Rank correlation

In many cases quantitative measurement is not possible because they are in qualitative form. for eg. we cannot measure beauty or intelligence quantitatively.

But it is possible to rank the individuals in some order. The correlation coefficient obtained from the ranks so obtained is called rank correlation.

That is, rank correlation is the correlation obtained from ranks.

Spearman's Rank correlation coefficient.

Spearman has devised a formula known as spearman's rank correlation coefficient to find the correlation coefficient from the rank. According to spearman's method, the formula for rank correlation coefficient is $1 - \frac{6\Sigma D^2}{n(n^2-1)}$, where D is the difference between ranks, and n, the number of items.

Prob(1) The ranking of 10 individuals at the start and at the finish of a course of a training are as follows,

individuals:	A	B	C	D	E	F	G	H	I	J
Rank before:	1	6	3	9	5	2	7	10	8	4
Rank after:	6	8	3	2	7	10	5	9	4	1

Calculate Spearman's rank correlation coefficient.

Rank before	Rank after	Rank difference (D)	D^2
1	6	5	25
6	8	2	4
3	3	0	0
9	2	7	49
5	7	2	4
2	10	8	64
7	5	2	4
10	9	1	1
8	4	4	16
4	1	3	9
			$\Sigma D^2 = 176$

Rank correlation coefficient $= 1 - \frac{6\Sigma D^2}{n(n^2-1)}$

$= 1 - \frac{6 \times 176}{10(10^2-1)}$

$= 1 - \frac{1056}{990} = 1 - 1.07$

$= -0.07$

Pool 2 Ten competitors in a beauty contest are ranked by 3 judges in the following order.

First judge: 1 6 5 10 3 2 4 9 7 8

second judge: 3 5 8 4 7 10 2 1 6 9

Third judge: 6 4 9 8 1 2 3 10 5 7

Use correlation coefficient to discuss which pair of judges have nearest approach to common tastes in beauty.

J_1	J_2	J_3	$D_1^2=(J_1-J_2)^2$	$D_2^2=(J_1-J_3)^2$	$D_3^3(J_2-J_3)^2$
4	3	6	4	25	9
6	5	4	1	4	1
5	8	9	9	16	1
10	4	8	36	4	16
3	7	1	16	4	36
2	10	2	64	0	64
4	2	3	4	1	1
9	1	10	64	1	81
7	6	5	1	4	1
8	9	7	1	1	4
			$\Sigma D_1^2=200$	$\Sigma D_2^2=60$	$\Sigma D_3^2=214$.

Rank correlation coeff b/w J_1 & J_2 $= \dfrac{1-6\Sigma D_1^2}{n(n^2-1)}$

$= 1-\dfrac{6\times200}{10(100-1)}$

$= 1-1.21 = -0.21$

Rank correlation coeff b/w J_1 & J_3 $= 1-\dfrac{6\Sigma D_2^2}{n(n^2-1)}$

$= 1-\dfrac{6\times60}{10(100-1)}$

$= 1-0.364 = 0.636$

Rank correlation coeff b/w J_2 & J_3 $= 1-\dfrac{6\Sigma D_3^2}{n(n^2-1)}$

$= 1-\dfrac{6\times214}{10(100-1)} = 1-1.30$

$= -0.30$.

The rank correlation coefficient in the case of I & III judges is greater than the other two pairs. Therefore I & III judges have highest similarity of thought and have the nearest approach to common taste in beauty.

Prob(3) Find the rank correlation coefficient b/w Poverty & overcrowding from the table below.

Town	A	B	C	D	E	F	G	H	I	J
Poverty	17	13	15	16	6	11	14	9	7	12
Overcrowding	36	46	35	24	12	18	27	22	2	8

Town	Poverty		Overcrowding		D^2 $(R_1-R_2)^2$
	Values	Rank (R_1)	Values	Rank (R_2)	
A	17	1	36	2	1
B	13	5	46	1	16
C	15	3	35	3	0
D	16	2	24	5	9
E	6	10	12	8	4
F	11	7	8	7	0
G	14	4	27	4	0
H	9	8	22	6	4
I	7	9	2	10	1
J	12	6	8	9	9

$\Sigma D^2 = 44$

Rank correlation coefficient $= 1 - \frac{6\Sigma D^2}{n(n^2-1)}$

$= 1 - \frac{6 \times 44}{10(100-1)} = 1 - \frac{264}{990}$

$= 1 - 0.267 = 0.733$

Repeated ranks (Tie in ranks)

When the values repeat in one or both the series, we added a correction factor $\Sigma \frac{m^3 - m}{12}$ to ΣD^2.

