UNIT II
MEASURES OF DISPERSION

Programme Educational Objectives

Our program will create graduates who:

1. *Will be recognized as a creative and an enterprising team leader.*
2. *Will be a flexible, adaptable and an ethical individual.*
3. *Will have a holistic approach to problem solving in the dynamic business environment.*

Research Methodology & Quantitative Techniques
Course Outcomes

CO1-Given a managerial problem and associated frequency distribution data, the student manager will be able to apply descriptive and inferential statistics to facilitate quick and rationale managerial decision making.

CO2-Given the data for two or more variables, the student manager will be able to estimate the strength of the relationship between two variables using ‘Karl Pearson’ and ‘Spearman’s Rank’ correlation coefficient.

CO3-Given the data for two or more variables, the student manager will be able to predict / forecast using as moving averages, regression and time series analysis.

CO4-Given a managerial problem, the student manager will be able to formulate it as 'research problem' and also will be able to suggest suitable research methodology to identify workable solutions.

CO5-Given a business Problem/situation, the student manager will be able to develop methods and instruments (questionnaire/ interview schedule) for collection and measurement of qualitative as well as quantitative data using primary and secondary sources from a given sampling framework.

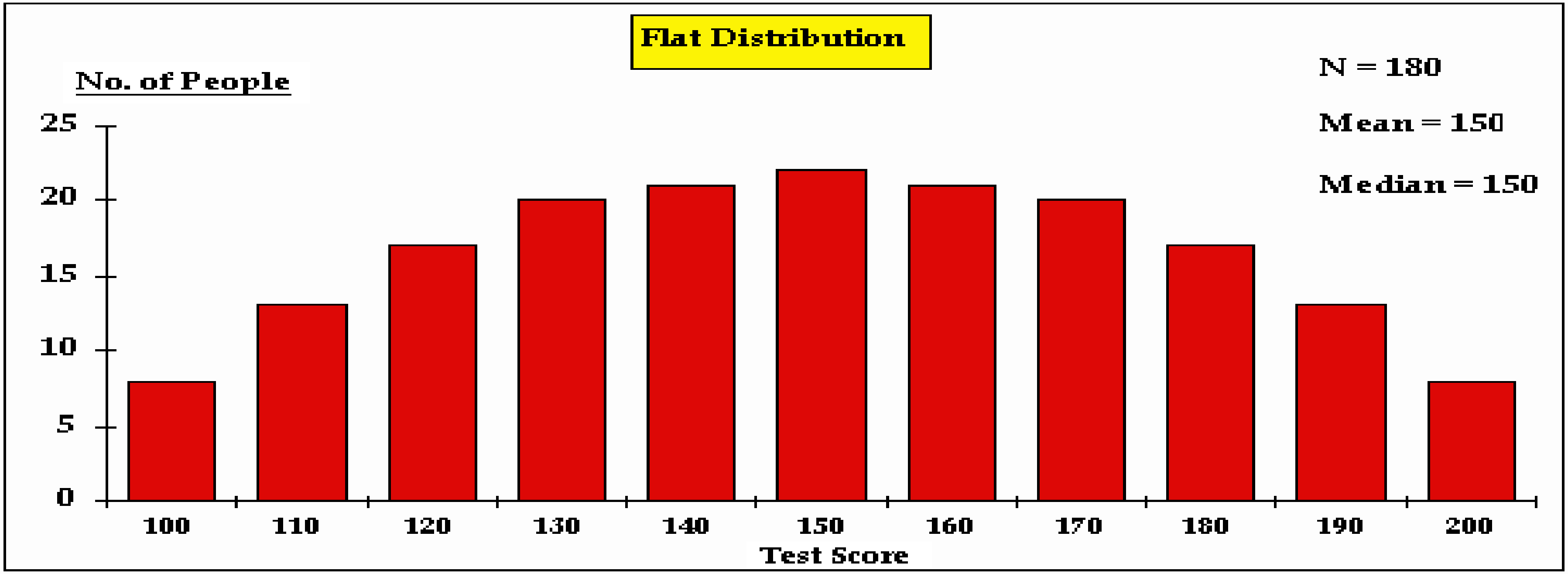
CO6-Given the sample statistics, the student manager will be able to apply Z, t and Chi-square tests to accept or reject the stated hypotheses for making sound decisions.

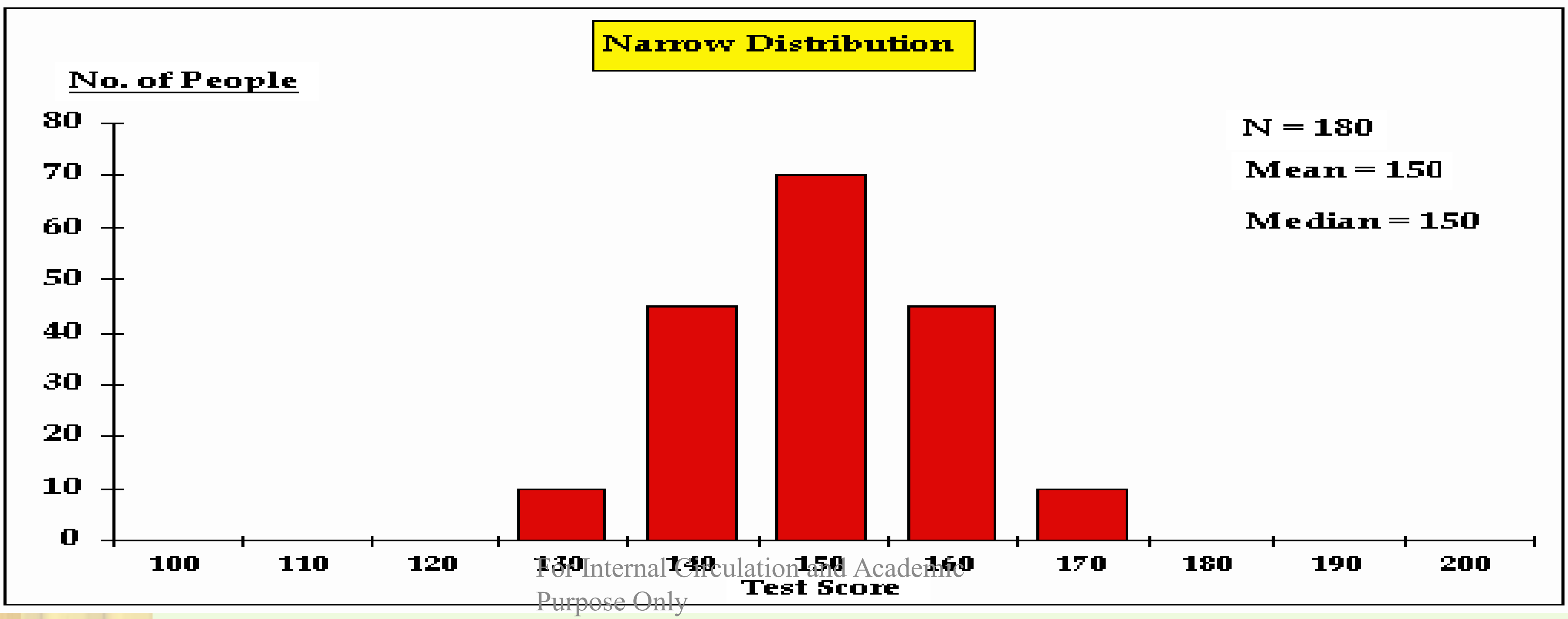
Learning Objectives

To understand the limitation of averages

To measures the extent to which the items vary from central value

To compare the data set in terms of variability, consistency by using absolute and relative measures


Flat Distribution
No. of People
N = 180
Mean = 150
Median = 150
25
20
15
10
5
0
100
110
120
130
140
150
160
170
180
190
200
Test Score


Narrow Distribution
No. of People
N = 180
Mean = 150
Median = 150
80
70
60
50
40
30
20
10
0
100
110
120
130
140
150
160
170
180
190
200
Test Score

WHAT IS DISPERSION?

Dispersion or spread is the degree of the scatter or variation of the variable about a central value.

Dispersion is the measure of the variations of the item

The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data

Measures of variability are usually used to indicate how tightly bunched the sample values are around the mean

PURPOSE OF MEASURING DISPERSION.

To judge the reliability of measures of central tendency

To make a comparative study of the variability of two series

To identify the causes of variability with a view to control it

To serve as a basis for further statistical analysis

RANGE.

*Absolute Measures of Dispersion: The measures of dispersion which are expressed in terms of **original units** of a data are termed as Absolute Measures.*

*Relative Measures of Dispersion: Relative measures of dispersion, are also known as coefficients of dispersion, are obtained as **ratios or percentages**. These are pure numbers **independent of the units** of measurement and used to compare two or more sets of data values.*

Absolute Measures • Range • Quartile Deviation • Mean Deviation • Standard Deviation

Relative Measure • Co-efficient of Range • Co-efficient of Quartile Deviation • Co-efficient of mean Deviation • co-efficient of Variation.

RANGE.

Range is the simplest possible measure of dispersion. It is the difference between the values of the extreme items of a series.

Range = L – S

Co-efficient of Range = (L-S)/(L+S)

1. The profits of a company for the last 8 years are given below. Calculate the Range and its Co-efficient

Year	1975	1976	1977	1978	1979	1980	1981	1982
Profits (in '000								230

(Answer: Range = 200, and its Co-efficient = 0.77)

2. Calculate Co-efficient of Range from the following data

Weekly Wages (Rs.)	50-60	60-70	70-80	80-90	90-100	100-110	110-120
No. of Laborers	50	45	45	40	35	30	30

First Method (By Taking Lower limits of first and upper limit of last interval)

Coefficient of Range $= (L-S)/(L+S)$

$= (120-50)/(120+50)$

$= 70/170 = 0.41$

Second Method (By Taking mid value of first and last interval)

Coefficient of Range $= (L-S)/(L+S)$

$= (115-55)/(115+55)$

$= 60/170 = 0.35$

It should be noted that in the calculation of Range only the values of the variable are taken into account and the ***frequencies are completely ignored****.*

INTER QUARTILE RANGE.

