B.Tech II Semester (2020 Batch)
PROBABILITY AND STATISTICS (20BM1104)
(For CSE-3 & CSE-4)

Unit – 5: Correlation and Regression

(*The method of least squares, curvilinear regression, multiple regressions, correlation (excluding causation)*)

Curve fitting: Computing a curve corresponding to a given set of points is called a curve fitting

Regression: A relation between independent and dependent variables obtained from a given set of points is called a regression

Simple regression: A relation between one dependent variable and one independent variable obtained from a given set of points is called a simple regression

Multiple regression: A relation between one dependent variable and two or more independent variables obtained from a given set of points is called a simple regression

Regression line of y on x: A relation of the form $y = a + bx$ is called a regression line of y on x

Regression line of x on y: A relation of the form $x = a + by$ is called a regression line of x on y

Regression curve of y on x_1, x_2**:** A relation of the form $y = a + bx_1 + cx_2$ is called a regression curve of y on x_1, x_2

Least Squares Method: The method of computing a curve (or regression curve) by using a given set of points such that the sum of the squares of deviations from the points to the curve along $y-$axis is minimum

Curve fitting by Least Squares Method:

1. To fit a straight line of the form $y = a + bx$, the Normal equation are given by
$$\sum y = na + b\sum x$$
$$\sum xy = a\sum x + b\sum x^2$$

2. To fit a straight line of the form $x = a + by$, the Normal equation are given by
$$\sum x = na + b\sum y$$
$$\sum xy = a\sum y + b\sum y^2$$

3. To fit an exponential curve of the form $y = ae^{bx}$,
First write $\ln y = \ln a + bx$ and then the Normal equation are given by
$$\sum \ln y = n\ln a + b\sum x$$
$$\sum x\ln y = \ln a\sum x + b\sum x^2$$

4. To fit an exponential curve of the form $y = ab^x$,
First write $\log y = \log a + x \log b$ and then the Normal equation are given by
$\sum \log y = n \log a + \log b \sum x$
$\sum x \log y = \log a \sum x + \log b \sum x^2$

5. To fit a power curve (geometric curve) of the form $y = ax^b$,
First write $\log y = \log a + b \log x$ and then the Normal equation are given by
$\sum \log y = n \log a + b \sum \log x$
$\sum \log x \log y = \log a \sum \log x + b \sum (\log x)^2$

6. To fit a parabola of 2nd degree (or quadratic curve) of the form $y = a + bx + cx^2$, the Normal equation are given by
$\sum y = na + b \sum x + c \sum x^2$
$\sum xy = a \sum x + b \sum x^2 + c \sum x^3$
$\sum x^2 y = a \sum x^2 + b \sum x^3 + c \sum x^4$

7. To fit a multiple regression curve of the form $z = a + bx + cy$, the Normal equation are given by
$\sum z = na + b \sum x + c \sum y$
$\sum xz = a \sum x + b \sum x^2 + c \sum xy$
$\sum yz = a \sum y + b \sum xy + c \sum y^2$

8. To fit a multiple regression curve of the form $y = a + bx_1 + cx_2$, the Normal equation are given by
$\sum y = na + b \sum x_1 + c \sum x_2$
$\sum x_1 y = a \sum x_1 + b \sum x_1^2 + c \sum x_1 x_2$
$\sum x_2 y = a \sum x_2 + b \sum x_1 x_2 + b \sum x_2^2$

Problems:

1. Fit a straight line $y = a + bx$ for the following data by least squares method

x	1	2	3	4	5
y	12	25	40	50	65

Solution: The normal equations for the straight line $y = a + bx$ are

$\sum y = na + b\sum x$ and $\sum xy = a\sum x + b\sum x^2$

Consider

x	y	x^2	xy
1	12	1	12
2	25	4	50
3	40	9	120
4	50	16	200
5	65	25	325
$\sum x = 15$	$\sum y = 192$	$\sum x^2 = 55$	$\sum xy = 707$

Here $\sum x = 15$, $\sum y = 192$, $\sum x^2 = 55$, $\sum xy = 707$ and $n = 5$

The normal equations becomes $192 = 5a + 15b \quad \cdots \quad \cdots \quad (1)$

and $707 = 15a + 55b \quad \cdots \quad \cdots \quad (2)$

Solving (1) and (2), $a = -0.9$ and $b = 13.1$

Hence the straight line is $y = -0.9 + 13.1x$

2. By the method of least squares, fit a straight line $y = a + bx$ for the following data

x	50	70	100	120
y	12	15	21	25

Solution: The normal equations for the straight line $y = a + bx$ are

$\sum y = na + b\sum x$ and $\sum xy = a\sum x + b\sum x^2$

Consider

x	y	x^2	xy
50	12	2500	600
70	15	4900	1050
100	21	10000	2100
120	25	14400	3000
$\sum x = 340$	$\sum y = 73$	$\sum x^2 = 31800$	$\sum xy = 6750$

Here $\sum x = 340$, $\sum y = 73$, $\sum x^2 = 31800$, $\sum xy = 6750$ and $n = 4$

The normal equations becomes $73 = 4a + 340b \quad \cdots \quad \cdots \quad \cdots \quad (1)$

and $6750 = 340a + 31800b \quad \cdots \quad \cdots \quad (2)$

Solving (1) and (2), $a = 2.2758$ and $b = 0.1879$

Hence the straight line is $y = 2.2758 + 0.1879x$

3. By the method of least squares, fit a straight line $x = a + by$ for the following data

x	12	15	21	25
y	50	70	100	120

Solution: The normal equations for the straight line $x = a + by$ are

$\sum x = na + b\sum y$ and $\sum xy = a\sum y + b\sum y^2$

Consider

x	y	y^2	xy
12	50	2500	600
15	70	4900	1050
21	100	10000	2100
25	120	14400	3000
$\sum x = 73$	$\sum y = 340$	$\sum y^2 = 31800$	$\sum xy = 6750$

Here $\sum x = 73, \sum y = 340, \sum y^2 = 31800, \sum xy = 6750$ and $n = 4$

The normal equations becomes $73 = 4a + 340b \quad \cdots \quad \cdots \quad \cdots \quad (1)$

and $6750 = 340a + 31800b \quad \cdots \quad \cdots \quad (2)$

Solving (1) and (2), $a = 2.2785$ and $b = 0.1879$

Hence the straight line is $x = 2.2785 + 0.1879y$

4. Fit an exponential curve $y = ae^{bx}$ for the following data

x	1	3	5	7	9
y	100	81	73	54	43

Solution: The curve is $y = ae^{bx}$

Taking log on both sides, $\log y = \log a + bx \log e$

That is, $Y = A + Bx$, where $Y = \log y, A = \log a, B = b\log e$

Now the normal equations for $Y = A + Bx$ are

$\sum Y = nA + B\sum x$ and $\sum xY = A\sum x + B\sum x^2$

Consider,

x	y	$Y = \log y$	x^2	xY
1	100	2	1	2
3	81	1.9085	9	5.7255
5	73	1.8633	25	9.3165
7	54	1.7324	49	12.1268
9	43	1.6335	81	14.7015
$\sum x = 25$		$\sum Y = 9.1377$	$\sum x^2 = 165$	$\sum xY = 43.8703$

