

# **CORRELATION ANALYSIS**

## **INSTRUCTOR: SANTOSH CHHATKULI**



Two important concept arise in the study of relationship between variables: **Correlation** and **Regression**.

**Correlation** relates to study of the relationship between two quantitatively measured variables. The relationship between two qualitatively measured variables is rather called Association.

**Regression** analysis related to study of the nature of relationship between variables. The ultimate objective of the regression analysis is to develop estimating equation which is used to predict the value of one variable with the knowledge of other variable(s).

# Correlation Study

**Correlation** is a measure of the linear relationship between two variables measured on a numerical scale. If the relationship between two variables is described by a straight line, then it is called **linear relationship**. If the relationship between variables is described by a curve line, then it is called **non-linear** or **curvilinear relationship**.

## Types of Correlation

1. **Positive Correlation:** If two variables change in the same direction (i.e. if value one variable increases the value of other variable also increases, or if one decreases, the other also decreases), then this is called a positive correlation. For example : Height and Weight.
2. **Negative Correlation:** If two variables change in the opposite direction ( i.e. if value one variable increases, the value of other variable decreases and vice versa), then the correlation is called a negative correlation. For example : Per capita income of state and Number of tuberculosis patients, GPA of students and TV hours.
3. **Zero Correlation:** It suggests no relationship between two variables. For example: Pulse rate and Height of the persons, Grade and Height.

# Methods of studying correlation

There are two method of studying correlation

1. Graphical method: Scatter Diagram
2. Mathematical method:
  - Karl Pearson's correlation coefficient
  - Spearman's correlation coefficient


# Graphical method

A first step that is usually useful in studying the relationship between two quantitatively measured variables is to prepare a **scatter diagram**. It is an exploratory tool for analyzing relationships between two variables.

A Scatter Diagram graphs pairs of numerical data, with one variable on each axis (in causal relationship, the independent variable is plotted on the horizontal axis and dependent variable is plotted on the vertical axis) to look for a relationship between them.

The data (values of two variables) is displayed as a collection of points or dots, each dot in the graph represents single pair of data.

If the variables are correlated, the points will fall along a line or curve. The better the correlation, the tighter the points will hug the line.

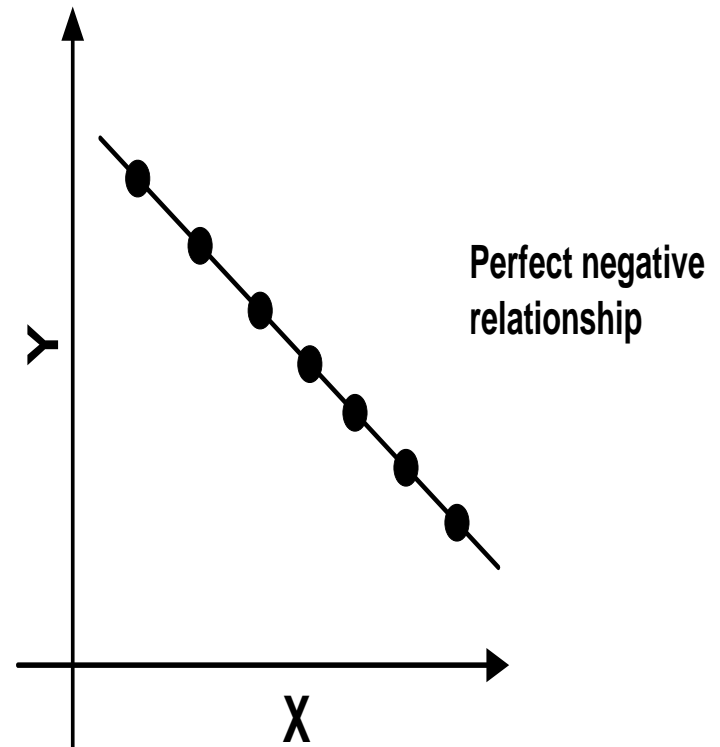
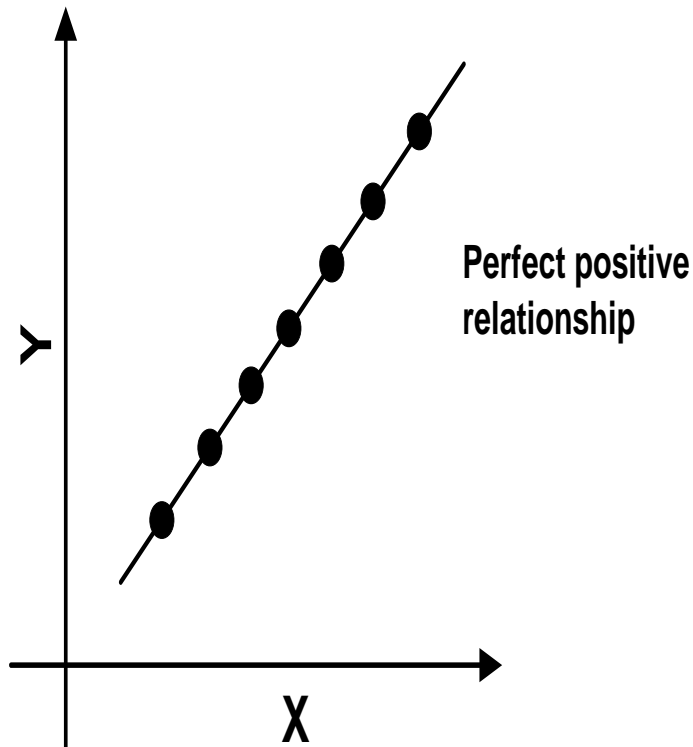