Q1 = the value of $\left(\frac{N+1}{4}\right)th$ *item*

Q3 = the value of $3\left(\frac{N+1}{4}\right)th$ *item.*

Quartile Deviation $= \frac{Q3-Q1}{2}$

Coefficient of Quartile Deviation = $\frac{Q3-Q1}{Q3+Q1}$

Find the Quartile Deviation and its Co-efficient from the following data, relating to the weekly of seven laborers

Weekly Wages (Rs.)	50	70	80	60	65	40	90

Answer: Quartile Deviation = Rs. 15, and its Co-efficient = 0.23

Calculate Quartile Deviation and its Co-efficient from the following data

Weight in Pounds	120	122	124	126	130	140	150	160
No. of Students	1	3	5	7	10	3	1	1

(Answer: Q1=124, Q3=130, Co-efficient of Q.D. = 0.0236)

QD in Continuous series.

Step 1: Calculate the class in which Q1 lies using formula N/4.

Step 2: Calculate the class in which Q3 lies using formula 3(N/4)

Step 3: $$Q1 = l_1 + \frac{l_2 - l_1}{f_1}(q1 - c)$$

Step 4: $$Q3 = l_1 + \frac{l_2 - l_1}{f_1}(q3 - c)$$

Step 5: Quartile Deviation $= \dfrac{Q3 - Q1}{2}$

Step 6: Coefficient Of QD $= \dfrac{Q3 - Q1}{Q3 + Q1}$

Calculate semi-inter quartile range and it's co-efficient from the following data

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Students	11	18	25	28	30	33	22	15	22

(Answer: Q.D. = 17.42 marks, co-efficient of Q.D. = 0.37)

Calculate Quartile Deviation and its relative measure

Variable	Frequency	Variable	Frequency
20-29	306	50-59	96
30-39	182	60-69	42
40-49	144	70-79	34

(Answer: Q.D. = 17.42 marks, co-efficient of Q.D. = 0.37)

Estimate an appropriate measure of dispersion of the following data

Income (Rs.)	No. of Persons	Income (Rs.)	No. of Persons
Less than 50	54	110-130	230
50-70	100	130-150	125
70-90	140	Above 150	51
90-110	300		

(Answer: Q.D. = 19.9 Rs.)

MEAN DEVIATION.

Mean deviation of a series is the arithmetic averages of the deviations of various items from a measure of central tendency (mean, median or mode).

$$\delta\bar{X} = \frac{\sum |d\bar{X}|}{N}$$

$$\delta M = \frac{\sum |dM|}{N}$$

$$\delta Z = \frac{\sum |dZ|}{N}$$

The following are the marks obtained by a batch of 9 students in a certain test. Calculate the mean deviation from mean and median.

SR. No.	Marks (out of 100)	SR. No.	Marks (out of 100)
1	68	6	38
2	49	7	59
3	32	8	66
4	21	9	41
5	54		

Mean = 428 / 9 = 47.55

Median = 49

X	X - MEAN	X - MEAN
68	20.45	20.45
49	1.45	1.45
32	-15.55	15.55
21	-26.55	26.55
54	6.45	6.45
38	-9.55	9.55
59	11.45	11.45
66	18.45	18.45
41	-6.55	6.55
		116.45

$$\delta\bar{X} = \frac{\sum | d\bar{X} |}{N}$$

$$\delta\bar{X} = \frac{116.45}{9}$$

$$\delta\bar{X} = 12.93$$

X	X - Median	X - Median
68	19	19
49	0	0
32	-17	17
21	-28	28
54	5	5
38	-11	11
59	10	10
66	17	17
41	-8	8
		115

$$\delta M = \frac{\Sigma \mid dM \mid}{N}$$

$$\delta M = \frac{115}{9}$$

$$\delta M = 12.77$$

Calculate Mean Deviation (from arithmetic Average) for the following values. Also, calculate its Coefficient. 4800, 4600, 4400, 4200, 4000.

(Answer: Mean Deviation = 240,
Co-efficient of Mean Deviation = 0.54)

Calculate Mean Deviation (from median) for the following values.

No. of Accidents	Persons having said no. of Accidents	No. of Accidents	Persons having said no. of Accidents
0	15	7	2
1	16	8	1
2	21	9	2
3	10	10	2
4	17	11	0
5	8	12	2
6	4		

STANDARD DEVIATION.

Standard Deviation is the square root of the arithmetic average of the squares of the deviations measured from the mean.

$$\sigma = \sqrt{\frac{\sum d^2}{N}}$$

Calculate the standard deviation of the heights of 10 students given below:

Height: 160, 160, 161, 162, 163, 163, 163, 164, 164, 170 (in cms)

X	d = X - 163	d^2
160	-3	9
160	-3	9
161	-2	4
162	-1	1
163	0	0
163	0	0
163	0	0
164	1	1
164	1	1
170	7	49
$\Sigma X = 1630$		$\sum d^2 = 74$

$$\sigma = \sqrt{\frac{\Sigma d^2}{N}}$$

$$\sqrt{\frac{74}{10}} \quad \sqrt{7.4}$$

= 2.72 cms

STANDARD DEVIATION – Method 2.

In this method we need not calculate the deviations. The formula used is as follows:

$$\sigma = \sqrt{\frac{\sum X^2 - (\sum X)^2 / N}{N}}$$

Calculate the standard deviation in the previous example using the above method.

X	X^2
160	25600
160	25600
161	25921
162	26244
163	26569
163	26569
163	26569
164	26896
164	26896
170	28900
$\Sigma X = 1630$	$\Sigma X^2 = 265764$

$$\sigma = \sqrt{\frac{\sum X^2 - (\sum X)^2 / N}{N}}$$

$$\sqrt{\frac{265764 - (1630)^2 / 10}{10}}$$

= 2.72 cms

STANDARD DEVIATION
Assumed Mean Method

In this method the formula used is as follows:

$$\sigma = \sqrt{\frac{\sum dX^2}{N} - \left(\frac{\sum dX}{N}\right)^2}$$

$$\sigma = \sqrt{\frac{\sum dX^2 - N(\bar{X} - A)2}{N}}$$

$$\sigma = \sqrt{\frac{\sum dX^2}{N} - (\bar{X} - A)}$$

Calculate the standard deviation in the previous example using the above method.

STANDARD DEVIATION - DISCRETE SERIES

In case of discrete series the formula used is as follows:

$$\sigma = \sqrt{\frac{\sum f d^2}{N}}$$

Calculate the standard deviation from the following data:

Size of item	Frequency	Size of item	Frequency
6	3	10	8
7	6	11	5
8	9	12	4
9	13		

X	**f**	**fX**	**d = X - 9**	**d²**	**fd²**
6	3	18	-3	9	27
7	6	42	-2	4	24
8	9	72	-1	1	9
9	13	117	0	0	0
10	8	80	1	1	8
11	5	55	2	4	20
12	4	48	3	9	36
	$\Sigma f = 48$	$\Sigma fX = 432$			$\Sigma fd^2 = 124$

Practice Example 1

Calculate standard deviation for the following distribution:

Values	10	20	30	40	50	60	70
Frequency	1	5	12	22	17	9	4

$$\sigma = \sqrt{\frac{\sum f dX^2}{N} - \left(\frac{f dX}{N}\right)^2}$$

(Answer: S.D. = 13.26)

Practice Example 2

The following table gives the number of finished articles turned out per day by different number of workers in a factory. Find the standard deviation of the daily output of finished articles.

No. of Articles	No. of Workers	No. of Articles	No. of Workers
18	3	23	17
19	7	24	13
20	11	25	8
21	14	26	5
22	18	27	4

(Answer: S.D. = 2.2 articles)

CONTINUOUS SERIES – Example

Calculate the standard deviation for the following table giving the age distribution of 542 members of the LOK SABHA.

Age	No. of Members	Age	No. of Members
20-30	3	60-70	140
30-40	61	70-80	51
40-50	132	80-90	2
50-60	153		

(Answer: S.D. = 11.9 years)

Let us ASSUME the MEAN AGE of the members as 45 years. Then we can use the following formula to calculate standard deviation:

$$\sigma = \sqrt{\frac{\sum f dX^2}{N} - \left(\frac{f dX}{N}\right)^2}$$

Age	X	f	dx	dx^2	fdx	fdx^2
20 – 30	25	3	-20	400	-60	1200
30 – 40	35	61	-10	100	-610	6100
40 – 50	45	132	0	0	0	0
50 – 60	55	153	10	100	1530	15300
60 – 70	65	140	20	400	2800	56000
70 – 80	75	51	30	900	1530	45900
80 – 90	85	2	40	1600	80	3200
		542			**5270**	**127700**

Practice Example 1

The following data relate to the age of a group of Govt. employees Calculate the standard deviation:

Age	50-55	45-50	40-45	35-40	30-35	25-30	20-25
No. of Employees	25	30	40	45	80	110	170

(Answer: S.D. = 9 years approx)

Practice Example 2

The following table relates to the profits and losses of 100 firms. Calculate the standard deviation of profits.

