1 set

~~8~~ a)

Correlation: Correlation means that the casual connection between two series or group of data.

Correlation analysis: Correlation analysis deals with the association between ~~two~~ variables.

There are three types of correlation

i) Simple correlation
ii) Partial correlation
iii) Multiple correlation.

Karl Pearson's Correlation co-efficient:

Karl Pearson correlation co-efficient is denoted by r,

$$r_{\bar{x}\bar{y}} \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2 \Sigma(y_i-\bar{y})^2}}$$

$$r_{xy} = r = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\Sigma x_i^2 - n\bar{x}^2)\ (\Sigma y_i^2 - n\bar{y}^2)}} = \text{———}$$

$-1 \le r \le +1$

~~r lies between -1 to 1.~~

$= \frac{\cancel{\Sigma x_i y_i - \Sigma x_i \Sigma y_i}}{n}$

Important properties of correlation co-efficient:

i) The value of r lie between -1 to 1

ii) It measures the magnitude and direction of statistical relationship between two variables.

iii) It is a symmetrical function of x and y. That is $r_{xy} = r_{yx}$

iv) $r = 0$ indicates no linear relation between x and y

v) $r = +1$ indicates a perfect positive relationship between x and Y.

vi) $r = -1$ indicates a perfect negative relationship between x and Y.

Interpretation of value of r:

rpret and show graphically when

$r=+1$, $r=-1$, $r=0$

10).

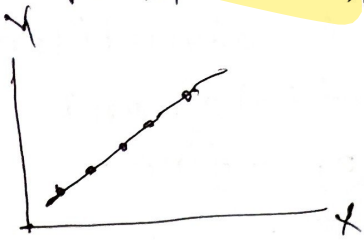

Figure-1: Scatter diagram showing $r=+1$

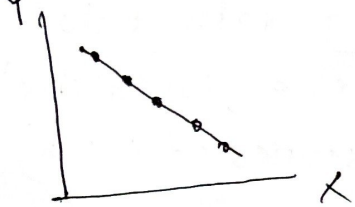

Figure-2 Scatter diagram showing $r=-1$

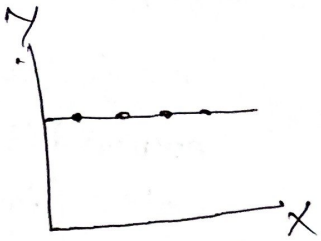

Figure-3 Scatter diagram showing $r=0$

Interpretation:

1. $r=+1$ indicates a perfect positive relationship between x and Y. In this case all the values of x and y fall in a straight line. (Figure-1).

2. $r=-1$ indicates a perfect negative relationship x and y. In this case all the values of x and y fall in a straight line. (Figure-2)

3. $r=0$, means there is no linear relationship between the variables x and y are shown in figure-3.

Math:

The following data relate to advertising expenditure (in lakhs of taka) and sales (in scores of taka) of a firm.

Advertising expenditure (in lakhs of TK)	10	12	15	20	23
Sales (in scores of TK)	14	17	23	21	25

i) Compute the correlation co-efficient between advertising expenditure and sales.

ii. Comment on the value of r.

Soln:

Advertising expenditure X	Sale: Y	x_i^2	y_i^2	$\Sigma x_i y_i$
10	14	100	196	140
12	17	144	289	204
15	23	225	529	345
20	21	400	441	420
23	25	529	525	575
$\Sigma x_i = 80$	$\Sigma y_i = 100$	$\Sigma x_i^2 = 1396$	$\Sigma y_i^2 = 2080$	$\Sigma x_i y_i = 1684$

$\bar{x} =$

We know, correlation co-efficient,

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

(P.T.O)

$$r = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}}$$

$$= \frac{1680 - 5 \times 16 \times 20}{\sqrt{(1396 - 5 \times 16^2)(2080 - 5 \times 20^2)}}$$

$$= \frac{80}{\sqrt{116 \times 80}}$$

$$r = \frac{80}{96.33} = .83$$

$\sum x_i = 80$

$\bar{x} = \frac{\sum x_i}{n}$

$= \frac{80}{5}$

$= 16$

$\sum y_i = 100$

$\bar{y} = \frac{\sum y_i}{n}$

$= \frac{100}{5}$

$= 20$

here $n = 5$

<u>Comment:</u> there is a strong positive correlation between expenditure and sale.

Rank Correlation: Suppose $1, 2, \ldots, n$ are assigned to the x observations in order of magnitude and similarly to the y observations. Then the simple correlation co-efficient between the two sets of ranks is called spearman's rank-correlation co-efficient. It is denoted by R.

The formula for computing the spearman's rank correlation co-efficient is

$$R = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

$$\therefore -1 \leq R \leq 1$$

Example: A social scientist wants to see whether there is any association between the intelligence and beauty among female students. To varify this, he randomly selected 6 female students from a class. The scores on intelligence and beauty are found as follows.

Student	A	B	C	D	E	F
Scores on Intelligence	80	75	90	70	65	60
Scores on beauty	65	70	60	75	85	80

Compute rank co-relation and comment.

