Param Pujya Dr. Babasaheb Ambedkar Smarak Samiti's

Dr. Ambedkar Institute of Management Studies & Research

Deeksha Bhoomi, Nagpur - 440010 (Maharashtra State) INDIA

NAAC Accredited with 'A' Grade

Tel: +91 712 6521204, 6521203 ,6501379
Email: info@daimsr.in

UNIT II
CORRELATION

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

Learn to identify the direction and strength of a correlation between two factors.

CORRELATION

- Correlation analysis is used to Measure strength of the association between two variables

  - Only concerned with strength of the relationship
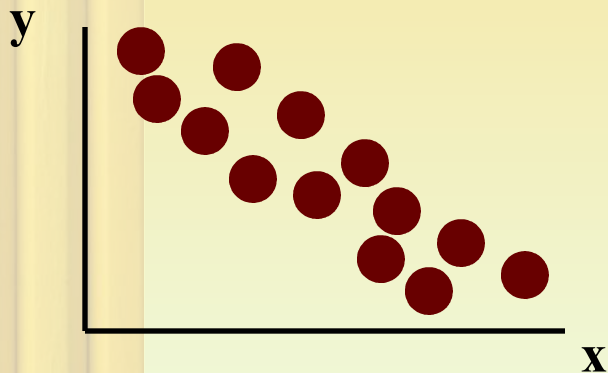
  - No causal effect is implied

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

SALES
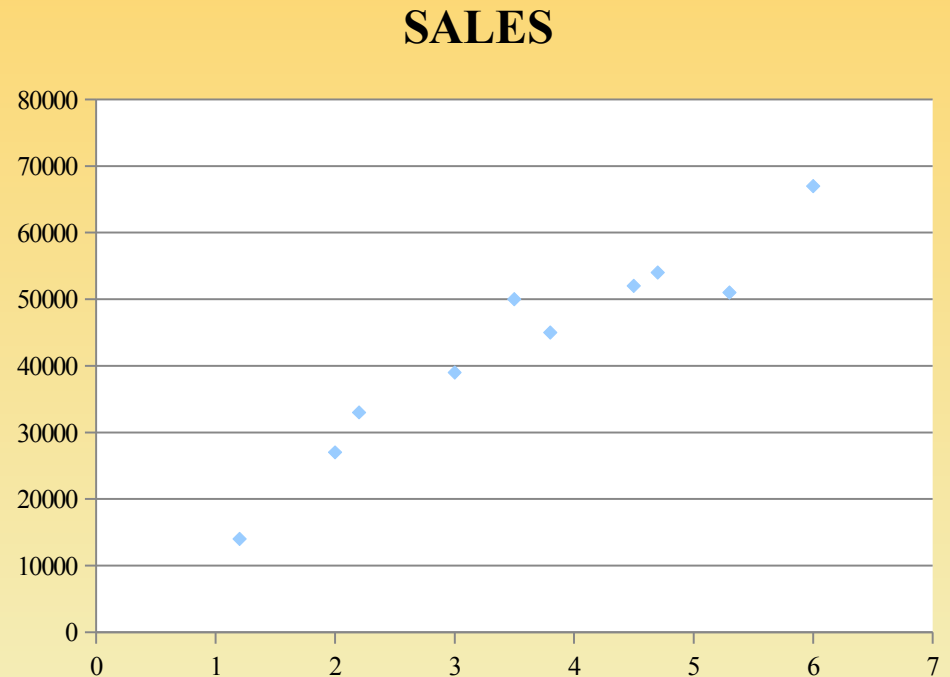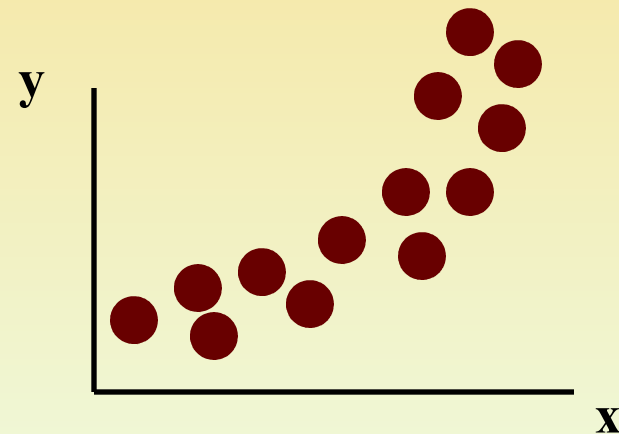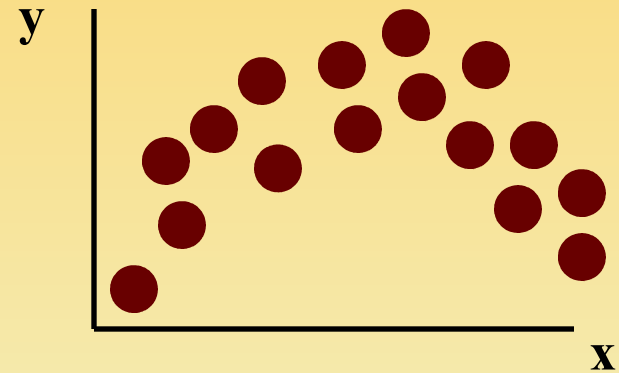
Scatter Plots

