PREFACE

In the curricular structure introduced by this University for students of Post Graduate Degree Programme, the opportunity to pursue Post Graduate course in Subjects introduced by this University is equally available to all learners. Instead of being guided by any presumption about ability level, it would perhaps stand to reason if receptivity of a learner is judged in the course of the learning process. That would be entirely in keeping with the objectives of open education which does not believe in artificial differentiation.

Keeping this in view, study materials of the Post Graduate level in different subjects are being prepared on the basis of a well laid-out syllabus. The course structure combines the best elements in the approved syllabi of Central and State Universities in respective subjects. It has been so designed as to be upgradable with the addition of new information as well as results of fresh thinking and analysis.

The accepted methodology of distance education has been followed in the preparation of these study materials. Co-operation in every form of experienced scholars is indispensable for a work of this kind. We, therefore, owe an enormous debt of gratitude to everyone whose tireless efforts went into the writing, editing and devising of a proper lay-out of the materials. Practically speaking, their role amounts to an involvement in invisible teaching. For, whoever makes use of these study materials would virtually derive the benefit of learning under their collective care without each being seen by the other.

The more a learner would seriously pursue these study materials, the easier it will be for him or her to reach out to larger horizons of a subject. Care has also been taken to make the language lucid and presentation attractive so that it may be rated as quality self-learning materials. If anything remains still obscure or difficult to follow, arrangements are there to come to terms with them through the counselling sessions regularly available at the network of study centres set up by the University.

Needless to add, a great part of these efforts is still experimental—in fact, pioneering in certain areas. Naturally, there is every possibility of some lapse or deficiency here and there. However, these do admit of rectification and further improvement in due course. On the whole, therefore, these study materials are expected to evoke wider appreciation the more they receive serious attention of all concerned.

Professor (Dr.) Subha Sankar Sarkar

Vice-Chancellor

First Revised Edition : May, 2017

Printed in accordance with the regulations and financial assistance of the Distance Education Bureau of the University Grants Commission.

POST GRADUATE : COMMERCE

[M. COM.]

: Subject Committee :
Members

1. Prof. Bhabatosh Banerjee
2. Prof. Ranajit Chakrabarty
3. Prof. Ujjwal Mallik
4. Dr. Madan Mohan Maji
5. Prof. Swagata Sen
6. Prof. Sudipti Banerjea

Paper-7

Modules : 1 & 2

Basic Statistical Concepts and Tools

Course Writer

Dr. Nishith Kumar Patra

Editor

Prof. Sharmila Banerjee

Netaji Subhas
Open University

Post Graduate : Commerce
M. Com-7

Module 1

Module 2

Unit 1 ❑ Central Tendency and Dispersion

Structure

1.0 Objectives

After the data set is obtained it becomes often necessary to have a single representative value which describes the entire set of observations. It is a central value of the data and the entire mass of data has a tendency to cluster around this central value. It gives a comprehensive view of the entire data. For example, the per capita income of a community indicates the standard of living of an average person in that community and is obtained by averaging the income of all the earning persons in the community.

But average alone cannot adequately describe a set of data unless all the observations are same. It is also necessary to describe the variability or the dispersion of the observations. So a measurement of dispersion or variation which measures the scatterness of the mass of data about some average, should also be considered. Thus, the degree of condensation of data is measured by the central tendency of the data and the degree of variability of the data is measured by the dispersion of data. For example, consider the following observations on income per hour for 3 persons in each of three communities : (i) Rs. 10, Rs. 10, Rs. 10, (ii) Rs. 5, Rs. 10, Rs. 15 and (iii) Rs. 2, Rs. 8, Rs. 20. All three samples have same average of Rs. 10 but have different dispersions. Therefore, to summarize a data set these two characteristics should be of prime importance.

1.1 Introduction

Statistics means collection of data for study of some enquiry in the plural sense and it means the body of methods that are used for the treatment of data like collection, organisation, presentation, analysis and interpretation of data investigated in the singular sense. Data are of two types : primary data and secondary data. Primary data are collected from the field of enquiry and secondary data are collected from the data collected already for some definite purpose and recorded in books, journals, reports, office records etc.

Utmost care must be taken in collecting and tabulating data and a thorough scrutiny must be done because they will form the foundation of statistical analysis. Any error in collection and tabulation of data will influence the results and decisions obtained from those analyses.

After collection of data they should be properly represented by any one of the three representations : (a) textual, (b) tabular and (c) graphical representations. Among these the first one is not good, the second one is the exact representation of data, useful for the literate persons and the third one is approximate, obtained by graphs and charts, useful for easy comparisons even for laymen. Government of every country collects data in relation to its people, its economy, its natural resources and its socio-political conditions to know the growth rate of population, per capita national income, own resources etc. Statistics and their interpretation are helpful for the proper organisation of business, commerce, agriculture and physical, biological and social sciences.

There are two types of data : qualitative and quantitative. The first type is called attribute and the second type is called variable. A variable is called a discrete variable

if it considers some isolated or integral, finite or countably infinite values and it is called continuous if it considers any value in a set of uncountable infinite values in an interval. Sex of a person, size of a family and age of a person are the examples of an attribute, discrete variable and continuous variable respectively. In case of continuous variable discreteness considered is completely artificial due to limitations of the measuring instruments.

To know the smoking habits of people in a village A of West Bengal each individual of that village has been asked whether he/she is a smoker or not and the data are then arranged in the following frequency table.

Table : 1.1

Result of smoking habit in the village A

Smoking habit	Number of people
Smoker	423
Non-Smoker	677
Total	1,100

Table 1.2 gives a typical frequency distribution of a discrete variable (obtained by usually counting and using tally marks). It is called a simple frequency distribution.

Table : 1.2

Frequency distribution table of the number of peas in 50 pea pods

Number of peas	(frequency) No. of pea pods
1	2
2	9
3	18
4	12
5	7
6	2
Total	50

If the number of observations in a data is sufficiently large, in order to represent these in a concise and interpretable form, grouped frequency table is considered. The values of the variable are grouped into class intervals.

One such example follows.

Table : 1.3

Frequency distribution of marks of students of a school in Madhyamik Examination, 2014

Marks	(frequency) No. of students
11-20	2
21-30	6
31-40	10
41-50	16
51-60	7
61-70	5
71-80	2
81-90	2
Total	50

We summarise data by classifying them in the form of a frequency distribution. Nature of the data will be described by some important features of data viz. the central tendency and dispersion of data. Their measures are considered below. The requisites of a good summary measure of different features of data are as follows :

It should be (i) easily understood, (ii) simple to compute, (iii) based on all the values, (iv) least affected by the extreme values, (v) rigidly defined, (vi) capable of algebraic treatment and (vii) least affected by sampling fluctuations.

Throughout this chapter of central tendency and dispersion we consider x_1, x_2, ..., x_i, ..., x_n as n values of the variable X, which may occur once with simple frequencies as follows :

Value of the variable X	frequency (f)
x_1	f_1
x_2	f_2
.	.
.	.
.	.
x_i	f_i
.	.
.	.
.	.
x_n	f_n
Total	N

or, according to the grouped frequency distribution as follows :

class of the variable x	mid-value	frequency (f)
$L_1 - U_1$	x_1	f_1
$L_2 - U_2$	x_2	f_2
...	...	...
$L_i - U_i$	x_i	f_i
...	...	...
$L_n - U_n$	x_n	f_n
Total		N

Here the mid-value (also called the class mark) of the i-th class is, $x_i = \frac{L_i + U_i}{2}$ for i = 1, 2, ..., n, L_i being the lower limit and U_i being the upper limit of the i-th class, i= 1, 2, ..., n.

The first two representations will be same when all frequencies are unity.

1.2 Central tendency and its measures

Generally, the data have a tendency to cluster around a central value. This tendency of data is called central tendency of data. The central tendency is measured by a single representative value which will describe the characteristic of the entire distribution and is obtained by condensing the entire mass of data in a single representative value. The value around which the data are clustered, represents a measure of central tendency or an average. Some of the important measures of central tendency of data are : mean, median and mode. Mean is of three types : (1) arithmetic mean, (2) geometric mean and (3) harmonic mean.

1.2.1 Arithmetic mean

If n values of the variable X are x_1, x_2, ..., x_n then the arithmetic mean of X is

$$\bar{x} = \frac{x_1 + x_2 + \ldots\ldots + x_n}{n} = \frac{1}{n}\sum_{1=1}^{n} x_i \ . \qquad \ldots\ldots\ldots (1)$$

For simple frequency distribution of the variable X, where x_i, the i-th value of X, has the frequency f_i, i = 1, 2, ..., n and N = total frequency = $\sum_{i=1}^{n} f_i$, the arithmetic mean of X is

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots\ldots + f_nx_n}{N} = \frac{1}{N}\sum_{i=1}^{n} f_ix_i \qquad \ldots\ldots\ldots (2)$$

For grouped frequency distribution of the variable x : i.e. for the i-th class having frequency f_i and the mid-value x_i, the arithmetic mean of x is

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i$$

$$= \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} \qquad \ldots (2')$$

It is to be remembered that the average defined in (1) is called the simple average while the average defined in (2') is called the weighted average, the weight being f_i, the class frequency of the i-th class.

First result will be achieved from the second result when $f_1 = f_2 = f_3 = ... = f_n = 1$.

IMPORTANT PROPERTIES OF THE ARITHMETIC MEAN :

(a) Sum of deviations of values of the variable from its arithmetic mean is always zero.

In symbols, (i) $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ for n values $x_1, x_2, ..., x_n$ of the variable x.

and (ii) $\sum_{i=1}^{n} f_i\,(x_i - \bar{x}) = 0$ for simple frequency distribution of the variable x, x_i being the mid-value of the i-th class, f_i being the frequency of that class.

Proof : (i) $\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x}$ (from 1)

$= 0$

(ii) From (2) $\sum_{i=1}^{n} f_i(x_i - \bar{x}) = \sum_{i=1}^{n} f_i x_i - \bar{x}\sum f_i = N\bar{x} - N\bar{x}$ (from 2)

$= 0$ where $N = \sum_{i=1}^{n} f_i$.

(b) If two variables X and Y are related by the expression y = a + bx where a and b are constants, then $\bar{y} = a + b\bar{x}$.

Proof : When the variable X has n values : $x_1, x_2, ..., x_n$ and variable Y has n values $y_1, y_2, ..., y_n$ so that $y_i = a + bx_i$ for i = 1, 2, ..., n,

then $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}(a + bx_i) = \sum a + b\sum_{i=1}^{n} x_i = na + b\sum_{i=1}^{n} x_i$.

Dividing both sides by n, we get $\bar{y} = a + b\bar{x}$.

For simple and grouped frequency distributions of x_i where x_i = the i-th value of the variable in simple frequency distribution and x_i = mid-value of the i-th class for grouped frequency distribution i = 1, 2, ..., n, with f_i as the corresponding frequency,

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{n} f_i y_i = \frac{1}{N}\sum_{i=1}^{n} f_i (a + bx_i)$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i a + \frac{1}{N}\sum_{i=1}^{n} f_i b x_i$$

$$= \frac{1}{N} a\sum_{i=1}^{n} f_i + b\frac{1}{N}\sum_{i=1}^{n} f_i x_i$$

$$= \frac{1}{N} aN + b\bar{x} = a + b\bar{x}.$$

(c) The arithmetic mean depends both on the change of origin and on the change of scale.

Proof : Let $u_i = \frac{x_i - c}{d}$ when c and d are constants, for i = 1, 2, ..., n

(c is called origin and d is called scale).

Let $x_1, x_2, ..., x_n$ be the values of the variable X and $u_1, u_2, ..., u_n$ be the respective values of the variable U satisfying $x_i = c + du_i$.

Naturally, then $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n}(c + du_i) = \frac{1}{n}\left[nc + d\sum_{i=1}^{n} u_i \right] = c + d\bar{u}$.

So, the arithmetic mean depends on the change of origin and on the change of scale.

Let x_i be the i-th value of the variable X with frequency f_i for i = 1, 2, ..., n such that total frequency is N. Then

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i = \frac{1}{N}\sum_{i=1}^{n} f_i (c + du_i) = \frac{1}{N}\left[cN + d\sum_{i=1}^{n} f_i u_i \right] = c + d\bar{u}.$$

So, the arithmetic mean depends on change of origin and on change of scale.

