UNIT 5 CORRELATION AND REGRESSION*

Structure

5.0 OBJECTIVES

After going through this unit you will be in a position to

- plot scatter diagram;
- compute correlation coefficient and state its properties;
- compute rank correlation;
- explain the concept of regression;
- explain the method of least squares;
- identify the limitations of linear regression;
- apply linear regression models to given data; and
- use the regression equation for prediction.

5.1 INTRODUCTION

The word 'bivariate' is used to describe situations in which two character are measured on each individual or item, the character being represented by two variables. For example, the measurement of height (X_i) and weight (Y_i) of students in a school. The subscript i in this case represents the student concerned.

* Prof. Kaustuva Barik, School of Social Sciences, Indira Gandhi National Open University.

Thus, for example, X_5, Y_5 represent the height and weight of the fifth student. Statistical data relating to simultaneous measurement of two variables are called bivariate data. The observation on each individual are paired, one for each variable (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n).

In statistical studies with several variables, there are generally two types of problems. In some problems it is of interest to study how the variables are interrelated; such problems are tackled by using correlation technique. For instance, an economist may be interested in studying the relationship between the stock prices of various companies; for this he may use correlation techniques. In other problems there is a variable y of basic interest and the problem is to find out what information the other variable provides on Y, such problems are tackled using regression techniques. For instance, an economist may be interested in studying what factors determine the pay of an employed person and in particular, he may be interested in exploring what role the factors such as education, experience, market demand, etc. play in determining the pay. In the above situation he may use regression techniques to set up a prediction formula for pay based on education, experience, etc.

5.2 SCATTER DIAGRAM

We first illustrate how the relationship between two variables is studied. A teacher is interested in studying the relationship between the performance in Statistics and Economics of a class of 20 students. For this he compiles the scores on these subjects of the students the last semester examination. Some data of this type are presented in Table 5.1.

Table 5.1: Scores of 20 Students in Statistics and Economics

Serial Number	Score in Statistics	Score in Economics	Serial Number	Score in Statistics	Score in Economics
1	82	64	11	76	58
2	70	40	12	76	66
3	34	35	13	92	72
4	80	48	14	72	46
5	66	54	15	64	44
6	84	56	16	86	76
7	74	62	17	84	52
8	84	66	18	60	40
9	60	58	19	82	60
10	86	82	20	90	60

A representation of data of this type on a graph is a useful device which will help us to understand the nature and form of the relationship between the two variables, whether there is a discernible relationship or not and if so whether it is linear of not. For this let us denote score in Economics by X and the score in Statistics by Y and plot the data of Table 5.1 on the x-y plane. It does not matter which is called X and which Y for this purpose. Such a plot is called *Scatter Plot* or *Scatter Diagram*. For data of Table 5.1 the scatter diagram is given in Fig. 5.1.

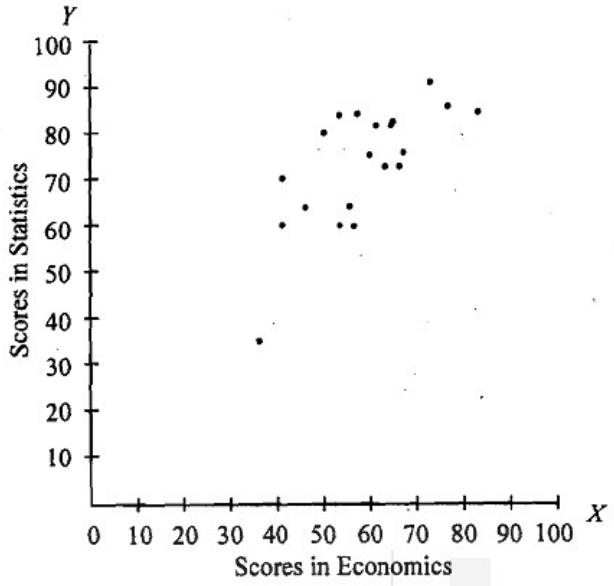

Fig. 5.1: Scatter Diagram of Scores in Statistics and Economics.

An inspection of Table 5.1 and Fig. 5.1 shows that there is a *positive relationship* between *x* and *y*. This means that larger values of *x* associated with larger values of *y* and smaller values of *y*. Further, the points seem to lie scattered around both sides of a straight line. Thus, it appears that a linear relationship exists between *x* and *y*. This relationship, however, in not *perfect* in the sense that there are deviations from such a relationship in the case of certain observations. It would indeed be useful to get a measure of the strength of this linear relationship.

5.3 COVARIANCE

In the case of a single variable we have learnt the concept of variance, which is defined as

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \ldots (5.1)$$

In the above we use a subscript *x* to specify that σ_2^2 represents the variance in *x*. In a similar manner we can represent σ_y^2 as the variance in *y* and σ_x and σ_y as the standard deviation in *x* and *y* respectively.

As you know, variance measures the dispersion from mean. In the case of bivariate data we have to reach a single figure which will present the deviation in both the variables from their respective means. For this purpose we use a concept termed covariance, which is defined as follows:

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \quad \ldots (5.2)$$

You may recall that standard deviation is always positive since it is defined as the positive square root of variance. In the case of covariance there are two terms $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ which represent the deviations in *x* from $\bar{X}$ and *Y* from $\bar{Y}$.

Moreover, $(X_i - \overline{X})$ can be positive or negative depending on whether x_i is less than or greater than $\overline{X}$. Similarly $(Y_i - \overline{Y})$ can be positive or negative. It is not necessary that whenever $(X_i - \overline{X})$ is positive $(Y_i - \overline{Y})$ will also be positive. Therefore, the product $(X_i - \overline{X})$ $(Y_i - \overline{Y})$ can be either positive or negative. A positive value for $(X_i - \overline{X})$ $(Y_i - \overline{Y})$ implies the whenever $X_i > \overline{X}$, we have $Y_i > \overline{Y}$. Thus a higher value of x_i is associated with a relatively higher value in y_i. On the other hand, $(X_i - \overline{X})(Y_i - \overline{Y}) < 0$ implies that a lower value in X_i is associated with a relatively higher value in y_i. when we sum it over all the observations and ivied by the number of observations, we may obtain a negative or positive value. Therefore, covariance can assume both positive and negative values.

When covariance between x and y is negative $(\sigma_{xy} < 0)$ we can say that the relationship could be inverse. Similarly, $(\sigma_{xy} < 0)$ implies a positive relationship between x and y. A major limitation of covariance is that it is not independent of unit of measurement. It means that if we change the unit of measurement of the variables we will get a difference value for σ_{xy}.

The computation of σ_{xy} as given in (5.2) often involves large numbers. Therefore, it is derived further as

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}(X_iY_i - \overline{X}Y_i - \overline{XY})$$

By further simplification we find that

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}X_iY_i - \frac{1}{n}\sum_{i=1}^{n}\overline{X}Y_i - \frac{1}{n}\sum_{i=1}^{n}X_i\overline{Y} + \frac{1}{n}\sum_{i=1}^{n}\overline{XY}$$

Since $\frac{1}{n}\sum_{i=1}^{n}\overline{X}Y_i = \frac{1}{n}X_i\overline{Y} = \overline{X}\overline{Y}$ we have

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}X_iY_i - \overline{X}\overline{Y} \qquad \ldots (5.3)$$

5.4 CORRELATION COEFFICIENT

The task before us is to measure the linear relationship between x and y. It is desirable to have this measure of strength of linear relationship independent of the scale chosen for measuring the variables. For instance, if we are measuring the relationship between height and weight, we should get the same measure whether height is measured in inches or centimetres and weight in pounds or kilograms. Similarly, if a variable is temperature, it should not matter whether it is recorded in Celsius or Fahrenheit.

This can be achieved by standardizing each variable, that is by considering $\frac{X-\overline{X}}{\sigma_x}$ and $\frac{Y-\overline{Y}}{\sigma_y}$ where $\overline{X}$ and $\overline{Y}$ are the means of X and Y respectively and σ_x and σ_y are standard deviations.

Let us denote these standardised variables by u and y respectively. Let us also use the notation (X_i, Y_i) to denote the score i[th] student in Economics and Statistics respectively, i ranging from 1 to n, the number of students, n being 20 in our example. Similarly, let (u_i, v_i) denote the standardised scores of i[th] student. Then recall the following formulae for mean and standard deviation:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n}\overline{X}_i; \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(\overline{X}_i - \overline{X})^2;$$

$$Y = \frac{1}{n}\sum_{i=1}^{n}X_i; \sigma_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

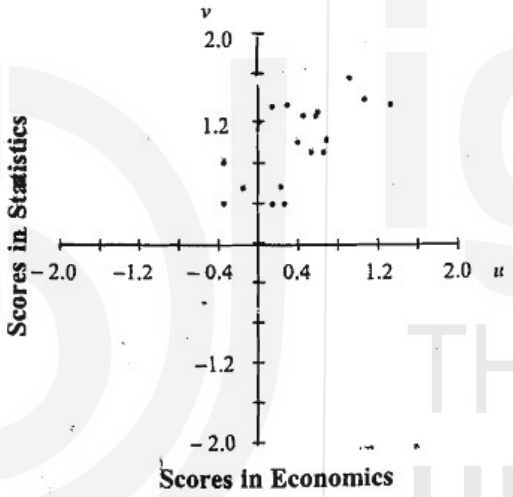

Fig. 5.2: Scatter Diagram of Standardised Scores in Statistics and Economics

Fig. 5.2 is the scatter diagram in terms of standardised variables u and v. Let us observe that in this example there is a positive association between the two scores. The larger one score is, the larger the other score also is; the smaller one score is the smaller the other score is, on the whole. In view of this, most of the points are either in the *first quadrant* or in the *third quadrant.* The first quadrant represents the cases where both scores are above their respective means and third quadrant represents the cases where both scores are below their respective means. There are only a very few points in second and fourth quadrants, which represent the cases where one score is above its mean and the other is below its mean. Thus the product of the u, v values is a suitable indicator of the strength of the relationship; this product is positive in the first and third quadrants and negative in the second and fourth. Thus the product of u, v averaged over all the points may be considered to be suitable measure of the strength of linear relationship between X and Y.

This measure is called the *correlation coefficient* between X and Y and is usually denoted by r_{xy} or simply by r, when it is clear what x and y in the context are. This is also called the *Pearson's Product-Moment Correlation Coefficient* to distinguish it from other types of correlation coefficients.

Thus the formula for r is

$$r = \frac{1}{n}\sum_{i=1}^{n} u_i v_i \qquad \ldots (5.4)$$

If we substitute the variables x and y in (5.4) above

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma_x}\right)\left(\frac{Y_i - \bar{Y}}{\sigma_y}\right) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sigma_x \sigma_y}$$

In the above expression, the term

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

is the *covariance* between x and y (σ_{xy}).

Thus, the formula for correlation coefficient is

$$r = \frac{\sigma_{xy}}{\sigma_x \times \sigma_y} \qquad \ldots (5.5)$$

Incorporating the formulae for $\bar{x}, \bar{y}, \sigma_x, \sigma_y$ it becomes

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \ldots (5.6)$$

Or, alternatively

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\left[\sqrt{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}\right]\left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]} \qquad \ldots (5.7)$$

Let us go back to the data given in Table 5.1 and work out the value of r. You can use any of the formulae (5.4), (5.5) or (5.7) to get the value of r. Since all the formulae are derived from the same concept we obtain the same value for r whichever formulae we use. For the data set in Table 5.1 we have calculated it by using (5.4) and (5.7). We construct Table 5.2 for this purpose.

Table 5.2: Calculation of Correlation Coefficient

Observation No.	X	Y	X^2	Y^2	XY
1	82	64	6724	4096	5248
2	70	40	4900	1600	2800
3	34	35	1156	1225	1190
4	80	48	6400	2304	3840
5	66	54	4356	2916	3564
6	84	56	7056	3136	4704
7	74	62	5476	3844	4588
8	84	66	7056	4356	5544
9	60	52	3600	2704	3120
10	86	82	7396	6724	7052
11	76	58	5776	3364	4408
12	76	66	5776	4356	5016
13	92	72	8464	5184	6624
14	72	46	5184	2116	3312
15	64	44	4096	1936	2816
16	86	76	7396	5776	6536
17	84	52	7056	2704	4368
18	60	40	3600	1600	2400
19	82	60	6724	3600	4920
20	90	60	8100	3600	5400
Total	1502	1133	116292	67141	87450

From Table 5.2 we note that

$$\sum_{i=1}^{20} X_i = 150; \overline{X} = 75.1;$$

$$\sum_{i=1}^{20} Y_i = 1133; \overline{Y} = 56.65;$$

$$\sum_{i=1}^{20} X_i^2 = 116292; \sigma_x^2 = \frac{1}{20}\left[116292 - \frac{1502^2}{20}\right] = 174.59; \sigma_x = 13.21;$$

$$\sum_{i=1}^{20} Y_i^2 = 67141; \sigma_x^2 = \frac{1}{20}\left[67141 - \frac{1133^2}{20}\right] = 147.83; \sigma_y = 12.16;$$

$$\sum X_i Y_i = 87450; \sigma_{xy} = \frac{1}{20}\left[87450 - \frac{1502 \times 1133}{20}\right] = 118.09$$

Thus, using formula given at (5.4), we have

$$r = \frac{118.09}{13.21 \times 12.16} = 0.735$$

Now let us use the formula 5.7. We have

$$r = \frac{20 \times 87450}{\sqrt{(20 \times 116292 - 1502^2)(20 \times 67141 - 1133^2)}} = 0.735$$

Thus we see that both the formulae provide the same value of the correlation coefficient *r.* You can check yourself that the same value of *r* is obtained by using the formula (5.5). For this purpose you will need values on

$\sum(X_i - \bar{X})^2, \sum(Y_i - \bar{Y})^2$ and $\sum(X_i - \bar{X})(Y_i - \bar{Y})$

Hence you can have five columns on

$(X_i - \bar{X}), (Y_i - \bar{Y}), (X_i - \bar{X})^2, (Y_i - \bar{Y})^2$ and $(X_i - \bar{X})(Y_i - \bar{Y})$ in a table and find the totals.

5.5 INTERPRETATION OF CORRELATION COEFFICIENT

It is a mathematical fact that the value of *r* as defined above lies between –1and + 1. The extreme values of –1and + 1are obtained only in situations where there is a *perfect linear relationship* between *X* and *Y.* The –1is obtained when this relationship is perfectly negative (i.e., inverse) and +1 when this is perfect positive (i.e., direct). The value of 0 is obtained when there is no linear relationship between *x* and *y*.

We can make some guess work about the sign and degree of the correlation coefficient from the scatter diagram. Fig. 5.3 gives example of scatter diagrams for various values of *r.* Fig. 5.3 (a) is a scatter diagram for the case $r = 0$; here there is no *linear* relationship between *x* and *y.* Fig. 5.3(b) is also and example of scatter diagram for the case $r = 0$; here there is discernible relationship between *X* and *Y* but it is not of the linear type. Here, initially, *Y* increases with *X* but later *Y* decreases as *X* increases resulting in a definitive quadratic relationship. But the correlation coefficient in the case is zero. Thus the correlation coefficient is only a measure of linear relationship. This sort of scatter diagram is obtained, if we plot, for instance, body weight (*Y)* of individuals against there are (*X*). Fig. 5.3.(c) is an example of a scatter diagram where there is a perfect positive linear relationship between *X* and *Y*. We get this sort of scatter diagram if we plot, for instance, height of individuals in inches (*X*) against their heights in centimetres (*Y*); in that case $Y = 2.54X$, which is a deterministic and perfect linear relationship. Figures 5.3(d) to 5.3(k) are scatter diagrams for other values of *r*. Form these scatter diagrams we get an idea of the nature of relationship and associated values of *r*.

From these is would seem that a value of 0.81 indicates a fair degree of linear relationship between scores in Statistics and Economics of these candidates. Such a quantification of relationship or association between variables is helpful for natural and social scientists to understand the phenomena they are investigating and explore these phenomena further. In an example of this sort, an educational psychologist may compute correlation coefficients between scores in various subjects and by further statistical analysis of the correlation coefficients and using psychological techniques may be able to form a theory as to what mental and other faculties are involved in making students good in various disciplines.

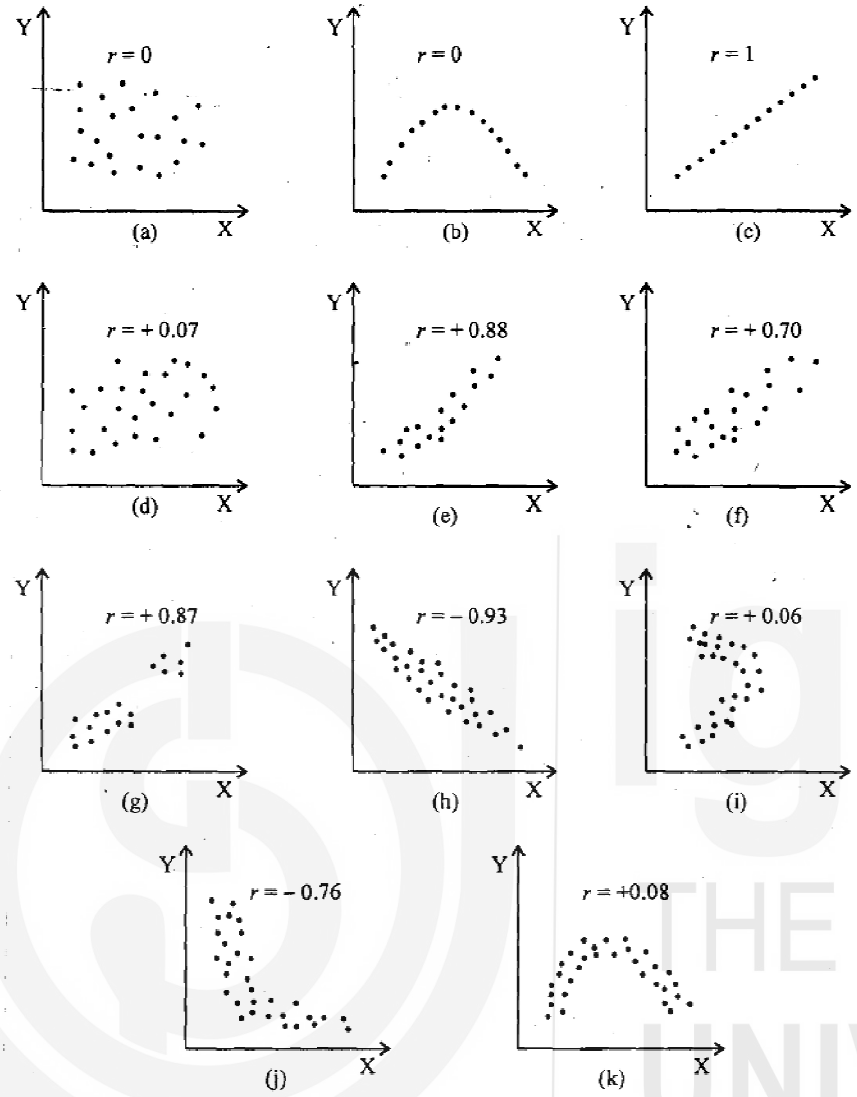

Fig. 5.3: Scatter Plots for Various Values of Correlation Coefficient

You should remember that

- Correlation coefficient shows the *linear relationship* between X and Y. Thus, even if there is a strong non-linear relationship between X and Y, correlation coefficient may be low.
- Correlation coefficient is independent of scale and origin. If we subtract some constant from one (or both) of the variables, correlation coefficient will remain unchanged from one (or both) of the variables by some constant, correlation coefficient will not change.
- Correlation coefficient varies between –1 and +1. It means that r cannot be smaller than –1 and cannot be greater than +1.

The existence of a linear relationship between two variables is not to be interpreted to mean a cause-effect relationship between the two.

For instance, if you work out the correlation between family expenditure on petrol and chocolates, you may find it to be fairly high indicating a fair degree of linear relationship. However, both of these are luxury items and richer families can afford them while poorer ones cannot. Thus the high correlation here is caused by the high correlation of each of the variables with family income. To consider another example, suppose for each of the last twenty years, you work out the average height of an Indian and the average time per week an Indian watches television; you are likely to find a positive correlation. This does not, however, imply that watching television increases one's height or that taller people tend to watch television longer. Both these variables have an increasing trend over time and this is reflected in the high correlation. This kind or correlation between two variables is caused by the effect of a third variable on each of them rather than a direct linear cause-effect of a third variable on each of them rather than a direct linear cause-effect relationship between them is called *spurious correlation.*

Another aspect of the computation of correlation coefficient that we should be aware of is that the correlation coefficient like any other quantity computed from sample, varies from sample to sample and these sample fluctuations should be taken into account in making use of the computed coefficient. We do not discuss these techniques here.

Whether the presence of a linear relationship between two variables and hence a high correlation between them is genuine or spurious, such a situation is helpful to *predict* one variable from the other.

Check Your Progress 1

1) Calculate r from the following given results:

   $n = 10; \sum X = 125; \sum X^2 = 1585; \sum Y = 80; \sum Y^2 = 650; \sum XY = 1007.$

   ……..

   ……..

   ……..

   ……..

   ……..

   ……..

2) Calculate the coefficient of correlation for the ages of husband and wife:

Age of husband	:	23	27	28	29	30	31	33	35	36	39
Age of wife	:	18	22	23	24	25	26	28	29	30	32

   ……..

   ……..

   ……..

   ……..

   ……..

   ……..

3) Specimens of similarly treated alloy steel containing various percentages of nickel are tested for toughness with the following results:

Toughness (arbitrary units)	47	50	52	52	54	56	58	59	60	60	62	64	65	66
Percentage of Nickel	2.7	2.7	2.8	2.8	2.9	3.2	3.2	3.3	3.4	3.5	3.6	3.7	3.7	3.8

Find the correlation coefficient between toughness and nickel content and comment on the result.

……..

……..

……..

……..

……..

……..

4) Determine the correlation coefficient between x and y.

x	:	5	7	9	11	13	15
y	:	1.7	2.4	2.8	3.4	3.7	4.4

……..

……..

……..

……..

……..

……..

5) The following table gives the saving bank deposits in billions of dollars and strikes and lock-outs, in thousands, over a number of years. Compute the correlation coefficient and comment on the result.

Saving deposits	:	5.1	5.4	5.5	5.9	6.4	6.0	7.2
Strikes and lock-outs	:	3.8	4.4	3.3	3.6	3.3	2.3	1.0

……..

……..

……..

……..

……..

……..

……..

5.6 RANK CORRELATION COEFFICIENT

The Pearson's product moment correlation coefficient (or simply, the correlation coefficient) described above is suitable if both the variables involved are measureable (numerical) and the relationship between the variables is linear. However, there are situations where variables are not numerical but various items can be ranked according to the characteristics (i.e., ordinal). Sometimes even when the original variables are measurable, they are converted into ranks and a measure of association is computed. Consider for instance the situation when two examiners are asked to judge ten candidates on the basis of an oral examination. In this case, it may be difficult to assign scores to candidates, but the examiners find it reasonably easy to rank the candidates in order of merit. Before using the resulted it may be advisable to find out if rankings are in reasonable concordance. For this, a measure of association between the ranks assigned by the two examiners may by computed. The Karl Pearson's correlation coefficient is not suitable in this situation. One may use the following called *Spearman's Rank Correlation Coefficient* for this purpose.

