtained by 10 students.

Roll No.	Marks (x)	$d = x - A = x - 74$
1	46	-28
2	64	-10
3	55	-19
4	47	-27
5	74	0
6	78	4
7	88	14
8	90	16
9	59	-15
10	63	-11
		$\Sigma d = -76$

$$\therefore \bar{X} = A + \frac{\Sigma d}{N}$$

$$= 74 - \frac{76}{10}$$

$$= 66.4 \;//$$

* Geometric mean :-

$$G.M = N\sqrt{x_1 + x_2 + \cdots + x_n} \quad \left(\sqrt[N]{x_1 + x_2 + \cdots + x_n}\right)$$

For individual observations,

$$G.M = \text{Antilog}\left(\frac{\Sigma \log x}{N}\right)$$

Eg:- Calculate geometric mean of 82, 72, 37, 750, 976

Sol^n:-

X	Log (X)
82	1.914
72	1.857
37	1.568
750	2.875
976	2.989

$$\Sigma \log(x) = 11.203$$

$$\therefore G.M = \text{Antilog}\left(\frac{\Sigma \log x}{N}\right)$$

$$= \text{Antilog}\left(\frac{11.203}{5}\right)$$

$$= \text{Antilog}(2.2406)$$

$$= \cancel{174.020} \; 174.020$$

Calculate geometric mean for the following data:

Production	145	135	149	146	150
No. of factories	6	4	2	3	1

The frequency of production in factories is 3.

Soln:-

Production (x)	Log x	No. of factories	$f \times \log(x)$
145	2.161	6	6.483
135	2.130	4	6.39
149	2.173	2	6.519
146	2.164	3	6.492
150	2.176	1	6.528
	$\Sigma \log x =$	$N = 16$	$\Sigma f \log(x) = 32.412$

$$G.M = \text{Antilog}\left(\frac{\Sigma f \log(x)}{N}\right)$$

$$= \text{Antilog}\left(\frac{32.412}{16}\right)$$

$$= \text{Antilog}\left(\overset{2.02575}{\cancel{6.482}}\right)$$

$$= 106.108$$

For Continuous series,

$$G.M = \text{Antilog}\left(\frac{\Sigma f \log m}{N}\right)$$

where, f = frequency

m = mid value of each class

Eg:- Find mid value of 10 - 13 $= \frac{10+13}{2} = 11.5$

20-30 $= \frac{20+30}{2} = 25$

Q. Find out the geometric mean of the following :

Yield of wheat (in tones)	No. of farms
10.5 - 13.5	9
13.5 - 16.5	19
16.5 - 19.5	23
19.5 - 22.5	7
22.5 - 25.5	4

(if frequency is not given consider it as No. of farms)

Soln:-

Here, f = frequency of no. of farms

m = midpoint of each class of yield of wheat

Midpoint (m)	Log m	f	f log m
12	1.079	9	9.711
15	1.176	19	22.344
18	1.255	23	28.865
21	1.322	7	9.254
24	1.380	4	5.520
		N = 62	$\Sigma f \log m = 75.694$

$$\therefore G.M = \text{Antilog}\left(\frac{\Sigma f \log m}{N}\right)$$

$$= \text{Antilog}\left(\frac{75.694}{62}\right)$$

$$= \text{Antilog}(1.221)$$

$$\therefore G.M = 16.629 \text{ tones}$$

* Harmonic Mean :-

→ It is the reciprocal of the arithmetic average of the reciprocal of values of various items in the variable.

→ Mathematically,

$$\boxed{H.M = \frac{N}{1/x_1 + 1/x_2 + \cdots + 1/x_n}}$$

$$\boxed{\therefore H.M = \frac{N}{\Sigma 1/x}}$$

Where, N = no. of observations

$x_1, x_2, \ldots, x_n$ = values in the observation.

Example, Compute Harmonic mean of the following data:

Family	1	2	3	4	5
Income (in Rs. 1000)	70	75	42	36	40

Solⁿ:- Let, x = income in Rs. Thousand

Family	Income (in Rs. 1000)	Reciprocal (1/x)
1	70	0.014
2	75	0.013
3	42	0.024
4	36	0.028
5	40	0.025
N = 5		Σ 1/x = 0.104.

$\therefore H.M = \frac{N}{\Sigma 1/x} = \frac{5}{0.104}$ = Rs. 48.077 Thousand.

* Harmonic Mean-Discrete Series :-

→ Here,

$$\boxed{H.M = \frac{N}{\Sigma f(1/x)}}$$

where, f = frequency.

Example:- Calculate Harmonic mean of the following data:

Size of Items	1	2	3	4	5
Frequency	9	6	5	8	2
Items Produced	8	7	9	11	10

Soln :-

Size of Items	Frequency (f)	Reciprocal (1/x)	$f(\frac{1}{x})$
1	9	0.125	1.125
2	6	0.0143	0.853
3	5	0.111	0.555
4	8	0.091	0.728
5	2	0.1	0.2

$\Sigma f(1/x) = 3.466$

$$\therefore H.M = \frac{N}{\Sigma f(1/x)}$$

$$= \frac{5}{3.466}$$

$$= 1.443 \ //$$

* PROBLEMS :-

* Harmonic Mean- Continuous Series :-

→ Here,

$$\boxed{H.M = \frac{N}{\Sigma f(\frac{1}{m})}}$$

where, m = midpoint of each class
f = frequency.
N = no. of observations based on frequency

Example :- Calculate Harmonic mean of the following data:

Marks	10-20	20-30	30-40	40-50	50-60
Frequency	8	20	35	45	55

Sol:

Midpoint (m)	Reciprocal (1/m)	frequency (f)	(f × 1/m)
15	0.067	8	0.5.336
25	0.040	20	0.800
35	0.286	35	~~10.010~~ 1.015
45	0.022	45	0.990
55	0.018	55	0.990
N=163			$\Sigma f(1/m) =$ ~~18.126~~ 4.331

$$\therefore H.M = \frac{N}{\Sigma f(1/m)}$$

$$= \frac{163}{4.331}$$

$$= 37.636$$

* **Median :-**

→ It is a middle value in a distribution.

→ Mathematically,

$$\boxed{M = \left[\frac{N+1}{2}\right]^{th} \text{ item.}}$$

Eg :

Sr. No.	1	2	3	4	5	6	7
Weight (in kg)	48	59	67	56	50	60	78

Find out the median of the weight of students.
(arrange weights in ascending order)

Sol^n :- Here, N = 7

48, 50, 56, 59, 60, 67, 78

$\therefore$ Median $= \left[\frac{N+1}{2}\right]^{th}$ item

$= \left[\frac{7+1}{2}\right]^{th}$ item

$\therefore$ Median $= 4^{th}$ item.

$\therefore$ The median ~~is~~ weight of students is ~~56~~ 59 kg.

* **Mode :-**

→ It is a value that has occured most frequently in a distribution.

→ Example : 1) In a distribution, we have the numbers
75, 86, 75, 83, 84, 75, 83.

Here, 83 has occured 2 times.
75 has occured 3 times

$\therefore$ Mode = 75 //

2) In a distribution, we have 70, 72, 73, 56, 37, 62, 55 and 85.
Calculate the Mode.

Sol^n :- Since, the numbers/values ~~are~~ have appeared once, there is no mode.

* Harmonic Mean- Continuous Series:-

→ Here,

$$\boxed{H.M = \frac{N}{\Sigma f\left(\frac{1}{m}\right)}}$$

where, m = midpoint of each class

f = frequency.

N = no. of observations based on frequency

Example: Calculate Harmonic mean of the following data:

Marks	10-20	20-30	30-40	40-50	50-60
Frequency	8	20	35	45	55

Sol:

Midpoint (m)	Reciprocal (1/m)	frequency (f)	(f × 1/m)
15	0.067	8	0.5336
25	0.040	20	0.800
35	0.286	35	~~10.010~~ 1.015
45	0.022	45	0.990
55	0.018	55	0.990
N=163			$\Sigma f(1/m)$ = ~~18.126~~ 4.331

$$\therefore H.M = \frac{N}{\Sigma f(1/m)}$$

$$= \frac{163}{4.331}$$

$$= 37.636$$

* **PROBLEMS :-**

- For Calculating median arrange the items in ascending order.

1) Find out the median marks scored by 9 students.

Roll No.	1	2	3	4	5	6	7	8	9
Marks	54	23	81	75	56	82	64	36	28

Soln :- Arranging marks in ascending order :

Marks	23	28	36	54	56	64	75	81	82
Roll No.	2	9	8	1	5	7	4	3	6

Here, $N = 9$

$\therefore$ Median $= \left[\frac{N+1}{2}\right]^{th}$ item.

$\therefore$ Median $= \left[\frac{9+1}{2}\right]^{th}$ item

$= 5^{th}$ item

$\therefore$ Median marks scored by students is 56 //

2) Find out the median of the following:

52, 53, 56, 37, 33, 60, 67, 61

Soln :- Arranging the values in ascending order, we have,

33, 37, 52, 53, 56, 60, 61, 67.

Here, $N = 8$

$\therefore$ Median $= \left[\frac{N+1}{2}\right]^{th}$ item

$= \left[\frac{8+1}{2}\right]^{th}$ item

$\therefore$ Median $= 4.5^{th}$ item

$\therefore$ Value of 4.5^{th} item $= \frac{53+56}{2}$

$= 54.5$

$\therefore$ Median $= 54.5$ //

* Range :-

→ Difference between the largest value and smallest value of the variables.

∴ Range = Largest value - Smallest value.

$$\boxed{\therefore R = L - S}$$

⇒ $\boxed{\text{Coefficient of Range} = \frac{L-S}{L+S}}$

→ Example: Find the range and its Coefficient of weights of 10 Students. The weights are 65, 19, 86, 15, 17, 18, 8, 4, 9, 7.

Soln:- Here,

Largest value (L) = 86

Smallest value (S) = 4

$\therefore$ Range (R) = L - S

= 86 - 4

$\therefore$ R = 82 //

$\therefore$ Coefficient of Range $= \frac{L-S}{L+S}$

$= \frac{86-4}{86+4}$

$= \frac{82}{90}$

$\therefore$ Coefficient of Range = 0.911 //

* Partition Values :-

→ If values are arranged in ascending order, then it can be divided into :-

a) Quartiles
b) Deciles
c) Percentiles

} Quantiles.

* Quartiles :-

→ 3 Values, i.e, Q_1, Q_2 and Q_3 of a data set.

→ First Quartile (Q_1) :- Lowest value greater than 1/4 th (25%) of the data.

