Department of Distance Education

Punjabi University, Patiala

Class : B.A. (Economics) Part-III **Semester : 6**
Paper : II (Quantitative Methods) **Unit : 2**
Medium : English

Lesson No.

2.1 : CORRELATION ANALYSIS

(KARL PEARSON'S AND SPEARMAN'S FORMULA)

2.2 : SIMPLE REGRESSION ANALYSIS

2.3 : INDEX NUMBERS

2.4 : TIME SERIES ANALYSIS

Department website : www.pbidde.org

LESSON NO. 2.1 AUTHOR : DR. S.L. AGGARWAL

CORRELATION ANALYSIS

2.1.1 Introduction

The measures of the central tendency and skewness throw light on the construction of a series. These measures are also used for comparison between two series of the same variab le. But this is not enough, sometimes. The term correlation indicates the relationship between two such variables in which with changes in the values of one variable, the values of the other variable also change.

2.1.2 Objectives

After completion of this lesson, you should be able to :

- Understand the meaning of Correlation.
- Know different types of correlation
- Compute Karl Pearson's coefficient fo correlation
- Compute the rank correlation coefficient

2.1.3 Meaning of Correlation

In this lesson meaning, types of correlationa nd differnt ways leading to the calculation of correlation will be examined. Correlation is a statistical technique which shows the relationship between two or more variables.

According to L.R. Conner, "If two or more quantities vary in sympathy so that movements in the one tend to be accompanied by corresponding movements in the others than, they are said to be correlated."

According to Ya Lun Chou, "Correlation analysis attempts to determine the degree of relationship between variables."

We can see the relationship between various pairs of variables like, age and blindness, income and consumption, height and weight, etc. In correlation, thus, we deal with bivariate distributions.

2.1.4 Types of Correlation :

Correlation is described or classified in several different ways. It can be classified into :

2.1.4.1 Positive or Negative Correlation :

On the basis of the direction of the change in the two variables, correlation can be-ve or +ve. If the change in both the variables is in the same direction i.e., if both increase simultaneously or decrease simultaneously, the correlation is said to be positive.

If the change in both the variables is in the opposite direction i.e. if one increases, other decreases, then correlation is said to negative. The following examples would illustrate the difference between positive and negative correlation.

1. Positive Correlation :

Price (Rs.)	:	10	11	12	13
Supply	:	100	130	140	160

2. Negative Correlation :

Price (Rs.)	:	10	11	12	13

Demand	:	100	90	85	82

2.1.4.2 Simple and Multiple Correlation :

If we study, correlation between two variables, it is called simple correlation. In case, we study relation in more than two variables, it is called multiple correlation.

2.1.4.3 Linear and Non-Linear Correlation :

On the basis of the ratio of change in the related variables, the correlation can be linear or non-linear.

If the amount of change in a variable is at a constant ratio to the change in the other variable, the correlation is said to be linear. This relationship is represented by the equation $y = a + bx$ when plotted on a graph paper, straight line is formed. This type of correlation is found only in physical sciences.

Illustration :

Price (Rs.)	:	10	11	12	13
Quantity Supplied	:	80	120	160	200

If the amount of change in one variable is not at constant ratio to the change in the other variable, the correlation is said to be non-linear. This type of correlation is generally found in social sciences.

Illustration :

Income (Rs.)	:	100	150	200	250
Consumption (Rs.)	:	70	100	120	130

Correlation is non-linear for every increase in income by Rs. 50, the consumption firstly increases by Rs.30, then by Rs. 20 and then by Rs. 10.

2.1.5 Properties of Correlation :

(i) Degree of correlation is indicated by coefficient of correlation. According to Prof. Karl Pearson, the coefficient of correlation varies between two limits i.e. ± I i.e. the maximum value of coefficient of correlation can be +1 and the minimum value of coefficient of correlation can be - 1. It means correlation lies between -1 and +1 i.e. $-1 \leq r \leq +1$.

(ii) Correlation is not dependent on origin and change of scale. That means that there is no difference between step-deviation and simple deviation methods.

2.1.6 Methods of measuring Correlation :

Following methods can be used to measure the correlation between two variables :

1. Scatter diagram method.
2. Graphic method
3. Karl Pearson's method.
4. Concurrent deviation method.

2.1.6.1 Karl Pearson's Method :

The scatter diagram and graphic methods are not free from drawbacks. These methods provide us only the direction of relationship and not its extent. Karl Pearson gave a formula in 1896 to calculate coefficient of correlation between two variables. It is an algebraic method. It tells us both the direction and amount of correlation between two variables. It is denoted by symbol 'r'.

The formula for Pearson's Correlation coefficient is.

Direct Method :

$$r = \frac{\sum xy}{N\sigma x\sigma y}$$

where r = Coefficient of Correlation.

$$x = \left(x - \bar{x}\right), y = \left(y - \bar{y}\right)$$

σx = S.D. of x-series = $\sqrt{\frac{\sum x^2}{N}}$

σy = S.D. of y-series = $\sqrt{\frac{\sum y^2}{N}}$

N = Number of pairs of observations.

Covariance of x and y = $\frac{\sum xy}{N}$

$$\text{or } r = \frac{\Sigma xy}{N\sqrt{\frac{\Sigma x^2}{N} \cdot \frac{\Sigma y^2}{N}}}$$

Thus $r = \frac{\sum xy}{N.\sigma x\sigma y}$

$$\text{or } r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

Example 1 :

Calculate Karl Person's Coefficient of correlation from the following data

Income (Rs.) : 10 12 18 24 23 27

Consumption (Rs.): 13 18 12 25 30 10

Solution :

Calculation of coefficient of 'r'

Income			**Consumption**			
(X)	$x=X-\overline{X}$	x^2	(Y)	$y=Y-\overline{Y}$	y^2	xy
	x -19			Y-18		
10	-9	81	13	-5	25	45
12	-7	49	18	0	0	0
18	-1	1	12	-6	36	+6
24	+5	25	25	+7	49	+35
23	+4	16	30	+12	144	+48
27	+8	64	10	-8	64	-64
$\Sigma X = 114$	$\Sigma x=0$	$\Sigma x^2=236$	$\Sigma Y=108$	$\Sigma y=0$	$\Sigma y^2=318$	$\Sigma xy = +70$

$n = 6$

Mean of variable $\overline{X} = \frac{\Sigma X}{N} = \frac{114}{6} = 19$

Mean of variable $\overline{Y} = \frac{\Sigma Y}{N} = \frac{108}{6} = 18$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

$$= \frac{70}{\sqrt{236 \times 318}} = \frac{70}{\sqrt{75046}} = \frac{70}{\sqrt{273.95}} = 0.255$$

Short Cut Method :

In direct method, the mean is a whole number. Therefore, method is simple. But where mean is in fraction, it will involve difficult calculations. To meet such a situation, short cut method is used. In this method, the deviations are calculated from assumed mean and following formula is used

to calculate the coefficient of correlation (r).

$$r = \frac{\sum dx.dy - \frac{(\sum dx.\sum dy)}{N}}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{N}}\sqrt{\sum dy^2 - \frac{(\sum dy)^2}{N}}}$$

Where,

$\sum dx$ = Sum of deviation of X series from its assumed mean i.e. $\sum(X-A_x)$

$\sum dy$ = Sum of deviation of Y series from its assumed mean i.e. $\sum(Y-A_y)$

$\sum dx^2$ = Sum of square of deviations of x from assumed mean i.e. $\sum(X-A_x)^2$

$\sum dy^2$ = Sum of square of deviations of y from assumed mean i.e. $\sum(Y-A_y)^2$

$\sum dxdy$ = Sum of Products of deviations of x and y series from their respective assumed means dx.dy = $\sum(X-A)\ (Y-A)$

Example 2 :

Calculate coefficient of correlation from the following data :

Experience :	(X)	16	12	18	4	3	10	5	12
Performance:	(Y)	23	22	24	17	19	20	18	21

Solution :

	A = 10			**A = 120**		
X	**dx = X - 10**	**dx^2**	**Y**	**dy = y - 20**	**dy^2**	**dx.dy**
16	+6	36	23	+3	9	18
12	+2	4	22	+2	4	+4
18	+8	64	24	+4	16	+32
4	-6	36	17	-3	9	+18
3	-7	49	19	-1	1	+7
10	0	0	20	0	0	0
5	-5	25	18	-2	4	+10
12	+2	4	21	+1	1	+2
n=8	$\sum dx=0$	$\sum dx^2=218$		$\sum dy = +4$	$\sum dy^2 = 44$	$\sum dxdy = +91$

$$r = \frac{\sum dxdy - \frac{\sum dx . \sum dy}{N}}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{N}} \sqrt{\sum dy^2 - \frac{(\sum dy)^2}{N}}}$$

$$r = \frac{91 - \frac{0 \times 4}{8}}{\sqrt{218 - \frac{(0)^2}{8}} \sqrt{44 - \frac{(4)^2}{8}}}$$

$$r = \frac{91}{\sqrt{9156}} \quad \frac{91}{95.687} = 0.951$$

2.1.6.2 Rank Correlation

Prof. C.E. Spearman has given a method of judging Correlation between two attributes which cannot be measured in quantitative terms such as beauty wisdom, honesty, intelligence, etc. In other words, it is used when data are of qualitative nature. In such cases coefficient of Correlation is calculated by using the following formula :

$$r_k = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

Where r_k = Coefficient of rank correlation.
n = number of pairs of items.
$\sum D^2$ = sum of squares of differences in Ranks.