Scatter Plots

Strong relationships
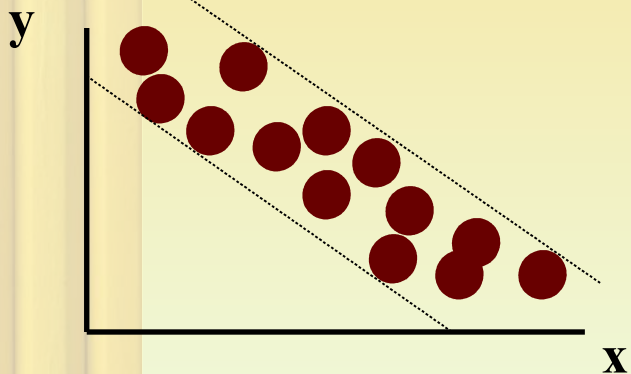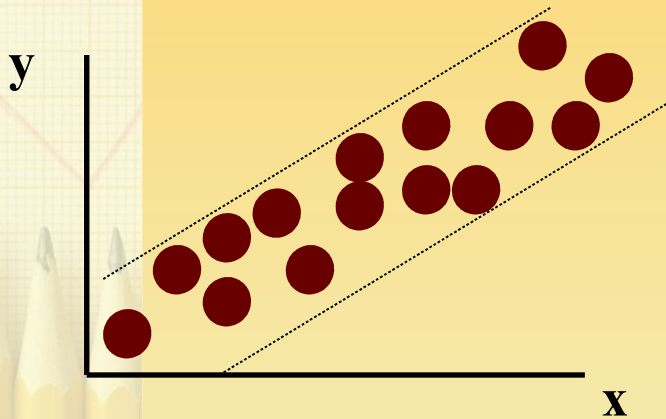
Weak relationships
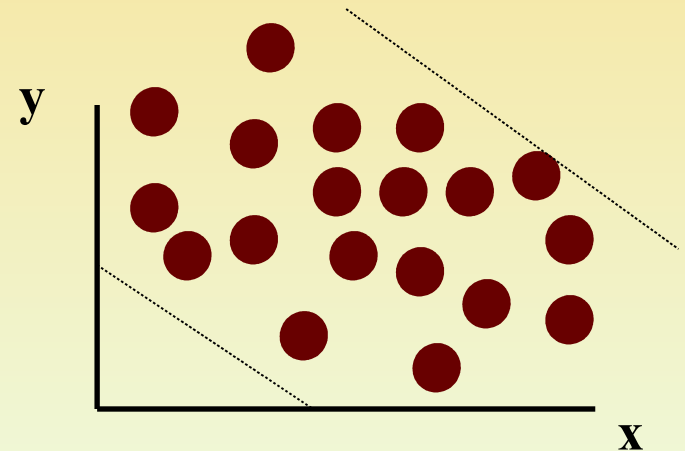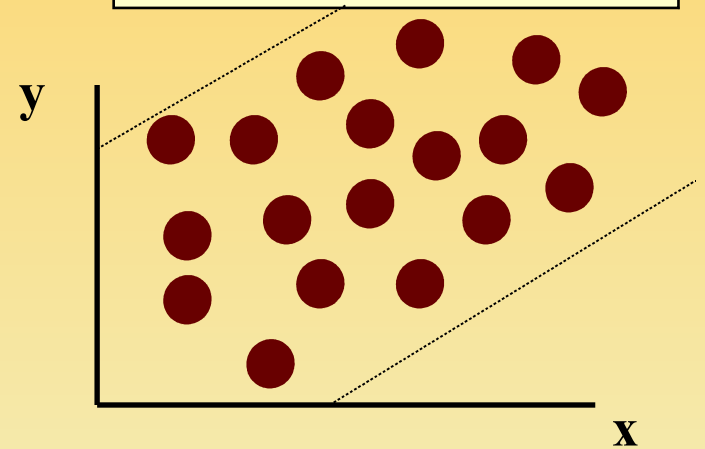
Scatter Plots