The formula for rank correlation coefficient thus obtained is, $1 - \frac{6\left(\Sigma D^2 + \frac{m^3-m}{12}\right)}{n(n^2-1)}$

where m stands for number of times each value repeats in any series.

Prob(1) Obtain the rank correlation coefficient for the following data.

X:	68	64	75	50	64	80	75	40	55	64
Y:	62	58	68	45	81	60	68	48	50	70

X	rank of x	Y	rank of Y	Rank difference (D)	D^2
68	4	62	5	1	
		58	7	1	
64	6	68	3.5	1	
75	2.5	45	10	1	
50	9	81	1	25	
64	6	60	6	25	
80	1	68	3.5	1	
75	2.5			1	
40	10	48	9	1	
		50	8	0	
55	8	70	2	16	
64	6				

$\Sigma D^2 = 72$

75 occurs 2 times $\therefore m = 2$; $m^3 - m = 2^3 - 2 = 6$

64 occurs 3 times $m = 3$; $m^3 - m = 3^3 - 3 = 24$

68 occurs 2 times $m = 2$ $m^3 - m = 2^3 - 2 = 6$.

Total $m^3 - m = \Sigma m^3 - m = 36$

$$\text{Rank Correlation coeff} = 1 - \frac{6\left(\Sigma D^2 + \frac{m^3 - m}{12}\right)}{n(n^2-1)}$$

$$= \frac{6 \times \left(72 + \frac{36}{12}\right)}{10(100-1)} = 6 \times \frac{75}{990} = \underline{\underline{545}}$$

Assigning ranks when item is repetitive

When a value repeats, rank is the average of ranks due for all of them if they are different.

Uses of Correlation

1. It helps to study the associations between 2 variables.
2. Correlation measures degree of relationships b/w 2 variables.
3. From the correlation coefficient, we can develop a measure called probable error and this measure indicates whether the correlation is significant or not.
4. Correlation ~~item~~ analysis helps to estimate future values.

Regression

Regression analysis means estimation or Prediction of the unknown value of one variable from the known value of the other variable.

Dependent & independent variables

In regression analysis there are two types of variables. The variable whose value is to be predicted is called dependent variable and the variable which is used for prediction is called independent variable.

Linear and non-linear regression

On the basis of proportion of changes in the variables, the regression can be classified into linear & non-linear.

If the given bivariate data are plotted on a graph, the points so obtained on the scatter diagram will more or less concentrate around a curve called 'curve of regression'.

If the regression curve is a straight line, say AB then there is linear regression b/w the variables under study. This relation can be expressed in the form $y = a + bx$.

If the curve of regression is not a straight line, then the regression is called non-linear regression.

Line of best fit (Regression line)

When the given bivariate data are plotted on a graph, we get a scatter diagram. If the points of the scatter diagram concentrate around a straight line, that line is called the line of best fit or regression line.

Methods of drawing regression lines

Regression lines can be drawn by two methods ① free hand curve method ② method of least squares.

free hand curve method - This is an easy method for obtaining regression line. The original data are plotted on a graph paper. The original data when plotted on a graph, gives a worm like curve, that depicts the general tendency of the data. Independent variable is taken along the X-axis & ~~dependent~~ dependent variable along the ~~vertical~~ Y-axis.

Method of least squares (curve fitting)

This method uses the principle of least squares to draw the regression lines.

The principle of least squares is that principle which states that the line of

best fit should be drawn in such a manner that the sum of the squares of difference b/w the known values of the dependent variable and the corresponding values of it, obtained from the line of best fit, should be least.

