Introduction to Statistical Methods

ISM Team

BITS
Pilani

Pilani|Dubai|Goa|Hyderabad

innovate achieve lead

BITS Pilani

Webinar-4

Correlation Analysis

(Correlation, Rank Correlation, Regression and Time Series)

(13 th March 2025)

Agenda

- **Correlation,**
- **Rank Correlation**
- **Regression Analysis**
- **Time Series**

Correlation Analysis

- **Definition**
- **Types of Correlation**
- **Methods of studying Correlation**

Methods of studying Correlation

There are various methods to know whether the two variables are correlated or not. They are

- Scatter diagram
- Karl Pearson's coefficient of correlation

Scatter Diagram

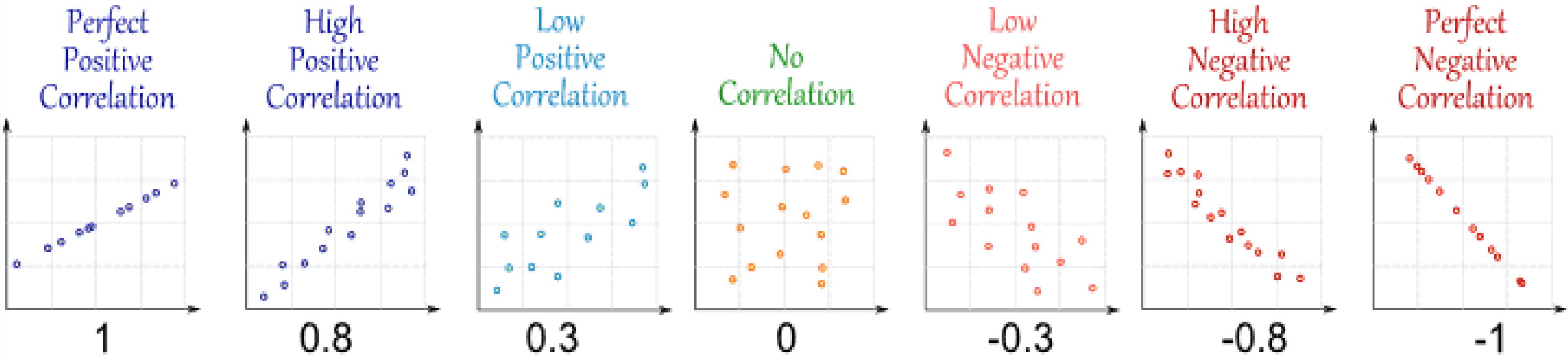
Procedure: given n pair of values (x1,y1) (x2,y2)…(xn,yn) of two variables X and Y.

- Take the independent variable on x axis and dependent variable on y axis. Then plot the n points on the graph.

- The diagram of dots obtained is called scatter diagram.
  - If the points reveal any upward or downward trend the variables are said to be correlated , otherwise un-correlated.
  - If the points are very closed to each other, a good amount of correlation exists.
  - Upward trend indicates +ve correlation and downward trend indicates –ve correlation.
- Limitation: By this method we cannot establish the exact degree of correlation between the variables.

Scattered diagram- Some specific cases

$$-1 \leq Correlation(x, y) \leq +1$$



Karl Pearson's coefficient of correlation

The extent relationship between the variables calculated with the help of statistical techniques known as correlation coefficient. It was developed by Karl Pearson. It is denoted by r(x,y) or r_{xy}

It always varies between +1 or -1

$-1 \leq r(x,y) \leq +1$

When r= +1 then it denotes perfect +ve correlation

r= 0 denotes no correlation

r= -1 perfect –ve correlation

Correlation coefficient between two random variables X and Y is as follows:

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

$$= \frac{E(X,Y) - E(X)E(Y)}{\sqrt{E(X^2) - [E(X)]^2} . \sqrt{E(\mathrm{Y}^2) - [E(\mathrm{Y})]^2}}$$

$$= \frac{\frac{1}{N}\sum(X_i - \bar{X})(\mathrm{Y}_i - \bar{Y})}{\sqrt{\frac{1}{N}\sum(X_i - \bar{X})^2} \sqrt{\frac{1}{N}\sum(\mathrm{Y}_i - \bar{Y})^2}}$$

$$= \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2} \sqrt{N\sum Y^2 - (\sum Y)^2}}$$

Properties of the coefficient of correlation

1. The coefficient of correlation lies between -1 and +1.
2. Correlation is independent of change of origin and scale.
3. The coefficient of correlation is the geometric mean of two regression coefficients

$$r_{xy} = \sqrt{b_{xy} b_{yx}}$$

1. The degree of relationship between two variables is symmetric

$$r_{xy} = r_{yx}$$

1.Calculate Karl Pearson's coefficient of correlation for the following data:

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

$N = 9$

Sol: Correlation coefficient

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

X	Y	$_X2$	$_Y2$	XY
9	15	81	225	135
8	16	64	256	128
7	14	49	196	98
6	13	36	169	78
5	11	25	121	55
4	12	16	144	48
3	10	9	100	30
2	8	4	64	16
1	9	1	81	9
$\sum X = 45$	$\sum Y = 108$	$\sum X^2 = 285$	$\sum Y^2 = 1356$	$\sum XY = 597$

N=9

Therefore, N=9, $\sum XY = 597, \sum X = 45, \sum X^2 = 285, \sum Y = 108, \sum Y^2 = 1356$

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

$$= \frac{9*597 - 45*108}{\sqrt{9*285 - (45)^2}\sqrt{9*1356 - (108)^2}}$$

$$= \frac{513}{\sqrt{540}\sqrt{540}} = \frac{513}{540}$$

$$= +0.95$$

There is a high degree of +ve correlation between variables X and Y.
i.e if x increases y also increases

2. Calculate Karl Pearson's 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

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

$$\bar{x} = \frac{\sum x}{n} = \frac{544}{8} = 68 \qquad \bar{y} = \frac{\sum y}{n} = \frac{552}{8} = 69$$

$$r_{xy} = \frac{n\sum uv - (\sum u)(\sum v)}{\sqrt{n\sum u^2 - (\sum u)^2}\sqrt{n\sum v^2 - (\sum v)^2}}$$

$$= \frac{8(24) - 0}{\sqrt{8(36) - 0}\sqrt{8(44) - 0}}$$

$$= 0.603$$

Spearman's rank correlation coefficient

- Karl Pearson's method is based on the assumption that the population being studied is normally distributed.
- When the population is not normal, then this method is not useful to find coefficient of correlation.
- Spearman in 1904 has given ranks to the observations according to size and he based the calculation on ranks rather than original observations to avoid the assumptions about population defined by Karl Pearson. This method is called as rank correlation. It is denoted as $R(x,y)$ or $\rho(x,y)$.
- To define the qualitative characteristics (ex: honesty, beauty etc..) in terms of quantitative measure this method is very useful.

