Correlation

- Quantitative measure of the relationship between two variables.
- If change in one variable affects a change in the other variable, the variables are said to be correlated.
- If the increase in one variable results in corresponding increase in the other variable, the correlation is said to be positive or direct
- If the decrease in one variable results in the corresponding decrease in the other variable the correlation is said to be positive or direct
- However if increase in one variable results in the decrease in other variable, or decrease in one variable results in the increase in the other variable, the correlation is said to be diverse or negative.

§ Karl Peason's coefficient of Correlation

As a measure of intensity or degree of linear relationship between two variables, Karl Pearson (1867–1936), a British Biometrician, developed a formula called Correlation coefficient

Correlation coefficient between two variables X and Y usually denoted by $r(X, Y)$ or simply r_{XY}, is defined as

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (1)$$

If (x_i, y_i), $i = 1, 2, \ldots n$ is the bivariate distribution, then

$$\sigma_{XY} = Cov(X,Y) = E\left[\{X - E(X)\}\{Y - E(Y)\}\right]$$

$$= \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y}) = \mu_{11}$$

$$\sigma_X^2 = E(X - E(X))^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$$

$$\sigma_Y^2 = E(Y - E(Y))^2 = \frac{1}{n}\sum (y_i - \bar{y})^2$$

We can use another form of formula ① for computational purpose, as

$$\begin{aligned} \mathrm{cov}(X,Y) &= \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n}\sum (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x}\bar{y}) \\ &= \frac{1}{n}\sum x_i y_i - \bar{y}\,\frac{1}{n}\sum x_i - \bar{x}\,\frac{1}{n}\sum y_i + \frac{1}{n}\cdot n\bar{x}\bar{y} \\ &= \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y} \\ &= \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y} \end{aligned}$$

$$\sigma_X^2 = \frac{1}{n}\sum (x_i - \bar{x})^2 = \frac{1}{n}\sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \frac{1}{n}\left(\sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2\right)$$

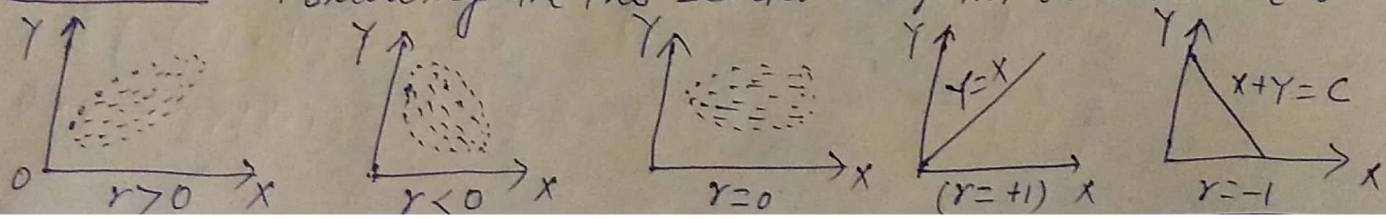
$$= \frac{1}{n}\left(\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right) = \frac{1}{n}\left(\sum x_i^2 - n\bar{x}^2\right)$$

$$= \frac{1}{n}\sum x_i^2 - \bar{x}^2$$

Similarly $\sigma_Y^2 = \frac{1}{n}\sum y_i^2 - \bar{y}^2$

$$r(X,Y) = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n}\sum x_i^2 - \bar{x}^2\right)\left(\frac{1}{n}\sum y_i^2 - \bar{y}^2\right)}}$$

Remark Following are the scatter diagram for different r



(2) $r(X, Y)$ is a measure of linear relationship between X and Y. For a non-linear relationship, however it is not very suitable.

(3) Karl Pearson's correlation coefficient is also called Product-moment correlation coefficient.

(4) Correlation coefficient cannot exceed unity numerically. It always lies between -1 and $+1$. If $r = +1$, the correlation is perfect and positive and if $r = -1$, the correlation is perfect and negative.

$$\boxed{-1 \leq r \leq 1}$$ (Prove it)

(5) Correlation coefficient is independent of change of origin and scale. (Prove it)

(6) If X and Y are random variables and a, b, c, d are any numbers provided only $a \neq 0$, $c \neq 0$, then

$$r(aX + b, cY + d) = \frac{ac}{|ac|} r(X, Y)$$

(7) Two independent variables are uncorrelated, but two uncorrelated variables need not necessarily be independent.

