Probability and Statistics

Unit - I

CURVE FIT   TING AND STATISTICAL METHODS

Curve Fitting: Curve fitting by the method of least squares and fitting of the curves of the form, $y = ax + b$, $y = ax^2 + bx + c$, $y = ae^{bx}$ and $y = ax^b$

Statistical Methods: Measures of central tendency and dispersion. Correlation-Karl Pearson's coefficient of correlation-problems. Regression analysis- lines of regression, problems. Rank correlation.

Least Square method

The method of finding a specific relation $y = f(x)$ for the data to satisfy as accurately as possible and such an equation is called the best fitting equation or the curve of best fit.

Fitting of a straight line y = a + bx

Consider a set of n given values (x, y) for fitting the straight line $y = a + bx$ where a and b are parameters to be determined. We find the parameters a and b using the normal equations

$$na + b\sum x = \sum y$$

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

Fitting of a second degree parabola y = a + bx + cx^2

Consider a set of n given values (x, y) for fitting the curve $y = a + bx + cx^2$ where a, b and c are parameters to be determined. We find the parameters a, b and c using the normal equations

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

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

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

Note: The normal equations for fitting a straight line or parabola can be written instantly from the desired equation of the curve as follows

We first apply summation ($\sum$) to the desired equation keeping the constants a, b and c outside the summation where the summation of pure constant terms like $\sum a, \sum b, \sum c$ are to be written as na, nb, nc respectively

We then multiply the given equation by the independent variable x and apply summation again. This will be sufficient for fitting a straight line. However in the case of parabola we must also multiply by x^2 and apply summation.

Fitting of a curve of the form $y = ab^x$

Consider $y = ab^x$ (1)

Taking log on both sides, we get

$\log_e y = \log_e a + x \log_e b$

or $Y = A + BX$ (2)

where $Y = \log_e y$, $A = \log_e a$, $B = \log_e b$ and X = x.

Which is the same as $y = a + bx$, the normal equations associated with equation (2) are as follows

$$nA + B\sum X = \sum Y \qquad (3)$$

$$A\sum X + B\sum X^2 = \sum XY \qquad (4)$$

Solving (3) and (4) we obtain A and B.

But we have $\log_e a = A \Rightarrow a = e^A$

and $\log_e b = B \Rightarrow b = e^B$

Substitution of the values of a and b in (1) give us the best fitting curve $y = ab^x$ in the least square sense.

Note: We can also fit curves of the form $y = ae^{bx}$ (Exponential curve), $y = ax^b$ (Geometric curve) in the similar way.

Working procedure for problems:

Method I:

Step 1: We first write the normal equations appropriate to the curve of fit.

Step 2: We prepare the relevant table and the find the values of the summation present in the normal equations. We substitute these values to arrive a system of equations in the unknown parameters.

Step 3: We find the parameters by solving and substitute in the given equation.

Fitting of a straight line *y* = a + b*x*

Example: Fit a straight line $y = a + bx$ in the least square sense for the data

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

Solution: The normal equations for $y = a + bx$ are given by

$$na + b\sum x = \sum y$$

$$a\sum x + b\sum x^2 = \sum xy \qquad \text{Here n = 8}$$

The relevant table is as follows

x	y	xy	x^2
1	1	1	1
3	2	6	9
4	4	16	16
6	4	24	36
8	5	40	64
9	7	63	81
11	8	88	121
14	9	126	196
$\sum x = 56$	$\sum y = 40$	$\sum xy = 364$	$\sum x^2 = 524$

The normal equations becomes

$$8a + 56b = 40 \qquad (1)$$

$$56a + 524b = 364 \qquad (2)$$

Solving (1) and (2) we get a = 0.52, b = 0.64

The equation $y = a + bx$ becomes $y = 0.52 + 0.64x$.

Example: Find a law of the form $y = a + bx$ for the following data

x	100	120	140	160	180	200
y	45	55	60	70	80	85

Solution: The normal equations for $y = a + bx$ are given by

$$na + b\sum x = \sum y$$

$$a\sum x + b\sum x^2 = \sum xy \qquad \text{Here n = 6}$$

The relevant table is as follows

x	y	xy	x^2
100	45	4500	10000
120	55	6600	14400
140	60	8400	19600
160	70	11200	25600
180	80	14400	32400
200	85	17000	40000
$\sum x = 900$	$\sum y = 395$	$\sum xy = 62100$	$\sum x^2 = 142000$

The normal equations becomes

$$6a + 900b = 395 \qquad (1)$$

$$900a + 142000b = 62100 \qquad (2)$$

Solving (1) and (2) we get a = 4.7619, b = 0.4071

The equation $y = a + bx$ becomes $y = 4.7619 + 0.4071x$

Example: Fit a straight line for the data given below using the method of least squares

x	1	2	3	4	6	8
y	2.4	3	3.6	4	5	6

Solution: The normal equations for $y = a + bx$ are given by

$$na + b\sum x = \sum y$$

$$a\sum x + b\sum x^2 = \sum xy \qquad \text{Here } n = 6$$

The relevant table is as follows

x	y	xy	x^2
1	2.4	2.4	1
2	3	6	4
3	3.6	10.8	9
4	4	16	16
6	5	30	36
8	6	48	64
$\sum x = 24$	$\sum y = 24$	$\sum xy = 113.2$	$\sum x^2 = 130$

The normal equations becomes

$$6a + 24b = 24 \qquad (1)$$

$$24a + 130b = 113.2 \qquad (2)$$

Solving (1) and (2) we get a = 1.9764, b = 0.5058

The equation $y = a + bx$ becomes $y = 1.9764 + 0.5058x$.

Example: Find a law of the form $y = a + bx$ for the following data

Year (x)	1911	1921	1931	1941	1951
Production (y)	8	10	12	10	6

Solution: The normal equations for $y = a + bx$ are given by

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$ Here n = 5

The relevant table is as follows

x	y	xy	x^2
1911	8	15288	3651921
1921	10	19210	3690241
1931	12	23172	3728761
1941	10	19410	3767481
1951	6	11706	3806401
$\sum x = 9655$	$\sum y = 46$	$\sum xy = 88786$	$\sum x^2 = 18644805$

The normal equations becomes

$$46 = 5a + 9655b \quad (1)$$

$$88786 = 9655a + 18644805b \quad (2)$$

Solving (1) and (2) we get a = 86.44, b = - 0.04

The equation $y = a + bx$ becomes $y = 86.44 - 0.04x$

Alternative Method

The normal equations for $y = a + bX$ are given by

$$\sum y = na + b\sum X$$

$$\sum Xy = a\sum X + b\sum X^2$$ Here n = 5 and $X = x - 1931$

The relevant table is as follows

x	$X = x - 1931$	y	Xy	X^2
1911	-20	8	-160	400
1921	-10	10	-100	100
1931	0	12	0	0
1941	10	10	100	100

1951	20	6	120	400
	$\sum X = 0$	$\sum y = 46$	$\sum Xy = -40$	$\sum X^2 = 1000$

The normal equations becomes

$$46 = 5a + 0b$$

$$a = \frac{46}{5} = 9.2$$

$$-40 = 0a + 1000b$$

$$b = \frac{40}{1000} = -0.04$$

The equation $y = a + bX$ becomes $y = 9.2 - 0.04X$

Put $X = x - 1931$

$$y = 9.2 - 0.04(x - 1931)$$

$$y = 86.44 - 0.04x$$

Examples:

1.Find the equation of the beat fitting straight line for the following data

x	1	2	3	4	5
y	14	13	9	5	2

2. Fit a straight line for the data given below using the method of least squares

x	0	1	2	3	4	5
y	9	8	24	28	26	20

3. Fit a straight line for the data given below using the method of least squares

x	62	64	65	69	70	71	72
y	65.7	66.8	67.2	69.3	69.8	70.5	70.9

4. Find a law of the form $y = a + bx$ for the following data

x	50	70	100	120
y	12	15	21	25

5. A simply supported beam carries a concentrated load P at its midpoint corresponding to various values of Pthe maximum deflection Y is measured and is given below

P	100	120	140	160	180	200
Y	0.45	0.55	0.60	0.70	0.80	0.85

Find a law of the form $Y = a + bP$ andhence estimate Y when P = 150.