→ Second Quartile (Q_2) :- Lowest value greater than half of the data (50%).

→ Third Quartile (Q_3) :- Lowest value greater than 3/4 th (75%) of the data.

→ Consider we have Raw or Ungrouped data.

→ Arrange the data in increasing order and the quartile deviation is given by,

$$\boxed{Q.D = \frac{Q_3 - Q_1}{2}}$$

where, $Q_3 = 3\left[\frac{n+1}{4}\right]^{th}$ item

$Q_1 = 1\left[\frac{n+1}{4}\right]^{th}$ item

Example : Compute quartiles for the data given below -

25, 18, 30, 8, 15, 5, 10, 35, 40, 45

Soln :- No. of items (n) = 10

Arranging the items in ascending order,

5, 8, 10, 15, 18, 25, 30, 35, 40, 45

$\therefore Q_1 = \left[\frac{n+1}{4}\right]^{th}$ item

$= \left[\frac{10+1}{4}\right]^{th}$ item

$= [2.75]^{th}$ item

$\boxed{= 2^{nd} \text{ item} + \frac{3}{4}(3^{rd} \text{ item} - 2^{nd} \text{ item})}$

$= 8 + \frac{3}{4}(10-8)$

$= 8 + 3/4 \times 2$

$\therefore Q_1 = 9.5$ //

$\therefore Q_3 = \frac{3}{\cancel{4}}\left[\frac{n+1}{4}\right]^{th}$ item

$= 3\left[\frac{10+1}{4}\right]^{th}$ item

$= 3(2.75)^{th}$ item

$= (8.25)^{th}$ item

$\boxed{= 8^{th} \text{ item} + \frac{1}{4}(9^{th} \text{ item} - 8^{th} \text{ item})}$

$= 35 + \frac{1}{4}(40-35)$

$= 35 + \frac{1}{4} \times 5$

$\therefore Q_3 = 36.25$ //

$\therefore Q_2 = \left[\frac{n+1}{2}\right]^{th}$ item

$= \left[\frac{10+1}{2}\right]^{th}$ item

$= (5.5)^{th}$ item

$\boxed{= 5^{th} \text{ item} + \frac{1}{2}(6^{th} \text{ item} - 5^{th} \text{ item})}$

$= 18 + \frac{1}{2}(25-18)$

$= 18 + 7/2$

$\therefore Q_2 = 21.5$ //

... deviation from mode

* Deciles :-

→ Decile is given by,

$$\boxed{D_i = \left[\frac{i(n+1)}{10}\right]^{th} \text{ observation.}}$$

where, n = no. of values in the data
i = decile class

→ Eg: Consider the data 5, 24, 36, 12, 20, 8. Compute 5th decile

Soln: Arranging data in ascending order,

5, 8, 12, 20, 24, 36.

Here, n = 6
i = 5

$$\therefore D_5 = \left[\frac{5(n+1)}{10}\right]^{th} \text{ observation}$$

$$= \left[\frac{5(6+1)}{10}\right]^{th} \text{ observation.}$$

$$= (3.5)^{th} \text{ observation}$$

$$= 3^{rd} \text{ item} + \frac{1}{2}\left(4^{th} \text{ item} - 3^{rd} \text{ item}\right)$$

$$= 12 + \frac{1}{2}(20-12)$$

$$= 12 + 1/2 \times 8$$

$$= 12 + 4$$

$$\therefore D_5 = 16$$ //

→ Decile is each of ten equal groups into which a population can be divided as per distribution.

* Percentile :-

→ Percentile is each of 100 equal groups into which a population can be divided as per distribution.

→ ∴ Percentile is given by,

$$\boxed{P_i = \left[\frac{i(n+1)}{100}\right]^{th} \text{ observation}}$$

where, n = no. of values in the data
i = Percentile class.

→ Example :- 1) Consider the data 5, 24, 36, 12, 20, 8. Compute P_{15}.

Soln :- Arranging data in ascending order,

5, 8, 12, 20, 24, 36

$n = 6$

$i = 15$

$\therefore P_{15} = \left[\frac{15(n+1)}{100}\right]^{th}$ observation

$\therefore P_{15} = \left[\frac{15(6+1)}{100}\right]^{th}$ observation

$\therefore P_{15} = (1.05)^{th}$ observation

$\therefore P_{15} = 1^{th}$ item $+ \frac{1}{20}$ (2^{nd} item $-$ 1^{st} item)

$= 5 + \frac{1}{20}(8-5)$

$= 5 + \frac{3}{20}$

$\therefore P_{15} = 5.15$ //

2) Compute quartiles, deciles and percentiles of the given data:

52, 72, 83, 94, 25, 36, 12, 48, 74, 84, 92, 7, 1, 5, 14, 16, 20, 35

Decile = D_4

Percentile = P_{20}

Soln :- No. of items (n) = 18

Arranging the items in ascending order,

1, 5, 7, 12, 14, 16, 20, 25, 35, 36, 48, 52, 72, 74, 83, 84, 92, 94

→ Quartiles :

$Q_1 = \left[\frac{n+1}{4}\right]^{th}$ item

$= \left[\frac{18+1}{4}\right]^{th}$ item

$= (4.75)^{th}$ item

$= 4^{th}$ item $+ \frac{3}{4}$ (5^{th} item $-$ 4^{th} item)

$= 12 + \frac{3}{4}(14-12)$

$= 12 + 3/4 \times 2$

$Q_1 = 13.5$ //

$$Q_2 = \left[\frac{n+1}{2}\right]^{th} \text{ item}$$
$$= \left[\frac{18+1}{2}\right]^{th} \text{ item}$$
$$= (9.5)^{th} \text{ item}$$
$$= 9^{th} \text{ item} + \frac{1}{2}(10^{th} \text{ item} - 9^{th} \text{ item})$$
$$= 35 + \frac{1}{2}(36-35)$$
$$= 35 + 1/2$$
$$\therefore Q_2 = 35.5$$

$$\therefore Q_3 = 3\left[\frac{n+1}{4}\right]^{th} \text{ item}$$
$$= 3\left[\frac{18+1}{4}\right]^{th} \text{ item}$$
$$= \cancel{3(9.5)}\ 3(4.75)^{th} \text{ item}$$
$$= (14.25)^{th} \text{ item}$$
$$= 14^{th} \text{ item} + 1/4(15^{th} \text{ item} - 14^{th} \text{ item})$$
$$= 74 + 1/4(83-74)$$
$$\therefore Q_3 = 76.25$$

→ Deciles : $i = 4$

$$\therefore D_4 = \left[\frac{4(n+1)}{10}\right]^{th} \text{ item}$$
$$= \left[\frac{4(18+1)}{10}\right]^{th} \text{ item}$$
$$= \left[\frac{4\times19}{10}\right]^{th} \text{ item}$$
$$= [7.6]^{th} \text{ item}$$
$$= 7^{th} \text{ item} + \frac{3}{5}(8^{th} \text{ item} - 9^{th} \text{ item})$$
$$= 20 + \frac{3}{5}(25-20)$$
$$= 20 + 3$$
$$= 23$$
$$\boxed{\therefore D_4 = 23}$$

→ Percentiles : $i = 20$

$$\therefore P_{20} = \left[\frac{20(n+1)}{100}\right]^{th} \text{ item}$$
$$= \left[\frac{20(18+1)}{100}\right]^{th} \text{ item}$$
$$= (3.8)^{th} \text{ item}$$
$$= 3^{rd} \text{ item} + \frac{4}{5}(4^{th} \text{ item} - 3^{rd} \text{ item})$$
$$= 7 + \frac{4}{5}(12-7)$$
$$= 7 + 4$$
$$= 11$$
$$\therefore P_{20} = 11$$

* Measures of Dispersion :-

→ Consider the data-set :

a) 5, 5, 5, 5, 5

b) 3, 4, 5, 6, 7

c) 1, 3, 5, 7, 9

} Mean = 5, Median = 5

→ Thus, average values, i.e Mean = Median.

→ Here, average fails to give any idea about the spread of observations in data set.

→ This spread or scatteredness is called as dispersion.

→ So, the measures of dispersion :

(i) Range

(ii) Mean deviation

(iii) Standard Deviation.

* Mean Deviation :-

→ It is the average of difference of values of items from some average of the series.

→ Technically the difference is termed as deviation.

→ In calculation of mean deviation, we ignore negative sign of the deviation.

→ Symbol : δ (delta)

→ Mean deviation from mean,

$$\boxed{\delta_{\bar{x}} = \frac{\Sigma |x_i - \bar{x}|}{n}}$$

... if deviations $|x_i - \bar{x}|$ are obtained from arithmetic average.

where, x_i = i^{th} values of variable X

$\bar{x}$ = Arithmetic average.

n = No. of items

→ Mean deviation from median,

$$\boxed{\delta_M = \frac{\Sigma |x_i - M|}{n}}$$

... if deviations $|x_i - M|$ are obtained from median.

where, M = Media.

* Mathematical Expectation :

→ Mean deviation from mode,

$$\boxed{\delta_Z = \frac{\Sigma |x_i - Z|}{n}}$$

... if deviations $|x_i - Z|$ are obtained from Mode.

where, Z = mode

* Standard Deviation :

→ Denoted by σ (Sigma).

→ It is the square root of average of squares of deviation, when such deviation for the values of individual items in a series are obtained from the arithmetic average.

→ Mathematically,

$$\boxed{S.D = \sigma = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n}}}$$

where, x_i = i^{th} values of variable x

$\bar{x}$ = Arithmetic average

n = No. of items.

→ In case of frequency distribution, if 'f_i' is the frequency of i^{th} item, then the standard deviation is given by,

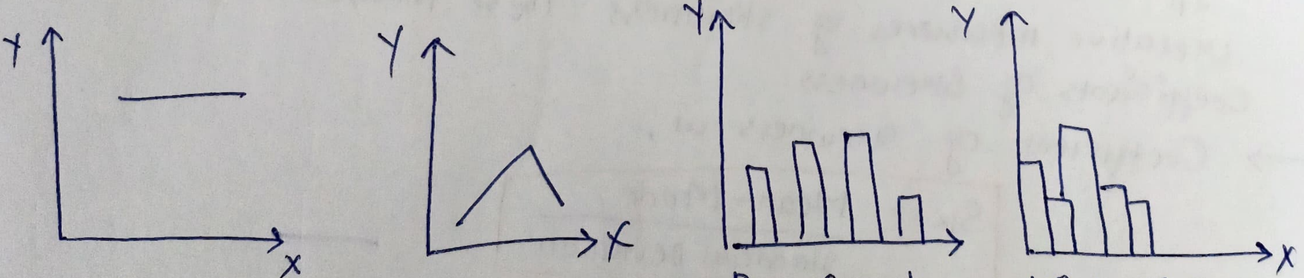
$$\boxed{\sigma = \sqrt{\frac{\Sigma (f_i (x_i - \bar{x})^2)}{\Sigma f_i}}}$$

* Measure of Skewness :

→ Skewness : Means Lack of Symmetry

→ Skewness gives the idea of the shape of distribution of data.