Here $\sum x = 25, \sum Y = 9.1377, \sum x^2 = 165, \sum xY = 43.8703$ and $n = 5$

The normal equations becomes $9.1377 = 5A + 25B \quad \cdots \quad \cdots \quad \cdots \quad (1)$

and $43.8703 = 25A + 165B \quad \cdots \quad \cdots \quad (2)$

Solving (1) and (2), $A = 2.0548$ and $B = -0.0455$ or

Therefore, $a = 10^A = 113.4488$ and $b = \dfrac{B}{\log e} = -0.1048$

Hence the required curve is $y = 113.4488e^{-0.1048x}$

5. Fit an exponential curve $y = ab^x$ for the following data

x	1	2	3	4	5
y	130	152.2	177.3	190.2	244.7

Solution: The curve is $y = ab^x$

Taking log on both sides, $\log y = \log a + x\log b$

That is, $Y = A + Bx$, where $Y = \log y, A = \log a, B = \log b$

Now the normal equations for $Y = A + Bx$ are

$\sum Y = nA + B\sum x$ and $\sum xY = A\sum x + B\sum x^2$

Consider,

x	y	$Y = \log y$	x^2	xY
1	130	2.1139	1	2.1139
2	152.2	2.1824	4	4.3648
3	177.3	2.2487	9	6.7461
4	190.2	2.2792	16	9.1168
5	244.7	2.3886	25	11.9432
$\sum x = 15$		$\sum Y = 11.2129$	$\sum x^2 = 55$	$\sum xY = 34.2849$

Here $\sum x = 15$, $\sum Y = 11.2129$, $\sum x^2 = 55$, $\sum xY = 34.2849$ and $n = 5$

The normal equations becomes $11.2129 = 5A + 15B \quad \cdots \quad \cdots \quad \cdots \quad (1)$

and $34.2849 = 15A + 55B \quad \cdots \quad \cdots \quad \cdots \quad (2)$

Solving (1) and (2), $A = 2.0487$ and $B = 0.0646$ or

Therefore, $a = 10^A = 111.8716$ and $b = 10^B = 1.1604$

Hence the required curve is $y = 111.8716\,(1.1604)^x$

6. Fit an exponential curve $y = ab^x$ for the following data

x	2	3	4	5	6
y	144	172.8	207.4	248.8	298.6

$\log y = \log a + x\log b$ or $Y = A + Bx$,where $Y = \log y, A = \log a, B = \log b$

The normal equations for $Y = A + Bx$ are

$\sum Y = nA + B\sum x$ and $\sum xY = A\sum x + B\sum x^2$

Here $\sum x = 20$, $\sum Y = \sum \log y = 11.5837$, $\sum x^2 = 90$, $\sum xY = \sum x\log y = 47.1266$ and $n = 5$

The normal equations becomes $11.5837 = 5A + 20B$ and $47.1266 = 20A + 90B$

On solving, $A = 2$ and $B = 0.0792$ or $a = 100$ and $b = 1.2$

Hence the exponential curve $y = 100(1.2)^x$

7. Fit a power curve $y = ax^b$ for the following data

x	1	2	3	4	5	6
y	2.98	4.26	5.21	6.10	6.80	7.50

Solution: The power curve is $y = ax^b$

Taking log on both sides, $\log y = \log a + b \log x$

That is, $Y = A + bX$, where $Y = \log y, X = \log x, A = \log a$

Now the normal equations for $Y = A + bX$ are

$\sum Y = nA + b\sum X$ and $\sum XY = A\sum X + b\sum X^2$

Consider,

x	y	$X = \log x$	$Y = \log y$	X^2	XY
1	2.98	0	0.4742	0	0
2	4.26	0.3010	0.6294	0.0906	0.1895
3	5.21	0.4771	0.7168	0.2276	0.3420
4	6.10	0.6021	0.7853	0.3625	0.4728
5	6.80	0.6990	0.8325	0.4886	0.5819
6	7.50	0.7782	0.8751	0.6055	0.6809
		$\sum X = 2.8573$	$\sum Y = 4.3134$	$\sum X^2 = 1.7748$	$\sum XY = 2.2671$

Here $\sum X = 2.8573, \sum Y = 4.3134 \sum X^2 = 1.7748, \sum XY = 2.2671$ and $n = 6$

The normal equations become $4.3134 = 6A + 2.8573b \quad \cdots \quad \cdots \quad \cdots \quad (1)$

and $2.2671 = 2.8573A + 1.7748b \quad \cdots \quad \cdots \quad (2)$

Solving (1) and (2), $A = 0.4740$ and $b = 0.5143$

Therefore, $a = 10^A = 2.9783$

Hence the curve is $y = 2.9783(x)^{0.5143}$

8. Determine the regression line of y on x for the following data

x	20	25	28	35	43
y	52	48	63	79	95

Solution: The regression line of y on x is given by $y = a + bx$

The normal equations for are

$\sum y = na + b\sum x$ and $\sum xy = a\sum x + b\sum x^2$

Consider

x	y	x^2	xy
20	52	400	1040
25	48	625	1200
28	63	784	1764
35	79	1225	2765
43	95	1849	4085
$\sum x = 151$	$\sum y = 337$	$\sum x^2 = 4883$	$\sum xy = 10854$

Here $\sum x = 151, \sum y = 337, \sum x^2 = 4883, \sum xy = 10854$ and $n = 5$

The normal equations become $337 = 5a + 151b \quad \cdots \quad \cdots \quad (1)$

and $10854 = 151a + 4883b \quad \cdots \quad \cdots \quad (2)$

Solving (1) and (2), $a = 4.0998$ and $b = 2.096$
Hence the regression line of y on x is $y = 4.0998 + 2.096x$

9. Determine the regression line of x on y for the following data

x	1.1	2.3	4.5	7.6
y	21	35	64	84

And estimate the value of x at $y = 4.8$

Solution: The regression line of x on y is given by $x = a + by$
The normal equations for the straight line $x = a + by$ are

$\sum x = na + b\sum y$ and $\sum xy = a\sum y + b\sum y^2$

Consider

x	y	y^2	xy
1.1	21	441	23.1
2.3	35	1225	80.5
4.5	64	4096	288
7.6	84	7056	638.4
$\sum x = 15.5$	$\sum y = 204$	$\sum y^2 = 12818$	$\sum xy = 1030$