No relationship
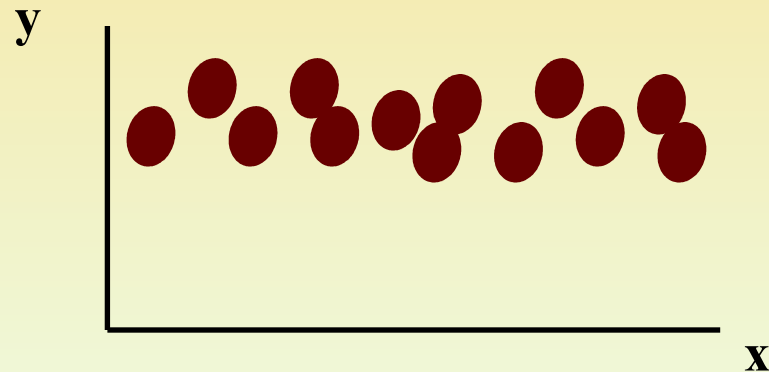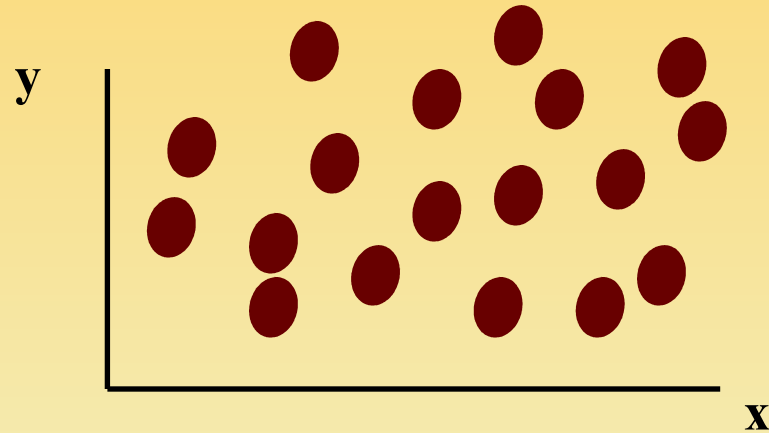
Correlation Coefficient