A scatter diagram is a graphical tool which can give us three types of information:

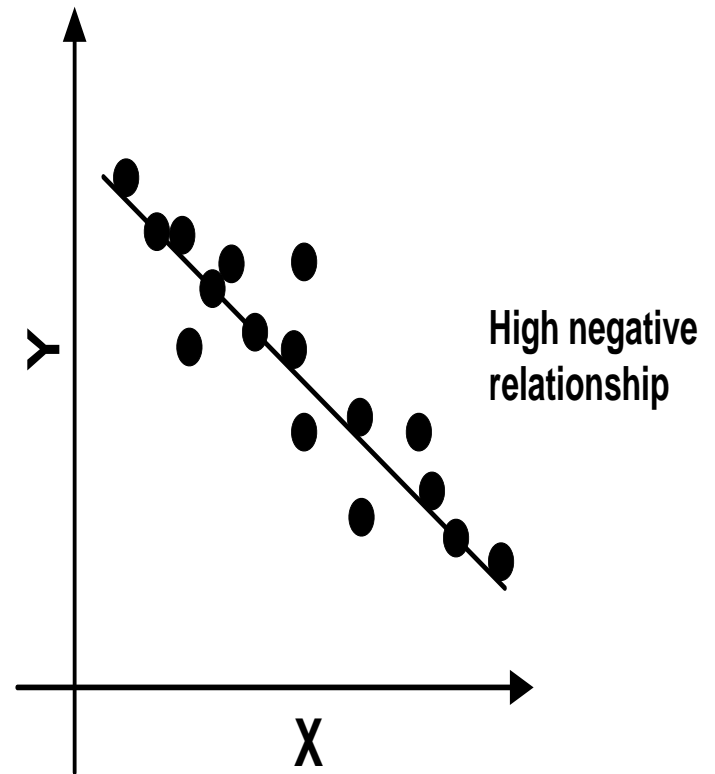
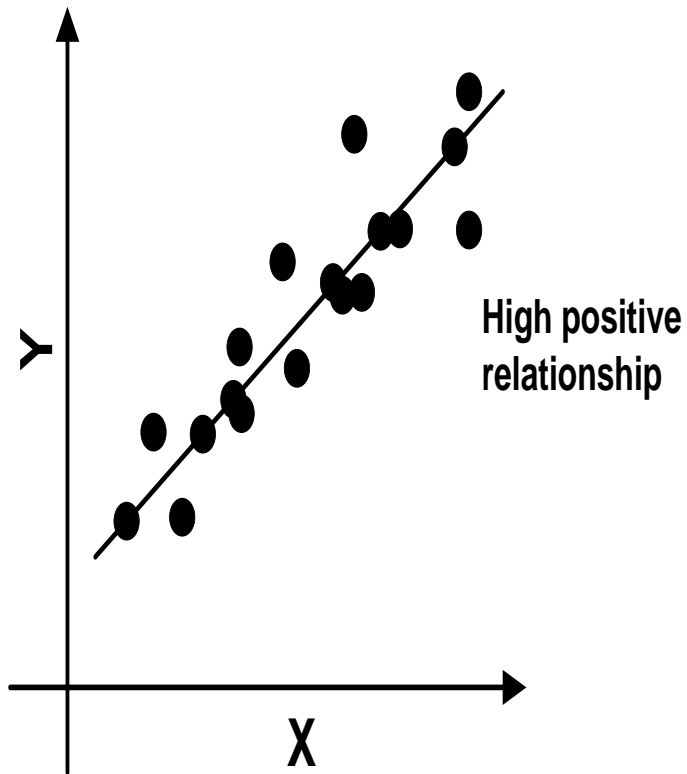
- Visually, we can look for patterns that indicate that the variables are related.
- If the variables are related, we can see what kind of line, or estimating equation, describes this relationship.
- We can trace presence of outliers or extreme values.