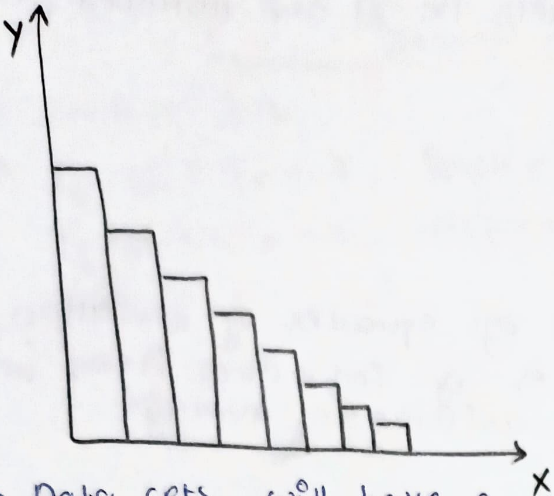
→ Consider the data sets where Skewness is not present and the data sets can be presented by line Chart or Histogram or bar graphs

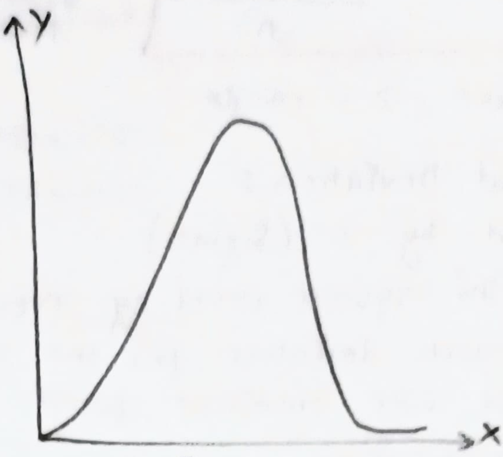


Line Chart Line Chart Bar Graph Histogram

SYMMETRIC DATA SETS

Skewed Data Sets Representation





→ Data sets will have a skewed distribution when,

$\boxed{\text{Mean} \neq \text{Median} \neq \text{Mode (and all three exists)}}$

→ If curve is stretched towards Right more, then we Positive Skewness.

→ If curve is stretched towards Left more, then it is Negative Skewness

→ Difference between Mean, Median or Mode provides a way of expressing skewness in a data-set.

→ If $\boxed{\text{Mode} < \text{Median} < \text{Mean}}$, then we have Positive skewness.

→ If $\boxed{\text{Mean} < \text{Median} < \text{Mode}}$, then we have Negative skewness.

→ Skewness is measured by (Mean - Mode).

→ In case of Mode is not defined, it can be estimated from Mean and Median.

→ For a moderately asymmetrical distribution,

$$\boxed{\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}}$$

→ In order to Compare skewness of 2 data sets, we obtain relative measures of skewness. These measures are called as Coefficients of Skewness.

→ Coefficient of skewness is,

$$\boxed{S_K = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}}$$

→ If Mode is ill-defined,

$$\boxed{S_K = \frac{3(\text{Mean}-\text{Median})}{\text{Standard Deviation}}}$$

→ Limits : ± 3.

→ If $0 < S_K < 3$, Positive Skewness

If $-3 < S_K < 0$, Negative Skewness.

* Population Method / Census Method :-

→ If secondary data is not available, primary data may be collected.

→ Primary data can be obtained by population method also known as census method.

→ In census method, data is collected for each and every unit of population.

→ Types of population :

a) Finite or Infinite Population :-

- Finite → Number of elements can be counted.

Eg : i) No. of students in a college

ii) No. of people in a village

iii) No. of schools or colleges in a city.

- Infinite → Number of elements cannot be counted

Eg : i) No. of stars in the sky

ii) No. of viewers in TV programme

iii) No. of readers in a newspaper.

b) Real or Hypothetical Population :-

- Real → Elements / items really exists

Eg : i) No. of factories in a district

ii) No. of people in city

- Hypothetical → Elements / items may not really exist.

Eg : i) Tossing a coin

→ Merits of Census Method :

1) Data is obtained from each and every unit of the population.

2) Data results are more representative, accurate and reliable

3) Characters of all the limits of the population can be studied.

4) Intensive study is possible.

→ **Demerits of Census Method :**

1) Requires more time and cost.
2) Not required for all types of research
3) Inefficient and inexperienced researcher may collect wrong information.

* Sampling :-

→ Sampling is the process of learning about the population on the basis of a sample draw from it.

→ Process of sampling :

a) Selecting the sample
b) Collecting the Information
c) Making an inference about the population.

→ **Merits :**

1) Organisation and administration of sample surveys are easy
2) It reduces cost, time and energy.
3) Gives Accurate result.
4) Detailed enquiry can be made.

→ **Demerits :**

1) Data is not obtained from each and every unit of population.
2) Reliable results are not possible.

* Methods of Sampling :-

Methods of Sampling

- Probability Sampling OR Random Sampling
  - Simple or Unrestricted Random Sample
  - Restricted Random Sampling
    - stratified samping
    - Systematic Sampling
    - cluster sampling
- Non-Probability Sampling OR Non-Random Method
  - Judgement Sampling
  - Quota Sampling
  - Convenience Sampling

* Probability or Random Sampling :-

→ Each item of the population has equal chance of being selected

→ Selection of sample items is independent of the person making the study.

a) Simple or Unrestricted Random Sampling :

- Sample drawn is such that each and every unit of the population has an equal and independent chance of being included in the sample.
- Data is collected using :

A) Lottery Method - All items in a population are numbered on separate slips of paper of same size and same color

- Slips are then folded and put in a box or container.
- Required no. of slips are selected from the box or container

B) Table of Random Numbers -

- Random table is used to select the sample of the population size is very large.
- Random Table is a table of digits generated through a random process.

b) Restricted Random Sampling :

1) Stratified Sampling method - Population to be sampled is subdivided into groups which are mutually exclusive and include all the items in the population.

A simple random sample is chosen independently from each of the group.

Here, a sample is a secured and greater accuracy is obtained.

2) Systematic Sampling / Quasi-Random Sampling -

- This method is used in the cases where a complete cost of the population from which the sample is to be drawn is available.
- List can be prepared in alphabetical, numerical, geographical or some other order.
- Items are serially numbered.
- Here, the first item is selected at random by lottery method.

- Then, the subsequent items are selected by taking k^{th} item from the list, where 'K' is sampling interval/sampling ratio.
- Thus, $\boxed{K = \frac{N}{n}}$ where, N = Population size/universe size
  n = Sample size

3) Multi-Stage Sampling or Cluster Sampling :
- Random selection is made of primary, intermidiate and final units from a given population.
- Cluster sampling is carried out in stages.
- For example :- Consider 5000 households from Maharashtra State.

  Stage 1 : Divide state into number of districts.

  Stage 2 : Divide districts into villages and take sample of villages randomly.

  Stage 3 : Select no. of households from each each village
- Gives flexibility in sampling method.
- Covers large area

* Non-Probability Sampling :-

a) Judgement Sampling / Purposive Sampling / Deliberate Sampling :
- Choice of sample items depends on judgement of investigator.

b) Quota Sampling :
- Population is divided into various quotas w.r.t some common character.
- Required sample is selected from the quota.

c) Convenience Sampling :
- Chunk method
- Chunk → fraction of the population is investigated which is selected neither by probability nor by judgement but by convenience.

* Size of Sample :-

→ No. of sampling units selected from population.

→ Size of sample should be optimum (not too large and not too small).

→ Factors considered for sample size :-

a) Size of the universe / population

b) Availability of the resources

c) Degree of accuracy : Larger the sample, degree of accuracy is greater

d) Homogeneity or Heterogeneity of the universe :-

Homogeneous = Sample size small

Heterogeneous = Large Sample

e) Nature of study

f) Method of Sampling

g) Nature of Respondents

* Sampling Errors :-

→ Difference in value of sample and population.

→ Types :

a) Biased Error -

- Any bias in selection, estimation in sampling.
- Occurs in Non-Random Sampling methods.

b) Unbiased Error -

- Result may vary with actual result of the population.
- The difference is called as unbiased error.

* Measurement of Errors :-

1) Absolute Error : It is the difference between approximate figure and the original figure.

$$\boxed{\therefore \text{Absolute Error} = \underset{(a)}{\text{Actual value}} - \underset{(e)}{\text{Estimate value}}}$$

Eg :- In a college, avg. estimated height of students is 160 cm and the actual avg. height is 162 cm.

$\therefore$ Absolute error $= 162 - 160$

$= 2$ cm //

2) Relative Error :- Ratio of absolute Error to the estimated value.

$$\text{Relative Error} = \frac{\text{Absolute Error}}{e}$$

$$\boxed{\therefore R.E = \frac{a-e}{e}}$$

Eg :- If actual value is 1250 and estimated value is 1200, th

$$R.E = \frac{1250-1200}{1200}$$

$\therefore R.E = 0.042$ //

* PROBLEMS :-

1) In a class there are 96 students with roll nos. from 1 to 96. It is desired to take a sample of 10 students. Use the systematic sampling method to determine sample size/ratio.

Soln :- $N = 96$ (Population size)

Sample size $= n = 10$

Thus, the sample ratio is,

$$\therefore K = \frac{N}{n}$$

$$\therefore K = \frac{96}{10}$$

$\therefore K = 9.6$ // ≈ 10

* Mathematical Expectation :-

→ If 'x' is a random variable, which assumes any one of the possible values $x_1, x_2, \dots, x_n$ with probabilities $P(x_1), P(x_2), \dots P(x_n)$.

→ Thus, mathematical expectation is given by,

$$\boxed{E(X) = \sum_{i=1}^{n} x_i P(x_i)}$$ ... if 'x' is a discrete random variable

→ For eg :

The PMF (Probability Mass Function) of X is,

x	0	1	2
$P(x)$	1/4	1/2	1/4

$$\therefore E(X) = \sum_{i=1}^{3} x_i P(x_i)$$

$$= x_1 P(x_1) + P(x_2) x_2 + x_3 P(x_3)$$

$$= 0 \times 1/4 + 1 \times 1/2 + 2 \times 1/4$$

$$= \frac{1}{2} + \frac{1}{2}$$

$$\therefore \underline{E(X) = 1}$$

→ Expectation is known as Mean.