Table 5.3: Ranks of 10 Candidates by two Examiners

S. No	Rank given by		Difference	
	Examiner I	*Examiner II*	D_i	D_i^2
1	6.0	6.5	–0.5	0.25
2	2.0	3.0	–1.0	1.00
3	8.5	6.5	2.0	4.00
4	1.0	1.0	0.0	0.00
5	10.0	2.0	8.0	64.00
6	3.0	4.0	–1.0	1.00
7	8.5	9.5	–1.0	1.00
8	4.0	5.0	–1.0	1.00
9	5.0	8.0	–3.0	9.00
10	7.0	9.5	–2.5	6.25
		$\sum D_i = 0$	$\sum D_1^2 = 87.50$	

Let us consider the data of Table 5.3. Here there are some ties; the tied cases are given the same rank in such a way their total is the same as when there is no tie. For example, when there are two cases with rank 6, each is given a rank of 6.5 and there is no case with rank either 6 or 7. Similarly, if there are three cases with rank 5, then each is given a rank of 6 and there is no case with rank 5 or 7. Spearman's rank correlation coefficient, called Spearman's Rho, denoted by ρ, is based on the difference D_i (i for i^{th} observation) between the two rankings. If the two rankings completely coincide, then D_i is zero for every case. The larger the value of D_i, the greater is the difference between the two rankings and smaller is the association. Thus, the association can be measured by considering the magnitudes of D_i. Since the sum of D_i is always zero, to find a single index on the basis of D_i values, we should remove the sign of D_i and consider only the magnitude. In Spearman's ρ, this is done by taking D_i^2.

However, the largeness or smallness of $\sum_{i=1}^{n} D_i^2$, where n is the number of cases, will depend on n. thus, in order to be able to interpret this value, we could create a ratio by dividing this sum by the largest possible value, which depends only on n, which is $\frac{n(n^2-1)}{6}$. However, $\frac{6\times\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$ is zero for perfect association and 2 for lack of association, i.e., perfect negative association, while we would like it to be other way around. So we subtract this ratio from 1. Thus

$$\rho = 1 - \frac{6\times\sum_{i=1}^{n} D_i^2}{n(n^2-1)} \qquad \ldots (5.8)$$

is defined as Spearman's rank correlation.

Let us calculated the value of ρ from the data given in Table 5.3.

$$\rho = -\frac{6\times 87.5}{10(10^2-1)} = 1 - \frac{525}{990} = 1 - 0.53 = 0.47.$$

Like Karl Pearson's coefficient of correlation the Spearman's rank correlation has a value +1 for perfect matching of ranks, –1for perfect mismatching of ranks and 0 for the lack of relation between the ranks.

There are other measures of association suitable for use when the variables are of nominal, ordinal and other types. We do not discuss them here.

Check Your Progress 2

1) In a contest, two judges ranked eight candidates A, B, C, D, E, F, G and H in order of their preference, as shown in the following table. Find the rank correlation coefficient.

	A	B	C	D	E	F	G	H
First Judge	5	2	8	1	4	6	3	7
Second Judge	4	5	7	3	2	8	1	6

…….

…….

……..…

……..…

……..…

……..…

……..

2) Compute the correlation coefficient of the following ranks of a group of students in two examinations. What conclusion do you draw from the results?

Roll Nos.	1	2	3	4	5	6	7	8	9	10
Rank in B. Com. Exam.	1	5	8	6	7	4	2	3	9	10
Rank in M. Com Exam.	2	1	5	7	6	3	4	8	10	9

…….....

…….....…….

…….....…..…

…….....…..…

…….....…..…

…….....…..…

3) Ten competitors in a musical contest were ranked by 3 judges A, B and C in the following order:

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

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

…….....…….....

…….....……..

…….....…..…

…….....…..…

…….....…..…

4) Ten students obtained the following marks in Mathematics and Statistics. Calculate the rank correlation coefficient.

Student (Roll No.)	1	2	3	4	5	6	7	8	9	10
Marks in Mathematics	78	36	98	25	75	82	90	62	65	39
Marks in Statistics	84	51	91	60	68	62	86	58	53	47

…….....

…….....….

…….....…

…….....…

…….....…

5.7 THE CONCEPT OF REGRESSION

In the previous section we noted that correlation coefficient does not reflect cause and effect relationship between two variables. Thus we cannot predict the value of one variable for a given value of the other variable. This limitation is removed by regression analysis. In regression analysis, the relationship between variables are expressed in the form of a mathematical equation. It is assumed that one variable is the cause and the other is the effect. You should remember that regression is a statistical tool which helps understand the relationship between variables and predicts the unknown values of the dependent variable from known values of the independent variable.

In regression analysis we have two types of variables: i) dependent (or explained) variable, and ii) independent (or explanatory) variable. As the name (explained and explanatory) suggests the dependent variable is explained by the independent variable.

In the simplest case of regression analysis there is one dependent variable and one independent variable. Let us assume that consumption expenditure of a household is related to the household income. For example, it can be postulated that as household income increases, expenditure also increases. Here consumption expenditure is the dependent variable and household income is the independent variable.

Usually we denote the dependent variable as Y and the independent variable as X. Suppose we took up a household survey and collected *n* pairs of observations in X and Y. The next step is to find out the nature of relationship between X and Y.

The relationship between X and Y can take many forms. The general practice is to express the relationship in terms of some mathematical equation. The simplest of these equations is the linear equation. This means that the relationship between X and Y is in the form of a straight line and is termed linear regression. When the equation represents curves (not a straight line) the regression is called non-linear or curvilinear.

Now the question arises, 'How do we identify the equation form?' There is no hard and fast rule as such. The form of the equation depends upon the reasoning and assumptions made by us. However, we may plot the X and Y variables on a graph paper to prepare a scatter diagram. From the scatter diagram, the location of the points on the graph paper helps in identifying the type of equation to be fitted. If the points are more or less in a straight line, then linear equation is assumed. On the other hand, if the points are not in a straight line and are in the form of a curve, a suitable non-linear equation (which resembles the scatter) is assumed.

We have to take another decision, that is, the identification of dependent and independent variables. This again depends on the logic put forth and purpose of analysis: whether 'Y depends on X' or 'X depends on Y'. Thus there can be two regression equations from the same set of data. These are i) Y is assumed to be

dependent on X (this is termed 'Y on X' line), and ii) X is assumed to be dependent on Y (this is termed 'X on Y' line).

Regression analysis can be extended to cases where one dependent variable is explained by a number of independent variables. Such a case is termed multiple regression. In advanced regression models there can be a number of both dependent as well as independent variables.

You may by now be wondering why the term 'regression', which means 'reduce'. This name is associated with a phenomenon that was observed in a study on the relationship between the stature of father (x) and son (y). It was observed that the average stature of sons of the tallest fathers has a tendency to be less than the average stature of these fathers. On the other hand, the average stature of sons of the shortest fathers has a tendency to be more than the average stature of these fathers. This phenomenon was called *regression towards the mean*. Although this appeared somewhat strange at that time, it was found later that this is due to natural variation within subgroups of a group and the same phenomenon occurred in most problems and data sets. The explanation is that many tall men come from families with average stature due to vagaries of natural variation and they produce sons who are shorter than them on the whole. A similar phenomenon takes place at the lower end of the scale.

5.8 LINEAR RELATIONSHIP: TWO-VARIABLES CASE

The simplest relationship between X and Y could perhaps be a linear *deterministic* function given by

$$Y_i = a + bX_i \qquad \ldots(5.9)$$

In the above equation X is the independent variable or explanatory variable and Y is the dependent variable or explained variable. You may recall that the subscript i represents the observation number, i ranges from 1 to n. Thus Y_1 is the first observation of the dependent variable, X_5 is the fifth observation of the independent variable, and so on.

Equation (5.9) implies that Y is completely determined by X and the parameters a and b. Suppose we have parameter values $a = 3$ and $b = 0.75$, then our linear equation is $Y = 3 + 0.75\ X$. From this equation we can find out the value of Y for given values of X. For example, when $X = 8$, we find that $Y = 9$. Thus if we have different values of X then we obtain corresponding Y values on the basis of (5.9). Again, if X_i is the same for two observations, then the value of Y_i will also be identical for both the observations. A plot of Y on X will show no deviation from the straight line with intercept 'a' and slope 'b'.

If we look into the deterministic model given by (5.9) we find that it may not be appropriate for describing economic interrelationship between variables. For example, let Y = consumption and X = income of households. Suppose you record your income and consumption for successive months.

For the months when your income is the same, do your consumption remain the same? The point we are trying to make is that economic relationship involves certain randomness.

Therefore, we assume the relationship between Y and X to be *stochastic* and add one error term in (5.9). Thus our stochastic model is

$$Y_i = a + bX_i + e_i \qquad \ldots(5.10)$$

where e_i is the error term. In real life situations e_i represents randomness in human behaviour and excluded variables, if any, in the model. Remember that the right hand side of (5.10) has two parts, viz., i) deterministic part (that is, $a + bX_i$), and ii) stochastic or randomness part (that is, e_i). Equation (5.10) implies that even if X_i remains the same for two observations, Y_i need not be the same because of different e_i. Thus, if we plot (5.10) on a graph paper the observations will not remain on a straight line.

Example 5.1

The amount of rainfall and agricultural production for ten years are given in Table 5.4.

Table 5.4: Rainfall and Agricultural Production

Rainfall (in mm.)	Agricultural production (in tonne)
60	33
62	37
65	38
71	42
73	42
75	45
81	49
85	52
88	55
90	57

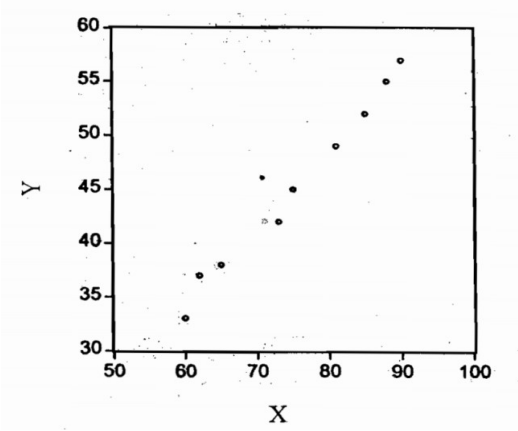

Fig. 5.4: Scatter Diagram

We plot the data on a graph paper. The scatter diagram looks something like Fig. 5.4. We observe from Fig. 5.4 that the points do not lie strictly on a straight line. But they show an upward rising tendency where a straight line can be fitted. Let us draw the regression line along with the scatter plot.

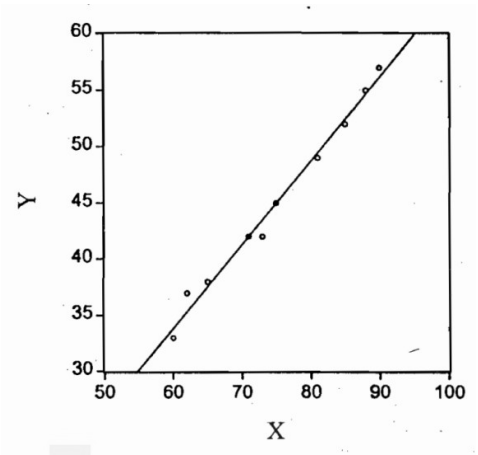

Fig. 5.5: Regression Line

The vertical difference between the regression line and the observations is the error e_i. The value corresponding to the regression line is called the predicted value or the expected value. On the other hand, the actual value of the dependent variable corresponding to a particular value of the independent variable is called the observed value. Thus 'error' is the difference between predicted value and observed value.

A question that arises is, 'How do we obtain the regression line? The procedure of fitting a straight line to the data is explained below.

5.9 MINIMISATION OF ERRORS

As mentioned earlier, a straight line can be represented by

$$Y_i = a + bX_i$$

where *b* is the *slope* and *a* is the *intercept* on y-axis. The location of a straight line depends on the value of *a* and *b*, called *parameters*. Therefore, the task before us is to *estimate* these parameters from the collected data. (You will learn more about the concept of estimation in Block 4). In order to obtain the line of best fit to the data we should find estimates of *a* and *b* in such a way that the error e_i is minimum.

In Fig. 5.4 these differences between observed and predicted values of Y are marked with straight lines from the observed points, parallel to y-axis, meeting the regression line. The lengths of these segments are the errors at the observed points.

Let us denote the *n* observations as before by (X_i, Y_i), $i = 1, 2, ..., n$. In Example 5.1 on agricultural production and rainfall, n=10.

Let us denote the predicted value of Y_i at X_i by $\hat{Y}_i$ (the notation $\hat{Y}_i$ is pronounced as 'Y_i-cap' or 'Y_i-hat'). Thus

$\hat{Y}_i = a + bX_i$, i = 1, 2, ..., n.

The error at the i^{th} point will then be

$$e_i = Y_i - \hat{Y}_i \qquad \ldots\ldots(5.11)$$

It would be nice if we can determine a and b in such a way that each of the e_i, i = 1, 2, ..., n is zero. But this is impossible unless it so happens that all the n points lie on a straight line, which is very unlikely. Thus we have to be content with minimising a combination of e_i, i = 1, 2, ..., n. What are the options before us?

- It is tempting to think that the total of all the e_i, i = 1, 2, ….., n, that is, $\sum_{i=1}^{n} e_i$ is a suitable choice. But it is not. Because, e_i for points above the line are positive and below the line are negative. Thus by having a combination of large positive and large negative errors, it is possible for $\sum_{i=1}^{n} e_i$ to be very small.
- A second possibility is that if we take $a = \bar{y}$ (the arithmetic mean of the Y_i's) and b = 0, $\sum_{i=1}^{n} e_i$ could be made zero. In this case, however, we do not need the value of X at all for prediction! The predicted value is the same irrespective of the observed value of X. This evidently is wrong.
- What then is wrong with the criterion $\sum_{i=1}^{n} e_i$? It takes into account the sign of e_i. What matters is the magnitude of the error and whether the error is on the positive side or negative side is really immaterial. Thus, the criterion $\sum_{i=1}^{n} |e_i|$ is a suitable criterion to minimise. Remember that $|e_i|$ means the absolute value of e_i. Thus, if e_i = 5 then $|e_i|$ = 5 and also if e_i = –5 then $|e_i|$=5. However, this option poses some computational problems.
- For theoretical and computational reasons, the criterion of *least squares* is preferred to the absolute value criterion. While in the absolute value criterion the sign of e_i is removed by taking its absolute value, in the *least squares criterion* it is done by squaring it. Remember that the squares of both 5 and –5 are 25. This device has been found to be mathematically and computationally more attractive.

We explain in detail the least squares method in the following section.

5.10 METHOD OF LEAST SQUARES

In the least squares method we minimise the sum of squares of the error terms, that is, $\sum_{i=1}^{n} e_i^2$.

From (5.9) we find that $e_i = Y_i - \hat{Y}_i$

which implies $e_i = Y_i - (a + bX_i) = Y_i - a - bX_i$.

Hence, $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$ …(5.12)

The next question is: How do we obtain the values of *a* and *b* to minimise (5.12)?

- Those of you who are familiar with the concept of differentiation will remember that the value of a function is minimum when the first derivative of the function is zero and second derivative is positive. Here we have to choose the value of *a* and *b*. Hence, $\sum_{i=1}^{n} e_i^2$ will be minimum when its partial derivatives with respect to *a* and *b* are zero. The partial derivatives of $\sum_{i=1}^{n} e_i^2$ are obtained as follows:

$$\frac{\partial \sum_i e_i^2}{\partial a} = \frac{\partial \sum_i (Y_i - a - bX_i)^2}{\partial a} = 2 \cdot (-1) \cdot \sum_i (Y_i - a - bX_i) \quad \ldots(5.13)$$

$$\frac{\partial \sum_i e_i^2}{\partial b} = \frac{\partial \sum_i (Y_i - a - bX_i)^2}{\partial b} = 2 \cdot (-X_i) \cdot \sum_i (Y_i - a - bX_i) \quad \ldots(5.14)$$

  By equating (5.13) and (5.14) to zero and re-arranging the terms we get the following two equations:

$$\sum_{i=1}^{n} Y_i = na + b\sum_{i=1}^{n} X_i \quad \ldots(5.15)$$

$$\sum_{i=1}^{n} X_i Y_i = a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 \quad \ldots(5.16)$$

  These two equations, (5.15) and (5.16), are called the *normal equations* of least squares. These are two simultaneous linear equations in two unknowns. These can be solved to obtain the values of *a* and *b*.

- Those of you who are not familiar with the concept of differentiation can use a rule of thumb (We suggest that you should learn the concept of differentiation, which is so much useful in Economics). We can say that the normal equations given at (5.15) and (5.16) are derived by multiplying the coefficients of *a* and *b* to the linear equation and summing over all observations. Here the linear equation is $Y_i = a + bX_i$. The first normal equation is simply the linear equation $Y_i = a + bX_i$ summed over all observations (since the coefficient of *a* is 1).

  $\sum Y_i = \sum a + \sum bX_i$ or $\sum Y_i = na + b\sum X_i$

The second normal equation is the linear equation multiplied by X_i (since the coefficient of b is X_i)

$$\sum X_i Y_i = \sum aX_i + \sum bX_i^2 \quad \text{or} \quad \sum X_i Y_i = a\sum X_i + b\sum X_i^2$$

After obtaining the normal equations we calculate the values of a and b from the set of data we have.

Example 5.2: Assume that quantity of agricultural production depends on the amount of rainfall and fit a linear regression to the data given in Example 5.1.

In this case dependent variable (Y) is quantity of agricultural production and independent variable (X) is amount of rainfall. The regression equation to be fitted is

$$Y_i = a + bX_i + e_i$$

For the above equation we find out the normal equations by the method of least squares. These equations are given at (5.15) and (5.16). Next we construct a table as follows:

Table 5.5: Computation of Regression Line

X_i	Y_i	X_i^2	X_iY_i	$\hat{Y}_i$	e_i
60	33	3600	1980	33.85	-0.85
62	37	3844	2294	35.34	1.66
65	38	4225	2470	37.57	0.43
71	42	5041	2982	42.03	-0.03
73	42	5329	3066	43.51	-1.51
75	45	5625	3375	45.00	0.00
81	49	6561	3969	49.46	-0.46
85	52	7225	4420	52.43	-0.43
88	55	7744	4840	54.66	0.34
90	57	8100	5130	56.15	0.85
Total $\sum_i X_i$ = 750	$\sum_i Y_i$ =450	$\sum_i X_i^2$ =57294	$\sum_i X_iY_i$ =34526	$\sum_i \hat{Y}_i$ =450	$\sum_i e_i$ =0

By substituting values from Table 5.5 in the normal equations (5.15) and (5.16) we get the following:

$$450 = 10a + 750b$$

$$34526 = 750a + 57294b$$

By solving these two equations we obtain $a = -10.73$ and $b = 0.743$.

So the regression line is $\hat{Y}_i = -10.73 + 0.743X_i$.

Notice that the sum of errors $\sum_i e_i$ for the estimated regression equation in zero (see the last column of Table 5.5).

The computation given in Table 5.5 often involves large numbers and poses difficulty. Hence we have a short-cut method for calculating the values of *a* and *b* from the normal equations.

Let us take

$x = X - \bar{X}$ and $y = Y - \bar{Y}$ where $\bar{X}$ and $\bar{Y}$ are the arithmetic means of X and Y respectively.

Hence $xy = (X - \bar{X})(Y - \bar{Y})$

By re-arranging terms in the normal equations we find that

$$b = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2} \qquad \ldots(5.17)$$

$$a = \bar{Y} - b\bar{X} \qquad \ldots(5.18)$$

You may recall that *covariance* is given by $\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$.

Moreover, variance of X is given by $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

Since $b = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2}$ we can say that $b = \frac{\sigma_{xy}}{\sigma_x^2}$ …(5.19)

Since these formulae are derived from the normal equations we get the same values for *a* and *b* in this method also. For the data given in Table 5.4 we compute the values of a and b by this method. For this purpose we construct Table 5.6.

Table 5.6: Computation of Regression Line (short-cut method)

X_i	Y_i	x_i	y_i	x_i^2	$x_i\ y_i$
60	33	-15	-12	225	180
62	37	-13	-8	169	104
65	38	-10	-7	100	70
71	42	-4	-3	16	12
73	42	-2	-3	4	6
75	45	0	0	0	0
81	49	6	4	36	24
85	52	10	7	100	70
88	55	13	10	169	130
90	57	15	12	225	180
Total = 750	450	0	0	1044	776

On the basis of Table 5.6 we find that

$$\overline{X} = \frac{750}{10} = 75 \quad \text{and} \quad \overline{Y} = \frac{450}{10} = 45$$

$$b = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2} = \frac{776}{1044} = 0.743$$

$$a = \overline{Y} - b\overline{X} = 45 - 0.743 \times 10 = -10.73$$

Thus the regression line in this method also $\hat{Y}_i = -10.73 + 0.743X_i$ …(5.20)

Coefficient *b* in (5.20) is called the regression coefficient. This coefficient reflects the amount of increase in Y when there is a unit increase in X. In regression equation (5.20) the coefficient *b* = 0.743 implies that if rainfall increases by 1 mm., agricultural production will increase 0.743 thousand tonne.

Regression coefficient is widely used. It is also an important tool of analysis. For example, if Y is aggregate consumption and X is aggregate income, *b* represents marginal propensity to consume (MPC).

5.11 PREDICTION

A major interest in studying regression lies in its ability to forecast. In Example 5.1 we assumed that the quantity of agricultural production is dependent on the amount of rainfall. We fitted a linear equation to the observed data and got the relationship

$$\hat{Y}_i = -10.73 + 0.743X_i$$

From this equation we can predict the quantity of agricultural output given the amount of rainfall. Thus when rainfall is 60 mm. agricultural production is $(-10.73 + 0.74 \times 60) = 33.85$ thousand tonnes. This figure is the *predicted value* on the basis of regression equation. In a similar manner we can find the predicted values of Y for different values of X.

Let us compare the predicted value with the observed value. From Table 5.4, where observed values are given, we find that when rainfall is 60 mm. agricultural production is 33 thousand tonnes. In fact, the predicted values $\hat{Y}_i$ for observed values of X are given in the fifth column of Table 5.5. Thus when rainfall is 60 mm. Predicted value is 33.85 thousand tonnes. Thus the error value e_i is –0.85 thousand tonne.

Now a question arises, 'Which one, between observed and predicted values, should we believe?' In other words, what will be the quantity of agricultural production if there is a rainfall of 60 mm. in future? On the basis of our regression line it is given to be 33.85 tonnes. And we accept this value because it is based on the overall data. The error of –0.85 is considered as a random fluctuation which may not be repeated.