Cont..

- Rank correlation coefficient

$$R = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Where di is the difference between two ranks.

If the ranks are repeated(equal ranks)then we use

$$R = \rho = \frac{1 - 6[\sum d^2{}_i + \frac{m_1(m_1^2-1)}{12} + \frac{m_2(m_2^2-1)}{12} + ...]}{n(n^2-1)}$$

Where mi stands for number of items whose ranks are common. If there are more than one group of items with common rank , this value is added as many times the number of such group

1. Note: $\Sigma d_i = 0$, if the data is in the form of ranks this method is useful.
2. If n>30 then it cannot give reliable value.
3. For calculating correlation for quantitative data this can be applicable, but slightly effective than Pearson correlation coefficient.

Problem 1) Calculate the rank correlation coefficient for the following data

X	60	34	40	54	57	35	36
Y	70	72	65	68	64	34	42

Solution: If the ranks are not given we have to assign the ranks for both the variables, assigning rank 1 to the lowest value or highest value. We have to follow same method for both variables.

X	Y	R_X	R_Y	$d_i = R_X - R_Y$	d_i^2
60	70	1	2	-1	1
34	72	7	1	6	36
40	65	4	4	0	0
54	68	3	3	0	0
57	64	2	5	-3	9
35	37	6	7	-1	1
36	42	5	6	-1	1
					$\sum d_i^2 = 48$

Here ranks are not repeated, $R = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$

n = number of observations = 7, $\sum d_i^2 = 48$

$$R = 1 - \frac{6*48}{7(7^2-1)} = \frac{1}{7} = 0.14$$

Therefore, rank correlation coefficients between X and Y is 0.14.

Problem 2) Calculate the rank coefficient of correlation for the following data

X	80	78	75	75	68	67	60	54
Y	12	13	14	14	14	16	15	17

Solution: For the given data we have to assign ranks. But in the above series there is more than one item with the same value.

In such cases same ranks are to be given to the repeated items. This same rank is the average of the ranks which these items would have assumed if they were different from one other.

X	Y	R_X	R_Y	$d_i=R_X - R_Y$	d^2_i
80	12	1	8	-7	49
78	13	2	7	-5	25
75	14	3.5	5	-1.5	2.25
75	14	3.5	5	-1.5	2.25
68	14	5	5	0	0
67	16	6	2	4	16
60	15	7	3	4	16
59	17	8	1	7	49
					$\sum d_i^2 = 159.5$

In x data
75 repeated 2
$m_1 = 2$
In y data 14 repeated
3 times
$m_2 = 3$

$$R = \rho = \frac{1 - 6[\sum d^2_i + \frac{m_1(m_1^2 - 1)}{12} + \frac{m_2(m_2^2 - 1)}{12} + ...]}{n(n^2 - 1)}$$

$$= 1 - \frac{6(159.5 + \frac{2(2^2 - 1)}{12} + \frac{3(3^2 - 1)}{12})}{8(8^2 - 1)}$$

$$= 1 - \frac{972}{504} = -0.929$$

$m_1 = 2$

$m_2 = 3$

[75 is repeated 2 times m1=2 and 14 is repeated 3 times m2=3

Regression analysis

- Let x and y are two variables for regression analysis then there will be two regression lines.
- Regression line of x on y =

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

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

- Regression line of y on x=

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

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

- The intersection of two lines x on y and y on x gives the mean values of x series and y series. i.e., and $\bar{x}$ $\bar{y}$

Properties of Regression coefficients

- Correlation coefficient is the geometric mean of the regression coefficients. $r = \sqrt{b_{xy}.b_{yx}}$

- If one regression coefficient is > unity then the other must be < unity.
- The arithmetic mean of the two regression coefficient is > the correlation coefficient.

Problems

- The folowing table gives the normal weight of kids during the first 8 years of life;

Age	0	1	2	3	5	6	7	8
Weight	5	7	8	10	15	17	20	22

i. Obtain two lines of regression
ii. Estimate the weight of kid at the age of 4 years

Solution : we know that regression line y on x is $y-\bar{y}=b_{yx}(x-\bar{x})$ Where $\bar{x},\bar{y}$ are means

$$b_{yx}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$$ is regression coefficient.

Line of regression x on y is $x-\bar{x}=b_{xy}(y-\bar{y})$

$$b_{xy}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(y_i-\bar{y})^2}$$

N=no of observations. $=8$

$$\bar{x}=\frac{\sum x}{N}=\frac{32}{8}=4$$

$$\bar{y}=\frac{\sum y}{N}=\frac{104}{8}=13$$

Age(x)	Weight(y)	$(x-\bar{x})=x-4$	$(y-\bar{y})=y-13$	$(x-\bar{x})^2=(x-4)^2$	$(y-\bar{y})^2=(y-13)^2$	$(x-\bar{x})(y-\bar{y})=(x-4)(y-13)$
0	5	-4	-8	16	64	32
1	7	-3	-6	9	36	18
2	8	-2	-5	4	25	10
3	10	-1	-3	1	9	3
5	15	1	2	1	4	2
6	17	2	4	4	16	8
7	20	3	7	9	49	21
8	22	4	9	16	81	36
Total=32	104	0	0	60	284	130

$\bar{x}=\frac{\Sigma x}{8}=\frac{32}{8}=4$ $\quad \Sigma x$ $\quad \Sigma y$ $\quad \bar{y}=\Sigma y/n=\frac{104}{8}=0$ $\quad 0$ $\quad 60$ $\quad 284$ $\quad 13$

From the table $\sum x=32, \sum y=104, \sum(x-\bar{x})=0, \sum(y-\bar{y})=0, \sum(x-\bar{x})^2=60$

$\sum(y-\bar{y})^2=284, \sum(x-\bar{x})(y-\bar{y})=130$

$$b_{yx} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{130}{60} = 2.167$$

$$b_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(y_i - \bar{y})^2} = \frac{130}{284} = 0.458$$

Line of regression y on x

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

$$y - 13 = 2.167(x - 4)$$

$$y - 13 = 2.167x - 8.668$$

$$y = 2.167x + 5.668$$

Sub $x = 4$

$\therefore y = 2.167(4) + 5.668$

Line of regression x on y

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

$$x - 4 = 0.458(y - 13)$$

$$x - 4 = 0.458y - 5.954$$

$$x = 0.458y - 1.954$$

Weight of a kid at the age of 4 is estimating y when x=4, we use y on x regression line:

$$y = 2.167x + 5.668$$

$$y = 2.167(4) + 5.668$$

$$y = 14.336$$

The weight of a kid at the age of 4 is 14.336.

