– Aditi Chaurasia (17BCS001)

*1) Descriptive Statistics

They are used to describe the basic features of the data & provide simple summaries about the sample and the measures. It can either represent entire or a sample of a population & present quantitative descriptions. It consists of two basic categories of measures: measures of central tendency and measures of dispersion. They help in understanding the features of a specific data set.

- Measure of central tendency:- It describe the center position of a distribution for a data set. The three types of estimates of central tendency are mean (i.e. average of all values), median (i.e. middle value of the set of values after listing all values in any numerical order (either ascending or descending)) and mode (most frequently occured value in the whole data set).
- Measure of dispersion:- It refers to the spread of the values around the central tendency. In otherwords, it describes how the data is distributed within the set. Range (i.e. difference between extreme values, viz. minimum & maximum value), standard deviation, mean deviation, quartile deviation. and variance are measures of variability.

→ **Application**

Ques) Following are test score values: 15, 20, 21, 20, 36, 15, 25, 15
Do the descriptive analysis of given data.

Soln. mean $= \frac{\sum_{i=1}^{n} x_i}{n}$, here, $n = 8$

$\therefore$ mean, $m = \frac{15+20+21+20+36+15+25+15}{8} = \frac{167}{8} = 20.875$

median, $m_d = 20$

since data in ascending order: 15, 15, 15, 20, 20, 21, 25, 36

$n = 8$ which is even $\therefore$ median $= \frac{\text{val}\left[\left(\frac{n^{th}}{2}\right) + \left(\frac{n^{th}}{2} + 1\right)\right]}{2} = \frac{20+20}{2}$

$= 20$

mode, $m_o = 15$ since it has occured maximum number of times i.e. 8 times.

Range, $R = 36 - 15 = 21$

x_i	$x_i - \bar{x}$ (mean deviation)	$(x_i - \bar{x})^2$
15	-5.875	34.515
20	-0.875	0.765
21	0.125	0.015
20	-0.875	0.765
36	15.125	228.765
15	-5.875	34.515
25	4.125	17.015
15	-5.875	34.515

$$\sum_{i=1}^{n=8} (x_i - \bar{x})^2 = 350.875$$

$$\therefore \text{variance} = \frac{1}{n-1} \sum_{i=1}^{n=8} (x - \bar{x}_i)^2 = \frac{1}{7}(350.875) = 50.125$$

$$\text{and std. deviation} = \sqrt{\text{variance}} = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{8} (x_i - \bar{x})^2} = 7.079$$

and Quartiles, Q_3 (third quartile) and Q_1 (first quartile).

given data (in ascending order): 15, 15, 15, 20, 20, 21, 25, 36

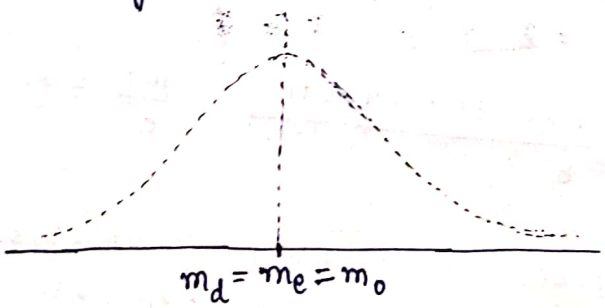
$\therefore Q_1 = (15+15)/2 = 15$

$Q_2 = (20+20)/2 = 20 = \text{median}$

$Q_3 = (21+25)/2 = 23$

Therefore, IQR (i.e. interquartile range) is $Q_3 - Q_1 = 23 - 15 = 8$

Note: • For symmetrical distribution of data, median = mean = mode



- Central tendency says that there is one number that best summarizes the entire set of measurements.
- Standard deviation is the measurement of average distance b/w each value and mean. A low s.d. indicates the data points tend to be close to the mean of the dataset & higher value of it indicates that data points are spread out over a wide range of values.