(d) (i) For two sets of data with means $\bar{x}_1$ and $\bar{x}_2$ having n_1 and n_2 observations respectively, the mean of the combined set of $(n_1 + n_2)$ observations is

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}.$$

(ii) For two sets of grouped data with means $\bar{x}_1$ and $\bar{x}_2$ having total frequencies N_1 and N_2 respectively, the combined arithmetic mean is

$$\bar{x} = \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2}{N_1 + N_2}.$$

Proof : (a) (i) Let the first and second group of n_1 and n_2 observations with respective means $\bar{x}_1$ and $\bar{x}_2$ be $x_{11}, x_{12}, ..., x_{1n_1}$ and $x_{21}, x_{22}, ..., x_{2n_2}$ respectively.

Then mean of the combined set is

$$\bar{x} = \frac{(x_{11} + x_{12} + \ldots + x_{1n_1}) + (x_{21} + x_{22} + \ldots + x_{2n_2})}{n_1 + n_2} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

since $\bar{x}_1 = \frac{x_{11} + x_{12} + \ldots + x_{1n_1}}{n_1}$ and $\bar{x}_2 = \frac{x_{21} + x_{22} + \ldots + x_{2n_2}}{n_1}$.

(ii) Let the first and second set of data be $x_{11}, x_{12}, \ldots, x_{1n_1}$ and $x_{21}, x_{22}, \ldots, x_{2n_2}$ and corresponding frequencies be $f_{11}, f_{12}, \ldots, f_{1n_1}$ and $f_{21}, f_{22}, \ldots, f_{2n_2}$ and total frequencies be N_1 and N_2 respectively.

Then $\bar{x}_1 = \frac{(f_{11}x_{11} + f_{12}x_{12} + f_{1n_1}x_{1n_1}) + (f_{21}x_{21} + f_{22}x_{22} + \ldots + f_{2n_2}x_{2n_2})}{(f_{11} + f_{12} + \ldots + f_{1n_1}) + (f_{21} + f_{22} + \ldots + f_{2n_2})}$

$$= \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2}$$

where $\bar{x}_1 = \frac{1}{N_1}\sum_{i=1}^{n_1} f_{1i}x_{1i}$, $\bar{x}_2 = \frac{1}{N_2}\sum_{i=1}^{n_2} f_{2i}x_{2i}$, $N_1 = \sum_{i=1}^{n_1} f_{1i}$ and $N_2 = \sum_{i=1}^{n_2} f_{2i}$.

Note (c) This change of origin and scale in X which defines $u_i = \frac{x_i - c}{d}$, i = 1, 2, ..., n can be used to simplify the calculation of the arithmetic mean when observations $x_1, x_2, \ldots, x_n$ are equispaced with length d.

(d.1) If there are k groups of observations with means $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ and corresponding total frequencies $N_1, N_2, \ldots, N_k$ respectively, then the mean of the combined group is

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2 + \ldots + N_k\bar{x}_k}{N_1 + N_2 + \ldots + N_k} = \frac{\sum_{i=1}^{k} N_i\bar{x}_i}{\sum_{i=1}^{k} N_i}.$$

(d.2) For two groups with N_1 and N_2 observations and corresponding means $\bar{x}_1$ and $\bar{x}_2$ the mean of the combined group $\bar{x}$ lies between $\bar{x}_1$ and $\bar{x}_2$.

Proof : Let $\bar{x}_1 \le \bar{x}_2$.

Then $\bar{x} - \bar{x}_1 = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} - \bar{x}_1 = \frac{(N_1\bar{x}_1 + N_2\bar{x}_2) - (N_1 + N_2)\bar{x}_1}{N_1 + N_2}$

$$= \frac{N_2(\bar{x}_2 - \bar{x}_1)}{N_1 + N_2} \ge 0$$

i.e., $\bar{x}_1 \le \bar{x}$. Again, $\bar{x}_2 - \bar{x} = \bar{x}_2 - \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} = \frac{N_1(\bar{x}_2 - \bar{x}_1)}{N_1 + N_2} \ge 0$ i.e., $\bar{x}_2 \ge \bar{x}$. Hence $\bar{x}_1 \le \bar{x} \le \bar{x}_2$.

Similarly, if $\bar{x}_2 \le \bar{x}_1$ one can show that $\bar{x}_2 \le \bar{x} \le \bar{x}_1$.

Merits : The arithmetic mean is the most commonly used measure of central tendency because (i) the formula is easily understood, (ii) it is easy to compute, (iii) it is based on all the observations, (iv) it is rigidly defined, (v) it is capable of algebraic treatment and (vi) it is least affected by sampling fluctuations.

Demerits : (i) It is affected by a very large or very small extreme values, (ii) if in a grouped frequency distribution any of the classes is open-ended then mid-value of that class cannot be determined and so mean can not be determined. In such a case the appropriate average is median which will be discussed now.

1.2.2 Median

The median of a set of values is the middlemost value when the values of a variable are arranged either in decreasing or in increasing order of magnitude. If n, the number of observations is odd, then the median is the middlemost observation i.e., the $\frac{n+1}{2}$th value of the ordered data. If n is even then the median is the average of two middlemost values i.e. the average of the $\frac{n}{2}$-th and the $\left(\frac{n}{2}+1\right)$-th values of the ordered data.

Suppose for a simple frequency distribution of a variable x (shown in the following table) the cumulative frequencies (C.F.), F (less than type) are calculated as

$F_1 = f_1$, $F_2 = f_1 + f_2$, $F_3 = f_1 + f_2 + f_3$, ..., $F_n = f_1 + f_2 + ... + f_n$ = total frequency = N. If N is odd, the least value of $F \geq \frac{N+1}{2}$, and if N is even then the least F's $\geq \frac{N}{2}$ & $\geq \frac{N}{2}+1$ are identified. If N is odd, median is that x for which least $F \geq \frac{N+1}{2}$ and if N is even the median is the simple average of two x's for which least F's $\geq \frac{N}{2}$ and $\geq \frac{N}{2}+1$.

variable (x)	x_1	x_2	x_3	.....	x_n	Total
frequency (f)	f_1	f_2	f_3	.....	f_n	N.
C.F. (less than type) (F)	F_1	F_2	F_3	.....	F_n	

For a grouped frequency distribution of a variable first determine the cumulative frequencies F (less than type) as follows :

Class of variable (x)	$L_1 - U_1$	$L_2 - U_2$	.....	$L_n - U_n$	Total
Frequency	f_1	f_2	.....	f_n	N
C.F. (less than type) (F)	F_1	F_2	.....	$F_n = N$	

where $F_1 = f_1$, $F_2 = f_1 + f_2$, $F_3 = f_1 + f_2 + f_3$, ..., $F_n = f_1 + f_2 + ... + f_n$ = total frequency, N, and L_i and U_i are the lower and upper boundaries of the i-th class of x, i.e., $L_{i+1} = U_i$ for i = 1, 2, ..., n. From the C.F. column least $F \geq \frac{N}{2}$ is identified and the corresponding class is called the median class.

The median is calculated from the following formula :

$$\text{median} = L + \frac{\frac{N}{2} - F_1}{f_m} \times c$$

where L = lower class boundary of the median class,

F_1= cumulative frequency of the class preceding the median class,

f_m= frequency of the median class,

c = class length of the median class (i.e., c = U – L).

Important property. If y is a linear function of x in the form y = c + dx then median of two variables is related as M_e (y) = c + d M_e (x) where $M_e(x)$ and $M_e(y)$ are the respective medians of variables x and y and c and d are constants.

Proof : Let n values $x_1, x_2, ..., x_n$ of the variable x be such that

$$x_1 \leq x_2 \leq ... \leq x_n. \qquad ... \text{ (i)}$$

Let $y = y_i = c + dx_i$ when $x = x_i$. Then

$y_1 \leq y_2 \leq .... \leq y_n$ when $d > 0$

and $y_1 \geq y_2 \geq .... \geq y_n$ when $d < 0$.

If x_k is the middle most value of the variable x, so is y_k for the variable y. So the median of y is = c + d $m_e(x)$. Similarly, it can be shown that if there are two middle most values of x and y say x_k, x_{k+1} and y_k and y_{k+1} where $y_k = c + dx_k$, $y_{k+1} = c + dx_{k+1}$. The median of y will be average of y_k and y_{k+1} and so median of y = $c + dm_e(x)$.

The generalised result in this context is, if y = h(x) is a monotonic function of x then $M_e(y) = h(M_e(x))$.

Merits : (i) Median is rigidly defined, (ii) it is easily understandable, (iii) it is simple to compute, (iv) it is not affected by the presence of a few extremely small or large values, (v) it can be calculated even if one or both the terminal classes of a grouped frequency distribution are open.

Demerits : (i) The median is not based on all the observations, (ii) it is not amenable to algebraic treatment, (iii) it may not be uniquely obtained in the case of an even number of observations.

1.2.3 Mode

Mode of a variable is the value of the variable having the highest frequency or frequency density according as the variable is discrete or continuous. There may exist more than one mode in a frequency distribution. The distribution having two modes is called the bimodal distribution, which is sometimes found.

In the frequency distribution of a discrete variable the mode can be obtained as the value of the variable for which the frequency is the highest.

In case of a continuous variable the class having highest frequency density is called modal class and it can easily be identified.

Let C_{m_o}, f_{m_o} be the class length and frequency of modal class L – U where L = lower boundary of the modal class and U = upper boundary of the modal class i.e., $C_{m_o} = U - L$. Let C_{m_o-1}, f_{m_o-1} be the class length and frequency of the class preceding the modal class and C_{m_o+1}, f_{m_o+1} be the class length and frequency of the class preceding the modal class. Then

$$\text{Mode} = M_o = L + \frac{\frac{f_{m_o}}{C_{m_o}} - \frac{f_{m_o-1}}{C_{m_o+1}}}{\frac{2f_{m_o}}{C_{m_o}} - \frac{f_{m_o-1}}{C_{m_o-1}} - \frac{f_{m_o+1}}{C_{m_o+1}}} \times C_m.$$

If three consecutive classes including the modal class in the middle are of equal length

$$\text{Mode} = M_o = L + \frac{f_{m_o} - f_{m_o-1}}{2f_{m_o} - f_{m_o-1} - f_{m_o+1}} \times c$$

where L = lower class boundary of the modal class having frequency f_{m_o}, f_{m_o-1} and f_{m_o+1} are frequencies of the classes just preceding and following the modal class.

Important property : If two variables x and y are linearly related by the relation $y = c + dx$ and mode of x is M_o, then mode of y will be $c + dM_o$, where c and d are constants.

Merits : (i) The mode is rigidly defined, (ii) the significance of mode is easily comprehensible, (iii) it is simple to compute, (iv) it is not affected by the presence of a few extremely small or large values, (v) it can be calculated even for a grouped frequency distribution having one or both of the terminal classes open.

Demerits : (i) The mode is not based on all the observations, (ii) it is not amenable to algebraic treatment, (iii) determination of exact value of mode may not be possible.

1.2.4 Geometric mean

The Geometric mean, in short, G.M. of n observations which are assumed to be positive, x_1, x_2, x_3, ..., x_n is defined to be the n-th root of the product of the values of these observations. Thus, the G.M. $= (x_1 x_2 x_3 \dots x_n)^{\frac{1}{n}}$.

Taking log on both sides we get

$$\log \text{G.M.} = \frac{1}{n}\left[\log x_1 + \log x_2 + \log x_3 + \cdots + \log x_n\right] = \frac{1}{n}\sum_{i=1}^{n} \log x_i .$$

For a simple frequency distribution of the variable X in which the i-th value x_i has frequency f_i, i = 1, 2, ..., n and total frequency = N, the geometric mean of x_1, x_2, x_3 ..., x_n is`

$$g_x = \left(\prod_{i=1}^{n} x_i^{f_i} \right)^{\frac{1}{N}} .$$

It can be calculated from the equation $\log g_x = \frac{1}{N}\sum_{i=1}^{n} f_i \log x_i$ (taking log on both sides of the above expression). Hence g_x = anti log $\frac{1}{N}\sum_{i=1}^{n} f_i \log x_i$. Thus the geometric mean is the anti log of the arithmetic mean of the logarithms of the observations.

If $f_1 = f_2 = ... = f_n = 1$ i.e., N = n, the geometric mean of n values x_1, x_2, ... x_n of the variable x is

$$g_x = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{N}} .$$

It can be calculated from the equation $\log g_x = \frac{1}{N}\sum_{i=1}^{n} \log x_i$ (taking log on both sides of the above expression).