The second question that comes to our mind is, 'Is the prediction valid for any value of X?' For example, we find from the regression equation that when rainfall is zero, agricultural production is –10.73 thousand tonne. But common sense tells us that agricultural production cannot be negative! Is there anything wrong with our regression equation? In fact, the regression equation here is estimated on the basis of rainfall data in the range of 60-90 mm. Thus prediction is be valid in this range of X. Our prediction should not be for far off values of X.

A third, question that arises here is, 'Will the predicted value come true?' This depends upon the *coefficient of determination*. If the coefficient of determination is closer to one, there is greater likelihood that the prediction will be realised. However, the predicted value is constrained by elements of randomness involved with human behaviour and other unforeseen factors.

5.12 RELATIONSHIP BETWEEN REGRESSION AND CORRELATION

In regression analysis the status of the two variables (X, Y) are different such that Y is the variable to be predicted and X is the variable, information on which is to be used. In the rainfall-agricultural production problem, it makes sense to predict agricultural production on the basis of rainfall and it would not make sense to try and predict rainfall on the basis of agricultural production. However, in the case of scores in Economics and Statistics (see Table 5.1), either one could be X and the other Y. Hence we consider the two prediction problems: (i) predicting Economics score (Y) from Statistics score (X); and (ii) predicting Statistics score (X) from Economics score (Y).

Thus, we can have two regression coefficients from a given set of data depending upon the choice of dependent and independent variables. These are:

a) Y on X line, $Y_i = a + bX_i$

b) X on Y line, $X_i = \alpha + \beta Y_i$

You may ask, 'What is the need for having two different lines? By rearrangement of terms of the Y on X line we obtain $X_i = -\frac{a}{b} + \frac{1}{b}Y_i$. Thus we should have $\alpha = -\frac{a}{b}$ and $\beta = \frac{1}{b}$. However, the observations are not on a straight line and the relation between X and Y is not a mathematical one. You may recall that estimates of the parameters are obtained by the method of least squares. Thus the regression line $\hat{Y}_i = a + bX_i$ is obtained by minimising $\sum_i (Y_i - a - bX_i)^2$ whereas the regression line $\hat{X}_i = \alpha + \beta Y_i$ is obtained by minimising $\sum_i (X_i - \alpha - \beta Y_i)^2$.

However, there is a relationship between the two regression coefficients b and β. We have noted earlier that $b = \frac{\sigma_{xy}}{\sigma_x^2}$. By a similar formula by interchanging the roles of X and Y we find $\beta = \frac{\sigma_{xy}}{\sigma_y^2}$. But by definition we notice that $\sigma_{xy} = \sigma_{yx}$.

Thus $b \times \beta = \frac{\sigma_{xy}^2}{\sigma_x^2 \times \sigma_y^2}$, which is the same as r^2.

This r^2 is called the *coefficient of determination*. Thus the product of the two regression coefficients of Y on X and X on Y is the square of the correlation coefficient. This gives a relationship between correlation and regression. Notice, however, that the coefficient of determination of either regression is the same, i.e., r^2; this means that although the two regression lines are different, their predictive powers are the same. Note that the coefficient of determination r^2 ranges between 0 and 1, i.e., the maximum value it can assume is unity and the minimum value is zero; it cannot be negative.

From the previous discussions, two points emerge clearly:

1) If the points in the scatter lie close to a straight line, then there is a strong relationship between X and Y and the correlation coefficient is high.
2) If the points in the scatter diagram lie close to a straight line, then the observed values and predicted values of Y by least squares are very close and the prediction errors $(Y_i - \hat{Y}_i)$ are small.

Thus, the prediction errors by least squares seem to be related to the correlation coefficient. We explain this relationship here. The sum of squares of errors at the various points upon using the least squares linear regression is $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$.

On the other hand, if we had not used the value of observed X to predict Y, then the prediction would be a constant, say, a. The best value of a by least squares criterion is such an a that minimises $\sum_{i=1}^{n}(Y_i - a)^2$; the solution to this a is seen to be $\bar{Y}$. Thus the sum of squares of errors of prediction at various points without using X is $\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$.

The ratio, $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \Big/ \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ can then be used as an index of how much has been gained by the use of X. In fact, this ratio is the coefficient of determination and same as r^2 mentioned above. Since both the numerator and denominator of this ratio are non-negative, the ratio is greater than or equal to zero.

Check Your Progress 3

1) From the following data find the coefficient of linear correlation between X and Y. Determine also the regression line of Y on X, and then make an estimate of the value of Y when $X = 12$.

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

.....

……

……

……

……

……

2) Obtain the lines of regression for the following data:

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

……

……

……

……

……

3) Find the two lines of regression from the following data:

Age of Husband (X)	2	2	2	2	3	2	2	4	2	1
	5	2	8	6	5	0	2	0	0	8
Age of Wife (Y)	1	1	2	1	2	1	1	2	1	1
	8	5	0	7	2	4	6	1	5	4

Hence estimate (i) age of husband when the age of wife is 19, (ii) the age of wife when the age of husband is 30.

……

……

……

……

……

4) From the following data, obtain the two regression equations :

Sales	:	91	97	108	121	67	124	51	73	111	57
Purchases	:	71	75	69	97	70	91	39	61	80	47

……

……

……

……

……

5) Obtain the equation of the line of regression of yield of rice (y) on water (x) from the data given in the following table :

Water in inches (x)	12	18	24	30	36	42	48
Yield in tons (y)	5.27	5.68	6.25	7.21	8.02	8.71	8.42

Estimate the most probable yield of rice for 40 inches of water.

……

……

……

……

……

5.13 MULTIPLE REGRESSION

So far we have considered the case of the dependent variable being explained by one independent variable. However, there are many cases where the dependent variable is explained by two or more independent variables. For example, yield of crops (Y) being explained by application of fertilizer (X_1) and irrigation water(X_2). This sort of models is termed multiple regression. Here, the equation that we consider is

$$Y = \alpha + \beta X_1 + \gamma X_2 + e \qquad \ldots(5.21)$$

Where Y is the explained variable, X_1 and X_2 are explanatory variables, and e is the error term. In order to make the presentation simple we have dropped the subscripts. A regression equation can be fitted to (5.21) by applying the method of least squares. Here also we minimise $\sum e^2$ and obtain the normal equations as follows:

$$\begin{aligned} \Sigma Y &= n\alpha + \beta\Sigma X_1 + \gamma\Sigma X_2 \\ \Sigma X_1 Y &= \alpha\Sigma X_1 + \beta\Sigma X_1^2 + \gamma\Sigma X_1 X_2 \\ \Sigma X_2 Y &= \alpha\Sigma X_2 + \beta\Sigma X_1 X_2 + \gamma\Sigma X_2^2 \end{aligned} \qquad \ldots (5.22)$$

By solving the above equations we obtain estimates for α, β and γ. The regression equation that we obtain is

$$\hat{Y} = \alpha + \beta X_1 + \gamma X_2 \qquad \ldots(5.23)$$

Remember that we obtain predicted or forecast values of Y (that is $\hat{Y}$) through (5.23) by applying various values for X_1 and X_2.

In the bivariate case (Y,X) we could plot the regression line on a graph paper. However, it is quite complex to plot the three variable case (Y, X_1, X_2) on graph paper because it will require three dimensions. However, the intuitive idea remains the same and we have to minimise the sum of errors. In fact when we add all the error terms ($e_1, e_2, \ldots\ldots e_n$) it sum up to zero.

In many cases the number of explanatory variables may be more than two. In such cases we have to follow the basic principle of least squares: minimize Σe^2. Thus if $Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots\ldots\ldots\ldots + a_n X_n + e$ then we have to minimize

$$\Sigma e^2 = \Sigma (Y - a_0 - a_1 X_1 - a_2 X_2 \ldots\ldots - a_n X_n)^2$$

and find out the normal equations.

Now a question arises, 'How many variables should be added in a regression equation?' It depends on our logic and what variables are considered to be important. Whether a variable is important or not can be identified on the basis of statistical tests also. These tests will be discussed later in Block 4.

We present a numerical example of multiple regression below.

Example 5.3

A student tries to explain the rent charged for housing near the University. She collects data on monthly rent, area of the house and distance of the house from the university campus and fits a linear regression model.

Rent (in Rs.'000) Y	Area (in sq.mt.) X_1	distance(in Km.) X_2
20	65	5.7
25	66	3.2
26	70	7.5
28	70	6.5
30	75	5.0
31	76	4.0
32	72	6.0
33	75	6.2
35	78	3.5
40	103	2.4

In the above example rent charged (Y) is the dependent variable while area of the house (X_1) and distance of the house from the university campus (X_2) are independent variables.

The steps involved in estimation of regression line are:

i) Find out the regression equation to be estimated. In this case it is given by $Y = \alpha + \beta X_1 + \gamma X_2 + e$.

ii) Find out the normal equations for the regression equation to be estimated. In this case the normal equations are

$$\Sigma Y = n\alpha + \beta \Sigma X_1 + \gamma \Sigma X_2$$
$$\Sigma X_1 Y = \alpha \Sigma X_1 + \beta \Sigma X_1^2 + \gamma \Sigma X_1 X_2$$
$$\Sigma X_2 Y = \alpha \Sigma X_2 + \beta \Sigma X_1 X_2 + \gamma \Sigma X_2^2$$

iii) Construct a table as given in Table 9.4.

iv) Put the values from the table in the normal equations.

v) Solve for the estimates of α, β and γ.

Table 5.7: Computation of Multiple Regression

Y	X_1	X_2	X_1Y	X_2Y	X_1^2	X_2^2	X_1X_2	$\hat{Y}$	e_i
20	65	5.7	1300	114	4225	32.49	370.5	25.49	-5.49
25	66	3.2	1650	80	4356	10.24	211.2	25.71	-0.71
26	70	7.5	1820	195	4900	56.25	525	27.94	-1.94
28	70	6.5	1960	182	4900	42.25	455	27.85	0.15
30	75	5	2250	150	5625	25	375	30.00	0.00
31	76	4	2356	124	5776	16	304	30.37	0.63
32	72	6	2304	192	5184	36	432	28.72	3.28
33	75	6.2	2475	204.6	5625	38.44	465	30.11	2.89
35	78	3.5	2730	122.5	6084	12.25	273	31.24	3.76
40	103	2.4	4120	96	10609	5.76	247.2	42.58	-2.58
300	750	50	225000	15000	562500	2500	37500	300	0

By applying the above mentioned steps we obtain the estimated regression line as

$\hat{Y} = -4.80 + 0.45X_1 + 0.09X_2$.

5.14 NON-LINEAR REGRESSION

The equation fitted in regression can be non-linear or curvilinear also. In fact, it can take numerous forms. A simpler form involving two variables is the quadratic form. The equation is

$Y = a + bX + cX^2$

There are three parameters here viz., a, b and c and the normal equations are:

$$\Sigma Y = n\alpha + b\Sigma X + c\Sigma X^2$$
$$\Sigma XY = \alpha\Sigma X + b\Sigma X^2 + c\Sigma X^3$$
$$\Sigma X^2 Y = \alpha\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4$$

By solving for these equation we obtain the values of a, b and c.

Certain non-linear equations can be transformed into linear equations by taking logarithms. Finding out the optimum values of the parameters from the transformed linear equations is the same as the process discussed in the previous section. We give below some of the frequently used non-linear equations and the respective transformed linear equations.

1) $Y = a\,c^{bx}$

By taking natural log (ln), it can be written as

$\ln Y = \ln a + bX$

or $Y' = \alpha + \beta X'$

Where, $Y' = \ln Y$, $\alpha = \ln a$, $X' = X$ and $\beta = b$

2) $Y = aX^b$

By taking logarithm (log), the equation can be transformed into

$\log Y = \log a + b \log X$

or $Y' = \alpha + \beta X'$

where, $Y' = \log Y$, $\alpha = \log a$, $\beta = b$ and $X' = \log X$

3) $Y = \dfrac{1}{a + bX}$

If we take $Y' = \dfrac{1}{Y}$ then

$Y' = a + bX$

4) $Y = a + b\sqrt{X}$

If we take $X' = \sqrt{X}$ then

$Y = a + bX'$

Once the non-linear equation is transformed, the fitting of a regression line is as per the method discussed in the beginning of this Unit.

We derive the normal equations and substitute the values calculated from the observed data. From the transformed parameters, the actual parameters can be obtained by making the reverse transformation.

Check Your Progress 4

1) Using the data on scores in Statistics and Economics of Table 5.1, compute the regression of y on x and x on y and check that the two lines are different. On the scatter diagram, plot both these regression lines. Check that the product of the regression coefficients is the square of the correlation coefficient.

……

……

……

……

……

2) Suppose that the least squares linear regression of family expenditure on clothing (Rs. y) on family annual income (Rs. x) has been found to be $y = 100 + 0.09x$, in the range $1000 < x < 100000$. Interpret this regression line. Predict the expenditure on the clothing of a family with an annual income of Rs. 10,000. What about families with annual income of Rs. 100 and Rs. 10,00,000?

……

……

……

……

……

5.15 LET US SUM UP

In this Unit we discussed an important statistical tool, that is, regression. In regression analysis we have two types of variables: dependent and independent. The dependent variable is explained by independent variables. The relationship between variable takes the form of a mathematical equation. Based on our logic, understanding and purpose of analysis we categorise variables and identify the equation form.

The regression coefficient enables us to make predictions for the dependent variable given the values of the independent variable. However, prediction remains more or less valid within the range of data used for analysis. If we attempt to predict for far off values of the independent variable we may get insensible values for the dependent variable.

5.16 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) + 0.47

2) + 0.996

3) + 0.98

4) + 0.995

5) – 0.84

Check Your Progress 2

1) 2/3

2) + 0.64

3) – 0.21, + 0.64, –0.30

4) + 0.82

Check Your Progress 3

1) + 0.98 ; $y = 0.64x + 0.54$; 8.2

2) $x = 0.95y - 6.4$; $y = 0.95x + 7.25$

3) $x = 2.23y - 12.70$; $y = 0.39x + 7.33$

(i) 29.6 (ii) 18.9

4) $y = 0.613x + 14.81$; $x = 1.360y - 5.2$

5) $y = 3.99 + 0.103x$; 8.11 tons

Check Your Progress 4

1) (i) $y = a + bx = 5.856 + 0.676x$

(ii) $x = \alpha + \beta y = 29.848 + 0.799y$

(iii) $r = 0.73$

(iv) $0.676 \times 0.799 = 0.54$

2) Expenditure on clothing, when family income is Rs. 10,000, is Rs. 1,000. In case of income below 1,000 or above 1,00,000 the regression line may not hold good. In between both these figures, one rupee increase in income increases expenditure on clothes by 9 paise.

UNIT 6 INDEX NUMBERS*

Structure

6.0 OBJECTIVES

After going through this Unit, you will be able to:

- define index numbers; and
- Construct and calculate them.

6.1 INTRODUCTION

An "index" in the common sense of the word is an "indicator" and not anything more than that. "Index numbers" or "indices" are forms of the plural, but they all mean the same thing.

An index number represents the general level of magnitude of the changes between two (or more) periods of time or places, in a number of variables taken as a whole. In this definition, the world "variable" refers to numerical variables which can be measured in quantity, such as the prices of commodities.

* Adapted from IGNOU study material of EEC 13: Elementary Statistical Methods and Survey Techniques, Unit 10 written by J Roy with modifications by Kaustuva Barik

For example, we may like to compare the price level of an article between 2010 and 2020 or between Mumbai and Kolkata. Let us consider the yield of rice in 2015 and in 2020 as 50,000 and 60,000 tons respectively. The year 2015 is taken as base for comparison of yields that is 2015 = 100. The corresponding figure for 2020 will be $\frac{60,000}{50,000} \times 100 = 120$. This is a single-commodity index number in its simplest form, being just a relative number. In practice, however, we deal usually with a number of commodities for the construction of an index.

Index numbers are ratios that are usually expressed as percentage in order to avoid awkward decimals. Thus if one commodity costs Rs. 45 in 2019 and Rs. 150 in 2020 the ratio would be $\frac{150}{45}$ or 3.33. If instead of this, we express the ratio into a percentage $\frac{150}{45} \times 3.33$, we say that the index is 333, based on 2019, which is 100.

6.2 STEPS IN CONSTRUCTION OF INDEX NUMBERS

Many government and private agencies are engaged in computation of index numbers or indices as they are often required for the purpose of forecasting business and economic conditions, providing general information, etc.

It is not always the case that the comparison should be over time, but most common types of index numbers measure changes over time. Similarly, index numbers may be constructed for studying changes in any variable such as intelligence, aptitude, efficiency, production, etc. However, the time series of prices is frequently used. Therefore, in the discussion below we will concentrate mostly on prices of commodities. The principles of construction are, however, quite general in nature, and may thus be applied to other areas of interest.

There are various uses of price index numbers. The *wholesale price index number* indicates the price changes taking place n wholesale markets. In the other hand, the *consumer price index number* or the *cost of living index number* tells us about the changes in the prices faced by an individual consumer. Its major application is in the calculation of dearness allowance so that real wage does not decrease; or in comparing the cost of living in, say, different regions. It is also used to measure change in purchasing power of money. The reciprocal of a general price index is known as *purchasing power of money* with reference to the base period. For example, if the price index number goes up to 150, it means that the same amount of money will be able to purchase 100/150 = 0.67 times or 67% of the volume of goods being purchased in the base period.

6.2.1 Selection of Base Period

Since index numbers measure relative changes, they are expressed with one selected situation (e.g., period, place, etc.) as 100. This is called the *base* or the starting point of the series of index numbers. For example, a date is first chosen and all changes are measured from it. The base may be one day such as with index or retail prices, the average of a year, or the average of a period. While selecting a *base period* the following aspects should be taken into consideration.

1) The base period must be "normal" in the sense that the data chosen are not affected by any irregular or abnormal situations such as natural calamities, war, etc. It is desirable to restrict comparisons to stable periods for achieving accuracy.

2) It should not be too back-dated as the patterns of trade, imports or consumer preferences may change considerably if the time-span is too long. A five to ten year interval is likely to be suitable for one base date, and after that the index becomes more and more outdated. Greater accuracy is attained for moderate short-run indices than for those covering greater span of time.

3) For indices dealing with economic data, the base period should have some economic significance.

6.2.2 Choice of a Suitable Average

An index number is basically the result of averaging a series of data (e.g., *price-relatives* of several commodities). There are, however, several ways of a averaging a series: mean (i.e., arithmetic mean), mode median, geometric mean and harmonic mean.

The question naturally arises as to which average to chose. The mode has the merit of simplicity, but may be indefinite. The median suffers from the same limitations. Moreover, neither of them takes into account the size of the items at each end of a distribution. The harmonic mean has very little practical applications to index numbers. As a result, mode, median and harmonic mean are not generally used in the calculation of index numbers. Thus, the arithmetic mean is most commonly used. However, the geometric mean is sometimes used despite its slight difficulty in calculation.

6.2.3 Selection of Items and their Numbers

The number and kinds of commodities to be included in the construction of an index numbers depend on the particular problem to be dealt with, economy and ease of calculations. Various practical considerations determine the number and kinds of items to be taken into account. For a wholesale price index, the number of commodities should be as large as possible. On the other hand, for an indexed meant to serve as a predictor of price movement rather than an indicator movement rather than an indicator of changes over time, a much smaller number of items may be adequate. Care should, however, be taken to ensure that items chosen are not too few which make the index unrepresentative of the general level. A fixed set of commodities need not also be used for a very long period as some items lose their importance with the passage of time and some new items gain in significance. In general, the commodities should be sensitive and representative of the various elements in the price system.

6.2.3 Collection of Data

As prices often vary from market to market, they should be collected at regular intervals from various representative markets. It is desirable to select shops which are visited by a cross section of customers. The reliability of the index depends greatly on the accuracy of the quotations given for each constituent item.

6.3 METHOD OF CONSTRUCTION OF INDEX NUMBER

Various methods of construction of index numbers are as follows:

1) Relative methods

   a) Simple average of relatives

   b) Weighted average of relatives

2) Aggregative methods

   a) Simple aggregative formula

   b) Weighted aggregative formula

      i) Lasperyres' index

      ii) Paasche's index

      iii) Edgeworth-Marshall's index

      iv) Fisher's Ideal index.

6.3.1 Relative Methods

If we record prices of a variety of commodities at a given date and at a later date record the prices of similar items, the change in price can be simply expressed as a percentage of the new compared with the old for each commodity.

This provides us with price relatives and if weights are available the next step will be to multiply the relatives by the weights. Finally, an index number can be produced if we add together the weighted relatives and calculate an average. It is unrealistic to assume that the consumption of each commodity has been equal. So most indices take account of the proportions of each item actually used. This method to weighting shows the relative importance of each in the series.

Given k commodities with base year prices of

$$P_{o1,}\ P_{o2,}.\ .\ .\ .\ P_{ok,}$$

and current prices of

$$P_{o1,}\ P_{o2,}.\ .\ .\ .\ P_{ok,}$$

the price relative for the i^{th} commodity will be $\frac{P_{ni}}{P_{0i}}$ where i = 1, 2, ……, k and the subscript 0 refers to the base year and subscript n refers to the current year.

a) Simple Average of Relatives

The arithmetic mean of the price relative is given by

$$\text{Index} = 100 \sum_{i=1}^{k} \frac{\left(P_{ni}/P_{oi}\right)}{k} \quad \ldots(6.1)$$

For simplicity we can omit the subscript 'i' and write

$$\text{index} = 100 \sum \left(\frac{P_n/P_o}{k}\right)$$

b) Weighted Average of Relatives

The most suitable weights to use are the value of each item, which is denoted by w_i for the i-th commodity. We may use the value of the base year quantities sold at the base year prices ($w_i = p_{oi}\ q_{oi}$) or current year quantities sold at current prices ($w_{1i} = p_{1i}\ q_{1i}$)

If we omit the subscript 'i' for simplicity, a weighted arithmetic mean of price relatives using **base year values as weights** is given by

$$\text{index} = \frac{\sum \frac{P_n}{P_0} \times w_0}{\sum w_0} \times 100 \quad \ldots(6.2)$$

You should note that the base year weighting preserves continuity, but it loses "up-to-dateness" in the course of time.

Example 6.1: The table below presents the average fares per railway journey. Using 2010 average = 100, calculations are made according to base year weights.

Class of Ticket	No. of passenger journeys in 2010 (in millions)	Fare		Weights $w_0 = p_0 q_0$	Price relatives $P = \frac{P_n}{P_0}.100$	$P.w_0$
		2010	2020			
Full Fare	23	12	60	276	500	138000
Excursions	25	6	30	150	500	75000
Festival	20	4	15	80	375	30000
Season tickets	32	5	14	160	280	44800
Total				**666**		**287800**

Applying formula (6.2), we get

Index for 2020 = $\frac{287800}{666} = 432.13.$

Using **Current year values ($\boldsymbol{w_n = P_n q_n}$) as weights,** the index is given by

$$\text{index} = \frac{\sum \frac{P_n}{P_0} \times w_n}{\sum w_n} \times 100 \quad \ldots(6.3)$$

Example 6.2: The table below shows the average fares per railway journey. Using 2010 average = 100, calculation are made according to current year weights.