A. When Ranks are given :

(i) Take the differences of the two ranks i.e. (R_1—R_2) and denote these differences by D.

(ii) Square these differences and obtain $\sum D^2$

(iii) Apply the formula

$$r_k = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

Example 3 :

Calculate Rank Correlation coefficient between the ranks given for X and Y variables :

X	:	2	1	4	3	5	7	6
Y	:	1	3	2	4	5	6	7

x (R_1)	y (R_2)	(R_1-R_2) D	D^2
2	1	1	1
1	3	-2	4
4	2	+2	4
3	4	-1	1
5	5	0	0
7	6	+1	1
6	7	-1	1
			$\sum D^2 = 12$

$$r_k = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

Here, $\sum D^2 = 12$ $N = 7$

$$r_k = 1 - \frac{6 \times 12}{7^3 - 7} = 1 - \frac{72}{343 - 7}$$

$$= 1 - \frac{72}{336}$$

$$= \frac{336 - 72}{336} = \frac{264}{336}$$

$r_k = 0.785$

B. When Ranks are not given :

When we are given the actual data and not the ranks, it will be necessary to assign the ranks. Ranks can be assigned by taking either highest value as I or the lowest value as 1. The same method is used in case of both the variables.

Example 4 :

Calculate the coefficient of correlation from the following data by Spearman's Rank difference method :

	Price of Sugar (Rs.)	Price of tea (Rs.)
	75	120
	60	150
	80	115
	81	110
6	50	140

Solution :

Price of Sugar (Rs.)	R_1	Price of tea (Rs.)	R_2	D R1-R2	$(R_1-R_2)^2 = D^2$
75	3	120	3	0	$0^2=0$
60	2	150	5	-3	$(-3)^2=9$
80	4	115	2	2	$2^2=4$
81	5	110	1	4	$4^2=16$
50	1	140	4	-3	$(-3)^2=9$
					$\sum D^2=38$

$$r_k = 1 - \frac{6\sum D^2}{N\left(N^2-1\right)}$$

$$r_k = 1 - \frac{6\times 38}{120} = 1 - \frac{228}{120} = \frac{120-228}{120}$$

$$r_k = \frac{-108}{120} = -0.9$$

C. Equal Ranks :

In such a case, it is necessary to give each individual an average rank.

$$r_k = 1 - \frac{6\left[\sum D^2 + \frac{1}{12}\left(m^3-m\right) + \frac{1}{12}\left(m^3-m\right) + ...\right]}{N^3-N}$$

where m = no. of items whose ranks are equal.

Example 5 :

Compute Speaman's Rank Correlation from the following data.

Marks in Economics : 50 60 65 70 75 40 70 80

Marks in Maths : 80 71 60 75 91 82 70 50

Solution :

Let the marks in Eco. be denoted as X and marks in Maths be denoted as Y

X	R_1	Y	R_2	$R_1-R_2=D$	$(R_1-R_2)^2 = D^2$
50	2	80	6	-4	16
60	3	71	4	-1	1
65	4	60	2	+2	4
70	5.5	75	5	+0.5	0.25
75	7	91	8	-1	1
40	1	82	7	-6	36
70	5.5	70	3	+2.5	6.25
80	8	50	1	+7	49
					$\sum D^2 = 113.5$

$$r_k = 1 - \frac{6\left[\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \ldots\ldots\ldots\ldots\right]}{N^3 - N}$$

$$= 1 - \frac{6\left[113.5 + \frac{1}{12}(2^3 - 2)\right]}{8^3 - 8}$$

$$= 1 - \frac{(113.5 + 0.5)}{504}$$

$$= 1 - \frac{6 \times 114}{504}$$

$$= 1 - \frac{114}{84} = \frac{84 - 114}{84}$$

$$r = -\frac{30}{84}$$

Exercise

1. Calculate the correlation coefficient for the following heights (in inches) of fathers (X) and their sons (Y) :

X	:	65	66	67	67	68	69	70	72
Y	:	67	68	55	68	72	72	69	71

2. Calculate rank coefficient of correlation :

X	:	80	78	75	75	75	68	67	60	59
Y	:	12	13	14	14	14	14	16	15	17

2.1.7 Summary

In this lesson the concept of correlation or the association between two variables has been discussed. Various types of correlation have been discribed. Karl Pearson correlation coefficient r quantifies the association between two variables. The correlation coefficient r may assume values between -1 and 1. The sign indicates whether the association is direct (+ve) or inverse (-ve). Value of r equal to one indicates perfect association while value of r equal to zero indicates no assoication. Further Spearman's rank correlation for data with ranks, without ranks and equal ranks have been outlined.

2.1.8 Further Readings

S.C. Gupta : Fundamentals of statistics
S.P. Gupta : Statistical Methods
S.L. Aggarwal: Quantitative Methods
S.L. Bhardwaj

2.1.9. List of Questions

2.1.9.1 Short Questions

(a) Explain the meaning of the concept of correlation.
(b) Define Karl Pearson's Coefficient of Correlation.
(c) What is Spearman's rank correlation coefficient?
(d) Point out the difference between Linear and Non-Linear Correlation.
(e) Illustrate Positive Correlation with examples.

2.1.9.2 Long Questions

(a) Compute Karl Pearson's Coefficient of correlation between X and Y from the following observations.

X	:	1	2	3	4
Y	:	1	4	9	16

(b) Define correlation. Explain various types of correlation with suitable examples.

(c) Calculate the Karl Pearson's coefficient of correlation for the follwing ages of husband and wives at the time of their marriage

Age of Husband (in years) :	23	27	28	28	28	30	30	33	35	38
Age of Wife (in years) :	18	20	22	27	21	29	27	29	28	29

(d) Two judges in a beauty competition rank the 12 enteries as following:

X :	1	2	3	4	5	6	7	8	9	10	11	12
Y:	12	9	6	10	3	5	4	7	8	2	11	1

What degree of agreement is there between two judges?

(e) Calculate Spearman's rank correlation coefficient

Advertisement cost (....RS) :	39	65	62	92	82	75	25	98	36	78
Sales (Lakhs Rs) :	47	53	58	86	62	68	60	91	51	84

LESSON NO. 2.2 AUTHOR : DR. VIPLA CHOPRA

REGRESSION ANALYSIS

Structure

2.2.1 Introduction

Regression is a statistical device for measuring or estimating relationship between variables. Regression means to revert or to return back. The term was first introduced by Sir Francis Galton in 1877. He found, in his study of the relationship between the heights of fathers and sons, that all fathers were likely to have tall sons and short fathers were likely to have short sons. However, the mean height of the sons of tall fathers was lower than the mean height of their fathers, and the mean height of the sons of short father, was higher than the mean height of their short fathers. He referred to this tendency to return to the mean height of all men as regression in his research paper.

In regression analysis we have to assume one variable as the independent variable and the other as dependent variable. After estimating the relationship between variables, one can predict the most likely values of dependent variable on the basis of given values of independent variable. Thus, it indicates the average relationship between two or more variables.

According to Ya-lun-Chau, "Regression analysis attempts to establish the nature of relationship between variables that is to study the functional

relationship between the variables and thereby provide a mechanism for prediction or forecasting."

2.2.2 Objectives

After completion of this lesson you will be able to :

- understand the nature of relationship between variables
- differentiate between correlation and regression
- estimate regression equations of Y on X and X on Y.
- describe the properties of regression coefficient.

2.2.3 Difference between Correlation and Regression :

1. The correlation Coefficient is a measure of degree of covariability between two variables while the regression establishes a functional relationship between dependent and independent variables so that the former can be predicted for a given value of the later.
2. Correlation merely ascertains the direction and degree of relationship between two variables, but it does not clearly specify as to which variable is the cause and which is the effect. But this cause and effect relationship is clearly indicated by regression analysis.

2.2.4 Regression Lines : The device used for estimating the value of one variable from the value of the other consists of a line through the points drawn in such a manner as to represent the average relationship between the two variables. Such a line is called line of regression. There are two regression lines. One line as the regression of X on Y and the other as the regression of Y on X.

2.2.4.1 Regression line of X on Y : The regression line of X on Y is formed by taking the most probable value of X for the given value of Y.

2.2.4.2 Regression line of Y on X : The regression line of Y on X is formed by taking the most probable value of Y for the given value of X.

(i) These two regression lines show the average relationship between two variables. If there is perfect correlation (i.e. $r = \pm 1$) both the lines will coincide i.e. there will be only one line (see Fig. 1, Fig. 2.).

(ii) In case $r = 0$, both the lines will cut each other at right angle i.e. parallel to X-axis and Y-axis (Fig. 3).

(iii) These lines cut each other at the point of means of X and Y (see Fig.4).

(iv) Nearer these lines are, greater will be extent of correlation between X and Y, (Fig.5, Fig.6).