Note: The points above are important for MCQ

Q(1) Calculated the correlation coefficient for the following height (in inches) of father (X) and their sons (Y):

X:	65	66	67	67	68	69	70	72
Y:	67	68	65	68	72	72	69	71

Solution

X	Y	X^2	Y^2	X·Y
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
Total 544	552	37028	38132	37560

$\overline{X} = \frac{1}{n}\Sigma X = \frac{544}{8} = 68$

$\overline{Y} = \frac{1}{n}\Sigma Y = \frac{552}{8} = 69$

$$r(X,Y) = \frac{\frac{1}{n}\Sigma XY - \overline{X}\,\overline{Y}}{\sqrt{\left(\frac{1}{n}\Sigma X^2 - \overline{X}^2\right)\left(\frac{1}{n}\Sigma Y^2 - \overline{Y}^2\right)}}$$

$$= \frac{\frac{1}{8} \times 37560 - 68 \times 69}{\sqrt{\left\{\frac{37028}{8} - (68)^2\right\}\left\{\frac{38132}{8} - (69)^2\right\}}}$$

$$= \frac{4695 - 4692}{\sqrt{(4628.5 - 4624)(4766.5 - 4761)}}$$

$$= \frac{3}{\sqrt{4.5 \times 5.5}} = \boxed{0.603}$$

Note: You need to be very precise in computation as the final answer is something evaluator will look for

Alternate method (It is based on the fact that correlation coefficient is independent of change of origin)

X	Y	$U = X - 68$	$V = Y - 69$	U^2	V^2	UV
65	67	−3	−2	9	4	6
66	68	−2	−1	4	1	2
67	65	−1	−4	1	16	4
67	68	−1	−1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
		$\Sigma U = 0$	$\Sigma V = 0$	$\Sigma U^2 = 36$	$\Sigma V^2 = 44$	$\Sigma UV = 24$

Note: Here we notice that ΣU is the sum of deviation about mean, so it is zero. It is an important observation to check error in computation

$\bar{U} = \frac{1}{n}\sum U = 0$, $\bar{V} = \frac{1}{n}\sum V = 0$

$$r = \frac{\frac{1}{n}\sum UV - \bar{U}\bar{V}}{\sqrt{\left(\frac{1}{n}\sum U^2 - \bar{U}^2\right)\left(\frac{1}{n}\sum V^2 - \bar{V}^2\right)}} = \frac{\frac{1}{8}\cdot 24 - 0\times 0}{\sqrt{\left(\frac{1}{8}\times 36 - 0^2\right)\left(\frac{1}{8}\times 44 - 0^2\right)}}$$

$$= \frac{3}{\sqrt{\frac{36}{8}\times\frac{44}{8}}} = \frac{3\times 8}{6\times 2\sqrt{11}} = \frac{2}{\sqrt{11}} = \boxed{0.603}$$

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

$n = 25$, $\sum X = 125$, $\sum X^2 = 650$, $\sum Y = 100$, $\sum Y^2 = 460$, $\sum XY = 508$.

It was, however, later ~~discussed~~ discovered at the time of checking that he had copied down two pairs as

X	Y
6	14
8	6

while the correct values were

X	Y
8	12
6	8

Obtain the correct value of correlation coefficient.

Sol: Corrected $\sum X = 125 - 6 - 8 + 8 + 6 = 125$

Corrected $\sum Y = 100 - 14 - 6 + 12 + 8 = 100$

Corrected $\sum X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$

Corrected $\sum Y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$

Corrected $\sum XY = 508 - 6\times 14 - 8\times 6 + 8\times 12 + 6\times 8 = 520$

$\bar{X} = \frac{1}{n}\sum X = \frac{1}{25}\times 125 = 5$, $\bar{Y} = \frac{1}{n}\sum Y = \frac{1}{25}\times 100 = 4$

$Cov(X,Y) = \frac{1}{n}\sum XY - \bar{X}\bar{Y} = \frac{1}{25}\times 520 - 5\times 4 = 4/5$

$\sigma_X^2 = \cancel{\frac{1}{25}x}\cdot\frac{1}{n}\sum X^2 - \bar{X}^2 = \frac{1}{25}\times 650 - 5^2 = 1$

$\sigma_Y^2 = \frac{1}{n}\sum Y^2 - \bar{Y}^2 = \frac{1}{25}\times 436 - 4^2 = 36/25$

Corrected $r(X,Y) = \dfrac{Cov(X,Y)}{\sigma_X \sigma_Y} = \dfrac{4/5}{1 \times 6/5} = \dfrac{2}{3}$
$= \boxed{0.67}$

Probable Error of Correlation coefficient

If r is the correlation coefficient in a sample of n pairs of observations, then its standard error (S.E) is given by

$$S.E(r) = \frac{1-r^2}{\sqrt{n}}$$

Probable error (P.E) of correlation coefficient is given by

$$P.E(r) = 0.6745 \times S.E(r)$$

$$= 0.6745 \frac{(1-r^2)}{\sqrt{n}}$$

If $r < P.E(r)$, correlation is not at all significant

If $r > 6\,P.E(r)$, it is definitely significant.