Here $\sum x = 15.5, \sum y = 204, \sum y^2 = 12818, \sum xy = 1030$ and $n = 4$

The normal equations become $15.5 = 4a + 204b \quad \cdots \quad \cdots \quad \cdots \quad (1)$

and $1030 = 204a + 12818b \quad \cdots \quad \cdots \quad (2)$

Solving (1) and (2), $a = -1.1849$ and $b = 0.0992$
Hence the regression line of x on y is $x = -1.1849 + 0.0992y$
Therefore, the value of x at $y = 4.8$ is given by $x = -1.1849 + 0.0992\,(4.8) = 0.70874$

10. Fit a second degree parabola $y = a + bx + cx^2$ for the following data

x	1	3	5	7	9
y	2	7	10	11	9

Solution: The curve is $y = a + bx + cx^2$
The normal equations are

$\sum y = na + b\sum x + c\sum x^2$,

$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$

$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$

Consider,

x	y	x^2	x^3	x^4	xy	x^2y
1	2	1	1	1	2	2
3	7	9	27	81	21	63
5	10	25	125	625	50	250
7	11	49	343	2401	77	539
9	9	81	729	6561	81	729
$\sum x = 25$	$\sum y = 39$	$\sum x^2 = 165$	$\sum x^3 = 1225$	$\sum x^4 = 9669$	$\sum xy = 231$	$\sum x^2 y = 1583$

Here $\sum x = 25, \sum y = 39, \sum x^2 = 165, \sum x^3 = 1225, \sum x^4 = 9669, \sum xy = 231$
$\sum x^2 y = 1583$ and $n = 5$

The normal equations $39 = 5a + 25b + 165c \quad \cdots \quad \cdots \quad \cdots \quad (1)$

$231 = 25a + 165b + 1225c \quad \cdots \quad \cdots \quad (2)$

and $1583 = 165a + 1225b + 9669c \quad \cdots \quad \cdots \quad (3)$

Solving (1), (2) and (3), $a = -1.5571$, $b = 3.7571$ and $c = -0.2857$

Hence the parabola is $y = -1.5571 + 3.7571x - 0.2857x^2$

11. Fit a second degree parabola $y = a + bx + cx^2$ for the following data

x	1.0	1.5	2.0	2.5	3.0	3.5	4.0
y	1.1	1.3	1.6	2.0	2.7	3.4	4.1

Solution: The curve is $y = a + bx + cx^2$

The normal equations are

$\sum y = na + b\sum x + c\sum x^2$,

$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$

$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$

Consider,

x	y	x^2	x^3	x^4	xy	x^2y
1.0	1.1	1	1	1	1.1	1.1
1.5	1.3	2.25	3.375	5.0625	1.95	2.925
2.0	1.6	4	8	16	3.2	6.4
2.5	2.0	6.25	15.625	39.0625	5	12.5
3.0	2.7	9	27	81	8.1	24.3
3.5	3.4	12.25	42.875	150.0625	11.9	41.65
4.0	4.1	16	64	256	0.4	1.6
$\sum x =$ 17.5	$\sum y =$ 12.2	$\sum x^2 =$ 50.75	$\sum x^3 =$ 161.875	$\sum x^4 =$ 548.1875	$\sum xy =$ 31.65	$\sum x^2 y =$ 90.475

The normal equations $12.2 = 7a + 17.5b + 50.75c \quad \cdots \quad \cdots \quad \cdots \quad (1)$

$31.65 = 17.5a + 50.75b + 161.875c \quad \cdots \quad \cdots \quad (2)$

and $90.475 = 50.75a + 161.875b + 548.1875c \quad \cdots \quad \cdots \quad (3)$

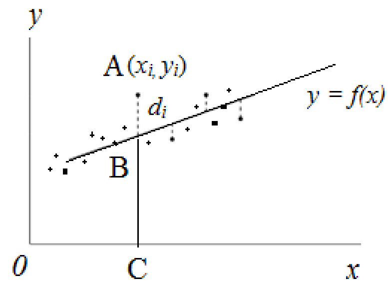
Solving (1), (2) and (3), $a = -2.3929$, $b = 3.7119$ and $c = -0.7095$

Hence the parabola is $y = -2.3929 + 3.7119x - 0.7095x^2$

12. Determine the Normal equations to fit a straight line of the form $y = a + bx$

Solution: Let $y = f(x)$, where $f(x) = a + bx$

Let $A = (x_i\ y_i)$ be any given data point



At $x = x_i$, the observed (given) value of y is y_i; that is $AC = y_i$

At $x = x_i$, the expected value of y is $f(x_i)$; that is $BC = f(x_i)$

Therefore, the deviation (d_i) at $x = x_i$ is given by $AB = AC - BC$

That is, $d_i = |y_i - f(x_i)|$ so that $d_i^2 = [y_i - f(x_i)]^2$

Sum of the squares of the deviations is given by $S = \sum d_i^2 = \sum [y_i - f(x_i)]^2 = \sum [y_i - (a + bx_i)]^2$

According least squares method S is minimum

To get S minimum, we need $\frac{\partial S}{\partial a} = 0$ and $\frac{\partial S}{\partial b} = 0$

Now $\frac{\partial S}{\partial a} = 0 \Rightarrow 2\sum [y_i - (a + bx_i)](1) = 0$

$\Rightarrow \sum [y_i - (a + bx_i)] = 0$

$\Rightarrow \sum y_i = \sum a + \sum bx_i$

$\Rightarrow \sum y_i = na + b\sum x_i \quad \cdots \quad \cdots \quad \cdots \quad (1)$

And $\frac{\partial S}{\partial b} = 0 \Rightarrow 2\sum [y_i - (a + bx_i)](x_i) = 0$

$\Rightarrow \sum [x_i y_i - (ax_i + bx_i^2)] = 0$

$\Rightarrow \sum x_i y_i = \sum ax_i + \sum bx_i^2$

$\Rightarrow \sum x_i y_i = a\sum x_i + b\sum x_i^2 \quad \cdots \quad \cdots \quad \cdots \quad (2)$