Problem: The following are the regression equations:

$$\left.\begin{aligned} 8x - 10y + 66 = 0 \\ 40x - 18y = 214 \end{aligned}\right\}$$

Find the regression line x on y and y on x and also find

i. $\bar{x}$ And $\bar{y}$
ii. The regression coefficients
iii. Correlation Coefficients

Solution:

$$8x - 10y + 66 = 0 \quad \text{.....(1)}$$

$$40x - 18y = 214 \quad \text{.....(2)}$$

i. By solving two regression lines we get $\overline{x}$ and $\overline{y}$

$\overline{x}$ = 13 and $\overline{y}$ = 17

ii. Rewrite the equations of regression to get the form x on y and y on x

From eq-(1) $8x - 10y + 66 = 0$

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

$$y = a + bx \Rightarrow y = \frac{66}{10} + \frac{8}{10}x \qquad y = a + bx$$

$$b_{yx} = \frac{8}{10} = \frac{4}{5}$$

Equation (2) $40x - 18y = 214$

$$\therefore x = \frac{18}{40}y + \frac{214}{40}$$

$$x = a + by \Rightarrow x = \frac{214}{40} + \frac{18}{40}y$$

$$\therefore b_{xy} = \frac{18}{40} = \frac{9}{20}$$

Regression coefficient x on y $= \frac{9}{20} = b_{xy}$

Regression coefficient y on x $= \frac{4}{5} = b_{yx}$

iii)Correlation coefficient $r = \pm\sqrt{b_{xy}.b_{yx}} = \pm\sqrt{\frac{9}{20}.\frac{4}{5}} = \pm\sqrt{\frac{9}{25}} = \pm 0.6$

Since both regression coefficients are positive ,r must be positive. Therefore r=0.6

Problem: **Calculate the coefficient of correlation from the following data:**

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

Also obtain the equation for linear regression and obtain an estimate of y when x=6.2

$\sum X$ =45 ; $\sum Y$ =108 ; $\sum X^2$ = 285; $\sum Y^2$=1356; $\sum XY$ =597; N=9

Var(X) = 6.67 ; $\sigma_X = \sqrt{6.67} = 2.58$

Var(Y) = 6.67 ; $\sigma_Y = \sqrt{6.67} = 2.58$

$\text{Cov}(X,Y) = \frac{\sum XY}{N} - \bar{X}\,\bar{Y} = 6.33$

$r = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = 0.95$

Regression Equation of X on Y is $X - \bar{X} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y})$: X = 0.95Y – 6.4

Regression Equation of Y on X is $Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$: Y = 0.95X +7.25

Estimate of Y when x =6.2 = 0.95 * 6.2 +7.25 = 13.14

Problem: **The two regression lines having there mean and standard deviation 31.6 , 38 and 3.72 , 6.31 and $\rho = -0.36$. Find the 2 Regression lines.**

$\bar{X}$ = 31.6 ; $\bar{Y}$ = 38

σ_X = 3.72 ; σ_Y = 6.31 ; r= -0.36

Regression Equation of Y on X is $Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \overline{X})$

Y =- 0.61 X + 57.28

Least square Method

Algebraically method-:

1. Least Square Method-:

The regression equation of X on Y is :

$$X = a + bY$$

Where,

X=Dependent variable

Y=Independent variable

The regression equation of Y on X is:

$$Y = a + bX$$

Where,

Y=Dependent variable

X=Independent variable

And the values of a and b in the above equations are found by the method of least of Squares-reference . The values of a and b are found with the help of normal equations given below:

(I)

$$\sum X = na + b\sum Y$$
$$\sum XY = a\sum Y + b\sum Y^2$$

(II)

$$\sum Y = na + b\sum X$$
$$\sum XY = a\sum X + b\sum X^2$$

Cont…

Example1-: From the following data obtain the two regression equations using the method of Least Squares.

X	3	2	7	4	8
Y	6	1	8	5	9

Solution-:

X	Y	XY	X^2	Y^2
3	6	18	9	36
2	1	2	4	1
7	8	56	49	64
4	5	20	16	25
8	9	72	64	81
$\sum X = 24$	$\sum Y = 29$	$\sum XY = 168$	$\sum X^2 = 142$	$\sum Y^2 = 207$

Cont…

$Y = a + bx$

$x = a + by$

$$\sum Y = na + b\sum X$$
$$\sum XY = a\sum X + b\sum X^2$$

} Normal eqns

Substitution the values from the table we get

29=5a+24b……(i)
168=24a+142b
84=12a+71b……..(ii)

Multiplying equation (i) by 12 and (ii) by 5

348=60a+288b……(iii)
420=60a+355b……(iv)

By solving equation(iii)and (iv) we get

a=0.66 and b=1.07

$y = 0.66 + 1.07x$

Cont…

By putting the value of a and b in the **Regression equation Y on X** we get

Y=0.66+1.07X

Now to find the regression equation of X on Y ,
The two normal equation are

$$\sum X = na + b\sum Y$$
$$\sum XY = a\sum Y + b\sum Y^2$$

$x = a + by$

Substituting the values in the equations we get

24=5a+29b……(i)
168=29a+207b……..(ii)

Multiplying equation (i)by 29 and in (ii) by 5 we get

a=0.49 and b=0.74

Substituting the values of a and b in the **Regression equation X and Y**

X=0.49+0.74Y

Non-linear least square approximation:

Parabola: let the equation of parabola to be fit be $y = a + bx + cx^2$ .....(1)

The normal equations are

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

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

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

Normal eqns

Solving these equations for a,b,c and substituting in (1) we get required parabola of best fit.