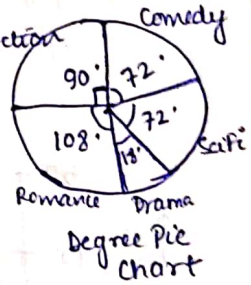
Graphical methods:

Graphic presentation is a simpler way of understanding the features and drawing comparisons. Graphical methods are analytical tools and more understandable to readers. Graphs are used to present time series (eg. histogram, line graph, etc.) and frequency distributions. It allows to present statistical data in an attractive manner & shows trends, & fluctuations of the data at a glance & also it ~~is~~ is much easier to find mean, median & mode values of the data. Following are graphical methods.

- Pie Chart: It is used to do univariate analysis. It is a circular, statistic graphic which is divided into slices to illustrate numerical proportion. Area is proportional to the quantity it represents. It is of two types: one based on percentage and other based on degree.

Ques.) Following is the data given below.

	Comedy	Action	Romance	Drama	Scifi	Total
	4	5	6	1	4	20
%age	$(4/20)\times100$ = 20%	$(5/20)\times100$ = 25%	$(6/20)\times100$ = 30%	$(1/20)\times100$ = 5%	$(4/20)\times100$ = 20%	100%
degree	$\frac{4}{20}\times360^\circ$ = 72°	$\frac{5}{20}\times360^\circ$ = 90°	$\frac{6}{20}\times360$ = 108°	$\frac{1}{20}\times360^\circ$ = 18°	$\frac{4}{20}\times360^\circ$ = 72°	360°



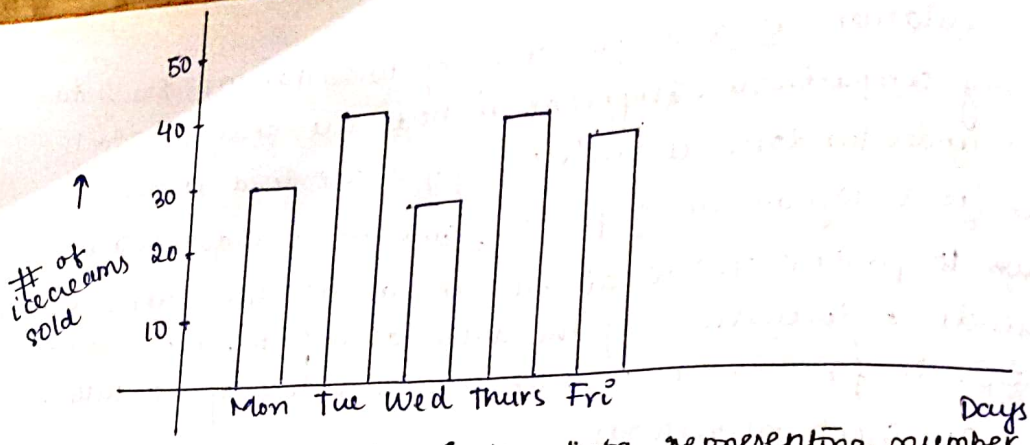
Degree Pie Chart

From the pie chart, it is easy to tell which movie genres are most liked & which are least liked, at a glance.

- Bar Plot: It is also called bar graph (or bar chart). It is a graphical display of data using bars of different heights. Unlike pie chart, it is used to compare different categories. Generally, there are two types of bar graph: (1) Grouped Bar graph and (2) Stacked Bar graph
It is used to show changes in set categories over time, location or sector.

Ques) 150 icecreams are sold in 5 working days. Following is the data given.

Day	Mon	Tue	Wed	Thurs	Fri
No. of icecreams sold	30	40	26	38	34



Ques) (comparison). Following is the data representing number of girls & boys in Class 5th to Class 9th. Using graphical representation, compare girls and boys strength.