Therefore (1) and (2) are the required normal equations

13. Determine the least squares regression equation of the form $y = a + bx_1 + cx_2$ for the following data

y	3	5	6	8	12	14
x_1	16	10	7	4	3	2
x_2	90	72	54	42	30	12

Solution: The equation is $y = a + bx_1 + cx_2$

The normal equations are

$\sum y = na + b\sum x_1 + c\sum x_2$,

$\sum x_1 y = a\sum x_1 + b\sum x_1^2 + c\sum x_1 x_2$

$\sum x_2 y = a\sum x_2 + b\sum x_1 x_2 + c\sum x_2^2$

Consider,

y	x_1	x_2	x_1^2	x_2^2	x_1x_2	x_1y	x_2y
3	16	90	256	8100	1440	48	270
5	10	72	100	5184	720	50	360
6	7	54	49	2916	378	42	324
8	4	42	16	1764	168	32	336
12	3	30	9	900	90	36	360
14	2	12	4	144	24	28	168
$\sum y =$ 48	$\sum x_1 =$ 42	$\sum x_2 =$ 300	$\sum x_1^2 =$ 434	$\sum x_2^2 =$ 19008	$\sum x_1x_2 =$ 2820	$\sum x_1y =$ 236	$\sum x_2y =$ 1818

The normal equations $48 = 6a + 42b + 300c \quad \cdots \quad \cdots \quad \cdots \quad (1)$

$236 = 42a + 434b + 2820c \quad \cdots \quad \cdots \quad (2)$

and $1818 = 300a + 2820b + 19008c \quad \cdots \quad \cdots \quad (3)$

Solving (1), (2) and (3), $a = 16.1067$, $b = 0.4270$ and $c = -0.2219$

Hence the regression equation is $y = 16.1067 + 0.4270x_1 - 0.2219x_2$

14. Determine the least squares regression equation of the form $z = a + bx + cy$ for the following data

z	16	19	23	20	26	23	28
x	1	2	3	4	5	6	7
y	4	5	7	2	6	1	4

Solution: The equation is $z = a + bx + cy$

The normal equations are

$\sum z = na + b\sum x + c\sum y$,

$\sum xz = a\sum x + b\sum x^2 + c\sum xy$

$\sum yz = a\sum y + b\sum xy + c\sum y^2$

Consider,

z	x	y	x^2	y^2	xy	xz	yz
16	1	4	1	16	4	16	64
19	2	5	4	25	10	38	95
23	3	7	9	49	21	69	161
20	4	2	16	4	8	80	40
26	5	6	25	36	30	130	156
23	6	1	36	1	6	138	23
28	7	4	49	16	28	196	112
$\sum z =$ 155	$\sum x =$ 28	$\sum y =$ 29	$\sum x^2 =$ 140	$\sum y^2 =$ 147	$\sum xy =$ 107	$\sum xz =$ 667	$\sum yz =$ 651

The normal equations $155 = 7a + 28b + 29c \quad \cdots \quad \cdots \quad \cdots$ (1)

$667 = 28a + 140b + 107c \quad \cdots \quad \cdots$ (2)

and $651 = 29a + 107b + 147c \quad \cdots \quad \cdots$ (3)

Solving (1), (2) and (3), $a = 10$, $b = 2$ and $c = 1$

Hence the regression equation is $z = 10 + 2x + y$

Exercise:

1. Fit a straight line $y = a + bx$ for the following data by least squares method

x	1	2	3	4	5	6
y	14	33	40	63	76	85

2. Fit a straight line $y = a + bx$ for the following data by least squares method

x	0	2	3	5	9
y	-3	4	3	8	15

3. Fit a straight line $x = a + by$ for the following data by least squares method

x	18.5	25.4	30	64.5	34.6	89.8	20.8
y	5	8	10	25	12	36	6

(Ans: $x = 7 + 2.3y$)

4. The following shows the improvement of eight students in a speed-reading program, and the number of weeks they have been in the program:

No.of weeks x	3	5	2	8	6	9	3	4
Speed gain (words/min.) y	86	118	49	193	164	232	73	109

Fit a straight line by the method of least squares

5. If p is the pull required to lift a load w by means of a pulley block, find a linear law of the form $p = mw + c$ using the data

p	12	15	21	25
w	50	70	100	120

6. Predict y at $x = 3.75$ by fitting power curve $y = a\,x^b$ for the following data

x	1	2	3	4	5	6
y	2.98	4.26	5.21	6.10	6.80	7.50

7. Fit a second degree parabola $y = a + bx + cx^2$ for the following data

x	2.5	3.6	4.6	5.2	6.8	7.2	8.9	9.2
y	1.8	2.6	4.8	6.2	8.9	4.2	2.9	4.5

(Ans: $y = -7.7504 + 4.422x - 0.3472x^2$)

8. Determine the regression line of y on x for the following data

x	50	60	70	90	100
y	65	51	40	26	08

9. Find Y when X_1 = 10 and X_2 = 6 from the least square regression equation of Y on X_1 and X_2 for the following data

Y	90	72	54	42	30	12
X_1	3	5	6	8	12	14
X_2	16	10	7	4	3	2

10. Fit a least-squares regression plane for the following data and also find y at $x_1 = 2.2$ and $x_2 = 90$.

y	5.3	7.8	7.4	9.8	10.8	9.1	8.1	7.2	6.5	12.6
x_1	1.5	2.5	0.5	1.2	2.6	0.3	2.4	2	0.7	1.6
x_2	66	87	69	141	93	105	111	78	66	123

Correlation: The relationship between two variables such that a change in one variable results in a $+$ve or -ve change in the other and also greater change in one variable results in corresponding greater change in the other is called a correlation. For a change in one variable, if there is a corresponding change in the other variable then the variables are called correlated.

Note:

(i) If the variables deviate in the same direction then the correlation is called direct or $+$ve correlation

(ii) If the variables deviate in the opposite direction then the correlation is called inverse or -ve correlation

Correlation Coefficient (or Karl Pearson coefficient of correlation): The numerical measurement of linear relationship between the variables x and y is called the coefficient of correlation of x and y and it is denoted by $r(x, y)$ or r

Note:

(i) The coefficient of correlation r is always lies between -1 and 1; that is, $-1 \le r \le 1$

(ii) If $r = 0$ then the variables are not correlated

(iii) If $r = 1$ then the variables are positively and perfectly correlated

(iv) If $r = -1$ then the variables are negatively and perfectly correlated

(v) If $0 < r < 1$ then the variables are positively and partially correlated

(vi) If $-1 < r < 0$ then the variables are negatively and partially correlated

Correlation formulas:

(i) Mean of x is given by $\bar{x} = \dfrac{\sum x}{n}$

(ii) Variance of x is given by $\sigma_x^2 = \dfrac{\sum (x-\bar{x})^2}{n}$ or $\sigma_x^2 = \dfrac{\sum x^2}{n} - (\bar{x})^2$

(iii) Covariance of x and y is given by $Cov(x,y) = \dfrac{\sum (x-\bar{x})(y-\bar{y})}{n}$ or $Cov(x,y) = \dfrac{\sum xy}{n} - (\bar{x})(\bar{y})$

(iv) Coefficient of correlation of x and y is given by $r = \dfrac{Cov(x,y)}{\sigma_x \sigma_y}$ or $r = \dfrac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}}$