Fitting of a second degree parabola $y = a + bx + cx^2$

Example: Fit a parabola of second degree $y = a + bx + cx^2$ for the data

x	0	1	2	3	4
y	1	1.8	1.3	2.5	2.3

Solution: The normal equations for $y = a + bx + cx^2$ are given by

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

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

$$a\sum x^2 + b\sum x^3 + c\sum x^4 = \sum x^2 y \qquad \text{Here } n = 5$$

The relevant table is as follows

x	y	xy	x^2	x^2y	x^3	x^4
0	1	0	0	0	0	0
1	1.8	1.8	1	1.8	1	1
2	1.3	2.6	4	5.2	8	16
3	2.5	7.5	9	22.5	27	81
4	2.3	9.2	16	36.8	64	256
$\sum x = 10$	$\sum y =$ 8.9	$\sum xy = 21.1$	$\sum x^2 = 30$	$\sum x^2 y =$ 66.3	$\sum x^3 = 100$	$\sum x^4 = 354$

The normal equations become

$$5a + 10b + 30c = 8.9 \qquad (1)$$

$$10a + 30b + 100c = 21.1 \qquad (2)$$

$$30a + 100b + 354c = 66.3 \qquad (3)$$

Solving (1), (2) and (3) we get a = 1.078, b = 0.414 and c = - 0.021

The equation $y = a + bx + cx^2$ becomes $y = 1.078 + 0.414x - 0.021x^2$.

Example: Fit a parabola $y = a + bx + cx^2$ by the method of least square for the data

x	2	4	6	8	10
y	3.07	12.85	31.47	57.38	91.29

Solution: The normal equations for $y = a + bx + cx^2$ are given by

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

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

$$a\sum x^2 + b\sum x^3 + c\sum x^4 = \sum x^2 y \qquad \text{Here } n = 5$$

The relevant table is as follows

x	y	xy	x^2	x^2y	x^3	x^4
2	3.07	6.14	4	12.28	8	16
4	12.85	51.4	16	205.6	64	256
6	31.47	188.82	36	1132.92	216	1296
8	57.38	459.04	64	3672.32	512	4096
10	91.29	912.9	100	9129	1000	10000
$\sum x = 30$	$\sum y =$ 196.06	$\sum xy =$ 1618.3	$\sum x^2 =$ 220	$\sum x^2 y =$ 14152.12	$\sum x^3 =$ 1800	$\sum x^4 =$ 15664

The normal equations become

$$5a + 30b + 220c = 196.06 \qquad (1)$$

$$30a + 220b + 1800c = 1618.3 \qquad (2)$$

$$220a + 1800b + 15664c = 14152.12 \qquad (3)$$

Solving (1), (2) and (3) we get a = 0.696, b = - 0.855 and c = 0.992

The equation $y = a + bx + cx^2$ becomes $y = 0.696 - 0.855x + 0.992x^2$.

Example: Fit a parabola $y = a + bx + cx^2$ by the method of least square to 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 normal equations for $y = a + bx + cx^2$ are given by

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

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

$$a\sum x^2 + b\sum x^3 + c\sum x^4 = \sum x^2 y \qquad \text{Here } n = 7$$

The relevant table is as follows

x	y	xy	x^2	x^2y	x^3	x^4
1.0	1.1	1.1	1	1.1	1	1
1.5	1.3	1.95	2.25	2.925	3.375	5.0625
2.0	1.6	3.2	4	6.4	8	16

2.5	2.0	5	6.25	12.5	15.625	39.0625
3.0	2.7	8.1	9	24.3	27	81
3.5	3.4	11.9	12.25	41.65	42.875	150.0625
4.0	4.1	16.4	16	65.6	64	256
$\sum x =$ 17.5	$\sum y =$ 16.2	$\sum xy = 47.65$	$\sum x^2 =$ 50.75	$\sum x^2 y =$ 154.475	$\sum x^3 =$ 161.875	$\sum x^4 =$ 548.1875

The normal equations become

$$7a + 17.5b + 50.75c = 16.2 \quad (1)$$

$$17.5a + 50.75b + 161.875c = 47.65 \quad (2)$$

$$50.75a + 161.875b + 548.1875c = 154.475 \quad (3)$$

Solving (1), (2) and (3) we get a = 1.0357, b = - 0.1928 and c = 0.2428

The equation $y = a + bx + cx^2$ becomes $y = 1.0357 - 0.1928x + 0.2428x^2$.

Example: Fit a parabola of second degree $y = a + bx + cx^2$ in the least square sense for the data

x	10	20	30	40	50	60
y	157	179	210	252	302	361

Solution: The normal equations for $y = a + bx + cx^2$ are given by

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

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

$$a\sum x^2 + b\sum x^3 + c\sum x^4 = \sum x^2 y \qquad \text{Here } n = 7$$

The relevant table is as follows

x	y	xy	x^2	x^2y	x^3	x^4
10	157	1570	100	15700	1000	10000
20	179	3580	400	71600	8000	160000
30	210	6300	900	189000	27000	810000
40	252	10080	1600	403200	64000	2560000
50	302	15100	2500	755000	125000	6250000
60	361	21660	3600	1299600	216000	12960000
$\sum x =$ 210	$\sum y =$ 1461	$\sum xy =$ 58290	$\sum x^2 =$ 9100	$\sum x^2 y =$ 2734100	$\sum x^3 =$ 441000	$\sum x^4 =$ 22750000

The normal equations become

$$6a + 210b + 9100c = 1461 \quad (1)$$

$$210a + 9100b + 441000c = 58290 \quad (2)$$

$$9100a + 441000b + 22750000c = 2734100 \qquad (3)$$

Solving (1), (2) and (3) we get a = 143.9, b = 0.8260 and c = 0.0466

The equation $y = a + bx + cx^2$ becomes $y = 143.9 + 0.826x + 0.0466x^2$.

Example: Fit a second degree parabola $y = a + bx + cx^2$ in the least square sense for the following data and hence estimate y at x = 6

x	1	2	3	4	5
y	10	12	13	16	19

Solution: The normal equations for $y = a + bx + cx^2$ are given by

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

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

$$a\sum x^2 + b\sum x^3 + c\sum x^4 = \sum x^2 y \qquad \text{Here } n = 5$$

The relevant table is as follows

x	y	xy	x^2	x^2y	x^3	x^4
1	10	10	1	10	1	1
2	12	24	4	48	8	16
3	13	39	9	117	27	81
4	16	64	16	256	64	256
5	19	95	25	475	125	625
$\sum x = 15$	$\sum y = 70$	$\sum xy = 232$	$\sum x^2 = 55$	$\sum x^2 y = 906$	$\sum x^3 = 225$	$\sum x^4 = 979$

The normal equations become

$$5a + 15b + 55c = 70 \qquad (1)$$

$$15a + 55b + 225c = 232 \qquad (2)$$

$$55a + 225b + 979c = 906 \qquad (3)$$

Solving (1), (2) and (3) we get a = 9.4, b = 0.4857 and c = 0.2857

The equation $y = a + bx + cx^2$ becomes $y = 9.4 + 0.4857x + 0.2857x^2$ (4)

Now y at x = 6, from (4)

$$y = 9.4 + 0.4857(6) + 0.2857(6)^2 = 22.5994$$

Examples:

1. Fit a parabola of second degree $y = a + bx + cx^2$ in the least square sense for the data

x	0	1	2	3	4	5	6
y	14	18	27	29	36	40	46

2. Fit a second degree polynomial of the form $y = a + bx + cx^2$ for the data

x	0	1	2	3	4	5
y	1	3	7	13	21	31

3. Fit a parabola of second degree $y = a + bx + cx^2$ in the least square sense for the data

x	1	2	3	4	5
y	25	28	33	39	46

4. Fit a parabola of second degree $y = a + bx + cx^2$ in the least square sense for the data

x	0	1	2	3	4
y	1	5	10	22	38

5. Fit a curve of the form $y = a_0 + a_1 x + a_2 x^2$ to the data

x	0	1	2	3	4
y	1	1.8	1.3	2.5	6.3

by the method of least squares

6. Fit a parabola $y = a + bx + cx^2$ to the data

x	1	2	3	4
y	1.7	1.8	2.3	3.2

by the method of least squares

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

x	-2	-1	0	1	2
y	-3.150	-1.390	0.620	2.886	5.378

8. Fit a parabola $y = a + bx + cx^2$ by the method of least square to the following data

x	-3	-2	-1	0	1	2	3
y	4.63	2.11	0.67	0.09	0.63	2.15	4.58

Fitting of a curve of the form $y = ab^x$, $y = ae^{bx}$ (Exponential curve), $y = ax^b$ (Geometric curve)

