Spearman's Rank Correlation Coefficient

Rank correlation coefficient is useful for finding correlation between any two qualitative characteristics.
For example: Beauty, Honesty, and Intelligence etc., which cannot be measured quantitatively but can be arranged serially in order of merit or proficiency possessing the two characteristics.

Suppose we associate the ranks to individuals or items in two series based on order of merit, the Spearman's Rank correlation coefficient r is given by

$$\rho = 1 - \left[\frac{6 \sum d^2}{n(n^2 - 1)} \right]$$

Where, Σd^2 = Sum of squares of differences of ranks between paired items in two series, n = Number of paired items

Problem: In a quantitative aptitude test, two judges rank the ten competitors in the following order.

Competitor	1	2	3	4	5	6	7	8	9	10
Ranking of judge I	4	5	2	7	8	1	6	9	3	10
Ranking of judge II	8	3	9	10	6	7	2	5	1	4

Is there any concordance between the two judges ?

Solution: Let Rx: Ranking by Judge I and Ry: Ranking by Judge II The Spearman's rank correlation coefficient is given by

$$\rho = 1 - \left[\frac{6 \sum d^2}{n(n^2 - 1)} \right]$$

Where, $\Sigma d^2 = (R_x - R_y)^2$ and n = number of competitors.

R_x	R_y	$d = R_x - R_y$	d^2
4	8	-4	16
5	3	2	4
2	9	-7	79
7	10	-3	9
8	6	2	4
1	7	-6	36
6	2	4	16
9	5	4	16
3	1	2	4
10	4	6	36
		TOT	190

$$\rho = 1 - \left[\frac{6(190)}{10(100-1)}\right]$$

$$= 1 - 1.1515$$

$$= -0.1515$$

We say that there is low degree of negative rank correlation between the two judges.

Problem : Twelve recruits were subjected to selection test to ascertain their suitability for a certain course of training. At the end of training they were given a proficiency test. The marks scored by the recruits are recorded below:

Recruit	1	2	3	4	5	6	7	8	9	10	11	12
Selection Test Score	44	49	52	54	47	76	65	60	63	58	50	67
Proficiency Test Scrore	48	55	45	60	43	80	58	50	77	46	47	65

calculate rank correlation coefficient and comment on your result

Solution: Let selection test score be a variable X and proficiency test score be a variable Y. We associate the ranks to the scores based on their magnitudes. The spearman's rank correlation coefficient is given by

$$\rho = 1 - \left[\frac{6 \sum d^2}{n(n^2 - 1)} \right]$$

Where, $\Sigma d^2 = (R_x - R_y)^2$ = sum of squares of differences between the ranks of observations X and Y

n = number of recruits.

Given,

X	Y	R_x	R_y	$d = R_x - R_y$	d^2
44	48	12	8	4	16
49	55	10	6	4	16
52	45	8	11	-3	9
54	60	7	4	3	9

47	43	11	12	-1	1
76	80	1	1	0	0
65	58	3	5	-2	4
60	50	5	7	-2	4
63	77	4	2	2	4
58	46	6	10	-4	16
50	47	9	9	0	0
67	65	2	3	-1	1

From the table, we have,

$$\sum d^2 = 80,\ n = 12$$

$$\rho = 1 - \left[\frac{6(80)}{12(144-1)}\right]$$

$$= 1 - 0.2797$$

$$= 0.7203$$

We say that there is high degree of positive rank correlation between the scores of selection and proficiency tests.

Partial and Multiple Correlation

Let us say that we find a correlation between these two factors. That is, as the bank balance increases, cholesterol level also increases.

But this is not a correct relationship as Cholesterol level can also increase as age increases. Also as age increases, the bank balance may also increase because a person can save from his salary over the years.

Thus there is age factor which influences both cholesterol level and bank balance. Suppose we want to know only the correlation between cholesterol and bank balance without the age influence, we could take persons from the same age group and thus control age, but if this is not possible we can statistically control the age factor and thus remove its influence on both cholesterol and bank balance. This if done is called partial correlation.