Profits	Number of Firms
5000 to 6000	8
4000 to 5000	12
3000 to 4000	30
2000 to 3000	10
1000 to 2000	5
0 to 1000	5
-1000 to 0	6
-2000 to -1000	8
-3000 to -2000	9

SD Rs. 2791

Practice Example 3

Calculate the standard deviation of the following data:

Age Under (Years)	10	20	30	40	50	60	70	80
No. of Persons dying	15	30	53	75	100	110	115	125

(Answer: S.D. = 19.7 years)

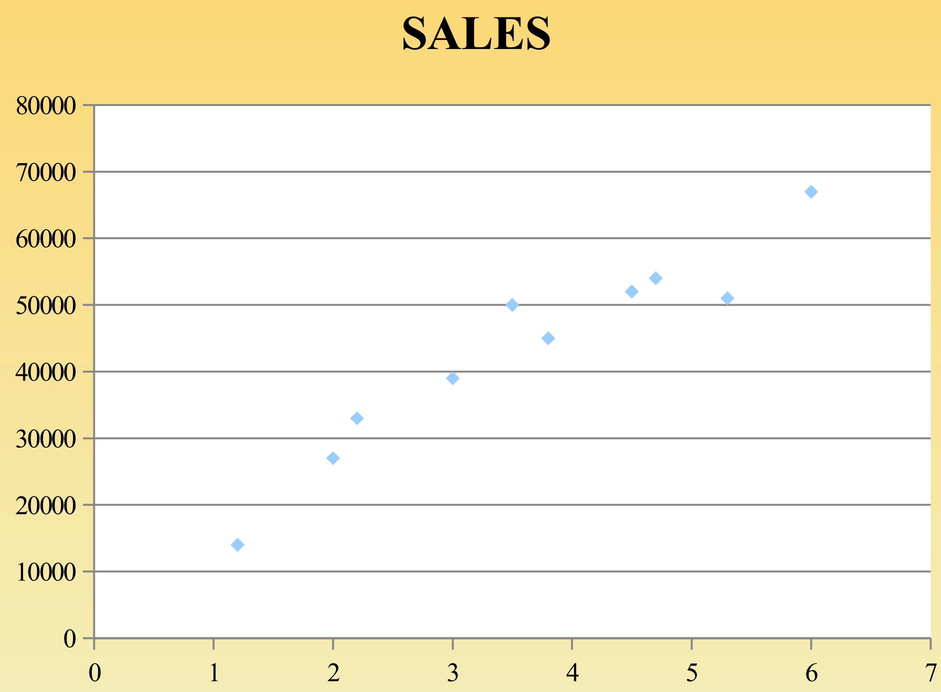
CORRELATION

CORRELATION

- Correlation analysis is used to Measure strength of the association between two variables
  - Only concerned with strength of the relationship
  - No causal effect is implied

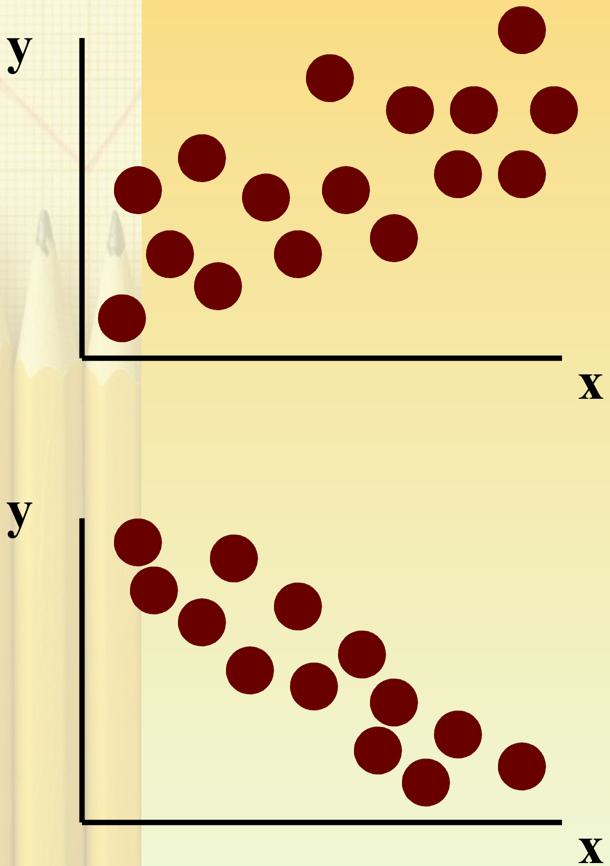
EXAMPLE

YEAR	ADV. BUDGET	SALES
2001	2	27000
2002	3	39000
2003	4.5	52000
2004	6	67000
2005	2.2	33000
2006	1.2	14000
2007	3.5	50000
2008	4.7	54000
2009	5.3	51000
2010	3.8	45000