Example: Fit a curve of the form $y = ab^x$ in the least square sense for the following data

x	0	2	4	5	7	10
y	100	120	256	390	710	1600

Solution: Consider $y = ab^x$

Take log on both sides

$\log_e y = \log_e a + x \log_e b$

Let us write this in the form

$Y = A + BX$

where $Y = \log_e y$, $A = \log_e a$, $B = \log_e b$, $X = x$

The associated normal equations are

$$nA + B\sum X = \sum Y$$

$$A\sum X + B\sum X^2 = \sum XY \qquad \text{Here n = 6}$$

The relevant table is as follows

$X = x$	y	$Y = \log_e y$	XY	X^2
0	100	4.6051	0	0
2	120	4.7874	9.5748	4
4	256	5.5451	22.1804	16
5	390	5.9661	29.8305	25
7	710	6.5652	45.9564	49
10	1600	7.3777	73.777	100
$\sum X = 28$		$\sum Y =$ 34.8466	$\sum XY =$ 181.3191	$\sum X^2 = 194$

The normal equations becomes

$$6A + 28B = 34.8466 \qquad (1)$$

$$28A + 194B = 181.3191 \qquad (2)$$

Solving (1) and (2) we get A = 4.4297, B = 0.2952

But $A = \log_e a \Rightarrow a = e^A \Rightarrow a = e^{4.4297} = 83.9062$

$B = \log_e b \Rightarrow b = e^B \Rightarrow b = e^{0.2952} = 1.3433$

Thus the required curve is $y = (83.9062)(1.3433)^x$

Example: Fit a curve of the form $y = ab^x$ in the least square sense for the following data

x	1	2	3	4	5	6	7
y	87	97	113	129	202	195	193

Solution: Consider $y = ab^x$

Take log on both sides

$\log_e y = \log_e a + x \log_e b$

Let us write this in the form

$Y = A + BX$

where $Y = \log_e y$, $A = \log_e a$, $B = \log_e b$, $X = x$

The associated normal equations are

$$nA + B\sum X = \sum Y$$

$$A\sum X + B\sum X^2 = \sum XY \qquad \text{Here } n = 7$$

The relevant table is as follows

$X = x$	y	$Y = \log_e y$	XY	X^2
1	87	4.4659	4.4659	1
2	97	4.5747	9.1494	4
3	113	4.7273	14.1819	9
4	129	4.8598	19.4392	16
5	202	5.3082	26.541	25
6	195	5.2729	31.6374	36
7	193	5.2626	36.8382	49
$\sum X = 28$		$\sum Y =$ 34.4714	$\sum XY =$ 142.253	$\sum X^2 = 140$

The normal equations becomes

$$7A + 28B = 34.4714 \qquad (1)$$

$$28A + 140B = 142.253 \qquad (2)$$

Solving (1) and (2) we get A = 4.3005, B = 0.1559

But $A = \log_e a \Rightarrow a = e^A \Rightarrow a = e^{4.3005} = 73.7366$

$B = \log_e b \Rightarrow b = e^B \Rightarrow b = e^{0.1559} = 1.1687$

Thus the required curve is $y = (73.7366)(1.1687)^x$

Example: Fit a curve of the form $y = ae^{bx}$ for the data

x	0	2	4
y	8.12	10	31.82

Solution: Consider $y = ae^{bx}$

Take log on both sides

$\log_e y = \log_e a + bx$

Let us write this in the form

$Y = A + BX$

where $Y = \log_e y$, $A = \log_e a$, $B = b$, $X = x$

The associated normal equations are

$$nA + B\sum X = \sum Y$$

$$A\sum X + B\sum X^2 = \sum XY$$ Here $n = 3$

The relevant table is as follows

$X = x$	y	$Y = \log_e y$	XY	X^2
0	8.12	2.0943	0	0
2	10	2.3025	4.605	4
4	31.82	3.4600	13.84	16
$\sum X = 6$		$\sum Y =$ 7.8568	$\sum XY =$ 18.445	$\sum X^2 = 20$

The normal equations becomes

$$3A + 6B = 7.8568 \qquad (1)$$

$$6A + 20B = 18.445 \qquad (2)$$

Solving (1) and (2) we get $A = 1.9360$, $B = 0.3414$

But $A = \log_e a \Rightarrow a = e^A \Rightarrow a = e^{1.936} = 6.9309$

$B = 0.3414 \Rightarrow b = 0.3414$

Thus the required curve is $y = (6.9309)e^{0.3414x}$

Example: Fit a curve of the form $y = ax^b$ for the data

x	1	2	3	4	5	6
y	2.98	4.26	5.21	6.1	6.8	7.5

Solution: Consider $y = ax^b$

Take log on both sides

$\log_e y = \log_e a + b\log_e x$

Let us write this in the form

$Y = A + BX$

where $Y = \log_e y$, $A = \log_e a$, $B = b$, $X = \log_e x$

The associated normal equations are

$$nA + B\sum X = \sum Y$$

$$A\sum X + B\sum X^2 = \sum XY \qquad \text{Here } n = 6$$

The relevant table is as follows

x	$X = \log_e x$	y	$Y = \log_e y$	XY	X^2
1	0	2.98	1.0919	0	0
2	0.6931	4.26	1.4492	1.0044	0.4803
3	1.0986	5.21	1.6505	1.8132	1.2069
4	1.3862	6.1	1.8082	2.5065	1.9215
5	1.6094	6.8	1.9169	3.0860	2.5901
6	1.7917	7.5	2.0149	3.6100	3.2101
	$\sum X = 6.579$		$\sum Y =$ 9.9316	$\sum XY =$ 12.0201	$\sum X^2 =$ 9.4089

The normal equations becomes

$6A + 6.579B = 9.9316$ (1)

$6.579A + 9.4089B = 12.0201$ (2)

Solving (1) and (2) we get A = 1.0907, B = 0.5148

But $A = \log_e a \Rightarrow a = e^A \Rightarrow a = e^{1.0907} = 2.9763$

$B = b \Rightarrow b = 0.5148$

Thus the required curve is $y = (2.9763)x^{0.5148}$

Example: At constant temperature, the pressure P and the volume V of a gas are connected by the relation PV^{γ} = constant. Find the best fitting equation of this form to the following data and estimate V when P = 4

P(Kg/Sq. cm	0.5	1.0	1.5	2.0	2.5	3.0
V(c.c)	1620	1000	750	620	520	460

Solution: The given relation is $PV^{\gamma} = k$, where k is a constant

Take log on both sides

$\log_e P + \gamma \log_e = \log_e k$

$\log_e P = \log_e k - \gamma \log_e V$

Let us write this in the form

$Y = A + BX$

where $Y = \log_e P$, $A = \log_e k$, $B = -\gamma$, $X = \log_e V$

The associated normal equations are

$$nA + B\sum X = \sum Y$$

$$A\sum X + B\sum X^2 = \sum XY \qquad \text{Here } n = 6$$

The relevant table is as follows

V	$X = \log_e V$	P	$Y = \log_e P$	XY	X^2
1620	7.3901	0.5	-0.6931	-5.1220	54.6735
1000	6.9077	1.0	0	0	47.7163
750	6.6200	1.5	0.4054	2.6837	43.8244
620	6.4297	2.0	0.6931	4.4564	41.3410
520	6.2538	2.5	0.9162	5.7297	39.1100
460	6.1312	3.0	1.0986	6.7357	37.5916
	$\sum X$ = 39.7325		$\sum Y$ = 2.4202	$\sum XY$ = 14.4835	$\sum X^2$ = 264.2568

The normal equations becomes

$$6A + 39.7325B = 2.4202 \qquad (1)$$

$$39.7325A + 264.2568B = 14.4835 \qquad (2)$$

Solving (1) and (2) we get A = 9.3297, B = - 1.3479

But $A = \log_e k \Rightarrow k = e^A \Rightarrow k = e^{9.3297} = 11267.7506$

$B = -\gamma \Rightarrow \gamma = -B = 1.3479$

Thus the required relation is $PV^{1.3479} = 11267.7506 \qquad (3)$

When P = 4, from equation (4), we have

$$(4)V^{1.3479} = 11267.7506$$

$$V^{1.3479} = \frac{11267.7506}{4} = 2816.9376$$

$$V^{1.3479} = 2816.9376 \Rightarrow V = (2816.9376)^{1/1.3479}$$

$$V = 362.5532$$

Examples:

1.Fit a curve of the form $y = ab^x$ in the least square sense for the following data

x	1	2	3	4	5	6	7	8
y	1.0	1.2	1.8	2.5	3.6	4.7	6.6	9.1

2. Fit a curve of the form $y = ab^x$ in the least square sense for the following data and hence estimate y when x = 8.

x	0	1	2	3	4	5	6
y	32	47	65	92	132	190	275

3. Fit a curve of the form $y = ae^{bx}$ for the data

x	5	6	7	8	9	10
y	133	55	23	7	2	2

Statistical Methods

Measure of Central Tendency:

A measure of central tendency describes a set of data by identifying the central position in the data set at a single value.