If there are three variables X_1, X_2 and X_3 there will be three coefficients of partial correlation, each studying the relationship between two variables when the third is held constant. If we denote by $r_{12.3}$ i.e., the coefficient of partial correlation between X_1 and X_2 keeping X_3 constant, it is calculated as

$$r_{12\,.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - {r_{13}}^2}\sqrt{1 - r_{23}^2}}$$

$$r_{13\,.2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{1 - {r_{12}}^2}\sqrt{1 - r_{23}^2}}$$

$$r_{23\,.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{1 - {r_{12}}^2}\sqrt{1 - r_{13}^2}}$$

Problem: In a trivariate distribution , it is found that $r_{12} = 0.7, r_{13} = 0.61$ and $r_{23} = 0.4$. Find the partial correlation coefficients.

Answer:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^{\ 2}}\sqrt{1 - r_{23}^{2}}} = \frac{0.7 - (0.61)(0.4)}{\sqrt{1 - (0.61)^2}\sqrt{1 - (0.4)^2}} = 0.628$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r_{12}^{\ 2}}\sqrt{1 - r_{23}^{2}}} = 0.504$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1 - r_{12}^{\ 2}}\sqrt{1 - r_{13}^{2}}} = -0.048$$

2. Is it possible to get the following from a set of experimental data?
$r_{12} = 0.6, r_{13} = 0.5$ and $r_{23} = 0.8$

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} = \frac{0.6 - (-0.5)(0.8)}{\sqrt{1 - (0.5)^2}\sqrt{1 - (0.8)^2}} = 1.923$$

Since the value of $r_{12.3}$ is greater than one, there is some inconsistency in the given data.

Multiple Correlation

Sometimes in psychology we have certain factors which are influenced by large number of variables.

For instance academic achievement will be affected by intelligence, work habit, extra coaching, socio economic status, etc.

To find out the correlation between academic achievement with various other factors as mentioned above can be done by Multiple Correlation.

The coefficient of multiple correlation with three variables X_1, X_2 and X_3 are $R_{1.23}, R_{2.13}$and $R_{3.12}$, is the coefficient of multiple correlation related to X_1 as a dependent variable and X_2, X_3 as two independent variables and it can be expressed in terms of r_{12}, r_{23} and r_{13} as

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}},$$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{13}^2}},$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{12}^2}}$$

Example:

1. The following zero-order correlation coefficients are given:
$r_{12} = 0.98, r_{13} = 0.44$ and $r_{23} = 0.54$. Calculate multiple correlation coefficient treating first variable as dependent and second and third variables as independent.

Solution:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.98)^2 + (0.44)^2 - 2(0.98)(0.54)(0.44)}{1 - (0.54)^2}} = 0.986$$

2. From the following data, obtain $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$

X_1	2	5	7	11
X_2	3	6	10	12
X_3	1	3	6	10

Solution:

We need r_{12}, r_{13} and r_{23} which are obtained from the following table:

S. No	X_1	X_2	X_3	$(X_1)^2$	$(X_2)^2$	$(X_3)^2$	$X_1 X_2$	$X_1 X_3$	$X_2 X_3$
1	2	3	1	4	9	1	6	2	3
2	5	6	3	25	36	9	30	15	18
3	7	10	6	49	100	36	70	42	60
4	11	12	10	121	144	100	132	110	120
TOT	25	31	20	199	289	146	238	169	201

Now we get the total correlation coefficient r_{12}, r_{23} and r_{13}

$$r_{12} = \frac{N(\sum X_1 X_2) - (\sum X_1)(\sum X_2)}{\sqrt{\{N(\sum X_1^2) - (\sum X_1)^2\}}\sqrt{\{N(\sum X_2^2) - (\sum X_2)^2\}}}$$

$$r_{12} = 0.97$$

$$r_{13} = \frac{N(\sum X_1 X_3) - (\sum X_1)(\sum X_3)}{\sqrt{\{N(\sum X_1^2) - (\sum X_1)^2\}}\sqrt{\{N(\sum X_3^2) - (\sum X_3)^2\}}}$$

$$r_{13} = 0.99$$

$$r_{23} = \frac{N(\sum X_2 X_3) - (\sum X_2)(\sum X_3)}{\sqrt{\{N(\sum X_2^2) - (\sum X_2)^2\}}\sqrt{\{N(\sum X_3^2) - (\sum X_3)^2\}}}$$

$$r_{23} = 0.97$$

Now, we calculate $R_{1.23}$

We have, $r_{12} = 0.97$, $r_{13} = 0.99$ and $r_{23} = 0.97$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}}$$