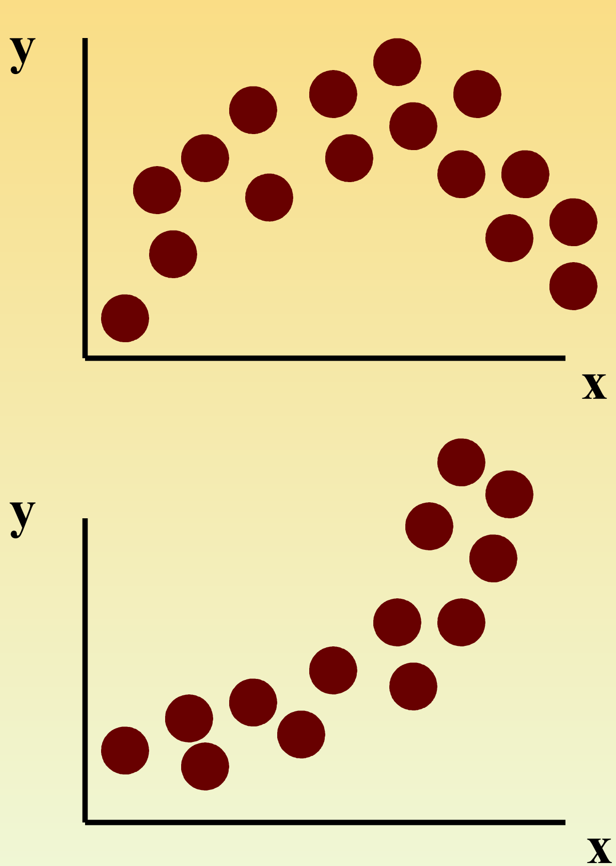


Scatter Plots

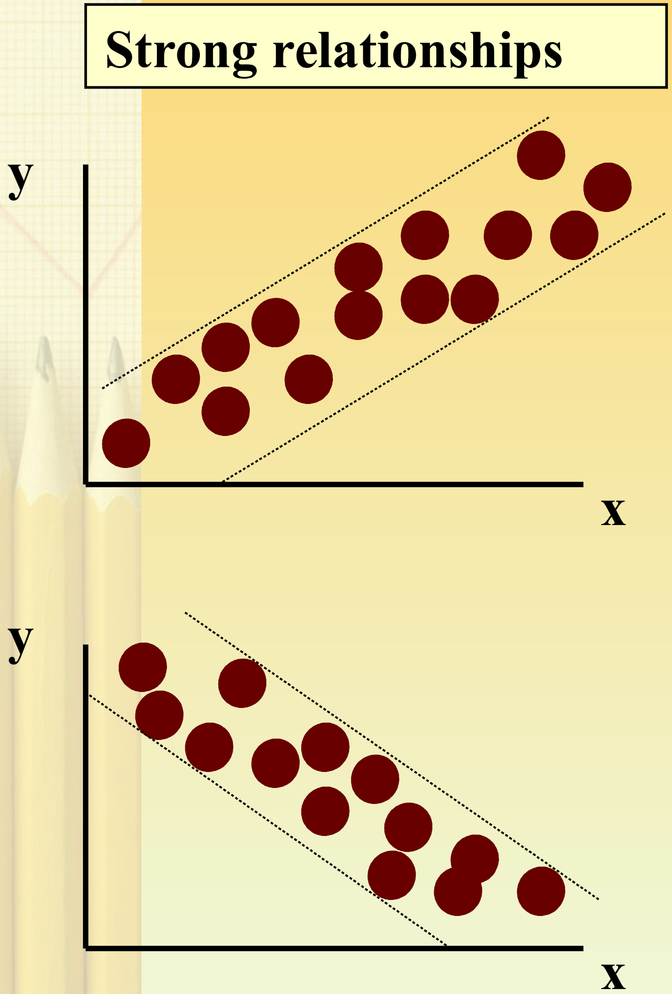
Linear relationships

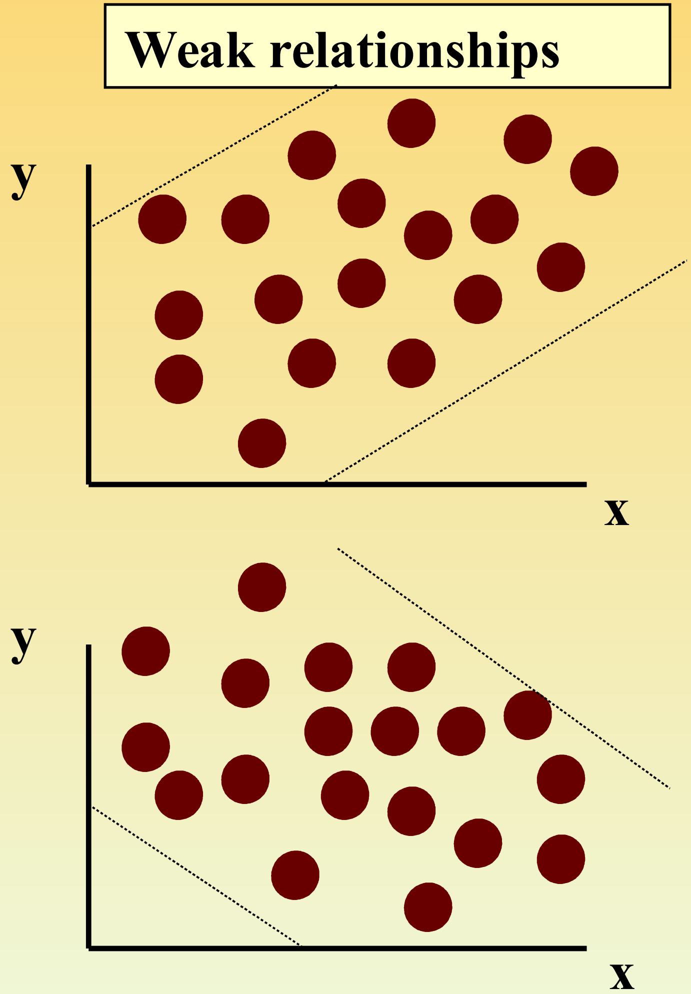


Curvilinear relationships



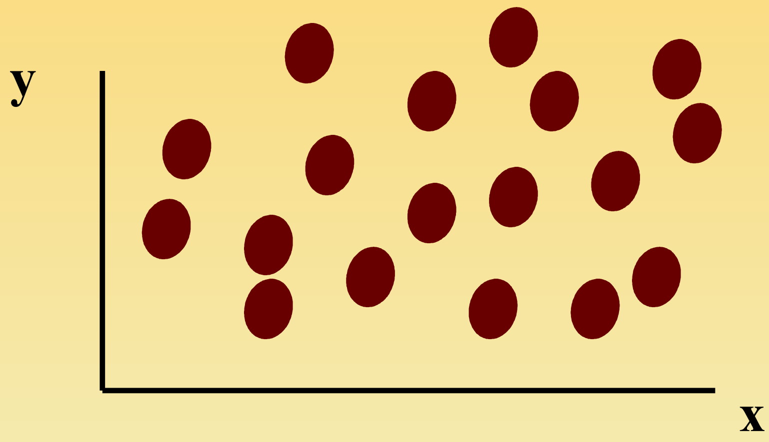
Scatter Plots

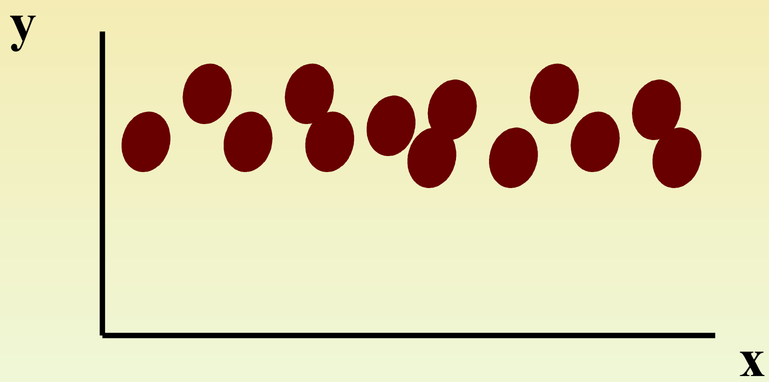




Scatter Plots

No relationship





Correlation Coefficient

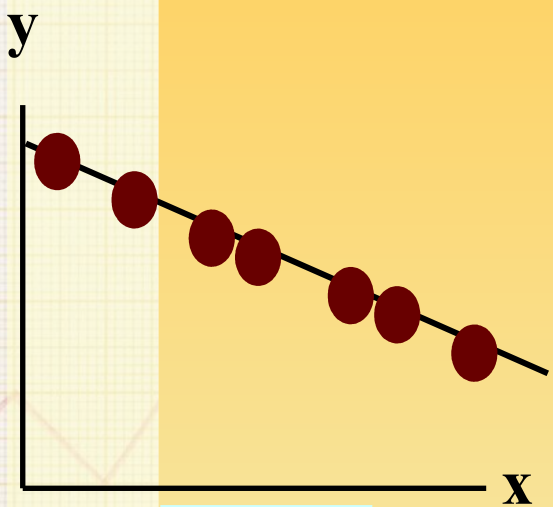
- The population correlation coefficient **ρ (rho)** measures the strength of the association between the variables
- The sample correlation coefficient **r** is an estimate of ρ and is used to measure the strength of the linear relationship in the sample observations

Features of ρ and r

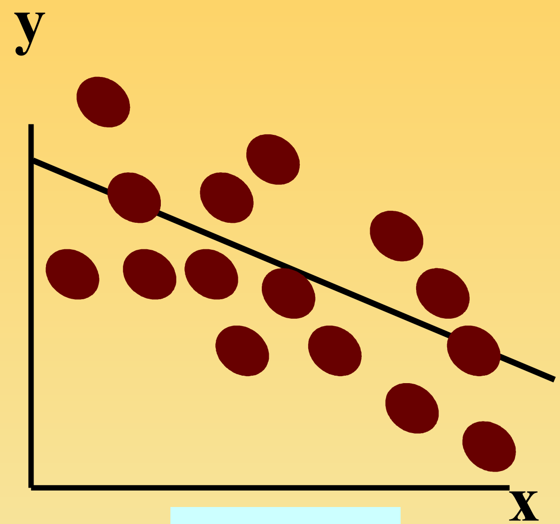
- Range between -1 and 1
- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker the linear relationship

Features of ρ and r

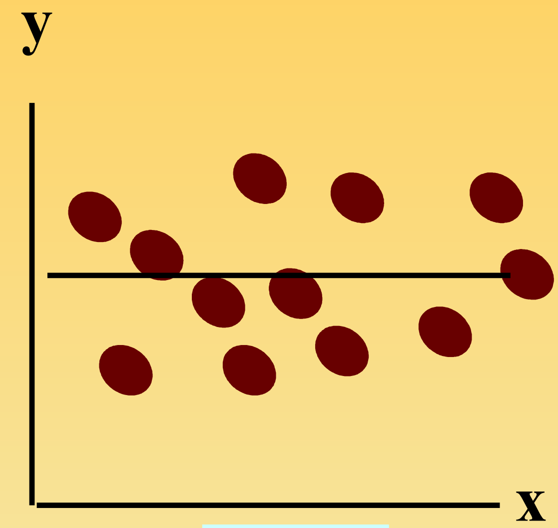
Value of r	Interpretation
+ or - 1	Perfect correlation between variables
+ or – 0.9 to 0.99	Very high degree of correlation
+ or – 0.7 to 0.9	High degree of correlation
+ or – 0.5 to 0.7	Moderate degree of correlation
+ or – 0.25 to 0.5	Low degree of correlation
+ or – 0 to 0.25	Very low degree of correlation
0	No correlation



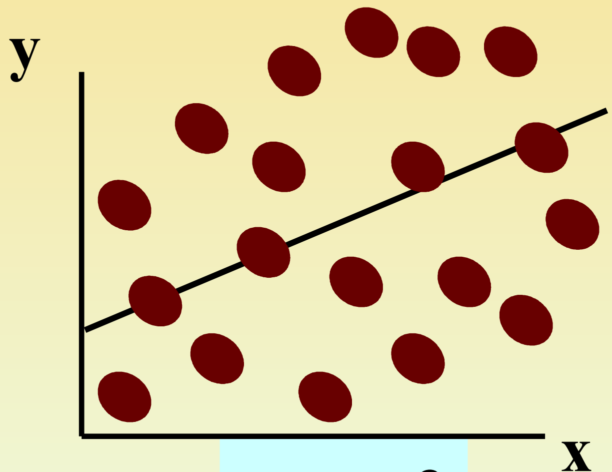
r = -1



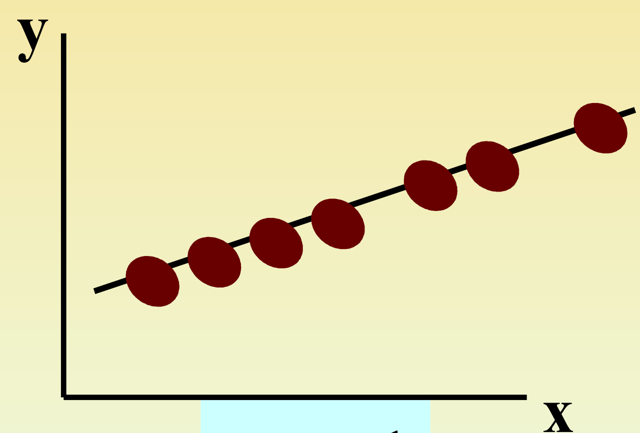
r = -.6



r = 0



r = +.3



r = +1

Karl Pearson's coefficient of correlation - r

$$r = \frac{\sum x y}{\sqrt{\sum x^2 \sum y^2}}$$

where,

r = Correlation Coefficient

$x = (X - \bar{X})$

$y = (Y - \bar{Y})$

From the data given below find out the Pearson's correlation coefficient for the given data and comment on the nature of correlation.

ROLL NUMBER OF STUDENT	MARKS IN MATHS (OUT OF 100)	MARKS IN STATISTICS (OUT OF 100)
1	50	60
2	70	85
3	40	52
4	30	38
5	80	90

Step 1

Calculating the Mean of first variable i.e. $\bar{X}$

Step 2

Calculating the Mean of second variable i.e. $\bar{Y}$

Step 3

Calculating the value of x = $X - \bar{X}$

Step 4

Calculating the value of y = $Y - \bar{Y}$

Step 5

Calculating the $\sum x y$

Step 6

Calculating the $\sum x^2$

Step 7

Calculating the $\sum y^2$

Step 8

Putting the values in the formula to calculate r

SR. NO.	X	Y	x = X - 54	x^2	y = Y - 65	y^2	XY
1	50	60	-4	16	-5	25	20
2	70	85	16	256	20	400	320
3	40	52	-14	196	-13	169	182
4	30	38	-24	576	-27	729	648
5	80	90	26	676	25	625	650
SUM (∑)	270	325		1720		1948	1820
MEAN	54	65					

$$r = \frac{1820}{\sqrt{1720 \times 1948}}$$

$$\mathbf{r = +\ 0.99}$$

Comments:

- The Pearson's correlation coefficient is +0.99 which indicates that there is a very high degree of positive correlation between the marks obtained in Maths and marks obtained in Statistics.
- In other words a student who scores high marks in Maths also scores high marks in Statistics whereas a student who scores low marks in Maths also score low marks in Statistics.

Example

Making use of the below data calculate the coefficient of correlation.

CASE	A	B	C	D	E	F	G	H
X1	10	6	9	10	12	13	11	9
X2	9	4	6	9	11	13	8	4

ANSWER: r = + 0.896

Example

*Making use of the below data calculate the coefficient of correlation. (**HINT: Since r is a pure number, changing the scale of series does not affect its values).***

CASE	A	B	C	D	E	F	G
X	10000	20000	30000	40000	50000	60000	70000
Y	3000	5000	6000	8000	10000	11000	13000

ANSWER: r = + 0.997

Example

From the data given below find out the Pearson's correlation coefficient for the given data and comment on the nature of correlation.

Date	BSE SENSEX closing	Gold prices (per 10 gms)
5th Sep 2012	17250	29000
6th Sep 2012	16800	29900
7th Sep 2012	17100	29500
8th Sep 2012	[illegible]	31000

Answer = - 0.95

Direct Method Karl Pearson's coefficient of correlation - r

In this method we need not calculate the deviations of items from mean.