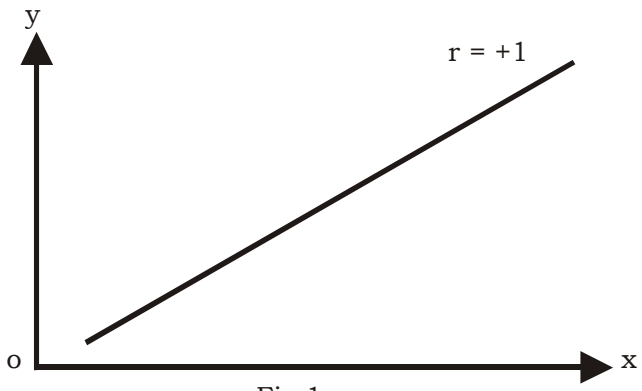


Fig.1

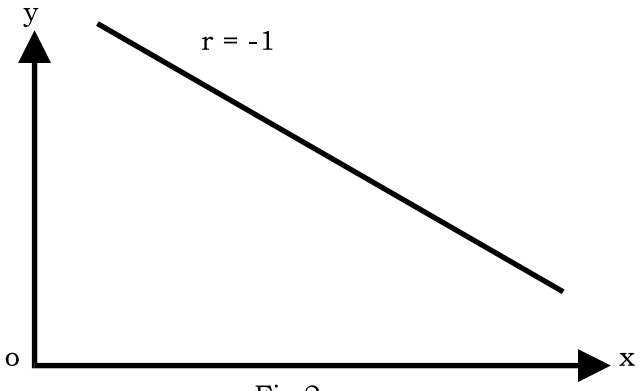


Fig.2

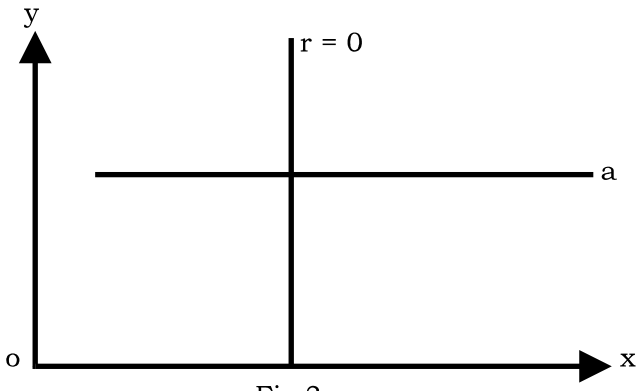


Fig.3

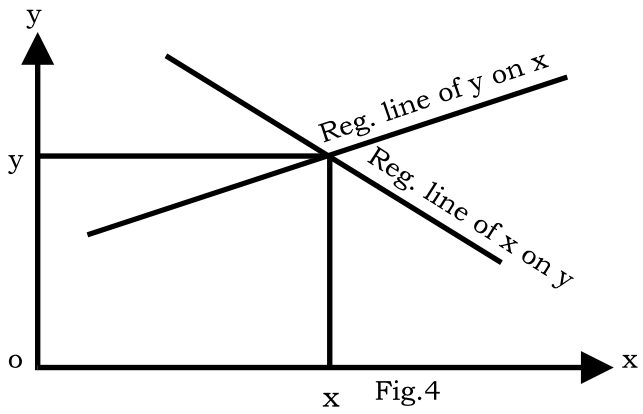


Fig.4

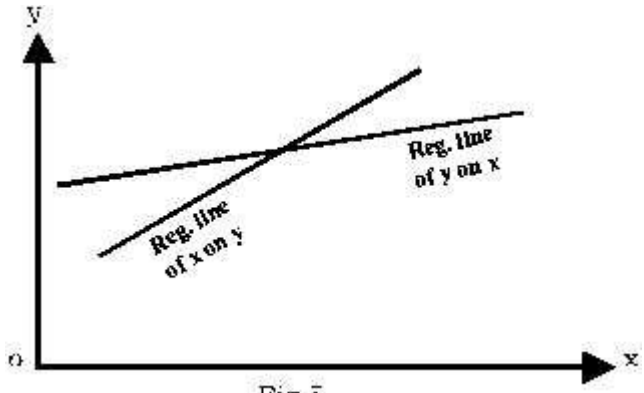


Fig.5

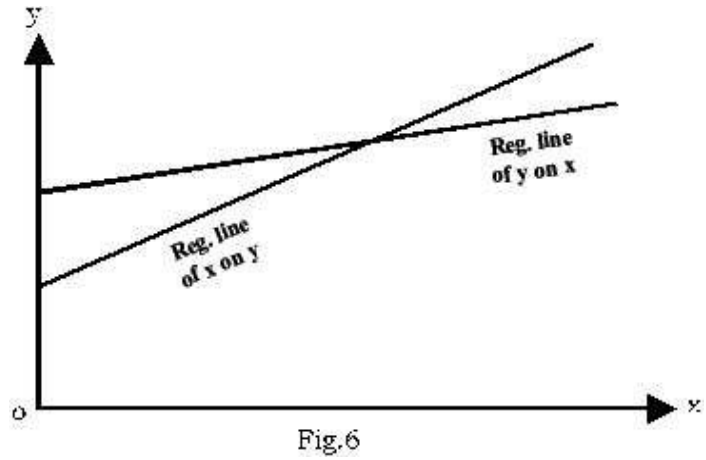
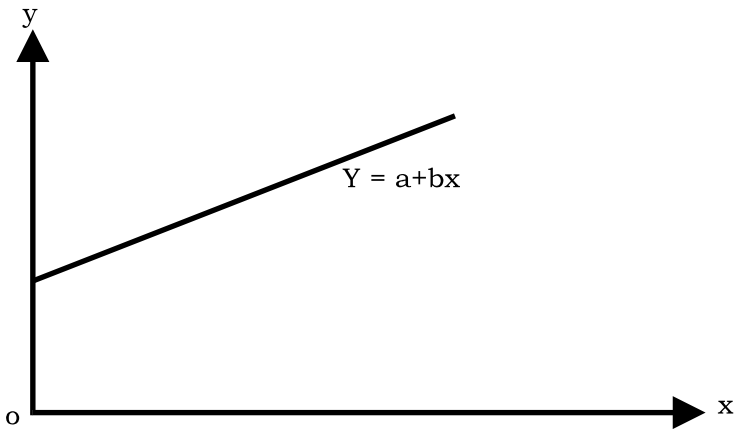


Fig.6

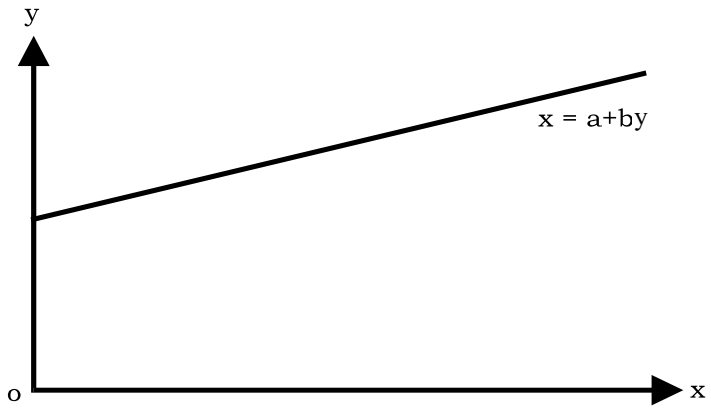
Assumptions : The regression lines are drawn on least square assumption which states that the sum of squares of the deviations of the observed 'Y' values from the fitted lines shall be minimum. The devitations from the points to the line of best fit can be measured in two ways—vertical, i.e. parallel to Y-axis and horizontal, i.e. parallel to X-axis. For minimising the total of the squares separately it is essential to have two regression lines.

Regression line of Y on X minimises total of the squares of the vertical deviations :

i. e. $\sum (y-y_c)^2$ is minimum.



Regression line of x on y
$\Sigma(x-x_c)^2$ is minimum



Regression line of x on y

2.2.5 Regression Equations :

Regression equations are the algebric expressions of the regression lines. There are two regression lines, so there will be two regression equations.

2.2.5.1 Regression equation of Y on X : Regression equation of Y on X describes the variation in the value of Y for the given changes in X. The regression equation of Y on X will be :

$y = a + bx$ where X and Y are variables, and a and b are constants.

The 'a' constant is the y-axis intercept, i.e., the point where regression line touches the y-axis.

The constant 'b' shows the slope of the line.

2.2.5.2 Regression equation of X on Y : Regression equation of X and Y describes the variation in the values of X for the given changes in Y. The regression equation of X on Y will be

$X = a + by$

Here, a tells us how high above the X-axis the regression line is started

b = slope of the line.

Normal Equations : Two normal equations have to be solved for finding the values of constants a and b and in the regression equation Y on X i.e. $Y = a + bx$.

They are :

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

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

and for regression equation X on Y i.e. $X = a + by$, they are

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

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

The coefficient of regressions are found by the formula :

Regression coefficient of X on Y, β_{xy} or $\beta_1 = r \frac{\sigma x}{\sigma y}$

Regression coefficient of Y on X or β_{yx} or $\beta_2 = r \frac{\sigma y}{\sigma x}$

where σx = Standard deviations of x series.

σy = Standard deviations of y series.

r = Coefficient of correlation of x and y series.

$$b_{xy} = r \frac{\sigma x}{\sigma y} = \frac{\sum xy}{N\sigma x \sigma y} \times \frac{\sigma x}{\sigma y} = \frac{\sum xy}{N6^2 y} = \frac{\sum xy}{\sum y^2}$$

Similarly, $b_{yx} = \frac{\sum xy}{\sum x^2}$

Example 1 :

From the following data obtain the two regression equations :

x	:	5	8	7	6	4
y	:	3	4	5	2	1

Solution :

	x	y	x^2	y^2	xy
	5	3	25	9	15
	8	4	64	16	32
	7	5	49	25	35
	6	2	36	4	12
	4	1	16	1	4
Total (∑)	30	15	190	55	98

Regression equation of Y on X : y = a + bx.