The commonly used measures of central value are mean, median and mode.

Arithmetic Mean:

If $x = \{x_1, x_2, x_3, ..., x_n\}$ are the set of all '*n*' values of a variate, then the Arithmetic Mean (simply mean) is given by

1) **Direct Method**

$$\bar{x} = \frac{\sum x_i}{n} \text{ and } \bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

2) **Step Deviation Method(Assumed Mean Method)**

$$\bar{x} = A + \frac{\sum f_i u_i}{\sum f_i} \times h \text{ where } u_i = \frac{x_i - A}{h}$$

3) **Continuous Series:**

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$ where x_i is the mid value of the i^{th} class interval.

Example: Calculate the arithmetic mean for the following data $7, 6, 8, 10, 13, 14$ by direct method

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

$$\bar{x} = \frac{7+6+8+10+13+14}{6} = 9.6667$$

Example: Calculate the Arithmetic Mean for the following series.

Marks	5	10	15	20	25	30
No. of students	20	43	75	76	72	45

Solution: Direct Method

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

Marks(x_i)	No. of students (f_i)	$x_1 f_i$
5	20	100
10	43	430
15	75	1125
20	76	1520
25	72	1800
30	45	1350
	$\sum f_i$ =331	$\sum f_i x_i$ =6325

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

$$\bar{x} = \frac{6325}{331} = 19.1087$$

4) Step Deviation Method (Assumed Mean Method)

$\bar{x} = A + \frac{\sum f_i u_i}{\sum f_i} \times h$ where $u_i = \frac{x_i - A}{h}$

Let A = 20 and here *h* = 5

Marks(x_i)	$u_i = \frac{x_i - A}{h} = \frac{x_i - 20}{5}$	No. of students (f_i)	$u_i f_i$
5	-3	20	-60
10	-2	43	-86
15	-1	75	-75
20	0	76	0
25	1	72	72
30	2	45	90
		$\sum f_i$ =331	$\sum f_i u_i$ =-59

$$\bar{x} = A + \frac{\sum f_i u_i}{\sum f_i} \times h$$

$$\bar{x} = 20 + \frac{-59}{331} \times 5 = 19.1087$$

Standard Deviation:

It is defined as the positive square-root of the arithmetic mean of the Square of the deviations of the given observation from their arithmetic mean.
The standard deviation is denoted by the Greek letter σ (sigma).

1) Calculation of Standard deviation - Individual Series:

There are two methods of calculating Standard deviation in an individual series.
a) Deviations taken from Actual mean

$$\sigma = \sqrt{\left(\frac{\sum x^2}{n}\right)} = \sqrt{\left(\frac{\sum (x-\bar{x})^2}{n}\right)}$$

b) Deviation taken from Assumed mean

$$\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2} \text{ Where } d = x - A$$

Example: Calculate the standard deviation from the following data 14, 22, 9, 15, 20, 17, 12, 11.

Solution: Standard deviation from actual mean

Value (x)	$x-\bar{x}$	$(x-\bar{x})^2$
14	-1	1
22	7	49
9	-6	36
15	0	0
20	5	25
17	2	4
12	-3	9
11	-4	16
$\sum x = 120$		$\sum(x-\bar{x})^2 = 140$

$$\bar{x} = \frac{120}{8} = 15$$

$$\sigma = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$
$$= \sqrt{\frac{140}{8}} = \sqrt{17.5} = 4.18.$$

Example: Find standard deviation for the following data. Number of seeds per fruit is given by $6, 7, 9, 11, 12, 14, 17, 20, 24, 30.$

Solution: Standard deviation from actual mean

Value (x)	$x-\bar{x}$	$(x-\bar{x})^2$
6	-9	81
7	-8	64
9	-6	36
11	-4	16
12	-3	9
14	-1	1
17	2	4
20	5	25
24	9	81
30	15	225
$\sum x = 150$		$\sum (x-\bar{x})^2 = 542$

$$\bar{x} = \frac{150}{10} = 15$$
$$\sigma = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$
$$= \sqrt{\frac{542}{10}} = \sqrt{54.2} = 7.362.$$

Example: The table below gives the marks obtained by 10 students in statistics. Calculate the standard deviation by assumed mean method

Student number	1	2	3	4	5	6	7	8	9	10
Marks(x)	43	48	65	5	31	60	37	48	8	59

Solution: Standard deviation from actual mean

Value (x)	$x-\overline{x}$	$\left(x-\overline{x}\right)^2$
43	2.6	6.76
48	7.6	57.76
65	24.6	605.16
5	-35.4	1253.16
31	-9.4	88.36
60	19.6	384.16
37	-3.4	11.56
48	7.6	57.76
8	-32.4	1049.76
59	18.6	345.96
$\sum x = 404$		$\sum\left(x-\overline{x}\right)^2 = 3859.9$

$$\overline{x} = \frac{404}{10} = 40.4$$

$$\sigma = \sqrt{\frac{\sum\left(x-\overline{x}\right)^2}{n}}$$

$$= \sqrt{\frac{3859.9}{10}} = \sqrt{385.99} = 19.6466.$$

Alternate Method:

Standard deviation from Assumed mean

$$\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2} \text{ Where } d = x - A$$

Let A = 40

Value (x)	$d = x - A$	d^2
43	3	9
48	8	64
65	25	625
5	-35	1225
31	-9	81
60	20	400
37	-3	9
48	8	64
8	-32	1024
59	19	361

$\sum x = 404$	$\sum d = 4$	$\sum d^2 = 3862$

$$\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2}$$

$$\sigma = \sqrt{\frac{3862}{10} - \left(\frac{4}{10}\right)^2} = \sqrt{386.2 - 0.16} = \sqrt{386.04} = 19.6479.$$

2) Calculation of standard deviation - Discrete Series:

There are three methods for calculating standard deviation in discrete series:

(a) Actual mean methods

$$\sigma = \sqrt{\left(\frac{\sum fd^2}{\sum f}\right)} \text{ Where } d = x - \bar{x}$$

(b) Assumed mean method

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2} \text{ Where } d = x - A$$

(c) Step-deviation method.

$$\sigma = \sqrt{\frac{\sum fd'^2}{\sum f} - \left(\frac{\sum fd'}{\sum f}\right)^2} \times C \text{ Where } d' = \frac{x - A}{C}$$

Example: Calculate the standard deviation from the following data

x	20	22	25	31	35	40	42	45
f	5	12	15	20	25	14	10	6

Solution: Standard deviation from assumed mean

x	f	$d = x - A$ ($A = 31$)	d^2	fd	fd^2
20	5	-11	121	-55	605
22	12	-9	81	-108	972
25	15	-6	36	-90	540
31	20	0	0	0	0
35	25	4	16	100	400
40	14	9	81	126	1134
42	10	11	121	110	1210
45	6	14	196	84	1176

	$\sum f = 107$			$\sum fd = 167$	$\sum fd^2 = 6037$

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2}$$

$$= \sqrt{\frac{6037}{107} - \left(\frac{167}{107}\right)^2} = \sqrt{56.42 - 2.44} = \sqrt{53.98} = 7.35$$

Example: Compute standard deviation from the following data by step deviation method

Marks	10	20	30	4	50	60
No. of students	8	12	20	10	25	3

Solution:

Marks (x)	f	$d' = \frac{x-30}{10}$	fd'	fd'^2
10	8	-2	-16	32
20	12	-1	-12	12
30	20	0	0	0
40	10	1	10	10
50	7	2	14	28
60	3	3	9	27
	$\sum f = 60$		$\sum fd' = 5$	$\sum fd'^2 = 109$

$$\sigma = \sqrt{\frac{\sum fd'^2}{\sum f} - \left(\frac{\sum fd'}{\sum f}\right)^2} \times C$$

$$\sigma = \sqrt{\frac{109}{60} - \left(\frac{5}{60}\right)^2} \times 10$$

$$= \sqrt{1.817 - 0.0069} \times 10 = 1.345 \times 10 = 13.45$$

Example: Calculate the standard deviation of the following

Size	6	7	30	9	1	11	12
Frequency	3	6	9	13	8	5	4

Solution: Standard deviation from assumed mean

x	f	$d = x - A$ ($A = 9$)	d^2	fd	fd^2
6	3	-3	9	-9	27
7	6	-2	4	-12	24

30	9	21	441	189	3969
9	13	0	0	0	0
1	8	-8	64	-64	512
11	5	2	4	10	20
12	4	3	9	12	36
	$\sum f = 48$			$\sum fd = 126$	$\sum fd^2 = 4588$

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2}$$

$$= \sqrt{\frac{4588}{48} - \left(\frac{126}{48}\right)^2} = \sqrt{95.5833 - 6.8906} = \sqrt{88.6927} = 9.4176$$

Example: Find the standard deviation for the following data.