$R_{1.23} = 0.99$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{13}^2}}$$

$R_{2.13} = 0.97$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{12}^2}}$$

$R_{3.12} = 0.99$

Regression

Definition: Regression is the measure of the average relationship between two or more variables in terms of the original units of data.

Regression Equation: The functional relationship of a dependent variable with one or more independent variable is called regression equation.
It is also called a prediction equation or estimating equation.

Note: The independent variable in regression analysis is called the "predictor" or "regressor" and the dependent variable is called the regressed variable.

Types of Regression:

- If there are only two variables under consideration, then the regression is called simple regression.
- If there are more than two variables under consideration, then the regression is called multiple regression.
- If there are more than two variables under consideration, and only the relation between two variables is established, after excluding the effect of the remaining variables, then the regression is called partial regression.
- If the relationship between x and y is non-linear, then the regression is a curvilinear regression.

There are certain guidelines for regression lines:

1) Use regression lines when there is a significant correlation to predict values.

2) Do not use if there is not a significant correlation.

3) Stay within the range of the data. For example, if the data is from 10 to 60, do not predict a value for 400.

Regression Equations (Linear Fit)

- Linear regression equation of y on x
- Linear regression equation of x on y

Equation of the Regression Line of Y on X

The regression line of Y on X is the best-fitting straight line for the observed pairs of values $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, based on the assumption that x is the independent variable and y is the dependent variable.

let the equation of the regression line of Y on X be assumed as

$y = ax + b.$ (1)

By the principle of least squares, the normal equations which give the values of a and b.

are $$\Sigma y_i = a \Sigma x_i + nb \quad (2)$$

and $$\Sigma x_i y_i = a \Sigma x^2{}_i + b \Sigma x_i \quad (3)$$

Dividing equation (2) by n, we get

$$\bar{y} = a \bar{x} + b \quad (4)$$

where $\bar{x} = E(X)$ and $\bar{y} = E(Y)$. (1)–(4) gives the required equation as

$$y - \bar{y} = a(x - \bar{x}) \tag{5}$$

Eliminating b between equations (2) and (3)

we get
$$a = \frac{n\Sigma x_i y_i - \Sigma x_i \cdot \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

$$= \frac{\frac{1}{n}\Sigma x_i y_i - \left(\frac{1}{n}\Sigma x_i\right)\cdot\left(\frac{1}{n}\Sigma y_i\right)}{\frac{1}{n}\Sigma x_i^2 - \left(\frac{1}{n}\Sigma x_i\right)^2}$$

or
$$a = \frac{E(XY) - E(X)\cdot E(Y)}{E\left(X^2\right) - E^2(X)} = \frac{Cov(X,Y)}{\sigma_x^2} \tag{6}$$

Using (6) in (5), we get the equation of the regression line of Y on X as

$$y - \bar{y} = \frac{Cov(X,Y)}{\sigma_x^2}(x - \bar{x}) \tag{7}$$

or
$$y - \bar{y} = \frac{r_{XY}\,\sigma_Y}{\sigma_X}(x - \bar{x}) \tag{8}$$

In a similar manner, assuming the equation of the regression line of X and Y as $x = ay + b$ and using the equations

$$\Sigma x_i = a\,\Sigma y_i + nb \text{ and } \Sigma x_i y_i = a\,\Sigma y^2{}_i + b\,\Sigma y_i,$$

we can get the equation of the regression line of X on Y as

$$x - \bar{x} = \frac{Cov(X,Y)}{\sigma_y^2}(y - \bar{y}) \tag{9}$$

or

$$x - \bar{x} = \frac{r_{XY}\,\sigma_X}{\sigma_Y}(y - \bar{y}) \tag{10}$$

Note 1. $\frac{r_{XY}\,\sigma_Y}{\sigma_X}$ *is called the regression coefficient of Y on X and denoted by* b_1 *or* b_{YX}. $\frac{r_{XY}\,\sigma_X}{\sigma_Y}$ *is called the regression coefficient of X on Y and denoted by* b_2 *or* b_{XY}.