Now two normal equations are :

$\sum y = Na + b\sum x \therefore 15 = 5a + 30b$ ..... (i)

$\sum xy = a\sum x + b\sum x^2 \therefore 98 = 30a + 190b$ ..... (ii)

Multiply (i) by 6, we get

$90 = 30a \pm 180b$ ..... (iii)

Substract (iii) from (ii)

Put b = .8 in (1) we get

15 = 5a + 30 (.8)

or 15 = 5a + 24

a =1.8

By putting the values of a and b in equation, y = a + bx.

y = 1.8 + .8x.

Regression equation of X on Y : x = a + by.

Two normal equations are :

$\sum x = Na + b\sum y \therefore 30 = 5a + 15b$ ..... (i)

$\sum xy = a\sum y + b\sum y^2 \therefore 98 = 15a + 55b$ ..... (ii)

Multiply (i) by 3 we get :

$90 = 15a + 45b$

Again $98 = 15a + 55b$ ..... (ii)

and $90 = 15a \pm 45b$ ..... (iii)

— — —

Subtracting (iii) from (ii) we get :

8= 10b

b =.8

Put b = .8 in (i) we get

30 = 5a + 15(.8)

or 30 = 5a + 12

$\therefore$ — 5a = -18 or a = 3.6.

Putting the value of a and b we get,

x = 3.6 + .8y.

Example 2 : Two regression equations are

3x + 2y - 26 = 0 and

6x + y - 31 = 0. Find $\bar{x}$, $\bar{y}$ and r.

Also determine σy if σx = 5

Solution :

Calculation of $\bar{x}$, $\bar{y}$

3x + 2y - 26 = 0 .....(i)

6x + y - 31 = 0 ..... (ii)

Multiply (i) by 2 we get,

6x + 4y - 52 = 0

Substract (ii) from (iii) we get 3y – 21 = 0

$$\text{or} \quad y = \frac{21}{3} = 7\left(\bar{y} = 7\right)$$

substituting value of Y in (i) :

3x + 2×7 - 26 = 0

3x + 14 - 26 = 0

3x - 12 $\therefore$ $\bar{x}$ = 4 ($\bar{x}$ = 4)

Calculation of 'r' eq (i) is y/x : 2y = 26 -3x.

y = 13 - 1.5 X or b_{yx} = – 1.5

x/y : 6x = 31 - Y

or x = 5.17 - .17y

or b_{xy} = -.17

$r = \pm\sqrt{b_{xy}.b_{yx}} = \sqrt{-1.5 \times -.17} = \sqrt{.255}$

r = –.5

Regression coefficient of y on x : b_{yx} = r. σy/σx

$$-1.5 = -.5\ \frac{\sigma y}{5}$$

or $-.5\sigma y = -7.5 \quad \therefore \sigma y = 15.$

2.2.6 Properties of Regression Coefficients :

1. The geometric mean between two regression coefficients is coefficient of correlation :

$$\sqrt{b_{xy} . b_{yx}}$$

2. If one regression coefficient is greater than unity other regression coefficient must be less than unity.
3. Regression coefficients are independent of origin but not of scale.

Example 3 :

If $b_{xy} = .8$ and $b_{yx} = .6$.

What would be the value of the coefficient of correlation.

Solution :

The value of coefficient of correlation is the geometric mean of the two regression coefficients. That is,

$$r = \sqrt{.8 \times .6} = \sqrt{.48}$$

Example 4 :

Two regression equations are given as

$x - 4y = -13$ and $9y - x = 53$ and $\sigma x = 12$

We have to find (a) Mean of X and mean of Y.

(b) Coefficient of Correlation.

Solution :

(a) Means of X and Y

$\bar{x}$ and $\bar{y}$ can be obtained by solving the given equations simultaneously for x and y.

The given equations are $x - 4y = -13$ ..... (i)

$-x + 9y = 53$ ..... (ii)

Adding (i) and (ii) we get $5y = 40$ or $y = 8$.

Put $y = 8$ in (i), $x - 32 = -13$ or $x = 19$

(b) Coefficient of correlations :

From the given equation, we do not know which regression equation is y on x and which regression equation is x on y. Hence, we have to take one equation as y on x and the other as x on y.

If $b_{yx} . b_{xy} \leq 1$, then our assumption is correct.

But if $b_{yx} . b_{xy} > 1$, then we change the assumption.
Let x - 4y = -13 is regression line of y on x.
and 9y - x = 53 is regression line of x on y.
From y on x equation.
x - 4y = -13
or 4y = x + 13

$$\therefore y = \frac{1}{4}x + \frac{13}{4}$$

Hence $b_{yx} = \frac{1}{4}$

From x on y equation.
9y - x = 53
x = 9y - 53
Hence $b_{xy} = 9$
Now $b_{yx} . b_{xy} = \frac{1}{4} \times 9 = \frac{9}{4} = 2.25 > 1$ This shows that our assumption is wrong.

Deviations taken from Assumed Means :

When actual means of x and y variables are in decimals, the calculations can be simplified by taking deviations from the assumed means.

The two regression equations are

$$x - \bar{x} = r \frac{\sigma x}{\sigma y}\left(y - \bar{y}\right)$$

$$\text{Where } r\frac{\sigma x}{\sigma y} = b_{xy} = \frac{\sum dxdy - \frac{\sum dxdy}{N}}{\sum dy^2 - \frac{\left(\sum dy\right)^2}{N}}$$

dx = x - A and dy = y - A.
Similarly, the regression equation of y on x is

$$y - \bar{y} = r\frac{\sigma y}{\sigma x}\left(x - \bar{x}\right)$$

$$r\,\frac{\sigma y}{\sigma x} = b_{yx} = \frac{\sum dxdy - \frac{\sum dx \sum dy}{N}}{\sum dx^2 - \frac{(\sum dx)^2}{N}}$$

Example 5 :

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

x	d_x= **x-5**	**dx^2**	**y**	**dy^2**	d_y=y-	**dxdy**
6	+1	1	9	2	4	2
2	–3	9	11	4	16	–12
10	+5	25	5	–2	4	–10
4	–1	1	8	1	1	–1
8	+3	9	7	0	0	0
$\Sigma x=30$	$\Sigma dx=5$	$\Sigma dx^2=45$	$\Sigma y=40$	$\Sigma dy=5$	$\Sigma dy^2=25$	$\Sigma dxdy=-21$

Regression equation of x on y : $x - \bar{x} = r\,\frac{\sigma x}{\sigma y}(y - \bar{y})$

$$\bar{x} = \frac{\sum x}{N} = \frac{30}{5} = 6 \quad \bar{y} = \frac{\sum y}{N} = \frac{40}{5} = 8$$

$$r\,\frac{\sigma x}{\sigma y} = \frac{N\sum d\,x\,dy - \sum d\,x\,\Sigma dy}{N\sum dy^2 - (\sum dy)^2}$$

$$= \frac{5(-21) - (5)\,5)}{(5)(25) - (5)^2} = -\frac{105-25}{125-25} = \frac{130}{100} = -1.3$$

x –6 = – 1.3 (y–8)

x–6 = –1.3y + 10.4 or x = 16.4 – 1.3y

Regression Equation of y on x : $y - \bar{y} = r\,\frac{\sigma y}{\sigma x}(x - \bar{x})$

$$r\frac{\sigma y}{\sigma x} = \frac{N\sum dxdy - \sum dx \sum dy}{N\sum dx^2 - (\sum dx)^2}$$

$$= \frac{5(-21) - (5)(5)}{(5)(45) - (5)^2} = \frac{-105-25}{225-25} = -.65 - \frac{130}{200}$$

$y - 8 = -0.65 (x - 6)$

$y - 8 = -0.65x + 3.9\ 0$

$y = -0.65x + 11.9$ $y = 11.9 - 0.65x$

EXERCISE

1. What is regression and what do regression lines indicate ?
2. From the following data obtain the two regression equations.

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

3. Two regression lines are $x + 4y + 3 = 0$ and $4x + 9y + 5 = 0$

Determine x, y and r.

2.2.7 Summary

In this liesson fundamentals of linear regression have been highlighted. Two regression lines namely y on x and x on y have been illustrated graphically. The algebric expression of these regression lines namely regression equations have been illustrated using examples.

2.2.8 Further Readings

1. Statistical Methods : S.P. Gupta
2. Statistical Methods : C.B. Gupta
3. Statistics : B.N. Gupta
4. Quantitative Methods : S.L. Aggarwal, S.L. Bhardwaj

2.2.8 List of Question

2.2.8.1 Short Questions

(a) Define regression.

(b) Why are there two lines of regression in two variable linear regression?

(c) Explain the concept of regression co-efficients.

(d) Explain the difference between correlation and regression

(e) If $b_{xy} = -\frac{1}{2}, b_{yx} = -\frac{7}{4}$, Find r.

2.2.8.2 Long Questions

(a) Find the lines of regression using least square technique.