Problem

1)Fit the second degree parabola to the following data

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

sol: let the equation of the parabola be

$$y = a + bx + cx^2 \ldots\ldots\ldots\ldots (1)$$

The normal equations are

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

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

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

x	y	xy	x^2	x^2y	x^3	x^4
0	1	0	0	0	0	0
1	5	5	1	5	1	1
2	10	20	4	40	8	16
3	22	66	9	198	27	81
4	38	152	16	608	64	256
$\sum x = 10$	$\sum y = 76$	$\sum xy = 152$	$\sum x^2 = 30$	$\sum x^2 y = 851$	$\sum x^3 = 100$	$\sum x^4 = 354$

Sub these values in the normal equations

$$76 = 5a + 10b + 30c$$
$$243 = 10a + 30b + 100c$$
$$851 = 30a + 100b + 354c$$

Solving above a=1.42, b=0.26, c=2.221

$y = a + bx + cx^2$

$y = 1.42 + 0.26x + 2.221$

Sub a, b, c in (1)

$y = 1.42 + 0.26x + 2.21x^2$ is the required parabola of the best fit.

Power curve: The power curve is $y = ax^b \ldots\ldots(1)$

Taking log on both sides

$$\log_{10} y = \log_{10} a + b \log_{10} x$$

$$Y = A + bX \ldots\ldots\ldots\ldots(2)$$

$$where, Y = \log_{10} y, A = \log_{10} a, X = \log_{10} x$$

(2) is a linear equation in X and Y
The normal equations are

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

$$\sum XY = A\sum X + b\sum X^2$$

Solving for A and b and substitute in (2).

Problem: Fit a curve $y = ax^b$ to the following data

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

Sol: $y = ax^b \ldots\ldots\ldots\ldots (1)$

Taking log on both sides

$$\log_{10} y = \log_{10} a + b \log_{10} x$$

$$Y = A + bX \ldots\ldots\ldots\ldots (2)$$

$$where, Y = \log_{10} y, A = \log_{10} a, X = \log_{10} x$$

This is a linear eq in X and Y

The normal eq s are

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

$$\sum XY = A \sum X + b \sum X^2 \qquad \ldots\ldots\ldots\ldots (3)$$

x	X=log x	y	Y=log y	XY	X^2
1	0	2.98	0.4742	0	0
2	0.3010	4.26	0.6294	0.1894	0.0906
3	0.4771	5.21	0.7168	0.3420	0.2276
4	0.6021	6.10	0.7853	0.4728	0.3625
5	0.6990	6.80	0.8325	0.5819	0.4886
6	0.7782	7.50	0.8751	0.6810	0.6056
TOTAL	$\sum X = 2.8574$		$\sum Y = 4.3133$	$\sum XY = 2.2671$	$\sum X^2 = 1.7749$

Substituting the above values in (3)

$4.3133 = 6A + 2.8574\ b$

$2.2671 = 2.8574\ A + 1.7749\ b$

Solving the equations for A and b

A = 0.4739 And b = 0.5143

$a = (10)^A = (10)^{0.4739} = 2.978$

Substituting a and b in (1)

$$y = 2.978x^{0.5143}$$

Exponential curve: $1) y = ae^{bx}, 2) y = ab^{x}$

1)The exponential curve $y = ae^{bx} \ldots\ldots\ldots (1)$

Taking log on both sides

$$\log_{10} y = \log_{10} a + bx \log_{10} e$$

$$Y = A + Bx \ldots\ldots\ldots\ldots (2)$$

$$where, Y = \log_{10} y, A = \log_{10} a, \mathrm{B} = b \log_{10} e$$

Equation (2) is linear equation in x and Y
So the normal equations are given by

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

$$\sum xY = A \sum x + B \sum x^{2}$$

Problem: Fit a curve of the form $y = ae^{bx}$ to the following data:

x	0	5	8	12	20
y	3.0	1.5	1.0	0.55	0.18

Sol: Let $y = ae^{bx} \ldots\ldots\ldots\ldots(1)$

Taking log on both sides

$$\log_{10} y = \log_{10} a + bx \log_{10} e$$

$$Y = A + Bx \ldots\ldots\ldots\ldots(2)$$

$$where, Y = \log_{10} y, A = \log_{10} a, \mathrm{B} = b \log_{10} e$$

Equation (2) is linear equation in x and Y
So the normal equations are given by

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

$$\sum xY = A\sum x + B\sum x^2 \ldots\ldots\ldots\ldots(3)$$

x	y	$Y = \log_{10} y$	x^2	xY
0	3.0	0.4771	0	0
5	1.5	0.1761	25	0.8805
8	1.0	0	64	0
12	0.55	-0.2596	144	-3.1152
20	0.18	-0.7447	400	-14.894
$\sum x = 45$		$\sum Y = -0.3511$	$\sum x^2 = 633$	$\sum xY = -17.1287$

Sub these values in (3) we get

$$5A + 45B = -0.3511$$

$$45A + 633B = -17.1287$$

Solving above for A and B
A=0.4815 and B=-0.0613

$$A = \log_{10} a = 0.4815$$

$$\therefore a = (10)^{A} = (10)^{0.4815} = 3.0304$$

$$B = b \log_{10} e = -0.0613$$

$$\therefore b = \frac{-0.0613}{0.4343} = -0.1411$$

Sub a and b values in (1) $y = 3.0304e^{-0.1411x}$

Time Series Analysis

- Importance of Time series
- Definition
- Components
- Secular trend in time series

Importance of Time series

- A very powerful tool for business forecasting
- Basis for understanding past behavior
- Can forecast future activities/planning for future operations
- Facilitates comparison
- Estimation of trade cycle

Examples:

Retail sales

Electronic power consumption

Stock price

Demand

Cont..

- Time series is arrangement of statistical data in accordance with time of occurrence.