* Variance :-

→ It is a measurement of the spread between numbers in a data set.

→ It measures how far each number in the set is from mean.

→ It is denoted by $V(X)$ or σ^2.

→ Square root of variance is standard deviation.

$$\boxed{\therefore S.D. (\sigma) = \sqrt{V(X)}}$$

→ Also,

$$V(X) = E[X - E(X)]^2$$

$$\boxed{\therefore V(X) = E(X^2) - [E(X)]^2}$$

* Simple Random Sampling Properties :-

→ Terminologies :

N = Population size

n = Sample size

X_i = value of character under study for i^{th} unit in the population.

x_i = Value of character under study for i^{th} unit in the sample.

→ Population Mean is given by,

$$\boxed{\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i}$$ (Also, denoted by μ)

→ Sample Mean is given by,

$$\boxed{\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

→ Population Variance,

$$\boxed{\sigma^2 = S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2}$$

It is also termed as population mean square.

→ Sample Variance,

$$\boxed{s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

It is also termed as Sample mean square.

* Simple Random Sampling with Replacement (SRSWR) :-

→ No. of samples that can be selected with replacement

$$\boxed{= N^n}$$

* Simple Random Sampling without Replacement (SRSWOR) :-

→ No. of samples that can be selected without replacement $= \boxed{{}^{N}C_n}$

Note :- ${}^{n}C_r = \frac{n!}{r!(n-r)!}$ //

* PROBLEMS :-

Q1. Suppose a population consists of 5 units i.e 4, 5, 7, 9, 10. How many samples of size 2 can be drawn from it? Give the possible samples.

Soln :- Here, $N = 5$

$n = 2$

a) SRSWR :-

No. of samples that can be selected with replacement $= N^n$

$= 5^2$

$= 25$

$\therefore$ The possible samples are given by.

(4,4), (4,5), (4,7), (4,9), (4,10), (5,4), (5,5), (5,7), (5,9), (5,10)
(7,4), (7,5), (7,7), (7,9), (7,10), (9,4), (9,5), (9,7), (9,9), (9,10),
(10,4), (10,5), (10,7), (10,9), (10,10).

b) SRSWOR :-

No. of samples that can be selected without replacement $= {}^N C_n$

$= {}^5 C_2 = \dfrac{5!}{2!\,(5-2)!}$

$= 10$

$\therefore$ The possible samples are given by,

(4,5), (4,7), (4,9), (4,10), (5,7), (5,9), (5,10), (7,9), (7,10),
(9,10).

Q2. Calculate the population Mean and population Standard deviation for the population : 2, 5, 8, 13.

Soln :- $N = 4$

$\therefore$ Population Mean $(\bar{x}) = \dfrac{1}{N} \sum_{i=1}^{N} x_i$

$= \dfrac{1}{4}(2+5+8+13)$

$= 7$

$\therefore$ Population variance $= \sigma^2 = \dfrac{1}{N-1} \sum_{i=1}^{4} (x_i - \bar{x})^2$

$= \dfrac{1}{3}\left[(-5)^2 + (-2)^2 + 1^2 + 6^2\right]$

$\therefore \sigma^2 = 22$

∴ Population standard deviation $= \sqrt{\text{Population variance}}$

$$= \sqrt{22}$$

$$\sigma = 4.690$$

Q3. A population of 5 units has the observations :-
7, 6, 8, 4, 10.
Random samples of 3 units are drawn such that no unit is repeated in a sample and the order of selection doesn't matter. Out of 5 units, there can be 10 samples each consisting of 3 units. Find the possible samples, their means and variances. Also find population mean and its variance.

Soln:- Population size (N) = 5
Sample size (n) = 3

The possible samples, their means and variances are as follows:-

Sample No.	1	2	3	4	5	6	7	8	9	10
Observation	(7,6,8)	(7,6,4)	(7,6,10)	(7,8,4)	(7,8,10)	(7,4,10)	(6,8,4)	(6,8,10)	(6,4,10)	(8,4,10)
$\bar{x}$	7	5.667	7.667	6.33	8.33	7	6	8	$20/3$	$22/3$
s^2	1	$\frac{7}{3}$	$\frac{13}{3}$	$\frac{13}{3}$	$\frac{7}{3}$	9	4	4	$\frac{28}{3}$	$\frac{28}{3}$

$$\text{Mean of } \bar{x} = \frac{7 + 17/3 + \frac{23}{3} + \frac{25}{3} + \frac{19}{3} + 7 + 6 + 8 + \frac{20}{3} + \frac{22}{3}}{10}$$

$$\therefore \text{Mean of } \bar{x} = 7.$$

Also, Population Mean,

$$\bar{X} = \frac{7+6+8+4+10}{5}$$

$$\boxed{\therefore \bar{X} = 7}$$

$$\text{Also, Population variance} = \frac{1}{4}\left[(7-7)^2 + (6-7)^2 + (8-7)^2 + (4-7)^2 + (10-7)^2\right]$$

$$\boxed{\therefore s^2 = 5}$$

Q.4 Consider the population :- 2, 5, 6, 13
Construct a sampling distribution of the sample mean when random samples of size 2 are selected from the population a) with replacement
b) without replacement. (Probability)
Find the mean and standard error of the distribution in each case. (Standard deviation)

Soln :- NOTE :-
→ Distribution means to find probability
→ Standard error is but standard deviation.

a) SRS WR :

Population size (N) = 4
Sample size (2) = 2

No. of possible samples with replacement $= N^n$
$= 4^2$
$= 16$ //

∴ The possible samples are. (N_1)
(2,5), (2,8), (2,13), (5,2), (5,5), (5,8), (5,13), (8,2), (8,5), (8,8)
(8,13), (13,2), (13,5), (13,8), (13,13), (2,2)

The sample means are :
2, 7/2, 5, 15/2, 7/2, 5, 13/2, 9, 5, 13/2, 8, 21/2, 15/2, 9, 21/2, 13,

The sampling distribution table of Sample Mean :-

$\bar{x}$	No. times mean occured (f)	Probability = f/N_1
2	1	~~1/4~~ 1/16
3.5	2	~~1/2~~ 1/8
5	3	~~3/4~~ 3/16
7.5	2	~~1/2~~ 1/8
6.5	2	~~1/2~~ 1/8
9	2	~~1/2~~ 1/8
8	1	~~1/4~~ 1/16
10.5	2	~~1/2~~ 1/8
13	1	~~1/4~~ 1/16

Here, f = no. of times the sample mean has occurred (frequency)
p = Probability of the occurrence of sample mean.

Here, the mean is,

$$E(\bar{x}) = \sum_{i=1}^{n} x_i P(x_i)$$

$$= 2\times\frac{1}{16} + 3.5\times\frac{1}{8} + 5\times\frac{3}{16} + 7.5\times\frac{1}{8} + 6.5\times\frac{1}{8} \times 9\times\frac{1}{8} + 8\times\frac{1}{16} + 10.5\times\frac{1}{8}$$

$$+ 13\times\frac{1}{16}$$

$$= 7$$

∴ Variance is,

$$V(X) = E(X^2) - [E(X)]^2$$

$$\therefore V(X) = \frac{229}{4} - 49$$

$$\therefore V(X) = \frac{33}{4}$$

$$\therefore \text{standard error } (\sigma) = \sqrt{V(X)} = \sqrt{\frac{33}{4}} = 2.872$$ //

b) SRSWOR:

No. of possible samples without replacement $= {}^{N}C_{n}$

$= {}^{4}C_{2}$

$= 6$

∴ The possible samples are : (N_2)

(2,5), (2,8), (2,13), (5,8), (5,13), (8,13)

∴ The sample means are :

7/2, 5, 15/2, 13/2, 9, 21/2

∴ The sampling distribution table is :

$\bar{x}$	No. of times mean occured (f)	Probability = f/N₂
3.5	1	1/6
5	1	1/6
7.5	1	1/6
6.5	1	1/6
9	1	1/6
10.5	1	1/6

Here, the mean is,

$$E(\bar{x}) = \sum_{i=1}^{n} x_i P(x_i)$$

$$= 3.5 \times \frac{1}{6} + 5 \times \frac{1}{6} + 7.5 \times \frac{1}{6} + 6.5 \times \frac{1}{6} + 9 \times \frac{1}{6} + 10.5 \times \frac{1}{6}$$

$$= (3.5 + 5 + 7.5 + 6.5 + 9 + 10.5) \times 1/6$$

$$= 7$$

∴ Variance is,

$$V(x) = E(x^2) - [E(x)]^2$$

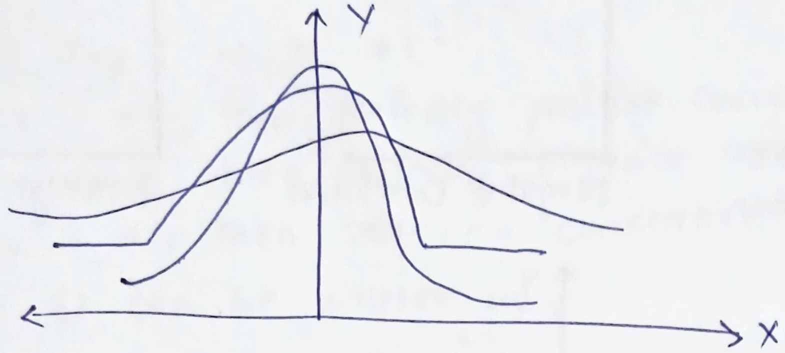
$$\therefore V(x) = \frac{109}{2} - 49$$

$$\therefore V(x) = \frac{11}{2}$$

∴ Standard error $(\sigma) = \sqrt{V(x)} = \sqrt{\frac{11}{2}} = 2.345$ //

* Kurtosis :-

→ It tells about the flatness of the distribution curve.