Waxy endospermic plants	7	8	9	10	11	12
Number of plants	13	13	18	17	15	14

Solution: Standard deviation from assumed mean

x	f	$d = x - A$ ($A = 9$)	d^2	fd	fd^2
7	13	-2	4	-26	52
8	13	-1	1	-13	13
9	18	0	0	0	0
10	17	1	1	17	17
11	15	2	4	30	60
12	14	3	9	42	126
	$\sum f = 90$			$\sum fd = 50$	$\sum fd^2 = 268$

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f} - \left(\frac{\sum fd}{\sum f}\right)^2}$$

$$= \sqrt{\frac{268}{90} - \left(\frac{50}{90}\right)^2} = \sqrt{2.9777 - 0.3086} = \sqrt{2.6691} = 1.6337$$

3) Calculation of standard Deviation-Continuous Series:

In the continuous series the method of calculating standard deviation is almost the same as in a discrete series. But in a continuous series, mid-values of the class intervals are to be found out. The step deviation method is widely used.

$$\sigma = \sqrt{\frac{\sum fd'^2}{\sum f} - \left(\frac{\sum fd'}{\sum f}\right)^2} \times C \text{ Where } d' = \frac{x - A}{C}$$

where C is the width of the interval

Example: The daily temperature recorded in a city in Russia in a year is given below. Calculate the standard deviation

Temperature C^0	No. of days
- 40 to - 30	10
- 30 to - 20	18
- 20 to - 10	30
- 10 to 0	42
0 to 10	65
10 to 20	180
20 to 30	20

Solution:

Temperature	Mid Value (m)	No. of days (f)	$d' = \frac{m-(-5)}{10}$	fd'	fd'^2
-40 to -30	-35	10	-3	-30	90
-30 to -20	-25	18	-2	-36	72
-20 to -10	-15	30	-1	-30	30
-10 to0	-5	42	0	0	0
0 to 10	5	65	1	65	65
10 to 20	15	180	2	360	720
20 to30	25	20	3	60	180

		$\sum f = 365$		$\sum fd' = 389$	$\sum fd'^2 = 1157$

$$\sigma = \sqrt{\frac{\sum fd'^2}{\sum f} - \left(\frac{\sum fd'}{\sum f}\right)^2} \times C$$

$$= \sqrt{\frac{1157}{365} - \left(\frac{389}{365}\right)^2} \times 10$$

$$= \sqrt{3.1699 - 1.1358} \times 10$$

$$= \sqrt{2.0341} \times 10 = 1.4262 \times 10 = 14.262$$

Example: Calculate the standard deviation from the following series

Class interval	5-15	15-25	25-35	35-45	45-55
Frequency	8	12	15	9	6

Solution:

Class interval	Mid Value (m)	Frequency (f)	$d' = \frac{m-30}{10}$	fd'	d'^2	fd'^2
5-15	10	8	-2	-16	4	32
15-25	20	12	-1	-12	1	12
25-35	30	15	0	0	0	0
35-45	40	9	1	9	1	9
45-55	50	6	2	12	4	24
		$\sum f = 50$		$\sum fd' = -7$		$\sum fd'^2 = 77$

$$\sigma = \sqrt{\frac{\sum fd'^2}{\sum f} - \left(\frac{\sum fd'}{\sum f}\right)^2} \times C$$

$$= \sqrt{\frac{77}{50} - \left(\frac{-7}{50}\right)^2} \times 10$$

$$= \sqrt{1.54 - 0.0196} \times 10$$

$$= \sqrt{1.5204} \times 10 = 1.233 \times 10 = 12.33$$

Example: Following is the distribution of persons according to different income groups. Calculate the standard deviation.

Income in Rs(100)	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of varieties	6	8	10	12	7	4	3

Solution:

Correlation:

Suppose two variables x and y are related in such a way that an increase in one is accompanied by an increase or decrease in the other. Such a relationship is called correlation (or covariation). If x and y increase or decrease together, then we say that x and y are positively (directly) correlated. On the other hand, if y decreases as x increases or vice-versa then we say that x and y are negatively (inversely) correlated.

For example, demand and price of a commodity are positively correlated, whereas supply and price are negatively correlated.

The numerical measure of correlation between two variables *x* and *y* is known as the co-efficient of correlation and it is defined as

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{n\sigma_x \sigma_y}$$

where *n* is the number of observations, $\bar{x} = \frac{\sum x}{n}$ is mean of x, $\bar{y} = \frac{\sum y}{n}$ is mean of y,

$\sigma_x = \sqrt{\frac{\sum (x-\bar{x})^2}{n}} = \sqrt{\frac{\sum x^2}{n} - (\bar{x})^2}$ is the standard deviation of x and

$\sigma_y = \sqrt{\frac{\sum (y-\bar{y})^2}{n}} = \sqrt{\frac{\sum y^2}{n} - (\bar{y})^2}$ is the standard deviation of y.

Alternate form (1):

If $X = x - \bar{x}$ and $Y = y - \bar{y}$

$$\sigma_x = \sqrt{\frac{\sum (x-\bar{x})^2}{n}} = \sqrt{\frac{\sum X^2}{n}}$$

$$\sigma_y = \sqrt{\frac{\sum (y-\bar{y})^2}{n}} = \sqrt{\frac{\sum Y^2}{n}}$$

or $\sigma_x \sigma_y = \sqrt{\frac{\sum X^2}{n}} \sqrt{\frac{\sum Y^2}{n}} \Rightarrow n\sigma_x \sigma_y = \sqrt{\sum X^2} \sqrt{\sum Y^2}$

Therefore $r = \dfrac{\sum XY}{\sqrt{\sum X^2}\sqrt{\sum Y^2}}$

Alternate form (2):

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x \sigma_y}$$

Property: The co-efficient of correlation numerically does not exceed unity.

Proof: We have to show that $-1 \le r \le 1$

Let $S = \dfrac{1}{2n} \sum \left(\dfrac{X}{\sigma_x} + \dfrac{Y}{\sigma_y} \right)^2$ and $S' = \dfrac{1}{2n} \sum \left(\dfrac{X}{\sigma_x} - \dfrac{Y}{\sigma_y} \right)^2$

where $X = x - \bar{x}$ and $Y = y - \bar{y}$

Obviously both S and S' are ≥ 0.

Now $S = \dfrac{1}{2n} \sum \left(\dfrac{X^2}{\sigma_x^2} + \dfrac{Y^2}{\sigma_y^2} + \dfrac{2XY}{\sigma_x \sigma_y} \right) \geq 0$

$$S = \frac{1}{2n} \left(\sum \frac{X^2}{\sigma_x^2} + \sum \frac{Y^2}{\sigma_y^2} + 2\sum \frac{XY}{\sigma_x \sigma_y} \right) \geq 0$$

$$S = \frac{1}{2}\left(\frac{1}{\sigma_x^2} \frac{\sum X^2}{n} + \frac{1}{\sigma_y^2} \frac{\sum Y^2}{n} + 2 \frac{\sum XY}{n\sigma_x \sigma_y} \right) \geq 0$$

$$S = \frac{1}{2}\left(\frac{1}{\sigma_x^2} \sigma_x^2 + \frac{1}{\sigma_y^2} \sigma_y^2 + 2r \right) \geq 0$$

$$S = \frac{1}{2}(1 + 1 + 2r) \geq 0$$

$$S = \frac{1}{2}(2 + 2r) \geq 0$$

$\Rightarrow 1 + r \geq 0$

$\Rightarrow -1 \leq r$ (1)

Similarly we can obtain $S' = \frac{1}{2}(2 - 2r) \geq 0$

$\Rightarrow 1 - r \geq 0$

$\Rightarrow r \leq 1$ (2)

From (1) and (2) $-1 \leq r \leq 1$

Regression

Regression is an estimation of one independent variable in terms of the other. If x and y are correlated, the best fitting straight line in the least square sense gives reasonably a good relation between x and y.