# Probable forms of Scatter Diagram

## Perfect Relationship

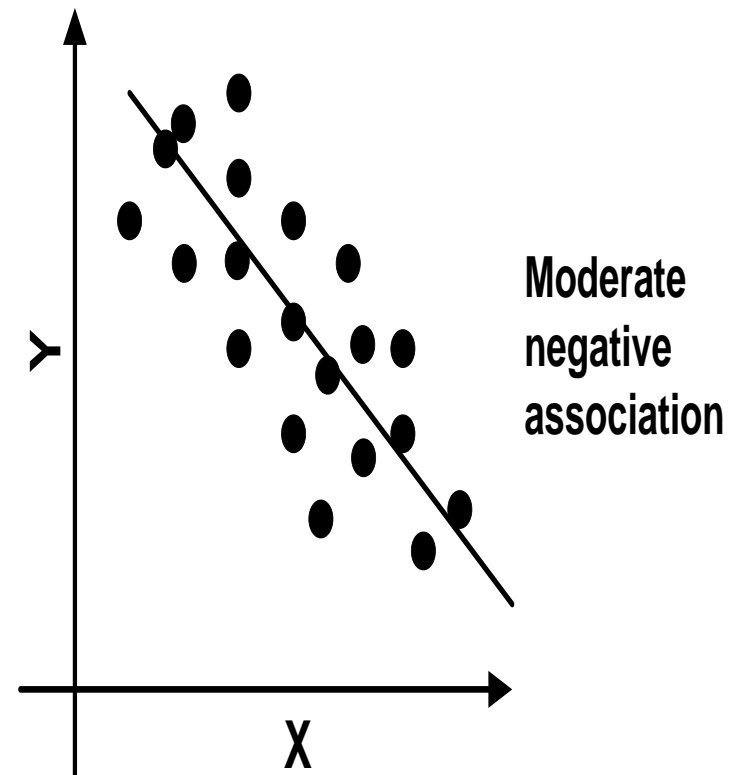
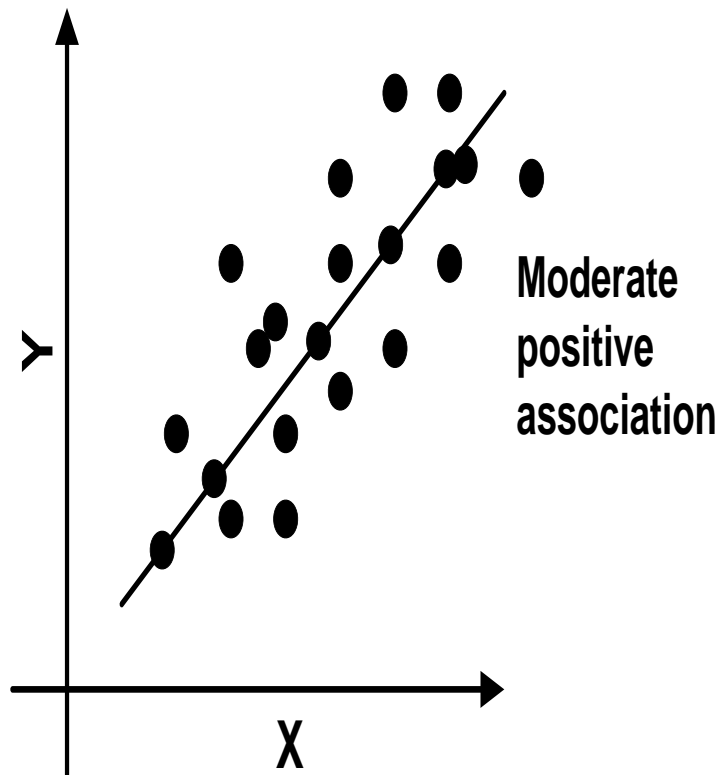


# Strong Relationship

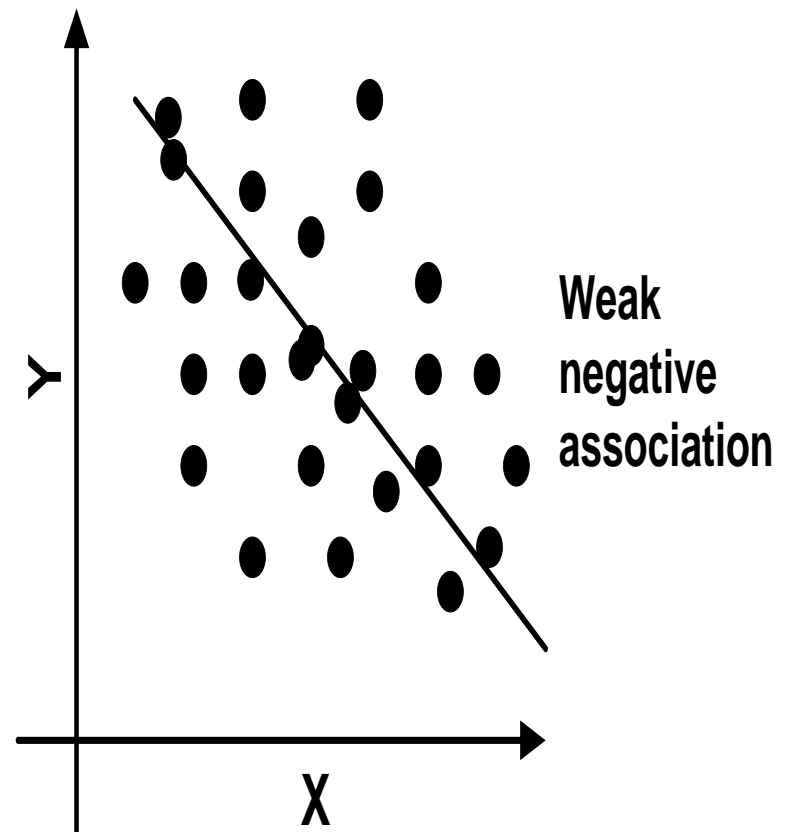
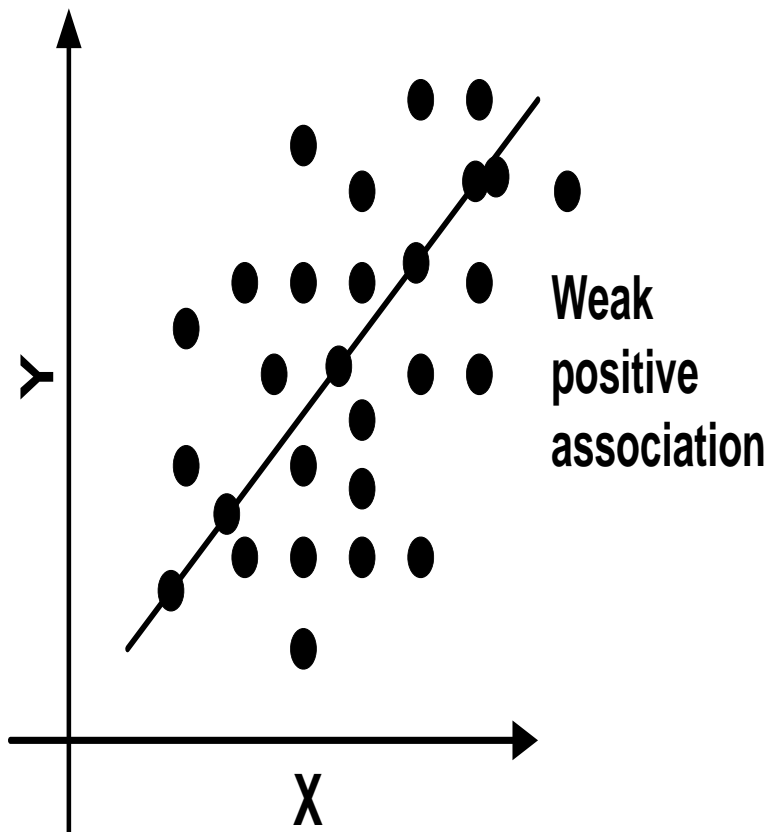