(b) From the following data, find out the most probable value of x when y = 12

	x series	y series
mean	25	22
standard deviation	4	5

coefficient of correlation between x and y = +.8

(c) Fit a straight line regression of y on x from the following data

X :	0	1	2	3	4	5	6
Y:	2	1	3	2	4	3	5

(d) Find the regression lines of y on x, and x on y, for the three pairs of observations

X = 1, 2, 3

Y = 2, 4, 5

From two regression equations. Caluculate the value of r

(e) Calculate the coefficient of regression for the distribution given below

X	:	8	6	4	7	5
Y	:	9	8	5	6	2

LESSON NO. : 2.3 AUTHOR : DR. PUSHPINDER SINGH GILL

INDEX NUMBERS

Structure

2.3.1 Introduction

An Italian named of G.R. Carli introduced index numbers in statistical studies. Index numbers are used in the fields of economics and business for making comparison of prices, of cost of production, of employment, of purchasing power of money or comparison of wages. In this lesson meaning, purpose, types of index numbers and problems relating to index numbers will be discussed in detail.

2.3.2 Objectives

After going through this lesson you will be able to

- know the meaning and purpose of index numbers
- realise the problems in the construction of index numbers.
- describe/construct various types of index numbers.
- judge the adequacy of index numbers

2.3.3 Meaning : An index number is a statistical measure designed to show changes in a variable or a group of related variables with respect to time, geographical location, income, or any characteristic. A collection of index numbers for different years, locations, etc. is sometimes called an index series.

2.3.3.1 Purpose of Index Numbers

Index numbers are used for making comparison. For example, with index numbers we can compare food or other living costs in a city during a given year in part of a country with that in another part. Although used mainly in business and economics, index numbers can be applied in many other fields. In education, for example, we can use index numbers to compare the relative intelligence of students in different locations or for different years.

Many government and private agencies are engaged in computing index numbers for purposes of forecasting business and economic conditions, providing general information , and so forth. Thus, we have wage indices, production indices, unemployment indices and many other. The best known is the cost of living index or consumer price index, prepared by the U.S. Bureau of Labour Statistics.

2.3.3.2 Problems related to Index Numbers

Several things can distort index numbers. The four most common causes of distrotions are :

1. **Limited Data :** Sometimes there is difficulty in finding suitable data to compute an index. Suppose the sales manager of colonial aircraft is interested in computing an index describing seasonal variations in the sale of company's small planes. If sales are reported only on an annual basis, he would be unable to determine the seasonal sales pattern.
2. **Incomparability of Indices :** Occurs when attempts are made to compare one index with another often there has been a basic change in what is being measured.
3. **Inappropriate weighing of factor :** Inappropriate weighing of factor can also distort an index. In developing a composite index, such as the consumer price index, we must consider changes in some variables to be more important than changes in others.
4. **Use of an Improper Base :** Selection of an improper base also leads to distribution of index numbers. The base period should be normal, i.e., it must be free from natural calamities like earthquakes, famines, floods etc. and other abnormalities like wars, booms, depressions etc. Further, the base period should not be too distant from the past.

2.3.4 Methods of Constructing Index Numbers

Broadly speaking, large number of formulae have been grouped under two heads:

(a) Unweighted aggregative index

(b) Weighted aggregative index

2.3.4.1 Unweighted aggregative index

The simplest form of a composite index is an unweighted aggregative index. Unweighted means that we add, or sum, all the values, the principal advantage of an unweighted aggregate index is its simplicity.

An unweighted aggregative index is calculated by adding all the elements in the composite for the given time period and then dividing this result by the sum of same elements during the base period.

$$\text{Price Index } P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$$

$$\text{Quantity Index } Q_{01} = \frac{\sum Q_1}{\sum Q_0} \times 100$$

Example 1 :

From the following data construct an index for 1994 taking 1993 as base :

Commodity	Price in 1993 (Rs.)	Price in 1994 (Rs.)
A	50	70
B	40	60
C	80	90
D	110	120
E	20	20
	$\Sigma P_0 = 300$	$\Sigma P_1 = 360$

$$P01 = \frac{\sum P_1}{\sum P_0} \times 100 = \frac{360}{300} \times 100 = 120$$

This means a net increase in the price of commodities to the extent of 20%

In unweighted index numbers, no importance is given to the relative importance of commodities. Quantities and prices are not taken into consideration simultaneously while constructing an index number.

2.3.4.2 Weighted Aggregated Index

As we know sometimes we have to attach greater importance to changes in some variables than to others when we compute an index. This weighting allows us to include more information than just the change in price over time. It also helps us to improve the accuracy of the general price level estimation based on our sample. The problem is to decide how much weight to attach to each of the variables in the sample.

We use P_0 for base year's price and P_1 for current year's price level. Similarly, for quantity index, meaning thereby 0 stands for base period and 1 stands for current period. The following methods to construct an index number are named after persons who have suggested them.

Lespeyre's Index, or the Base-year method (1871)

$$P_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100$$

Laspeyres index attempts to answer the question : "What is the change in aggregate value of the base period list of goods when valued at given period prices?" This index is very widely used in practical work.

The main limitation of the Lespeyres method is that it does not take into account the consumption pattern. It has an upward bias. When prices increase there is tendency to reduce the consumption of higher priced items. Hence, by using base year weights, too much weight will be given to these items which have increased price the most. Similarly, when prices decline, consumers shift their purchases to those items which decline the most. By using base period weights, too little weight is given to those items which decrease most in price, again overstanding the index.

Passche's Index or Given-year Method

$$P_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_1} \times 100$$

In general, this formula answers the question : 'What would be the value of the given period list of goods when valued at base period prices?" The difficulty in computing the Passche index in practice is that revised weights, or quantities, must be computed each year or each period, additing to the data collection expense in the preparation of the index. For this reason, the Passche index is not used frequently in practice where the number of the commodities is large.

Fisher's Ideal Index

$$P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100 \quad \text{or } P_{01} = \sqrt{L \times P}$$

It is clear that Fisher's ideal index is the geometric mean of Laspeyres and Passche's Indices.

The formula is known as 'Ideal' because of the' following reasons :

(i) It is based on the geometric mean which is theoretically considered to be the best average of constructing index numbers. Thus, it is free from

bias.

(ii) It takes into account both current year as well as base year prices and quantities.

(iii) It satisfies both the time reversal test as well as factor reversal test.

2.3.5 Tests of Adequacy :

2.3.5.1 Time Reversal Test : Reversal Test is a test to determine whether a given method will work both ways in time, forward and backward. In the words of Fisher, "The test is that the formula for calculating the index number should be such that it will give the same ratio between one point of comparison and the other, no matter which of the two is taken as base."

Symbolically, the following equation should be satisfied :

$P_{01} \times P_{10} = 1$

The test is not satisfied by Laspeyres and Passche method. Fisher's ideal index satisfies the test.

$$P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}}$$

$$P_{10} = \sqrt{\frac{\sum P_0 q_1}{\sum P_1 q_1} \times \frac{\sum P_0 q_0}{\sum P_1 q_0}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \times \frac{\sum P_0 q_1}{\sum P_1 q_1} \times \frac{\sum P_0 q_0}{\sum P_1 q_0}} = \sqrt{1} = 1$$

2.3.5.2 Factor Reversal Test holds that the product of a price index and the quantity index should be equal to the corresponding value index. In the words of Fisher, "Just as each formula should permit the interchange of two times and quantities without giving inconsistent results so it ought to permit interchaning the prices and quantities without giving inconsistent results, i.e., the two results multiplied together should give the true value ratio."

Symbolically.

$$P_{01} \times Q_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_0}$$

$$P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}}$$

$$Q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum P_0 q_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}}$$

= Value Index

$$P_{01} \times Q_{01} = \sqrt{\frac{(\sum P_1 q_1)^2}{(\sum P_0 q_0)^2}} = \frac{\sum P_1 q_1}{\sum P_0 q_0} = \text{Value Index}$$

All the above mentioned formulae are for fixed base index, i.e., the base remains the same throughout the series of the index.

Example : 2

Construct index numbers of price from the following data by applying :

1. Laspeyre's Method
2. Passche's Method
3. Fisher's Ideal Method.

	1993		1994	
Commodity	Price	Qty	Price	Qty
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

Solution :

	1993		1994					
Commodity	P_0	q_0	P_1	q_1	$P_1 q_0$	$P_0 q_0$	$P_1 q_1$	$P_0 q_1$
	2	8	4	6	32	16	24	12
	5	10	6	5	60	50	30	25
	4	14	5	10	70	56	50	40
	2	19	2	13	38	38	26	26
					$\Sigma P_1 q_0 = 200$	$\Sigma P_0 q_0 = 160$	$\Sigma P_1 q_1 = 130$	$\Sigma P_0 q_1 = 103$

1. Laspeyre's Method : $P_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100$

$$P_{01} = \frac{200}{160} \times 100 = 125$$

2. Passche's Method : $P_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_1} \times 100$

$$= \frac{130}{103} \times 100 = 126.21$$

3. Fisher's Ideal Method : $P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100$

$$= \sqrt{\frac{200 \times 130}{160 \times 103}} \times 100$$

$$= \sqrt{1.578} \times 100$$

$$= 1.256 \times 100$$

$$= 125.62$$

2.3.6 Errors in Interpretation

1. Generalisation from a specific index

One of the most common misinterpretations of an index is generalisation of the results. The Consumer Price Index measures how prices of a particular combination of goods purchased by moderate-income urban India has changed. Despite its specific definition, the consumer price Index is frequently described as reflecting the cost of living for all Indians. Although it is related to the cost of living to some degree, to day that it measures the changes in the cost of living is not correct.

2. Lack of general knowledge regarding published indices

Part of the problem leading to the first error is lack of knowledge of what the various published indices measure, all the well-known indices are accompained by detailed statements concerning measurement.