The best fitting straight line of the form y = ax + b (x being the independent variable) is called the regression line of y on x and x = ay +b (y being the independent variable) is called the regression line of x on y.

The regression line of y on x

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

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

where $b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\sum XY}{\sum X^2}$,

$X = x - \bar{x}$ and $Y = y - \bar{y}$

The regression line of x on y

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

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

where $b_{xy}=r\dfrac{\sigma_x}{\sigma_y}=\dfrac{\sum XY}{\sum Y^2}$,

Note: The values $b_{yx}=r\dfrac{\sigma_y}{\sigma_x}$ and $b_{xy}=r\dfrac{\sigma_x}{\sigma_y}$ are known as the regression co-efficients. Their product is equal to r^2

i.e., $r=\sqrt{b_{xy}\times b_{yx}}$

Example: Show that θ is the angle between the lines of regression then $\tan\theta=\dfrac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\left(\dfrac{1-r^2}{r}\right)$

Solution: We know that if θ is acute ~~th~~e angle between the lines $y=m_1x+c_1$ and $y=m_2x+c_2$ is given by $\tan\theta=\dfrac{m_2-m_1}{1+m_1m_2}$

We have the lines of regression

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

and $x-\bar{x}=r\dfrac{\sigma_x}{\sigma_y}(y-\bar{y})$ we write this equation as

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

Slope of equations (1) and (2) are respectively given by

$$m_1=r\frac{\sigma_y}{\sigma_x} \text{ and } m_2=\frac{\sigma_y}{r\sigma_x}$$

Substituting these in the formula for tanθ, we have

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

$$= \frac{\dfrac{\sigma_y}{\sigma_x}\left(\dfrac{1-r^2}{r}\right)}{\dfrac{\sigma_x^2 + \sigma_y^2}{\sigma_x^2}} = \frac{\sigma_x\sigma_y\left(\dfrac{1-r^2}{r}\right)}{\sigma_x^2 + \sigma_y^2}$$

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

Example: Calculate the co-efficient of correlation and obtain the lines of regression for the following data

x	1	2	3	4	5	6	7	8	9
y	9	8	10	12	11	13	14	16	15

Obtain an estimate for *y* which corresponds to *x* = 6.2.

Solution: Here n = 9

$$\bar{x} = \frac{\sum x}{n} = \frac{45}{9} = 5 \text{ and } \bar{y} = \frac{\sum y}{n} = \frac{108}{9} = 12$$

We prepare the following table

x	$X = x - \bar{x}$	X^2	y	$Y = y - \bar{y}$	Y^2	XY
1	-4	16	9	-3	9	12
2	-3	9	8	-4	16	12
3	-2	4	10	-2	4	4
4	-1	1	12	0	0	0
5	0	0	11	-1	1	0

6	1	1	13	1	1	1
7	2	4	14	2	4	4
8	3	9	16	4	16	12
9	4	16	15	3	9	12
$\sum x = 45$		$\sum X^2 = 60$	$\sum y = 108$		$\sum Y^2 = 60$	$\sum XY = 57$

Now $r = \dfrac{\sum XY}{\sqrt{\sum X^2}\sqrt{\sum Y^2}}$

$$r = \frac{57}{\sqrt{60}\sqrt{60}} = \frac{57}{60} = 0.95$$

$$\sigma_x^2 = \frac{\sum X^2}{n} = \frac{60}{9} = 6.6667$$

$$\sigma_x = 2.582$$

$$\sigma_y^2 = \frac{\sum Y^2}{n} = \frac{60}{9} = 6.6667$$

$$\sigma_y = 2.582$$

Therefore, the regression co-efficients are

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.95 \times \frac{2.582}{2.582} = 0.95$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.95 \times \frac{2.582}{2.582} = 0.95$$

Therefore, the line of regression of y on x is

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

$$y - 12 = 0.95(x - 5)$$

$$y = 0.95x + 7.25 \qquad (1)$$

The line of regression of x on y is

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

$$x - 5 = 0.95(y - 12)$$

$$x = 0.95y - 6.4 \qquad (2)$$

When x = 6.2 we find from equation (1) that y = 13.14.

Example: Find the correlation co-efficient and the regression lines for the following data

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

Find the best estimate for y when x = 3.5 and the best estimate for x when y = 3.5.

Solution: Here n = 9

$\bar{x} = \frac{\sum x}{n} = \frac{15}{5} = 3$ and $\bar{y} = \frac{\sum y}{n} = \frac{25}{5} = 5$

We prepare the following table

x	$X = x - \bar{x}$	X^2	y	$Y = y - \bar{y}$	Y^2	XY
1	-2	4	2	-3	9	6
2	-1	1	5	0	0	0
3	0	0	3	-2	4	0
4	1	1	8	3	9	3

5	2	4	7	2	4	4
$\sum x = 15$		$\sum X^2 = 10$	$\sum y = 25$		$\sum Y^2 =$ 26	$\sum XY = 13$

Now $r = \dfrac{\sum XY}{\sqrt{\sum X^2}\sqrt{\sum Y^2}}$

$$r = \frac{13}{\sqrt{10}\sqrt{26}} = \frac{13}{\sqrt{10 \times 26}} = 0.8$$

$$\sigma_x^2 = \frac{\sum X^2}{n} = \frac{10}{5} = 2$$

$$\sigma_x = 1.4142$$

$$\sigma_y^2 = \frac{\sum Y^2}{n} = \frac{26}{5} = 5.2$$

$$\sigma_y = 2.2803$$

Therefore, the regression co-efficients are

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.8 \times \frac{2.2803}{1.4142} = 1.2899$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.8 \times \frac{1.4142}{2.2803} = 0.4961$$

Therefore, the line of regression of y on x is

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

$$y - 5 = 1.2899(x - 3)$$

$$y = 1.2899x + 1.1303 \qquad (1)$$

The line of regression of x on y is

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

$$x - 3 = 0.4961(y - 5)$$

$$x = 0.4961y + 0.5195 \qquad (2)$$

estimate for y when x = 3.5 and the best estimate for x when y = 3.5.

When x = 3.5 we find from equation (1) that y = 5.64495

When y = 3.5 we find from equation (2) that x = 2.25585

Example: Find the correlation coefficient between x and y and regression lines y on x and x on y from the following data

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

Solution: Here n = 8

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

We prepare the following table

x	$X = x - \bar{x}$	X^2	y	$Y = y - \bar{y}$	Y^2	XY
78	3.5	12.25	125	-0.75	0.5625	-2.625
89	14.5	210.25	137	11.25	126.5625	163.125

97	22.5	506.25	156	30.25	915.0625	680.625
69	-5.5	30.25	112	-13.25	175.5625	72.875
59	-15.5	240.25	107	-18.75	351.5625	290.625
79	4.5	20.25	138	12.25	150.0625	55.125
68	-6.5	42.25	123	-2.75	7.5625	17.875
57	-17.5	306.25	108	-17.75	315.0625	310.625
$\sum x =$ 596		$\sum X^2 =$ 1368	$\sum y =$ 1006		$\sum Y^2$ =2055.5	$\sum XY =$ 1596.25

Now $r = \dfrac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$

$$r = \frac{1596.25}{\sqrt{1368}\sqrt{2055.5}} = \frac{1596.25}{1676.87} = 0.9519 \approx 0.95$$

$$\sigma_x^2 = \frac{\sum X^2}{n} = \frac{1368}{8} = 171$$

$$\sigma_x = 13.0766$$

$$\sigma_y^2 = \frac{\sum Y^2}{n} = \frac{2055.5}{8} = 256.9375$$

$$\sigma_y = 16.0292$$

Therefore, the regression co-efficients are

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.95 \times \frac{16.0292}{13.0766} = 1.1645$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.95 \times \frac{13.0766}{16.0292} = 0.7750$$