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

Example

Making use of the below data calculate the coefficient of correlation by Direct Method.

CASE	A	B	C	D	E	F	G	H
X1	10	6	9	10	12	13	11	9
X2	9	4	6	9	11	13	8	4

ANSWER: r = + 0.896

Shortcut or Assumed Mean Method for calculation of - r

When actual means are in fractions, the calculations of the correlation coefficient become complicated. So assume a mean and use the below formula:

$$r = \frac{N \sum dx\,dy - \{(\sum dx) \times (\sum dy)\}}{\sqrt{N \sum dx^2 - (\sum dx)^2}\ \sqrt{N \sum dy^2 - (\sum dy)^2}}$$

Calculate the coefficient of correlation between X and Y from the following data. Assume 69 and 112 as the mean value for series X and Y respectively.

X	78	89	99	60	59	79	68	61
Y	125	137	156	112	107	136	123	108

X	dx (X-69)	dx^2	Y	dy (Y-112)	dy^2	dx dy
78	9	81	125	13	169	117
89	20	400	137	25	625	500
99	30	900	156	44	1936	1320
60	-9	81	112	0	0	0
59	-10	100	107	-5	25	50
79	10	100	136	24	576	240
68	-1	1	123	11	121	-11
61	-8	64	108	-4	16	32
$\sum X = 593$	$\sum dx = 41$	$\sum dx^2 = 1727$	$\sum Y = 1004$	$\sum dy = 108$	$\sum dy^2 = 3468$	$\sum dxdy = 2248$

$$r = \frac{N \sum dx\, dy - \{(\sum dx) \times (\sum dy)\}}{\sqrt{N \sum dx^2 - (\sum dx)^2}\ \sqrt{N \sum dy^2 - (\sum dy)^2}}$$

$$r = \frac{8 \times 2248 - \{(41) \times (108)\}}{\sqrt{8 \times 1727 - (41)^2}\ \sqrt{8 \times 3468 - (108)^2}}$$

$$r = 0.97$$

Calculate Karl Pearson's correlation coefficient from the advertisement cost and sales as per data given below:

Advertisement Cost	**39**	**65**	**62**	**90**	**82**	**75**	**25**	**98**	**36**	**78**
sales	**47**	**53**	**58**	**86**	**62**	**68**	**60**	**91**	**51**	**84**

(Answer r = +0.7804)

A computer while calculating correlation coefficient between variables X and Y from 25 pairs of observations obtained the following results:

$N = 25 \qquad \Sigma X = 125 \qquad \Sigma X^2 = 650$

$\Sigma Y = 100 \qquad \Sigma Y^2 = 460 \qquad \Sigma XY = 508$

It was, however, discovered at the time of checking that two pairs of observation were not correctly copied. They were taken as (6, 14) and (8, 6) while the correct values were (8, 12) and (6, 8). Prove that the correct value of correlation coefficient should be 2/3.

Following are the results of B Com exam:

Calculate the coefficient of correlation between age and successful candidates in examination.

Age of students	13-14	14-15	15-16	16-17	17-18	18-19	19-20	20-21	21-22	22-23
Students appeared for exam	200	300	100	50	150	400	250	150	25	75
Successful students	124	180	65	34	99	252	145	81	12	33

(Answer r = - 0.7747)

Calculation of PE – Probable Error

Probable Error helps in interpreting the value of Coefficient of Correlation. It tells us about the reliability of the value of Coefficient of Correlation. It is obtained by the following formula:

$$P.E.r = 0.6745\ \frac{1 - r^2}{\sqrt{N}}$$

- If the value of r is less than PE then r is not at all significant i.e. there is no evidence of correlation.
- **If value of r is 6 times the value PE** then r is significant i.e. there is a certain evidence of correlation.
- By adding and subtracting the PE from r we get the upper and lower limits within which the coefficient of correlation of population is expected to lie.

EXAMPLE

Calculate Probable Error and the limits of correlation in population given that r=0.8 and number of pairs in observed sample is 16.

$$P.E.r = 0.06$$

The following table gives the distribution of items of production and also the relatively defective items among them, according to size groups. Find the correlation coefficient between **size** and **defect quality** and its probable error.

Size Group	15-16	16-17	17-18	18-19	19-20	20-21
Number of Items	200	270	340	360	400	300
Number of defective Items	150	162	170	180	180	114

Calculation of Coefficient of Determination

The **square of the coefficient of correlation** is known as **Coefficient of Determination.**

$$r^2$$

It tells us about what amount of variation in dependent variable has been explained by independent variable.

Assume that the coefficient of correlation between rainfall and per acre yield of rice is 0.8. Find out the coefficient of determination and comment on its value.

Coefficient of determination = r2 = $(0.8)^2$ = 0.64

Comments:

0.64 Coefficient of Determination means that 64% variation in per acre yield of rice (Dependent Variable) is explained by rainfall (Independent Variable). 36% variation is unexplained by rainfall; it may be due to other factors such as use of fertilizers, soil and seed quality, etc.

Following table gives the results of an examination. Calculate Karl Pearson's Correlation Coefficient and its probable error? Also comment if the value of correlation coefficient obtained is significant or not.

Age	13-14	14-15	15-16	16-17	17-18
percent age of failures	39	40	43	43	36
Age	**18-19**	**19-20**	**20-21**	**21-22**	
percent age of failures	39	48	44	56	

r = + 0.658 and PEr = 0.127

Calculation of Correlation Coefficient in Grouped Data

Calculate the coefficient of correlation for the following data.

AGE OF HUSBANDS	AGE OF WIVES					
	10-20	20-30	30-40	40-50	50-60	TOTAL
15-25	6	3	-	-	-	9
25-35	3	15	11	-	-	29
35-45	-	11	14	7	-	32
45-55	-	-	6	12	3	21
55-65	-	-	-	3	6	9
TOTAL	9	29	31	22	9	100

FORMULA

$$r = \frac{N \sum f dx\, dy - \{(\sum fdx) \times (\sum fdy)\}}{\sqrt{N \sum f dx^2 - (\sum fdx)^2}\ \sqrt{N \sum f dy^2 - (\sum fdy)^2}}$$

			X	10-20	20-30	30-40	40-50	50-60	Total			
			MV	15	25	35	45	55				
		dX (X-35)		-20	-10	0	10	20				
Y	MV	dY Y-40		-2	-1	0	1	2		fdY	fdY^2	fdxdy
15-25	20	-20	-2	6	3	-	-	-	9	-18	36	30
25-35	30	-10	-1	3	15	11	-	-	29	-29	29	21
35-45	40	0	0	-	11	14	7	-	32	0	0	0
45-55	50	10	1	-	-	6	12	3	21	21	21	18
55-65	60	20	2	-	-	-	3	6	9	18	36	30
Total				9	29	31	22	9	100	$\Sigma fdY = -8$	$\Sigma fdY^2 = 122$	
			fdX	-18	-29	0	22	18	$\Sigma fdX = -7$			

$$r = \frac{100 \times 99 (-7 \times -8)}{\sqrt{100 \times 123 - (-7)^2}\ \sqrt{100 \times 122 - (-8)^2}}$$

$$r = +0.8073$$

RANK CORRELATION (SPEARMAN'S CORRELATION COEFFICIENT)

$$R = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

When Ranks are given:

- Take the differences of two ranks (R1 – R2) and denote these differences by D.
- Square these differences and obtain the total ∑ D2
- Apply the formula given above.

EXAMPLE

The ranking of 10 students in two subjects A and B are given below. Calculate the Spearman's Correlation Coefficient.

A	6	5	3	10	2	4	9	7	8	1
B	3	8	4	9	1	6	10	7	5	2

R1	R2	(R1 - R2) = D	D^2
6	3	3	9
5	8	-3	9
3	4	-1	1
10	9	1	1
2	1	1	1
4	6	-2	4
9	10	-1	1
7	7	0	0
8	5	3	9
1	2	-1	1
		$\sum D^2 =$	36

$$R = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

$$R = 1 - \frac{6 \times 36}{10\ (10^2 - 1)}$$

$$R = 0.782$$

Two ladies were asked to rank 7 different types of lipsticks. The ranks given by them are as follows: Calculate Spearman’s rank correlation coefficient.

Lipsticks	A	B	C	D	E	F	G
Neelu	2	1	4	3	5	7	6
Neena	1	3	2	4	5	6	7

(Answer: 0.786)

Ten competitors in a beauty contest are ranked by three judges in the following order.

1st judge	1	6	5	10	3	2	4	9	7	8
2nd judge	3	5	8	4	7	10	2	1	6	9
3rd judge	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to determine which pair of judges has the nearest approach to common tastes in beauty.

(Answer: Pair of 1 and 3, R= 0.636)

Where Ranks are NOT given:

- Assign the ranks by taking either the highest value or lowest value as 1 and find out the R.

Calculate the Spearman's coefficient of correlation between marks given to ten students by judges X and Y in a certain competitive examination.

Student No.	1	2	3	4	5	6	7	8	9	10
Marks by judge X	52	53	42	60	45	41	37	38	25	27
Marks by judge Y	65	68	43	38	77	48	35	30	25	50

Where Ranks are EQUAL:

- In some cases it may be found necessary to rank two or more entries as equal. In such a case it is customary to give each individual an average rank.
- ***Thus, if two entries are ranked equal at 5th place they are each given the rank (5+6)/2 = 5.5 while if 3 are ranked equal at 5th place they are given the rank (5+6+7)/3 = 6.***
- Where equal ranks are assigned to some entries an adjustment in the formula for calculating the rank coefficient of correlation is made.
- **The adjustment consists of adding 1/12 (m^3 – m) to the value of ∑ D^2.**
- Here, m stands for number of items whose ranks are common.

EXAMPLE

Obtain the rank correlation coefficient between the variables X and Y from the following pairs of observed values.