3. Effect of time span on an Index

Factors related to an index tend to change with time. In particular, the appropriate weights tend to change. Thus, unless the weights are changed accordingly, the index becomes less reliable.

4. Quality Index

One frequent criticism of index numbers is that they do not reflect changes in the quality of the items they measure. If the quality has indeed changed, then the index either understates or overstates the price-level changes.

Thus, the index numbers are to be constructed carefully, keeping all the above

mentioned problems and errors in mind.

2.3.7 Fixed-Base and Chain-Base Index

As stated earlier, all the above mentioned formulae for the construction of index numbers are fixed base metod. As time lapses conditions which were once important become less significant and it becomes difficult to compare accurately present conditions with those of a remote period. New items may have to be included and old ones may have to be deleted in order to make the index more representative. In such cases, it may be desirable to use chain base index, when this method is used, the comparisons are not made with a fixed base, rather the base changes from year to year. For example, for 1994, 1993 will be the base; for 1993, 1992 will be the base , and so on.

$$\text{Chain Index for current year} = \frac{\text{Average link relative of current year} \times \text{Chain Index of previous year}}{100}$$

Steps :

(i) Express the figures for each year as percentages of the preceding year The results so obtained are called link relatives.

(ii) Chain together these percentages by successive multipication to form a chain index of previous year divided by 100

The link relatives obtained in step (i) facilitate comparisons from one year to another, i.e., between closely situated periods in which the q's are not likely to have changed much the chain indices obtained in step (ii) by a process of chaining binery comparisons facilitate long-term comparisons.

Example : 3

From the following data of the wholesale prices of wheat for ten years. construct index number taking (a) 1985 as base, and (b) by chain base method.

Year :	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
Price of Wheat : (As per 40 kg.)	50	60	62	65	70	78	82	84	88	90

Solution

Fixed-Base Indices :

Year	**Price of Wheat**	**Index number** (1985 = 100)
1985	50	100
1986	60	$\frac{60}{50} \times 100 = 120$
1987	62	$\frac{62}{50} \times 100 = 124$

1988	65	$\frac{65}{50} \times 100=130$
1989	70	$\frac{70}{50} \times 100=140$
1990	78	$\frac{78}{50} \times 100=156$
1991	82	$\frac{82}{50} \times 100=164$
1992	84	$\frac{84}{50} \times 100=168$
1993	88	$\frac{88}{50} \times 100=176$
1994	90	$\frac{90}{50} \times 100=180$

Chain Indices

Year	Price of Wheat	Link Relatives	Chain Indices
1985	50	100.00	100.00
1986	60	$\frac{60}{50} \times 100 =120$	$\frac{120x100}{100}=120$
1987	62	$\frac{62}{60} \times 100=103.33$	$\frac{103.33x120}{100}=124$
1988	65	$\frac{65}{62}x100=104.84$	$\frac{104.84x124}{100}=130$
1989	70	$\frac{70}{65}$ x 100= 107.69	$\frac{107.69x130}{100}=140$
1990	78	$\frac{78}{70}$ x 100= 111.43	$\frac{111.43x140}{100}=156$
1991	82	$\frac{82}{78}$ x100 = 105.44	$\frac{105.44x156}{100}=164$
1992	84	$\frac{84}{82}$ x100 = 102.44	$\frac{102.44x164}{100}=168$

1993	88	$\frac{88}{84}$ x100 =104.76	$\frac{104.76x168}{100} = 176$
1994	90	$\frac{90}{80}$ x100 = 102.27	$\frac{102.27x176}{100} = 180$

Example 4:

Compute the chain base index number with 1990 as base from the following table giving the average wholesale prices of the commodities A,B and C for the years 1991 to 1995.

	Average wholesale price (in Rs.)				
Commodity	1991	1992	1993	1994	1995
A	20	16	28	35	21
B	25	30	24	36	45
C	20	25	30	24	30

Solution **COMPUTATION OF CHAIN INDICES**

	Relatives based on preceding year				
Commodity					
	1991	1992	1993	1994	1995
A	100	$\frac{16}{20}x100 = 80$	$\frac{28}{16}x100 = 175$	$\frac{35}{28}x100 = 125$	$\frac{21}{35}x100 = 60$
B	100	$\frac{30}{25}x100 = 120$	$\frac{24}{30}x100 = 80$	$\frac{36}{24}x100 = 150$	$\frac{45}{36}x100 = 125$
C	100	$\frac{25}{20}x100 = 125$	$\frac{30}{25}x100 = 120$	$\frac{24}{30}x100 = 80$	$\frac{30}{24}x100 = 125$
Total of Link relatives	300	325	375	355	310
Average of Link relatives	100	108.33	125	118.33	103.33
Chain Index	100	$\frac{108.33x100}{100}$ = 108.33	$\frac{125x108.33}{100}$ =135.41	$\frac{118.33x135.41}{100}$ 160.23	$\frac{103.33x160.23}{100}$ 165.57

Merits of the Chain Base Method

1. The chain base method has a great significance in practice because in economic and business data we are more often concerned with making comparisons with the previous period and not with any distant past. The link relatives obtained by Chain base method serve this purpose.
2. Chain base method permits the introduction of new commodities and the deletion of old ones without necessitating either of them.

2.3.8 Summary

In this lesson the concept of index numbers and problems faced in the construction of index numebrs have been highlighted. Besides steps in the costruction of various types of index numebrs have been explained alongwith illustrations. Both fixed base and chain base methods in the construction of index number have been explained the end.

2.3.9 Further Readings

S.P. Gupta : Statistical Methods
S.C. Gupta : Fundamentals of Statistics
S.L Aggarwal : Quantitative Methods

2.3.10 List of Questions

15.10.1 Short Questions

(a) What is an index number?
(b) Explain time reversal test and factor reversal test.
(c) Explain some merits of chain base method.
(d) Give the formula for the calculation of Laspeyre's index number.
(e) State two tests for a good index number

2.3.10.2 Long Questions

(a) What are index numbers ? How are they constructed? Discuss the limitations of index numbers.
(b) What is meant by revesibility of an index number? Discribe the time and factor reversal tests in the theory of index numbers. Give a formula which satisfies both these tests.
(c) Construct index number of price from the following data by applying :
   1. Laspeyre's method.
   2. Paasche's method
   3. Fisher's method

Commodities	Base Year		Current Year	
	Price	Quantity	Price	Quantitiy
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

(d) Calculate the quantity index number using Fisher's formula for the following data and show that it satisfies the time reversal test

Commodity	1973		1974	
	Price	Quantity	Price	Quantity
A	6	70	8	120
B	8	90	10	100
C	12	140	16	280

(e) From the chain base index numbers given below, prepare fixed base index numbers

Years :	**1963**	**1964**	**1965**	**1966**	**1967**
Index numbers :	**80**	**110**	**120**	**105**	**125**

LESSON NO. : 2.4 **AUTHOR : DR. VIPLA CHOPRA**

TIME SERIES ANALYSIS

Stucture

2.4.1 Introdcution

Time series data are of great importance in the field of Economics and Commerce In economics many economic variables vary with the change of time like agricultural production, industrial production, imports, exports, prices etc. Economists have to arrange these data on the basis of time of their occurrence for the purpose of their proper analysis and interpretation. In this lesson time series data will be analysed using different methods.

2.4.2 Objectives :

After completion of this lesson you should be able to

- learn the meaning of time series data
- Know the various components of time series
- determine the trend by using differnt methods

2.4.3 Meaning of Time Series : Statistical data can be arranged in a number of ways according to magnitude or size, according to place of occurrence or geographical

location and according to the time of their occurrence, they form a time series. Thus time series is the arrangement of statistical data in chronological order - daily, weekly, fortnightly, monthly or yearly. The data on the population of India is a time series data where time interval between two successive figures is 10 years. Similarly figures of national income, agricultural and industrial production, etc., are available on yearly basis. Time series are of particular importance in the field of business and economics because variables like price, wages, production, sales, profits, etc., vary from one time period to another. Businessmen and economists are greatly interested in the study of such time series because this study helps them in making forecasts, and planning for the future.

2.4.4 Components of a Time Series :

When the data are arranged on the basis of time of their occurrence, often they show fluctuations from time to time - from day to day, from week to week, from month to month, and from year to year. These fluctuations are caused by a constantly working composite force. This composite force has four components, commonly known as the components of a time series which are as follows :

1. Secular trend or long term movements (T)
2. Periodic movements or short term fluctuations.
   These comprise of -
   (i) Seasonal fluctuations (S)
   (ii) Cyclical fluctuations (C) and
   (iii) Irregualr or random fluctuations (I)

The value of a variable over a period of time changes due to the combined impact of these four components. Therefore, it is necessary to isolate and measure the separate effects of these forces in a given time series. The process of analysis of time series aims at achieving this objective.

2.4.4.1. Secular Trend :

Secular trend or simply trend is the general tendency of the data to increase or decrease or stagnate over a long period of time. Most of the business and economic time series would reveal a tendency to increase or to decrease over a number of years. For example data regarding agricultural production, industrial production, population, bank deposits, deficit financing, etc., show that, in general, these magnitudes have been rising over a fairly long period. As opposed to this a time series may also reveal a declining trend, eg, with the improved medical facilities, the death rate is likely to show a declining trend.

According to A.E. Waugh, secular trend is, “that irreversible movement which continues, in general, in the same direction for a considerable period of time.” It

should be noted that trend refers to only smooth, regular , long term movement of the data and has nothing to do with sudden and erratic movements either in upward or downward direction.