(v) $r = \dfrac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x \sigma_y}$ or $r = \dfrac{\sigma_{x+y}^2 - \sigma_x^2 - \sigma_y^2}{2\sigma_x \sigma_y}$

(vi) Regression line of y on x is given by $y - \bar{y} = r\dfrac{\sigma_y}{\sigma_x}(x - \bar{x})$

The slope of the regression line of y on x is called regression coefficient of y on x

It is denoted by b_{yx} and is given by $b_{yx} = r\dfrac{\sigma_y}{\sigma_x}$

(vii) Regression line of x on y is given by $x - \bar{x} = r\dfrac{\sigma_x}{\sigma_y}(y - \bar{y})$

The slope of the regression line of x on y is called regression coefficient of x on y

It is denoted by b_{xy} and is given by $b_{xy} = r\dfrac{\sigma_x}{\sigma_y}$

(viii) Both the regression lines passes through the point $(\bar{x}, \bar{y})$

(ix) The Geometric Mean of the regression coefficients is r; that is $r^2 = b_{xy} \times b_{yx} = r\dfrac{\sigma_x}{\sigma_y} \times r\dfrac{\sigma_y}{\sigma_x}$

Problems:

1. Determine the coefficient of correlation for the following data

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

Solution: The coefficient of correlation $r = \dfrac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}}$

Here $\bar{x} = \dfrac{\sum x}{n} = \dfrac{56}{8} = 7$ and $\bar{y} = \dfrac{\sum y}{n} = \dfrac{40}{8} = 5$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	1	- 6	- 4	24	36	16
3	2	- 4	- 3	12	16	9
4	4	- 3	- 1	3	9	1
6	4	- 1	- 1	1	1	1
8	5	1	0	0	1	0
9	7	2	2	4	4	4
11	8	4	3	12	16	9
14	9	7	4	28	49	16
$\sum x =$ 56	$\sum y =$ 40			$\sum(x-\bar{x})(y-\bar{y}) =$ 84	$\sum(x-\bar{x})^2 =$ 132	$\sum(y-\bar{y})^2 =$ 56

Observe that, $\sum(x-\bar{x})^2 = 132$, $\sum(y-\bar{y})^2 = 56$, $\sum(x-\bar{x})(y-\bar{y}) = 84$

Therefore, the coefficient of correlation $r = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} = \dfrac{84}{\sqrt{(132)(56)}} = 0.977$

2. Determine the coefficient of correlation for the following data

x	78	36	98	25	75	82	90	62	65	39
y	84	51	91	60	68	62	86	58	53	47

Solution: The coefficient of correlation $r = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$

Here $\bar{x} = \dfrac{\sum x}{n} = \dfrac{650}{10} = 65$ and $\bar{y} = \dfrac{\sum y}{n} = \dfrac{660}{10} = 66$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
78	84	13	18	234	169	324
36	51	- 29	-15	435	841	225
98	91	33	25	825	1089	625
25	60	- 40	-6	240	1600	36
75	68	10	2	20	100	4
82	62	17	-4	-68	289	16
90	86	25	20	500	625	400
62	58	- 3	-8	24	9	64
65	53	0	-13	0	0	169
39	47	- 26	-19	494	676	361
$\sum x =$ 650	$\sum y =$ 660			$\sum(x-\bar{x})(y-\bar{y}) =$ 2704	$\sum(x-\bar{x})^2 =$ 5398	$\sum(y-\bar{y})^2 =$ 2224

Observe that, $\sum(x-\bar{x})^2 = 5398$, $\sum(y-\bar{y})^2 = 2224$, $\sum(x-\bar{x})(y-\bar{y}) = 2704$

Therefore, the coefficient of correlation $r = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} = \dfrac{2704}{\sqrt{(5398)(2224)}} = 0.7804$

3. Determine the coefficient of correlation for the following data

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

Solution: The coefficient of correlation $r = \dfrac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}}$

Here $\bar{x} = \dfrac{\sum x}{n} = \dfrac{544}{8} = 68$ and $\bar{y} = \dfrac{\sum y}{n} = \dfrac{552}{8} = 69$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
65	67	-3	-2	6	9	4
66	68	-2	-1	2	4	1
67	65	-1	-4	4	1	16
67	68	-1	-1	1	1	1
68	72	0	3	0	0	9
69	72	1	3	3	1	9
70	69	2	0	0	4	0
72	71	4	2	8	16	4
$\sum x =$ 544	$\sum y =$ 552			$\sum (x-\bar{x})(y-\bar{y}) =$ 24	$\sum (x-\bar{x})^2 =$ 36	$\sum (y-\bar{y})^2 =$ 44

Observe that, $\sum (x-\bar{x})^2 = 36$, $\sum (y-\bar{y})^2 = 44$, $\sum (x-\bar{x})(y-\bar{y}) = 24$

Therefore, the coefficient of correlation $r = \dfrac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}} = \dfrac{24}{\sqrt{(36)(44)}} = 0.603$

4. From the following data

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

Determine (i) $\bar{x}$ and $\bar{y}$ (ii) σ_x and σ_y (iii) $Cov(x, y)$ (iv) the correlation coefficient between x and y

Solution: We know that $\bar{x} = \dfrac{\sum x}{n}$, $\sigma_x^2 = \dfrac{\sum (x-\bar{x})^2}{n}$, $Cov(x, y) = \dfrac{\sum (x-\bar{x})(y-\bar{y})}{n}$

and $r = \dfrac{Cov(x, y)}{\sigma_x \sigma_y}$

Consider,

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
65	67	-3	-2	6	9	4
66	68	-2	-1	2	4	1
67	65	-1	-4	4	1	16
67	68	-1	-1	1	1	1
68	72	0	3	0	0	9
69	72	1	3	3	1	9
70	69	2	0	0	4	0
72	71	4	2	8	16	4
$\sum x =$ 544	$\sum y =$ 552			$\sum(x-\bar{x})(y-\bar{y}) =$ 24	$\sum(x-\bar{x})^2 =$ 36	$\sum(y-\bar{y})^2 =$ 44

(i) $\bar{x} = \dfrac{\sum x}{n} = \dfrac{544}{8} = 68$ and $\bar{y} = \dfrac{\sum y}{n} = \dfrac{552}{8} = 69$

(ii) $\sigma_x^2 = \dfrac{\sum(x-\bar{x})^2}{n} = \dfrac{36}{8} = 4.5$, $\sigma_x = \sqrt{4.5} = 2.1213$

$\sigma_y^2 = \dfrac{\sum(y-\bar{y})^2}{n} = \dfrac{44}{8} = 5.5$, $\sigma_y = \sqrt{5.5} = 2.3452$