Important properties : 1. If two variables x and y are related by y = bx then geometric mean of y = b × geometric mean of x.

Proof : $g_y = \left(\prod_{i=1}^{n} y_i^{f_i} \right)^{\frac{1}{N}} = \left(\prod_{i=1}^{n} (bx_i)^{f_i} \right)^{\frac{1}{N}} = \left(b^{\Sigma f_i} \prod_{i=1}^{n} x_i^{f_i} \right)^{\frac{1}{N}}$

$$= b\left(\prod_{i=1}^{n} x_i^{f_i} \right)^{\frac{1}{N}} = b.g_x \text{ [since } y_i = bx_i, \text{ i = 1, 2, ..., n.]}$$

2. For two groups with respective number of observations N_1 and N_2 and geometric means G_1 and G_2 respectively, the geometric mean, G, of the combined group is

$$G = \left(G_1^{N_1} G_2^{N_2} \right)^{\frac{1}{N_1+N_2}}.$$

Proof : Suppose for the two frequency distributions with respective total frequencies N_1 and N_2 the first set of observations are $x_{11}, x_{12}, ..., x_{1n_1}$ with frequencies $f_{11}, f_{12}, ..., f_{1n_1}$ respectively and the second set of observations are $x_{21}, x_{22}, ... x_{2n_2}$ with frequencies $f_{21}, f_{22}, ..., f_{2n_2}$ respectively.

Let $N_1 = \sum_{i=1}^{n_1} f_{1i}$ and $N_2 = \sum_{i=1}^{n_2} f_{2i}$. The combined geometric mean, G, can be calculated as

$$G = \left(\prod_{i=1}^{n_1} x_{1i}^{f_{1i}} \prod_{i=1}^{n_2} x_{2i}^{f_{2i}} \right)^{\frac{1}{N_1+N_2}} \text{ or } G = \left(G_1^{N_1} . G_2^{N_2} \right)^{\frac{1}{N_1+N_2}}$$

where $G_1 = \left(\prod_{i=1}^{n_1} x_{1i}^{f_{1i}} \right)^{\frac{1}{N_1}}$ and $G_2 = \left(\prod_{i=1}^{n_2} x_{2i}^{f_{2i}} \right)^{\frac{1}{N_2}}$

Generalisation : For k groups with geometric means $G_1, G_2, ..., G_k$ and corresponding total frequencies $N_1, N_2, ..., N_k$, the geometric mean of the combined group is

$$G = \left(G_1^{N_1} G_2^{N_2} \ldots G_k^{N_k} \right)^{\frac{1}{N_1+N_2+\ldots+N_k}}.$$

Merits : It is based on all the observations. It is rigidly defined. It is useful in averaging ratios and percentages in determining the ratio of change. When the observations are in geometric progression, the geometric mean is the suitable average. It is capable of algebraic treatment. It gives less weight to large items and more weight to smaller items than the arithmetic mean.

Demerits : It is difficult to compute and to interpret. If the series has any negative value then its computation is impossible. If a series has one value zero, the G.M. will be zero.

1.2.5 Harmonic mean

Harmonic Mean, in short, H.M., of a number of observations is defined to be the reciprocal of the arithmetic mean of the reciprocals of the observations. So for finding out the H.M. we are to find, first of all, the reciprocals of the observations, then the

A.M. of those new observations and then the reciprocal of this A.M. will be the required H.M. For raw data the H.M. will be $\frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$.

For a frequency distribution where the variable X takes values $x_1, x_2, ..., x_n$ with frequencies $f_1, f_2, ..., f_n$ respectively, the harmonic mean, H.M., is

$$\text{H.M.} = \frac{N}{\sum_{i=1}^{n} \frac{f_i}{x_i}}.$$

For the simple case where $f_1 = f_2 = ... = f_n = 1$, i.e. for n values $x_1, x_2, ..., x_n$ the harmonic mean will be

$$\text{H.M.} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

Important properties. 1. For two variables X and Y related by y = bx, the harmonic mean of the variable y = b × harmonic mean of the variable X where b is some constant.

Proof : Let the variable X take values $x_1, x_2, ..., x_n$ with frequencies $f_1, f_2, ..., f_n$ respectively. Then the variable Y takes values $y_1 = bx_1, y_2 = bx_2, ..., y_n = bx_n$, with frequencies $f_1, f_2, ..., f_n$ respectively. In both cases total frequency $= N = \sum_{i=1}^{n} f_i$. Then for the harmonic means H_x and H_y of variables X and Y respectively,

$$H_y = \frac{N}{\sum_{i=1}^{n} \frac{f_i}{y_i}} = \frac{N}{\sum_{i=1}^{n} \frac{f_i}{bx_i}} = \frac{N}{\frac{1}{b}\sum_{i=1}^{n} \frac{f_i}{x_i}} = b \times \frac{N}{\sum_{i=1}^{n} \frac{f_i}{x_i}} = bH_x.$$

2. For two groups with number of observations N_1 and N_2 and harmonic means H_1 and H_2 respectively, the grouped harmonic mean, H, is obtained from the relation $\frac{N_1 + N_2}{H} = \frac{N_1}{H_1} + \frac{N_2}{H_2}$.

Proof : Let the two frequency distributions have values $x_{11}, x_{12}, ..., x_{1n_1}$ with frequencies $f_{11}, f_{12}, ..., f_{1n_1}$ respectively and values $x_{21}, x_{22}, ..., x_{2n_2}$ with frequencies $f_{21}, f_{22}, ..., f_{2n_2}$ respectively. Then $N_1 = \sum_{i=1}^{n_1} f_{1i}$ and $N_2 = \sum_{i=1}^{n_2} f_{2i}$.

The grouped harmonic mean (H) and harmonic means of the i-th group (H_i) for i = 1, 2 are

$$H = \frac{N_1 + N_2}{\sum_{i=1}^{n_1} \frac{f_{1i}}{x_{1i}} + \sum_{i=1}^{n_2} \frac{f_{2i}}{x_{2i}}},\ H_1 = \frac{N_1}{\sum_{i=1}^{n_1} \frac{f_{1i}}{x_{1i}}} \text{ and } H_2 = \frac{N_2}{\sum_{i=1}^{n_2} \frac{f_{2i}}{x_{2i}}}$$

So $\frac{N_1 + N_2}{H} = \sum_{i=1}^{n_1} \frac{f_{1i}}{x_{1i}} + \sum_{i=1}^{n_2} \frac{f_{2i}}{x_{2i}} = \frac{N_1}{H_1} + \frac{N_2}{H_2}$.

Generalisation : For k groups with harmonic means H_1, H_2, ..., H_k and total frequencies N_1, N_2, ..., N_k respectively, the grouped harmonic mean, H, can be obtained from the following result :

$$\frac{N_1 + N_2 + \ldots + N_k}{H} = \frac{N_1}{H_1} + \frac{N_2}{H_2} + \ldots + \frac{N_k}{H_k}.$$

Merits. Harmonic mean is based on all the items of the series. It is capable of algebraic treatment. In problems relating to time, distance and speed it gives better result than other averages.

Demerits. It is not easily understood. It is difficult to compute. This is not desirable generally and as such it is not of much use in analysis of economic data.

1.2.6 Other positional measures

Besides median, there are measures which divide an ordered data into equal parts. Among these the most important measures are quartiles, deciles and percentiles. Quartiles are those values which divide the ordered data into 4 equal parts, deciles divide the ordered data into 10 equal parts and percentiles divide the ordered data into 100 equal parts. Thus there are 3 quartiles : Q_1, Q_2 and Q_3 the first, second and the third quartiles corresponding to which cumulative frequencies (less than type) are $\frac{N}{4}$, $\frac{2N}{4}$ and $\frac{3N}{4}$, N being the total frequency. There are 9 deciles : D_1, D_2, ..., D_9 with corresponding cumulative frequencies (less than type) as $\frac{N}{10}$, $\frac{2N}{10}$, ... $\frac{9N}{10}$. Similarly, there are 99 percentiles : P_1, P_2 ..., P_{99} with the corresponding cumulative frequencies (less than type) as $\frac{N}{100}$, $\frac{2N}{100}$, ... $\frac{99N}{100}$. From these definitions it is easy to note that Q_2, D_5 and P_{50} are same as the median, ie, $Q_2 = D_5 = P_{50}$ = median.

To determine the quartiles, deciles and percentiles we first determine the class L – U of the variable in which the particular quartile (or decile or percentile) lies. Then we use the formula

$$Q_i = L + \frac{\frac{iN}{4} - F}{f_{Q_i}} \times c,\ i = 1,2,3;\ D_i = L + \frac{\frac{iN}{10} - F}{f_{D_i}} \times c,\ i = 1,2,\ldots,9$$

and $P_i = L + \frac{\frac{iN}{100} - F}{f_{P_i}} \times c$, i = 1, 2, ..., 99 where L = Lower boundary of the classes in which the i-th quartile, Q_i (or i-th Decile, D_i or i-th percentile, P_i) lies, c = class length of that class, f_{Q_i}, f_{D_i}, f_{P_i} are the frequencies of the classes in which Q_i, D_i and P_i lies, F = cumulative frequency (less than type) of the preceding class of the class containing Q_i or D_i or P_i respectively.

1.3 Dispersion and its Measures

Generally, the values of a variable differ among themselves and they are concentrated around a central location. This characteristic is known as the central tendency of the data. The characteristic of the data, that is, how the data or the values of the variable are dispersed around some central value, is known as dispersion. The measures of dispersion are : range, quartile deviation, mean deviation and standard deviation. These measures are absolute measures of dispersion. Relative measures of dispersion are : (a) coefficient of variation, (b) coefficient of quartile deviation and (c) coefficient of mean deviation.

1.3.1 Range

It is the difference between the highest and the lowest values in a set of observations.

This measure is the simplest possible measure of dispersion. It is very easy to calculate. This measure is not based on all the observations. This measure fails to tell the characteristics of the distribution within two extreme observations. It cannot be computed in case of open end classes. Range is a passive measure of dispersion. However, it is used in statistical quality control (SQC) technique.

Range of the set of observations 10, 12, 20, 8, 4 is 20 – 4 = 16.

Important property. Let y = a + bx be the relation between two variables x and y. Then range (y) = |b| Range (x) where a and b are constants.

Proof. Maximum y = a + b × maximum of × if b > 0
= a + b × minimum of × if b < 0.
Minimum y = a + b × minimum of × if b > 0
= a + b × maximum of × if b < 0.

$\therefore$ Range (y) = maximum y – minimum y = b (maximum x – minimum x) if $b > 0$

and = –b (maximum x – minimum x) if $b < 0$

So Range (y) = $|b|$ Range (x).

Note : This property shows that the range is independent of change of origin but depends on the change of scale.

1.3.2 Mean deviation

It is the mean of absolute deviations of the given values of the variable from some average. Let $x_1, x_2, ..., x_n$ be the given values of a variable x and c be the chosen average, then $Md_c(x)$, the mean deviation of the variable *x* about c is calculated as

$$MD_c(x) = \frac{1}{n}\sum_{i=1}^{n}|x_i - c|.$$

If $c = \bar{x}$, the mean deviation about mean is

$$MD_{\bar{x}}(x) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|.$$

If $x = \tilde{x}$, the median, $\quad MD_{\tilde{x}}(x) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \tilde{x}|.$

Again, if $x_1, x_2, ..., x_n$ are the given values of a variable *x* and $f_1, f_2, ..., f_n$ are the corresponding frequencies, then

$$MD_c(x) = \frac{1}{N}\sum_{i=1}^{n}|x_i - c| \text{ and } MD_{\bar{x}}(x) = \frac{1}{N}\sum_{i=1}^{n} f_i\,|x_i - \bar{x}|$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i$ and N = total frequency $= \sum_{i=1}^{n} f_i$.

Important properties :

1. If two variables x and y are related as y = a + bx where a and b are constants then

$$MD_{\bar{y}}(y) = |b| MD_{\bar{x}}(x).$$

Proof. Suppose y takes the values $y_i = a + bx_i$ where $x = x_i$ for i = 1, 2, ..., n.

Let $x_1, x_2, ..., x_n$ have respective frequencies $f_1, f_2, ..., f_n$ so that $N = \sum_{i=1}^{n} f_i$.