Class of Ticket	No. of passenger journeys in 2010 (in millions)	Fare		Weights $w_n = p_n q_n$	Price relatives $P = \frac{P_n}{P_0}.100$	$P.w_n$
		2010	2020			
Full Fare	25	12	60	1500	500	750000
Excursions	26	6	30	780	500	390000
Festivals	9	4	15	135	375	50630
Season-tickets	27	5	14	378	280	105800
Total				**2793**		**1296430**

Applying formula (6.3), we get

$$\text{index} = \frac{1296430}{2793} = 464.17$$

6.3.2 Aggregative Methods

In this method, the aggregate (sum-total) of the prices of all commodities in the current or given year is expressed as a percentage of the same in the base year. Thus, in the case of **simple aggregative index,** we have:

$$\text{index number} = \frac{\text{aggregate prices in current year}}{\text{aggregate prices in the base year}} \times 100$$

$$= \frac{P_{n1} + P_{n2} + \ldots\ldots\ldots\ldots P_{nk}}{P_{01} + P_{02} + \ldots\ldots\ldots\ldots P_{0k}} \times 100$$

$$\frac{\sum P_{ni}}{\sum P_{0i}} \times 100 = \frac{\sum P_n}{\sum P_0} \times 100 \qquad \ldots (6.4)$$

where the summation symbol $\left(\sum_{i=1}^{k}\right)$ extends over all the selected commodities numbering *k*. On the other hand, in the case of **weighted aggregative index** we have,

$$\text{General index} = \frac{p_{n1} + p_{n2}q_2 + \ldots\ldots\ldots\ldots p_{nk}\,q_k}{p_{01}q_1 + p_{02}q_2 + \ldots\ldots\ldots\ldots p_{0k}q_k} \times 100$$

$$\frac{\sum p_{ni}q_i}{\sum p_{0i}\,q_i} \times 100$$

$$\text{or simply} = \frac{\sum p_n\,q}{\sum p_0 q} \times 100 \qquad \ldots (6.5)$$

The weights used should be actual quantities bought or sold, and these are kept unchanged until such time as the index requires to be revised.

There are many formulae for weighted aggregative index, but depending on the type of weights used, we discuss four indices which are commonly used.

a) Laspeyres' index

If we use base period quantities (q_0) as the weights in the general weighed aggregative index formula (6.5), we get what is known as Laspeyre's formula (L).

$$L = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100 \quad \ldots (6.6)$$

It can be seen that this index has fixed base year quantity as weights (q_0) and is equivalent to arithmetic mean of price relatives given at formula (6.2). Thus, we can also write (6.6) as

$$L = \frac{\sum \frac{p_n}{p_0} \times p_0 q_0}{\sum p_0 q_0} \times 100$$

b) **Paasche's index**

If we use current year quantities (q_n) as weights in the general aggregative index formula(6.5), we get what get what is known as Paasche's formula(P).

$$P = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100 \quad \ldots (6.7)$$

Where q_n(actually $q_{n1}, q_{n2}, \ldots . q_{nk}$)are the quantities bought or sold in the current period.

c) **Fisher's Ideal Index**

An index number obtained as geometric mean (i.e., square root of the product) of indices obtained by Laspeyres' and Paasche's formulae, satisfies certain important properties (to be discussed later), is known as the Fisher's ideal formula.

$$F = \sqrt{L \times P} = \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0}} \times \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100 \quad \ldots (6.8)$$

d) **Edgeworth-Marshall Index**

If the mean of the base period and the current period quantities is used as weight, i.e.,

$w = \frac{1}{2}(q_0 + q_n)$, we get a compromise formula of 'Edgeworth-Marshall index'.

$$I = \frac{\sum p_n (q_0 + q_n)/2}{\sum p_0 (q_0 + q_n)/2} \times 100$$

$$= \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)} \times 100 \quad \ldots (6.9)$$

We take some hypothetical data and calculate the above indices from it (see Table 6.1).

Table 6.1: Illustrative Calculation of Laspeyres' Edgeworth-Marshall]s and Fisher's Indices

Item	Base Year		Current Year		(p_0q_0)	(p_nq_0)	(p_0q_n)	(p_nq_n)
	Price (p_0)	Quantity (q_0)	Price (p_n)	Quantity (q_n)				
A	20	7	25	9	140	175	180	225
B	42	6	40	8	252	240	336	320
C	30	17	25	4	510	425	120	100
D	8	15	14	10	120	210	80	140
E	10	8	13	5	80	104	50	65
Total					**1102**	**1154**	**766**	**850**

From the above table we calculate various indices as follows:

1) Lasperyres' price index $= \frac{\sum P_n q_0}{\sum P_n q_0} \times 100 = \frac{1154}{1102} \times 100 =$

$$104.72 = 105$$

2) Paasche's price index $= \frac{\sum P_n q_n}{P_0 q_n} \times 100 = \frac{850}{766} \times 100 =$

$$110.97 = 111$$

3) Edgewroth-Marshall's price index $= \frac{\sum P_n q_0 + \sum P_n q_n}{\sum P_0 q_0 + \sum P_0 q_n} \times$

$$100 = \frac{1154+850}{1102+766} \times 100 =$$

$$\frac{2004}{1868} \times 100 = 107.28 = 107$$

4) Fisher's ideal index $= \sqrt{\frac{\sum P_n q_0}{\sum P_0 q_0} \frac{\sum P_n q_n}{\sum P_0 q_n}} \times 100 =$

$$\sqrt{[(L) \times (P)]} =$$

$$\sqrt{(104.72 \times 110.97)} = 107.8 = 108$$

Note that for the same price change, different formulae provide different values. Moreover, when prices are increasing, Laspeyres' index gives the lowest value while Paasche's index gives the highest value. *Therefore, it is often said that Laspeyres' index is an under-estimate while Paasche's index is an over-estimate of true price change.*

6.3.3 Quantity or Volume Index Numbers

We can get a quantity or volume index number, which measures and permits

comparison of quantities of goods, from corresponding price index number formulae simply by replacing *p.* by *q* and *q* and *q* by *p.*

1) Quantity relative $= \frac{q_n}{q_0} \times 100$

2) Arithmetic Mean (A.M.) of quantity of relatives = $100 \sum \left(\frac{q_n}{q_0}\right) / k$

3) Weighted A.M. of quantity relative index:

a) Base year weights: $\frac{\sum(q_n/q_0)}{\sum w_0} \times 100 \qquad (\text{where } w_0 = p_0 q_0)$

b) Current year weights: $\frac{\sum(q_n/q_0)\times w_n}{\sum w_n} \times 100 \qquad (\text{where } w_n = p_n q_n)$

4) Simple aggregative quantity index $= \frac{\sum q_n}{\sum q_0} \times 100$

5) Laspeyres' quantity index $= \frac{\sum q_n P_0}{\sum q_0 P_0} \times 100$

6) Paasche's quantity index $= \frac{\sum q_n P_n}{\sum q_0 P_n} \times 100$

7) Fisher's ideal index $= \sqrt{\frac{\sum q_n P_0}{\sum q_0 P_0} \frac{\sum q_n P_n}{\sum q_0 P_n}} \times 100$

8) Edgeworth-Marshall's index $= \frac{\sum q_n (P_0 + P_n)}{\sum q_0 (P_0 + P_n)} \times 100$

Check Your Progress 1

1) What do index numbers seek to measure?

……

……

……

……

……

……

2) Discuss the various problems involved in construction of index numbers with particular reference to price indices.

……

……

……

……

……

3) The following are the prices of six different commodities for 2020 and 2021. Compute the price index by (a) aggregative method, and (b) average of price relatives method by using arithmetic mean.

Commodities	Price in 2020 (Rs.)	Price in 2021 (Rs.)
A	40	50
B	50	60
C	20	30
D	50	70
E	80	80
F	100	110

……

……

……

……

……

4) Calculate Fisher's Ideal Index Number from the following group of items.

	Base Year		Current Year	
Item No.	Price (in Rs.)	Quantity (in Kg.)	Price (in Rs.)	Quantity (in Kg.)
1	4	1.0	3	4
2	8	1.5	7	5

……

……

……

……

……

5) Calculate Laspeyres' and Paasche's Index Number from the following data.

	Base Year		Current Year	
Item	Quantity	Price	Quantity	Price
Bread	6.0	40	7.0	30
Meat	4.0	45	5.0	50
Tea	0.5	90	1.5	40

……

……

……

……

6.4 MERITS OF THE VARIOUS AGGREGATIVE MEASURES

The different index numbers serve different purposes and, therefore, the appropriateness of a particular index number depends on the purpose at hadn.

The Laspeyres' index calculation is simpler, since this uses the base period quantities as weighs which are not difficult to get and the denominator needs calculating only once. But in this index a rise in pries tends to *overstated,* since it does not take into account corresponding falls in demand or changes in output. Indices such as Paasche's on the other hand, use current period quantities as weights which are difficult o get and the weights need to be constructed afresh for every year. Moreover, Paasche's index tends to *understate* the rise in prices because it uses current weights. The Laspeyres' index is probably more commonly used, since it is convenient to employ fixed weights. But with the passage of time the weights are rendered out of date. For example, in 1995 the number of mobile phones in Odisha was nil. In 2020 there are more mobile phones than the number of land line connections. The Paasche's index uses the preferable current weights, but since-up-to-date information on quantity of goods produced or consumed or marketed or distributed are not readily obtained, the Laspeyres' index has a great advantage.

6.5 TESTS FOR INDEX NUMBERS

A perfect index number, which measures the change in the level of a phenomenon from one period to another, should satisfy certain tests. There are three major tests of index numbers: (1) Time reversal test, (2) Factor reversal test, and (3) Circular test.

6.5.1 The Time Reversal Test

According to this test, if we reverse the time subscripts (such as 0 and n) of a price (or quantity) index the result should be the reciprocal of the original index.

Symbolically,

$$I_{0n} \times I_{0n} = 1$$

where I_{0n} = index number for period n with the base period 0.

I_{n0} = index number for period *0* with the base period n.

If from 2010 to 2020 the price changes from Rs. 4 to Rs. 16, the price in 2020 are 400 percent of the price in 2010, and the price in 2010 is 25 percent of the price in 2020. The product of the two price relatives is $4 \times 0.25 = 1$. The test is based on the analogy that the principle, which hold good for a single commodity, should, also be true for the index number as a whole.

There are five methods which do satisfy the time reversal test. These are:

1) Simple geometric mean of price relative

2) Aggregative indices with fixed weights

3) Edgeworth-Marshall formula

4) Weighted geometric mean of price relatives if fixed weights are used

5) Fisher's ideal index

Fisher's ideal index $F = \sqrt{\frac{\sum P_n q_0}{\sum P_0 q_0} \times \frac{\sum P_n q_n}{\sum P_0 q_n}}$

If time subscripts are reversed,

$F' = \sqrt{\frac{\sum P_0 q_n}{\sum P_n q_n} \times \frac{\sum P_0 q_0}{\sum P_n q_0}}$

Since $F \times F' = 1$ the test is satisfied

6.5.2 The Factor Reversal Test

With the usual notations, a "value index" formula is given by

$$I_v = \frac{\sum P_n q_n}{\sum P_0 q_n}$$

Now, for example, Laspeyres' index for prices and quantities are given respectively by

$$I_p = \frac{\sum P_n q_0}{\sum P_0 q_n} \text{ and } I_q = \frac{\sum q_n P_0}{\sum q_0 P_0}$$

The factor reversal test desires that $I_p . I_q = I_v$

But for Laspeyres' index

$$I_p . I_q = \frac{\sum (P_n q_0)(\sum P_n q_0)}{\sum (P_0 q_0)^2} = I_v$$

On the other hand, Fisher's ideal index satisfies this test, as shown below.

$$I_p = \sqrt{\frac{\sum P_n q_0}{\sum P_n q_0} \times \frac{\sum P_n q_n}{\sum P_0 q_n}}$$

$$I_q = \sqrt{\frac{\sum q_n P_0}{\sum q_n P_0} \times \frac{\sum q_n P_n}{\sum q_0 P_n}}$$

$$I_p . I_q = \sqrt{\frac{\sum P_n q_0}{\sum P_0 q_0} \times \frac{\sum P_n q_n}{\sum P_0 q_n} \times \frac{\sum q_n P_0}{\sum q_0 P_0} \times \frac{\sum q_n P_n}{\sum q_0 P_n}}$$

$$= \sqrt{\frac{\sum P_n q_n}{\sum P_0 q_0} \times \frac{\sum P_n q_n}{\sum q_0 P_0}} = \frac{\sum P_n q_n}{\sum P_0 q_0} = I_v$$

To understand this principle further, we take the following example.

If the price and quantity per unit of an item changed in 2020, as compared to 2010, from Rs. 16 to Rs. 32 and from 100 units to 200 units respectively, then the price and quantity in 2020 would both be 200% or 2.00 times the price and quantity in 2010. The values (product of price and quantity) would be Rs. 1600 in 2010 and Rs. 6400 in 2020, so that the value ratio is 6400/1600 = 4.00 Thus, we verify that 2.00 × 2.00 = 4.00, that is, the product of price ratio and quantity ratio is equal to the value ratio.

Only the Fisher's ideal index satisfies this test.

Example 6.3 we show with the following data that the Fisher's ideal index satisfies the factor reversal test:

Item	Base Year		Current Year		(p_0q_0)	(p_nq_0)	(p_0q_n)	(p_nq_n)
	Price (p_0)	Quantity (q_0)	Price (p_n)	Quantity (q_n)				
I	6	50	10	56	300	500	336	560
II	2	100	2	120	200	200	240	240
III	4	60	6	60	240	360	240	360
IV	10	30	12	24	300	360	240	288
V	8	40	12	36	320	480	288	432
Total					**1360**	**1900**	**1344**	**1880**

Let us calculate the following from the data given in the above table.

Price Ratio: $I_p = \sqrt{\frac{\sum P_n q_0}{\sum P_n q_0} \times \frac{\sum P_n q_n}{\sum P_0 q_n}} = \sqrt{\frac{1900}{1360} \times \frac{1880}{1344}}$

Quantity Ratio: $I_p = \sqrt{\frac{\sum q_n P_0}{\sum q_n P_0} \times \frac{\sum q_n P_n}{\sum q_0 P_n}} = \sqrt{\frac{1344}{1360} \times \frac{1880}{1900}}$

Value Ratio: $I_v = \frac{\sum P_n q_n}{\sum P_0 q_0} = \frac{1880}{1360}$

$I_p.I_q = \sqrt{\frac{1900}{1360} \times \frac{1880}{1344} \times \frac{1340}{1360} \times \frac{1880}{1900}} = \sqrt{\frac{1880}{1360} \times \frac{1880}{1360}} = \frac{1880}{1360}$

$= I_v$ which shows that the test is satisfied.

6.5.3 Chain Index Number and Circular Test

Two types of base periods are used for the construction of index numbers, namely, (a) fixed base, (b) chain base. Most commonly used indices use fixed base method. This method cannot take into account any changes in price or quantity in any other year. It fails to include new commodities gaining importance at a later date or exclude commodities losing significance in course of time. These problems can be overcome by chain index numbers.

Using a suitable index number formula (say, Laspeyres' index), link indices, defined as follows, are first calculated: Link index = Index number with previous period as base. The chain index is obtained by multiplying link indices progressively. Thus, the chain index number I_{on} for period n with base period 0 is given by

$I_{01} = I_{01}$

$I_{02} = I_{01} \times I_{12}$

$I_{03} = I_{01} \times I_{12} \times I_{23} = I_{02} \times I_{23}$

……..

…….

$I_{0n} = I_{01} \times I_{12} \times I_{23}$ ……..$\times I_{(n-1)n} = I_{(n-1)} \times I_{(n-1)n}$

Example 6.4 The calculation of chain index numbers is illustrated with reference to the following data:

Year	Link index	Chain index (Base 2010 =100)
2010	100	100
2011	$I_{01} = 80$	$100 \times \frac{80}{100} = 80$
2012	$I_{12} = 120$	$80 \times \frac{120}{100} = 96$
2013	$I_{23} = 75$	$96 \times \frac{75}{100} = 72$

Thus, the chain index numbers for the years 2011 to 2013 with 2010 as the base are 80, 96 and 72 respectively.

Circular Test: The circular test is an extension of time reversal test over a number of years. It states that the chain index for the year 2013, calculated above, starting from the base year 2010 will be same as the index number directly calculated with fixed base period of 2010. In symbols,

$I_{01} = I_{12} \times \ldots\ldots\times I_{(n-1)n} \times = I_{n-0)} = 1$. (Notice that $I_{0n} = \frac{1}{I_{n0}}$

Considering an aggregate index with fixed weights

$$\frac{\sum P_1 q}{\sum P_0 q}$$

We can illustrate the test as follows:

With base period 0, we can trace the above formula from 1 to 3 years:

$$\frac{\sum P_1 q}{\sum P_0 q} \times \frac{\sum P_2 q}{\sum P_1 q} \times \frac{\sum P_3 q}{\sum P_2 q} \times \frac{\sum P_0 q}{\sum P_3 q} = 1$$

The formulae satisfying the requirements of circular test are:

1) Simple aggregative index
2) Simple geometric mean of relatives
3) Weighted aggregative index (such as Laspeyres' index with constant weights)
4) Weighted geometric mean of relatives with constant weights.

Fisher's ideal index does not satisfy this test. It has been proved that no index satisfies both the factor reversal and the circular test.

Check your progress 2

1) Compute the chain index number with 2010 prices as the base from the following table giving the average wholesale prices of commodities A, B and C for years 2010-2014.

Commodity	**Average whole sale Price (in Rs.)**				
	2010	**2011**	**2012**	**2013**	**2014**
A	20	16	28	35	21
B	25	30	24	36	45
C	20	25	30	24	30

……...…
……...…
……...…
……...…
……..
……..……

2) Construct Fisher's ideal Index number from the following data and show that it satisfies Factor and Time Reversal Tests.

Commodities	Base Year		Current Year	
	Price Per unit	Expenditure (Rs.)	Price per unit	Expenditure (Rs.)
A	2	40	5	75
B	4	16	8	40
C	1	10	2	24
D	5	25	10	60

……

……

……

……

6.6 COST OF LIVING INDEX (CLI) OR CONSUMER PRICE INDEX (CPI)

This is an index of changes in the prices of goods and services commonly consumed by a homogeneous group of people, such as families of industrial workers. The major items of consumption that are considered for the construction of CLI are:

1) Food
2) Fuel and Light
3) Clothing
4) House rent
5) Miscellaneous

The common method for obtaining the consumption basket is to conduct a family living survey among the population group for which the index is to be constructed. Prices of selected items are also collected from various retail markets used by consumers in question. It may be noted that each of the above broad groups contains several sub groups. Thus, 'food' includes cereals, pulses, oils, meat, fish, egg, spices, vegetables, fruits, non-alcoholic beverages, etc. 'Miscellaneous' includes such items as medical care, education, transport, recreation, gifts and many others. When more than one price quotation is collected for a single commodity, a simple average is taken. Index number is constructed for each of the five groups using weighted average of the price group; the weighs used are proportional to the expenditure on the consumed item buy an average family. Next, the overall index (CLI) is computed as an weighted

average of group indices, the weights being again the proportional expenditure on different groups (e.g., 50 per cent on food).

Using Laspeyres' formula

Cost of living Index: $I = \frac{\sum w\left(\frac{P_n}{P_0} \times 100\right)}{\sum w}$ where $w = \frac{P_0 q_0}{\sum P_0 q_0}$

The CLI or consumer price index (CPI) numbers have significant practical implications and extensive public use. Its use as a wage regulator is the most important. The dearness allowance (DA) of employees is primarily determined by this index. When wages or incomes are divided by corresponding CLI, the effect of changes in prices (inflation) is eliminated. This is known as the process of deflation, which is used to find 'real wages' or 'real income' As mentioned earlier the reciprocal of CLI gives us the purchasing power of money.

Example 6.5: Construction of an index for food

Item	Prices			Weights	
	P_n	P_0	$P = (P_n \times P_0)$	w	$P \times w$
Rice	50	40	125.0	30	3750.0
Wheat	45	30	150.0	20	3000.0
Pulses	60	40	150.0	10	1500.0
Sugar	40	20	200.0	5	1000.0
Oil	75	60	125.0	15	1875.0
Potato	60	50	120.0	15	1800.0
Fish	200	150	133.3	5	666.5
Total				**100**	**13591.5**

Index (food) $= \frac{\sum w \times (P_n + P_0)}{\sum w} \times 100 = \frac{\sum Pw}{\sum w} \times 100$

$= \frac{13591.5}{100} = 135.915 = 135.92$

Example 6.6: Construction of a final Cost of Living index number.

Item	Weight	Index (Percentage Expenditure)	Weight × Index
Food	45	130	5850
Clothing	15	140	2100
Housing	20	170	3400
Fuel	5	110	550
Misc.	15	125	1875
Total	**100**		**13591.5**

Cost of Living Index $= \frac{13,775}{100} = 137.75 = 138$

Check your progress 3

1) Calculate a number which will indicate the percentage change in volume of traffic from October 2019 to October 2020, when account is taken of the relative values of the different types of traffic.

Type of traffic	Tons('000)		Receipts(Rs.'000)
	Oct. 2019	Oct. 2020	Oct. 2019
Merchandise	1246	1206	776
Fuel	4794	4229	562

……

……

……

……

……

2) Compute Paasche's price index number for 2020 with 2015 as the base from the following data:

Commodity	Unit	Price(Rs.) per unit		Quantities sold	
		2015	2020	2015	2020
A	kg.	4	5	95	120
B	kg.	60	70	118	13
C	kg.	35	40	50	70

……

……

……

……

……

3) From the following data, compute the Laspeyres' price index number for 2021 with 2019 as base:

Item	Price(Rs.)		Total Value(Rs.)	
	2019	2021	2019	
A	12.50	14.00	112.50	
B	10.50	12.00		126.00
C	15.00	14.00		105.00
D	9.40	11.20		47.00

……

……

……

……

……

4) Calculate Marshall-Edgeworth index number from the following data:

	2010		2015	
Commodity	**Price**	**Quantity**	**Price**	**Quantity**
Rice	9.3	100	4.5	90
Wheat	6.4	11	3.7	10
Jowar	5.1	5	2.7	3

……

……

……

……

.....

.....

6.7 WORKED OUT EXMAPLES

In this section, we shall provide worked out examples so as to further familiarize you with the topic.