Soln: we know, spearman's rank correlation co-efficient is

$$R = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}$$

Let us start ranking from the ~~heigl~~ highest value for both ~~valu~~ variables are shown in the given table. $[d_i^2 = (R_1 - R_2)^2]$

Student	Scores on Intelligence: X	Scores on beauty: Y	Rank on intelligence: R_1	Rank on beauty: R_2	$d^2 = (R_1 - R_2)^2$
A	80	65	2	5	9
B	75	70	3	4	1
C	90	60	1	6	25
D	70	75	4	3	1
E	65	85	5	1	16
F	60	80	6	2	16

$n = 6$

$\Sigma d_i^2 = 68$

$$\therefore R = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 68}{6 \times (6^2-1)}$$

$$= 1 - \frac{6 \times 68}{6 \times 35} = -.94$$

Comment: there exists almost perfect negative relationship between the intelligence and beauty

* Regression Analysis:

Regression analysis is a statistical technique for investigating and modeling the relationship between variables with a vision to predict the value of dependent variable for some known value of independent variable.

objective/ ~~pourp~~ purposes of regression analysis:

i. Establish a functional relationship between dependent and independent variables.

ii. predic the value of dependent variable for a given of independent variable.

Simple linear regression model:

A regression model that ~~contains~~ involves one independent variable is called simple linear regression model.

Example: $Y = B_0 + B_1 x + \varepsilon$

where B_0 = intercept

B_1 = slope.

ε = random error component

Multiple linear regression model:

A regression model that involves more than one independent variable is called multiple regression model.

Example: $Y = B_0 + B_1 x_1 + B_2 x_2 + \cdots + B_k x_k$

Dependent variable	Independent variable
Response variable	Regressor Variable
Predictand "	Explanatory "

Some situations where regression analysis are appropriate.

- A company might wish to improve its marketing process. After collecting data on the demand for a product, the product price and the advertising expenditure in promoting the product the company might use regression analysis to develop an equation to predic the future demand on the basis of price and advertising.

- An Agronomist may be interested to study the dependence of paddy on temperature, rainfall, amount of fertilizer and soil fertility. A multiple regression seems appropriate for this purpose to forecast the average yeild on the basis of explanatory variables.

* the following data give the test scores made by nine salesmen during the last year of a big departmental store.

~~Test score:~~

Test scores: y	
sales (in lakh taka): x	

The least squares method:

The least squares method is a statistical procedure to find the best fit for a set of data set. This method minimize the sum of squared errors.

The least squares parameters are

$$\text{slope} = \hat{B}_1 = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2}$$

$$= \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2}$$

$$\text{Intercept} = \hat{B}_0 = \bar{y} - \hat{B}_1\bar{x}$$

so, the straight line fit is

$$\hat{y} = \hat{B}_0 + \hat{B}_1 x$$

Example: the following data give the test scores and sales made by nine salesmen during the last year of a big departmental store.

Test scores: y	14	19	24	21	26	22	15	20	19
Sales (in lakh Taka): x	31	36	48	37	50	45	33	41	39

i) Find the regression line of test scores on sales

ii) Find the test score when the sales is ~~taka. 40 ta~~ tk. 40 Lakh.

Soln: The fitted regression line is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

~~Sales~~ Sales (x)	Test scores (y)	$x_i y_i$	x_i^2	y_i^2
31	14	434	961	196
36	19	684	1296	361
48	24	1152	2304	576
37	21	777	1369	441
50	26	1300	2500	676
45	2	990	2025	484
33	15	495	1089	225
41	20	820	1681	400
39	19	741	1521	361
$\Sigma x_i = 360$	$\Sigma y_i = 180$	$\Sigma x_i y_i = 7393$	$\Sigma x_i^2 = 14746$	$\Sigma y_i^2 = 3720$

here, $n = 9$

$\cancel{\bar{x} = \frac{\Sigma x_i}{n} = \frac{180}{9} = 20}$; $\cancel{\bar{y} = \frac{\Sigma y_i}{n} = \frac{360}{9} = 4}$

$\cancel{\bar{y} = \frac{\Sigma y_i}{n} = \frac{360}{9}}$

$$\text{slope} = \hat{B}_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

$$= \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2}$$

$$= \frac{7393 - 9 \times 20 \times 40}{14746 - 9 \times 20^2}$$

$$= \frac{196}{14746 - 14400}$$

$$= \frac{196}{346} = .566$$
$$= .57$$

$n = 9$

$\bar{x} = \frac{\Sigma x_i}{n}$

$= \frac{360}{9}$

$= 40$

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

$= \frac{180}{9}$

$= 20$

20
20
400

intercept $= \hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x}$

$= 20 - .57 \times 40$

$= 20 - 22.8$

$= -2.8$

so the fitted line is

i) $\hat{y} = \hat{B}_0 + \hat{B}_1 x$

$= -2.8 + .57x$

(ii) when $x = 40$, then

$\hat{y} = -2.8 + .57 \times 40$

$= 20$

H.T

The following table gives the ages and blood pressure of 7 women:

Age: x	56	42	35	46	49	70	62
Blood pressure: y	147	124	117	127	145	155	158

i) Calculate correlation co-efficient.

ii) Find the fitted regression line of blood pressure on age.

iii) Estimate the value of blood pressure when age is 48 years.

Residual: The residual is defined as the difference between observed value (y_i) and fitted value ($\hat{y}_i$)

Residual are denoted by

$$e_i = y_i - \hat{y}_i$$