Class	5th	6th	7th	8th	9th
# of girls	30	35	20	25	40
# of boys	20	40	25	35	20

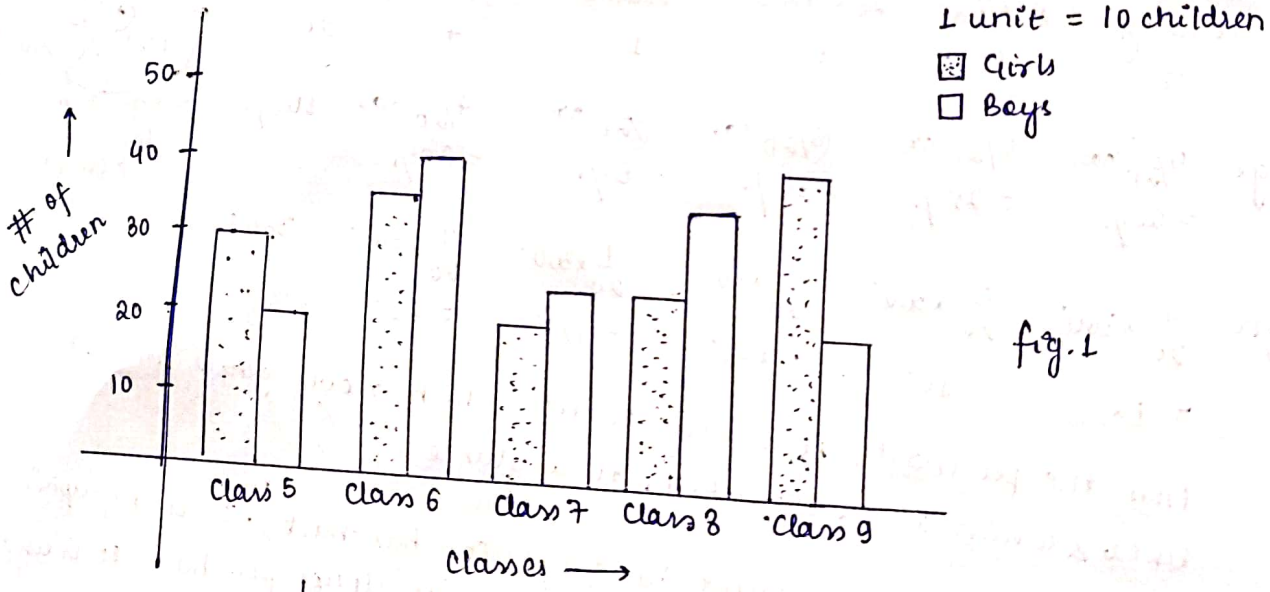
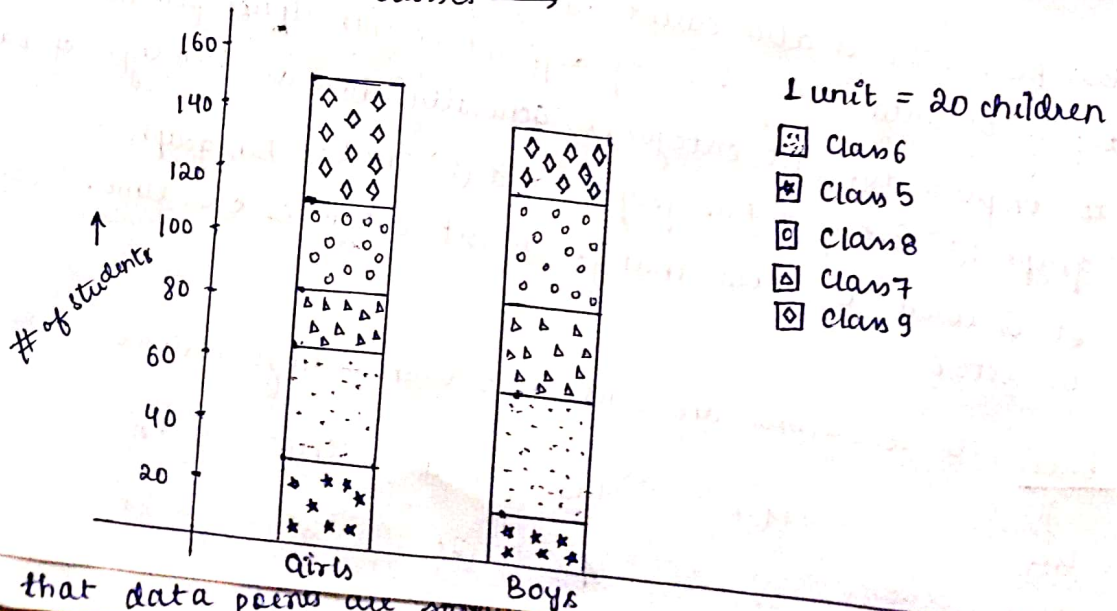