Rank Correlation

Rank correlation coefficient is calculated in the case, when a group of n individuals are arranged in order of merit or proficiency of two characteristics A and B.

Let (x_i, y_i); $i = 1, 2, \ldots n$ be the ranks of ith individual in two characteristics A and B respectively

The rank correlation, $\rho = 1 - \dfrac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$ ——— (1)

where $d_i = x_i - y_i$

This formula to calculate the rank correlation is called Spearman's formula,

We always have $\sum d_i = \sum (x_i - y_i) = \sum x_i - \sum y_i$
$= n(\bar{x} - \bar{y}) = 0$

Ex (1) The ranks of same 16 students in Mathematics and Physics are as follows. Two numbers within brackets denote the ranks of the students in Mathematics and Physics:
(1, 1), (2, 10), (3, 3), (4, 4), (5, 5), (6, 7), (7, 2), (8, 6), (9, 8)
(10, 11), (11, 15), (12, 9), (13, 14), (14, 12), (15, 16), (16, 13)
Calculate the rank correlation coefficient for proficiencies of this group in Mathematics and Physics.

Sol:

Marks in maths(X)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
Marks in physics(Y)	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13	
$d = X - Y$	0	-8	0	0	0	-1	5	2	1	-1	-4	3	-1	2	-1	3	0
d^2	0	64	0	0	0	1	25	4	1	1	16	9	1	4	1	9	136

The rank correlation coefficient is given by

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 136}{16(16^2-1)} = 1 - \frac{6 \times 136}{16 \times 255}$$

$$= 1 - \frac{1}{5} = \frac{4}{5} = 0.8$$

Remark: d is calculated by taking the difference in X and Y. $\sum d = 0$, provides a check for correctness of computation at that stage.

Here in the above problem, there is no tied rank.

Q(2) Ten competitors in a musical test were ranked by three judges A, B and C in the following order:

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

Using rank correlation method discuss which pair of judges has the nearest approach to common liking in music.

Sol:

Using rank correlation method, discuss which pair of judges has the nearest approach to common likings in music.

Solution. Here $n = 10$

Ranks by A (X)	Ranks by B (Y)	Ranks by C (Z)	$d_1 = X - Y$	$d_2 = X - Z$	$d_3 = Y - Z$	d_1^2	d_2^2	d_3^2
1	3	6	−2	−5	−3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	−3	−4	−1	9	16	1
10	4	8	6	2	−4	36	4	16
3	7	1	−4	2	6	16	4	36
2	10	2	−8	0	8	64	0	64
4	2	3	2	1	−1	4	1	1
9	1	10	8	−1	−9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	−1	1	2	1	1	4
Total			$\Sigma d_1 = 0$	$\Sigma d_2 = 0$	$\Sigma d_3 = 0$	$\Sigma d_1^2 = 200$	$\Sigma d_2^2 = 60$	$\Sigma d_3^2 = 214$

$$\rho(X, Y) = 1 - \frac{6\Sigma d_1^2}{n(n^2-1)} = 1 - \frac{6 \times 200}{10 \times 99} = 1 - \frac{40}{33} = -\frac{7}{33}$$

$$\rho(X, Z) = 1 - \frac{6\Sigma d_2^2}{n(n^2-1)} = 1 - \frac{6 \times 60}{10 \times 99} = 1 - \frac{4}{11} = \frac{7}{11}$$

$$\rho(Y, Z) = 1 - \frac{6\Sigma d_3^2}{n(n^2-1)} = 1 - \frac{6 \times 214}{10 \times 99} = 1 - \frac{214}{165} = -\frac{49}{165}.$$

Since $\rho(X, Z)$ is maximum, we conclude that the pair of judges A and C has the nearest approach to common likings in music.

10.7.3. Repeated Ranks

$\int$ Rank Correlation coefficient for Repeated rank:

Q) 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

Sol [Before the solution, first point you need to notice that data is not given in the 'rank' form. So we need to assign the ranks first. But here assigning the rank is not very direct as the values are repeated so see the solution carefully and convince yourself with the process]. Formula to calculate,

$$\rho = 1 - \frac{6(\Sigma d^2 + T_X + T_Y)}{n(n^2-1)}$$

Solution.