X	50	55	65	50	55	60	50	65	70	75
Y	110	110	115	125	140	115	130	120	115	160

Number of repetitions in series X

50 is repeated 3 times (m=3)

55 is repeated 2 times (m=2)

65 is repeated 2 times (m=2)

Number of repetitions in series Y

115 is repeated 3 times (m=3)

110 is repeated 2 times (m=2)

$$R = 1 - \frac{6\left[\sum D^2 + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right)\right]}{N(N^2 - 1)}$$

EXAMPLE

Obtain the rank correlation coefficient between the variables X and Y from the following pairs of observed values.

X	**15**	**10**	**20**	**28**	**12**	**10**	**16**	**18**
Y	**16**	**14**	**10**	**12**	**11**	**15**	**18**	**12**

ANSWER: R = - 0.369

EXAMPLE

Obtain the rank correlation coefficient between the variables X and Y from the following pairs of observed values.

X	40	50	60	60	80	50	70	60
Y	80	120	160	170	130	200	210	130

ANSWER: R = 0.429

EXAMPLE

Obtain the rank correlation coefficient from the following data.

Sr No.	1	2	3	4	5	6	7	8	9	10
Rank Diff.	-2	?	-1	+3	+2	0	-4	+3	+3	-2

ANSWER: R = 0.636

EXAMPLE

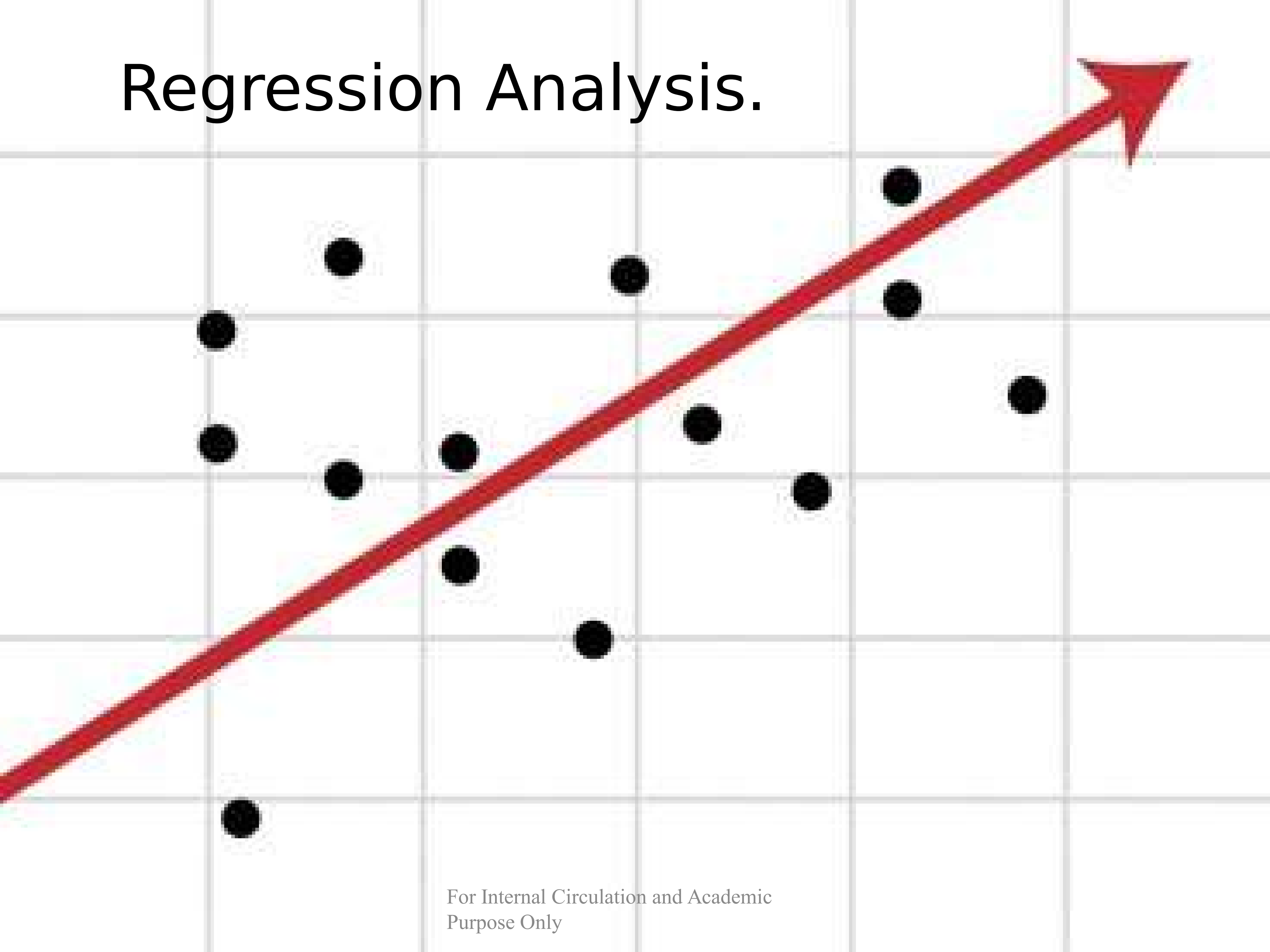
The coefficient of rank correlation of marks in two subjects for a group of 10 students was found to be 0.5.

It was later noticed that the difference in ranks in two subjects of one student was wrongly taken as 3 instead of 7.

Find the correct value of Rank Correlation.

ANSWER: $R = 0.258$

Regression Analysis.



Correlation & Regression coefficients.

$$r = \sqrt{b_{xy} \, X \, b_{yx}}$$

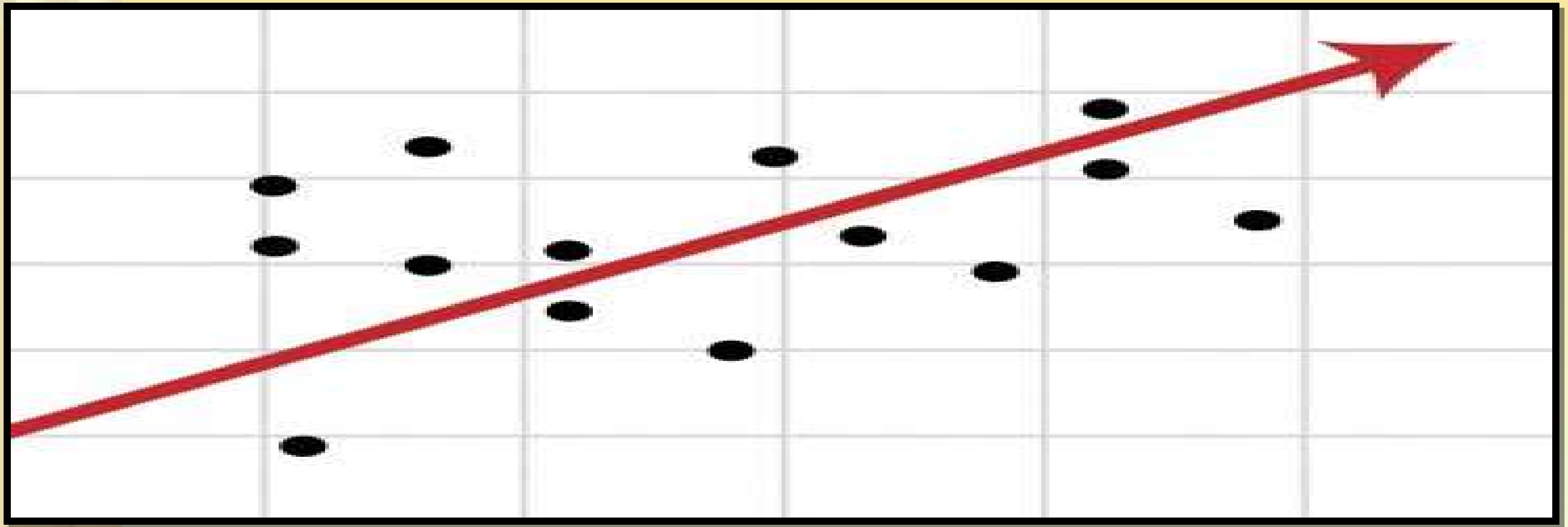
Find the value of the correlation coefficient if regression coefficients of Y on X and X on Y are 0.46 and 0.8 respectively.

Answer : r = 0.606

Method of LEAST SQUARES.

The method of least squares is a mathematical technique.

It is used to obtain the equation of a line which best fits the given data.



Method of LEAST SQUARES.

Equation of a straight line is $y = a + bx$

Normal Equations for obtaining the values of a and b are as follows

(i) $\sum y = Na + b \sum x$

(ii) $\sum xy = a \sum x + b \sum x^2$

EXAMPLE.

Fit a straight line of Y on X from the following data:

X	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Y	**2**	**1**	**3**	**2**	**4**	**3**	**5**

SOLUTION.

	X	Y	X^2	XY
	0	2	0	0
	1	1	1	1
	2	3	4	6
	3	2	9	6
	4	4	16	16
	5	3	25	15
	6	5	36	30
SUM	21	20	91	74

$\sum y = Na + b \sum x$

$\sum xy = a \sum x +$

$20 = 7a + 21b$

$74 = 21a +$

$20 = 7a + 21b$

Multiply by 3

$60 = 21a + 63b$

SOLUTION.

74 = 21a + 91b
minus
60 = 21a + 63b

14 = 28 b
b = 0.5

Putting value of b = 0.5 in above equations;
a = 1.357

Y = 1.357 + 0.5 X
Equation of the line which fits the given data.

EXAMPLE.

Fit a straight line of Y on X from the following data:

X	**0**	**2**	**4**	**6**	**8**	**10**	**12**
Y	**4**	**2**	**6**	**4**	**8**	**6**	**10**

EXAMPLE.

Fit a straight line of X on Y from the following data:

X	**0**	**2**	**4**	**6**	**8**	**10**	**12**
Y	**4**	**2**	**6**	**4**	**8**	**6**	**10**

Normal Equations for obtaining the values of a and b are as follows

(i) $\sum X = Na + b \sum Y$

(ii) $\sum XY = a \sum Y + b \sum Y^2$

EXAMPLE.