fig. 1

fig. 2



that data points are spread

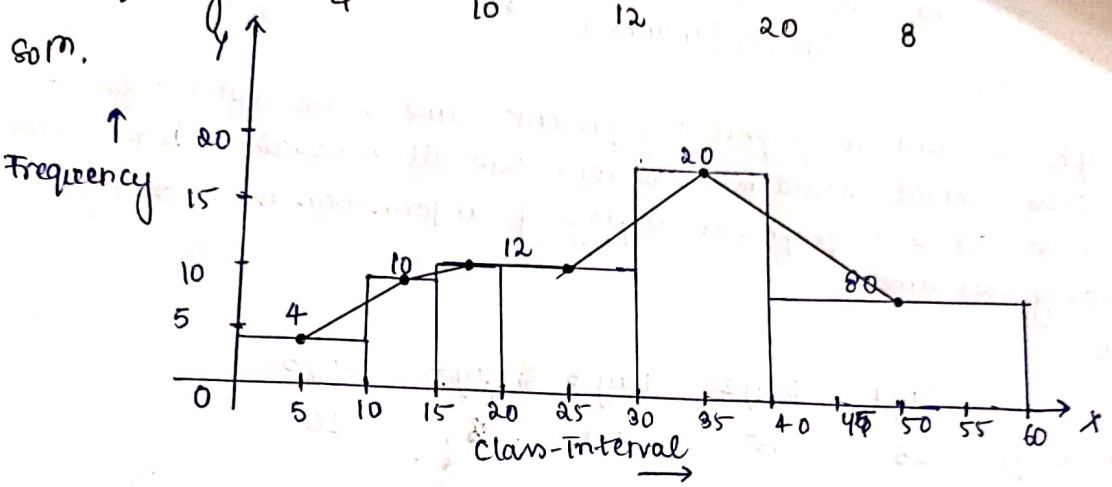
Histogram and Frequency Polygon (or Histograph)

Both are used to represent frequency distribution of given dataset & are used for continuous data. A histogram is a graph of a grouped frequency distribution & height of each bar represents frequency density of the class. It is used to compare two variables unlike histogram. (freq. polygon)

Ques). Plot a histogram & histograph for following data:

Class Interval	0-10	10-15	15-30	30-40	40-60
Frequency	4	10	12	20	8

Sol^n.



black line represents the frequency polygon for given continuous data.

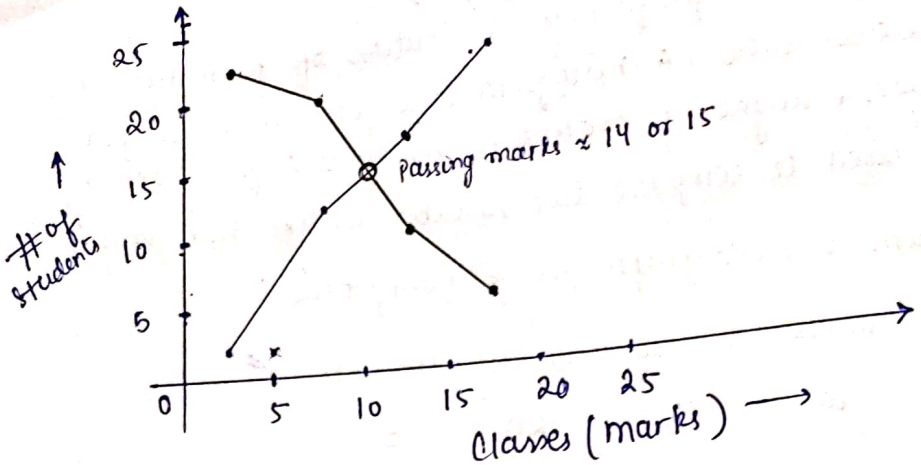
4) Cumulative Frequency Curve or Ogive

It is the graphical representation of a cumulative frequency distribution. Ogives are of two types: less than ogive & more than ogive. This type of graph is generally used to find cut-offs points for various competitive examinations.

Note: Intersection points of both types of ogive gives median value of the dataset.

Ques) Find the passing mark for students from the following data.