(iii) $Cov(x,y) = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{n} = \dfrac{24}{8} = 3$

(iv) $r = \dfrac{Cov(x,y)}{\sigma_x \sigma_y} = \dfrac{3}{2.1213 \times 2.3452} = 0.6030$

5. Find the correlation coefficient between x and y from the following data

x	78	89	97	69	59	79	68	57
y	125	137	156	112	107	138	123	108

Solution: The coefficient of correlation $r = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$

Here $\bar{x} = \dfrac{\sum x}{n} = \dfrac{596}{8} = 74.5$ and $\bar{y} = \dfrac{\sum y}{n} = \dfrac{1006}{8} = 125.75$

Consider,

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
78	125	3.5	-0.75	-2.625	12.25	0.5625
89	137	14.5	11.25	163.125	210.25	126.5625
97	156	22.5	30.25	680.625	506.25	915.0625
69	112	-5.5	-13.75	75.625	30.25	189.0625
59	107	-15.5	-18.75	290.625	240.25	351.5625
79	138	4.5	12.25	55.125	20.25	150.0625
68	123	-6.5	-2.75	17.875	42.25	7.5625
57	108	-17.5	-17.75	310.625	306.25	315.0625
$\sum x =$ 596	$\sum y =$ 1006			$\sum(x-\bar{x})(y-\bar{y}) =$ 1591	$\sum(x-\bar{x})^2 =$ 1368	$\sum(y-\bar{y})^2 =$ 2055.5

Observe that, $\sum(x-\bar{x})^2 = 1368$, $\sum(y-\bar{y})^2 = 2055.5$, $\sum(x-\bar{x})(y-\bar{y}) = 1591$

Therefore, the coefficient of correlation $r = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} = \dfrac{1591}{\sqrt{(1368)(2055.5)}} = 0.9488$

6. From the following data

x	78	89	97	69	59	79	68	57
y	125	137	156	112	107	138	123	108

Determine

(i) $\bar{x}$ and $\bar{y}$

(ii) σ_x and σ_y

(iii) $Cov(x, y)$

(iv) the correlation coefficient between x and y

(v) two regression lines

Solution: We know that $\bar{x} = \dfrac{\sum x}{n}$, $\sigma_x^2 = \dfrac{\sum(x-\bar{x})^2}{n}$, $Cov(x, y) = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{n}$

and $r = \dfrac{Cov(x, y)}{\sigma_x \sigma_y}$

Consider,

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
78	125	3.5	-0.75	-2.625	12.25	0.5625
89	137	14.5	11.25	163.125	210.25	126.5625
97	156	22.5	30.25	680.625	506.25	915.0625
69	112	-5.5	-13.75	75.625	30.25	189.0625
59	107	-15.5	-18.75	290.625	240.25	351.5625
79	138	4.5	12.25	55.125	20.25	150.0625
68	123	-6.5	-2.75	17.875	42.25	7.5625
57	108	-17.5	-17.75	310.625	306.25	315.0625
$\sum x =$ 596	$\sum y =$ 1006			$\sum(x-\bar{x})(y-\bar{y}) =$ 1591	$\sum(x-\bar{x})^2 =$ 1368	$\sum(y-\bar{y})^2 =$ 2055.5

(i) $\bar{x} = \dfrac{\sum x}{n} = \dfrac{596}{8} = 74.5$ and $\bar{y} = \dfrac{\sum y}{n} = \dfrac{1006}{8} = 125.75$

(ii) $\sigma_x^2 = \dfrac{\sum(x-\bar{x})^2}{n} = \dfrac{1368}{8} = 171$, $\sigma_x = \sqrt{171} = 13.0767$

$\sigma_y^2 = \dfrac{\sum(y-\bar{y})^2}{n} = \dfrac{2055.5}{8} = 256.9375$, $\sigma_y = \sqrt{256.9375} = 16.0293$

(iii) $Cov(x, y) = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{n} = \dfrac{1591}{8} = 198.875$

(iv) $r = \dfrac{Cov(x, y)}{\sigma_x \sigma_y} = \dfrac{198.875}{13.0767 \times 16.0293} = 0.9488$

(v) The regression line of y on x is given by $y-\bar{y}=r\dfrac{\sigma_y}{\sigma_x}(x-\bar{x})$

That is, $y-125.75=(0.9488)\,\dfrac{16.0293}{13.0767}\,(x-74.5)$

$\Rightarrow y-125.75=1.1630\,(x-74.5)$

$\Rightarrow y-125.75=1.1630x-86.6435$

$\Rightarrow y=39.1065+1.163x$

And the regression line of x on y is given by $x-\bar{x}=r\dfrac{\sigma_x}{\sigma_y}(y-\bar{y})$

That is, $x-74.5=(0.9488)\,\dfrac{13.0767}{16.0293}\,(y-125.75)$

$\Rightarrow x-74.5=0.7740\,(y-125.75)$

$\Rightarrow x-74.5=0.774y-97.3305$

$\Rightarrow x=-22.8305+0.774y$

7. The two regression equations of the variables x and y are $y-0.399x-6.934=0$ and $x-1.212y+2.461=0$. Find (i) mean of x (ii) mean of y (iii) correlation coefficient between x and y

Solution: Solving the given equations, $x=11.5083$ and $y=11.5258$

Therefore, $\bar{x}=11.5083$ and $\bar{y}=11.5258$

The regression coefficient of y on x is given by $r\dfrac{\sigma_y}{\sigma_x}=0.399$

The regression coefficient of x on y is given by $r\dfrac{\sigma_x}{\sigma_y}=1.212$

Correlation coefficient $r=\sqrt{\left(r\dfrac{\sigma_y}{\sigma_x}\right)\left(r\dfrac{\sigma_x}{\sigma_y}\right)}=\sqrt{(0.399)(1.212)}=0.6954$

(r is positive since both the regression coefficients are positive)

8. The two regression equations of the variables x and y are $x=19.13-0.87y$ and $y=11.64-0.50x$. Find (i) mean of x (ii) mean of y (iii) correlation coefficient between x and y

Solution: Solving the given equations, $x=15.9349$ and $y=3.6726$

Therefore, $\bar{x}=15.9349$ and $\bar{y}=3.6726$

The regression coefficient of y on x is given by $r\dfrac{\sigma_y}{\sigma_x}=-0.50$

The regression coefficient of x on y is given by $r\dfrac{\sigma_x}{\sigma_y}=-0.87$

Correlation coefficient $r=\sqrt{\left(r\dfrac{\sigma_y}{\sigma_x}\right)\left(r\dfrac{\sigma_x}{\sigma_y}\right)}=\sqrt{(-0.50)(-0.87)}=-0.66$

(r is negative since both the regression coefficients are negative)