- 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

y

x

r = -1

y

x

r = -.6

y

x

r = 0

y

x

r = +.3

y

x
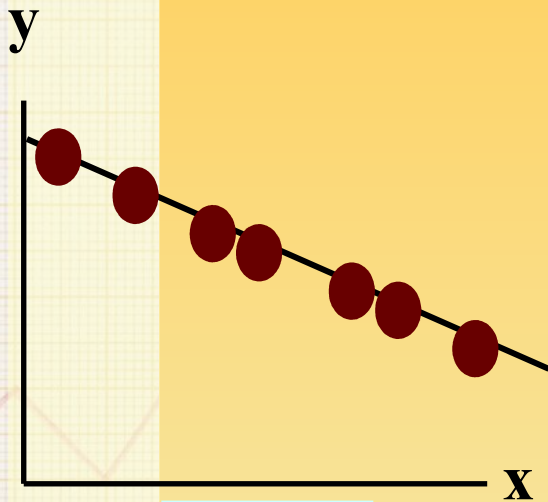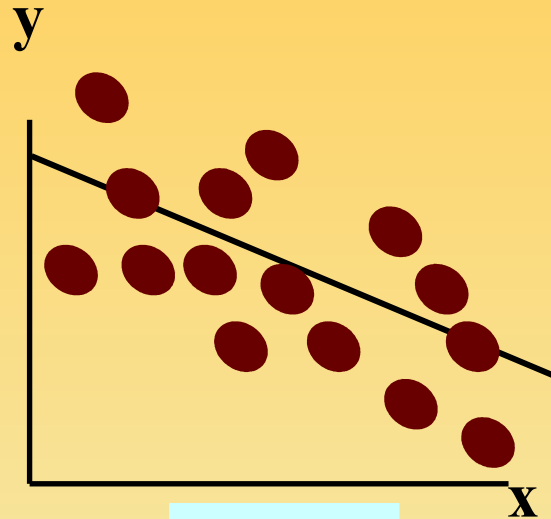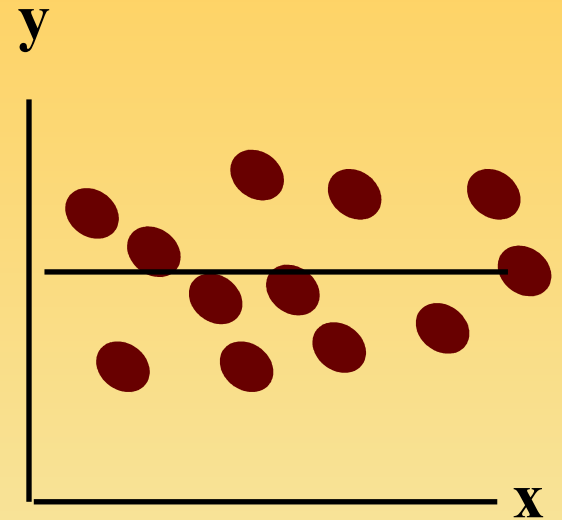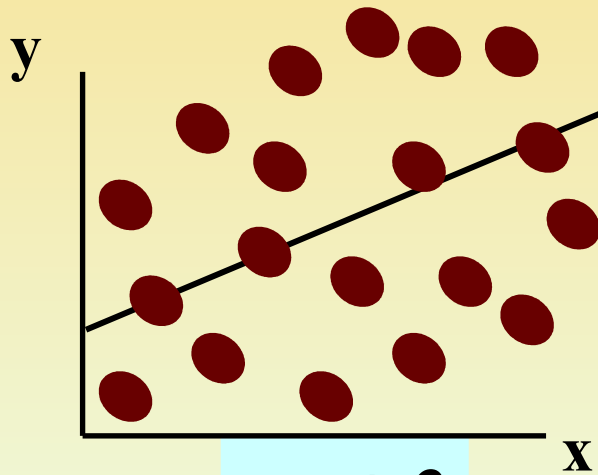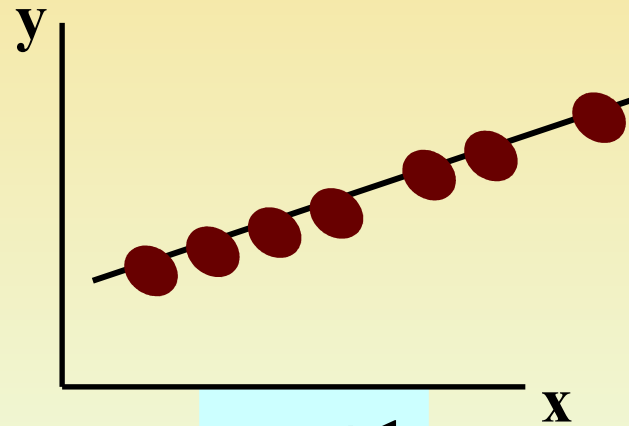
r = +1

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 (Σ)	270	325		1720		1948	1820
MEAN	54	65					

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

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		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)	dx²	Y	dy (Y-112)	dy²	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
∑X= 593	∑dx =41	∑dx²= 1727	∑Y= 1004	∑dy= 108	∑dy²= 3468	∑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 ΣX = 125 ΣX² = 650