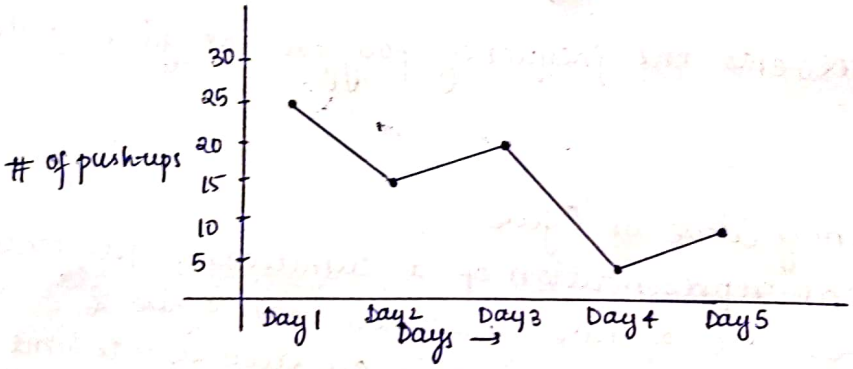
marks	0-5	5-10	10-15	15-20
# of students	2	10	5	5
Cum. frequency	2	12	17	22 (in ascending order)
cum freq (D.O.) descending order	22	20	10	5



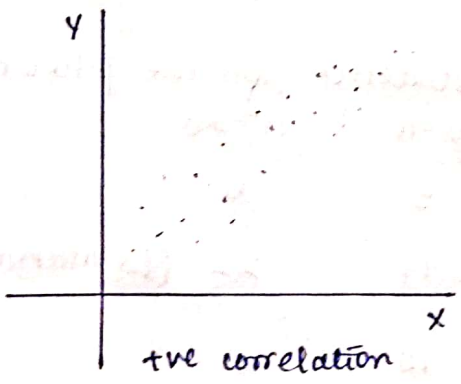
5) line Graph is used to graphically present time series distribution of data. Class interval should be continuous but all intervals need not have same width. It is a graphical display of information that changes continuously over time.

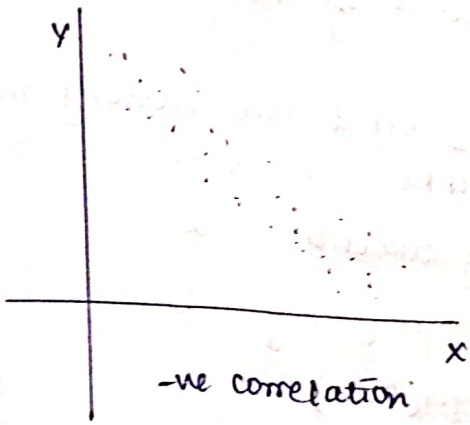
example.

Day	Day 1	Day 2	Day 3	Day 4	Day 5
# of push-ups.	25	15	20	5	10



6) Scatter Plot (or Dot Chart) enable us to find the nature of the relationship between the variables. If plotted points are scattered a lot then the relationship between the variable is lesser (or weaker).





Other plots are stem-leaf plot, box-plot (useful in finding outliers), etc.

) Coefficient of variation:

It is a measure of relative variability of a frequency distribution or a probability distribution. It is defined as the ratio of the standard deviation to the mean. It is also known as relative std. deviation, and is used to compare variability between different measures. The lower its value, more consistent the data will be.

CV for a population $= \frac{\sigma}{\mu} \times 100\ \%$.

CV for a sample $= \frac{s}{\bar{x}} \times 100\ \%$.

Ques) Which player is better: Player A or Player B.

	Player A	Player B
mean	50.1	45.8
std. dev.	11.2	12.9

Soln. $(CV)_A = \frac{\text{std. dev.}}{\text{mean}} \times 100 = \frac{11.2}{50.1} \times 100 = 22.35\ \%$.

and $(CV)_B = \frac{s}{\bar{x}} \times 100 = \frac{12.9}{45.8} \times 100 = 28.166\ \%$.

Clearly, $(CV)_A < (CV)_B \Rightarrow$ Player A is more consistent than Player B. Hence, Player A is a better player.

*4) Covariance:

It indicates the direction of the linear relationship between variables. It is a measure of the joint variability of two random variables i.e. It measures the degree to which two variables are linearly associated. In other words, it measures how changes in one variable is associated with changes in other variable. It can range from $-\infty$ to $+\infty$.