There are 2 lines of regression.

- while estimating the value of y for any given value of x, we take y as dependent variable and x as independent variable. Then we get line of regression of y on x.
- Similarly for estimating x for any given value of y, we use regression of x on y. Here x is dependent variable and y is independent variable. Thus there are 2 regression lines.

Regression equations

Regression equations are the equations of the regression lines. It is mathematical relation between dependent & independent variables.

The two regression equations,

(1) Regression equation of y on x
(2) Regression equation of x on y.

are not reversible or interchangable.

The two regression equations are,

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

(Regression equation of y on x)

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

(Regression equation of x on y).

Prob (1) From the following data of the age of husband and age of wife, form the two regression equations and calculate the husband's age when wife's age is 16.

Husband's age: 36 23 27 28 29 29 30 31 33 35
Wife's age: 29 18 20 22 27 21 29 27 29 28

Also find the age of wife when husband's age is 40.

x	y	xy	x^2	y^2
36	29	1044	1296	841
23	18	414	529	324
27	20	540	729	400
28	22	616	784	484
28	27	756	784	729
29	21	609	841	441
30	29	870	900	841
31	27	837	961	729
33	29	957	1089	841
35	28	980	1225	784
300	250	7623	9138	6414

$\bar{x} = \frac{\Sigma x}{n} = \frac{300}{10} = 30$ & $\bar{y} = \frac{\Sigma y}{n} = \frac{250}{10} = 25$

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x \times \Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{10 \times 7623 - (300 \times 250)}{10 \times 9138 - (300)^2}$$

$$= \frac{76230}{91380} - \frac{75000}{90000} = \frac{1230}{1380} = 0.89$$

$$b_{xy} = \frac{n\Sigma xy - (\Sigma x \times \Sigma y)}{n\Sigma y^2 - (\Sigma y)^2} = \frac{10 \times 7623 - (300 \times 250)}{10 \times 6414 - (250)^2}$$

$$= 0.75.$$

Equation of the line of regression of y on x is,

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

$$y - 25 = 0.89(x - 30)$$

$$y = 0.89x - 26.7 + 25$$

$$y = 0.89x - 1.7$$

equation of the line of regression of x on y is,

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

$$x - 30 = 0.75(y - 25)$$

$$x = 0.75y - 18.75 + 30$$

$$x = 0.75y + 11.25$$

(i) when wife's age(y) is 16, husband's age (x) is obtained by putting $y = 16$, on the equation of x on y,

$\therefore x = 0.75 \times (16) + 11.25 = 23.25$

Husband's age = 23.25

(ii) when husband's age(x) is 40, wife's age (y) is obtained by putting $x = 40$ in the

equation of y on x,

$$y = 0.89 \times 40 - 1.7 = 33.9.$$

$\therefore$ Wife's age $= \underline{33.9}$

Relation b/w Correlation Coefficient & Regression coefficient

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \quad \& \quad b_{xy} = r\frac{\sigma_x}{\sigma_y}.$$

(Problem) You are given the following data,

	x	y
A.M	36	85
S.D	11	8

Correlation coefficient b/w x & $y = 0.66$.

(i) Find the two regression equations

(ii) Estimate the value of x when $y = 75$.

Given $\bar{x} = 36$, $\bar{y} = 85$, $\sigma_x = 11$, $\sigma_y = 8$, $r = 0.66$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{0.66 \times 8}{11} = 0.48$$

$$b_{xy} = \frac{0.66 \times 11}{8} = 0.91$$

Equation of regression of y on x is,

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

$$y - 85 = 0.48(x - 36) = 0.48x - 17.28$$

$$\therefore y = 0.48x + 67.72$$

Equation of regression of x on y is,

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

$$x - 36 = 0.91(y - 85) = 0.91y - 77.35$$

$$x = 0.91y - 41.35$$

when $y = 75$,

$$x = 0.91 \times 75 - 41.35 = 26.9$$

How to get b_{xy} & b_{yx} from regression equn

When the regression equation of y on x is expressed in the form $y = ax + b$ then 'a' is b_{yx}.