Therefore, the line of regression of y on x is

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

$$y - 125.75 = 1.1645(x - 74.5)$$

$$y = 1.1645x + 38.9947 \qquad (1)$$

The line of regression of y on x is

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

$$x - 74.5 = 0.775(y - 125.75)$$

$$x = 0.775y - 22.9562 \qquad (2)$$

Example: Obtain the lines of regression and hence find the co-efficient of correlation for the following data

x	1	3	4	2	5	8	9	10	13	15
y	8	6	10	8	12	16	16	10	32	32

Solution: Here n = 10

$\bar{x} = \frac{\sum x}{n} = \frac{70}{10} = 7$ and $\bar{y} = \frac{\sum y}{n} = \frac{150}{10} = 15$

We prepare the following table

x	$X = x - \bar{x}$	X^2	y	$Y = y - \bar{y}$	Y^2	XY
1	-6	36	8	-7	49	42

3	-4	16	6	-9	81	36
4	-3	9	10	-5	25	15
2	-5	25	8	-7	49	35
5	-2	4	12	-3	9	6
8	1	1	16	1	1	1
9	2	4	16	1	1	2
10	3	9	10	-5	25	-15
13	6	36	32	17	289	102
15	8	64	32	17	289	136
$\sum x = 70$		$\sum X^2 = 204$	$\sum y = 150$		$\sum Y^2 = 818$	$\sum XY = 360$

The regression co-efficients are

$$b_{yx} = \frac{\sum XY}{\sum X^2} = \frac{360}{204} = 1.7647 \quad b_{xy} = \frac{\sum XY}{\sum Y^2} = \frac{360}{818} = 0.44$$

Therefore the line of regression of y on x is

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

$$y - 15 = 1.7647(x - 7)$$

$$y = 1.7647x + 2.6471 \qquad (1)$$

The line of regression of x on y is

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

$$x - 7 = 0.44(y - 15)$$

$$x = 0.44y + 0.4 \qquad (2)$$

Co-efficient of correlation is

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{0.44 \times 1.7647} = 0.8812$$

Example: Find the co-efficient of correlation by obtaining the lines of regression

x	1	2	3	4	5	6	7
y	9	8	10	12	11	13	14

Solution: Here n = 7

$$\bar{x} = \frac{\sum x}{n} = \frac{28}{7} = 4 \text{ and } \bar{y} = \frac{\sum y}{n} = \frac{77}{7} = 11$$

We prepare the following table

x	$X = x - \bar{x}$	X^2	y	$Y = y - \bar{y}$	Y^2	XY
1	-3	9	9	-2	4	6
2	-2	4	8	-3	9	6
3	-1	1	10	-1	1	1
4	0	0	12	1	1	0
5	1	1	11	0	0	0
6	2	4	13	2	4	4
7	3	9	14	3	9	9
$\sum x = 28$		$\sum X^2 = 28$	$\sum y = 77$		$\sum Y^2 = 28$	$\sum XY = 26$

The regression co-efficients are

$$b_{yx} = \frac{\sum XY}{\sum X^2} = \frac{26}{28} = 0.9285$$

$$b_{xy} = \frac{\sum XY}{\sum Y^2} = \frac{26}{28} = 0.9285$$

Therefore, the line of regression of y on x is

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

$$y - 11 = 0.9285(x - 4)$$

$$y = 0.9285x + 7.286 \qquad (1)$$

The line of regression of x on y is

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

$$x-4=0.9285(y-11)$$

$$x=0.9285y-6.2135 \qquad (2)$$

Co-efficient of correlation is

$$r=\sqrt{b_{xy}\times b_{yx}}=\sqrt{0.9285\times 0.9285}=0.9285$$

Example: Find the co-efficient of correlation by obtaining the lines of regression

x	80	45	55	56	58	60	65	68	70	75	85
y	52	56	50	48	60	62	64	65	70	74	90

Example: A person while calculating the co-efficient of correlation between two variables x, y from a set of 25 observations obtain the following results $\sum x = 125$, $\sum y = 100$, $\sum xy = 508$, $\sum x^2 = 650$, $\sum y^2 = 460$. But it was later found that the pair of values $\begin{matrix} x:8 & 6 \\ y:12 & 8 \end{matrix}$ where wrongly copied as $\begin{matrix} x:8 & 6 \\ y:14 & 6 \end{matrix}$.

Obtain the correct value for the correlation co-efficient.

Solution: Computation for wrong values is as follows

x	y	xy	x^2	y^2
6	14	84	36	196
8	6	48	64	36
$\sum x = 14$	$\sum y = 20$	$\sum xy = 132$	$\sum x^2 = 100$	$\sum y^2 = 132$

Computation for correct values is as follows

x	y	xy	x^2	y^2
8	12	96	64	144

6	8	48	36	64
$\sum x = 14$	$\sum y = 20$	$\sum xy = 144$	$\sum x^2 = 100$	$\sum y^2 = 208$

It may be observed that the summations $\sum x$, $\sum y$, $\sum x^2$ are unchanged even after the correction. However, we have

correct $\sum xy = 508 - 132 + 144 = 520$

correct $\sum y^2 = 460 - 232 + 208 = 436$

Therefore correct values of the mean and standard deviation of x and y are as follows

$$\bar{x} = \frac{\sum x}{n} = \frac{125}{25} = 5 \qquad \bar{y} = \frac{\sum y}{n} = \frac{100}{25} = 4$$

$$\sigma_x = \sqrt{\frac{\sum x^2}{n} - (\bar{x})^2} = \sqrt{\frac{650}{25} - 5^2} = \sqrt{1} = 1$$

$$\sigma_y = \sqrt{\frac{\sum y^2}{n} - (\bar{y})^2} = \sqrt{\frac{436}{25} - 4^2} = \sqrt{1.44} = 1.2$$

We have $r = \dfrac{\sum (x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}$

$$= \frac{\sum xy - \sum x\bar{y} - \sum \bar{x}y + \sum \bar{x}\,\bar{y}}{n\sigma_x\sigma_y}$$

$$= \frac{1}{\sigma_x\sigma_y}\left\{\frac{\sum xy}{n} - \frac{\bar{y}\sum x}{n} - \frac{\bar{x}\sum y}{n} + \frac{n\bar{x}\,\bar{y}}{n}\right\}$$

$$= \frac{1}{\sigma_x\sigma_y}\left\{\frac{\sum xy}{n} - \bar{x}\,\bar{y} - \bar{x}\,\bar{y} + \bar{x}\,\bar{y}\right\}$$

$$= \frac{1}{\sigma_x\sigma_y}\left\{\frac{\sum xy}{n} - \bar{x}\,\bar{y}\right\}$$

$$= \frac{1}{(1)(1.2)}\left\{\frac{520}{25} - (5\times 4)\right\} = 0.6666 \approx 0.67$$

Hence the correct value of r = 0.67.

Example: The two lines of regression for the variables x and y are given by x = 19.3 – 0.87y and y = 11.64 – 0.90x. Find

(i) the mean values of x and y

(ii) co-efficient of correlation between x and y.

Solution: The given regression lines are

$$x = 19.3 - 0.87y \Rightarrow x + 0.87y = 19.3 \qquad (1)$$

$$y = 11.64 - 0.90x \Rightarrow 0.90x + y = 11.64 \qquad (2)$$

(i) The mean values of x and y

Since two regression lines always intersect at a point $(\bar{x}, \bar{y})$ where $\bar{x}$ is the mean value of x and $\bar{y}$ is the mean value of y.

Solving (1) and (2), we get $\bar{x} = 42.2728, \bar{y} = -26.4055$

(ii) Co-efficient of correlation between x and y

To find coefficient of correlation rearrange the the regression line in such a way that the coefficient of dependent variable is less than one at least in one equation

$$y = 11.64 - 0.90x$$

$$x = 19.3 - 0.87y$$

$$b_{yx} = -0.90, b_{xy} = -0.87$$

Hence the coefficient of correlation between x and y is given by

$$r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{(-0.90) \times (-0.87)} = 0.8848$$

Example: In a partially destroyed lab record of analysis of correlation data, the following results are available, variance of *x* is 9. Regression equations are 8x – 10y + 66 = 0 and 40x – 18y – 214 = 0. Find $\bar{x}, \bar{y}, \sigma_y$ and correlation co-efficient.

Solution: The given regression lines are

$$8x - 10y = -66 \qquad (1)$$

$$40x - 18y = 214 \qquad (2)$$

(i) The mean values of *x* and *y*

Since two regression lines always intersect at a point $(\bar{x}, \bar{y})$ where $\bar{x}$ is the mean value of *x* and $\bar{y}$ is the mean value of *y*.