Example1.7: Construction of Price Index

Item	**Unit**	**Price per unit(Rs.)**		
		2010 (P_0)	**2020 (P_n)**	**($P_n \div P_0$)×100**
Rice	quintal	100	220.00	220
Wheat	kg.	1.50	2.40	160
Fish	kg.	15.00	28.00	187
Bread	lb.	0.60	1.35	225
Milk	litre	2.50	4.00	160
Total		**119.60**	**255.75**	**952**

a) **Aggregative Method**

Index number for 2020 (base 2010 =100)

$$\frac{\text{Average price per unit in 2020}}{\text{Average price unit in 2010}} \times 100$$

$= \frac{\sum P_n/k}{\sum P_0/k} \times 100 = \frac{255.75}{119.60} \times 100 = 213.84$

b) **Method of Price Relative**

Index number for 2020 (base 2010 =100)

$= \frac{\sum\left(\frac{P_n}{P_0} \times 100\right)}{k}$

$= {}^{952}/_{5} = 190.4$

Example 6.8: Calculate price index numbers from the following information, using (a) weighted aggregative formula, and (b) weighted arithmetic mean of price relatives:

Item	Unit	Price per unit(Rs.)		
		Base year	Current year	Weight
A	quintal	85	115	19
B	kg.	15	15	25
C	dozen	45	61	40
D	litre	55	100	20
E	Lb	17	23	21

Calculation for Index numbers

Item	P_0	P_n	w	P_0w	$P_nwI = (P_n \div P_0) \times 100$	Iw	
A	85	115	19	1615	2185	135.3	2570.7
B	15	20	25	375	500	133.3	3332.5
C	45	61	40	1800	2440	135.6	5424.0
D	55	100	20	1100	2000	181.8	3636.0
E	17	23	21	357	483	135.3	2841.3
Total			**125**	**5247**	**7608**		**17804.5**

a) Weighted aggregative index $= \frac{\sum P_n w}{\sum P_0 w} \times 100 = \frac{7608}{5247} \times 100 = 144.99 = 145$

b) Weighted arithmetic mean of price relatives

$= \frac{\sum iw}{\sum w} = \frac{17804.5}{125} = 142.44$

Example 6.9: Given below are the data on prices of some consumer goods and the weights attached to the various commodities. Calculate price index numbers

for the year 2021 (base 2020 = 100), using (a) simple average, and (b) weighted average of price relatives.

		Price (Rs.)		
Commodity	**Unit**	**2020**	**2021**	**Weights**
Wheat	Kg.	0.05	0.75	2
Milk	Litre	0.60	0.75	5
Egg	Dozen	2.00	2.40	4
Sugar	Kg.	1.80	2.10	8
Shoes	Pair	8.00	10.00	1

Calculations for price relative index.

Commodity	**Unit**	P_0	P_n	$I=(P_n \div P_0)\times 100$	w	I
Wheat	kg.	0.50	0.75	150	2	300
Milk	Litre	0.60	0.75	125	5	625
Egg	Dozen	2.00	2.40	120	4	480
Sugar	kg.	1.80	2.10	117	8	936
Shoes	Pair	8.00	10.00	125	1	125
Total				**637**	**20**	**2446**

a) Simple average of price relative index = $\frac{\Sigma\left(P_n/P_0\right)\times 100}{k} = \frac{637}{5} = 127.4$

b) Weighted average of price relative index = $\frac{\Sigma Iw}{\Sigma w} = \frac{2466}{20} = 123.3$

Example 6.10: On the basis of the following data, calculate the wholesale price index number for the five groups combined (Base: 2015-16 = 100).

Group	**Weight**	**Index no. for the week ending 31.01.2021**
Food	50	241
Liquor and tobacco	2	221
Fuel, power, light and lubricants	3	204
Industrial raw materials	16	256
Manufactured commodities	29	179

We compute: General index = $\frac{\Sigma Iw}{\Sigma w}$

where I = Group index, and w = Group weight

Group	weight (w)	Group Index(I)	$I \times w$
Food	50	241	12050
Liquor and tobacco	2	221	442
Fuel, power, light and lubricants	3	204	612
Industrial raw materials	16	256	4096
Manufactured commodities	29	179	5191
Total	**100**		**22391**

Index number of wholesale price = $\frac{22391}{100} = 223.91$

Example 6.11: Annual production (in million tons) of four commodities are given below:

Commodity	Production			Weight
	2015	2019	2020	
A	160	200	216	20
B	24	42	45	30
C	50	72	68	13
D	120	168	156	17

Calculate quantity index numbers for the years 2019 and 2020 with 2015 as base year, using (a) simple arithmetic mean, and (b) weighted arithmetic mean of the relatives.

Quantity relatives for 2019 with base year 2015 (=100)

$I = \left({}^{q_n}/_{q_0}\right) \times 100 = \left({}^{q_{54}}/_{q_{50}}\right) \times 100$

Commodity A: $\frac{200}{160} \times 100 = 125$

Commodity B: $\frac{42}{24} \times 100 = 175$

Commodity C: $\frac{72}{50} \times 100 = 144$

Commodity D: $\frac{168}{120} \times 100 = 140$

Quantity relatives for 2020 with 2015 = 100

$I = \left({}^{q_{54}}/_{q_{50}}\right) \times 100$

Commodity A: $\frac{216}{160} \times 100 = 135$

Commodity B: $\frac{45}{24} \times 100 = 187.5$

Commodity C: $\frac{68}{50} \times 100 = 136$

Commodity D: $\frac{156}{120} \times 100 = 130$

Commodity	Quantity relatives(I)		Weight	$I \times w$	
	2019	2020	w	2019	2010
A	125	135.0	20	2500	2700
B	175	187.5	30	5250	5625
C	144	136.0	13	1872	1768
D	140	130.0	17	2380	2210
Total	**584**	**588.5**	**80**	**12002**	**12303**

a) Simple arithmetic mean of quantity relatives $= \frac{\sum\left(q_n/q_0\right)}{k} \times 100$

(where k = number of commodities)

Index number for 2019 = $\frac{584}{4} = 146$

Index number for 2020 = $\frac{588.5}{4} = 147$

b) Weighted arithmetic mean of quantity relatives $\frac{\sum Iw}{\sum w}$

Index number for 2019 = $\frac{12002}{80} = 150$

Index number for 2020 = $\frac{12303}{80} = 154$

Example 6.12: From the following price(p) and quantity (q) data, compute Fisher's ideal index number.

Commodity	2015 (Base Year)		2020 (Current Year)	
	Price	Quantity	Price	Quantity
A	12	10	17	10
B	14	9	16	11
C	11	12	13	10

Calculations for Fisher's ideal index:

Commodity	P_0	q_0	P_n	q_n	$P_0 q_0$	$P_n q_n$	$P_0 q_n$	$P_n q_n$
A	12	10	17	10	120	170	120	170
B	14	9	16	11	126	144	154	176
C	11	12	13	10	132	156	110	130
Total					**378**	**470**	**384**	**476**

Laspeyres' price index = $\frac{\sum P_n q_0}{\sum P_n q_0} \times 100 = \frac{470}{378} \times 100 = 124.34 = 124$

Paasche's price index = $\frac{\sum P_n q_n}{\sum P_0 q_n} \times 100 = \frac{476}{384} \times 100 = 123.96 = 124$

Fisher's ideal index = $\sqrt{L \times P} = \sqrt{124 \times 124} = 124$.

6.8 LET US SUM UP

In this unit you have been introduced to the concepts and methods involved in the construction of Index Numbers. You have been shown how to use the Laspeyres', Paasche's and Fisher's formulae for calculating price as well as quantity indices. You also know now how to measure changes in consumer price or cost of living.

6.9 ANSWERS AND HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) and (2): Do it yourself.

3) Simple Aggregative Index Number = 117.14

Average of Price Relative Method = 122.9

4) 84.2

5) Laspeyres' Index Number = 86.02

Paasche's Index Number = 81.25

Check Your Progress 2

1) 108.33. 135.41, 160.23, 165.56

2) Do it yourself

Check Your Progress 3

1) We find quantity for Oct. 2020 with Oct. 2019 as base. The required index may be obtained as the weighted arithmetic mean of quantity relatives, using the receipts in 2019 as weights.

Type of traffic	q_0	q_n *(w)*	**Weight**	**Quantity** $(P_n \div q_0)\times 100$	**(4)×(5)**
(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Merchandise	1246	1206	776	97	75272
Minerals	1125	981	252	87	21924
Fuel	4794	4229	562	88	49456
Total		**1590**			**146652**

Quantity index = $\frac{\Sigma\left(q_n/q_0\right)\times 100\times w}{\Sigma w} = \frac{146652}{1590} = 92$

2) Calculation for Paasch's price index

Commodity	P_0	P_n	q_0	q_n	$P_0 q_n$	$P_n q_n$
A	4	5	95	120	480	600
B	60	70	118	130	7800	-----
C	35	40	50	70	2450	2800
Total					**1590**	**146652**

Paasche's Price index $= \frac{\sum P_n q_n}{\sum P_0 q_n} \times 100 = \frac{12500}{10730} \times 100 = 116$

3) We are given the base price(P_0), current price (P_n) and value in the base year ($P_0 q_0$). To find base year quantity (q_0), we can use the relation

$$q_0 = \frac{P_0 q_0}{P_0}$$

Using P_0, P_n and q_0, we can find Laspeyres' index as

$$L = \frac{\sum P_n q_0}{\sum P_0 q_0} \times 100$$

Calculation for Laspeyres' price index

Item	P_0	P_n	$P_0 q_0$	P_n	$P_n q_0$
A	12.50	14.00	112.50	9	126.00
B	10.50	12.00	126.00	12	144.00
C	15.00	14.00	105.00 70	7	98.00
D	9.40	11.20	47.00	5	56.00
Total			**390.50**		**424.00**

Laspeyres' price index $= \frac{\sum P_n q_0}{P_0 q_0} \times 100 = \frac{42.00}{390.0} \times 100 = 109.$

4) Marshall-Edgeworth index $= \frac{\sum P_n (q_0 + q_n)}{\sum P_0 (q_0 + q_n)} \times 100$

$$= \frac{\sum P_n q_0 + \sum P_n q_n}{\sum P_0 q_0 + \sum P_0 q_n} \times 100$$

Let us take 2010 as base and 2015 as current year.

Commodity	P_0	q_0	P_n	q_n	$P_0 q_0$	$P_0 q_n$	$P_n q_0$	$P_n q_n$
A	9.3	100	4.5	90	930.0	837.0	450.0	405.0
B	6.4	11	3.7	10	70.4	64.0	40.7	37.0
C	5.1	5	2.7	3	25.5	15.3	13.5	8.1
Total					**1025.9**	**916.3**	**504.2**	**450.1**

Required Index $= \frac{504.2 + 450.1}{1025.9 + 916.3} \times 100 = 49.1.$

UNIT 7 DETERMINISTIC TIME SERIES AND FORECASTING*

Structure

7.0 OBJECTIVES

After going through this Unit, you will be able to

- construct a trend line for time series data;
- compute moving averages; and
- calculate various measures of seasonal variations.

7.1 INTRODUCTION

A time series is a set of observations on a variable measured at successive points of time. Usually the variable values are recorded over equal time intervals-yearly, quarterly monthly, etc. Generally the term 'time series' refers to economic data, but it equally applies to quantitative data collected in other fields also. The time series of National Income, Agricultural Income, and Agricultural Production are based on annual observations.

* Adapted from IGNOU study material of EEC 13: Elementary Statistical Methods and Survey Techniques, Unit 11 written by S Bandopadhyay with modifications by K. Barik

Other examples of time series are yield of a crop in different years, population of a country over different points of time, sales of a departmental store during different seasons of the year, quarterly exports of tea, etc. For these types of data one of the variables is time, denoted by t_j and the other which is dependent on time (such as yield, population, sale or export) is represented by y_t. We will analyse some of these series with the help of the methodology to be developed in this Unit.

7.2 PROBLEMS AND OBJECTS OF STUDY OF TIME SERIES DATA

A study of time series data reveals that in general the observed values of the dependent variable (y_t) change over time. These changes are due to interaction of several forces such as increase in population, change in production techniques, change of tastes and habits of people, variations in climate, etc. Part of these changes is long term while others may be seasonal or cyclical. One of the main objectives of study of time series data is to isolate and measure the effects of various components. This analysis helps us in understanding the past behavior and predicting the future. Such predication is of utmost importance to an economist or a producer who can plan his production much ahead of sales.

7.2.1 Components of Time Series

A graphical representation of time series data reveals changes over time (in rather exceptional cases the series exhibits no change during the period of observation). However, these changes are not totally haphazard or random and at least a part of it can be explained. Some of the movements are periodic in nature while some others show persistent growth or decline. Along with these some unpredictable movements, random in nature, are also found to be mixed up. Again, not every series shows all the movements. It is assumed that the general series has four important components.

i) Secular or long-term trend(T)

ii) Seasonal variation (S)

iii) Cyclical fluctuation (C)

iv) Irregular or random movement (I)

In the classical approach, it is assumed that the observed value y_t may be represented either as the product of the above components

i.e., $y_t = T \times S \times C \times I$ (multiplicative model)
or, as the sum of components
$y_t = T + S + C + I$ (additive model)

Although the additive model facilitates easier calculation, the multiplicative model has been most widely used in analysis of time series.

a) Secular Trend

By secular trend we mean the smooth, regular, long-term changes in the series when observed over a period of time. Some series may exhibit an upward trend, some series a downward trend while some others may remain more or less constant over time. The upward trend of a series may be caused by factors such as increase in population and improvement in techniques of production. For example, the pattern of growth of many industries follows closely that of population growth of the country. Again the advances in technology may give rise to upward movement of most of the economics time series. But not all time series will exhibit growth. Some may show decline while some others may show fluctuations. The time series of crude death rates of a country is likely to show a declining trend.

b) Seasonal Variations

The graphs of most of the time series reveal that a large number of fluctuations are imposed on the trend. By seasonal variation we mean the periodic movement in a time series where the period is not longer than one year. A periodic movement is that which repeats at regular intervals or periods of time. For example, the sales of cold drinks increase during summer and decrease during winter, sales of garments are maximum during some seasons of the year, say during May or festivals, the number of passengers carried by buses has a peak during office hours, the number of books borrowed from a library has a peak during some days of the week, etc. The factors which contribute to this type of fluctuations are the climatic changes of different seasons, customs and habits which people follow at different times.

C) Cyclical Fluctuations

By cyclical fluctuations we mean oscillatory movements of a time series, where the period of oscillation, called the cycle, is more than a year. It includes those factors leading to alternating periods of expansion and contraction that characterize most economic and business series. Sometimes these fluctuations are highly irregular with respect to their Sometimes these fluctuations are highly irregular with respect to their shape, amplitude and direction. But the phenomena they reflect – the periods of depression, recovery, boom and collapse-have been observed in virtually all time series dealing with business and economics data.

d) Irregular Movement

The irregular movement includes component all factors not classifiable elsewhere. Thus factors such as work stoppage, elections, wars, fire may affect a particular time series; this category of movement includes all types of variations not accounted for by secular trend, seasonal or cyclical fluctuations. Unfortunately, factors of these kinds are frequently indistinguishable from

cyclical factors and as such in some discussions cyclical and irregular components are combined together.

7.2.2 Construction of Time Series: An Example

As an illustration we prepare a time series according to the multiplicative model. Table 7.1 presents trend, seasonal and cyclical-irregular components of a hypothetical series.

Table 7.1: Hypothetical Time Series and its Components (Quarterly)

Year	Components				
	Quarter	Series (y_t)	Trend (T)	Seasonal ($100S$)	Cyclical- Irregular ($100CI$)
1	I	79	80	120	82
	II	58	85	80	85
	III	84	90	92	102
	IV	107	95	108	105
2	I	130	100	120	108
	II	93	105	80	132
	III	121	110	92	120
	IV	161	115	108	130
3	I	216	120	120	150
	II	132	125	80	132
	III	150	130	92	125
	IV	163	135	108	112
4	I	176	140	120	105
	II	112	145	80	97
	III	128	150	92	93
	IV	142	155	108	85
5	I	134	160	120	70
	II	86	165	80	65
	III	94	170	92	60
	IV	104	175	108	55

In Table 7.1 the series is represented by a multiplicative model, such that

$$y_t = T \times S \times C \times I.$$

Thus the observation 79 (of I quarter of 1st year) $= 80 \times \frac{120}{100} \times \frac{82}{100}$.

Similarly, 112 (of II quarter of 4th year) $= 145 \times \frac{80}{100} \times \frac{97}{100}$.

Thus, each quarterly figure (y_t) is the product of the secular trend (T), the seasonal index (S), cyclical and the irregular component (CI). Such a synthetic composition looks very much like an actual time series and has encouraged use of the model as the basis for the analysis of time series data.

7.3 MEASUREMENT OF TREND

At times we are interested to know the trend movement in a time series. In such circumstances, we have to eliminate the effects of other components (seasonal, cyclical and irregular) from the series.

Two important methods of measuring trend are the 'Moving Averages Method', and the 'Method of Fitting Polynomials'. In moving average method, secular trend is obtained by smoothing out fluctuations by the process of averaging. In the latter, a polynomial of suitable degree is chosen either for the original variables or for its transformed variable and its constants are determined by the method of least squares. The choice of the degree of the polynomial can be made by plotting the data on a graph paper where different scales, arithmetic, semi-logarithmic or double-logarithmic scales may be used. Measurement of trend is necessary for studying the behavior of the time series and for forecasting the future.

7.3.1 Moving Averages Method

This is a simple method of smoothing out fluctuations of a series by calculating a number of averages covering overlapping periods of the series. The first step consists in selecting proper period of the moving average. If the period chosen is 3 years, the moving averages are obtained by calculating a series of mean values of three consecutive values covering overlapping periods of the series. Denoting the original series by y_1, y_2, y_3,….., the mean of the first three values, given by $(y_1 + y_2 + y_3)/3$, is placed at the midpoint of the period covering first three years. This is the first moving average value. The second moving average value is obtained by calculating the mean of the values covering the period from second to fourth year. This is given by $(y_2 + y_3 + y_4)/3$ and is placed at the mid-p0int of the period covering second to fourth year. This process is repeated. It is clear that some of the values for the years at the beginning as well as at the end cannot be obtained by this method.

Two cases may be distinguished, viz., when the period of moving average is odd and when it is even. If the period is odd (for example, if the period is three years), the first moving average is placed at the second year, the second moving average is placed at the third year and so on.

If, however, the period is even (for example four years), the moving average value fall between two consecutive years and 'centering' is necessary for getting trend values for various years.

As an illustration, let us consider a schematic representation for the calculation of centered 4-year moving averages. Here we will present two methods- the direct method (Table 7.2), as well as the short-cut method (Table 7.3).

Table 7.2 Calculation of centered 4-year moving averages (Direct Method)

Year	y_1	4-year Moving Total	4-year Moving Average	Centered Moving Total	Centered 4-year moving
(1)	(2)	(3)	(4)	(5)	(6)
1	y_1			-	-
2	y_2			-	-
		$y_1 + y_1 + y_1 + y_1 = T_1$	$T_1 / 4$		
3	y_3			$(T_1 + T_2)/4$	$(T_1 + T_2)/8$
		$y_2 + y_3 + y_4 + y_5 = T_2$	$T_2 / 4$		
4	y_4			$(T_2 + T_3)/4$	$(T_2 + T_3)/8$
		$y_3 + y_4 + y_5 + y_6 = T_3$	$T_3 / 4$		
5	y_5			$(T_3 + T_4)/4$	$(T_3 + T_4)/8$
		$y_4 + y_5 + y_6 + y_7 = T_4$	$T_4 / 4$		
6	y_6			-	-
7	y_7			-	-

In the above illustration, the period of moving averages is 4 years. Both in the direct and in the short-cut method col. 3 shows the 4-year moving totals. The first value (T_1) is placed between the second and the third year, the second moving total (T_2) is placed between the third and the fourth year and so on. The centered 4-year moving averages are placed at the third year, fourth year, by taking a further 2 item moving average in the direct method (Table 7.2.) In the short cut method (Table 7.3), the calculation of the 4-year moving average is omitted (as shown in col. 4 of Table 7.2 in the direct method). Instead, the 2-item moving totals of the 4-year moving averages are obtained (col. 4 and 5).

You should note that for a 4-year moving average, the procedure for centering leaves out 4/2 = 2 years at the end of the series each.

Table 7.3 Calculation of centered 4-year moving averages (Short Method)

Year	y_1	4-year Moving Total (M.T.)	2-item moving total (M.T.)	Centered 4-year moving average (M.A.)
(1)	(2)	(3)	(4)	(5)
1	y_1		-	-
2	y_2		-	-
		$y_1 + y_2 + y_3 + y_4 = T_1$		
3	y_3		$(T_1 + T_2)$	$(T_1 + T_2)/8$
		$y_2 + y_3 + y_4 + y_5 = T_2$		
4	y_4		$(T_2 + T_3)$	$(T_2 + T_3)/8$
		$y_3 + y_4 + y_5 + y_6 = T_3$		
5	y_5		$(T_3 + T_4)$	$(T_3 + T_4)/8$
		$y_4 + y_5 + y_6 + y_7 = T_4$		
6	y_6		-	-
7	y_7		-	-

7.3.2 Suitability of Moving Averages

Moving average method is simple to apply but the success of this method depends on the proper choice of the period. Moving average with a period exactly equal to or a multiple of the period exactly equal to or a multiple of the period of the cycle present in the series will completely eliminate the cyclical component and give an estimate of the trend. This method is flexible but some trend values at the beginning and at the end of the series have to be left out and their number increases with increase in the period of the moving average. Again as moving average assumes no law of change; the method cannot be used for forecasting future trend.

7.3.3 Examples of Moving Averages

Example 7.3.1: Calculate the three and five year moving averages of the following data:

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Production ('000 tons)	18	19	20	22	20	19	22	24	25	24	25	26

Steps of Calculation

1) In Table 7.3.1 the figures in col. 3 are obtained as the sum of three consecutive values of col. 2. Thus the first moving total (M.T.) is 57 = 18 + 19 + 20 and is placed against 2001. The second moving total 61 = 19 + 20 + 22 is placed against 2002.

2) The three-year moving average (M.A.) in col. 4 is obtained by dividing the corresponding three-year moving total in col. 3 by 3, the period of the moving average. Thus 57 ÷ 3 = 19, 61 ÷ 3 = 20.3, etc.

3) The five-year moving totals in col.5 are obtained as the sum of five consecutive values in col.2. Thus the first moving total against the year 2002 is 99 = 18 + 19 + 20 + 22 + 20.

4) The five-year moving total in col. 5 by 5. Thus moving average for 2005 is 107 ÷ 5 = 21.4.

Table 7.3.1: Calculation of (I) 3-year Average (II) 5-year Moving Average

Year	Production	3-year M.T.	3-year M.A.	5-year M.T.	5-year M.A.
(1)	(2)	(3)	(4)	(5)	(6)
2000	18	-	-	-	-
2001	19	57	19.0	-	-
2002	20	61	20.3	99	19.8
2003	22	62	20.77	100	20.6
2004	20	61	20.3	103	20.6
2005	19	61	20.3	107	21.4
2006	22	65	21.7	110	22.0
2007	24	71	23.7	114	22.8
2008	25	73	24.3	120	24.0
2009	22	74	24.7	124	24.8
2010	25	75	25.0	-	-
2011	26	-	-	-	-

Note that for 3-year centered moving averages $\frac{3-1}{2} = 1$ year, and for 5-year centered moving averages $\frac{5-1}{2} = 2$ years, respectively, are left out both at the beginning and the end of the series.