Then $MD_{\bar{y}}(y) = \frac{1}{n}\sum_i f_i|y_i - \bar{y}| = \frac{1}{n}\sum_{i=1}^{n} f_i|a + bx_i - a - b\bar{x}|$

$$= \frac{|b|}{n}\sum_{i=1}^{n} f_i\,|x_i - \bar{x}| = |b| MD_{\bar{x}}(x).$$

Note. This property shows that the mean deviation is independent of change of origin but depends on change of scale.

2. An important property of the mean deviation is that it is least when measured about median.

Proof. Let x be a variable assuming the values $x_1, x_2, ..., x_n$ and c be an arbitrary value such that $\sum_{x_i<c} x_i = S_1$, $\sum_{x_i>c} x_i = S_2$. Let there be n_1 values less than c and n_2 values greater than c.

Let m be the median, $\sum_{x_i<m} x_i = S_1$ and $\sum_{x_i>m} x_i = S_2$. Let MD_c and MD_m be the mean deviations of x about c and m respectively.

$$nMD_c = \sum_{i=1}^{n} |x_i - c| = \sum_{x_i>c} (x_i - c) + \sum_{x_i<c} (c - x_i) = S_2 - n_2 c + n_1 c - S_1$$

$$= S_2 - S_1 + (n_1 - n_2)\, c.$$

$$nMD_m = \sum_{i=1}^{n} |x_i - m| = \sum_{x_i>m} (x_i - m) + \sum_{x_i<m} (m - x_i)$$

$$= S_2' - S_1'$$

since same number of values are less than m and greater than m, i.e. $\sum_{x_i>m} m = \sum_{x_i<m} m$.

There may be 3 cases : (i) c < m, (ii) c > m, (iii) c = m.

Case (i). c < m.

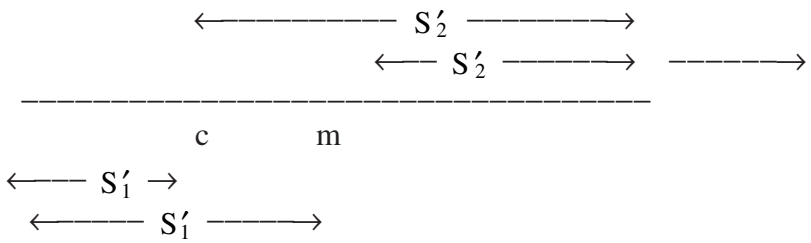


Fig. 1.1

$$nMD_c - nMD_m = S_2 - S_1 + (n_1 - n_2)c - (S_2' - S_1')$$

$$= (S_2 - S_2') + (S_1' - S_1) + (n_1 - n_2)c$$

Now, $S_2 - S_2' > \left(n_2 - \frac{n}{2}\right)c$ and $S_1' - S_1 > \left(\frac{n}{2} - n_1\right)c$

$$nMD_c - nMD_m > \left(n_2 - \frac{n}{2}\right)c + \left(\frac{n}{2} - n_1\right)c + (n_1 - n_2)c = 0$$

i.e. $MD_m < MD_c$

Case (ii) $c > m$.

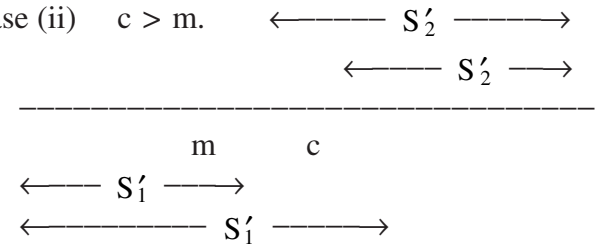


Fig. 1.2

$$S_2' - S_2 < \left(\frac{n}{2} - n_2\right)c \text{ and } S_1 - S_1' < \left(n_1 - \frac{n}{2}\right)c$$

$$\text{So } nMD_c - nMD_m = S_2 - S_1 + (n_1 - n_2)c - (S_2' - S_1') = -(S_1 - S_1') - (S_2' - S_2)$$

$$+(n_1 - n_2)c > -\left(\frac{n}{2} - n_2\right)c - \left(n_1 - \frac{n}{2}\right)c + (n_1 - n_2)c = 0$$

i.e. $MD_m < MD_c$

Case (iii) $c = m$ then $n(MD_c - MD_m) = n(MD_m - MD_m) = 0$

Then $MD_m = MD_c$

Thus from 3 cases, $MD_c \geq MD_m$.

Hence mean deviation about median is the least.

Merits and demerits : It is simple to understand and easy to compute. It is based on all the observations. It is not capable of further algebraic treatment. It is less affected by the extreme values than the standard deviation.

1.3.3 Standard Deviation

Standard deviation is the best, most important and widely used measure of dispersion. This is the positive root-mean square of deviations about mean i.e., if the variable x has n values $x_1, x_2, ..., x_n$, then the standard deviation $\sigma = +\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$, σ being the standard deviation.

The standard deviation, by definition, can not be negative. The square of the standard deviation is called the variance, var (x).

Thus, var (x) = $\sigma^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$ for raw data.

$= \frac{1}{N}\sum f_i(x_i - \bar{x})^2$ for grouped data.

For a simple frequency distribution with values $x_1, x_2, ..., x_n$ or a grouped frequency distribution with mid-values $x_1, x_2, ..., x_n$, having frequencies $f_1, f_2, ..., f_n$, the standard deviation, $\sigma = +\sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2}$.

The alternative formula for calculating σ can be derived as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i}^{n} f_i(x_i^2 + \bar{x}^2 - 2x_i\bar{x})$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 + \bar{x}^2 - 2\bar{x}\frac{1}{N}\sum_{i=1}^{n} f_i x_i$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i x_i^2 - \bar{x}^2, \text{ since } \bar{x} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i.$$

= Mean of squares of the observations – square of mean of the observations.

If $f_1 = f_2 = ... = f_n = 1$ then $N = n$ and

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \text{ where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Important properties

1. If all the values of a variable are equal then the standard deviation of the variable is zero.

Proof. Suppose all the values of a variable x are equal to c. Then their arithmetic mean $= \frac{nc}{n} = c$.

The standard deviation, $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(c-c)^2} = 0.$

Note : If standard deviation is zero then all the values of the variable are same.

2. If $y = a + bx$ be the relation between two variables x and y, then $\sigma_y = |b| \sigma_x$ where σ_x and σ_y are standard deviations of variables x and y respectively and a and b are constants.

Proof. Let x be a variable with values $x_1, x_2, ..., x_n$ with respective frequencies $f_1, f_2, ..., f_n$ and y be a variable with values $y_1, y_2, ..., y_n$ so that $y_i = a + bx_i$, i= 1, 2, ..., n.

$$\text{Now, } \bar{y} = \frac{1}{N}\sum_{i=1}^{n} f_i y_i = \frac{1}{N}\sum_{i} f_i (a + bx_i) = \frac{1}{N}\left[aN + b\sum_{i=1}^{n} f_i x_i \right] = a + b\bar{x}$$

$$\text{and } \sigma_y = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i(y_i - \bar{y})^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i(a + bx_i - a - b\bar{x})^2}$$

$$= \sqrt{\frac{b^2}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2} = |b|\sigma_x.$$

The result shows that the standard deviation is independent of change of origin but depends on change of scale.

Note : If $x_i = \frac{y_i - a}{b}$ and σ_x = standard deviation of x, then σ_y = standard deviation of y = $|b|\,\sigma_x$. Here a is origin and b is scale.

3. Let two groups of observations of sizes n_1 and n_2 be $x_{11}, x_{12}, ..., x_{1n_1}$, and $x_{21}, x_{22}, ..., x_{2n_2}$, with standard deviations σ_1 and σ_2 respectively.

Then the variance of the combined group is

$$\sigma^2 = \frac{n_1\sigma_1^{\,2} + n_2\sigma_2^{\,2} + n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2}{n_1 + n_2}$$

where $\sigma_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2$,

$$\sigma_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2,$$

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}x_{1i}\,,\ \bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}x_{2i} \text{ and } \bar{x} = \text{combined mean} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}.$$

Proof. $(n_1 + n_2)\sigma^2 = \sum_{i=1}^{n_1}(x_{1i} - \bar{x})^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x})^2$

$$= \sum_{i=1}^{n_1}[(x_{1i} - \bar{x}_1) + (\bar{x}_1 - \bar{x})]^2 + \sum_{i=1}^{n_2}[(x_{2i} - \bar{x}_2) + (\bar{x}_2 - \bar{x})]^2$$

$$= \sum_{i=1}^{n_1}[(x_{1i} - \bar{x}_1)^2 + (\bar{x}_1 - \bar{x})^2 + 2(\bar{x}_1 - \bar{x})(x_{1i} - \bar{x}_1)]$$

$$+ \sum_{i=1}^{n_2}\left[(x_{2i} - \bar{x}_2)^2 + (\bar{x}_2 - \bar{x})^2 + 2(x_2 - \bar{x})(x_{2i} - \bar{x}_2)\right]$$

$$= \sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + n_1(\bar{x}_1 - \bar{x})^2 + 2(\bar{x}_1 - \bar{x})\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)$$

$$+ \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2 + n_2(\bar{x}_2 - \bar{x})^2 + 2(\bar{x}_2 - \bar{x})\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)$$

$$= n_1\sigma_1^2 + n_1(\bar{x}_1 - \bar{x})^2 + 0 + n_2\sigma_2^2 + n_2(\bar{x}_2 - \bar{x})^2 + 0$$

$$\left(\textbf{Since } \sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1) = n_1\bar{x}_1 - n_1\bar{x}_1 = 0 \text{ and } \sum_{i=1}^{n_2}(x_{2i}-\bar{x}_2) = n_2\bar{x}_2 - n_2\bar{x}_2 = 0.\right)$$

$$= n_1\sigma_1^2 + n_2\sigma_2^2 + n_1(\bar{x}_1-\bar{x})^2 + n_2(\bar{x}_2-\bar{x})^2$$

$$\textbf{or} \qquad \sigma^2 = \frac{n_1\sigma_1^{\,2} + n_2\sigma_2^{\,2} + n_1(\bar{x}_1-\bar{x})^2 + n_2(\bar{x}_2-\bar{x})^2}{n_1+n_2}.$$

This is also called the formula for combined or composite variance. From this the formula for combined or composite standard deviation can be derived.

Note. 1. As $\bar{x} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2}$, $\bar{x}_1 - \bar{x} = \bar{x}_1 - \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2} = \dfrac{n_2(\bar{x}_1-\bar{x}_2)}{n_1+n_2}$ and

$$\bar{x}_2 - \bar{x} = \bar{x}_2 - \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2} = \frac{n_1(\bar{x}_2-\bar{x}_1)}{n_1+n_2}.$$

Therefore, $n_1(\bar{x}_1-\bar{x})^2 + n_2(\bar{x}_2-\bar{x})^2 = \dfrac{n_1n_2^2(\bar{x}_1-\bar{x}_2)^2}{(n_1+n_2)^2} + \dfrac{n_2n_1^2(\bar{x}_1-\bar{x}_2)^2}{(n_1+n_2)^2}$

$$= \frac{n_1n_2(\bar{x}_1-\bar{x}_2)^2(n_1+n_2)}{(n_1+n_2)^2} = \frac{n_1n_2(\bar{x}_1-\bar{x}_2)^2}{(n_1+n_2)}$$

So the expression for σ^2 can be written as

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1+n_2} + \frac{n_1n_2(\bar{x}_1-\bar{x}_2)^2}{(n_1+n_2)^2}.$$

Note. 2. Let there be k groups of values, where the i-th group has n_i values with mean $\bar{x}_i$ and standard deviation σ_i, i = 1, 2, ..., k. The variance, σ^2, of the combined group is obtained from the following expression :

$$\left(\sum_{i=1}^{k} n_i\right)\sigma^2 = \sum_{i=1}^{k} n_i\sigma_i^{\,2} + \sum_{i=1}^{k} n_i(\bar{x}_i-\bar{x})^2.$$

Result 1. Standard deviation is the smallest root mean square deviation.