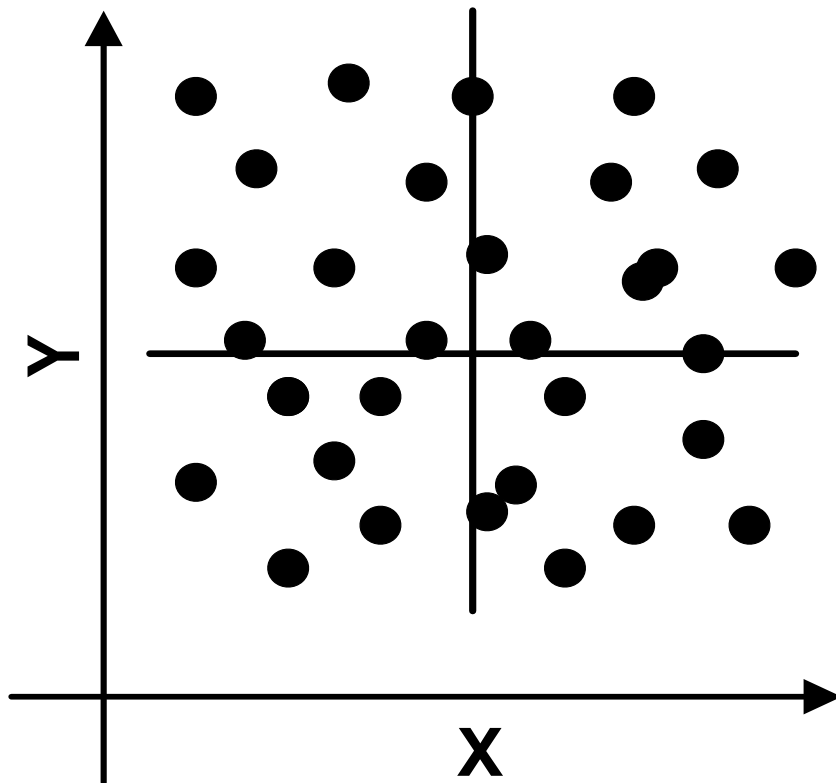
# Moderate Relationship



# Weak Relationship



# No Relationship



**No relationship**

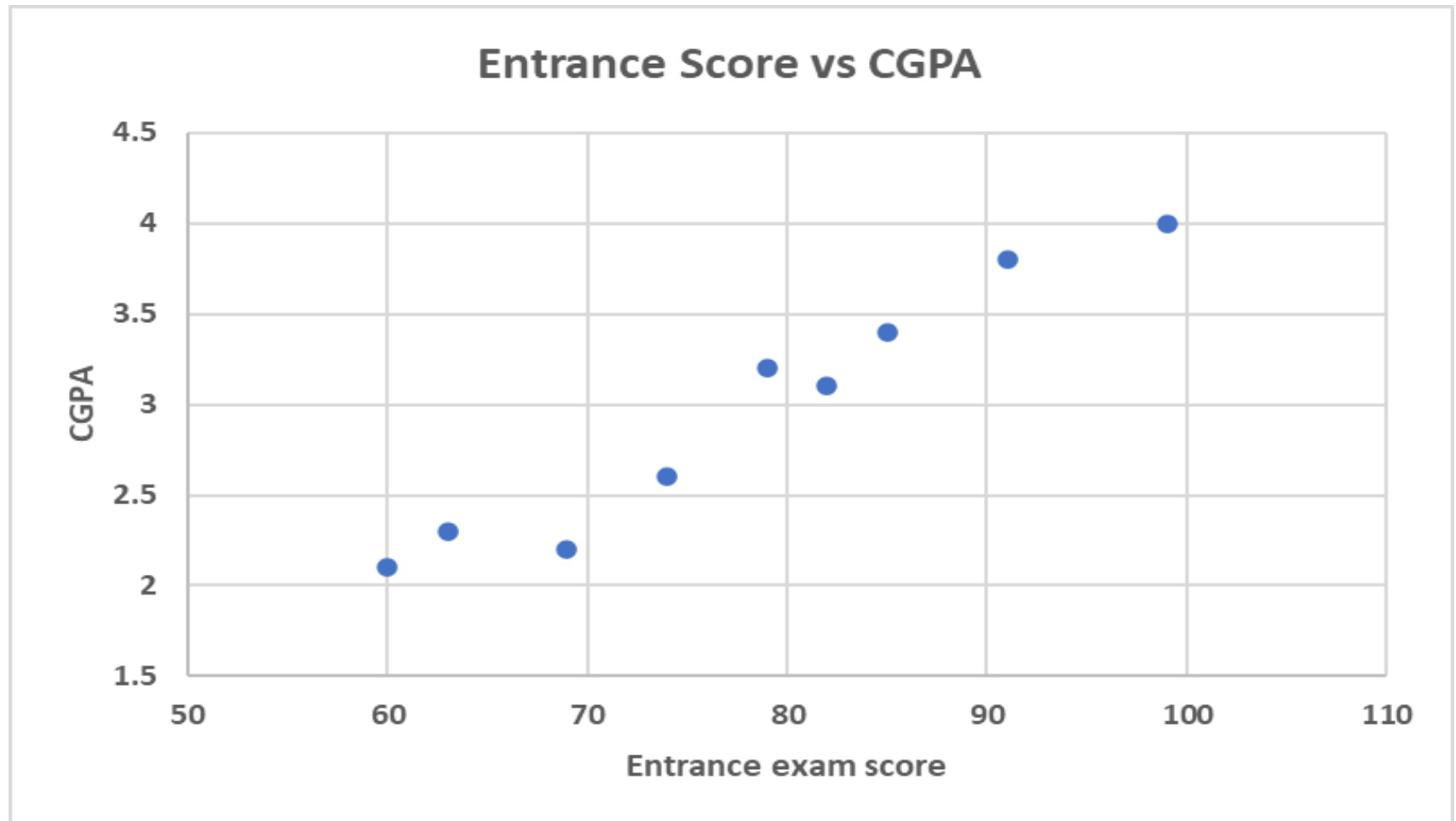
# Numerical example

The scores of a sample of students on entrance examination and cumulative grade point average (GPA) at graduation is given below:

Student	A	B	C	D	E	F	G	H	I
Entrance Score	74	69	85	63	82	60	79	91	99
CGPA	2.6	2.2	3.4	2.3	3.1	2.1	3.2	3.8	4.0

1. Draw a scatter diagram of CGPA vs Entrance score
2. Comment on the type and strength of relationship

# Scatter diagram of entrance score vs CGPA



# Karl Pearson's correlation coefficient

It is a mathematical measure of relationship between two quantitatively measured variables. It gives both degree and type of relationship.

The Karl Pearson's correlation coefficient based on sample is given by,

$$r_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} \dots(i)$$

where,

$r_{XY}$  =Karl Pearson's correlation coefficient

$Cov(X, Y)$ =Covariance between variables X and Y.

$Var(X)$ =Variance of X

$Var(Y)$ =Variance of Y

# Definitional Formula

Since,

$$\text{Sample } Cov(X, Y) = \frac{1}{n-1} \sum (X - \bar{X})(Y - \bar{Y}) \dots(\text{ii})$$

$$\text{Sample } Var(X) = \frac{1}{n-1} \sum (X - \bar{X})^2 \dots(\text{iii})$$

$$\text{Sample } Var(Y) = \frac{1}{n-1} \sum (Y - \bar{Y})^2 \dots(\text{iv})$$

Substituting the values in the equation (i), we get

$$r_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{\sum (Y - \bar{Y})^2}} \dots(\text{v})$$

If we write  $x = (X - \bar{X})$  and  $y = (Y - \bar{Y})$  then equation (v) reduces to:

$$r_{XY} = \frac{\sum xy}{\sum x^2 \sum y^2} \dots(\text{vi})$$

The equations (v) and (vi) are called definitional formula because they are derived from the definition.

# Computational Formula

The Pearson's product moment formula as given above is called definitional formula. It involves many calculations and it is tedious because mean has to be calculated first and most of the time mean is calculated as fractional number.

For computational purpose we use following formula:

$$r_{XY} = \frac{n \sum XY - \sum X \cdot \sum Y}{\sqrt{n \sum X^2 - (\sum X)^2} \sqrt{n \sum Y^2 - (\sum Y)^2}} \dots \text{(vii)}$$



# Properties of Pearson's $r$

1. The sample correlation coefficient  $r$  is a pure number without units.
2. The value of  $r$  always lies between -1 and +1. i.e.  $-1 \leq r_{XY} \leq +1$ . Hence -1 and +1 are limits of  $r$ .
3. The value of  $r$  is unchanged if either  $X$  or  $Y$  is multiplied by a constant or if a constant is added.

# Interpretation of Pearson's r

- If  $r = +1$ , then it suggests perfect positive correlation between two variables X and Y. Both variables rise and fall in same proportion.
- If  $r = -1$ , then there is perfect negative correlation between X and Y. When one variable rises, the other variable falls in same proportion.
- If  $r = 0$ , the two variables are statistically independent.
- The positive value of r indicates a positive relationship and high positive value of r indicate stronger relationship.
- The negative value of r indicates a negative relationship and closer the value of r to -1, stronger the negative relationship is.