Example 7.3.2: Compute trend values for the following time series using 4-yearly moving averages.

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Yield (qntls.)	52	54	55	57	58	61	63	66	67	70

Solution:

Table 7.3.2(a): Calculation of 4-year moving average (Direct Method)

Year	Yield	4-year M.T.	3-year M.A.	2 item M.T. of col.4(centered)	(Centered) 4-year M.A.
(1)	(2)	(3)	(4)	(5)	(6)
2009	52	-	-	-	-
2010	54	218	54.50	-	-
2011	55	224	56.00	110.50	55.250
2012	57	231	57.75	113.75	56.875
2013	58	239	59.75	1177.50	58.750
2014	61	248	62.00	121.75	60.875
2015	63	257	64.25	126.25	63.125
2016	66	266	66.50	130.75	65.375
2017	67	-	-	-	-
2018	70	-	-	-	-

Table 7.3.2(b): Calculation of 4-year Moving Average (Shortcut Method)

Year	Yield	4-year M.T.	2-item M.A	(Centered) 4-year M.A
(1)	(2)	(3)	(4)	(5)
2009	52	-	-	-
2010	54	218	-	-
2011	55	224	442	55.250
2012	57	231	455	56.875
2013	58	239	470	58.750
2014	61	248	487	60.875
2015	63	257	505	63.125
2016	66	266	523	65.375
2017	67	-	-	-
2018	70	-	-	-

Steps of Calculation (Direct Method) – see Table 7.3.2(a)

1) Col. 3 is the sum of four consecutive values in col.2

   Thus, 52 + 54 + 55 + 57 =218, 54 + 55 + 57 + 58 = 224.

2) Col. 4 = col. 3÷4. Thus 218 ÷ 4 = 54.5, 224 ÷ 4 = 56.

3) Col. 5 = Sum of two consecutive values in col.4.

   Thus, 54.5 + 56.0 = 110.5, 5.00 + 57.75 = 113.75.

4) Col. 6 = col. 5÷2. Thus 110.5÷2 = 55.25.

Steps of calculation (shortcut method) – see Table 7.3.2 (b)

1) Col. 4 is the sum of two consecutive values in col.3.

   Thus, 218 + 224 = 442, 224 + 231 = 455.

2) Col.5 = col.4 ÷ 8. Thus 442 ÷ 8 = 55.25

Example 7.3.3

Find trend values for the following series using a 3-year weighted moving average with weights 1, 2, 1.

Year	1	2	3	4	5	6
Value	2	3	5	6	8	11

Solution:

Table 7.3.3: Calculation of 3-year weighted moving average

Year	Value	3-year weighted moving total (M.T.)	3-year weighted moving average (M.A)
(1)	(2)	(3)	(4)
1	2	-	-
2	3	13	3.25
3	5	19	4.75
4	6	25	6.25
5	8	33	8.25
6	11	-	-

Step of calculation

1) Col. 3 figures are the weighted moving totals of col.2 figures with weights 1, 2, 1.

Thus $1 \times 2 + 2 \times 3 + 1 \times 5 = 13$

$1 \times 3 + 2 \times 5 + 1 \times 6 = 19$

2) Col.4 = col.3 ÷ (sum of weights, i.e., 4)

Thus $13 \div 4 = 3.25$, $19 \div 4 = 4.75$

Example 7.3.4: Calculate the 4-quarter moving average for the following time series data

Quarter	**Year**			
	2015	2016	2017	2018
1	62	66	72	79
2	58	60	67	74
3	72	74	80	88
4	60	64	69	77

Solution:

Year	Quarter	Value	4-quarter (M.T.)	Centered (M.T.)	4-quarter (M.A.)
(1)	(2)	(3)	(4)	(5)	(6)
2015	1	62	-	-	-
	2	58	-	-	-
			259	-	-
	3	72		508	63.50
			256		
	4	60		514	64.25
			258		
2016	1	66		518	64.75
			260		
	2	60		524	65.50
			264		
	3	74		534	66.75
			270		
	4	64		547	68.38
			277		
2017	1	72		560	70.00
			283		
	2	67		571	71.38
			288		
	3	80		583	72.88
			295		
	4	69		597	74.63
			302		
2018	1	79		612	76.50
			310		
	2	74		628	78.50
			318	-	-
	3	88	-	-	-
	4	77	-	-	-

Check Your Progress 1

1) Given below is data on index of production for the period 2011 to 2020.

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Index of Production	109.2	119.8	129.7	140.8	153.8	152.2	152.6	163.0	175.3	184.3

1) Fit the trend line and predict the index of production for the year 2012 by 3-year moving averages method.

……...…

……...……..….

……..……..…

……...……

……...……

2) Bring out the advantages and disadvantages of moving averages method.

……...…

……...……..….

……..……..…

……...……

……...……

7.4 METHOD OF FITTING POLYNOMIALS

Method of fitting polynomials is perhaps the best and the most objective method of determining trend. Here an appropriate type of polynomial is selected for trend and the constants appearing in the trend equation are determined on the basis of the given time series data.

The choice of an appropriate polynomial is facilitated by a graphical representation of the data for which, apart from the usual arithmetic scales, semi-logarithmic or doubly-logarithmic scales may be used. If the plotted data show approximately a straight line tendency on an ordinary graph paper, the used is $Y = a + bx$ (straight line or first degree polynomial).

If they show a straight line on a semi-logarithmic graph paper, the equation is log $Y = a + bx$, which is obtained by taking logarithm of $Y = A.B^x$ (exponential function). Note that $a = \log A$ and $b = \log b$.

Some times a second or a third degree polynomial may also be fitted.

$Y = a + bx + cx^2$ (second degree polynomial or parabola)

$Y = a + bx + cx^2 + dx^3$ (third degree polynomial)

The constants appearing in the above equations (such as *a,b,c…*) are obtained by applying the principle of "least squares", as in regressions (see unit 5). This states that the values of the constants will be such as to make the sum of squares of the deviations

$\sum(y - Y)^2$ minimum,

where y = observed value,

Y = expected value obtained from the trend equation of the type

$Y = a + bx$

Or,

$Y = a + bx + cx^2$ etc., and the summation is taken over all the observations.

In the case of straight line fitted by the method of "least squares", the constants a and b are determined from the following normal equations:

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

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

where n is the number of years covered.

Similarly in the case of parabola or second degree polynomial the constants *a, b* and *c* are determined from the three normal equations

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

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

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

Rule for writing down the normal equations

To get the first normal equation, multiply each observation by coefficient of a in that equation and take sum over all the n observations.

Thus for straight line $y = a + bx$, as the coefficient of a is 1, the first normal equation is $\sum y = na + b\sum x$.

For the second normal equation, multiply each observation by the coefficient of b in that equation and take sum over all the n observations. In the case of straight line, coefficient of b is x. So, the second normal equation is $\sum xy = a\sum x + b\sum x^2$.

Now we will consider trend fitting for periods covering odd (Table 7.4) and even (Table 7.5) number of years taking the first degree polynomial.

Case I: Odd number of years (n = 5)

Table 7.4

Year	y	x	x^2	xy
(1)	(2)	(3)	(4)	(5)
1	y_1	-2	4	
2	y_2	-1	1	
3	y_3	0	0	
4	y_4	1	1	
5	y_5	2	4	
total	Sy	0	10	Sxy

The normal equations are:

$\sum y = 5a + b\sum x = 5a$ (Since $\sum x = 0$)

$\sum xy = a\sum x + b\sum x^2 = 10b$

$\therefore a = \frac{\sum y}{5},\ b = \frac{\sum xy}{10}$

where the origin (x = 0) is taken at the midpoint of the interval covered by 5 years, i.e., at the third year and unit of time = 1 year. In real life situations the actual values of y'_1 s are considered. Hence $\sum x$ and $\sum xy$ will be known.

Case II: Even number of years (n = 6)

Table 7.5

Year	y	x	x^2	xy
(1)	(2)	(3)	(4)	(5)
1	y_1	-5	25	
2	y_2	-3	9	
3	y_3	-1	1	
4	y_4	1	1	
5	y_5	3	9	
6	Y_6	5	25	
total	$\sum y$	0	70	$\sum xy$

The constants a and b will be obtained from the following equations

$\sum x = 6a$

$\sum xy = 70b$

Here the origin (x = 0) will be in the middle of 3rd and 4th year and the unit of x =6 months.

7.4.1 Suitability of Least Squares Method

Trend lines are used for description of the growth or decline of the time series and as an aid to the study of the long-term trend of the economy. The method of fitting polynomials completely eliminates personal bias and trend values for all the given periods can be obtained. This is, however, not possible with moving average method. But the choice of the type of the polynomial curve is arbitrary and one cannot be sure whether a linear or parabolic curve will represent the trend best. The choice of the trend equation may itself lead to a bias. It is, however, possible to get some idea of the pattern of trend from the scatter diagram of the data.

7.4.2 Examples of Least Squares Method

Example 7.4.1

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

1) The data given below give the index of industrial production from 1961 to 1970

years	2005	2006	2007	2008	2009	2010	2011
Production	81	92	100	105	112	120	120

Estimate the production for 2012.

Solution:

Here the number of years is odd (n = 7) Let $y = a+bx$ be the equation of the straight line trend with origin (x = 0) at 2008 and one unit of x = 1 year. The least squares normal equations are (see unit 5)

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

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

Thus, substituting the values of $\sum y, \sum x$ y, $\sum x$ and $\sum x^2$ from the above table in the normal equations, we get

$7a = 736$, so $a = 105.1$

$28b = 203$, so $b = 7.21$

The trend equation is

$Y = 105.1+7.21x$, with origin at 2008 and unit of x =1 year. The value of x for 2012 would be 4.

Table 7.4.1: Fitting straight Line Trend

Year	Production (y)	x	x^2	xy
(1)	**(2)**	**(3)**	**(4)**	**(5)**
2005	81	-3	9	-243
2006	92	-2	4	-184
2007	100	-1	1	-100
2008	105	0	0	0
2009	112	1	1	112
2010	120	2	4	240
2011	126	3	9	378
Total	736	0	28	203

Hence, using the trend equation the estimate for 2012 is $Y = 105.1 + 4 \times 7.21 = 133.94$.

Example: 7.4.2

Fit a straight line trend to the following time series data:

Year	2010	2011	2012	2013	2014	2015
Profits:	3.1	3.3	3.6	3.2	3.7	3.9

Estimate the profit for 2016.

Solution:

Here the number of years is even ($n = 6$). Let $y = a+bx$ be the trend equation with origin.

($n = 0$) mid-way between 2012 and 2013 and unit of $x = 6$ months.

The normal equations are

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

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

Table 7.4.2: Fitting straight Line Trend

Year	**Profit (*y*) (Rs. lakhs)**	x	x^2	*xy*
(1)	(2)	(3)	(4)	(5)
2010	3.1	-5	25	-15.5
2011	3.3	-3	9	-9.9
2012	3.6	-1	1	-3.6
2013	3.2	1	1	0.0
2014	3.7	3	9	11.1
2015	3.9	5	25	19.5
Total	**20.8**	**0**	**70**	**4.8**

So, substituting the values of $\sum y, \sum xy, \sum x,$ and $\sum x^2$ from the above table in the normal equations, we get

$6a = 20.8$, or $a = 3.47$

$70b = 4.8$, or $b = 0.07$

The trend equation is

$Y = 3.47 + 0.07x$, with origin at the middle of 2012 and 2013 and unit of $x = 6$ months.

For 2016, x would be 7.

So, estimate for 2016 is

$Y = 3.47 + 0.07 \times 7 = 3.47 + 0.49 = 3.96$

Hence the estimated profit for 2016 is Rs. 3.96 lakhs.

Example: 7.4.3

The sales of a company (in thousands of rupees) for the year 2010 to 2016 are given in the following table. Fit an exponential trend ($Y = A,B^x$) and estimate the sales for 2017.

Year	2010	2011	2012	2013	2014	2015	2016
Sales	32	47	65	92	132	190	275

Solution:

Here the number of years is odd ($n = 7$). Taking log of both sides of the given equation, we can write $\log Y = \log A + x \log B$. Let $a = \log A$ and $b = \log B$. Thus we can write

$\log Y = a + bx$.

Further, we take origin (x = 0) at 2013 and one unit of x = 1 year. The least squares normal equations are:

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

$$\sum x\log y = a\sum x + b\sum x^2$$

Table 7.4.3: Fitting Straight Line Trend

Year	**Sales (y)**	x	x^2	**log$y$**	**x. log$y$**
2010	32	-3	9	1.5051	4.5153
2011	47	-2	4	1.6721	3.3442
2012	65	-1	1	1.8129	1.8129
2013	92	0	0	1.9638	0
2014	132	1	1	2.1206	2.1206
2015	190	2	4	2.2788	4.5576
2016	275	3	9	2.4398	7.3119
Total	**833**	**0**	**28**	**13.7931**	**4.3237**

So, substituting the values of $\sum \log y, \sum x.\log y, \sum x,$ and $\sum x^2$ from the above table in the normal equations, we get

$7a = 13.7931$, or $a = 1.97$

$28b = 4.3237$, or $b = 0.154$

Thus, the fitted function is log$y$ = 1.97 + 0.154x or Y = antilog(1.97 + 0.154x).

For 2017, x would be 4.

Thus, the estimated value for 2017 is

Y = antilog (1.97+0.154×4) = antilog 2.586 = 385.48.

A case with even number of years can be attempted as in the fitting of a straight line (see Example 7.4.2).

Example: 7.4.4

The following table shows the production of cement in India during 2002 to 2008.

Fit a second degree polynomial to the data.

Year	2012	2013	2014	2015	2016	2017	2018
Production	23.7	27.1	30.2	33.1	36.4	39.3	45.0

Solution:

Here the number of years is odd (n = 7). Let $y = a + bx + cx^2$ be the trend equation with origin (x = 0) at 2015 and unit of x = 1year. The normal equation is:

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

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

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

Table 7.4.4: Fitting Second Degree Polynomial

Year	y	x	x^2	x^3	x^4	xy	x^2y
2012	23.7	-3	9	-27	81	-71.1	213.3
2013	27.1	-2	4	-8	16	-54.2	108.4
2014	30.2	-1	1	-1	1	-30.2	30.2
2015	33.1	0	0	0	0	0.0	0.0
2016	36.4	1	1	1	1	36.4	36.4
2017	39.3	2	4	8	16	78.6	157.2
2018	45.0	3	9	27	81	135.0	405.0
Total	**234.8**	0	28	0	196	**94.5**	**950.5**

Substituting the values from the table in the normal equations

$7a + 28c = 234.8$

$28b = 94.5$

$28a + 196c = 950.5$

Solving these three equations, simultaneously, we get

$a = 33$

$b = 3.37$

$c = 0.134$

Hence, the second degree polynomial is

$Y = 33 + 3.37x + 0.134x^2$,

with origin ($x = 0$) at 2015 and unit of $x = 1$ year.

Example: 7.4.5

Fit a second degree polynomial to the following data. Estimate the trend value for 2012.

Year	2006	2007	2008	2009	2010	2011
Annual Indian Imports (10^8 Rs.)	23.7	27.1	30.2	33.1	36.4	39.3

Solution:

Here the number of years is even ($n = 6$). Let $y = a + bx + cx^2$ be the trend equation with origin ($x = 0$) mid-way between 2008 and 2009 and unit of $x = 6$ months. The normal equations are $\sum y = na + b\sum x + c\sum x^2$

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

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

Table 7.4.5: Fitting Second Degree Polynomial

Years	y	x	x^2	x^3	x^4	xy	x^2y
2006	507	-5	25	-125	625	-2535	12675
2007	602	-3	9	-27	81	-1806	5418
2008	681	-1	1	-1	1	-681	681
2009	914	1	1	1	1	914	914
2010	1255	3	9	27	81	3765	11295
2011	1361	5	25	125	625	6805	34025
Total	5320	0	70	0	1414	6462	65008

Substituting the values from the table in the normal equations

$6a + 70c = 5320$

$70b = 6462$

$70a + 1414c = 65008$

Solving the above three equations simultaneously, we get

$a = 829.2$, $b = 92.31$ and $c = 4.924$.

The second degree polynomials is

$Y = 829.2 + 92.31x + 4.924x^2$,

with origin ($x = 0$) mid-way between 2008 and 2009 and unit of $x = 6$ months.

For 2012, x would be 7.

Therefore, estimate for 2012 is

$Y = 829.2 + 92.31 \times 7 = 3.47 + 4.924 \times (7)^2$

$= 829.2+646.17+241.28 = 1716.65.$

7.5 MONTHLY OR QUATRELY TREND VALUES FROM ANNUAL DATA

In a time series, annual data may be available in different forms such as (1) monthly or quarterly averages fo02r each year, and (ii) annual totals.

If the trend equation is fitted to the monthly or quarterly data, there will be no difficulty in obtaining monthly or quarterly values. However, it is not advisable to fit a trend line by the method of least square to the monthly or quarterly data. The trend line is usually fitted to the annual data and then the trend values may be obtained for different months (or quarters). The method of obtaining monthly (or quarterly) trend equation is discussed below.

Let $y = a+bx$ be an annual trend equation. If we divide both sides of this equation by 12, we get a monthly average equation. Thus $\frac{y}{12} = \frac{a}{12} + \frac{b}{12}x$ is a monthly average equation. Denoting by $= \frac{y}{12}, A = \frac{a}{12}$ and $B = \frac{b}{12}$, we can write the monthly average equation as $Y = A+Bx$. Here, Y is monthly and average and B denotes change in monthly average per unit change in x,i.e., 1 year.

To get a monthly equation, we have to determine the corresponding rate of change of Y. since B is the average monthly change in Y per year, $\frac{B}{12}$ will denote average change per month. Thus, the monthly equation can be written as $Y = A + \frac{B}{12}x$ or $Y = \frac{a}{12} + \frac{b}{144}x$, where x denotes month rather than year.

Similarly, we can write $Y = \frac{a}{4} + \frac{b}{4}x$ as the quarterly average equation, where a unit of x denotes one year, and $Y = \frac{a}{4} + \frac{b}{16}x$ as the quarterly average equation, where a unit of x denotes one quarter.

Shifting of Origin

We define a, in the equation $Y = a + bx$, as the value of trend in the year of origin. Thus, with the shifting of origin the value of a changes. Let us assume that the year of origin. (i.e., $x = 0$) is 2015 and we want to change it to 2018.

We note that $x = 3$ for 2018, therefore, trend for 2018 = $a + 3b$. Treating this as constant term in the trend equation, $y = (a + 3b) + bx$ becomes the new trend equation with 2018 as origin.

Example 7.5.1

The trend equation for certain production data is $Y = 150 + 24x$ (y = annual production in thousand tons and x = time with origin at 2008, unit of x = 1 year). Estimate the trend value for May 2013.

Solution: The monthly trend equation is

$$Y = \frac{150}{12} + \frac{24}{144}x = 12.5 + 0.167x,$$

where Y = monthly production, unit of x = 1 month and origin at 2008, i.e., 30th June 2008.

To estimate the trend for May 2013, we substitute $x = 58.5$ in the above equation. Thus, we get $Y = 12.5 + 0.167 \times 58.5 = 22.25$ ('000 tons)

Example 7.5.2

The trend equation fitted to quarterly average sales for 7 years is given by $y = 250 + 20x$ (unit of x = 1 year, origin = 30th June 2010). Estimate the trend value for the first quarter of 2013 (January-March).

Solution: Here the quarterly average refers to average per quarter for each year. The quarterly trend equation is given by $Y = 250 + \frac{20}{4}x$, where Y = quarterly sales, x = 1 quarter and origin at 30th June 2010.

The interval between 30th June 2010 and the 1st quarter of 2013 are 10.5 quarters. Thus, to obtain the trend for 1st quarter of 2013, we substitute $x = 10.5$ in the above equation.

Hence, the required trend is $Y = 250 + 5 \times 10.5 = 302.5$.

Check Your Progress 2

1) Fit a straight line trend to the following data and show how to obtain the monthly trend values from the trend line fitted to the given time series. Obtain two such monthly values.

Year	2010	2011	2012	2013	2014
Monthly Production: (in '000 tons)	38	40	41	45	47

……

……

……

……

……

……

……

2) The trend equation for certain production data is $y = 240 + 48x$ (y = annual production in tons, x = time with origin at 2010, unit of x = 1 year). Estimate the trend for October 2016.

……...…

……...……..….

……..……..…

……...……

……...……

3) The trend equation fitted to quarterly average sales data is given by $y = 60 + 8x$ (unit of x = 1year, origin = 30th June, 2018). Estimate the trend value for first quarter (Jan-March.) of 2020.

……...…

……...……..….

……..……..…

……...……

……...……

……...……

7.6 MEASUREMENT OF SEASONAL VARIATIONS

There are a number of methods of measuring seasonal variations in time series, depending on how the other components such as cyclical, trend and irregular movements are present in it. For simplicity, we shall consider seasonal variations in monthly or quarterly data only, but the procedure for weekly or daily data will be similar. It may be mentioned here that annual data contains no seasonal variation. The application of the methods, to be discussed below, will give us 4(or 12) numbers for quarterly (or monthly) data. These will be termed as seasonal indices and are normally expressed as percentages. A figure of a particular quarter (or month) indicates whether that quarter is above or below the normal quarter. For example, a value of 80 for particular quarter indicates that the business for exports or sales (say) during that quarter is slack and it is below the normal quarter by 20%. We will consider only the multiplicative model for the measurement of seasonal variations. The main methods for the measurement of seasonal variation are given below.

1) Method of Simple Average

2) Ratio to Trend Method

3) Ratio to Moving Average Method

7.6.1 Method of Simple Average

This method assumes that the time series variable y is made up of only two components, viz., seasonal (S) and irregular or random component (I). Thus, we can write

$Y = S.I$

When we take average of y values for each month or quarter of all the years, the irregular component gets eliminated and we are left with seasonal component. We will illustrate this method in Table 7.6.