Proof. Suppose a variable x assumes n values $x_1, x_2, ..., x_n$. The root mean square deviation about an arbitrary value c is

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-c)^2}$$

Now, $\sum_{i=1}^{n}(x_i-c)^2 = \sum_{i=1}^{n}\left[(x_i-\bar{x})+(\bar{x}-c)\right]^2$

$$= \sum_{i=1}^{n}\left[(x_i-\bar{x})^2 + (\bar{x}-c)^2 + 2(\bar{x}-c)(x_i-\bar{x})\right].$$

$$= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - c)^2 + 2(\bar{x} - c)\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - c)^2$$

$$[\text{since } \sum_{i=1}^{n}(x_i - \bar{x}) = n\bar{x} - n\bar{x} = 0.]$$

Therefore, $\sum_{i=1}^{n}(x_i - c)^2 \geq \sum_{i=1}^{n}(x_i - \bar{x})^2$ [since $n > 0$ and $(\bar{x} - c)^2 \geq 0$]

and equality sign holds when $c = \bar{x}$.

$$\therefore \quad \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2} \geq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

So the standard deviation is the smallest root mean square deviation.

Result 2. Standard deviation cannot be smaller than the mean deviation about mean.

Proof. Let $y_1, y_2, ..., y_n$ be n real values and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Then $\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \geq 0$ or $\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 \geq 0$ or $\frac{1}{n}\sum_{i=1}^{n} y_i^2 \geq \bar{y}^2$.

Consider $y_i = |x_i - \bar{x}|$ for i = 1,2,3,…, n.

Then $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq \left(\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|\right)^2$ or $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \geq \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$.

i.e. $s.d. \geq MD_{\bar{x}}$.

The equality sign holds either when all the values of the variable are equal or when the variable takes only two distinct values.

Merits and demerits :

The standard deviation is the best measure of dispersion. It is based on all the observations. It is amenable to algebraic treatment and is least affected by sampling fluctuation than most other measures of dispersion. It is rigidly defined and easily understood. It is easy to compute. It is affected by large or small extreme values.

1.3.4 Quartile deviation (Q.D.)

This is a measure of dispersion using quartiles. If Q_1 and Q_3 are the first and third quartiles respectively, then $Q.D. = \frac{1}{2}(Q_3 - Q_1)$. It is also called semi-inter-quartile range (SIQR).

1.3.5 Relative Measures of Dispersion

These measures are independent of units of measurement. These measures are used to compare the consistency of two distributions expressed in different units or some times even in the same unit. Some relative measures of dispersion and their expressions are :

Coefficient of variation $= \frac{\text{standard deviation}}{\text{mean}}$.

Coefficient of quartile deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$.

Coefficient of mean deviation $= \frac{\text{mean deviation about mean}}{\text{mean}}$ or $\frac{\text{mean deviation about median}}{\text{median}}$. These values are multiplied by 100 to get measurements in percentages.

1.4 Worked out examples

Example 1. Calculate arithmetic mean, geometric mean, harmonic mean, median, range and mean deviation about median, standard deviation, quartile deviation, coefficient of variation, coefficient of mean deviation of the following data : 50, 54, 72, 82, 72, 93

Solution : $A.M. = \frac{50+54+72+82+72+93}{6} = \frac{423}{6} = 70.5$

Let G. M. $= G$. Then $\log G = \frac{1}{6}[\log 50 + \log 54 + 2\log 72 + \log 82 + \log 93]$

$\left[\text{since } G = \sqrt[6]{50 \times 54 \times 72 \times 82 \times 72 \times 93}\right]$

or log G = (1/6) [1.6990 + 1.7324 + 2 × 1.8573 + 1.9138 + 1.9685] $= \frac{11.0283}{6}$

= 1.83805

So G = antilog 1.83805 = 68.8732.

H.M. $= \frac{6}{\frac{1}{50}+\frac{1}{54}+\frac{2}{72}+\frac{1}{82}+\frac{1}{93}} = \frac{6}{.02+.0185+.0278+.0122+.0108}$

$= \frac{6}{.0893} = 67.1892$

When the values are arranged in increasing order they become

50, 54, 72, 72, 82, 93 . . . . . (1)

Then the two middle most values are 72 and 72. So median $= \frac{72+72}{2} = 72$

Mode = value which occurs most no. of times = 72

Range = maximum value – minimum value = 93 – 50 = 43.

Mean deviation about median $= \frac{1}{6}[|50-72|+|54-72|+2\,x\,|72-72|+|82-72|+$ $|93-72|]$

$$= \frac{1}{6}[22+18+0+10+21] = \frac{71}{6} = 11.8333$$

$$\text{s.d.} = \sqrt{\frac{1}{6}\left[50^2+54^2+2\times 72^2+82^2+93^2\right]-(70.5)^2}$$

$$= \sqrt{\frac{1}{6}\left[2500+2916+10368+6724+8649\right]-4970.25}$$

$$= \sqrt{\frac{31157}{6}-4970.25} = \sqrt{5192.8333-4970.25} = \sqrt{222.5833} = 14.92$$

Now $\frac{n}{4} = \frac{6}{4} = 1\cdot 5$, $\frac{3n}{4} = \frac{18}{4} = 4\cdot 5$. So the 2nd value in ordered series (1) is Q_1 i.e., $Q_1 = 54$ and the 5th value in ordered series (1) is Q_3 i.e., $Q_3 = 82$

$\therefore$ Quartile deviation $\frac{Q_3 - Q_1}{2} = \frac{82-54}{2} = \frac{28}{2} = 14.$

Coefficient of variation $= \frac{\text{s.d}}{\text{mean}} = \frac{14\cdot 92}{70\cdot 5} = 0.2116$. Coefficient of mean deviation $= \frac{11\cdot 8333}{72} = 0.1644.$

Example 2. Calculate arithmetic mean, geometric mean, harmonic mean, median, mode, standard deviation, mean deviation about median, range, quartile deviation, coefficient of mean deviation of the following frequency distribution.

Marks	10	20	30	40	50	60
No. of students :	8	12	20	10	7	3

Solution.

							Total
Marks (x)	10	20	30	40	50	60	
No. of students (f)	8	12	20	10	7	3	60
fx	80	240	600	400	350	180	1850
f log x	8	15.6124	29.5424	16.0206	11.8928	5.3345	86.4027
$\frac{f}{x}$	.8	.6	.6667	.25	.14	.05	2.5067
Less than type cumulative freq	8	20	40	50	57	60	
fx^2	800	4800	18000	16000	17500	10800	67900
$f\|x-30\|$	160	120	0	100	140	90	610

$\bar{x}$ = arithmetic mean $= \frac{\sum fx}{N} = \frac{1850}{60} = 30.83.$

$\log G = \frac{1}{N}\sum f \log x = \frac{86 \cdot 4027}{60} = 1.440045.$ So G = geometric mean = 27.5451.

$H = \frac{N}{\sum \frac{f}{x}} = \frac{60}{2 \cdot 5067} = 23 \cdot 9359.$

Median = average of marks for which cumulative frequency (less than type) are 30 and 31

= average of marks of the 30th and the 31st student

$= \frac{30+30}{2} = 30.$

Mode = mark for which frequency is maximum = 30.

$s.d. = \sqrt{\frac{1}{N}\sum fx^2 - \bar{x}^2} = \sqrt{\frac{67900}{60} - (30 \cdot 83)^2} = \sqrt{1131 \cdot 6667 - 950 \cdot 4889}$

$= \sqrt{181 \cdot 1778} = 13 \cdot 46.$

Mean deviation about median $= \frac{1}{N}\sum f|x-30| = \frac{610}{60} = 10 \cdot 17.$

Range = 60 – 10 = 50.

1st Quartile, Q_1 = Value for which CF (less than type) is $\frac{N}{4}$ (i.e., 15) = 20.

3rd Quartile, Q_3 = Value for which CF (less than type) is $\frac{3N}{4}$ (i.e. 45) = 40.

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2} = \frac{40 - 20}{2} = 10.$$

$$\text{Coefficient of variation} = \frac{\text{s.d.}}{\text{mean}} \times 100 = \frac{13 \cdot 46}{30 \cdot 83} \times 100 = 43 \cdot 66\%.$$

$$\text{Coefficient of mean deviation} = \frac{\text{mean deviation about median}}{\text{median}} = \frac{\frac{1}{N}\sum f|x - 30|}{30}$$

$$= \frac{10 \cdot 17}{30} = 0 \cdot 339 = 33 \cdot 9\%.$$

Example 3. Calculate arithmetic mean, geometric mean, harmonic mean, median, mode, standard deviation, mean deviation about mean, quartile deviation, coefficient of variation, coefficient of mean deviation, coefficient of quartile deviation from the following frequency distribution of height (in cm.) of 70 persons.

Heights (in cm.)	126-135	136-145	146-155	156-165	166-175	176-185
No. of persons	7	10	14	23	12	4

Solution. Here variable is height (in cm.) of a person and frequency is number of persons.

Class of heights (cm.)	126-135	136-145	146-155	156-165	166-175	176-185	Total
Class boundaries (cm.)	125.5-135.5	135.5-145.5	145.5-155.5	155.5-165.5	165.5-175.5	175.5-185.5	
mid-point (x)	130.5	140.5	150.5	160.5	170.5	180.5	
Frequency (f)	7	10	14	23	12	4	70
cum. freq (less than type) (F)	7	17	31	54	66	70	
$u = \frac{x - 150 \cdot 5}{10}$	–2	–1	0	1	2	3	
fu	–14	–10	0	23	24	12	35
fu^2	28	10	0	23	48	36	145
f log x	14.80927	21.47676	30.48551	50.72593	26.78069	9.02591	153.30407
$\frac{f}{x}$	0.05364	0.07117	0.09302	0.14330	0.07038	0.02216	0.45367
f |x–155.5|	175	150	70	115	180	100	790

$$\bar{x} = \text{a.m.} = 150 \cdot 5 + 10 \times \frac{\sum fu}{N} = 150 \cdot 5 + 10 \times \frac{35}{70} = 150 \cdot 5 + 5 = 155 \cdot 5 \text{ cm.}$$

$$\log G = \frac{1}{N}\sum_{i=1}^{n} f_i \log x_i = \frac{153\cdot 30407}{70} = 2\cdot 19006 \;\therefore\; G = \text{g.m.} = \text{antilog } 2.19006$$

$$= 154.90306 \text{ cm.},$$

$$\text{H.M.} = H = \frac{N}{\sum_i \frac{f_i}{x_i}} = \frac{70}{\cdot 45367} = 154\cdot 2972\,\text{cm}.$$

Median class is 155.5 –165.5 and its frequency is $f_m = 23$

$$\text{Median} = L + \frac{\frac{N}{2} - F}{f_m} \times c = 155\cdot 5 + \frac{35-31}{23}\times 10 = 155\cdot 5 + \frac{40}{23} = 155\cdot 5 + 1\cdot 74$$

$$= 157\cdot 24 \text{ cm.}$$

Modal class is 155.5 – 165.5. Its frequency $f_{m_o} = 23$, frequency of former and later classes are :

$$f_{m_o} = 12 \text{ and } f_{m_o-1} = 14.\ \text{Mode} = L + \frac{f_{m_o} - f_{m_o-1}}{2f_{m_o} - f_{m_o-1} - f_{m_o+1}} \times c$$

$$= 155\cdot 5 + \frac{23-14}{46-14-12}\times 10$$

$$\therefore \quad \text{Mode} = 155\cdot 5 + \frac{90}{20} = 155\cdot 5 + 4\cdot 5 = 160\,\text{cm}.$$

$$\text{Standard Deviation} = \sqrt{\frac{1}{N}\sum f_i u_i^{\,2} - \bar{u}^2} \times b = \sqrt{\frac{145}{70} - \left(\frac{35}{70}\right)^2} \times 10$$

$$= \sqrt{2\cdot 07143 - 0\cdot 25} \times 10$$

$$= \sqrt{1\cdot 82143} \times 10 = 1\cdot 34960 \times 10 = 13\cdot 496 \text{ cm}.$$

Mean deviation about mean

$$= \frac{1}{N}\sum f\,|x - \text{mean}| = \frac{1}{70}\sum f\,|x - 155\cdot 5| = \frac{790}{70} = 11\cdot 2857 \text{ cm}.$$

First quartile class is 145.5 – 155.5, 3rd quartile class is 155.5 – 165.5

$$\text{since } \frac{N}{4} = 17\cdot 5 < 31 \text{ and } \frac{3N}{4} = 52\cdot 5 < 54.$$

$$Q_1 = L + \frac{\frac{N}{4} - F}{f_{Q_1}} \times 10 = 145\cdot 5 + \frac{17\cdot 5 - 17}{14}\times 10 = 145\cdot 5 + \frac{5}{14} = 145\cdot 5 + \cdot 36 = 145\cdot 86\,\text{cm}.$$

$$Q_3 = L + \frac{\frac{3N}{4} - F}{f_{Q_3}} \times 10 = 155 \cdot 5 + \frac{52 \cdot 5 - 31}{23} \times 10 = 155 \cdot 5 + \frac{215}{23} = 155 \cdot 5 + 9 \cdot 35 = 164 \cdot 85 \text{cm.}$$

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2} = \frac{164 \cdot 85 - 145 \cdot 86}{2} = \frac{18 \cdot 99}{2} = 9 \cdot 495 \text{cm.}$$

$$\text{Coefficient of variation} = \frac{\text{s.d.}}{\text{mean}} \times 100 = \frac{13 \cdot 496 \times 100}{155 \cdot 5} = 0 \cdot 0868 \times 100 = 8 \cdot 68\%.$$

$$\text{Coefficient of mean deviation} = \frac{\text{mean deviation about mean}}{\text{mean}} = \frac{11 \cdot 2857}{155 \cdot 5}$$

$$= \cdot 0726 = 7 \cdot 26\%.$$

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{164 \cdot 85 - 145 \cdot 86}{164 \cdot 85 + 145 \cdot 86} = \frac{18 \cdot 99}{310 \cdot 71}$$

$$= \cdot 0611 = 6 \cdot 11\%$$

Example 4. A sample of size 10 has mean 3 and standard deviation (s.d.) 2. Another sample of size 15 has mean 8 and s.d. 4. If the two samples are taken together, find the mean and the s.d. of the combined sample.