- If two random variables are independent, the covariance will be zero.
- If variables are inversely related (that is, always moves in opposite direction then covariance will be negative.
- If greater values of one variable corresponds with greater value of second variable & then same holds for lesser values then covariance will be positive.

Covariance for a population, $cov(x,y) = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{N}$

Covariance for a sample, $Cov(X,Y) = \frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{(N-1)}$

here, X & Y are random variables & N is total no. of observations.

$\bar{x}$ and $\bar{y}$ are expected value (i.e. mean) of the random variables, X & Y, respectively.

Ques) Calculate the covariance for the following dataset.

x: 2.1, 2.5, 3.6, 4.0

y: 8, 10, 12, 14

Soln.

x	$-(\bar{x}-x_i)$	y	$(\bar{y}-y_i)$	$(\bar{x}-x_i)(\bar{y}-y_i)$
2.1	-1	8	-3	3
2.5	-0.6	10	-1	0.6
3.6	0.5	12	+1	0.5
4.0	0.9	14	+3	2.7
$\sum_{i=1}^{4} x_i = 12.2$		$\sum_{i=1}^{4} y_i = 44$		$\sum_{i=1}^{4}(\bar{x}-x_i)(\bar{y}-y_i) = 6.8$

$\therefore \bar{x} = \frac{\sum_{i=1}^{4} x_i}{n} = \frac{12.2}{4} \approx 3.1$ and $\bar{y} = \frac{\sum_{i=1}^{4} y_i}{n} = \frac{44}{4} = 11$

$\therefore cov(x,y) = \frac{\sum_{i=1}^{4}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$

$= \frac{6.8}{3} = 2.267 > 0$

$\therefore$ X and Y are positively related.

Note: The value of covariance is affected by the change in scale of the variables.

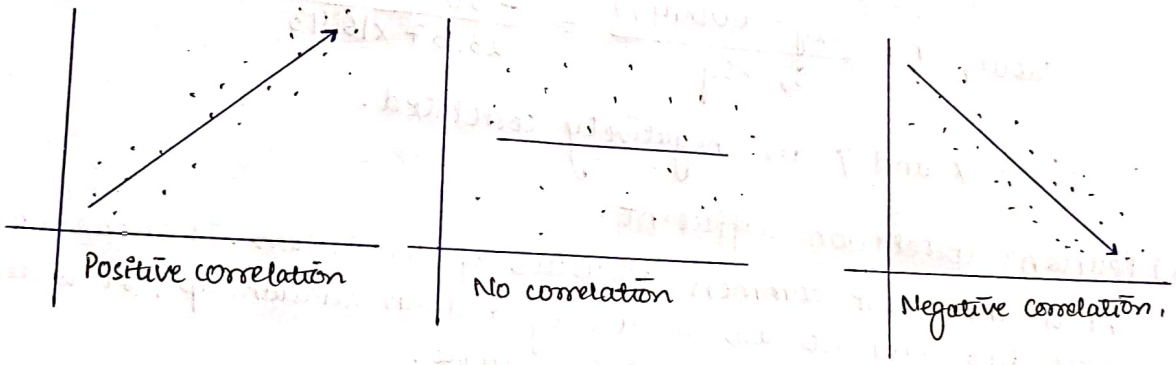
*5) Correlation and Correlation coefficient.

Unlike covariance, correlation measures both the strength and direction of the linear relationship between two

coefficient of two variable by dividing the covariance of those variables by the product of the standard deviations of the same values.) Its value ranges from -1 to +1 (since they are standardized values). It is dimensionless and its value is not influenced by the change in scale of the values unlike covariance.

Correlation coefficients are used to measure how strong a relationship is between two variables. The two common types of correlation coefficient are Pearson's correlation (or curved correlation coefficient), denoted by 'r', and Spearman's correlation (or rank correlation coefficient), denoted by 'R'.