2. *Clearly* $b_1\, b_2 = r_{XY}^2$, *i.e.,* r_{XY} *is the geometric mean of* b_1 *and* b_2.

$$\therefore \qquad r_{XY} = \pm\sqrt{b_1\, b_2}$$

The sign of r_{XY} *is the same as that of* b_1 *or* b_2, *as* $b_1 = r_{xy}\frac{\sigma_Y}{\sigma_X}$ *and*

$b_2 = r_{XY}\frac{\sigma_Y}{\sigma_X}$ *have the same sign as* r_{XY} ($\because$ σ_X *and* σ_Y *are positive).*

Also $$\frac{b_1}{b_2} = \frac{\sigma_Y^2}{\sigma_X^2}$$

3. *When there is perfect linear correlation between X and Y, viz., when* $r_{XY} = \pm 1$*, the two regression lines coincide.*
4. *The point of intersection of the two regression lines is clearly the point whose co-ordinates are* $(\bar{x}, \bar{y})$.
5. *When there is no linear correlation between X and Y, viz., when* $r_{XY} = 0$*, the equations of the regression lines become* $y = \bar{y}$ *and* $x = \bar{x}$*, which are at right angles.*

Problem 1: For the following data, find the regression line of y on x.

x	1	2	3	4	5	8	10
y	9	8	10	12	14	16	15

Solution 1: $\overline{x} = \frac{\sum x_i}{n} = \frac{33}{7} = 4.714$ and $\overline{y} = \frac{\sum y_i}{n} = \frac{84}{7} = 12.$

x	y	xy	x^2
1	9	9	1
2	8	16	4
3	10	30	9
4	12	48	16
5	14	70	25
8	16	128	64
10	15	150	100

$$b_{yx} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = 0.867$$

The regression equation of y on x is

$$\begin{aligned} y - \overline{y} &= b_{yx}(x - \overline{x}) \\ \Rightarrow y - 12 &= 0.867(x - 4.714) \\ \Rightarrow y &= 0.867x + 7.9129. \end{aligned}$$

Problem 2: From the following data, fit two regression equations by finding actual means (of x and y), i.e., by the actual mean method.

x	1	2	3	4	5	6	7
y	2	4	7	6	5	6	5

Solution 2: $\overline{x} = \frac{\sum x_i}{n} = \frac{28}{7} = 4$ and $\overline{y} = \frac{\sum y_i}{n} = \frac{35}{7} = 5$.

x	y	$X = x - \overline{x}$	$Y = y - \overline{y}$	X^2	Y^2	XY
1	2	-3	-3	9	9	9
2	4	-2	-1	4	4	2
3	7	-1	2	1	1	-2
4	6	0	1	0	0	0
5	5	1	0	1	1	0
6	6	2	1	4	4	2
7	5	3	0	9	9	0
28	35	0	0	28	16	11

$$b_{yx} = \frac{\sum X_i Y_i}{\sum X_i^2} = \frac{11}{28} = 0.3928$$

$$b_{xy} = \frac{\sum X_i Y_i}{\sum Y_i^2} = \frac{11}{16} = 0.6875$$

The regression equation of y on x is

$$\begin{aligned} y - \overline{y} &= b_{yx}(x - \overline{x}) \\ \Rightarrow y - 5 &= 0.3928(x - 4) \\ \Rightarrow y &= 0.393x + 3.428. \end{aligned}$$

The regression equation of x on y is

$$\begin{aligned} x - \overline{x} &= b_{xy}(y - \overline{y}) \\ \Rightarrow x - 4 &= 0.6875(y - 5) \\ \Rightarrow x &= 0.688y + 0.56. \end{aligned}$$