Table 7.6 Illustration of the Method of Simple Average

Years	Quarters			
	I	II	II	IV
1	y_1	Y_2	Y_3	Y_4
2	Y_5	Y_6	Y_7	Y_8
3	Y_9	y_{10}	y_{11}	y_{12}
4	y_{13}	y_{14}	y_{15}	y_{16}
5	y_{17}	y_{18}	y_{19}	Y_{20}
Total	T_1	T_2	T_3	T_4
Average	A_1	A_2	A_3	A_4
S.I.	s_1	s_2	s_3	s_4
S.I. (adjusted)	S_1	S_2	S_3	S_4

Explanatory notes:

a) $T_1 = y_1 + y_5 + y_9 + y_{13} + y_{17}$ is the total of y values of first quarter of each year. Similarly, T_2, T_3 and T_4, are the totals of second, third and fourth quarters of each year respectively.

b) A_i is the i^{th} quarter average $= \frac{T_i}{n}$, where i = 1, 2, 3, 4, and n denotes the number of years.

c) G is defined as the grand average $= \frac{\sum A_i}{4}$.

d) $s_i = \frac{A_i}{G} \times 100, i = 1, 2, 3, 4.$

e) $s = s_1 + s_2 + s_3 + s_4$

f) S_1, S_2, S_3, and S_4, are the seasonal indices for the first, second, third and the fourth quarters respectively, where $S_i = \frac{s_i}{s} \times 400,\ i = 1, 2, 3, 4$. Note that the sum of these 4 index numbers must be equal to 400. Further, $S_i = s_i$ if s = 400.

g) For a time series with monthly data, the sum of 12 seasonal indices, one for each month, must be equal to 1200.

Example: 7.6.1

Compute seasonal indices for the following data by the Method of Simple Average.

Years	Quarters			
	I	II	III	IV
1992	72	68	80	70
1993	76	70	82	74
1994	74	66	84	80
1995	76	74	84	78
1996	78	74	86	82

Solution:

Table 7.6.1: Calculation of Seasonal Indices

Years	Quarters			
	I	II	III	IV
1992	72	68	80	70
1993	76	70	82	74
1994	74	66	84	80
1995	76	74	84	78
1996	78	74	86	82
Total	3776	352	416	384
Average	75.2	70.4	83.2	76.8
S.I.	43	92.15	108.90	100.52

Explanatory Notes:

Grand Average $G = \frac{A_1+A_2+A_3+A_4}{4} = \frac{75.2+7\ .4+83.2+7\ .8}{4} = 76.4$

Seasonal Index for Quarter I, i.e., $S_1 = 98.43$

Seasonal Index for Quarter II, i.e., $S_2 = 92.15$

Seasonal Index for Quarter III, i.e., $S_3 = 108.90$

Seasonal Index for Quarter IV, i.e., $S_4 = 100.52$

Since the sum of these indices = 400, no adjustment is needed.

7.6.2 Ratio to Trend Method

If the data contain trend to an appreciable extent, we first find an appropriate trend equation to determine the trend for various quarters or months. Usually the monthly or quarterly trend values are obtained from the quarterly (or monthly) average trend equation. The trend is then eliminated by expressing the original y values as percentages of the corresponding trend values. This method is based upon the assumption that cyclical variations are either not market or completely absent.

Symbolically, we can write

$$\frac{y}{T} \times 100 = \frac{TSI}{T} \times 100 = SI \times 100$$

From this, the irregular component can be eliminated by the use of Simple Average Method.

Example: 7.6.2

The following table shows the sales (9n'000 Rs.) in a departmental store for five different years. Obtain the seasonal indices by Ratio to Trend Method.

Years	Quarters			
	I	II	III	IV
2000	502	1632	605	362
2001	526	1700	680	390
2002	556	1820	780	422
2003	590	1955	888	464
2004	632	2110	1002	515

Solution:

Let us fit a straight line trend to the data on quarterly averages. The trend equation fitted to quarterly averages $y = a + bx$, where y denotes quarterly average of the year and the unit of x = 1 year. The table below has been constructed from the given data by computing the averages of 4 quarters of each year.

Table 7.6.2(a): Fitting Linear Trend

Years	Quarters			
	y	x	x^2	xy
2000	775	-2	4	-1550
2001	824	-1	4	-824
2002	894	0	0	0
2003	974	1	1	9774
2004	1065	2	4	2130
Total	4532	0	10	730

The normal equations are

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

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

Substituting values from the above table into the normal equations for linear trend, we get

$5a = 4532$ or $a = 906.4$

$10b = 730$ or $b = 73$.

Thus, the quarterly average trend equation is $T = 906.4 + 73x$ (origin: 2002, unit of x = 1 year)

Note that we are using T instead of Y (used earlier in fitting of trends) From this, we can write the quarterly trend equation as

$T = 906.4 + \frac{73}{4}.x = 906.4 + 18.25x$

(origin: 30th June, 2002 unit of x = 1 quarter)

Shifting the origin to 3rd quarter (mid-point of third quarter) of 2002, the quarterly trend equation becomes

$T = 906.4 + 18.25\ (x + 0.) = 906.4 + 9.125 + 18.25x$

$= 9.125 + 18.25x.$

Putting appropriate values of x, we can get the trend values (T) for various quarters. The next step is to express the original values (y) as percentage of trend i.e., $(y \div T)\times 100$, giving "trend ratios". The trend values along with the trend ratios are shown in Table 7.6.2(b).

Table 7.6.2(b): Calculation of Trend Ratios

Year	Quarter	x	$T = 915.525 + 18.25x$	y	$(y \div T)\times 100$
2000	I	-10	733.0	502	68
	II	-9	7751.3	1632	217
	III	-8	769.5	605	779
	IV	-7	787.8	362	46
2001	I	-6	806.0	526	65
	II	-5	824.3	1700	206
	III	-4	842.5	680	81
	IV	-3	860.8	390	45
2002	I	-2	879.0	556	63
	II	-1	897.3	1820	203
	III	0	915.5	780	85
	IV	1	933.8	422	45
2003	I	2	952.0	590	62
	II	3	970.3	1955	201
	III	4	988.5	888	90
	IV	5	1006.8	464	46
2004	I	6	1025.0	632	62
	II	7	1943.3	2110	202
	III	8	1061.5	1002	94
	IV	9	10779.8	515	48

The trend ratios are now arranged by quarters and the seasonal indices are calculated by the method of simple averages.

Table 7.6.2(c) : Calculation of Seasonal Indices

Years	Quarters			
	I	II	III	IV
2000	68	217	79	46
2001	65	206	81	45
2002	63	203	85	45
2003	62	201	90	46
2004	62	202	94	48
Total	320	1029	429	230
Average	64.0	205.8	85.8	46.0
S.I.	63.74	209.98	85.46	45.82

7.6.3 Ratio to Moving Average Method

This method is used when the time series data is composed of all the four components. We know that moving averages, with period equal to the periodic variations, and completely eliminates those variations. Thus, if we take 4 period moving average (M) of quarterly data (or 12 period moving averages of monthly data), the seasonal variations (and some irregular movements) are eliminated from the original (y) values. Further, by expressing y as a percentage to moving average, i.e., ($y \div M$)$\times$ 100, we get values consisting of seasonal and irregular components. The seasonal component is, then, isolated from irregular component by the use of the Method of Simple Average.

Symbolically, we can write

$$\frac{y}{M} \times 100 = \frac{TCSI}{TCI'} = SI"$$

Example: 7.6.3

Use the Ratio to Moving Average Method to calculate seasonal indices for the following data.

Year	Summer	Monsoon	Autumn	Winter
2009	30	81	62	119
2010	33	104	86	171
2011	42	153	99	221
2012	56	172	129	235
2013	67	201	136	302

Solution:

Table 7.6.3: Calculation of Seasonal Indices by Ratio to Moving Average Method

Year	Quarter	y	4-period M.T.	Centered Total	4-period M.A. (M)	$\overline{(y \div M)\times 100}$
2009	Sum	30	-	-	-	-
	Mon	81	--	-	--	-
			292	-	-	-
	Aut	62		597	73.38	84.50
			295			
	Win	119		613	76.63	155.30
			318			
2010	Sum	33		660	82.50	40.00
			342			
	Mon	104		736	92.00	113.04
			394			
	Aut	86		796	99.63	86.32
			403			
	Win	171		855	106.88	160.00
			452			
2011	Sum	42		917	114.63	36.64
			465			
	Mon	153		980	122.50	124.90
			515			
	Aut	99		1044	130.50	75.86
			529			
	Win	221		1077	134.63	164.16
			548			
2012	Sum	56		1126	140.75	39.79
			578			
	Mon	172		1170	146.25	117.61
			592			
	Aut	129		1195	149.38	86.36
			603			
	Win	235		1235	154.38	152.23
			632			
2013	Sum	67		1271	158.88	42.17
			639			
	Mon	201		1345	168.13	119.55
			706	-	-	-
	Aut	136	-	-	-	-
	Win	302	-	-	-	-

The moving ratios are now arranged by quarters and the seasonal indices are calculated by the method of simple averages.

Year	Quarters			
	Summer	Monsoon	Autumn	Winter
2009	-	-	84.50	155.30
2010	40.00	113.04	86.32	160.00
2011	36.64	124.90	75.86	164.16
2012	39.79	117.61	86.36	152.23
2013	42.17	119.55	-	-
Total	158.60	475.10	333.04	631.69
Average	39.65	118.78	83.26	157.92
S.I.	36.69	118.89	83.34	158.08

Check Your Progress 1

1) The following data represent the production of finished steel tins for the years 1992 to 1995:

Production of Finished Steel Tins (000 tons)

Year	Jan.	Feb.	Mar	Apr.	May.	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
1992	420	414	502	365	368	332	390	396	429	417	422	496
1993	491	466	516	337	342	360	409	402	372	391	394	446
1994	463	465	478	310	325	406	415	437	438	445	430	416
1995	502	487	536	404	418	429	489	492	475	456	476	476

Compute the seasonal indices by the Method of Simple Averages.

……...……

...……..……...……..…

……..……..

……..……

…...……...…

…….....….

2) The following table gives the production of steel in India from 1992 to 1995 (in ′000) over the different quarters.

Year	1st quarter	2nd quarter	3rd quarter	4th quarter
1992	1336	1065	1215	1335
1993	1463	1039	1183	1161
1994	1306	1041	1290	1321
1995	1525	1251	1456	1408

Obtain seasonal indices by the method of Ratio to Trend, assuming linear trend.

……………………………………………………………………………..…………
...……………………………………………………………….……….…..…….…
………………………………………………………………………………………..
………………………………………………………………………………..……
…...……………………………………………………………………………...…
………………………………………...…

3) Given the following quarterly sales figures in thousands of rupees for the years 2006 to 2009. Find the specific seasonal by the method of moving averages.

Years	Quarters			
	I	II	III	IV
2006	290	280	285	310
2007	320	305	310	330
2008	340	321	320	340
2009	370	360	362	380

……………………………………………………………………………..…………
...……………………………………………………………….……….…..…….…
………………………………………………………………………………………..
………………………………………………………………………………..……
…...……………………………………………………………………………...…
………………………………………...…

4) The seasonal indices for the sales of garments of a particular type in a certain shop are given below:

Quarter	*Seasonal Index*
Jan-Mar	97
Apr-Jun	85
Jul-Sep	83
Oct-Dec	135

If the total sales in the first quarter of a year are Rs. 15,000, determine how much worth of garments of this type should be kept in stocky by the shop owner to meet the demand for each of the other three quarters of the year?

……...……

...……..…..……..…….…

……..……..

……...……

……...……

7.7 LET US SUM UP

Time series is a set of observation on a variable recorded over time – usually at equal intervals. The change in the variable concerned can be explained to some extent on the basis of components of time series. These components are trend, seasonal variations, cyclical fluctuations and random movements. The observed value of the variable may be represented either as the product of the aforesaid components (multiplicative model) or as the sum of the components (additive model).

Trend can be measured by the method of moving averages and fitting equations by the method of least squares. Once we estimate the trend, we can predict future values and also estimate the monthly or quarterly values from the annual trend.

Seasonal variation can be estimated by three methods: Method of Simple Averages, Ratio to Trend Method and Ratio to Moving Average Method. The method of simple average is used to average out the irregular component. The ratio to trend method can be used if cyclical variations are supposed to be absent while the ratio to moving average method is recommended when the time series variable is composed of all the four components.

7.10 ANSWERS OR HINTS OT CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Read Example 7.3.1 and answer.

2) Read Example 7.3.2 and answer.

Check Your Progress 2

1) $y = 42.2 + 2.3x$, origin: 2012, unit of x = 1 year (monthly average equation)

$y = 42.3 + 0.19x$, origin: July, 2012, unit of x = 1 month (monthly equation)

Estimate for March 2011 (x = -16) = 39.23

Estimate for September 2013 (x = 14) = 44.98

2) 45.17 tons

3) 73

Check Your Progress 3

1) 10957, 107.00, 118.69, 82.71, 84.87, 89.19, 99.47, 100.87, 100.11, 99.82, 100.58, 107.12

2) 112.27, 86.62, 100.21, 100.90

3) 104.2, 97.9, 96.5, 101.4

4) Rs. 13144; 12835; 20876.

UNIT 8 VITAL STATISTICS*

Structure

8.0 OBJECTIVES

After going through this Unit, you will able to

- explain the sources of data in vital statistics;
- calculate various vital rates;
- explain the procedure of construction of life tables; and
- appreciate the application and limitations of life tables.

* Adapted from IGNOU study material of EEC 13: Elementary Statistical Methods and Survey Techniques, Unit 12 written by C G Naidu with modifications by Kaustuva Barik.

8.1 INTRODUCTION

Vital statistics is mainly concerned with the factors contributing to population growth. Some of these factors are birth rates, death rates, expectancy of life, and migration. As you go through this Unit you will be in a position to appreciate the importance and applications of vital statistics in economics. The main objectives of this Unit are to introduce some of the basic concepts of vital statistics, the data sources, how to measure various ratios, and what are the applications of these ratios in projecting the population, calculating life expectancy, uses in actuarial science, etc.

8.2 DATA SOURCES

The data for vital statistics are usually collected through the following four methods; viz., Registration, Census, Survey and Sample Registration System. We discuss these methods below:

i) **Registration Method:** This method consists of continuous and permanent recording of births, deaths, marriages, migration, etc. Many countries including India have made registration of births and deaths compulsory under the law. The registration office issues a certificate on registration of a birth or death. Although, the registration method in simple and effective it suffers from the problem that all the births and deaths that are occurring are not registered. This is because the law has not been enforced strictly, particularly, in rural India.

ii) **Census:** Almost all the countries in the world conduct census periodically to enumerate their population. The census provides the vital statistics information such as age, sex, marital status, education level occupation, religion, etc. However, these information pertain to the census years only (once in ten years). The data for the years other than census years are not available.

iii) **Surveys:** Surveys are conducted in areas where the registration method is not effective or not functioning properly. The surveys enable us to have required vital statistics of these regions.

iv) **Sample Registration System (SRS):** This is a large scale demographic survey conducted in India for providing reliable annual estimates of birth rate, death rate and other fertility and mortality indicators at the national and sub-national levels. The field investigation consists of continuous enumeration of births and deaths by a resident part-time enumerator, generally a school teacher followed by an independent survey every six months by an official. The data obtained through these operations are matched. The unmatched and partially matched events are re-verified in the field and thereafter an unduplicated count of births and deaths is obtained. The SRS was initiated by the office of the Registration General, India on a pilot basis in a few selected states in 1964-65. It became fully operational during 1969-70 covering about 3700 sample units. Thereafter the sample size

has been periodically increased. The frame was recently updated in 2014 comprising 8861 sample units.

8.3 USES IF VITAL STATISTICS

Vital statistics are useful in many spheres of human activity. Some important uses of vital statistics are as follow.

1) The vital statistics help us in understanding how the population profile of a country or a region within the country is changing. The profile of the population is in terms of age, sex, religion, births, deaths, migration, marriages, etc. these statistics help us in predicting the future population structure of a country or region.
2) The estimation of population trends and projections help the policy planners and administrators for better planning and evaluation of economic and social development programmes. For example, the transportation infrastructure is influenced directly by the number of people in an area.
3) The mortality statistics help us to improve the health, conditions of the communities. For example, statistics on communicable diseases help the health authorities to improve the sanitary conditions of the area affected and improve medical facilities.
4) Actuarial science including life insurance is based on vital statistics. You will learn more about it in section 12.4 of this Unit.

8.4 MEASUREMENT OF POPULATION

The total population of a country is usually expressed at a particular point of time. For example, the latest census in India was conducted in 2011 and the total population of the country was found to be 1.21 billion on 31.3.2011. The total population measured at a census is usually considered as accurate. As you may be aware, the census in India are conducted once in 10 years. The inter-censal data are estimated using the following methods.

8.4.1 Linear Interpolation Method

The estimation of total population for a given inter-censal year can be calculated using the following formula:

$$P_t = P_0 + \frac{n}{N}(P_1 - P_0) \qquad \ldots (8.1)$$

where, P_1 is estimated population at a given inter-censal year t

P_0 is population in the previous census

P_1 is population in the succeeding census

n is the number of years between the given year and the previous census year

N is the number of years between the two census years

The above method provides us a good estimate at a constant rate between the inter-censal years.

Example 8.1: The total population of India in 1991 census was 846 million and in 2001 census was 1027 million. Calculate the total population of India in 1996.

Here, $P_0 = 846$, $P_1 = 1027$, $N = 10$, $n = 5$

Therefore, $P_{1996} = 846 + \frac{5}{10}(1027 - 846) = 936.5$ million.

The limitation of the above method is that is that we can estimate the population only for the years between two census years. We cannot have the estimates for the future years.

8.4.2 Using Compound Growth Rate Formula

Normally it was observed that the population growth takes place in a geometrical progression. In case the base year population and the population compound growth rate (between the base census year and succeeding census year) are known, we can estimate the total population for a given year using the following formula.

$P_1 = P_0(1+r)^n$

where, r is the compound growth rate (between the base census year and succeeding census year)

n is the number of years from the base year (usually previous census year)

P_0 is the base year (usually previous census year

P_t is the estimated population at a given year t from the base year

Example 8.2: The population of a small town in 2001 was 50500. The compound growth rate of the population of that town between 2001 and 2011 was 0.025. Estimate the population of the town for the year 2015 (assuming that the population growth rate will be the same beyond 2011).

Here, we are given $P_0 = 50500$, $r = 0.025$, and $n = 14$ (since 2015 – 2001 = 14)

Therefore, $P_{2015} = 50500(1 + 0.025)^{14} = 71355$

8.4.3 Natural Increase and Net Migration Method

You know that the census gives us the total population. Similarly, the total number of births, deaths, and migration statistics are obtained from registrars. The population of an area increase by:

i) Natural increase (that is, total number of births – total number of deaths)

ii) Ne migration (that is, total number of people immigrated to the area – total number of people emigrated out of the area).

The population for a given period is calculated using the following formula.
$P_t = P_0 + (B - D) + (I - E)$

Where, P_t is the estimated population at a given year t form the base year (usually previous census year)

P_0 is the base year (usually previous census year)

B and D are the total number of births and deaths respectively during the base year to the year t.

I and E are the total number of immigrants and emigrants respectively during the base year to the year t.

Example 8.3: The population of a small town in 2011 census was 22000. From 2011 to 2013 the number of births, deaths, immigrants and emigrants are 800, 150, 2500 and 1500 respectively. Find the total population of the town in 2013.

Here, $P_0 = 22000$, B = 800, D = 150, I = 2500, E = 1500

Therefore, $P_{2013} = 22000 + (800 - 150) + (2500 - 1500)$

$= 23650$

Check Your Progress I

The following table gives information on mid-year total population of India and annual compound growth rates of population.

Year	Population (Crores)	Period	Compound growth rate (%)
1950	36.99	1950-60	1.9
1960	44.59	1960-70	2.2
1970	55.50	1970-80	2.1
1980	68.70	1980-90	2.0
1990	84.17	1990-2000	1.8
2000	100.27	2000-2010	1.4
2010	113.17	2010-20	1.1
2020	132.61		

Sources: US Census Bureau: IDB Summary Demographic Data for India, 2020

Note that the compound growth rates are in terms of percentage. Divide it by 100 to get the required r. For example, for the period 1950-60 the compound growth rate is 1.9%. Therefore, r = 1.9/100 = 0.019.

On the basis of the above table answer the questions below:

1) Find the mid-year population for the following years using linear interpolation method.

Year	Mid-year population
1954	
1966	
1973	
1985	
1998	
2005	
2018	

2) Find the mid-year population for the following years using compound growth rate method.

Year	Mid-year population
1954	
1966	
1973	
1985	
1998	
2005	
2018	

3) Compare the above two methods and draw your conclusions.

……

……

……

……

……

4) Briefly explain different data sources for vital statistics.

……

……

……

……

……

5) What are the important uses of vital statistics?

……

……

……

……

……

8.5 VITAL RATES

Normally, the data on vital statistics are available in the form of number of births, number of deaths, etc. In order to have a meaningful utility of these data we generally transform this data into some vital rates or ratios. Number of births or deaths in year per 100 persons is usually low and would result in small fractions. The changes in these ratios would also be not perceptible. In order to avoid this

problem, vital rates are expressed on the basis of per thousand persons. In this section you will learn some important vital rates.

8.5.1 Crude Birth Rate

The crude birth rate is defined as the number of births per 1000 population in a specific community or region. To calculate the crude birth rate we use the following formula:

$$\text{Crude birth rate} = \frac{\text{Annual Number births (in a community or region)}}{\text{Annual mid year population (of the community or region)}} \times 1000$$

The crude birth rate tells us at what rate the births are occurring in a region or community.

Example 8.4: The mid-year population and number of births occurred of a tribal community in Madhya Pradesh in 2020 are 40,000 and 1200 respectively. Find the crude birth rate.

Here, we have 2020 mid-year population = 40000 and the 2020 number of births = 1200

$$\text{Crude birth rate} = \frac{1200}{40000} \times 1000$$

= 30 per 1000 persons per annum

8.5.2 Crude Death Rate

The crude death rate is defined as the number of deaths per 1000 population in a specific age group or sex group or community or region. To calculate the crude death rate we use the following formula.

$$\text{Crude Death Rate} = \frac{\text{Annual number of deaths (in a specific age group or sex group or community or region)}}{\text{Annual mid year population (of the specific age group or sex group or community or region)}} \times 100$$

The crude death rate tells us at what rate the deaths are happening in a age group, sex group or region or community.

Example 8.5: The mid-year population and the number of deaths registered in 2020 for a town in Maharashtra among females are 25000 and 245 respectively. Find the crude death rate.

Here, we have 2020 mid year female population = 25000 and the number of deaths in 2020 = 245.

$$\text{Crude death rate (females)} = \frac{245}{25000} \times 1000$$

= 9.8 per 1000 persons per annum among females.

8.5.3 Crude Rate of Natural Increase

The crude rate of natural increase is defined as the rate at which a population increases in a given year because of a surplus of births over deaths expressed as per 1000 persons.

The annual natural increase is measured as: annual number of births-annual number of deaths.

The formula for calculating the crude rate of natural increase is

$$\text{Crude rate of natural increase} = \frac{\text{Annual natural increase}}{\text{Annual mid year population}} \times 1000$$

= crude birth rate – crude death rate

The crude rate of natural increase for a given year tells us at what rate natural increase has added the population over the year.

Example 8.6: In India the crude birth rate and crude death rates in 2016 are 21.4 and 7.5 respectively. Find the crude rate of natural increase.