Correlation coefficient (r)	Degree of association
$\pm 1$	Perfect
$\pm 0.7$ to $\pm 1.0$	Strong
$\pm 0.4$ to $\pm 0.7$	Moderate
$\pm 0.2$ to $\pm 0.4$	Weak
$\pm 0.1$ to $\pm 0.2$	Negligible
0	No association

# Limitation of Pearson's $r$

1. It quantifies only the strength of the linear relationship between two variables  $X$  and  $Y$ . If  $X$  and  $Y$  have a curvilinear relationship it will not provide a valid measure of this association.
2. It is highly sensitive to the extreme values present in the data set.
3. We should not compare correlation with causation. It doesn't tell us about cause and effect relationship between two variables. High correlation between  $X$  and  $Y$  doesn't necessarily imply that  $X$  causes  $Y$  or that  $Y$  causes  $X$  to change.

# Computing r (previous example)

Solution:

Y = Cumulative grade point average (CGPA)

X = Entrance exam score

The Karl Pearson's correlation coefficient is given by,

$$r_{XY} = \frac{n \sum XY - \sum X \cdot \sum Y}{\sqrt{n \sum X^2 - (\sum X)^2} \sqrt{n \sum Y^2 - (\sum Y)^2}}$$

# Computation table

Student	X	Y	X2	Y2	XY
A	74	2.6	5476	6.76	192.4
B	69	2.2	4761	4.84	151.8
C	85	3.4	7225	11.56	289
D	63	2.3	3969	5.29	144.9
E	82	3.1	6724	9.61	254.2
F	60	2.1	3600	4.41	126
G	79	3.2	6241	10.24	252.8
H	91	3.8	8281	14.44	345.8
I	99	4	9801	16	396
Total	<b>702</b>	<b>26.7</b>	<b>56078</b>	<b>83.15</b>	<b>2152.9</b>

The Karl Pearson's correlation coefficient is given by,

$$\begin{aligned} r_{XY} &= \frac{n \sum XY - \sum X \cdot \sum Y}{\sqrt{n \sum X^2 - (\sum X)^2} \sqrt{n \sum Y^2 - (\sum Y)^2}} \\ &= \frac{9 \cdot 2152.9 - 702 \cdot 26.7}{\sqrt{9 \cdot 56078 - 702^2} \sqrt{9 \cdot 83.15 - 26.7^2}} \\ &= 0.9741 \end{aligned}$$

**Conclusion:** There is a strong positive correlation between CGPA and Entrance exam score. It means that entrance exam score is directly predicting Cumulative GPA of student.

# Computing r using coded method

If variable X and Y are transformed to new variable U and V so that  $U = X - a$  and  $V = Y - b$ , then  $r_{XY}$  is computed by using following formula:

$$r_{XY} = \frac{n \sum UV - \sum U * \sum V}{\sqrt{n \sum U^2 - (\sum U)^2} \sqrt{n \sum V^2 - (\sum V)^2}}$$

# Computation table

Student	X	Y	$U = X - 80$	$V = Y - 3$	U2	V2	UV
A	74	2.6	-6	-0.4	36	0.16	2.4
B	69	2.2	-11	-0.8	121	0.64	8.8
C	85	3.4	5	0.4	25	0.16	2
D	63	2.3	-17	-0.7	289	0.49	11.9
E	82	3.1	2	0.1	4	0.01	0.2
F	60	2.1	-20	-0.9	400	0.81	18
G	79	3.2	-1	0.2	1	0.04	-0.2
H	91	3.8	11	0.8	121	0.64	8.8
I	99	4	19	1	361	1	19
Total	<b>702</b>	<b>26.7</b>	-18	-0.3	1358	3.95	70.9



*Now,*

$$\begin{aligned} r_{XY} &= \frac{n \sum UV - \sum U * \sum V}{\sqrt{n \sum U^2 - (\sum U)^2} \sqrt{n \sum V^2 - (\sum V)^2}} \\ &= \frac{9*70.9 - (-18)*(-0.3)}{\sqrt{9*1358 - (-18)^2} \sqrt{9*3.95 - (-0.3)^2}} \\ &= 0.974073 \end{aligned}$$

Both coded and non-coded method gives the same result, but coded method is computationally bit easier.

## Computing Karl Pearson's correlation coefficient in bivariate frequency distribution

For X and Y are presented in bi-variate frequency table then Karl Pearson's correlation coefficient is computed by following formula:

$$r_{XY} = \frac{n \sum fXY - \sum fX * \sum fY}{\sqrt{n \sum fX^2 - (\sum fX)^2} \sqrt{n \sum fY^2 - (\sum fY)^2}}$$

# Coded Method

If the two variables X and Y are transformed to two new variables U and V respectively by following rule

$$U = \frac{X-a}{h}$$
$$V = \frac{Y-b}{k}$$

where a and b are assumed mean of X and Y series and h and k are scaling factor, then Karl Pearson's correlation is given by,

$$r_{XY} = \frac{n \sum fUV - \sum fU * \sum fV}{\sqrt{n \sum fU^2 - (\sum fU)^2} \sqrt{n \sum fV^2 - (\sum fV)^2}}$$

# Spearman's correlation coefficient

Value of the Karl Pearson's correlation coefficient is markedly influenced by extreme values and thus does not provide a good description of the relationship between two variables X and Y when their distributions are skewed or contain outlying values. In such situation Pearson's correlation coefficient can't be used to measure the relationship between two variables rather values of both variables are converted to rankings and their rank correlation is calculated.