From the following data obtain the Regression equations of Y on X and X on Y

X	**6**	**2**	**10**	**4**	**8**
Y	**9**	**11**	**5**	**8**	**7**

Answer:

$$Y = 11.9 - 0.65X$$

$$X = 16.4 - 1.3Y$$

Derivation of Lines of Regression directly from the data. (DEVIATION FROM ACTUAL MEAN)

The equation of regression line of X on Y is

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$X - \bar{X} = \frac{r\,\sigma x}{\sigma y}(Y - \bar{Y})$$

The equation of regression line of Y on X is

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$Y - \bar{Y} = \frac{r\,\sigma y}{\sigma x}(X - \bar{X})$$

Derivation of Lines of Regression directly from the data. (DEVIATION FROM ACTUAL MEAN)

Regression Coefficient of X on Y is

$$b_{xy} = \frac{\sum x\,y}{\sum y^2}$$

Regression Coefficient of Y on X is

$$b_{yx} = \frac{\sum x\,y}{\sum x^2}$$

In these formulas x and y are deviations from mean.

Regression Coefficients directly from the data.

The regression Coefficient of X on Y is

$$b_{xy} = \frac{r\,\sigma x}{\sigma y} = \frac{\sum XY - N\ \bar{X}\ \bar{Y}}{\sum Y^2 - N\,(\bar{Y})^2}$$

The regression Coefficient of Y on X is

$$b_{yx} = \frac{r\,\sigma y}{\sigma x} = \frac{\sum XY - N\ \bar{X}\ \bar{Y}}{\sum X^2 - N\,(\bar{X})^2}$$

EXAMPLE.

From the following data obtain the Regression equations using deviation of means method.

X	**6**	**2**	**10**	**4**	**8**
Y	**9**	**11**	**5**	**8**	**7**

Answer:

Y = 11.9 – 0.65 X

X = 16.4 – 1.3 Y

X on Y

X	x	x^2	Y	y	y^2	xy
6	0	0	9	1	1	0
2	- 4	16	11	3	9	- 12
10	4	16	5	-3	9	-12
4	-2	4	8	0	0	0
8	2	4	7	-1	1	-2
30	0	40	40	0	20	-26

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$b_{xy} = \frac{\sum xy}{y^2}$$

$$b_{xy} = \frac{-26}{20} = -1.3$$

X – 6 = -1.3 (Y – 8)

X – 6 = -1.3Y + 10.4

X = -1.3Y + 16.4

$$\bar{X} = \frac{\sum X}{N} = \frac{30}{5} = 6$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{40}{5} = 8$$

EXAMPLE.

Following data relate to the scores of 9 salesmen of a company in an intelligence test and weekly sales in thousands. If the intelligence test score of a salesman is 65 what would be his weekly expected sales?

Test Score	**50**	**60**	**50**	**60**	**80**	**50**	**80**	**40**	**70**
Weekly Sales	**30**	**60**	**40**	**50**	**60**	**30**	**70**	**50**	**60**

Answer: Rs. 53750

EXAMPLE.

Following table shows ages (X) and blood pressure (Y) of 8 persons. Find the expected blood pressure of a person whose age is 49 years.

X	**52**	**63**	**45**	**36**	**72**	**65**	**47**	**25**
Y	**62**	**53**	**51**	**25**	**79**	**43**	**60**	**33**

Answer: 49.5

EXAMPLE.

In a correlation study the following values are obtained. Find the two regression equations that are associated with above values.

	X	Y
Mean	**65**	**67**
Std. Deviation	**2.5**	**3.5**
Correlation Coefficient	**0.8**	

Answer:

X = 26.72 + 0.57 Y

Y = -5.8 + 11.2 X

Properties of Regression Coefficients.

1. Both regression coefficients will have the same sign i.e. positive or negative
2. Square root of the product of two regression coefficients will give Correlation coefficient.
3. If one regression coefficient is more than 1 then the other will be less than 1.
4. The regression coefficients will have the same sign as that of the correlation coefficient.
5. Arithmetic mean of Regression coefficient is greater than correlation coefficient.
6. The two regression lines intersect at coordinates which are mean of series X and series Y.

Derivation of Lines of Regression directly from the data. (DEVIATION FROM ASSUMED MEAN)

Regression Coefficient of X on Y is

$$b_{xy} = r\frac{\sigma x}{\sigma y} = \frac{N\sum dx\,dy - (\sum dx \, X \sum dy)}{N\sum dy^2 - (\sum dy)^2}$$

Regression Coefficient of Y on X is

$$b_{yx} = r\frac{\sigma y}{\sigma x} = \frac{N\sum dx\,dy - (\sum dx \, X \sum dy)}{N\sum dx^2 - (\sum dx)^2}$$

In these formulas dx and dy are deviations from ASSUMED mean.

EXAMPLE.

From the following data obtain the Regression equations by taking deviation of series X from 5 and of series Y from 7.

X	**6**	**2**	**10**	**4**	**8**
Y	**9**	**11**	**5**	**8**	**7**

Answer:

Y = 11.9 – 0.65 X

X = 16.4 – 1.3 Y

EXAMPLE.

A panel of two judges P and Q graded eight dance performances as below. However, Judge Q was absent during the 7th performance. What might have been his ranking if judge Q had been present?

P	**46**	**42**	**44**	**40**	**43**	**41**	**37**	**45**
Q	**40**	**38**	**36**	**35**	**39**	**37**	**?**	**41**

Answer:
$Y = +0.75\ X + 5.75$
33.5 marks

EXAMPLE.

From the following regression equations find the mean values of X and Y series.

$$8X - 10Y = -66$$

$$40X - 18Y = 214$$

Hint: Regression lines (X on Y and Y on X) cut each other at the point of mean.

Answer:
Mean of X series = 13
Mean of Y series = 17

EXAMPLE.

In a partially destroyed laboratory record of analysis of correlation data, only the following results are legible:

Variance of X = 9

Regression equations

$8X – 10Y + 66 = 0$

$40X – 18Y = 214$

Find out :

1. The mean values of X and Y
2. Coefficient of correlation between X and Y
3. Standard Deviation of Y

Solution.

To find the mean values of X and Y we will solve the given equations:

Regression equations given:

(i) $8X - 10Y = -66$

(ii) $40X - 18Y = 214$

Multiplying equation (i) by 5 we get

(i) $40X - 50Y = -330$

Subtracting equation (ii) from equation (i) we get:

$-32\,Y = -544$

$Y = -544 / -32 = 17$ (Mean of Y = 17)

Solution.

Substituting the value of Y=17 in equation (i)

(i) $8X - 10Y = -66$

$8X - 10\,(17) = -66$

$8X - 170 = -66$

$8X = 104$

$X = 104 / 8 = 13$ (Mean of $X = 13$)

To find the Correlation coefficient (r) we must find the regression coefficients (b_{xy} and b_{yx}). However, we don't know which equation is X on Y and which equation is Y on X. So lets assume equation (i) as X on Y

Solution.

Assuming equation (i) as X on Y we will try to find bxy

(i) $8X - 10Y = -66$

$8X = -66 + 10Y$

$X = (-66/8) + (10/8)\ Y$

$X = a + bxy\ Y$

bxy= 10 / 8 = 1.25

From equation (ii) we can find out byx

(ii) $40X - 18Y = 214$

$Y = -(214/18) + (40/18)\ X$

bxy= 40 / 18 = 2.22

Solution.

Here we can see that both the regression coefficients are greater than 1 which is not possible. Therefore, our assumption that equation (i) is equation of X on Y is wrong.

Equation (i) is actually equation of Y on X. So we can write the equation as:

$-10 Y = - 8X -66$

$Y = (8/10) X + 6.6$ which means that $byx = 0.8$

From equation (ii) i.e. X on Y we get

$40X = 214 + 18Y$ or $X = (214/40) + (18/40) Y$ which means that $bxy = 18/40 = 0.45$

Solution.

We know that Correlation coefficient is square root of product of Regression coefficients:

r = square root of 0.8 X 0.45 = 0.6

We can further calculate standard deviation of Y using the formula of Regression Coefficient.

Standard Deviation of Y = 4

STANDARD ERROR OF ESTIMATES.

- The standard error of estimate measures the accuracy of the estimated figures.
- Smaller the value of standard error, closer will be the dots to the regression line and better will be the estimates.
- If standard error of estimate is zero then the is no variation about the line and correlation is perfect.

$$S_{xy} = \sqrt{\frac{\sum (X - X_c)^2}{N}}$$

$$S_{xy} = \sigma x \sqrt{1 - r^2}$$

$$S_{xy} = \sqrt{\frac{\sum X^2 - a\sum X - b\sum XY}{N}}$$

$$S_{yx} = \sqrt{\frac{\sum (Y - Y_c)^2}{N}}$$

$$S_{yx} = \sigma y \sqrt{1 - r^2}$$

$$S_{yx} = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{N}}$$

Uses of Regression analysis.

1. Regression line facilitates to predict the values of dependent variable from the given value of independent variable.
2. Standard Error facilitates to obtain a measure of error involved in using the regression line as a basis of estimation.
3. Regression coefficients help us to calculate coefficient of correlation and coefficient of determination,
4. Regression analysis is a highly useful tool in Economics and business.

TIME SERIES & FORECASTING.

- **Components of Time Series.**
- **Trend - Moving averages, semi-averages and least-squares.**
- **Seasonal variation, cyclic variation and irregular variation.**
- **Index numbers, calculation of seasonal indices.**
- **Additive and multiplicative models.**
- **Forecasting, Non linear trend – second degree parabolic trends**

TIME SERIES.

A Time Series is a set of observations taken at specified times, usually at equal intervals.

Mathematically, a time series is defined by values Y_1, Y_2, ... of a variable at times t_1, t_2,... Thus Y is a function of t symbolized by $Y = F(t)$.

UTILITY OF
TIME SERIES ANALYSIS.

Helps in understanding past behaviour.
Helps in planning future operations.
Helps in evaluating current accomplishments.
Facilitates comparison.

COMPONENTS OF A TIME SERIES.