CALCULATIONS FOR RANK CORRELATION

X	Y	Rank X (x)	Rank Y (y)	$d = x - y$	d^2
68	62	4	5	–1	1
64	58	6	7	–1	1
75	68	2·5	3·5	–1	1
50	45	9	10	–1	1
64	81	6	1	5	25
80	60	1	6	–5	25
75	68	2·5	3·5	–1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
				$\Sigma d = 0$	$\Sigma d^2 = 72$

In the X-series we see that the value 75 occurs 2 times. The common rank given to these values is 2·5 which is the average of 2 and 3, the ranks which these values would have taken if they were different. The next value 68, then gets the next rank which is 4. Again we see that value 64 occurs thrice. The common rank given to it is 6 which is the average of 5, 6 and 7. Similarly in the Y-series, the value 68 occurs twice and its common rank is 3·5 which is the average of 3 and 4. As a result of these common rankings, the formula for 'ρ' has to be corrected. To Σd^2 we add $\frac{m(m^2-1)}{12}$ for each value repeated, where m is the number of times a value occurs. In the X-series the correction is to be applied twice, once for the value 75 which occurs twice ($m = 2$) and then for the value 64 which occurs thrice ($m = 3$). The total correction for the X-series is : $\frac{2(4-1)}{12} + \frac{3(9-1)}{12} = \frac{5}{2}$. Similarly, this correction for the Y-series is $\frac{2(4-1)}{12} = \frac{1}{2}$, as the value 68 occurs twice.

$$\therefore \qquad \rho = 1 - \frac{6\left(\Sigma d^2 + \frac{5}{2} + \frac{1}{2}\right)}{n(n^2-1)} = 1 - \frac{6(72+3)}{10 \times 99} = 0·545.$$

Calculate the correlation coefficient.

Solution.

CORRELATION TABLE

v	Mid-value	u / Age (X) / Marks (Y)	-1 / 18	0 / 19	1 / 20	2 / 21	Total $g(v)$	$vg(v)$	$v^2g(v)$	$\sum\limits_u uv\, f(u, v)$
-2	15	10—20	4 (8)	2 (0)	2 (−4)		8	−16	32	4
-1	25	20—30	5 (5)	4 (0)	6 (−6)	4 (−8)	19	−19	19	−9
0	35	30—40	6 (0)	8 (0)	10 (0)	11 (0)	35	0	0	0
1	45	40—50	4 (−4)	4 (0)	6 (6)	8 (16)	22	22	22	18
2	55	50—60		2 (0)	4 (8)	4 (16)	10	20	40	24
3	65	60—70		2 (0)	3 (9)	1 (6)	6	18	54	15
		Total $f(u)$	19	22	31	28	100	25	167	52
		$u\,f(u)$	−19	0	31	56	68			
		$u^2f(u)$	19	0	31	112	162			
		$\sum\limits_v uv f(u, v)$	9	0	13	30	52			

Let $\quad U = X - 19, \quad V = \{(Y - 35)/10\}$

$$\bar{u} = \frac{1}{N} \sum_u uf(u) = \frac{68}{100} = 0{\cdot}68, \quad \bar{v} = \frac{1}{N} \sum_v v\,g(v) = \frac{25}{100} = 0{\cdot}25$$

$$\text{Cov}(u, v) = \frac{1}{N} \sum_u \sum_v uv\, f(u, v) - \bar{u}\,\bar{v} = \frac{1}{100} \times 52 - 0{\cdot}68 \times 0{\cdot}25 = 0{\cdot}35$$

$$\sigma_U^2 = \frac{1}{N} \sum_u u^2 f(u) - \bar{u}^2 = \frac{162}{100} - (0{\cdot}68)^2 = 1{\cdot}1576$$

$$\sigma_V^2 = \frac{1}{N} \sum_v v^2 g(v) - \bar{v}^2 = \frac{167}{100} - (0{\cdot}25)^2 = 1{\cdot}6075$$

$$\therefore \quad r(U, V) = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = \frac{0{\cdot}35}{\sqrt{1{\cdot}1576 \times 1{\cdot}6075}} = 0{\cdot}25$$

Since correlation coefficient is independent of change of origin and scale,

$$r(X, Y) = r(U, V) = 0{\cdot}25.$$

Remark. Figures in circles in the table are the product terms $uvf(u, v)$.

Scanned with CamScanner