The Spearman's rank correlation coefficient is denoted by  $r_S$  is simply the Pearson's correlation coefficient between ranked values of X and Y.

Spearman's rank correlation coefficient is given by

$$r_S = \frac{\sum (R_X - \bar{R}_X)(R_Y - \bar{R}_Y)}{\sqrt{\sum (R_X - \bar{R}_X)^2} \sqrt{\sum (R_Y - \bar{R}_Y)^2}}$$

where

$R_X$  = rank of the X

$R_Y$  = rank of the Y

$\bar{R}_X$  = mean ranks for the X

$\bar{R}_Y$  = mean rank for the Y

# Computational Formula

An easier method of computing sample Spearman's rank correlation coefficient is given by,

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

Where,

$n$  = number pairs of data points in the sample

$d$  = difference between the rank of  $X$  and the rank of  $Y$ . If ranking is done correctly is always zero.

The Spearman's rank correlation coefficient also ranges from -1 to +1, so the interpretation of is same as Pearson's  $r$ .

# Spearman's rank correlation (Previous example)

Solution:

Y = Cumulative grade point average (CGPA)

X = Entrance exam score

The Spearman's rank correlation coefficient is given by,

$$r_S = 1 - \frac{6 \sum d^2}{n(n^2-1)}$$

Where,

n = number pairs of data points in the sample

d = difference between the rank of X and the rank of Y.

Now,

$$r_S = 1 - \frac{6*4}{9(9^2-1)}$$

# Computation table

Student ID	Entrance Score	CGPA	Rank of X (Rx)	Rank of Y (Ry)	Diff of ranks (d = Rx – Ry)	d <sup>2</sup>
A	74	2.6	4	4	0	0
B	69	2.2	3	2	1	1
C	85	3.4	7	7	0	0
D	63	2.3	2	3	-1	1
E	82	3.1	6	5	1	1
F	60	2.1	1	1	0	0
J	79	3.2	5	6	-1	1
H	91	3.8	8	8	0	0
I	99	4	9	9	0	0
					0	4

Now Spearman's rank correlation coefficient is given by,

$$r_s = 1 - \frac{6*4}{9(9^2-1)} \\ = 0.9667$$

Conclusion: There is a strong positive correlation between Entrance score and CGPA. It means that entrance score is directly predicting CGPA of students at the end of the course.



# Test of significance of correlation

A simpler, approximate method to assess whether a correlation coefficient is significant or not is the use of probable error.

The probable error of  $r$  is given by,

$$P.E. (r) = 0.6745 \times \frac{1 - r^2}{\sqrt{n}}$$

Where:

$r$  = sample correlation coefficient

$n$  = no. of data pairs in the sample or sample size

Test of significance:

1. If  $|r| < P.E. (r)$ , the correlation coefficient is insignificant
2. If  $|r| > 6 \times P.E. (r)$ , the correlation coefficient is significant
3. If  $P.E. (r) \leq |r| \leq 6 \times P.E. (r)$  test is inconclusive.

## Probable range of population correlation coefficient

The probable error (PE) of the correlation coefficient( $r$ ) is used to estimate the range within which the true population correlation coefficient ( $\rho$ ) is likely to lie.

1. The range  $r \pm P.E.(r)$  gives the 50 % confidence interval for true correlation coefficient  $\rho$
2. The range  $r \pm 2 P.E.(r)$  gives the 95 % confidence interval for true correlation coefficient  $\rho$
3. The range  $r \pm 3 P.E.(r)$  gives the 99 % confidence for the true correlation coefficient  $\rho$

# Example

Suppose a sample of 50 data pairs shows a sample correlation coefficient of  $r = 0.6$ .

The probable error is given by,

$$\begin{aligned} P.E. (r) &= 0.6745 \times \frac{1 - r^2}{\sqrt{n}} \\ &= 0.6745 \times \frac{1 - 0.6^2}{\sqrt{50}} = 0.062 \end{aligned}$$

Here the  $r$  is greater than  $P.E.(r)$ . In order to sure about significance of  $r$ , we need to calculate  $6 \times P.E.(r)$

Here,

$$6 \times P.E. (r) = 6 \times 0.06105 = 0.3663$$

Since, correlation coefficient ( $r$ ) is greater  $6 \times P.E. (r)$  the sample correlation is significant.

# Confidence Limits

S.N.	Limits	Confidence Probability
1	$r \pm P.E.(r)$ $= 0.6 \pm 0.06105$ $= 0.661, 0.539$	There is 50 % chance the true correlation is between 0.539 and 0.661.
2	$r \pm 2 P.E.(r)$ $= 0.6 \pm 2 \times 0.06105$ $= 0.7221, 0.4779$	There is 95 % chance the true correlation is between 0.4779 and 0.7221.
3	$r \pm 3 P.E.(r)$ $= 0.6 \pm 3 \times 0.06105$ $= 0.7832, 0.4169$	There is 99 % chance the true correlation is between 0.4169 and 0.7832

# Correction for tied data

If there are tied data, then they are assigned average ranks. In such case  $r_S$  is better calculated by formula.

$$(r_S)_C = 1 - \frac{6 \times \text{Adjusted } \sum d^2}{n(n^2-1)}$$

Where,

$$\text{Adjusted } \sum d^2 = \text{Unadjusted } \sum d^2 + \sum t_X + \sum t_Y$$

$$\sum t_X = \frac{\sum t_i(t_i^2-1)}{12}$$

$t_i$  = the number of tied values of X in a group of tie

$$\sum t_Y = \frac{\sum t_i(t_i^2-1)}{12}$$

$t_i$  = the number of tied values of Y in a group of ties.