Solving (1) and (2), we get $\bar{x} = 13, \bar{y} = 17$

(ii) Co-efficient of correlation between *x* and *y*

To find coefficient of correlation rearrange the the regression line in such a way that the coefficient of dependent variable is less than one at least in one equation

$$y = \frac{8}{10}x + \frac{66}{10}$$

$$x = \frac{9}{20}y + \frac{107}{20}$$

$$b_{yx} = \frac{8}{10} = 0.80, b_{xy} = \frac{9}{20} = 0.45$$

Hence the coefficient of correlation between *x* and *y* is given by

$$r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{(0.80) \times (0.45)} = 0.6$$

(iii) Standard deviation of *y*

Given that variance of *x* = *V* = $\sigma_x^2 = 9 \Rightarrow \sigma_x = 3$

Standard deviation of y = $\sigma_y = \frac{b_{yx} \times \sigma_x}{r} = \frac{0.8 \times 3}{0.6} = 4$

Examples:

1. Calculate the co-efficient of correlation and obtain the lines of regression for the following data

x	3	5	6	9	10	12	15	20	22	28
y	10	12	15	18	20	22	27	30	32	34

2. Find the correlation coefficient and the regression lines y on x and x on y for the following data

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

3.Find the correlation coefficient and the regression lines y on x and x on y for the following data

x	2	4	6	8	10
y	5	7	9	8	11

4. Find the co-efficient of correlation by obtaining the lines of regression

x	17	18	19	19	20	20	21	22	21	23
y	12	16	14	11	15	19	22	15	16	20

Rank Correlation

A group of n individuals may be arranged in order to merit with respect to some characteristics. The same group would give different orders for different characteristics. Considering the orders corresponding to two characteristics A and B, the correlation between these n pairs of ranks is called rank correlation in the characteristics A and B for that group of individuals.

Let x_i, y_i be the ranks of the ith individuals in A and B respectively. Assuming that no two individuals are bracketed equal in either case, each of the variables taking the values 1, 2, 3, 4, …..n, we have

Rank correlation between A and B is given by

$$\rho = 1 - \frac{6\sum d_i^2}{n^3 - n}$$

where $d_i = x_i - y_i$ difference between the ranks of the ith individuals in A and B respectively.

Example: 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 ρ.

Solution: If $d_i = x_i - y_i$, then $d_i = -5, 2, -4, 2, 2, 0, 1, -1, 2, 1$

$$\sum d_i^2 = 25 + 4 + 16 + 4 + 4 + 0 + 1 + 1 + 4 + 1 = 60$$

Hence $\rho = 1 - \frac{6\sum d_i^2}{n^3 - n} = 1 - \frac{6 \times 60}{10^3 - 10} = 1 - \frac{360}{990} = 0.6$ nearly.

Example: Three judges, *A*, *B*, *C*, give the following ranks. Find which pair of judges has common approach

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

Solution:

A(x_i)	B(y_i)	C(z_i)	$d_i(x_i, y_i) = x_i - y_i$	$d_i^2(x_i, y_i)$	$d_i(y_i, z_i) = y_i - z_i$	$d_i^2(y_i, z_i)$	$d_i(x_i, z_i) = x_i - z_i$	$d_i^2(x_i, z_i)$
1	3	6	-2	4	-3	9	-5	25
6	5	4	1	1	1	1	2	4
5	8	9	-3	9	-1	1	-4	16
10	4	8	6	36	-4	16	2	4
3	7	1	-4	16	6	36	2	4

2	10	2	-8	64	8	64	0	0
4	2	3	2	4	-1	1	1	1
9	1	10	8	64	-9	81	-1	1
7	6	5	1	1	1	1	2	4
8	9	7	-1	1	2	4	1	1
				$\sum d_i^2(x_i, y_i) = 200$		$\sum d_i^2(y_i, z_i) = 214$		$\sum d_i^2(x_i, z_i) = 60$

Rank correlation between A and B is

$$\rho(x,y) = 1 - \frac{6\sum d_i^2\left(x_i, y_i\right)}{n^3 - n} = 1 - \frac{6\times 200}{10^3 - 10} = 1 - \frac{1200}{990} = -0.2121$$

Rank correlation between B and C is

$$\rho(y,z) = 1 - \frac{6\sum d_i^2\left(y_i, z_i\right)}{n^3 - n} = 1 - \frac{6\times 210}{10^3 - 10} = 1 - \frac{1260}{990} = -0.2727$$

Rank correlation between A and B is

$$\rho(x,z) = 1 - \frac{6\sum d_i^2\left(x_i, z_i\right)}{n^3 - n} = 1 - \frac{6\times 60}{10^3 - 10} = 1 - \frac{360}{990} = 0.6363$$

Since $\rho(x,z) = 0.6363$ is maximum, the pair of judges A and C have the nearest common approach.

Example: Calculate the rank correlation coefficient from the following data showing ranks of 10 students in two subjects

Maths	3	8	9	2	7	10	4	6	1	5
Physics	4	9	10	1	8	7	3	4	2	6

Solution: Solution: If $d_i = x_i - y_i$, then $d_i = -1, -1, -1, 1, -1, 3, 1, 2, -1, -1$

$$\sum d_i^2 = 1+1+1+1+1+9+1+4+1+1 = 21$$

Hence $\rho = 1 - \frac{6\sum d_i^2}{n^3 - n} = 1 - \frac{6 \times 21}{10^3 - 10} = 1 - \frac{126}{990} = 0.8727$.

The Spearman rank correlation for repeated ranks is given by

$$\rho = 1 - \frac{6\left\{\sum d^2 + \frac{m_1(m_1^2 - 1)}{12} + \frac{m_2(m_2^2 - 1)}{12} + \dots\right\}}{n^3 - n}$$

Where $m_1, m_2, \dots$ are the number of items whose ranks are common.

Example: Find rank correlation for the following data

x	56	42	72	36	63	47	55	49	38	42	68	60
y	147	125	160	118	149	128	150	145	115	140	152	155

Solution:

x	Rank of x (x_i)	y	Rank of y (y_i)	$d_i = x_i - y_i$	d_i^2
56	8	147	7	1	1
42	3.5	125	3	0.5	0.25
72	12	160	12	0	0
36	1	118	2	-1	1
63	10	149	8	2	4
47	5	128	4	1	1
55	7	150	9	-2	4
49	6	145	6	0	0

38	2	115	1	1	1
42	3.5	140	5	-1.5	2.25
68	11	152	10	1	1
60	9	155	11	-2	4
					$\sum d_i^2 = 19.5$

The Spearman rank correlation for repeated ranks is given by

$$\rho = 1 - \frac{6\left\{\sum d^2 + \frac{m_1(m_1^2 - 1)}{12}\right\}}{n^3 - n}$$

where $m_1 = 2$ is the number of items whose ranks are common

$$\rho = 1 - \frac{6\left\{19.5 + \frac{2(2^2 - 1)}{12}\right\}}{12^3 - 12} = 1 - \frac{120}{1716} = 0.93$$

Example: Find rank correlation for the following data showing rank of 10 students in two tests

Student	A	B	C	D	E	F	G	H	I	J
Test 1	70	68	67	55	60	60	75	63	60	72
Test 2	65	65	80	60	68	58	75	63	60	70

Solution:

Student	Test 1	Rank (x_i)	Test 2	Rank (y_i)	$d_i = x_i - y_i$	d_i^2
A	70	8	65	5.5	2.5	6.25
B	68	7	65	5.5	1.5	2.25
C	67	6	80	10	-4	16
D	55	1	60	2.5	-1.5	2.25

E	60	3	68	7	-4	16
F	60	3	58	1	2	4
G	75	10	75	9	1	1
H	63	5	63	4	1	1
I	60	3	60	2.5	0.5	0.25
J	72	9	70	8	1	1
						$\sum d_i^2 = 50$

The Spearman rank correlation for repeated ranks is given by

$$\rho = 1 - \frac{6\left\{\sum d^2 + \frac{m_1(m_1^2-1)}{12} + \frac{m_2(m_2^2-1)}{12} + \frac{m_3(m_3^2-1)}{12}\right\}}{n^3 - n}$$

where $m_1 = 3, m_2 = 2, m_3 = 2$ are the number of items whose ranks are common

$$\rho = 1 - \frac{6\left\{50 + \frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12} + \frac{2(2^2-1)}{12}\right\}}{10^3 - 10}$$

$$\rho = 1 - \frac{318}{990} = 0.6787$$