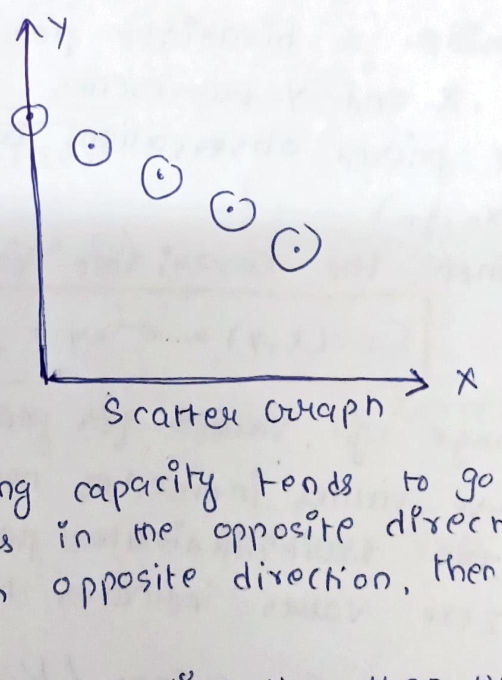
* Correlation :-

→ Example :

Consider, X = no. of cigarettes smoked in a day

Y = Lung capacity

X	Y
0	45
5	42
10	33
15	31
20	29



Scatter Graph

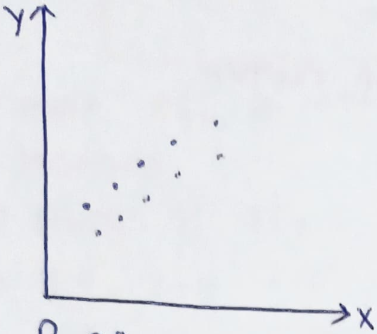
→ As the smoking goes up, the lung capacity tends to go down.

→ Thus, the two variable change is in the opposite direction.

→ If x and y change values in opposite direction, then we have Negative Correlation.

→ If x and y change values in same direction, then we have Positive correlation.

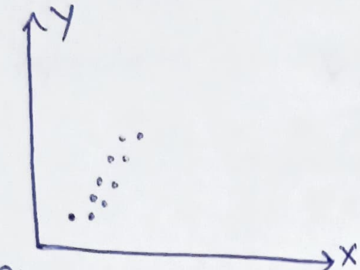
→ Following are the scatter plots for various types of relationships:



Positive Correlation

Negative Correlation.

Weakly Correlation



Strongly Correlated

Not Correlated.

* Covariance :-

→ Consider a bivariate population.

→ Let, X and Y represents the two variables.

→ The paired observations are given by $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

→ Then, the covariance is given as,

$$\boxed{Cov(X,Y) = \sigma_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}$$

→ Range of values for covariance : $-\infty$ to $+\infty$

→ -ve values indicates negative relationship

+ve value indicates positive relationship

zero values indicates No relationship. {Sometimes Non-linear}

* Correlation Coefficient / Karl Person's Coefficient of Correlation :-

→ Mathematically,

$$\boxed{r_{xy} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}} \rightarrow ①$$

where, σ_x = Standard deviation of X

σ_y = Standard deviation of Y

→ Substituting $Cov(X,Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ in eq^n ①

$$\therefore \boxed{r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y}} \rightarrow (2)$$

→ Here r_{xy} is not affected by change in scale or change in location.

→ Range of r_{xy} : -1 to +1

→ If $r_{xy} = +1$, then perfectly positive correlation

If $r_{xy} = -1$, then perfectly negative correlation

If $r_{xy} = 0$, then zero/no correlation.

→ Equation (2) can be written as,

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} \rightarrow (3)$$

→ By shifting the origin and changing the scale, we have,

$$u = \frac{x-a}{c}, \quad v = \frac{y-b}{d}$$

$$r_{uv} = \frac{n\sum uv - \sum u \sum v}{\sqrt{n\sum u^2 - (\sum u)^2}\sqrt{n\sum v^2 - (\sum v)^2}} \rightarrow (4)$$

→ Then, $\boxed{r_{xy} = r_{uv}}$ (from (3) and (4))

* Spearman's Rank Correlation :-

→ It is the technique of determining the degree of correlation between two variables.

→ It is used when ranks are given to the different values of variables.

→ Here, $\boxed{r_s = 1 - \left[\frac{6\sum d_i^2}{n(n^2-1)}\right]}$

Where, d_i = difference between ranks of i^{th} pair of two variables.

n = No. of pairs of observations.

Q. Calculate the correlation coefficient between the income and expenditure of a worker from the following data:

Month	Jan	Feb	March	April	May	June
Income (in Rs.)	451	459	461	461	463	467
Expenditure (in Rs.)	433	437	441	451	455	451

Soln:- Let, income be denoted by 'x' and expenditure be denoted by 'y'

Karl Pearson's coefficient of correlation is,

$$r_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \cdot \Sigma(y-\bar{y})^2}}$$

Now, $\bar{x} = \frac{\Sigma x}{n} = 460.333$

$\bar{y} = \frac{\Sigma y}{n} = 444.667$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
451	433	-9.333	-11.667	87.105	136.119	108.888
459	437	-1.33	-7.667	1.769	58.783	10.197
461	441	0.667	-3.667	0.445	13.447	-2.446
461	451	0.667	6.333	0.445	40.107	4.224
463	455	2.67	10.333	7.129	106.771	27.589
467	451	6.667	6.333	44.449	40.107	42.222
Total				141.342	395.334	190.674

Q.3 Calculate Karl Person's coefficient of correlation between X and Y from the following data:

X	39	65	62	90	82	75	25	98	36	78
Y	47	53	58	86	62	68	60	91	51	84

Soln:- $n = 10$

Karl Pearson's coefficient of correlation is,

$$r_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \cdot \Sigma(y-\bar{y})^2}}$$

Now, $\bar{x} = \frac{\Sigma x}{n} = \frac{294}{5} = \cancel{58.8}\ 65$

$\bar{y} = \frac{\Sigma y}{n} = 66$

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
39	47	-19.8	-19	392.04	361	
65	53	6.2	-13	38.44	169	
62	58	3.2	-8	10.24	64	
90	86	31.2	20	973.44	400	
82	62	23.2	-4	538.24	16	
75	68	16.2	2	262.44	4	
25	60	-33.8	-6	1142.44	36	
98	91	39.2	25	1536.64	625	
36	51	-22.8	-15	519.84	225	
78	84	19.2	18	368.64	324	
				5398	2224	2704

$$\therefore r_{xy} = \frac{108.888 + 10.197 - 2.446 + 4.224 + 27.589 + 42.222}{\sqrt{(87.105 + 1.769 + 0.445 + 0.445 + 7.129 + 44.449)(136.119 + 58.753 + 13.447 + 40.107 + 106.771 + 40.107)}}$$

$$= \frac{190.674}{\sqrt{141.342 \times 395.334}}$$

$$\therefore r_{xy} = 0.806$$

Q. Coefficient of correlation between two variables X and Y is 0.48. The covariance is 36. Variance of X is 16. Find standard deviation of Y.

Soln: Now,

$$r_{xy} = 0.48$$

$$Cov(X,Y) = 36$$

$$V(X) = 16$$

$$\therefore \sigma_x = \sqrt{V(X)} = \sqrt{16} = 4$$

Then,

$$r_{xy} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

$$\therefore \sigma_y = \frac{Cov(X,Y)}{\sigma_x \cdot r_{xy}}$$

$$= \frac{36}{4 \times 0.48}$$

$$\therefore \sigma_y = 18.75$$

∴ The standard deviation of Y is 18.75

* Lines of Regression :-

→ Regression : Stepping back to average value

→ Dependent variable : whose value is to be predicted (explained variable)

→ Independent variable : where value is used for prediction (Predictive / Explanator)

→ Regression eqn of y on x is,

$$\boxed{y - \bar{y} = b_{yx} (x - \bar{x})}$$

where, $b_{yx} = r \frac{\sigma y}{\sigma x}$

→ Regression eqn of x on y is,

$$\boxed{x - \bar{x} = b_{xy} (y - \bar{y})}$$

where, $b_{xy} = r \frac{\sigma y}{\sigma x}$

→ b_{yx} and b_{xy} = regression coefficients

r = Correlation coefficient

→ ~~other~~ Now, $r^2 = b_{xy} b_{yx}$

$\therefore r = \pm \sqrt{b_{xy} \cdot b_{yx}}$

→ If b_{xy} and b_{yx} = +ve then r = +ve

If b_{xy} and b_{yx} = -ve then r = -ve

Also,

$$b_{yx} = \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2}$$ and

$$b_{xy} = \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma y^2 - (\Sigma y)^2}$$

If origin is shifted and scale is changed,

$u = \frac{x - a}{c}$, $v = \frac{y - b}{d}$

$\therefore b_{vu} = \frac{n \Sigma uv - \Sigma u \Sigma v}{n \Sigma u^2 - (\Sigma u)^2}$ and $b_{uv} = \frac{n \Sigma uv - \Sigma u \Sigma v}{n \Sigma v^2 - (\Sigma v)^2}$

Also,

$b_{yx} = b_{vu} \times \frac{d}{c}$ and $b_{xy} = b_{uv} \times \frac{c}{d}$

Q. By using following data, find the lines of regression :

$n = 10$, $\Sigma x = 250$, $\Sigma y = 300$, $\Sigma xy = 4900$, $\Sigma x^2 = 6500$, $\Sigma y^2 = 10000$

Soln :- Given data is,

$n = 10$

$\Sigma x = 250$

$\Sigma y = 300$

$\Sigma xy = 4900$

$\Sigma x^2 = 6500$

$\Sigma y^2 = 10000$

Now, $\bar{x} = \frac{\Sigma x}{n} = \frac{250}{10} = 25$

$\bar{y} = \frac{\Sigma y}{n} = \frac{300}{10} = 30$

Regression coefficients are,

$$b_{yx} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{10 \times 4900 - 250 \times 300}{10 \times 6500 - (250)^2}$$

$$= \cancel{\frac{-52}{5}}\ 1.6$$

$$b_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma y^2 - (\Sigma y)^2}$$

$$= \frac{10 \times 4900 - 250 \times 300}{10 \times 10000 - (300)^2}$$

$$= \cancel{\frac{-13}{25}}\ 0.4$$

Regression eqn of y on x is

$y - \bar{y} = b_{yx}(x - \bar{x})$

$y - 30 = 1.6(x - 25)$

$\therefore y - 30 = 1.6x - 40$

$\therefore y - 30 = 1.6x - 40$

$\boxed{\therefore\ y = 1.6x - 10}$

Regression eqn of x on y is.

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$\therefore x - 25 = 0.4(y - 30)$$

$$\therefore x - 25 = 0.4y - 12$$

$$\boxed{\therefore x = 0.4y + 13}$$

Q. If the tangent of the angle made by the lines of regression of y and x is 0.6. Find correlation coefficient between x and y if $\sigma_y = 2\sigma_x$

Soln :- Given data is,

$$b_{yx} = 0.6$$

$$\sigma_y = 2\sigma_x$$

{Line of regression is,

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$b_{yx} = 0.6$ (tangent of the angle)}

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$\therefore r = b_{yx} \cdot \frac{\sigma_x}{\sigma_y}$$

$$= 0.6 \cdot \frac{\sigma_x}{2\sigma_x}$$

$$\boxed{\therefore r = 0.3}$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

Q. The equations of two lines of regression obtained in a Correlation analysis are :

$2x+3y-3=0$ and $x+2y-5=0$.