Solution :

For the first sample size $= n_1 = 10.$

mean $= \bar{x}_1 = 3$

s.d. $= s_1 = 2.$

For the second sample size $= n_2 = 15$, mean $= 8$, s.d. $= s_2 = 4$

$$\text{For combined sample mean} = \bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{10 \times 3 + 15 \times 8}{10 + 15} = \frac{150}{25} = 6.$$

$$\text{s.d.} = s = \sqrt{\frac{n_1 {s_1}^2 + n_2 {s_2}^2 + n_1 (\bar{x}_1 - \bar{x})^2 + n_2 (\bar{x}_2 - \bar{x})^2}{n_1 + n_2}}$$

$$= \sqrt{\frac{10 \times 2^2 + 15 \times 4^2 + 10(3-6)^2 + 15(8-6)^2}{10 + 15}} = \sqrt{\frac{40 + 240 + 90 + 60}{25}} = \sqrt{\frac{430}{25}}$$

$$= \sqrt{17 \cdot 2} = 4 \cdot 15$$

Example 5. For n positive real observations prove that A.M. ≥ G.M. ≥ H.M.

Solution : For two observations x_1 and x_2 which are positive and real, we may write, $\left(\sqrt{x_1} - \sqrt{x_2}\right)^2 \geq 0$ since the square of a real quantity can not be negative, the equality sign will hold good when $x_1 = x_2$.

i.e., $x_1 + x_2 - 2\sqrt{x_1 x_2} \geq 0$.

Hence $x_1 + x_2 \geq 2\sqrt{x_1 x_2}$ or $\frac{x_1 + x_2}{2} \geq \sqrt{x_1 x_2}$ ... (1)

So for n = 2, A.M. ≥ G.M., the equality sign will hold good when $x_1 = x_2$.

Now, for four positive real observations x_1, x_2, x_3 and x_4

$$\frac{\frac{x_1 + x_2}{2} + \frac{x_3 + x_4}{2}}{2} \geq \sqrt{\frac{x_1 + x_2}{2} \cdot \frac{x_3 + x_4}{2}} \text{ and } \frac{x_1 + x_2}{2} \geq \sqrt{x_1 x_2}$$

$$\text{and } \frac{x_3 + x_4}{2} \geq \sqrt{x_3 x_4}$$

i.e. $\frac{x_1 + x_2 + x_3 + x_4}{4} \geq \sqrt{\sqrt{x_1 x_2} \cdot \sqrt{x_3 x_4}} = \left(x_1 x_2 x_3 x_4\right)^{\frac{1}{4}}$.

Thus for n = 4, A.M. ≥ G.M., the equality sign will be valid when $x_1 = x_2 = x_3 = x_4$.

Proceeding in this way A.M. ≥ G.M. can be shown for n = 2, 4, 8, 16 etc. So when $n = 2^m$, where m is a positive integer, A.M. ≥ G.M. for n positive real observations x_1, x_2, ..., x_n.

Now, consider $2^{m-1} < n < 2^m$ i.e. n lies between 2^{m-1} and 2^m.

Consider $2^m = N$ and $A = \frac{x_1 + x_2 + \ldots + x_n}{n}$ = A.M. of the n given values.

Consider N real values of which first n are x_1, x_2, ..., x_n and last (N – n) values are equal to A.

Since for 2^m observations A.M. ≥ G.M. we may write

$$\frac{x_1 + x_2 + \ldots + x_n + (N-n)A}{N} \geq (x_1 x_2 \ldots x_n \cdot A \cdot A \ldots\ldots A)^{\frac{1}{N}}$$

i.e. $\frac{nA + (N-n)A}{N} \geq \left(G^n A^{N-n}\right)^{\frac{1}{N}}$

where G = geometric mean of x_1, x_2, ..., $x_n = (x_1 x_2, x_3 \ldots x_n)^{\frac{1}{n}}$,

i.e., $G^n = x_1 x_2 \ldots x_n$.

So $\frac{NA}{N} \geq \left(G^n A^{N-n}\right)^{\frac{1}{N}}$

or $A^N \geq G^n A^{N-n}$

or $\frac{A^N}{A^{N-n}} \geq G^n$

or $A^n \geq G^n$

or $A \geq G$.

So when $2^{m-1} < n < 2^m$, then also A.M. $\geq$ G.M.

Hence A.M. $\geq$ G.M. for any number of positive and real observations. ... (i)

As $x_1, x_2, ..., x_n$ are real and positive, $\frac{1}{x_1}, \frac{1}{x_2}, \ldots\ldots, \frac{1}{x_n}$ are also real and positive. Then for any positive integral value of n, A.M. $\geq$ G.M.

i.e. $$\frac{\frac{1}{x_1} + \frac{1}{x_2} + \ldots\ldots + \frac{1}{x_n}}{n} \geq \left(\frac{1}{x_1} \cdot \frac{1}{x_2} \ldots\ldots \frac{1}{xn}\right)^{\frac{1}{n}} = \frac{1}{\left(x_1 x_2 \ldots x_n\right)^{\frac{1}{n}}}$$

or $$\frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots\ldots + \frac{1}{x_n}} \leq \left(x_1 x_2 \ldots x_n\right)^{\frac{1}{n}}$$

or H.M. $\leq$ G.M. ... (ii)

Combining these two results, (i) and (ii) A.M. $\geq$ G.M. $\geq$ H.M. holds for any n real positive observations.

Example 6. Two workers on the same job show the following results over a long period of time :

	Worker A	Worker B
Mean time of completing the job (in minutes)	30	25
Standard Deviation (in minutes)	6	4

(a) Which worker appears to be more consistent in time to complete the job?

(b) Which worker appears to be faster in completing the job?

Solution : The consistency of a worker can be judged by computing the coefficient of variation. So let us compute the C.V. of worker A and B.

(a) Coefficient of variation for worker A is $\frac{\text{s.d.}}{\text{mean}} \times 100 = \frac{6}{30} \times 100 = 20\%$ and coefficient of variation for worker B is $\frac{\text{s.d.}}{\text{mean}} \times 100 = \frac{4}{25} \times 100 = 16\%$. As $16\% < 20\%$, worker B appears to be more consistent.

(b) Worker B takes 25 minutes on an average as against 30 minutes taken by worker A to complete the job. So worker B appears to be faster in completing the job.

1.5 Summary

This chapter discussed the concepts of central tendency and dispersion of data, their different measures along with formulae, important properties, their merits and demerits and some useful results. Illustrated examples are given to use these measures and their formulae.

1.6 Exercises

1. Explain what is meant by central tendency of data. What are the common measures of central tendency?
2. Define Arithmetic mean, Median and Mode and discuss their merits and demerits as measures of central tendency.
3. Define Arithmetic mean, Geometric mean and Harmonic mean and compare their relative advantages and disadvantages.
4. Prove that for two positive real quantities

   (i) A.M. $\geq$ G.M. $\geq$ H.M., (ii) G.M. = $\sqrt{\text{A.M.} \times \text{H.M.}}$.
5. Explain what you mean by central tendency of a frequency distribution. Compare mean, median and mode as measures of central tendency.
6. Prove that for n positive and real observations

   A.M. $\geq$ G.M. $\geq$ H.M.

   When will the equality sign hold?
7. Show that the combined arithmetic mean $(\bar{x})$ of the two groups lies between arithmetic means $\bar{x}_1$ and $\bar{x}_2$ of two groups.

8. If a variate assumes n values a, ar, ar^2, ..., ar^{n-1} ($r < 1$) and with equal frequencies, then find arithmetic mean (A), geometric mean (G) and harmonic mean (H) and show that $AH = G^2$.

9. Show that the sum of the deviations of observations about mean is zero.

10. Find A.M., G.M., H.M., Median and Mode of the following numbers :

    (i) 4, 6, 8, 8, 12, 72, (ii) 4, 6, 6, 12, 16.

11. Find the A.M., G.M., H.M., Median, Mode, First and Third quartiles of the following data :

Daily wages (Rs) :	10	15	20	25	30	35
No. of workers :	5	12	16	14	10	2

12. Find A.M., Median, First quartile and Mode of the following distribution :

Marks :	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
No. of students :	10	20	35	25	10

13. From the following cumulative frequency distribution of marks of 22 students in Mathematics, calculate Arithmetic mean, Median, Mode and First quartile.

Marks :	Below 10	Below 20	Below 30	Below 40	Below 50
No. of students :	10	20	35	25	10

14. The frequency distribution of expenditure of 1000 families is given below :

Expenditure (Rs) :	40 – 59	60 – 79	80 – 99	100 – 119	120 – 129
No. of families :	50	?	500	?	50

    The mean and median of the distribution are equal to Rs. 87.50. Determine the missing frequencies.

15. Explain with suitable examples the term 'dispersion'. State the absolute and relative measures of dispersion.

16. State the requisites of a satisfactory measure of dispersion and in their light examine any two common measures of dispersion.

17. What do you mean by dispersion of a frequency distribution? Compare range, mean deviation and standard deviation as measures of dispersion.

18. Prove that standard deviation is independent of change of origin but is dependent on change of scale.

19. For simple frequency distribution prove that the root mean square deviation is least when deviations are taken about mean.

20. The size, mean and variances of one set of values are $n_1, \bar{x}_1, s_1^2$ and those of another set of values are $n_2, \bar{x}_2, s_2^2$ respectively. When the two sets are pooled together, find the mean and variance of the combined set.

21. Find the range, mean deviation about median and standard deviation of the following observations :

    13, 84, 68, 24, 96, 109, 84, 27.

22. Prove that the mean deviation about median is least.

23. Find the mean deviation about arithmetic mean (μ), quartile deviation and standard deviation (σ) from the following data :

Weight (in kg.)	35.0–39.9	40.0–44.9	45.0–49.9	50–54.9	55–59.9	60–64.9	65–69.9
No. of students	5	16	30	23	17	8	1

    (i) Find the number of students weighing between $\mu \pm 2\sigma$.

    (ii) Find also the coefficient of variation of their weights.

24. The number of employees, average wages per employee and the standard deviations of the wages per employee for two factories are given below :

	Factory A	Factory B
Number of employees :	50	100
Average wage per employee per month (Rs) :	120	85
s. d. of wages per employee per month (Rs) :	3	4

    (a) Which factory has more total monthly wage?

    (b) In which factory there is less variation in the distribution of wages per employee?

    (c) Suppose in factory B, the wage of an employee was wrongly noted as Rs. 120.00 instead of Rs. 100. What would be the correct standard deviation for factory B?

25. Show that the standard deviation, s, of a set of observations $x_1, x_2, ..., x_n$ is given by

$$s^2 = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2, \; i \neq j.$$

    Hence or otherwise examine the consequences of (i) adding a constant value and (ii) multiplying a constant value to all the observations.