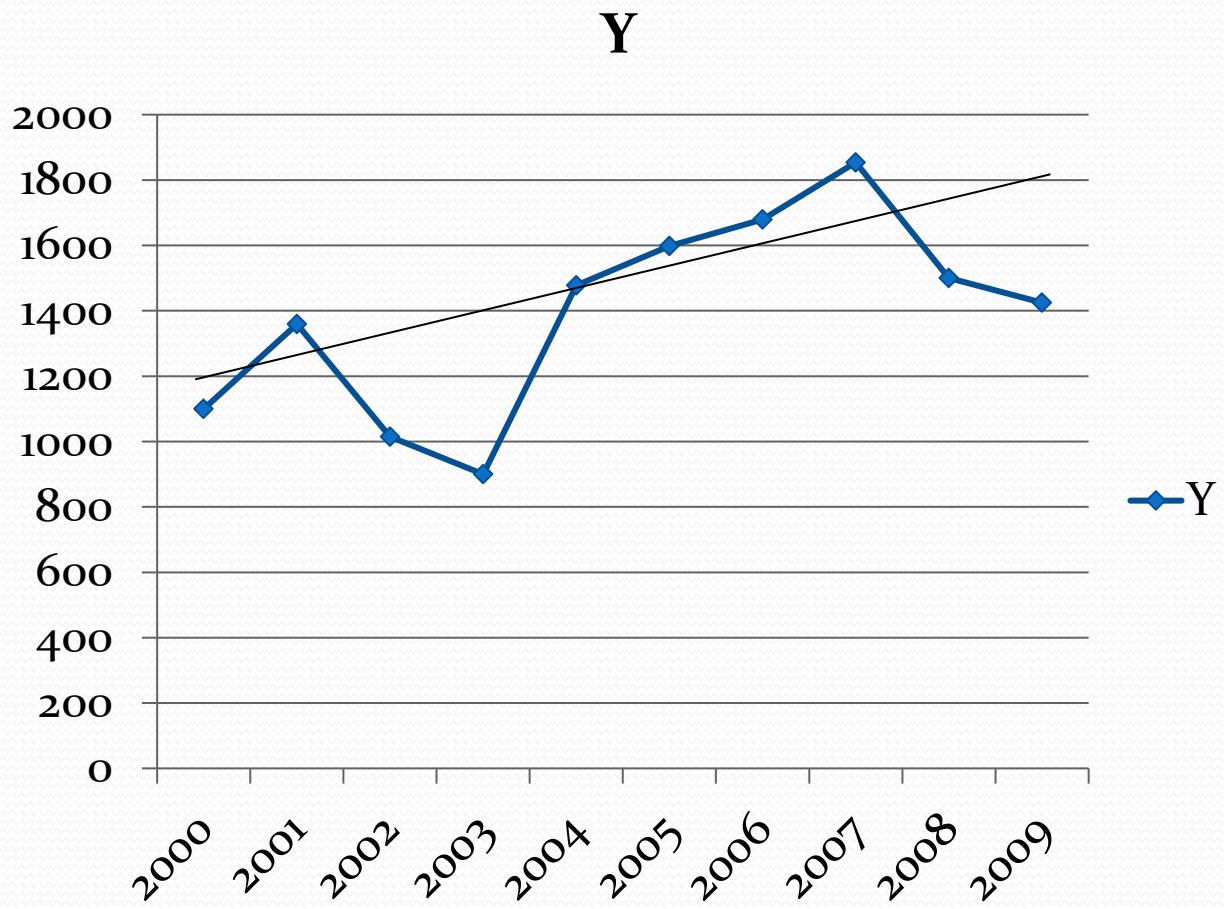
Time	Rice production (thousand tones)
2011	12
2012	15
2013	18
2014	16
2015	22
2016	25
2017	23

MESUREMENT OF Secular Trend Methods

1. Graphical Method (Free hand method)
2. Moving Averages Method
3. Least square Method

Graphical Method

Year	Y
2000	1100
2001	1359
2002	1015
2003	900
2004	1478
2005	1598
2006	1679
2007	1854
2008	1500
2009	1425