# Example: Tied case

A corporation administers an aptitude test to all new sales representatives. Management is interested in the extent to which this test is able to predict sales representatives' eventual success. The accompanying table records average weekly sales (in thousands of dollars) and aptitude test scores for a random sample of eight representatives.

Weekly sales:	10	12	28	24	18	16	15	12
Test score:	55	60	85	75	80	85	65	60

- (a) Compute Spearman's rank correlation coefficient
- (b) Interpret the result

# Solution

Here,

Y = Weekly Sales (in thousands of dollars)

X = Test Score

n = 8

There are some tied observations in X series and as well as in Y-series. Hence, the formulas need adjustment.

The Spearman's rank correlation coefficient is given by,

$$(r_S)_C = 1 - \frac{6 \times \text{Adjusted } \sum d^2}{n(n^2 - 1)}$$

# Computation Table

S.N.	Weekly Sales (Y)	Test Score (X)	$R_x$	$R_y$	$d = R_x - R_y$	$d^2$
1	10	55	1	1	0	0
2	12	60	2.5	2.5	0	0
3	28	85	7.5	8	- 0.5	0.25
4	24	75	5	7	- 2	4
5	18	80	6	6	0	0
6	16	85	7.5	5	2.5	6.25
7	15	65	4	4	0	0
8	12	60	2.5	2.5	0	0
Total						$\sum d^2 = 10.5$



$$n = 8$$

$$\sum d^2 = \text{Unadjusted sum of square of difference} \\ = 10.5$$

$$\sum T_X = \text{Adjustment due to tied observations in X} \\ = \frac{\sum t_i(t_i^2 - 1)}{12} = \frac{2(2^2 - 1)}{12} + \frac{2(2^2 - 1)}{12} = 1$$

$$\sum T_Y = \text{Adjustment due to tied observations in Y} \\ = \frac{\sum t_i(t_i^2 - 1)}{12} = \frac{2(2^2 - 1)}{12} = 0.5$$

Now,

$$\begin{aligned}\text{Unadjusted } \sum d^2 &= \sum d^2 + \sum T_X + \sum T_Y \\ &= 10.5 + 1 + 0.5 \\ &= 12\end{aligned}$$

Finally,

$$\begin{aligned}(r_S)_C &= 1 - \frac{6 \times \text{Adjusted } \sum d^2}{n(n^2-1)} \\ &= 1 - \frac{6 \times 12}{10(10^2-1)} = 1 - 0.07273 = 0.93\end{aligned}$$

Conclusion: there is a strong positive correlation between weekly sales and aptitude test score of salesperson. It means that aptitude test score is a good indicator of likely weekly sales by salesperson.

# Spearman's rank correlation for already ranked data

# Ten programmers in a coding competition were judged by three experts and their ranks are given below:

Programmer	A	B	C	D	E	F	G	H	I	J
Rank by Judge I	1	6	5	10	3	2	4	9	7	8
Rank by Judge II	3	5	8	4	7	10	2	1	6	9
Rank by Judge III	6	4	9	8	1	2	3	10	5	7

Use the Spearman's rank correlation coefficient to find the pair of judges who have the nearest approach in coding

# Computational Table

Programmer	Rank by judge (R1)	Rank by judge (R2)	Rank by judge (R3)	d1 = R1-R2	d2 = R1-R3	d3 = R2-R3	d1^2	d2^2	d3^3
A	1	3	6	-2	-5	-3	4	25	9
B	6	5	4	1	2	1	1	4	1
C	5	8	9	-3	-4	-1	9	16	1
D	10	4	8	6	2	-4	36	4	16
E	3	7	1	-4	2	6	16	4	36
F	2	10	2	-8	0	8	64	0	64
G	4	2	3	2	1	-1	4	1	1
H	9	1	10	8	-1	-9	64	1	81
I	7	6	5	1	2	1	1	4	1
J	8	9	7	-1	1	2	1	1	4
				0	0	0	200	60	214

Rank correlation coefficient between Judge 1 and Judge 2

$$r_{S(12)} = 1 - \frac{6 \sum d_1^2}{n(n^2-1)} = 1 - \frac{6 * 200}{10(10^2-1)} = - 0.2121$$

Rank correlation coefficient between Judge 1 and Judge 3

$$r_{S(13)} = 1 - \frac{6 \sum d_2^2}{n(n^2-1)} = 1 - \frac{6 * 60}{10(10^2-1)} = + 0.6364$$

Rank correlation coefficient between Judge 2 and Judge 3

$$r_{S(23)} = 1 - \frac{6 \sum d_3^2}{n(n^2-1)} = 1 - \frac{6 * 214}{10(10^2-1)} = - 0.2969$$

Conclusion: There is a close agreement in between judge 1 and 3, in judging the programming competition.