- $r = 1$ indicates a strong positive relationship (i.e. perfectly correlated)
- $r = -1$ indicates a strong negative relationship (i.e. inversely related)
- $r = 0$ indicates no relationship at all (nonsense correlation)
- $r \geq 0.50$ indicates significant correlation



→ Sample correlation coefficient, $r = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2 \sum_{i=1}^{N} (y_i - \bar{y})^2}}$

where S_{xy} is sample covariance of x & y, S_x is sample std. deviation of x & S_y is sample std. deviation of y.

→ Population correlation coefficient, $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

where σ_{xy} is covariance of x and y populations & σ_x & σ_y are std. deviation of x and y, respectively.

Qeee) Compute the correlation coefficient for the following data.

X :	20	50	70	40	10
Y :	40	20	10	20	50

Soln. $\bar{X} = 190/5 = 38$ and $\bar{Y} = 140/5 = 28$

$X - \bar{X}$	$Y - \bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$
-18	12	216
12	-8	-960
32	-18	-576
2	-8	-16
28	22	-616

$$\sum_{i=1}^{5} (x_i - \bar{x})(y_i - \bar{y}) = -1520$$

and $S_x = \sqrt{\frac{2280}{4}} = 23.87$, $S_y = \sqrt{\frac{1080}{4}} = 16.43$

and $\text{cov}(x, y) = \frac{-1520}{4} = -380$

Now, $r = \frac{S_{xy} (= \text{cov}(x,y))}{S_x \cdot S_y} = \frac{-380}{23.87 \times 16.43} = -0.9689$

$\therefore$ x and y are negatively correlated.

*6) Pearson's correlation coefficient.

It is the most common measure of correlation. It shows how well two data sets are related by a linear relationship. It is denoted by ρ (rho) for population & 'r' for sample.

$$r = \frac{S_{xy}}{S_x \cdot S_y}$$

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

Limitations: Curved coefficient correlation will not give any information about the slope of line (showing linear relationship.). It will only tell us whether there is a relationship between the two data sets.

* $r = \frac{S_{xy}}{S_x \cdot S_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$ can be also written as

* $r = \frac{n\left(\sum_{i=1}^{n} xy\right) - \left(\sum_{i=1}^{n} x\right)\left(\sum_{i=1}^{n} y\right)}{\sqrt{\left[n\sum_{i=1}^{n} x^2 - \left(\sum_{i=1}^{n} x\right)^2\right]\left[n\sum_{i=1}^{n} y^2 - \left(\sum_{i=1}^{n} y\right)^2\right]}}$

The previous question is an application of curved correlation coefficient.

Ques) Find the value of correlation coefficient from the following data.

Subject	Age (x)	Glucose level (Y)	x^2	y^2	xy
1	43	99	1849	9801	4257
2	21	65	441	4225	1365
3	25	79	625	6241	1975
4	42	75	1764	5625	3150
5	57	87	3249	7569	4959
6	59	81	3481	6561	4779
	$\Sigma x = 247$	$\Sigma y = 486$	$\Sigma x^2 = 11409$	$\Sigma y^2 = 40022$	$\Sigma xy = 20485$

$$\therefore r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

here $n = 6$

$$\therefore r = \frac{6 \times 20485 - (247 \times 486)}{\sqrt{[6 \times 11409 - (247)^2][6 \times 40022 - (486)^2]}}$$

$$= \frac{122910 - 120042}{\sqrt{7445 \times 3936}}$$

$$= \frac{2868}{5413.27} = 0.5298 \geq 0.50$$

$\therefore$ Variables, Age and Glucose level are significantly correlated (and also has positive correlation).

*7) Rank Correlation Coefficient.

It is denoted by 'R' and indicates association between ranks. That's why it is called Spearman's rank order correlation.

- $R = +1$ indicates a perfect positive association of ranks.
- $R = 0$ indicates no association between ranks.
- $R = -1$ indicates a perfect negative association of ranks.

Ques) Calculate rank correlation coefficient for the following table.

Then the regression line of y on x is

$$y = a + bx$$

here, y is dependent on x, and 'a' & 'b' are parameters.

There are two methods using which we can calculate parameters value.

(1) Least square method.

It uses two normal equations which are given below.

$$\Sigma y_i = na + b\Sigma x_i$$

$$\Sigma x_i y_i = b\Sigma x_i^2 + a\Sigma x_i$$

On solving these two eqn(s) we will get 'a' & 'b' values.