Find the mean values of x and y obtain the value of Correlation Coefficient and Variance of y; given that variance of x is 12.

Ans Since, $\bar{x}$ and $\bar{y}$ intersect the lines of regression then,

$$2\bar{x}+3\bar{y}=3 \quad —①$$
$$\bar{x}+2\bar{y}=5 \quad —②$$

solving ① and ②,

$$\bar{x}=-9$$
$$\bar{y}=7$$

Assume $\bar{x}=9$ and $\bar{y}=7$

Regression eq^n of y on x is,

$$2x+3y-3=0$$
$$\therefore 3y=-2x+3$$
$$\boxed{\therefore y=\frac{-2}{3}x+1}$$

Here, $b_{yx}=-\frac{2}{3}$

Regression eq^n of x on y is,

$$x+2y-5=0$$
$$\boxed{\therefore x=-2y+5}$$

Here, $b_{xy}=-2$

$$\therefore r=\pm\sqrt{b_{xy}\cdot b_{yx}}$$
$$=\pm\sqrt{-2\times(-2/3)}$$
$$=\pm\sqrt{\frac{4}{3}}$$

$\therefore r=\pm 1.155$ (lies between $-1,0,1$)

$\therefore r=\pm 1$

$\boxed{\therefore r=-1}$ since b_{xy} and b_{yx} are negative.

Now,

$\sigma_{x^2} = 12$

$\therefore \sigma x = \pm\sqrt{12}$

$\therefore \sigma x = \pm 3.464$

Case 1 : $\sigma x = 3.464$

$\therefore b_{yx} = r\frac{\sigma y}{\sigma x}$

$\therefore \sigma y = \frac{b_{yx}\sigma x}{r}$

$\therefore \sigma y = \frac{-2/3 \times 3.464}{-1}$

$\therefore \sigma y = 2.309$

$\boxed{\therefore \sigma y^2 = 5.334}$

Case 2 : $\sigma x = -3.464$

$\therefore b_{xy} = r\frac{\sigma x}{\sigma y}$

$\therefore \sigma y = \frac{r\sigma x}{b_{xy}}$

$\therefore \sigma y = \frac{-1 \times -3.464}{-2}$

$\therefore \sigma y = -2.309$

$\boxed{\therefore \sigma y^2 = 5.334}$

Case 3 : $\sigma x = 3.464$

$\therefore b_{xy} = r\frac{\sigma x}{\sigma y}$

$\therefore \sigma y = \frac{r\sigma x}{b_{xy}}$

$\therefore \sigma y = \frac{-1 \times 3.464}{-2}$

$\therefore \sigma y = 1.732$

$\boxed{\therefore \sigma y^2 = 3}$

Case 4 : $\sigma x = -3.464$

$\therefore b_{yx} = r\frac{\sigma y}{\sigma x}$

$\therefore \sigma y = \frac{b_{yx} \times \sigma x}{r}$

$\therefore \sigma y = \frac{-2/3 \times -3.464}{-1}$

$\therefore \sigma y = -17.32$

$\boxed{\therefore \sigma y^2 = 3}$

Q. Find the estimated value of y when $x = 12.2$ from given data :-

	x	y
Mean	7.6	14.8
Std. deviation	3.6	2.5

$r = 0.99$

Soln :- Given data is :-

$x = 12.2$

$\bar{x} = 7.6$

$\bar{y} = 14.8$

$\sigma x = 3.6$

$\sigma y = 2.5$

$r = 0.99$

Regression eqn of y on x is

$\therefore (y - \bar{y}) = byx(x - \bar{x})$

where,

$$byx = r \frac{\sigma y}{\sigma x}$$

$$\therefore byx = 0.99 \times \frac{2.5}{3.6}$$

$\therefore byx = 0.6875$

$\therefore y - 14.8 = 0.686 (12.2 - 7.6)$

$\therefore y = 17.963$

* Multiple Linear Regression :-

Let $x_1, x_2, x_3 \ldots$ be the random variables

Then, linear scholarship of y as a function of $x_1, x_2, x_3, \ldots$ is,

$$y = a + b_1 x_1 + b_2 x_2 + \ldots b_k x_k + \text{error}.$$

where, $a = b_0 =$ intercept / Constant term

$b_1, b_2, \ldots b_K$ = regression slopes / regression coefficients.

Q. Suppose we have following data:

y	x_1	x_2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11

Find a multiple linear regression mode for the dataset. Also find x_1^2, x_2^2, x_1y, x_2y and x_1x_2

Sol :- Mean of y:

$$\bar{y} = \frac{\Sigma y}{n}$$

$$= \frac{140+155+159+179+192+200+212+215}{8}$$

$$= \frac{363}{2} = 181.5$$

Mean of x_1:

$$\bar{x_1} = \frac{\Sigma x_1}{n}$$

$$= \frac{60+62+67+70+71+72+75+78}{8}$$

$$= 69.375$$

Mean of x_2:

$$\bar{X}_2 = \frac{\Sigma x_2}{n}$$

$$= 18.125$$

$\therefore \Sigma y = 1452$

$\therefore \Sigma x_1 = \cancel{477}\ 555$

$\therefore \Sigma x_2 = 145$

x_1^2	x_2^2	$x_1 y$	$x_2 y$	$x_1 x_2$
3600	484	8400 ~~10890~~	3080	1320
3844	625	9610	3875	1550
4489	576	10653	3816	1608
4900	400	12530	3580	1400
5041	225	13632	2880	1065
5184	196	14400	2800	1008
5625	196	15900	2968	1050
6084	121	16770	2365	858
$\Sigma x_1^2 = 38767$	2823	101895	25364	9859

Regression sums are given by:

$$\Sigma x_1^2 = \Sigma x_1^2 - \frac{(\Sigma x_1)^2}{n}$$

$$= 38767 - \frac{(555)^2}{8}$$

$$\boxed{\therefore \Sigma x_1^2 = 263.875}$$

$$\Sigma x_2^2 = \Sigma x_2^2 - \frac{(\Sigma x_2)^2}{n}$$

$$= 2823 - \frac{(145)^2}{8}$$

$$\therefore \Sigma x_2^2 = 194.875$$

$$\Sigma x_1 y = \Sigma x_1 \times y - \frac{\Sigma x_1 \Sigma y}{n}$$

$$= 101895 - \frac{555 \times 1452}{8}$$

$$\therefore \Sigma x_1 y = 1162.5$$

$$\therefore \Sigma x_2 y = \Sigma x_2 y - \frac{\Sigma x_2 \Sigma y}{n}$$

$$= 25364 - \frac{145 \times 1452}{8}$$

$$\boxed{\therefore \Sigma x_2 y = -953.5}$$

$$\Sigma x_1 x_2 = \Sigma x_1 x_2 - \frac{\Sigma x_1 \Sigma x_2}{n}$$

$$= 9859 - \frac{555 \times 145}{8}$$

$$\therefore \Sigma x_1 x_2 = -200.375$$

Now, regression coefficients,

$$b_1 = \frac{(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 \cdot y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$= 3.148 //$$

$$b_2 = \frac{(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$= -1.656 //$$

For b_0 :-

$$b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2$$

$$\therefore b_0 = 181.5 - 3.148 \times 69.375 - (-1.650)\, 145$$

$$= -6.876 //$$

$\therefore$ The regression of eqn of line is

$$y = b_0 + b_1 x_1 + b_2 x_2$$

$$\therefore y = -6.876 + 3.148\, x_1 - 1.656\, x_2$$

Q. Find multiple linear regression eqn of y on x_1 and x_2.

y	4	6	7	9	13	15
x_1	15	12	8	6	4	3
x_2	30	24	20	14	10	4

Soln:- Mean of y :

$\bar{y} = \frac{\Sigma y}{n}$ $\Sigma y = 54$

$\boxed{\therefore \bar{y} = 9}$ $\Sigma x_1 = 48$

$\Sigma x_2 = 102$

Mean of x_1 :

$\bar{x_1} = \frac{\Sigma x_1}{n}$

$\boxed{\therefore \bar{x_1} = 8}$

Mean of x_2 :

$\bar{x_2} = \frac{\Sigma x_2}{n}$

$\boxed{\therefore \bar{x_2} = 17}$

x_1^2	x_2^2	$x_1 y$	$x_2 y$	$x_1 x_2$
225	900	60	120	450
144	576	72	144	288
64	400	56	140	160
36	196	54	126	84
16	100	52	130	40
9	16	45	60	12
$\Sigma =$ 494	2188	339	720	1034

Regression sums are given by

$\Sigma x_1^2 = \Sigma x_1^2 - \frac{(\Sigma x_1)^2}{n}$ $\Sigma x_2^2 = \Sigma x_2^2 - \frac{(\Sigma x_2)^2}{n}$

$= 483.333$ $= 2139.833$

110 454

$$\therefore \Sigma x_1 y = \Sigma X_1 y - \frac{\Sigma X_1 \Sigma y}{n}$$

$$= -117 \quad (-93)$$

$$\therefore \Sigma x_2 y = \Sigma X_2 y - \frac{\Sigma X_2 \Sigma y}{n}$$

$$= -198$$

$$\therefore \Sigma x_1 x_2 = \Sigma X_1 X_2 - \frac{\Sigma X_1 \Sigma X_2}{n}$$

$$= 218$$

Now, regression coefficients,

$$b_1 = \frac{(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$= -0.209 \quad (0.389)$$

$$b_2 = \frac{(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$= -0.071 \quad (0.623)$$

For b_0 :-

$$b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2$$

$$= \cancel{3}9 - (-0.209)8 - (-0.071)\cancel{\bar{x}_2}\ 17$$

$$= 11.879 \quad (16.479)$$

$\therefore$ The multiple linear regression eq^n is:

$$y = b_0 + b_1 X_1 + b_2 X_2$$

$$\therefore y = 11.879 - 0.209 X_1 - 0.071 X_2$$

* Partial Regression Coefficients :-

Consider variables X_1, X_2, X_3. Partial correlation between X_2 and X_1 keeping X_3 as constant,

$$\therefore r_{12\cdot3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\cdot\sqrt{1-r_{23}^2}}$$