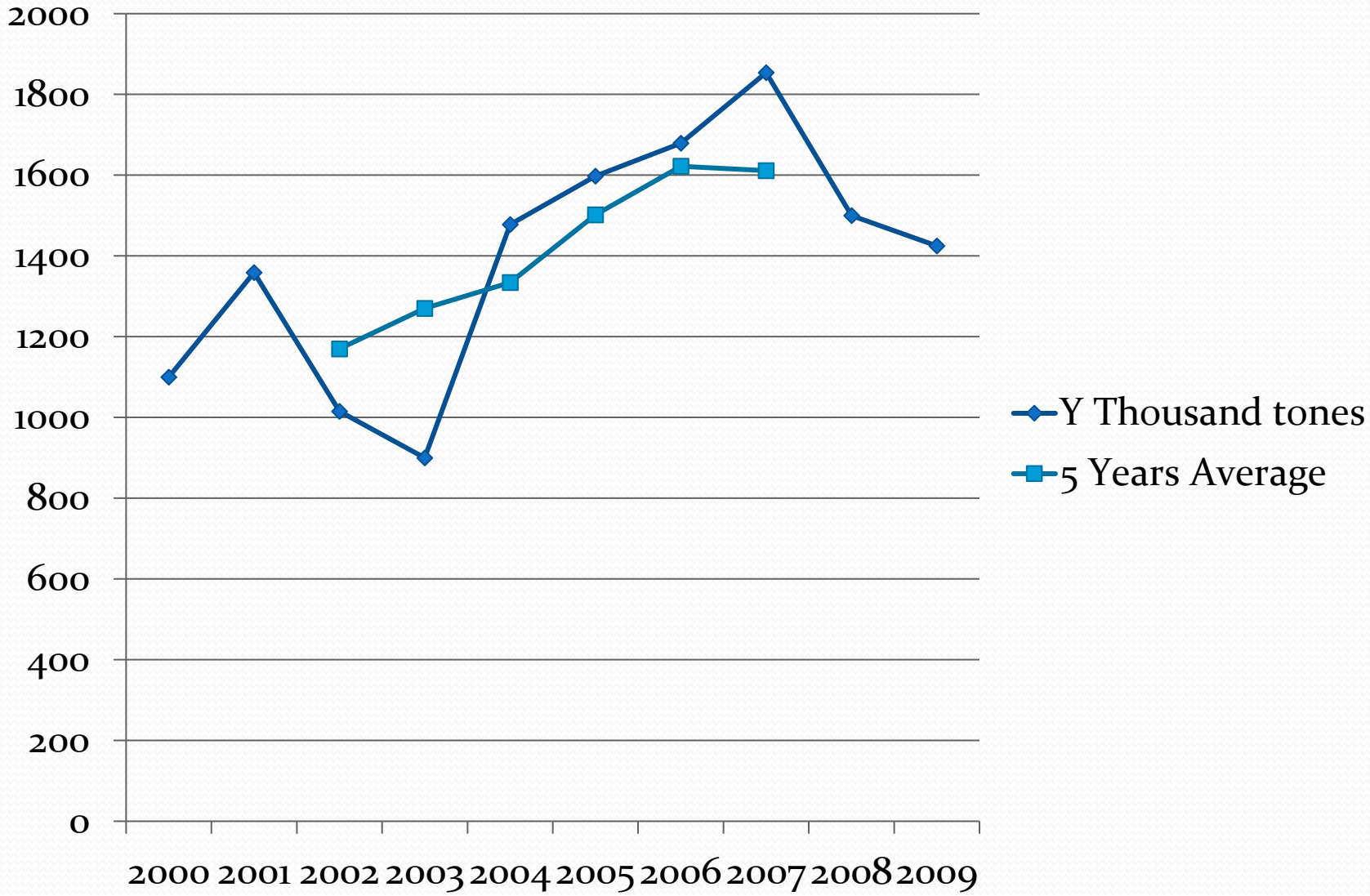


Moving averages method:

Find out the trend for 5 years moving averages to the following data

Year	Y
2000	1100
2001	1359
2002	1015
2003	900
2004	1478
2005	1598
2006	1679
2007	1854
2008	1500
2009	1425

Year(time)	Y in Thousand tones	5 year moving total	5 years average
2000	1100		
2001	1359		
2002	1015	5852	5852/5=1170
2003	900	6350	6350/5=1270
2004	1478	6670	6670/5=1334
2005	1598	7509	7509/5=1502
2006	1679	8109	8109/5= 1622
2007	1854	8056	8056/5=1611
2008	1500		
2009	1425		


2000
1800
1600
1400
1200
1000
800
600
400
200
0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
Y Thousand tones
5 Years Average

Measurement of Trend (T):

The trend component (T) of time series is determined using graphical method, semi-average method, Moving-average method and method of least squares.

Example : Calculate the moving average for the following data for 3 year cycle.

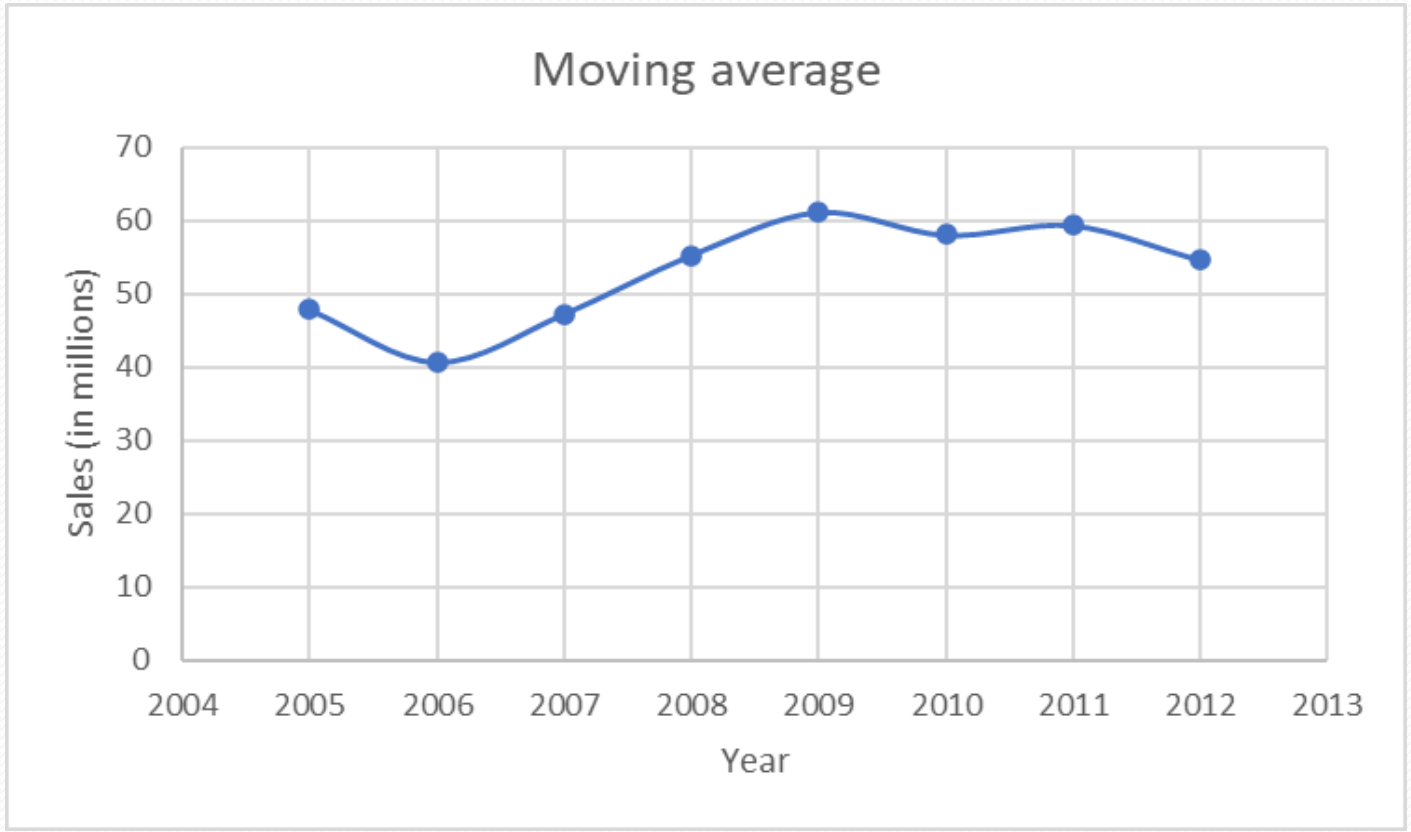
Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Sales	55.76	45.56	42.11	34.33	65.31	66.20	51.95	55.97	70.12	37.8

Solution: The data of sales of commodities in millions is given and the moving average for 3 year cycle is to be obtained.