Crude rate of natural increase = 21.4 – 7.5

= 13.9 per 1000 per annum

8.5.4 Rate of Net Migration

Migration is defined as the movement of people across a specific boundary of a region for the purpose of establishing a new or sent permanent residence. Immigrants are those who have come into the region and eminigrants are those who have moved out of the region.

The annual net migration is defined as: Annual number of immigrants – annual number of emigrants.

The formula for calculating the annual rate of net migration is:

$$\text{Annual rate of net migration} = \frac{\text{Annua net migration}}{\text{Annual mid year population}} \times 100$$

The annual rate of net migration tells us at what rate the net migration has added to the population over the course of the year.

Example 8.7: In 2020 for a region the annual number of immigrants, emigrants, and mid-year population are given as 6500, 5200 and 667700 respectively. Find the annual rate of net migration.

Here, we have the number of immigrants = 6500

the number of emigrants = 5200

mid-year population = 66700

Annual net migration = 6500 – 5200 = 1300

$$\text{Annual rate of net migration} = \frac{1300}{66700} \times 1000 = 19.7 \text{ per 1000 per annum}$$

8.5.5 Rate of Total Increase

The total increase in population in measured as:

$$\text{Annual natrual increase} + \text{annual net migration.}$$

$$\text{Rate of total increase} = \frac{\text{Annual total increase}}{\text{Annual mid year population}} \times 100$$

= crude rate of natural increase + rate of net migration.

The rate of total increase for a given year tells us the rate at which the population has increased over the year.

Example 8.8: The annual natural increase, annual net migration, and annual mid-year population in 2018 for a region are recorded as 1500, 500 and 50000 respectively. Find the rate of total increase.

Here, we have

Annual natural increase = 1500

Annual net migration = 500

Mid-year population = 50000

Annual total increase = 1500 + 500 = 2000

$$\text{Rate of total increase} = \frac{2000}{50000} \times 1000$$

= 40 per 1000 per annum.

8.5.6 Infant Mortality Rate

The infant mortality rate is defined as the member of deaths of infants (less than one year old) per 1000 live births in a given year. The formula to calculate the infant mortality rate is given as:

$$\text{Infant mortality rate} = \frac{\text{Annual infant deaths (of males or females or total)}}{\text{(Annual live births (of males or females or total)}} \times 100$$

The infant mortality rate tells us for a given year the chances of a birth failing to survive one year life. The infant mortality rates can be calculated separately for males and females.

Example 8.9: In 2019 for a small town the total number of live births and infant deaths among females are recorded as 3000 and 25 respectively. Find the infant mortality rate among females.

Here, we have

Annual live female births = 3000

Annual infant deaths = 25

$$\text{Infant mortality rate} = \frac{15}{3000} \times 1000$$

= 8.33 per 1000 per annum

Check Your Progress 2

The provisional estimates of crude birth rate, crude death rate, natural growth rate and infant mortality rate in India for the year 2018 are as follows:

Vital rate	Total	Rural	Urban
Birth rate	20.0	21.6	16.7
Death rate	6.2	6.7	5.1
Natural growth rate	13.8	14.9	11.6
Infant mortality rate	32	36	23

Sources: Sample Registration System Bulletin, May 2020

Observe that all the vital rates are higher in rural areas than in urban areas. Write one most significant reason for each of the following:

1) The birth rate in rural areas is high because:

.....

.....

.....

.....

2) The death rate in urban areas is low because:

.....

.....

.....

.....

3) The infant mortality rate in rural areas is high because:

.....

.....

.....

.....

8.6 LIFE TABLES

The life expectancy is defined as the average number of additional years a person could expect to live if the current mortality trends continue for the rest of that person's life. A life table is a tabular display of life expectancy and probability of dying at each age or age group for a given population, according to the age-specific death rates prevailing at that time.

The life table gives us an organized complete picture of a population's mortality. We can explain it with an example. We start with a group (usually called 'cohort') of 100,000 female births and estimate the number which will survive to every age or age group, if they are subjected to the existing mortality condition. We can say, for example, that out of 100,000 initial female births 95,000 will reach the age of 15 years, 92,500 the age of 25 years, and so on, and the mean age at which all 100,000 will die is 72 years.

The construction of a life table is a simple process. It involves the following steps that are repeated for each age group.

i) **Age interval (x to x + n):** The period of life between two exact ages. The exact age (x) represents the lower limit of each age interval, beginning with 0 and incrementing to 1, 5, 10, 15 and so on upto 100+ (which is an open interval). The first and second age groups are usually '<1' and '1-4' and the last age group is '100+' whereas the rest of the age groups are of equal size, like '5-9', '10-14', '15-19', …, ' 95-99'.

ii) **Width of the age interval (n_x):** This is the number of years in the age interval (x to x+n). Usually, the first value is 1 (interval <1), the second is 4

(i.e., 1-4) and the remaining values are 5 (namely, 5-9, 10-14, … 95-99). The last value is an exception, which is again taken as 1 (100+).

iii) **Number of deaths recorded in the age interval (d^x):** This column presents the number of persons dying in that age group during the year corresponding to the life table.

iv) **Number of persons in the age interval (P_x):** This column indicates the number of persons in the age interval during the year corresponding to the life table.

v) **Separation factor ($_na_x$)**: This represents the average number of years lived by those who die between age x and (x+n). Although, it is necessary in calculations, this factor is not typically presented as a column of the life table. Each person living in the interval (x to x+n). In a complete life table, a value 0.5 (that is, half of one year) is valid from the age of 5 years. For a simpler calculation, it is assumed that those who die in the 5 year age intervals of a life table live on average 2.5 years. However, remember that the value of the fraction depends on the mortality pattern over the entire interval and not the mortality rate for any single year. In addition, since a large portion of infant deaths occur in the first few weeks of life, this value is much smaller in the <1 and 1-4 age groups.

Similarly, the death rates in the last three groups (namely, 91-94, 95-99, and 100+) are very high. Therefore, the value of the separation factor is small in the age group 91-94 and 95-99. In the last age group (100+) since the death is certain we have taken the separation factor as 1.

Calculation of the separation factor is easy if the date of birth and the date of death are available. For the purpose of constructing a life table the separation factor will be given in the table. When they are not, values from model life tables, such as those tabulated by Coale and Demney shown in Table 8.1 can be utilised for and the rest are taken as 0.5 years for every year in the group interval (that is 2.5 in year interval).

Table 8.1: Separation Factors for Ages < 1 and 1-4

		Separation factor for age < 1			Separation factor for ages 1-4		
	Zones	Men	Women	Both sexes	Mean	Women	Both sexes
Infant Mortality Rate >0.100	North (1)	0.33	0.35	0.3500	1.558	1.570	1.5700
	East (2)	0.29	0.31	0.3100	1.313	1.324	1.3240
	South (3)	0.33	035	0.3500	1.240	1.239	1.2390
	West (4)	0.33	0.35	0.3500	1.352	1.361	1.3610
Infant Mortality Rate <0.100	North (1)	0.0425	0.05	0.05	1.859	1.733	1.7330
	East (2)	0.0025	0.01	0.01	1.614	1.487	1.4870
	South (3)	0.0425	0.05	0.05	1.541	1.402	1.4020
	West (4)	0.0425	0.05	0.05	1.653	1.524	1.5240

Source: Coale, Ansley J. and Demeny P. (1966) *Regional Model Life Tables and Stable Population,* Princeton University Press.

Notes: (1) Iceland, Norway and Switzerland; (2) Austria, Czechoslovakia, North-central Italy, Poland and Hungary; (3) South Italy, Portugal and Spain; (4) Rest, of the World.

vi) **Central Mortality ($_nM_x$):** This column results from dividing the number of deaths in the age interval x to x+n (column d_x) by the number of people in this age group (column P_x).

$$_nM_x = \frac{d_x}{P_x}$$

vii) **Probability of dying between the ages x and x+n ($_nq_x$):** The probabilities of dying are calculated based on the age-specific mortality rates for each age group. This column is interpreted as the probability of dying between the ages for the person who has survived up to age x. For the last age group of the table, where death is unavoidable, the probability of dying is 1. For other age groups, the calculation is more complicated. The formula for calculation is given below:

$$_nq_x = \frac{n_x \times {}_nM_x}{1 + (n_x - {}_na_x) \times {}_nM_x}$$

viii) **Probability of survival between the ages x and x+n ($_np_x$):** It is interpreted as the probability of a person who reaches age x to reach the exact age x + n alive. The formula for calculation is given below:

$$_np_x = 1 - {}^nq_x$$

Since, it is $1 - {}_nq_x$, we normally do not show this as a separate column in the life table.

ix) **Survivors to exact age x($_nl_x$):** This column indicates the number of persons living in the age group x to x + n out of the initial cohort which is usually taken as 100,000.

x) **Deaths between the exact ages x and x+n ($_xd_x$):** This is calculated using the following formula.

$$_nd_x = {}_nl_x \times {}_nq_x$$

xi) **Number of years lived by the total of the cohort of 100, 000 births in the interval x to x + n ($_nL_x$):** Each member of the cohort who survives the interval x to x+n contributes n years to L, while each member who dies in the interval x and x+n contributes the average number of years lived by those who die in this period (that is, the separation factor of deaths $_na_x$). The $_nL_x$ is calculated using the following formula.

$$_nL_x = n_x \times {}_nl_{x+n} + {}_na_x \times {}_nd_x$$

where, $_nl_{x+n} = {}_nl_x \times {}_np_x$

or,

$$_nl_{x+n} = {}_nl_x - {}_nd_x$$

xii) **Total years lived after exact age $x({}_nT_x)$:** This number is essential for the calculation of life expectancy. It indicates the total number of years lived by the survivor ${}_nl_x$ between the anniversary x and the extinction of the whole generation. The value of the first row of ${}_nT_x$ is the total number of years lived by the cohort until death of its last component.

$${}_nT_x = \text{Sum of } {}_nL_x \text{ (from last row of } {}_nL_x \text{ to the current row of } {}_nL_x)$$

xiii) **Life expectancy at age $x({}_ne_x)$:** Among all the indicators provided by the life table, the most widely used is the life expectancy $({}_ne_x)$ which represents the average number of years lived by a generation of newborns under given mortality conditions.

Table 8.2 below provides the basic information required for construction of a life table. The data pertains to Indian females in 2000. Let us construct the life table.

Table 8.2: Basic Information

Age	Number of deaths (d_x)	Number of people (p_x)	n_x	Separation factor $({}_na_x)$
<1	788471	11655599	1	0.1
1-4	430704	44728827	4	1.6
5-9	137870	54725561	5	2.5
10-14	69159	52128201	5	2.5
15-19	100055	48475620	5	2.5
20-24	119360	42745630	5	2.5
25-29	116085	39848328	5	2.5
30-34	109226	35983667	5	2.5
35-39	102540	31934500	5	2.5
40-44	124848	27744053	5	2.5
45-49	150315	23125487	5	25
50-54	172910	19212249	5	2.5
55-59	226553	16258203	5	2.5
60-64	288036	13715985	5	2.5
65-69	354148	10813430	5	2.5
70-74	368365	7554310	5	2.5
75-79	335430	4615527	5	2.5
80-84	252665	2332329	5	2.5
85-89	130278	817817	5	2.5
90-94	42440	183658	5	2
95-99	8199	24796	5	2
100+	915	1961	1	+
All ages	4428572	488625738		

Source: World Health Organisation

Using the formulas given earlier the following life table is constructed.

Table 8.3: Life Table

Age	n_x	$_na_x$	$_nM_x$	$_nq_x$	$_np_x$	$_nl_x$	$_nd_x$	$_nL_x$	$_nT_x$	$_ne_x$
<1	1	0.1	0.06765	0.6377	0.93623	100000	6376.52	94261.1	6268416	62.6842
1 to 4	4	1.6	0.0063	0.3765	0.96235	93623.5	3524.63	366.35	6174155	65.9467
5 to 9	5	2.5	0.00252	0.1252	0.98748	90098.8	1127.83	447675	5808120	64.4639
10 to 14	5	2.5	0.00133	0.00661	0.99339	88971	588.245	443384	5360446	60.2493
15 to 19	5	2.5	0.00206	0.01027	0.98973	88382.8	907.437	439645	4917061	55.6337
20 to 24	5	2.5	0.00279	0.1386	0.98614	87475.3	1212.83	434345	4477416	51.1849
25 to 29	5	2.5	0.00291	0.01446	0.98554	86262.5	1247.4	428194	4043071	46.8694
30 to 34	5	2.5	0.00304	0.01506	0.98494	85015.1	1280.57	421874	3614877	42.5204
35 to 39	5	2.5	0.00321	0.01593	0.98407	83734.5	1333.63	415339	3193003	38.1324
40 to 44	5	2.5	0.0045	0.02225	0.97775	82400.9	1833.39	407421	2777655	33.7092
45 to 49	5	2.5	0.0065	0.03198	0.96802	80567.5	2576.57	396396	2370243	29.4193
50 to 54	5	2.5	0.0090	0.04401	0.95599	77990.9	3432.36	381374	1973847	25.3087
55 to 59	5	2.5	0.01393	0.06733	0.93267	74558.6	5019.87	360243	1592474	21.3587
60 to 64	5	2.5	0.0210	0.09976	0.90024	69538.7	6937.35	330350	1232230	17.7201
65 to 69	5	2.5	0.03275	0.15136	0.84864	62601.3	9476.39	289318	901880	14.4067
70 to 74	5	2.5	0.04876	0.21732	0.78268	53126	11545.3	236767	612562	11.5304
75 to 79	5	2.5	0.07267	0.3075	0.6925	41580.7	12786.2	175938	375795	9.03774
80 to 84	5	2.5	0.10833	0.42622	0.57378	28794.5	12272.9	1132900	199857	6.94081
85 to 89	5	2.5	0.1593	0.56964	0.43036	16521.6	9411.37	59079.6	86567	5.23963
90 to 94	5	2	0.23108	0.68236	0.31764	7110.23	4851.74	20996	27487.4	3.8659
95 t0 99	5	2	0.33067	0.82999	0.17001	2258.5	1874.53	5668.9	6491.49	2.87425
100+	1	+	0.46678	1	0	383.969	383.969	822.593	822.593	2.14235

Life expectancy always decreases from the first row of the table to the last row, with the exception of the second row and sometimes the third row (age group/5-9), which can be greater than the first row (age group/<1) in countries with high infant mortality. It is generally observed that for a given population, life expectancy is greater in women than in men and overall life expectancy should be approximately between the two. However, in countries where the maternal mortality is high the general living conditions of women are worse, life expectancy among women is lower than men.

8.7 APPLICATIONS OF LIFE TABLES

The life table is widely used in demographic, actuarial, social and health studies. The principal objective of a life table is to calculate life expectancy at birth and at other ages. However, life table provides interesting demographic data which have various applications. In this section you will learn the applications of the life table.

8.7.1 Calculation of Probability of Surviving and Dying

While constructing life table you have learnt that ${}_nq_x$ is the *probability of dying* between the two ages (x, x+n) for the person who has survived up to age x. For example, let us consider the row corresponding to age group 30-34 years in Table 8.3. The probability of dying (females) between 30 to 34 years of age, for those who have survived up to 30 years of age, is 0.01506 $({}_nq_x)$.It means that out of every 100,000 Indian females who have survived the age of 30 years, 1506 (= 100, 000 ×0.01506) will die between the age 30 and 34 years. Secondly, ${}_np_x$ tells us the *probability of living* between the two ages (x, 30-34 years/ x + n) of survival is (1 – 0.01506) = 0.98494 $({}_nq_x)$.That means out of every 100,000 Indian females who have survived the age of 30 years, 98494 will survive in the age group 30-34 years.

Thirdly, we can calculate the *probability at birth* of a person dying between ages 0-4 years. This is given by the number of original births dying $({}_nd_x)$ between the ages 0-4 years, divided by the number of original births (usually 100000). In our example, ${}_nd_x = 1281$ and the probability is 0.01281 (= 1281/100000). This probability tells us that on and average out of every 100,000 female births in India (subject to mortality in 2000), 1281 females will die between the ages 0-4 years.

8.7.2 Uses in Actuarial Science

Life tables have important applications in actuarial science especially in the field of life assurance. Life tables form the basis for determining the rates of premiums necessary to various amount of life assurance. Life tables provide the actuarial science with a sound foundation, converting the insurance business from a mere gambling in the human lives to the ability to offer well calculated safeguard in the event of death.

Actually, the calculations involved in the fixation of premium amounts in Life assurance are very complex, but the underlying principles are simple. Let us consider a few examples.

Example 8.10: According to mortality conditions in India for the year 2000, what annual premium would an Indian female have to pay on a whole life policy worth Rs. 100,000 if this life was assured at birth, assuming that the assurance office earns no income on its funds?

Let the premium be Rs. X per annum. Since a female on the average can be expected to live 62.7 years, over her life time she will have paid Rs. x × 62.7 in premiums. This will have to be equal to the value of the policy Rs. 100000. Therefore, Rs. x × 62.7 = 100000 and x = 100000/62.7 = Rs. 1594.90.

Example 8.11: In the above example if the policy was taken at the age of 25 years, then find the annual premium.

If the policy was taken at age 25 then the total premium paid will be Rs. x × 46.9 for 46.9 years expectation of life at 25 years age. Then the annual premium must be x = 100000/46.9 = Rs. 2132.20.

Example 8.12: In example 8.10 if the policy is an endowment policy, taken at 30 years of age and payable up to 50 years of age or prior deaths. What is the annual premium to be paid?

If the policy is an endowment policy, taken out say at 30 years payable up to 50 years or prior death, we should proceed on a somewhat different method. From Table 8.3 we know that 850155 $({}_{n}l_{x})$ survivors at age 30 live 1600530 years (= 415339 + 407421 + 396396 + 381374) $({}_{n}L_{x})$ years between ages of 30 and 50. Consequently, on the average a total of Rs. x × (1600530/85015) premiums will be collected and hence the annual premium must be Rs. 100000 / 18.83 = Rs. 5311.67.

8.7.3 Other Applications of Life Tables

Apart from its uses in insurance life tables is useful in undertaking comparative analysis of mortality conditions across countries or region. We discuss some of the applications of life table below:

i) **Calculation of mortality due to specific causes:** Life tables for different groups of population such as sex (male/female), age distribution (different age groups), and religion are calculated for comparisons. The mortality statistics may prompt us to find the specific causes of deaths in different groups of population.

Table 8.4: Life Expectancy at Birth: Select Countries – 2019

	Males years	Females years
Australia(a)	80.7	84.9
Canada	79.9	84.1
China	74.5	79.5
France	79.7	85.9
Germany	78.6	83.3
Hong Kong (SAR of China)	82.3	87.7
India	68.2	70.7
Indonesia	69.4	73.7
Italy	81.2	85.6
Japan	81.3	87.3
Korea, Republic of	79.7	85.7
Netherlands	80.3	83.4
New Zealand	80.2	83.6
Papua New Guinea	63.0	65.6
Singapore	81.0	85.4
United Kingdom	79.5	83.1
United States of America	76.1	81.1

Source: www.worlddata.info

ii) **Comparison of mortality conditions:** The life expectancy at birth and other ages are the best indices of mortality. These indices considerably vary from place to place and time to time. Over time, in most countries, life expectancy has increased steadily due to improved health facilities. As mentioned earlier, female life expectancy is higher than male life expectancy expect even though female maternal mortality is high. Table 8.4

above explains the life expectancy for males and females in some selected countries.

iii) **Population projections:** Life tables have also been used in preparation of population projections by age and sex. That is, in estimating what the size of the population will be at some future date.

8.7.4 Limitations of Life Tables

Life tables are based on demographic data collected from sources such as census and SRS. Therefore, life tale estimates have all the disadvantages of any statistical measure based on population censuses and vital records. Data on ages and mortality registration may be incomplete or biased. Infant mortality weight heavily on life expectancy, which means that under-reporting of this indicator, a habitual fact in many countries can have an important effect on the result of the tables. Also, important differences in specific age/sex groups with high mortality may be overlooked, since this would have little effect on the overall life expectancy.

Constructing life tables for small populations, at the local or sub-regional level, is generally not recommended, since migratory movements affect the population structure more than at the regional or national levels. In these cases, a very small number of deaths can be obtained, which may produce imprecise calculations of the table's columns.

Check Your Progress 3

Read the Life Table given in Table 8.3 in the text. Now interpret the values in the life table by answering the following questions.

1) What is the probability of a female child in India in 2000 would die before reaching 1 year of age?

.....

.....

.....

.....

2) How many years is a female born in 2000 in India expected to live?

.....

.....

.....

.....

3) What is the probability of dying of a female between 15 and 20 years of age?

.....

.....

.....

.....

4) What is the mortality rate between 15 and 20 years of age?

.....

.....

.....

.....

5) What is the probability that a female reaching 15 years of age reaches 20?

.....

.....

.....

.....

6) How many additional years is a female between 15 and 20 years of age in 2000 in India expected to live?

.....

.....

.....

8.8 LET US SUM UP

Vital statistics is mainly concerned with births and deaths. The reliability of vital rates depends upon the effectiveness of the registration system. Incompleteness of registration of births and deaths, in spite of the laws, has made it difficult to give a correct picture of birth and death rates.

Life tables present the mortality and survival experience of a whole population and permit evaluation of its affect on specific groups and over different periods. It is a simple instrument that is easily constructed with data collected routinely.

It is important to keep in mind that life tables are constructed based on population data from censuses and mortality registries. Therefore, the quality of the data affects the validity of the life table.

8.11 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1)

Year	Mid-year population
1950	40.03
1966	51.14
1973	59.46
1985	76.44
1998	97.05
2005	106.72
2018	128.72

2)

Year	Mid-year population
1954	39.88
1966	50.81
1973	59.07
1985	75.85
1998	97.08
2005	107.49
2018	122.55

3) The estimated mid-year populations using linear interpolation method are slightly less than the method using compound growth rate method.

4) See Section 8.2 and answer.

5) See Section 8.3 and answer.

Check Your Progress 2

1) The birth rate in rural areas is high because of the lack of awareness among the people on the family planning methods and its need.

2) The death rate in urban areas is low because of the improved health facilities in towns and cities.

3) The infant mortality rate in rural areas is high because of the lack of health facilities in rural areas and malnutrition among mothers.

Check Your Progress 3

1) The Probability for a female under 1 to die in India in 2000 $\left({}_1q_0\right)$ is 0.06377.
2) The number of years that a female born in 2000 in India expected to live $\left({}_1e_0\right)$ is 62.68 years.
3) The probability of a female dying between 15 and 20 years of age group $\left({}_5q_{15}\right)$ is 0.01027.
4) The mortality rate between 15 and 20 years age group $\left({}_5M_{15}\right)$ is 0.00206.
5) The probability that a female in the 15-19 age group reaches 20-24 years age group $\left({}_5q_{15}\right)$ is 0.98973.
6) The life expectancy of a female in the age group 15-20 years in 2000 in India $\left({}_{15}e_{15}\right)$ is 55.63 years.