ΣY = 100 ΣY² = 460 ΣXY = 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
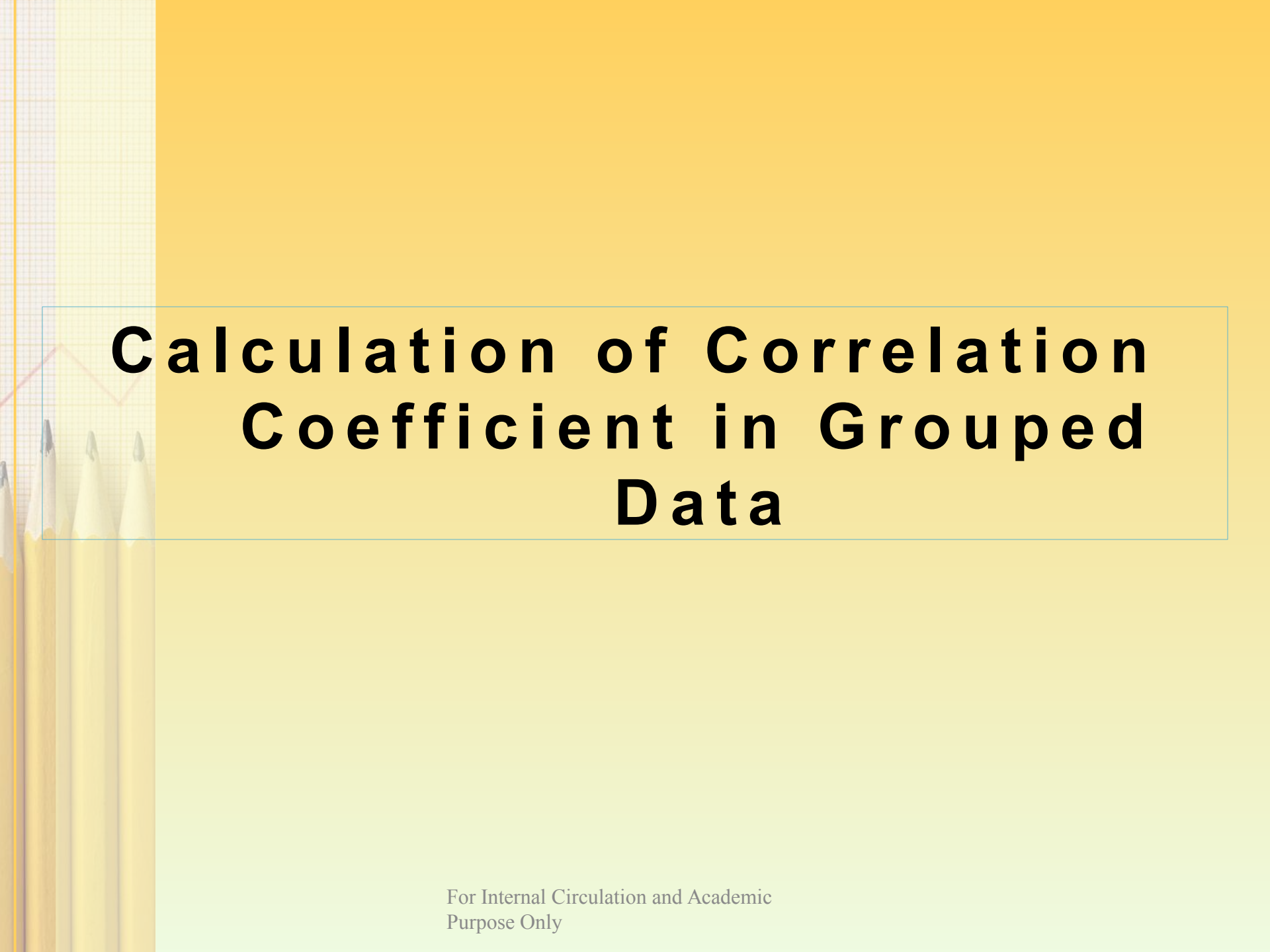
Comments:

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

F O R M U L A

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

Y	MV	dY Y-40	dX (X-35)	X 10-20 MV 15 -20 -2	20-30 25 -10 -1	30-40 35 0 0	40-50 45 10 1	50-60 55 20 2	Total	fdY	fdY²	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	ΣfdY = -8	ΣfdY² = 122	
			fdX	-18	-29	0	22	18	ΣfdX = -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)

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

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

References and Suggested Readings

Fundamentals of Statistics by S.C. Gupta

Statistics Methods by S.P.Gupta