9. Establish the formula $r = \dfrac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$

Solution: Consider,

$$\sigma_{x-y}^2 = \frac{1}{n}\sum\left[(x-y)-(\bar{x}-\bar{y})\right]^2 = \frac{1}{n}\sum\left[(x-\bar{x})-(y-\bar{y})\right]^2$$

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

$$= \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y$$

Therefore, $r = \dfrac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$

10. Establish the formula $r = \dfrac{\sigma_{x+y}^2 - \sigma_x^2 - \sigma_y^2}{2\sigma_x\sigma_y}$

Solution: Consider,

$$\sigma_{x+y}^2 = \frac{1}{n}\sum\left[(x+y)-(\bar{x}+\bar{y})\right]^2 = \frac{1}{n}\sum\left[(x-\bar{x})+(y-\bar{y})\right]^2$$

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

$$= \sigma_x^2 + \sigma_y^2 + 2r\sigma_x\sigma_y$$

Therefore, $r = \dfrac{\sigma_{x+y}^2 - \sigma_x^2 - \sigma_y^2}{2\sigma_x\sigma_y}$

11. Use the formula $r = \dfrac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$ to compute the correlation coefficient to the following data

x	78	89	97	69	59	79	68	57
y	125	137	156	112	107	138	123	108

Solution: Consider,

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$z = x-y$	$z-\bar{z}$	$(z-\bar{z})^2$
78	125	3.5	-0.8	12.25	0.64	-47	4.25	18.0625
89	137	14.5	11.2	210.25	125.44	-48	3.25	10.5625
97	156	22.5	30.2	506.25	912.04	-59	-7.75	60.0625
69	112	-5.5	-13.8	30.25	190.44	-43	8.25	68.0625
59	107	-15.5	-18.8	240.25	353.44	-48	3.25	10.5625
79	138	4.5	12.2	20.25	148.84	-59	-7.75	60.0625
68	123	-6.5	-2.8	42.25	7.84	-55	-3.75	14.0625
57	108	-17.5	-17.8	306.25	316.84	-51	0.25	0.0625
$\sum x =$ 596	$\sum y =$ 1006			$\sum(x-\bar{x})^2 = 1368$	$\sum(y-\bar{y})^2 = 2055.52$	$\sum z = -410$		$\sum(z-\bar{z})^2 = 241.5$

(i) $\bar{x} = \frac{\sum x}{n} = \frac{596}{8} = 74.5,\ \bar{y} = \frac{\sum y}{n} = \frac{1006}{8} = 125.75$ and $\bar{z} = x - y = \frac{\sum z}{n} = \frac{-410}{8} = -51.25$

(ii) $\sigma_x^2 = \frac{\sum (x-\bar{x})^2}{n} = \frac{1368}{8} = 171, \quad \sigma_x = \sqrt{171} = 13.0767$

$\sigma_y^2 = \frac{\sum (y-\bar{y})^2}{n} = \frac{2055.52}{8} = 256.94, \quad \sigma_y = \sqrt{256.94} = 16.0293$

$\sigma_{x-y}^2 = \sigma_z^2 = \frac{\sum (z-\bar{z})^2}{n} = \frac{241.5}{8} = 30.1875, \quad \sigma_{x-y} = \sigma_z = 30.1875 = 5.4943$

(iii) $r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x \sigma_y} = \frac{171 + 256.94 - 30.1875}{2 \times 13.0767 \times 16.0293} = 0.9488$

12. If θ is the angle between the two regression lines, prove that $\tan\theta = \frac{r^2-1}{r}\left(\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right)$

Solution: We know that the two regression lines

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \quad \cdots \quad \cdots \quad (1)$$

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \quad \cdots \quad \cdots \quad (2)$$

From (2), $y - \bar{y} = \frac{1}{r}\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

The slope of the line (1), $m_1 = r\frac{\sigma_y}{\sigma_x}$

The slope of the line (2), $m_2 = \frac{1}{r}\frac{\sigma_y}{\sigma_x}$

Therefore, $\tan\theta = \frac{m_1 - m_2}{1 + m_1 m_2} = \frac{r\frac{\sigma_y}{\sigma_x} - \frac{1}{r}\frac{\sigma_y}{\sigma_x}}{1 + \left(r\frac{\sigma_y}{\sigma_x}\right)\left(\frac{1}{r}\frac{\sigma_y}{\sigma_x}\right)} = \frac{\left(\frac{r^2-1}{r}\right)\frac{\sigma_y}{\sigma_x}}{\left(\frac{\sigma_x^2 + \sigma_y^2}{\sigma_x^2}\right)} = \frac{r^2-1}{r}\left(\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right)$

If θ is acute angle then $\tan\theta$ is positive, and therefore $\tan\theta = \left|\frac{r^2-1}{r}\right|\left(\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right)$

Exercise:

1. Determine the correlation coefficient for the following data

x	11.1	10.3	12	15.1	13.7	18.5	17.3	14.2	14.8	15.3
y	10.9	14.2	13.8	21.5	13.2	21.1	16.4	19.3	17.4	19.0

2. Compute the correlation coefficient to the following data

x	62	56	36	66	25	75	82	78
y	58	44	51	58	60	68	62	84

3. Compute the correlation coefficient to the following data

x	8	1	5	4	7
y	3	4	0	2	1

4. From the following data

x	50	60	70	90	100
y	65	51	40	26	08

Determine
(i) $\bar{x}$ and $\bar{y}$
(ii) σ_x and σ_y
(iii) $Cov(x, y)$
(iv) the correlation coefficient between x and y
(v) two regression lines

5. The equations of two regression lines obtained in a correlation analysis are $4x - 5y + 33 = 0$ and $20x - 9y = 107$. Compute (i) mean of x (ii) mean of y (iii) correlation coefficient between x and y

6. Psychological tests of intelligence and engineering ability were applied to 10 students. Here is a record of ungrouped data showing intelligence ratio (IR) and engineering ratio (ER). Calculate the coefficient of correlation

Student	A	B	C	D	E	F	G	H	I	J
IR	105	104	102	101	100	99	98	96	93	92
ER	101	103	100	98	95	96	104	92	97	94

$\bar{x} = 99, \bar{y} = 98, \sum(x - \bar{x})^2 = 170, \sum(y - \bar{y})^2 = 140, \sum(x - \bar{x})(y - \bar{y}) = 92$

Correlation coefficient $r = \dfrac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2}\sqrt{\sum(y - \bar{y})^2}} = 0.59$

7. Use the formula $r = \dfrac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$ to compute the correlation coefficient to the following data

X	62	56	36	66	25	75	82	78
Y	58	44	51	58	60	68	62	84

8. Given that $\bar{x} = 31.6, \bar{y} = 38, \sigma_x = 3.72, \sigma_y = 6.31$ and $r = -0.36$. Determine the two regression lines

Rank Correlation: The correlation between the ranks of the variables x and y is called the rank correlation

Rank Correlation Coefficient (or Spearman Rank Correlation Coefficient): It is denoted by $\rho(x, y)$ or ρ and is given by $\rho = 1 - \frac{6\sum d^2}{n\,(n^2-1)}$, where d is the difference between the ranks of corresponding values of x, y and n is the number of pairs of data points.