Procedure to calculate the Simple Moving average :

- Construct a table of 10 rows and 5 columns.
- Write the Year data in column 1 and sales data in column 2.
- Name the third column as column of differences, fourth column as three year moving totals and fifth column as three year moving averages.
- Omit the first row in third column and second row element of third column is difference in 4^{th} row element and first row element of second column. Like wise all elements are obtained.
- The fifth column is obtained by taking the taking average by dividing three year moving total with time interval.

Year	Sales	Column of difference	Three year moving totals	Three year moving Average
2005	55.76			
2006	45.56	-21.43	143.43	47.81
2007	42.11	19.75	122	40.66667
2008	34.33	24.09	141.75	47.25
2009	65.31	17.62	165.84	55.28
2010	66.2	-9.34	183.46	61.15333
2011	51.95	3.92	174.12	58.04
2012	55.97	-14.15	178.04	59.34667
2013	70.12			
2014	37.8			

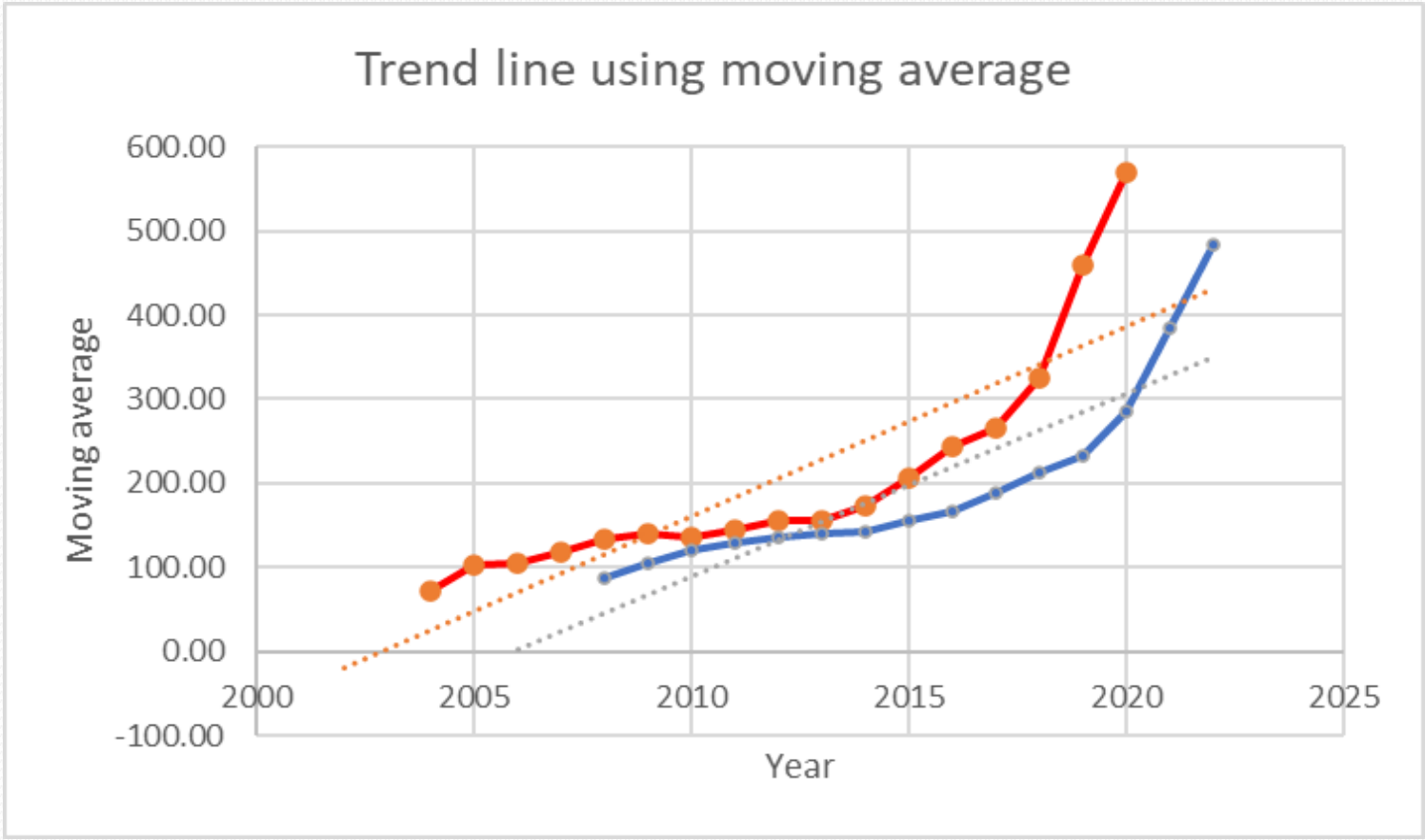

Moving average
70
60
50
40
30
20
10
0
Sales (in millions)
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
Year

Example : Determine the trends in stock value for over 20years with the following data:

Year	2022	2021	2020	2019	2018	2017	2016	2015	2014	2013
Stock value	870.55	893.6	477.25	293.48	311.72	321.57	220.7	176.59	191.09	120.96
Year	2012	2011	2010	2009	2008	2007	2006	2005	2004	2003
Stock value	155.3	136.99	173.29	141.77	72.42	181.54	94.84	103.25	72.71	67.36

Solution : The moving average method for 5 year and 7 year duration is implemented and estimated as given below.

Year	Stock value	5 year moving total	5 year moving average	7 year moving Total	7 Year moving average
2022	870.55				
2021	893.6				
2020	477.25	2846.60	569.32		
2019	293.48	2297.62	459.52	3388.87	484.12
2018	311.72	1624.72	324.94	2694.91	384.99
2017	321.57	1324.06	264.81	1992.40	284.63
2016	220.7	1221.67	244.33	1636.11	233.73
2015	176.59	1030.91	206.18	1497.93	213.99
2014	191.09	864.64	172.93	1323.20	189.03
2013	120.96	780.93	156.19	1174.92	167.85
2012	155.3	777.63	155.53	1095.99	156.57
2011	136.99	728.31	145.66	991.82	141.69
2010	173.29	679.77	135.95	982.27	140.32
2009	141.77	706.01	141.20	956.15	136.59
2008	72.42	663.86	132.77	904.10	129.16
2007	181.54	593.82	118.76	839.82	119.97
2006	94.84	524.76	104.95	733.89	104.84
2005	103.25	519.70	103.94	617.71	88.24
2004	72.71	363.75	72.75		
2003	67.36				
2002	25.59				


Trend line using moving average
600.00
500.00
400.00
300.00
200.00
100.00
0.00
-100.00
Moving average
2000
2005
2010
2015
2020
2025
Year

Example 2: Fit a trend line to the following data by the lease square method:

Year	2020.00	2018	2016	2014	2012	2010	2008
Stock value	477.25	311.72	220.7	191.09	155.3	173.29	72.42

Solution : As the data set has 7 values, the middle value of 2014 is taken and Variable X is taken as t-2014 value. The stock value of the company is taken as Y and XY values & X^2 values are computed.