26. A person was asked to calculate the mean and s. d. of 50 observations. He noticed that 4 of the 50 observations were zeroes and hence ignored them and calculated the mean and s. d. of the remaining 46 observations as 42.96 and 13.8812. Find the correct values of mean and s. d. of all the 50 observations.

27. For a set of 250 observations on a certain variable x the mean and standard deviation are 65.7 and 4.4 respectively. However, after scrutinising the data it is found that two observations 71 and 83, had been wrongly recorded as 91 and 80. Obtain the correct values of the mean and standard deviation.

28. Let S and R be the standard deviation and range of a set of n values of a variable x respectively. Show that

$$\frac{R^2}{2n} \leq s^2 \leq \frac{R^2}{4}.$$

When does the equality hold?

29. What are the desiderata of a good measure of central tendency? Compare the mean, the median and the mode in the light of these desiderata. Why is the arithmetic mean called the best measure of central tendency?

30. Give some examples where the geometric mean or the harmonic mean would be the appropriate type of average.

31. What is meant by relative dispersion? Define the coefficient of variation and explain its uses.

32. Obtain the mean and standard deviation of the first n natural numbers.

33. The number of runs scored by cricketers X and Y during a test series consisting of 5 test macthes has been shown below for each of the 10 innings :

Cricketer X : 5, 26, 97, 76, 112, 89, 6, 106, 24, 16.

Cricketer Y : 51, 47, 36, 60, 58, 39, 44, 42, 71, 50.

Make a comparative study of their batting performance.

34. Prove that for any set of values $x_1, x_2, x_3, ..., x_n$,

$$x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2 \geq \frac{(x_1 + x_2 + x_3 + \cdots + x_n)^2}{n}.$$

35. Standard deviations of two series are 15 and 18 and their coefficients of variation are 75% and 90% respectively. Find their arithmetic means.

36. The arithmetic mean and the standard deviation of a set of 9 items are 43 and 5 respectively. If an item of value 63 is added to the set, find the mean and the standard deviation of all the items.

37. Define ‘arithmetic mean’ and ‘geometric mean’ and mention their relative advantages and disadvantages.

38. Let the mean monthly wages per worker of two firms employing 40 and 60 workers be Rs. 850 and Rs. 880 respectively. What is the mean monthly wages per worker for the two firms taken together?

39. The average marks obtained in an examination by two groups of students were found to be 75 and 85 respectively. Determine the ratio of students in the two groups, if the average marks for all students was 80.

40. Find the arithmetic mean and standard deviation of the first ten natural numbers.

41. Find the harmonic mean of the following distribution :

x	2	4	6	8
f	3	5	4	2

42. How, in your opinion, will an average change when all the values of the variable increase or decrease (i) by the same amount, (ii) in the same proportion?

43. The arithmetic mean of two observations is 127.5 and their geometric mean is 60. Find (i) their harmonic mean and (ii) the two observations.

44. The mean marks of 100 students was found to be 40. Later on it was discovered that a score of 53 was misread as 83. Find the corrected mean corresponding to the corrected score.

45. Out of 400 observations, 100 observations have the value one and the rest of the observations are zero. Find the mean and S.D. of 400 observation taken together.

46. For a distribution of 85 observations, the mean and standard deviation were found to be 57 and 3.5 respectively. On checking it was discovered that two observations which should correctly read as 36 and 69, had been wrongly recorded as 46 and 76 respectively. Calculate the correct standard diviation.

47. A group of 100 items have mean 55 and S.D. 6. If the mean and S.D. of 40 of the items be 61 and 4.5 respectively, find the mean and S.D. of the other 42 items.

48. The scores of two golfers for 10 rounds each are :

A : 58 59 60 54 65 66 52 75 69 52
B : 84 56 92 65 86 78 44 54 78 68

Which player may be regarded as the more consistent one?

1.7 Suggested Readings

1. Gun A. M ; Gupta, M. K. and Dasgupta, B. *Fundamentals of Statistics* Vol. 1, The World Press Pvt. Ltd. 2002, Kolkata.
2. Chaudhuri, S. B. *Elementary Statistics* Vol. 1, Shraddha Prakashani, 1986.
3. Kenney, J. F. and Keeping, E. S. *Mathematics of Statistics Part 1*, Van Norstrand 1954 and Affiliated East West Press.
4. Gupta, S. C. and Kapoor, V. K. *Fundamentals of Mathematical Statistics Vol. 1,* Sultan Chand and Sons, New Delhi, 1989.
5. Yule, G. U. and Kendall M. G. *Introduction to the Theory of Statistics*, Charles Griffin, 1953.
6. Mills, F. G. *Statistical Methods,* H. Holt 1955.

Unit 2 ❑ Correlation and Regression Analysis

Structure

2.0 Objectives

In Chapter 1 different characteristics of a single variable have been considered. But in many situations it is seen that two or more variables are so related that values

of one variable are influenced by the values of another variable or by the values of several other variables. In case of two variables, increase or decrease in values of one variable may influence the increase or decrease in the values of other variables, i.e. movement of one variable may influence the movement of several other variables. In that case the two variables are said to be correlated. The degree of relationship between the variables under consideration is measured through correlation analysis. This idea can be extended to the case when movement of values of a set of variables may influence the movement of another single variable. Then multiple correlation analysis is considered to measure the degree of association of one variable with the other remaining variables. So the degree of association of the variables, their measurements, properties, predicted relations are discussed clearly alongwith worked out examples.

2.1 Introduction

Correlation of two variables measures the degree of relationship between them. If the values of the two variables vary in the same direction, the correlation is said to be 'positive'. If the values vary in opposite direction the correlation is negative. For examples, height and weight of a child, income and expenditure of a family etc. are positively correlated. If the variations are in opposite direction then the correlation is negative and the two variables are said to be negatively correlated. For example, price and demand, volume and pressure are negatively correlated. If the variations can not be described as in above two cases then the two variables are said to have zero correlation or they are said to be uncorrelated.

If corresponding to unit change in one variable constant change (approximate or exact) occurs in the other variable then the correlation between the two variables is said to be linear otherwise it may be non-linear. This linear correlation is measured by the correlation coefficient, r, which represents the correlation coefficient between two variables. This linear correlation between the two variables is called simple correlation.

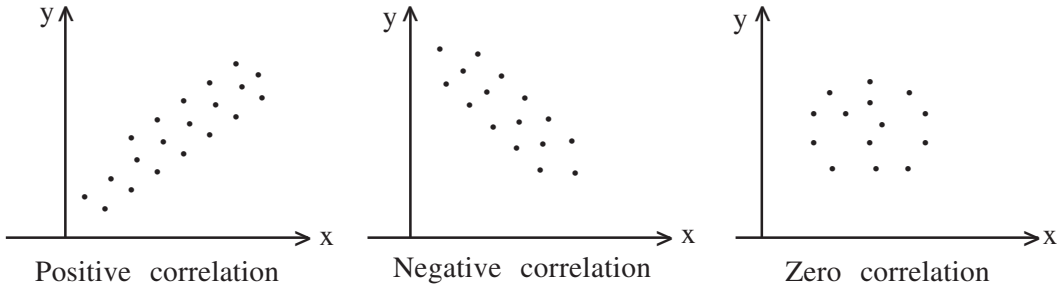
The joint distribution of three or more variables is known as multivariate distribution.

In multiple regression linear dependence of some variable on two or more other variables is considered.

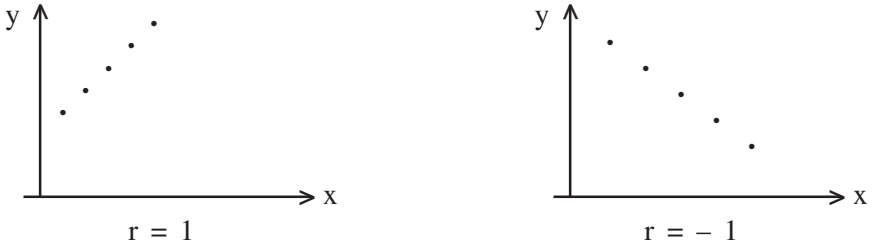
2.2 Scatter Diagram

For bivariate data each pair of values of variables x and y are plotted as a dot or point on a graph paper. This diagram of bivariate data is called a dot diagram or a scatter diagram.

If the scatter diagram shows some trend either upward (i.e., increase of one variable results in increase of the other variable) or downward (i.e., increase of one variable gives decrease of the other variable) then the variables are said to be correlated. In the first case variables are positively correlated while in the second case they are negatively correlated.



If the points lie exactly on a straight line and the slope of the line is positive, then the scatter diagram shows positive perfect correlation and in that case r = + 1. If the points lie on a straight line and the slope of the line is negative then the scatter diagram shows negative perfect correlation i.e., r = – 1.



2.3 Correlation Coefficient (r)

In case of linear correlation a measure of correlation between two variables x and y is called the correlation coefficient between x and y and it is represented by r_{xy} or r_{yx} or, in short, r where

$$r = \frac{Cov(x,y)}{\sqrt{V(x)V(y)}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where Cov (x, y) = covariance between x and y $= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, V(x) = variance of $x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ and V(y) = variance of $y = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$. This r is called Pearson's product moment correlation coefficient between two variables x and y. This formula can be simplified to

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}x_i y_i - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2 - \bar{y}^2}} = \frac{\sum_i x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\sum_i x_i^2 - n\bar{x}^2}\sqrt{\sum_i y_i^2 - n\bar{y}^2}}.$$

2.3.1 Properties

1. As the numerator and the denominator of r has same unit, the correlation coefficient is independent of unit. Also as divisor n is used for both numerator and denominator of r, the correlation coefficient is independent of number of pairs of values of x and y.

2. Correlation coefficient is independent of change of origin and change of scale.

Proof : If $u = \frac{x-a}{b}$ and $v = \frac{y-c}{d}$ where a, b, c and d are arbitrary constants, a and c are called origin, while b and d are called scale.

Since, $u_i = \frac{x_i - a}{b}, v_i = \frac{y_i - c}{d}$ for i = 1, 2, ..., n, then $\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{n}\sum_i (a + bu_i)$ $= a + b\bar{u}$ and similarly, $\bar{y} = c + d\,\bar{v}$, so that

$$x_i - \bar{x} = b(u_i - \bar{u}), y_i - \bar{y} = d(v_i - \bar{v}) \text{ for } i = 1, 2, ..., n.$$

Thus $Cov(x,y) = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{bd}{n}\sum_i (u_i - \bar{u})(v_i - \bar{v}) = bd\,Cov(u,v).$

$$V(x) = \frac{1}{n}\sum_i (x_i - \bar{x})^2 = \frac{b^2}{n}\sum_i (u_i - \bar{u})^2 = b^2 V(u).$$

Similarly, V (y) = d^2 V (v).

Thus $r_{xy} = \frac{Cov(x,y)}{\sqrt{V(x)V(y)}} = \frac{bd\,Cov(u,v)}{\sqrt{b^2V(u)d^2V(v)}} = \frac{bd}{|b||d|} r_{uv}$.

If b and d are of same sign, $r_{xy} = r_{uv}$ while if b and d are of opposite sign, $r_{xy} = -r_{uv}$. So the result shows that the correlation coefficient is independent of change of origin and change of scale.

3. $-1 \le r \le 1$.

Proof : For n pairs of values of x and y i.e. (x_1, y_1), (x_2, y_2), ... , (x_n, y_n),

consider $u_i = \frac{x_i - \bar{x}}{s_x}$ and $v_i = \frac{y_i - \bar{y}}{s_y}$ for i = 1, 2, ..., n, where $\bar{x} = \frac{1}{n}\sum_i x_i$,

$\bar{y} = \frac{1}{n}\sum_i y_i, s_x^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$ and $s_y^2 = \frac{1}{n}\sum_i (y_i - \bar{y})^2$.

Then $\sum_i u_i^2 = \frac{\sum_i (x_i - \bar{x})^2}{s_x^2} = \frac{ns_x^2}{s_x^2} = n, \sum_i v_i^2 = \frac{\sum_i (y_i - \bar{y})^2}{s_y^2} = \frac{ns_y^2}{s_y^2} = n$ and

$$\sum_i u_i v_i = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{s_x\ s_y} = nr$$

Now $\sum_i (u_i \pm v_i)^2 = \sum_i u_i^2 + \sum_i v_i^2 \pm 2\sum_i u_i v_i = n + n \pm 2nr.$

$= 2n(1 \pm r).$

Now $\sum (u_i \pm v_i)^2 \ge 0$ since the square of a real quantity can not be negative.