(2) Deviation method

The standard form of the regression equation of y on x is

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

where b_{yx} is regression coefficient for y on x regression line

and $b_{yx} = \frac{r\sigma_y}{\sigma_x}$, r = correlation coefficient of x & y,

σ_x = std. deviation of x & σ_y = std. deviation of y.

Similarly, for case 2: when x is dependent variable & Y is independent variable

Then, the regression line of x on y is

$$x = a + by$$

here, x is dependent on y & 'a' and 'b' are parameters.

(1) Least square method.

$$\Sigma x_i = na + b\Sigma y_i$$

$$\Sigma x_i y_i = a\Sigma y_i + b\Sigma y_i^2$$

On solving these 2 equations, we will get 'a' & 'b' values. Put these values in eqn $x = a + by$ & get equation of regression line x on y.

(2) Deviation method.

The standard form of the regression equation of x on y is

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

where b_{xy} is regression coefficient for line x on y

and $b_{xy} = \frac{r \cdot \sigma_x}{\sigma_y}$, r = correlation coefficient,

σ_x : Std. deviation of x and σ_y : Std. deviation of y.

- Properties of regression lines.

1. The value of the regression coefficient does not change.
2. There are two lines of regression. Both these lines are known to intersect at a specific point $[\bar{x}, \bar{y}]$ i.e. (sample mean of x, sample mean of y). Hence, it is the solution for both the equations of x & y.
3. The correlation coefficient between the two variables x & y is the geometric mean of both the coefficients i.e. $r = \pm\sqrt{b_{yx} + b_{xy}}$

Ques) Find the least square regression line $\hat{y} = a + bx$ for the following set of data $\{(-1,0), (0,2), (1,4), (2,5)\}$. Also find regression line $x = a + by$ using deviation method.

Soln.

x	y	xy	x^2
-1	0	0	1
0	2	0	0
1	4	4	1
2	5	10	4
$\Sigma x = 2$	$\Sigma y = 11$	$\Sigma xy = 14$	$\Sigma x^2 = 6$

we have, $\Sigma y = na + b\Sigma x$ —(1)

$\Sigma xy = a\Sigma x + b\Sigma x^2$ —(2)

$\therefore$ Eqn (1) and (2) becomes

$11 = 4a + 2b$ —(3)

and $14 = 2a + 6b \Rightarrow 7 = a + 3b \Rightarrow a = 7 - 3b$ —(4)

Put value of 'a' from (4) in eqn (3), we get

$11 = 4(7 - 3b) + 2b$

$\Rightarrow 11 = 28 - 12b + 2b \Rightarrow 10b = 17 \Rightarrow b = 1.7$

$\therefore a = 7 - 3 \times 1.7 = 7 - 5.1 = 1.9$

$\therefore$ reg line y on x is $y = 1.9 - 1.7x$

for reg. line x on y,

X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$
-1	0	-1.5	-2.75	4.125
0	2	-0.5	-0.75	0.375
1	4	0.5	1.25	0.625
2	5	1.5	2.25	3.375
$\Sigma X = 2$	$\Sigma Y = 11$			$\Sigma(x-\bar{x})(y-\bar{y}) = 8.5$

mean of x, $\bar{x} = 2/4 = 0.5$

and $\bar{Y} = 11/4 = 2.75$

$$\sigma_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{5}{3}} \quad \text{and} \quad \sigma_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}} = \sqrt{\frac{14.75}{3}}$$

$$\text{and } r = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} = \frac{8.5}{\sqrt{5 \times 14.75}} = \frac{8.5}{8.587} = 0.9898$$

$$\text{Now, } b_{xy} = \frac{r \cdot \sigma_x}{\sigma_y} = \frac{0.9898 \times \sqrt{5}}{\sqrt{14.75}}$$

$$= \frac{0.9898}{\sqrt{2.95}} = \frac{0.9898}{1.7175} = 0.5763$$

$\therefore$ reg. line is $(y-\bar{y})\, b_{xy} = x - \bar{x}$

i.e. $(x - 0.5) = 0.5763\,(y - 2.75)$

$x = 0.5763\,y - 1.584 + 0.5$

$x = 0.57y - 1.084$