Year+	Stock value (Y)	X=2014-t	XY	X^2
2020	477.25	-6.00	-2863.50	36.00
2018	311.72	-4.00	-1246.88	16.00
2016	220.7	-2.00	-441.40	4.00
2014	191.09	$\sum x$ 0.00	$\sum xy$ 0.00	$\sum x^2$ 0.00
2012	155.3	2.00	310.60	4.00
2010	173.29	4.00	693.16	16.00
2008	72.42	6.00	434.52	36.00
	1601.77	=0.00	-3113.50	112.00

Let the equation of trend line be Y=a+bX. We have to find the values of a and b such that the trend line satisfies the given data set. The normal equations for estimating ‘a’ and ‘b’ values are :

$\sum y$ = n*a +b* $\sum x$

=> 1601.77 = 7*a + b*0

=> a =1601.77/7 = 228.824

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

3113.50 = a*0 +b*112

=> b= 3113.50/112

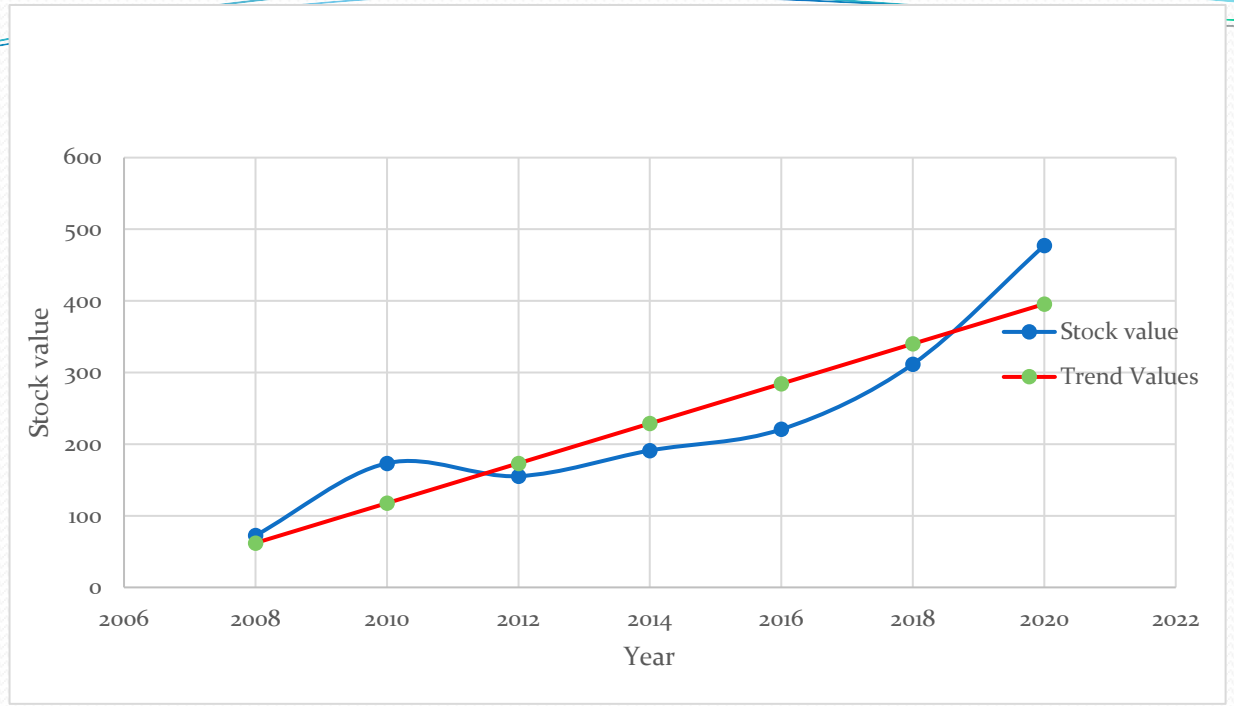
=> b= -27.79

Hence the trend line is Y=a +b*X

Y=228.824 – 27.79 * X

Hence trend values are computed and tabulated as shown below.

Year	Stock value	X=2014-t	XY	X^2	Trend Values
2020	477.25	-6.00	-2863.50	36.00	395.62
2018	311.72	-4.00	-1246.88	16.00	340.02
2016	220.7	-2.00	-441.40	4.00	284.42
2014	191.09	0.00	0.00	0.00	228.82
2012	155.3	2.00	310.60	4.00	173.23
2010	173.29	4.00	693.16	16.00	117.63
2008	72.42	6.00	434.52	36.00	62.03



With the obtained trend line , we can estimate the values in future also. If we want to know the future value of the stock in the year 2025, let us use the obtained equation Y = 228.824 -27.79 *X.

The central value is taken from 2014 so, X= 2014-2025 = -11, hence Y= 228.824 -27.79*(-11) = 534.61.

We can recommend this stock to be purchased by the consumer assuming if same linear trend continues then the net asset value of the stock will be about 534.61

Practice Problems:

1. Calculate the coefficient of Correlation:

X	105	109	102	101	100	99	98	96	93	92
Y	101	103	100	98	95	96	104	92	97	94

2. Obtain the rank correlation coefficient for the following data:

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

3. Find the equation of line of regression of y on x and x on y for the data:

x	5	2	1	4	3
y	5	8	4	2	10

Practice Problems:

4. Find the regression equations for the following data:

Age of husband: x	36	23	27	28	28	29	30	31	33	35
Age of wife: y	29	18	20	22	27	21	29	27	29	28

5. Fit a second degree parabola to the following data:

X	0	1	2	3	4
Y	1	5	10	22	38

Thank you