Similarly

$$r_{13\cdot2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\cdot\sqrt{1-r_{23}^2}}$$

and $$r_{23\cdot1} = \frac{r_{23} - r_{12}\cdot r_{13}}{\sqrt{1-r_{12}^2}\cdot\sqrt{1-r_{13}^2}}$$

Thus, Partial Regression Coefficients are :-

$$b_{12\cdot3} = \frac{\sigma_1}{\sigma_2} \times \frac{r_{12} - r_{13}r_{23}}{1-r_{23}^2} \qquad b_{23\cdot1} = \frac{r_{23} - r_{12}r_{13}}{1-r_{13}^2} \times \frac{\sigma_2}{\sigma_3}$$

$$b_{13\cdot2} = \frac{\sigma_1}{\sigma_3} \times \frac{r_{13} - r_{12}r_{23}}{1-r_{23}^2}$$

* The regression equation of X_1 on X_2 and X_3 (we assume $b_0 = 0$)

$$X_1 = b_{12\cdot3}X_2 + b_{13\cdot2}X_3 \quad \text{——— (1)}$$

The regression equation of X_2 on X_1 and X_3 is

$$X_2 = b_{12\cdot3}X_1 + b_{23\cdot1}X_3$$

$$b_{21\cdot3} = \frac{\sigma_2}{\sigma_1} \times \frac{r_{21} - r_{23}r_{13}}{1-r_{13}^2}$$

$$b_{23\cdot1} = \frac{\sigma_2}{\sigma_3} \times \frac{r_{23} - r_{21}r_{31}}{1-r_{31}^2}$$

Q. Given :-

$r_{12} = 0.8$, $r_{13} = 0.6$, $r_{23} = 0.5$

$\sigma_1 = 10$, $\sigma_2 = 8$, $\sigma_3 = 5$

Determine regression eqn of

(i) X_1 on X_2 and X_3

(ii) X_2 on X_1 and X_3

Solⁿ :- The partial regression coefficients are,

$$b_{12.3} = \frac{\sigma_1}{\sigma_2} \times \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2}$$

$$= \frac{10}{8} \times \frac{0.8 - (0.6 \times 0.5)}{1 - 0.5^2}$$

$$= 0.833$$

$$b_{13.2} = \frac{\sigma_1}{\sigma_3} \times \frac{r_{13} - r_{12} r_{32}}{1 - r_{32}^2}$$

$$= \frac{10}{5} \times \frac{0.6 - 0.8 \times 0.5}{1 - 0.5^2}$$

$$= 0.533$$

∴ The regression eqn of X_1 on X_2 and X_3

$$\therefore X_1 = b_{12.3} X_2 + X_{13.2} X_3$$

$$\boxed{\therefore X_1 = 0.833 X_2 + 0.533 X_3}$$

$$b_{21.3} = \frac{\sigma_2}{\sigma_1} \times \frac{r_{21} - r_{23} r_{13}}{1 - r_{13}^2} = \frac{8}{10} \times \frac{0.8 - 0.5 \times 0.6}{1 - 0.6^2}$$

$$= 0.625$$

$$b_{23.1} = \frac{\sigma_2}{\sigma_3} \times \frac{r_{23} - r_{21} r_{31}}{1 - r_{31}^2} = \frac{8}{5} \times \frac{0.5 - (0.8 \times 0.6)}{1 - 0.6^2}$$

$$= \cancel{0.674}\ 0.05$$

∴ The regression eqn of X_2 on X_1 and X_3 is,

$$X_2 = b_{21.3} X_1 + b_{23.1} X_3$$

$$\therefore X_2 = 0.625 X_1 + \underset{0.05}{\cancel{0.674}} X_3$$

Hypothesis

Q. Given :-

* $r_{12} = 0.28$, $r_{23} = 0.49$, $r_{13} = 0.5$

$\sigma_1 = 2.7$, $\sigma_2 = 2.4$, $\sigma_3 = 2.7$

* Determine regression eq^n of,

X_3 on X_1 and X_2

Sol^n :- The regression eq^n will be,

* $X_3 = b_{31.2} X_1 + b_{32.1} X_2$

$$= \frac{r_{31} - r_{32} r_{12}}{1 - r_{12}^2} \times \frac{\sigma_3}{\sigma_1} X_1 + \frac{\sigma_3}{\sigma_2} \times \frac{r_{32} - r_{31} r_{21}}{1 - r_{21}^2} X_2$$

$$= \frac{2.7}{2.7} \times \frac{0.5 - 0.49 \times 0.28}{1 - 0.28^2} X_1 + \frac{2.7}{2.4} \times \frac{0.49 - 0.5 \times 0.28}{1 - 0.28^2} X_2$$

$\therefore X_3 = 0.394 X_1 + 0.427 X_2$

Testing of Hypothesis

* Hypothesis :-

It is a statement which is true until proven false.

* NULL Hypothesis :-

It is denoted by H_0. There is no difference between sample parameter and population parameter.

* Alternate / Alternative Hypothesis :-

It is denoted by H_1/ H_a. It is negation of null Hypothesis.
If Null hypothesis is rejected means alternate hypothesis is accepted.
This rejection or acceptance of decision depends upon z/t

$$z \text{ or } t = \frac{\text{Difference}}{\text{Standard Error}}$$

* Types of Error's :-

1) Hypothesis = True, but Test rejects it (Type I Error)
2) Hypothesis = False, but Test Accepts it (Type II Error)
3) Hypothesis = True and our Test Accepts it (correct decision)
4) Hypothesis = False and our Test Rejects it. (correct decision)

* Type I errors :-

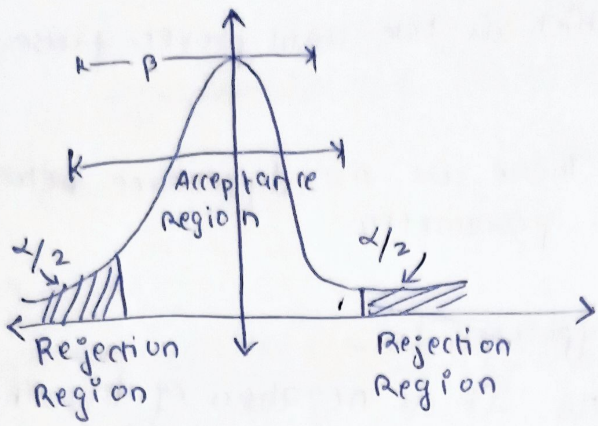
→ H_0 is rejected when it is true
→ Let, α = Probability of commiting a type-I error.

* Type II errors :-

→ H_0 is accepted when it is False.
→ Let β = Probability of committing a Type II error.

$$\boxed{\beta = 1 - \alpha}$$

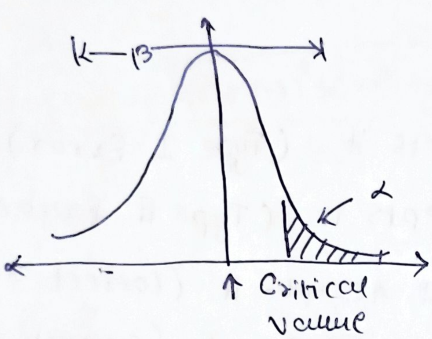
* Two Tailed test :-



$H_0 : \mu = \mu_0$

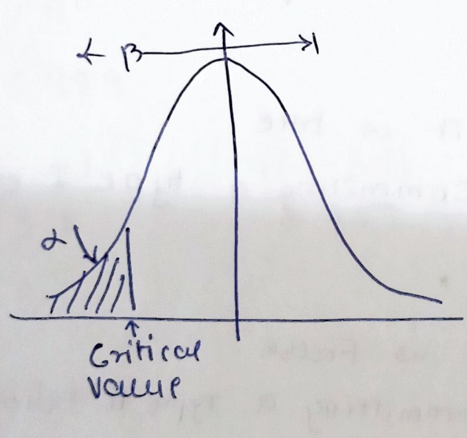
$H_1 : \mu \neq \mu_0$

* One Tailed test :-



$H_0 : \mu = \mu_0$

$H_1 : \mu > \mu_0$



$H_0 : \mu = \mu_0$

$H_1 : \mu < \mu_0$

Q. 10 individuals are choosen at random from a population and their heights are found in inches. The heights are 63, 63, 64, 65, 66, 69, 69, 70, 70, 71. Discuss the suggestion that the mean and height of population is 65 given for 9 degrees of freedom at 5% level of significance the value is 2.262.

Soln :- Given data :-

$n = 10$

$\mu = 65$

$$\bar{x} = \frac{63+63+64+65+66+69+69+70+70+71}{10}$$

$\therefore \bar{x} = 67$

H_0 = There is no difference between sample and population mean.

$$\therefore S = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{(-4)^2+(-4)^2+(-3)^2+(-2)^2+(-1)^2+(2)^2+(2)^2+(3)^2+(3)^2+(4)^2}{9}}$$

$= 3.127$

$$\therefore S.E = \frac{S}{\sqrt{n}} = \frac{3.127}{\sqrt{10}}$$

$\therefore S.E = 0.989$

$$\therefore t = \frac{|\bar{x}-\mu|}{S.E} = \frac{|67-65|}{0.989} = 2.022$$

$\therefore \nu = 9$

$\therefore t_{0.05} = 2.262$

$\therefore t < t_{0.05}$ therefore H_0 = Accepted.

$\therefore$ Mean height of population is 65.

Q8 A

Subtype I :-

Here, difference = $\bar{x} - \mu$, $t = \frac{|\bar{x} - \mu|}{S.E}$

where, $\bar{x}$ = mean of sample

μ = mean of population

Case 1 : Std. deviation is not given directly

$$s = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}}$$ ... std deviation

$$\therefore S.E = \frac{s}{\sqrt{n}}$$

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

$$\therefore S.E = \frac{s}{\sqrt{n-1}}$$

Case 2 : std. deviation is given.

$$S.E = \frac{s}{\sqrt{n-1}}$$

Also,

$d.f = \nu = n-1$ ... degree of freedom.

Confidence limits for population means are,

a) $\bar{x} \pm t_{0.05} \times S.E$ [For 95%]

b) $\bar{x} \pm t_{0.01} \times S.E$ [For 99%]

Q.1 Random sample of size 16 has a mean of ~~63~~ 53. Sum of squares of deviation observed from their mean is 135. Can this sample be regarded as drawn from the population having a mean of 56.