Repeated Values: If an item of x or y is repeated m times, then we give the average rank for the repeated items and add the factor $\frac{m\,(m^2-1)}{12}$ to $\sum d^2$ in the formula of ρ.

Problems:

1. Determine 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

Solution: The rank correlation of x, y is given by $\rho = 1 - \frac{6\sum d^2}{n\,(n^2-1)}$

The values of x in decreasing order: 80, 75, 75, 68, 64, 64, 64, 55, 50, 40

The values of y in decreasing order: 81, 70, 68, 68, 62, 60, 58, 50, 48, 45

Consider,

x	y	Rank x	Rank y	$d =$ Rank $x -$ Rank 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
					$\sum d^2 = 72$

Here, in the values of x, 75 is repeated 2 times and 64 is repeated 3 times

And in the values of y, 68 is repeated 2 times

Therefore the correction factor is given by

$$\frac{\sum m\,(m^2-1)}{12} = \frac{2(2^2-1)}{12} + \frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12} = \frac{1}{2} + 2 + \frac{1}{2} = 3$$

Now, $n = 10$, $\sum d^2 = 72$ and $\sum d^2 + \frac{\sum m\,(m^2-1)}{12} = 72 + 3 = 75$

Hence the Rank correlation coefficient,

$$\rho = 1 - \frac{6\sum d^2}{n\,(n^2-1)} = 1 - \frac{6\,(75)}{10\,(10^2-1)} = 1 - \frac{450}{990} = 1 - 0.4545 = 0.5455$$

2. Determine the rank correlation coefficient for the following data

x	10	15	12	17	13	16	24	14	22
y	30	42	45	46	33	34	40	35	39

Solution: The rank correlation of x, y is given by $\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$

Consider,

x	y	Rank x	Rank y	$d = \text{Rank } x - \text{Rank } y$	d^2
10	30	9	9	0	0
15	42	5	3	2	4
12	45	8	2	6	36
17	46	3	1	2	4
13	33	7	8	-1	1
16	34	4	7	-3	9
24	40	1	4	-3	9
14	39	6	5	1	1
22	35	2	6	-4	16
					$\sum d^2 = 80$

Here, there are no repetitions in the values of x and y

Now, $n = 9$, $\sum d^2 = 80$

Hence the Rank correlation coefficient,

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(80)}{9(9^2-1)} = 1 - \frac{480}{720} = 1 - 0.6667 = 0.3333$$

3. Ten participants in a contest are ranked by two judges as follows

x	1	6	5	10	3	2	4	9	7	8
y	6	4	9	8	1	2	3	10	5	7

Calculate the rank correlation coefficient

Solution: The rank correlation of x, y is given by $\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$

Consider,

x	y	Rank x	Rank y	$d = \text{Rank } x - \text{Rank } y$	d^2
1	6	1	6	-5	25
6	4	6	4	2	4
5	9	5	9	-4	16
10	8	10	8	2	4
3	1	3	1	2	4
2	2	2	2	0	0
4	3	4	3	1	1
9	10	9	10	-1	1
7	5	7	5	2	4
8	7	8	7	1	1
					$\sum d^2 = 60$

Here, there are no repetitions in the values of x and y

Now, $n = 10$, $\sum d^2 = 60$

Hence the Rank correlation coefficient,

$$\rho = 1 - \frac{6\sum d^2}{n\,(n^2-1)} = 1 - \frac{6\,(60)}{10\,(10^2-1)} = 1 - \frac{360}{990} = 1 - 0.3636 = 0.6364$$

4. Determine the rank correlation coefficient for the following data

x	5	10	6	3	19	5	6	12	8	2	10	19
y	8	3	2	9	12	3	17	18	22	12	17	20

Solution: The rank correlation of x, y is given by $\rho = 1 - \dfrac{6\sum d^2}{n\,(n^2-1)}$

Consider,

x	y	Rank x	Rank y	d = Rank x – Rank y	d^2
5	8	9.5	9	0.5	0.25
10	3	4.5	10.5	-6	36
6	2	7.5	12	-4.5	20.25
3	9	11	8	3	9
19	12	1.5	6.5	-5	25
5	3	9.5	10.5	-1	1
6	17	7.5	4.5	3	9
12	18	3	3	0	0
8	22	6	1	5	25
2	12	12	6.5	5.5	30.25
10	17	4.5	4.5	0	0
19	20	1.5	2	-0.5	0.25
					$\sum d^2 = 156$

Here, in the values of x, 19 is repeated 2 times, 10 is repeated 2 times, 6 is repeated 2 times, and 5 is repeated 2 times

And in the values of y, 17 is repeated 2 times, 12 is repeated 2 times and 3 is repeated 2 times,

Therefore the correction factor is given by

$$\frac{\sum m\,(m^2-1)}{12} = 7 \times \frac{2(2^2-1)}{12} = \frac{7}{2} = 3.5$$

Now, $n = 12$, $\sum d^2 = 156$ and $\sum d^2 + \dfrac{\sum m\,(m^2-1)}{12} = 156 + 3.5 = 159.5$

Hence the Rank correlation coefficient,

$$\rho = 1 - \frac{6\sum d^2}{n\,(n^2-1)} = 1 - \frac{6\,(159.5)}{12\,(12^2-1)} = 1 - \frac{957}{1716} = 1 - 0.5577 = 0.4423$$

Exercise:

1. Determine the rank correlation coefficient for the following data

x	8	3	9	2	7	10	4	6	1	5
y	9	5	10	1	8	7	3	4	2	6

Ans: $n = 10$, $\sum d^2 = 24$ and $\rho = 0.8545$

2. Determine the rank correlation coefficient for the following data

x	78	56	36	66	25	75	82	62
y	84	44	57	58	60	68	62	58

Ans: $n = 10$, $\sum d^2 = 28.5$, $\dfrac{\sum m\,(m^2 - 1)}{12} = \dfrac{1}{2}$ and $\rho = 0.655$

3. Determine the rank correlation coefficient for the following data

x	65	63	67	64	68	62	70	66	68	67	69	71
y	68	66	68	65	69	66	68	65	71	67	68	70

Ans: $n = 12$, $\sum d^2 = 72.5$, $\dfrac{\sum m\,(m^2 - 1)}{12} = 7$ and $\rho = 0.722$