Similarly when the regression equation y x on y is expressed in the form $x = cy + d$ 'c' is b_{xy}.

Prob ① Find b_{yx} if $2x + 4y - 5 = 0$ is the equn of y on x.

$$2x + 4y - 5 = 0.$$

$$4y = -2x + 5$$

or $y = -0.5x + 1.25$

here $a = -0.5$ $\therefore b_{yx} = -0.5$

② Find b_{xy} if $3x + 2y + 4 = 0$ is the equn of x on y.

$$3x + 2y + 4 = 0.$$

$$3x = -2y - 4$$

$x = -.67y - 1.33$

Hence $r = -0.67$

How to get $\bar{x}$ & $\bar{y}$ from the Regression equations?

Solve the two Regression equations. The values of x & y obtained are means of x & y.

Prob Find the means of variables x & y, given the following.

Regression of y on x : $2y - x - 50 = 0$

Regression of x on y : $3y - 2x - 10 = 0$.

The means of x & y are obtained by solving the 2 Regression equations.

$$2y - x - 50 = 0 \quad \text{— (1)}$$
$$3y - 2x - 10 = 0. \quad \text{— (2)}$$

(1) × 2.

$$4y - 2x - 100 = 0 \quad \text{— (3)}$$
$$3y - 2x - 10 = 0. \quad \text{— (2)}$$

(3) − (2)

$$y - 90 = 0$$
$$y = 90.$$

Substituting in (1)

$$2 \times 90 - x - 50 = 0.$$
$$180 - x - 50 = 0$$
$$x = 130$$

Hence $\bar{x} = 130$ & $\bar{y} = 90$

How to identify the two regression equations?

By supposing one of the equations as the regression of y on x and the other as x on y, we can obtain the Regression Coefficients. If the Product of these two is numerically not greater than 1, then our Supposition is true. If ~~the~~ $b_{yx} \times b_{xy}$ is numerically greater than 1, our Supposition is wrong.

Prob Out of the 2 lines given by

$x + 2y - 5 = 0$ & $2x + 3y - 8 = 0$.

which one is the regression line of x on y?

$x + 2y - 5 = 0$ ——— (1)

$2x + 3y - 8 = 0$ ——— (2).

Let us assume that eqn(1) as the regression of x on y and eqn(2) as the Regression of y on x,

(1) becomes,

$x = -2y + 5$ $\therefore b_{xy} = -2$.

(2) becomes,

$y = -\frac{2}{3}x + \frac{8}{3}$ $\therefore b_{yx} = -2/3$.

$b_{xy} \times b_{yx} = -2 \times -\frac{2}{3} = \frac{4}{3} = 1.33 > 1$

$\therefore$ Our assumption is wrong.

So (1) is the line of Regression of y on x and (2) is the line of Regression of x on y.

Note

$$b_{yx} \times b_{xy} = r\frac{\sigma_y}{\sigma_x} \times r\frac{\sigma_x}{\sigma_y} = r^2.$$

$$r = \sqrt{b_{yx} \cdot b_{xy}}$$

H.W (1) The ranking of 10 students in two subjects A & B are as follows. find the rank correlation coefficient.

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

(2) Judges x & y given the marks of 10 candidates in a beauty contest. find the rank correlation coefficient.

Candidate:	A	B	C	D	E	F	G	H	I	J
Judge x:	50	60	70	65	80	85	90	92	40	96
Judge y:	60	70	75	60	80	82	86	90	50	95

(3) The following data given below obtain the regression equation of y on x and regression equation of x on y.

Income (x):	120	90	83	150	130	140	110	95	75	105
Expenditure (y):	40	36	40	45	40	44	45	38	50	35

(4) Given the following data find what will be the probable yield when the rainfall is 29.

mean	rainfall	Production (yield)
S.D	5	40
	3	60

Correlation b/w rainfall & production is 0.8

(5) If the two regression equations are $5x - 4y + 20 = 0$ & $2x - 5y + 110 = 0$. Standard deviation of $x = 10$; obtain

(i) $\bar{x}$ & $\bar{y}$ (ii) r (3) σ_y