Soln:- Given data:-

$n = 16$

$\bar{x} = 53$

$\Sigma(x-\bar{x})^2 = 135$

$\mu = 56$

H_0 = There is no difference between mean of sample and population.

$$\therefore S = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{135}{16-1}} = 3$$

$$\therefore S.E = \frac{S}{\sqrt{n}} = \frac{3}{\sqrt{16}} = \frac{3}{4} = 0.75$$

$$\therefore t = \frac{|\bar{x}-\mu|}{S.E} = \frac{|53-56|}{0.75}$$

$\therefore t = 4$

$\therefore \nu = n-1 = 16-1 = 15$

If mean of sample < mean of population i.e $(\bar{x} < \mu)$ the test is a two tailed tested.

$t_{0.05} = 2.131$

$\because t > t_{0.05}$

$\therefore H_0$ = Rejected $\therefore$ No the sample cannot be drawn from the population having a mean of 56.

Q.2 The me
146.3
sales i
standar
success
Soln:- Given
Differ
μ
$\bar{x}$
n
Sta
H_0 =
$\therefore$ S
$\therefore t$ =
$\therefore t$
$\therefore \nu$
If
test
$\therefore t_0$
$\therefore$
$\therefore$ H
$\therefore$ Th

Page No.
Date

Q.2 The mean weekly sales of powder in a supermarket was 146.3 Kg. After a special advertising campaign. Mean weekly sales in 22 of its branches increased 153.7 Kgs with a standard deviation of 17.2 Kg. Was the advertising campaign successful.

Soln: Given data:

Difference = 153.7 − 146.3 = 7.4 Kg

$\mu = 146.3$ Kg

$\bar{x} = 153.7$ Kg

$n = 22$

Std. deviation = 17.2 Kg.

Ho = ~~Was~~ the advertising campaign is not successful.

$$\therefore S.E = \frac{S}{\sqrt{n-1}}$$

$$= \frac{17.2}{\sqrt{22-1}} = 3.753$$

$$\therefore t = \frac{\text{Difference}}{S.E} = \frac{7.4}{3.753}$$

$\therefore t = 1.972$

$\therefore \nu = 22 - 1 = 21$

If mean of sample > mean of population i.e ($\bar{x} > \mu$) the test is a one tailed test.

$\therefore t_{0.05} = 1.721$

$\therefore t > t_{0.05}$

$\therefore$ Ho = Rejected

$\therefore$ The advertising campaign is successful.

Q. A random of 10 gives mean 6.2 and S.D of 10.24. Can it be reasonably be regarded as a sample drawn from large population having mean 5.4.

Soln:- Given data :-

$n = 10$

$\bar{x} = 6.2$

$\mu = 5.4$

$s = 10.24$

H_0: There is no difference between sample & population mean.

$$S.E = \frac{s}{\sqrt{n-1}} = \frac{10.24}{\sqrt{10-1}} = 3.413$$

$$\therefore t = \frac{|\bar{x}-\mu|}{S.E} = \frac{|6.2 - 5.4|}{3.413}$$

$\therefore t = 0.234$

$\therefore \nu = n-1 = 10-1 = 9$

$\rightarrow \bar{x} > \mu$ $\therefore$ Consider one-tailed test

$\therefore t_{0.05} = 1.833$

$\therefore t < t_{0.05}$

$\therefore H_0$ = Accepted

Subtype-II : Difference between mean of two samples ($\bar{x}_1$ and $\bar{x}_2$)

Here, $\bar{x}_1$ and $\bar{x}_2$ are independent.

$$\therefore t = \frac{|\bar{x}_1 - \bar{x}_2|}{S.E}$$

where, $S.E = \sigma \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

where, $\sigma = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$

Also, $df = \nu = n_1 + n_2 - 2$

If s_1 and s_2 are not given directly, then,

$$\sigma = \sqrt{\frac{\Sigma(x_i - \bar{x}_1)^2 + \Sigma(x_i - \bar{x}_2)^2}{n_1+n_2-2}}$$

Q. Mean light of a sample of 10 electric bulbs is 1456 hrs with a S.D of 423 hrs. The second sample of 17 electric bulbs choosen from a different batch has a mean of 1280 hrs with a S.D of 398 hrs. Is there a significant difference between a mean of two samples.

Soln :- Given data :-

$n_1 = 10$ ~~hrs~~ $\quad n_2 = 17$

$s_1 = 423$ hrs $\quad s_1 = 398$ hrs

$\bar{x}_1 = 1456$ hrs $\quad \bar{x}_2 = 1280$ hrs

Ho = There is no difference between a mean of two samples.

$S.E = \sigma \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

$\sigma = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$

$= \sqrt{\frac{9(423)^2 + 16(398)^2}{10+17-2}}$

$= 407.177$

$\therefore S.E = 407.177 \times \sqrt{\frac{1}{10} + \frac{1}{17}}$

$\therefore S.E = 162.271$

$\therefore t = \frac{|1456 - 1280|}{162.271}$

$\therefore t = 1.085$

$\therefore \nu = 10 + 17 - 2 = 25$

Two tailed test as $\bar{x}_1 \neq x_2$

$\therefore t_{0.025} = \cancel{1.708}\ 2.060$

$\therefore t < t_{0.025}$

$\therefore$ Ho = Accepted.

Q. For a random samples of 10 persons fed on diet A. Increase in weight for certain period were 10, 6, 16, 17, 13, 12, 8, 14, 15, 9 kg for a another sample of 12 persons fed on deit B. Increase in weights for same period were 7, 13, 22, 17, 15, 12, 14, 18, 8, 21, 23, 10 kgs. Test whether two diets differ significantly as regards to increase in weights.

Solnº:- Given data :-

$n_1 = 10$ $\quad \bar{x}_1 = 12$

$n_2 = 12$ $\quad \bar{x}_2 = 15$

x_1	$(x_1-\bar{x}_1)^2$	x_2	$(x_2-\bar{x}_2)^2$
		7	64
10	4	13	4
6	36	22	49
16	16	17	4
17	25	15	0
13	1	12	9
12	0	14	1
8	16	18	9
14	4	8	49
15	9	21	36
9	9	23	64
		10	25
	$\Sigma = 120$		$\Sigma = 314$

$$\therefore \sigma = \sqrt{\frac{\Sigma(x_1-\bar{x}_1)^2 + \Sigma(x_2-\bar{x}_2)^2}{n_1+n_2-2}} = \sqrt{\frac{120+314}{10+12-2}}$$

$$\therefore \sigma = 4.658$$

$$\therefore S.E = \sigma \times \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = 4.658 \times \sqrt{\frac{1}{10}+\frac{1}{12}} = 1.994$$

$$\therefore t = \frac{|\bar{x}_1 - \bar{x}_2|}{S.E} = \frac{|12-15|}{1.994} = 1.504$$

$\nu = 10+12-2 = 20$

As $\bar{x}_1 \neq \bar{x}_2$ it is two tailed test

$\therefore t_{0.025} = 2.086$ Column 6

$\therefore t < t_{0.025}$ $\qquad$ One tailed test (columns)

$\therefore H_0 =$ Accepted.

Subtype I(c) :- Testing the significance of the difference between means of two samples (Dependent)

$$t = \frac{|\bar{d}|}{S.E}$$

where, $S.E = \frac{S}{\sqrt{n}}$

where, S = std. deviation of differences

and $S = \sqrt{\frac{\Sigma(d-\bar{d})^2}{n-1}}$

where, $d = x_2 - x_1$ = difference between two samples

and $\bar{d}$ = mean of difference.

Also, degree of freedom is,

$\nu = n-1$

Q. To verify whether a course in mathematics improve performance or not, two test were given to 12 participants One before and other after the course. Marks of the test before the course were 44, 40, 61, 52, 32, ~~44~~ 44, 70, 41, 67, 72, 53 and 72. Marks of the test after the course were 53, 38, 69, 57, 46, 39, 73, 48, 73, 74, 60 and 70. Determine whether the course was useful or not.

Soln :- Given data :

$n = 10$

H_0 = Course is not useful. $\bar{d} = 4.33$

x_1	x_2	d	$d-\bar{d}$	$(d-\bar{d})^2$
44	53	9	4.667	21.781
40	38	-2	-6.333	40.107
61	69	8	3.667	13.447
52	57	5	0.667	0.445
32	46	14	9.667	93.451

x_1	x_2	d	$d-\bar{d}$	$(d-\bar{d})^2$
44	39	-5	9.333	87.105
70	73	3	1.333	1.777
41	48	7	2.667	7.113
67	73	6	1.667	2.279
72	74	2	-2.333	5.443
53	60	7	2.667	7.113
72	70	-2	-6.333	40.107

$$S = \sqrt{\frac{\Sigma(d-\bar{d})^2}{n-1}}$$

$$\therefore S = \sqrt{\frac{320.657}{12-1}} = 5.4$$

$$\therefore S.E = \frac{S}{\sqrt{n}} = \frac{5.4}{\sqrt{12}} = 1.558$$

$$\therefore t = \frac{|\bar{d}|}{S.E} = \frac{4.33}{1.558} = 2.779$$

$$\upsilon = n-1 = 12-1 = 11$$

$\therefore$ It is the one sided test since the performance improves after course.

$\therefore t_{0.05} = 1.796$

$\therefore t > t_{0.05}$

$\therefore H_0 =$ Rejected.

$\therefore$ The course is useful.

Q. A drug is given to 10 patients and the increase in their BP was recorded as 3, 6, -2, 4, -3, 4, 6, 0, 0, 2. Is it reasonable to believe that drug has no effect on change in B.P

solⁿ:-

d	$d-\bar{d}$	$(d-\bar{d})^2$
3	1	1
6	4	16
-2	-4	+16
4	2	4
-3	-5	25
4	2	4
6	4	16
0	-2	4
0	-2	4
2	0	0
$\Sigma d = 20$	$\Sigma(d-\bar{d}) = 0$	$\Sigma(d-\bar{d})^2 = 90$

$\bar{d} = \frac{\Sigma d}{n} = \frac{20}{10} = 2$

H_0 = Drug has effect on change in B.P

$\therefore s = \sqrt{\frac{\Sigma(d-\bar{d})^2}{n-1}} = \sqrt{\frac{90}{9}} = \sqrt{10}$

$\therefore s = 3.162$

$\therefore S.E = \frac{s}{\sqrt{n}} = \frac{3.162}{\sqrt{10}} = 1$

$\therefore t = \frac{|\bar{d}|}{S.E} = \frac{2}{1} = 2$

$\therefore \nu = n-1 = 10-1 = 9$

It is the two sided test.

$\therefore t_{0.05} = 2.262$

$\therefore t < t_{0.05}$

$\therefore H_0$ = Accepted.

$\therefore$ It is reasonable to say that drug has effect on change in B.P