Secular Trend	T
Seasonal Variations	S
Cyclical Variations	C
Irregular Variations	I

$Y = T+S+C+I$ or $Y = TxSxCxI$

Additive Model Multiplicative Model

Measurement of Trend

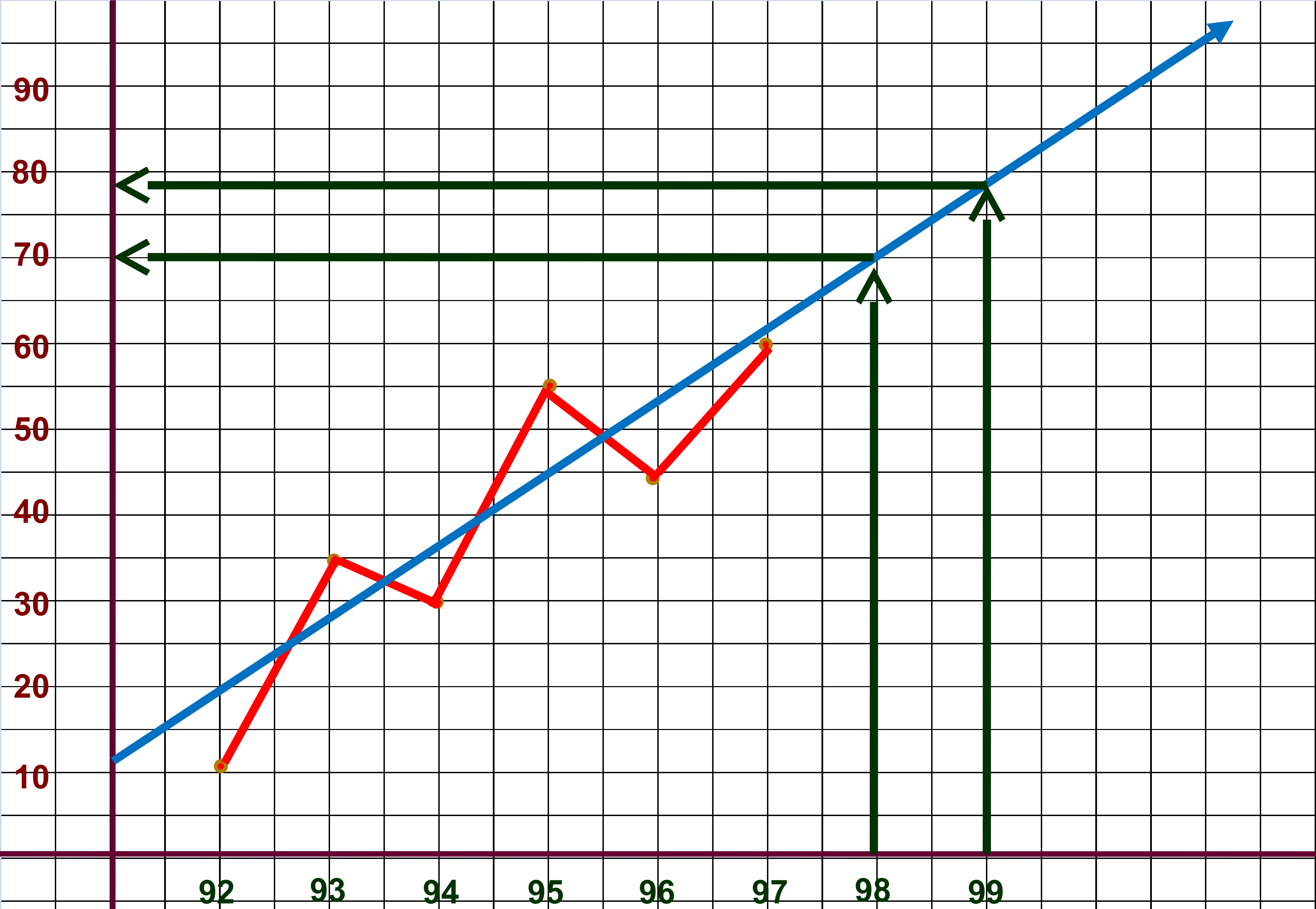
Freehand or Graphical method
Semi-average method
Moving average method
Least squares method.

Freehand or Graphical method

1. Plot the time series on a graph paper.
2. Examine carefully the direction of dots.
3. Draw a straight line according to personal judgement.

Fit a trend line to the following data using the Freehand method and predict values for 1998 & 1999.

YEAR	SUGAR PRODUCTION (Million Tonnes)
1992	10
1993	35
1994	30
1995	55
1996	45
1997	60


90
80
70
60
50
40
30
20
10
92
93
94
95
96
97
98
99

Fit a trend line to the following data using the Freehand method and predict values for 2009 & 2010.

YEAR	Sales (Millions)
2001	5
2002	15
2003	10
2004	25
2005	30
2006	20
2007	35
2008	45

Method of SEMI AVERAGES

1. Divide the data in two equal parts. In case of odd years, omit the middle year.
2. Obtain the average of each part.
3. Plot the two points against the midpoint of class interval on a graph.
4. Joint the two points to get a trend line.

Fit a trend line to the following data using the semi averages method and predict values for 2009 & 2010.

YEAR	Sales (Millions)
1993	102
1994	105
1995	114
1996	110
1997	108
1998	116
1999	112

Method of MOVING AVERAGES

There can be two ways to calculate moving averages.

1. 3 year, 5 year or 7 year moving averages. These are called odd year moving averages.

Or

2. 2 year, 4 year, 6 year or 8 year moving averages.

There is a slight difference in these two ways.

Calculate the 3 year moving averages of the production figures given below.

YEAR	PRODUCTION	YEAR	PRODUCTION
1985	15	1993	63
1986	21	1994	70
1987	30	1995	74
1988	36	1996	82
1989	42	1997	90
1990	46	1998	95
1991	50	1999	102
1992	56		

Construct 5 year moving averages of the number of students studying in a college.

YEAR	No. of students	YEAR	No. of students
1990	332	1995	405
1991	317	1996	410
1992	357	1997	427
1993	392	1998	405
1994	402	1999	438

Calculate the trend values by taking 4 year moving averages.

YEAR	VALUE	YEAR	VALUE
1984	12	1991	100
1985	25	1992	82
1986	39	1993	65
1987	54	1994	49
1988	70	1995	34
1989	87	1996	20
1990	105	1997	7

WEIGHTED MOVING AVERAGES

Generally weighted moving average is used to forecast trend figures.

WMA gives higher weightage to recent figures.

Calculate the trend values using 3 year WMA for the following data. Weights are to be assigned in order 1, 2, 3.

YEAR	SALES	YEAR	SALES
2001	10	2008	18
2002	12	2009	20
2003	12	2010	18
2004	14	2011	24
2005	16	2012	28
2006	18		
2007	22		

Year	Sales	WT	Wtd Sales	3 Y WMT	3 Y WMA
01	10	1	10	----	----
02	12	2	24	70	11.66
03	12	3	36	74	12.33
04	14	1	14	82	13.66
05	16	2	32	100	16.66
06	18	3	54	108	18
07	22	1	22	112	18.66
08	18	2	36	118	19.66
09	20	3	60	114	19
10	18	1	18	126	21
11	24	2	48	150	25
12	28	3	84	----	----

Calculate the trend values using 5 year WMA for the following data. Weights are to be assigned in order 1, 2, 2, 3, 3.

YEAR	SALES	YEAR	SALES
1990	18	1997	32
1991	20	1998	28
1992	21	1999	36
1993	26	2000	34
1994	22	2001	35
1995	24	2002	44
1996	30	2003	46
		2004	42

LEAST SQUARES METHOD

EQUATION OF SRTAIGHT TREND LINE

$$Y = a + b X$$

Normal Equations for obtaining the values of a and b are as follows

(i) $\sum Y = Na + b \sum X$

(ii) $\sum XY = a \sum X + b \sum X^2$

N = Number of years,

X = Converted value for years.

(i) $\sum Y = Na + b \sum X$

(ii) $\sum XY = a \sum X + b \sum X^2$

If we take the middle year as year of origin

then $\sum X = 0$.

Then a = $(\sum Y / N)$ = Mean of Y

AND

Putting the value of $\sum X = 0$ in equation (ii).

Then b = $(\sum XY / \sum X^2)$

Fit a straight line trend for the following series and Estimate the values for 1997

YEAR	Production
1990	60
1991	72
1992	75
1993	65
1994	80
1995	85
1996	95

$$Y = 76 + 4.857\,X$$

$$Y_{1997} = 95.428$$

Fit a straight line trend for the following series and Estimate the values for 1998

YEAR	Production
1989	38
1990	40
1991	65
1992	72
1993	69
1994	60
1995	87
1996	95

$Y = 65.75 + 3.667\ X$

$Y_{1997} = 106.087$

CALCULATION OF SEASONAL INDEX

There are 4 methods of computing seasonal component of time series:

1. Simple Average Method
2. Ratio to Trend Method
3. Ration to Moving Average Method
4. Link Relative Method

We will study only the first method...

Simple Average Method

1. Find the Quarterly totals.
2. Find Quarterly averages for each quarter.
3. Find grand average of quarterly averages.
4. Find the seasonal index of each quarter by dividing its quarterly average by grand average.

The given table shows trend free figures of quarterly sales made by a mega mall. Find the seasonal indices.

YEAR	I	II	III	IV
2003	39	20	60	85
2004	45	23	62	90
2005	60	32	76	100
2006	47	35	65	85

The following time series data on consumption of cold drinks contains only seasonal and irregular variations. Construct indices for seasonal variations using simple arithmetic mean.

YEAR	I	II	III	IV
2003	39	20	60	85
2004	45	23	62	90
2005	60	32	76	100
2006	47	35	65	85

Following data gives monthly production figures. Find monthly seasonal indices.

Yr	01	02	03	04	05
Jan	31	34	36	39	42
Feb	28	30	32	34	37
Mar	27	29	28	34	37
Apr	25	26	26	31	33
May	23	24	25	29	31
Jun	21	23	25	27	29
Jul	22	24	28	28	30
Aug	24	26	30	30	33
Sep	26	28	34	32	35
Oct	30	32	36	36	39
Nov	32	34	39	38	42
Dec	34	36	40	41	45

References and Suggested Readings

Fundamentals of Statistics by S.C. Gupta

Statistics Methods by S.P.Gupta