Uses of Trend :

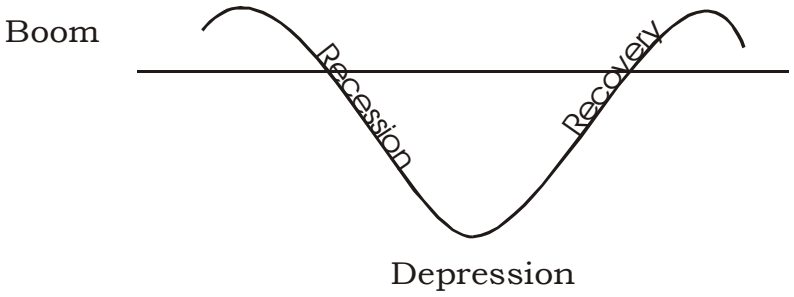
(i) The study of trend enables us to have a general idea about the pattern of behaviour of the phenomenon under consideration. This helps in business fore casting, and planning future operations.

(ii) In order to analyse the influence of other factors, the trend may first be measured and then eliminated from the observed values.

(iii) Trend values of two or more time series can be used for their comparison.

2.4.4.2. Periodic Movements/Short-term fluctuations :

Periodic movements, also known as oscillatory movements, repeat themselves after a regular interval of time. This time is known as the period of oscillation. These oscillations are shown in the following figure :



The oscillatory movement are termed as seasonal variations if their period is less than or equal to one year, and as cyclical variations if the period is greater than one year. Thus in a time - series data where only annual figures are given, there are no seasonal variations. Most of the time series relating to business and economics are influenced by seasonal forces, e.g., time - series relating to agricultural production and sales of agricultural produce, bank deposits, sales and profits in a departmental store, etc. Seasonal variations have two main causes.

(i) Climate and (ii) Customs.

The changes in climatic conditions affect the value of time series variable. For example, the sale of woolen garments is generally at its peak in the month of November because of the beginning of winter season. Similarly, timely rainfall may increase agricultural output, prices of agricultural commodities are lowest during their harvesting season, etc., reflect the effect of climatic conditions on the value of

time series variable. The customs and traditions of the people also give rise to the seasonal variations in time series. For example, the sale of ornaments may be highest during marriage season, sale of sweets during Diwali, etc., are variations that are the results of customs and traditions of the people.

Cyclical variations are revealed by most of the economic and business time series and, therefore, are also termed as trade (or business) cycles. Any trade cycle has four phases which are respectively known as boom, recession, depressin and recovery phases. These phases are shown in figure above. Various phases repeat themselves regularly one after another in the given sequence. The time interval between two identical phases is known as the period of cyclical variations. The period is always greater than one year. Normally, the period of cyclical variations lies between 3 to 10 years. In times of prosperity, production, sales, employment and other economic activities are high; in times of depression, the opposite is true. Thus a study of the cyclical fluctuations helps business executives in the formulation of policies aimed at stablising the level of business activity.

2.4.4.3 Random/irragular variations :

In addition to the influence of long term and short term forces, every time series is subjected to occasional influences, which may occur just once, or several times, but without any pattern or regularity. The variations they produce are, therefore, called irregular or random fluctuations. Wars, earthquakes, floods, strikes, lock outs fires and such other unforeseen or unforeseeable events are typical causes of erratic fluctuations.

2.4.5 Measurement of Trend :

Giving long-term series, we wish to determine and present the direction which it takes - is it growing or declining ? The various methods that can be used for determing trend are :

(i) Moving average method
(ii) Method of least square
(iii) Freehand or graphic method
(iv) Semi - average method

In the present lesson we will discuss the first two methods.

2.4.5.1 Moving Average Method

Moving average method is quite simple and is used for smoothing the fluctuations in curves. The trend values obtained by this method are very much accurate. According to Herbert Arkin and Raymond R. Colton, “A moving average of a time series is a new series obtained by finding out successively the average of a number of the original successive items choosen on the basis of periodicity of fluctuations dropping off one item and adding on the next at each stage” Moving

average is a series of arithmetic averages of variate values of a sequence of a fixed number of years. The first thing to be decided in this procedure is the period of moving average. The moving average may be for three, four, five, six years and so on according to the size and the periodicity of fluctuations of the data. The main purpose of moving average is to obtain trend by eliminating or in any case reducing to the minimum all other variations. This purpose can be achieved by selecting the moving average period as equivalent to the period of cycle. Selection of period is very significant Given the set of values of the variable (y) : y_1, y_2, y_3 ..... at times t_1, t_2, t_3 ..... the moving average of period N is defined and given by the sequence of the arithmetic averages.

$$\frac{y_1 + y_2 \ldots\ldots\ldots y_n}{N}, \quad \frac{y_2 + y_3 + \ldots\ldots\ldots y_n + 1}{N}, \quad \frac{y_3 + y_4 + \ldots\ldots\ldots y_n + 2}{N},$$

and so on.

Odd Period of Moving Average :

Suppose moving average is to be calculated for three years. First, we will take the total of the first three years and then compute the average of the first three values by dividing the total by 3. Then, the average is placed against the middle one of the three years. Now leave the first year figure and then take the average of the next three years figure and place it against the middle year of these three years. We proceed in this way, taking the average after leaving one preceding the other. We continue until we exhaust the series.

Example-1 :

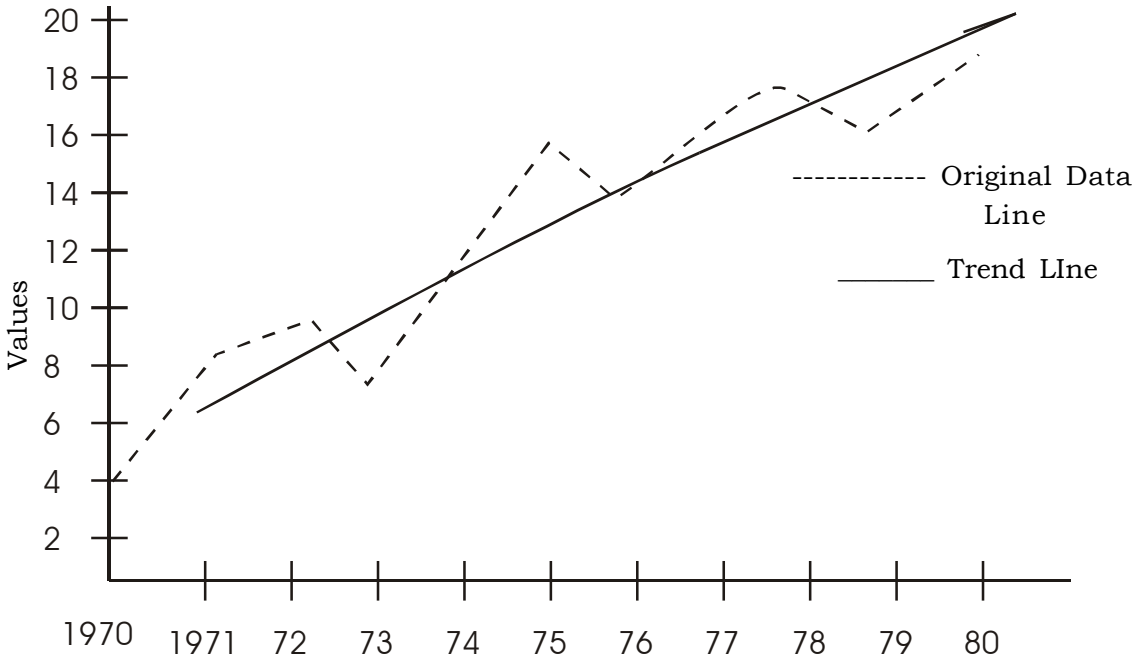
Compute a 3 yearly moving average of the following data and fit trend :

Years :	1970	1971	72	73	74	75	76	77	78	79	80
Values:	4	8	9	7	11	15	11	13	18	16	20

Solution :

Years	Values	3 yrs' M. Total	3 yr's MAv. (Trend)
1970	4		
1971	8	21	7
1972	9	24	8
1973	7	27	9
1974	11	33	11
1975	15	37	12.33
1976	11	39	13
1977	13	42	14
1978	18	47	15.67
1979	16	54	18
1980	20		

After obtaining trend values the original data as well as the moving average is plotted on a graph to determine the direction of trend.



Similarly we can calculate moving average for any odd number of years e.g., five or seven years etc.

Even Period of Moving Average :

If the moving average is to be calculated for even number of years say, four or six years, the procedure will be different. For four years' moving average, we will take the total of first four years and will calculate the average.This average will be placed in between second and third year, i.e., in the middle of four years. Leaving the first, calculate the average of next four years and place it in the middle of the four years, and so on. Then we will calculate the average of the moving averages already calculated. The process is known as 'centering' and always consists of taking a two period moving average of the moving averages.

There is another method also for centering the moving averages. Suppose we are calculating 4 years moving average, then total of 4 years shall be obtained and of these totals we will again take 2 yearly totals and divide these totals by 8. We will place the average of moving average of two groups before the middle of the moving

average. This average will be in front of a particular year of the original data.

Example 2: From the data given below calculate 4 yearly moving average.

Years :1980 81 82 83 84 85 86 87 88 89 90 91

Sale 7 8 9 11 10 12 8 6 5 10 5 9

(million tons)

Year	X	Y Sale (in million tonnes)	4 Yearly Total	4 Yearly Moving average	4 Yearly Moving average Centered
1980	0	7			
1981	1	8	35	8.75	9.1
1982	2	9	38	9.50	10.0
1983	3	11	42	10.50	10.4
1984	4	10	41	10.25	9.6
1985	5	12	36	9.00	8.4
1986	6	8	31	7.75	7.5
1987	7	6	29	7.25	6.9
1988	8	5	26	6.50	6.9
1989	9	10	29	7.25	
1990	10	5			
1991	11	9			

Example 3: Find the 4 yearly moving averages from the following data (i) by centering the averages and (i) by centering the totals :

Years	:	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
Production (in tonnes)	:	75	85	98	90	95	108	124	140	150	160

(i) Computation of the 4 yearly moving averages by centering the averages.

Year	Production	4 Yearly moving totals	4 Yearly moving average	Moving total of moving averages in twos	4 Yearly Moving average centered (col. 5÷2)
1	2	3	4	5	6
1989	75				
1990	85	348	87.00	179.00	89.50
1991	98	368	92.00	189.75	94.87
1992	90	391	97.75	202.00	101.00
1993	95	417	104.25	221.00	110.50
1994	108	467	116.75	247.25	123.63
1995	124	522	130.50	274.00	137.00
1996	140	574	143.50		
1997	150				
1998	160				

(ii) Computation of the 4 yearly Moving Averages by Centering the totals.

Year	Production	4 Yearly totals	Centering of the two adjacement totals	4 Yearly moving averages centered (col. 4÷8)
1	2	3	4	5
1989	75			
1990	85			
1991	98	348	716	89.50
1992	90	368	759	94.87
1993	95	391	808	101.00
1994	108	417	884	110.50
1995	124	467	989	123.63
1996	140	522	1096	137.00
1997	150	574		
1998	160			

Merits of Moving Average :

1. It is very simple to understand and easy to workout.
2. It is quite flexible in nature. In this method we can very easily add few more items to a series without affecting the trend values already obtained.
3. It reduces, at least, the cyclical variations to a great extent.
4. It can be used for determining the trend of the seasonal, cyclical and irregular fluctuations in the same manner in which the trend of the time series is determined.

Demerits of Moving Average :

1. The trend values for all the years cannot be computed .A number of trend values at the beginning and end of the series are not found.
2. Another shortcoming of the method is that it is difficult to determine the proper period of moving average.
3. Moving averages are calculated by using the arithmetic average. Hence, it is affected by the extreme values which is the major defect of the arithmetic average.
4. It cannot be used for forecasting.
5. Under this method, we cannot eliminate the irregular fluctuations in toto.

Despite these limitations, this method is best for computing trend in those cases where the periodicity of the data is clear cut.

2.4.5.2 Method of Least Squares :

This is the most widely used method and provides us with a mathematical device to obtain an objective fit to the trend of a given time series. This method is so

called because a trend line computed by this method is such that the sum of the squares of the deviation betwen the original data, and the corresponding computed trend values is the minimum. This method can be used to fit either a straight line trend or a parabolic trend.

Straight Line Trend : The straight line trend has an equation of the type.

$$y_c = a + b\,x$$

where y_c denotes the trend value of the dependent variable i.e. of the y series, x is the independent variable i.e. time unit of x series, a and b are constants, a denoting the value of y_c when x = O and b denoting the change in the value of y_c for a unit change in x variable.

In order to determine the value of the constants a and b, the following normal equations are to be solved.

$\Sigma y = Na + b\,\Sigma x$

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

Where N represents the number of years in the series.

By taking deviations of the time variable from its mid value or mid ponts (in case of even number of items) if the value of Σx could be made zero the values of a and b can be computed directly as under :

$$a = \frac{\sum y}{N} \quad \text{and} \quad b = \frac{\sum xy}{\sum x^2}$$

Example 4 : Fit a straight line trend to the following data and estimate the likely profit for the year 1986. Also calculate various trend values.

Year : 1977 1978 1979 1980 1981 1982 1983

profit : 60 72 75 65 80 85 95

(in lacs of Rs.)

Years (t)	y	x=t-1980	xy	x^2	Trend Values
1977	60	-3	-180	9	61.42
1978	72	-2	-144	4	66.42
1979	75	-1	-75	1	71.14
1980	65	0	0	0	76.00
1981	80	1	80	1	80.86
1982	85	2	170	4	85.72
1983	95	3	285	9	90.58
Total	532	0	136	28	

$$a = \frac{\sum y}{N} = \frac{532}{7} = 76 \text{ (n = 7, the no. of observations)}$$

and $b = \frac{136}{28} = 4.86$

Thus the fitted line of trend is $y = 76 + 4.86x$

Example 5: Fit a straight line trend by the method of least squares to the following data:

Year :	1973	1974	1975	1976	1977	1978	1979
Sales: (in Lakh of Rs.)	38	41	45	48	52	56	63

Solution :

Year	Sales (in Lakh of Rs.)	x	x^2	xy	Trend Values yc
1973	38	1	1	38	37
1974	41	2	4	82	41
1975	45	3	9	135	45
1976	48	4	16	192	49
1977	52	5	25	260	53
1978	56	6	36	336	57
1979	63	7	49	441	61
N = 7	$\Sigma y = 343$	$\Sigma x = 28$	$\Sigma x^2 = 140$	$\Sigma xy = 1484$	

$\Sigma y = Na + b \Sigma x$ or $343 = 7a+28b$ ..... (i)

$\Sigma xy = a \Sigma x + b \Sigma x^2$ or $1484 = 28a + 140b$ ..... (ii)

Multiply (i) by 4, we get

$1372 = 28a + 112\ b$ -------(i)

$1484 = 28a + 140\ b$ -------(ii)

$-112 = -28b$ or $-b = \frac{112}{28}$ $\therefore b = 4$

$343 = 7a + 28 \times 4$

$a = \frac{231}{7} = 33$

Hence the trend equation $y_c = a + bx$ will become

$y_c = 33 + 4x$

The value of y when x = 1 , $33 + 4 \times 1 = 37$

The value of y when x = 2 , $33 + 4 \times 2 = 41$

,, ,, ,, x = 3 , 33 + 4 × 3 = 45
,, ,, ,, x = 4 , 33 + 4 × 4 = 49
,, ,, ,, x = 5 , 33 + 4 × 5 = 53
,, ,, ,, x = 6 , 33 + 4 × 6 = 57
,, ,, ,, x = 7 , 33 + 4 × 7 = 61

Example 6: Given below are the data relating to the yearly sales of a retail shop :

Year :	1992	1993	1994	1995	1996	1997	1998
Sales :	120	130	135	125	145	150	140

(in '000 Rs.)

Fit a straight line by the method of least square and compute the trend values.

Solutions :

Year	Sales	Time Dvns t-1995			Trend Values $y_c = 135+3.93x$
t	y	x	xy	x^2	T
1992	120	-3	-360	9	123.21
1993	130	-2	-260	4	127.14
1994	135	-1	-135	1	131.07
1995	125	0	0	0	135.00
1996	145	1	145	1	138.93
1997	150	2	300	4	142.86
1998	140	3	420	9	146.79
Total	945	Σx= 0	Σxy = 110	28	N = 7

This indicates that 1995 is the middle year from which the deviations of the other years have been taken.

The straight line equation is given by y = a + bx.

Where $a = \frac{\sum y}{N} = \frac{945}{7} = 135$, $b = \frac{\sum xy}{\sum x^2} = \frac{110}{28} = 3.93$ approx.

$\therefore y_c = 135 + 3.93\ x$

Computation of Trend Values :

For 1992, when x = -3, $y_c = 123.21$
1993, when x = -2, $y_c = 127.14$
1994, when x = -1, $y_c = 131.07$
1995, when x = 0, $y_c = 135.00$
1996, when x = 1, $y_c = 138.93$
1997, when x = 2, $y_c = 142.86$
1998, when x = 3, $y_c = 146.79$

2.4.6 Summary

Time series analysis helps us to determine the type and nature of the variations in the data. In this lesson besides defining time series, various components of time series have been described. Moving average method and method of least square have been explained to determine the trend.

2.4.7 Further Readings

S.L. Aggarwal : Quantitative Methods
S.L. Bhardwaj
S.C. Gupta : Fundamentals of statistics
S.P. Gupta : Statistical Methods

2.4.8 List of Questions

16.8.1 Short Questions

(a) What is a time series?
(b) What is 'Secular Trend'?
(c) State different components of a time series.
(d) Explain 'Cyclical variations'.
(e) State two causes of seasonal variations

2.4.8.2 Long Questions

(a) What is a time series ? Quote some of its definitions and explain its essential characteristics.

(b) Explain, in brief the various components of a time series.

(c) Fit a straight line trend by the method of least square to the following data :

Years :	1989	90	91	92	93	94	95	96	97	98	99
Profits : (in '000 Rs.)	17	20	19	26	24	40	35	55	50	74	69

(d) Calculate the 3 yearly and 5 yearly moving averages for the following time series:

Year :	1988	89	90	91	92	93	94	95	96	97	98
Prodn.: (in quintals)	500	540	550	530	520	560	600	640	620	610	640

(e) Fit a tend line to the following data by the method of least square :

Year :	1986	1988	1990	1992	1994	1996	1998
Prodn.:	25	27	32	36	44	55	69

Also, estimate the figure of production in 1999 and find the trend values as well.

Type Setting :
Department of Distance Education, Punjabi University, Patiala.