We have $1 \pm r \ge 0$. This gives $1 \ge r$ and $r \ge -1$.

So $-1 \le r \le 1$.

Note : From the above, $1 \pm r = \frac{1}{2n}\sum_{2n} (u_i + v_i)^2$

or $r = -1 + \frac{1}{2n}\sum_i \left(\frac{x_i - \bar{x}}{s_x} + \frac{y_i - \bar{y}}{s_y}\right)^2$ and $r = 1 - \frac{1}{2n}\sum_i \left(\frac{x_i - \bar{x}}{s_x} - \frac{y_i - \bar{y}}{s_y}\right)^2$

considering plus and minus signs.

When r = 1, variables x and y are perfectly positively correlated and when r = –1, variables x and y are perfectly negatively correlated.

2.3.2 Bivariate Frequency Distribution

When the number of pairs of values of two variables x and y are considerably large then a two-way table representing bivariate frequency distribution can be formed by representing classes of one variable row-wise and classes of the other variable column-wise. f_{ij} denotes the frequency of observations whose x values belong to the i-th class of variable x and y values belong to the j-th class of variable y, i = 1, ..., m & j = 1, 2, ..., n.

Bivariate frequency distribution table of x and y

classes of variable y		classes of variable x						
		$x_1 - x_2$	$x_2 - x_3$	...	$x_i - x_{i+1}$	...	$x_m - x_{m+1}$	Total
	$y_1 - y_2$	f_{11}	f_{21}	...	f_{i1}	...	f_{m1}	$f_{.1}$
	$y_2 - y_3$	f_{12}	f_{22}	...	f_{i2}	...	f_{m2}	$f_{.2}$
	...	. .	. .	...	...	...	...	...
	$y_j - y_{j+1}$	f_{1j}	f_{2j}	...	f_{ij}	...	f_{mj}	$f_{.j}$
	...	. .	. .	...	...	...	. .	...
	$y_n - y_{n+1}$	f_{1n}	f_{2n}	...	f_{in}	...	f_{mn}	$f_{.n}$
	Total	$f_{1.}$	$f_{2.}$	...	$f_{i.}$	...	$f_{m.}$	N

For example, a bivariate frequency distribution of ages of husbands and wives (in years) is given below. Here the total number of pairs observed is 30.

Age of wives in years

Age of husbands in years	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80	Total
20 – 30	3						3
30 – 40	3	3					6
40 – 50		3	1				4
50 – 60		2	6	1			9
60 – 70				2	3	1	6
70 – 80					1	1	2
Total	6	8	7	3	4	2	30

2.3.3 Calculation of coefficient of correlation from Grouped Data

If, in a bivariate distribution, the data are fairly large then these are classified in the form of a two-way frequency table known as bi-variate table. The values of each variable are grouped into various classes. The number of classes in respect of each variable need not be same. Suppose there are m classes for the values of the variable x and n classes for the values of the variable y. In such a situation there will be m × n cells in the two-way table. Let x_i be the mid-point of the i-th class of x and y_j be the mid-point of the j-th class of y. These two classes define the (i, j)-th cell of the bivariate frequency table. In this cell the frequency will be denoted by f_{ij}. The coefficient of correlation is obtained by the formula

$$r_{xy} = \frac{\frac{1}{n}\sum_{i,j}(x_i - \bar{x})(y_j - \bar{y})f_{ij}}{\left\{\frac{1}{n}\sum_{i}(x_i - \bar{x})^2 f_{i0}\right\}^{\frac{1}{2}}\left\{\frac{1}{n}\sum_{j}(y_j - \bar{y})^2 f_{0j}\right\}^{\frac{1}{2}}}$$

where $f_{i0} = \sum_j f_{ij}$, $f_{0j} = \sum_i f_{ij}$ and $n = \sum_{i,j} f_{ij}$.

Naturally, $\bar{x} = \frac{1}{n}\sum_i x_i f_{i0}$ and $\bar{y} = \frac{1}{n}\sum_j y_j f_{0j}$.

For reducing computational labour, changes of origin and scale for both x and y may be taken. After these the new variables become

$$u_i = \frac{x_i - a}{b} \text{ and } v_j = \frac{y_j - c}{d}$$

where a, b, c and d are arbitrarily chosen constants, a and c being known as origin while b and d being known as scale.

However, it will be found advantageous to take as bases two class marks a and c, somewhere in the middle of the range of x and in the middle of the range of y respectively, and as units, the widths of the corresponding class intervals, say b and d. But the coefficient of correlation between two variables depends neither upon the origin, nor upon the scale. Hence

$$r_{xy} = r_{Vv} = \frac{\frac{1}{n}\sum_{i,j}(u_i - \bar{u})(v_j - \bar{v})f_{ij}}{\left\{\frac{1}{n}\sum_{i}(u_i - \bar{u})^2 f_{i0}\right\}^{\frac{1}{2}}\left\{\frac{1}{n}\sum_{i}(v_j - \bar{v})^2 f_{0j}\right\}^{\frac{1}{2}}}$$

$$= \frac{\sum_{i,j} u_i v_j f_{ij} - n\bar{u}.\bar{v}}{\left\{\sum_{i} u_i^2 f_{i0} - n\bar{u}^2\right\}^{\frac{1}{2}}\left\{\sum_{j} v_j^2 f_{0j} - n\bar{v}^2\right\}^{\frac{1}{2}}}$$

$$= \frac{n\sum_{i,j} u_i v_j f_{ij} - \left(\sum_{i} u_i f_{i0}\right)\left(\sum_{j} v_j f_{0j}\right)}{\left\{n\sum_{i} u_i^2 f_{i0} - \left(\sum u_i f_{i0}\right)^2\right\}^{\frac{1}{2}}\left\{n\sum_{j} v_j^2 f_{0j} - \left(\sum_{i} u_j f_{0j}\right)^2\right\}^{\frac{1}{2}}}$$

since $\bar{u} - \frac{1}{n}\sum_{i} u_i f_{i0}$ and $\bar{v} = \frac{1}{n}\sum_{j} v_i f_{0j}$.

The calculation of $\sum_{i,j} u_i v_j f_{ij}$ is done in two stages. First, we may calculate for different fixed values of j, $\sum_{i} u_i f_{ij} = u_j$ and then we obtain the sum $\sum_{j} v_j u_j$ which gives $\sum_{j} u_j \sum_{i} u_i f_{ij} = \sum_{i,j} u_i v_j f_{ij}$.

Alternatively, we may calculate, for various fixed values of i, $\sum_{j} v_j f_{ij} = V_i$ and at the next stage $\sum_{i} u_i V_i$ which equals $\sum_{i,j} u_i v_j f_{ij}$.

It is to be noted that the identity between $\sum_{i} u_i V_i$ and $\sum_{j} v_j u_j$ serves as a useful check on the calculations.

There are other checks also :

$$\sum_i u_i = \sum_{i,j} v_j f_{ij} = \sum_j v_j \sum_i f_{ij} = \sum_j v_j f_{0j}$$

and $\sum_j u_j = \sum_i u_i f_{i0}$

Now, $\sum_{i,j} u_i v_j f_{ij} = \sum_i u_i V_i = \sum_j v_j u_j$

Hence we finally, get

$$r_{xy} = r_{uv} = \frac{n\sum_i u_i v_i - \left(\sum_j u_j\right)\left(\sum_i v_i\right)}{\left\{n\sum_i u_i^2 f_{i0} - \left(\sum u_j\right)^2\right\}^{\frac{1}{2}}\left\{n\sum_j v_j^2 f_{0j} - \left(\sum_i v_i\right)^2\right\}^{\frac{1}{2}}}.$$

Naturally, this formula is used for calculating coefficient of correlation when grouped frequency distribution is given.

Example : The following table (bivariate table) shows the average daily wages (x) and the daily expenditures (y) of a group of workers in rural Bengal. Compute the coefficient of correlation between these two variables.

Expenditures (Rs.) / Wages (Rs.)	10-15	15-20	20-25	25-30	Total
16-18	3	1	–	–	4
18-20	1	2	3	–	6
20-22	2	6	4	3	15
22-24	–	7	4	6	17
24-26	2	10	12	8	32
26-28	–	6	8	6	20
28-30	–	2	3	1	6
Total	8	34	34	24	100

Solution : Let us denote the mid-values of the x-class intervals by x and the mid-values of the y-class intervals by y. Further, we assume that

$$u_i = \frac{x_i - 17{\cdot}5}{5} \text{ and } v_j = \frac{y_i - 23}{2}.$$

Y \ X			10-15	15-20	20-25	25-30				
			12.5	17.5	22.55	27.5	f_{0j}	$f_{0j}u_j$	$f_{0j}u_i^2$	$\sum fuv$
y		v \ u	-1	0	1	2				
16-18	17	–3	3(9)	1(0)	–(0)	–(0)	4	–12	36	9
18-20	19	–2	1(2)	2(0)	3(–6)	–(0)	6	–12	24	–4
20-22	21	–1	2(2)	6(0)	4(–4)	3(–6)	15	–15	15	–8
22-24	23	0	–(0)	7(0)	4(0)	6(0)	17	0	0	0
24-26	25	1	2(-2)	10(0)	12(12)	8(16)	32	32	32	26
26-28	27	2	–(0)	6(0)	8(16)	6(24)	20	40	80	40
28-30	29	3	–(0)	2(0)	3(9)	1(6)	6	18	54	15
		f_{i0}	8	34	34	24	N=100	$\sum f_{uv}$=51	$\sum fu^2$=241	$\sum f_{uv}$=78
		$v_i f_{i0}$	–8	0	34	48	$\sum fv$=74	CHECKED		
		$v_i^2 f_{i0}$	8	0	34	96	$\sum fv^2$=138			
		$\sum f_{uv}$	11	0	27	40	$\sum f_{uv}$=78			

The items in f_{i0} indicate column total of the cell frequencies. Similarly, the items in f_{0j} indicate row total of the cell frequencies. Naturally, $\sum$ row totals = $\sum$ column totals = N = total cell frequencies = 100. The items in f_{i0} row have been multiplied by v_i and the sum has been found to be 74. Similarly, items in f_{0j} columns have been multiplied by u_j and the sum has been found to be 51. Items in row $v_i^2 f_{i0}$ and those in column $f_{0j}u_i^2$ have been filled in the usual procedure. Next, we have multiplied each cell frequency with corresponding values of u and v and we have listed the products in brackets in the same cell. In this way the last row and last column have been filled in. Naturally, the last row total = the last column total and this is an important check.

We have thus found that

$\sum fuv = 78$, $\sum fu = 51$, $\sum fu^2 = 241$, $\sum fv = 74$, $\sum fv^2 = 138$

Thus $$r = \frac{100 \times 78 - 51 \times 74}{\sqrt{\{100 \times 241 - 51^2\}\{100 \times 138 - 74^2\}}}$$

$$= \frac{7800 - 3774}{\sqrt{(24100 - 2601)} \times \sqrt{(13800 - 5476)}}$$

$$= \frac{4026}{\sqrt{21499} \times \sqrt{8324}}$$

$$= \frac{4026}{146.63 \times 91.23}$$

= 0.301 (approximately).

The result shows that there exists a positive correlation between wage and expenditure but the relation is relatively weak.

2.4 Regression Analysis

Regression is the average functional relationship between two or more variables. Regression analysis measures the average relationship between two or more variables.

If there exists a relationship between two variables, then the change in one of them (the independent variable, say, x) will imply change in the other variable (the dependent variable, say, y). If the pairs of observations on x and y, namely (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) are plotted on a graph paper, a best fitting curve can be drawn through the plotted points. If the relationship is approximately linear and straight line of best fit is drawn, then it is called the regression line of y on x. Similarly, another regression line can be drawn using y as the independent variable and x the dependent variable and this is called the regression line of x on y.

2.4.1 Regression Lines

As we consider two variables x and y having n points (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) the regression line of y on x is $Y - \bar{y} = b_{yx}(x - \bar{x})$ and the regression line of x on y is $X - \bar{x} = b_{xy}(y - \bar{y})$ where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, b_{yx} = \frac{Cov(x,y)}{V(x)} = r_{xy}\frac{s_y}{s_x}$ = regression coefficient of y on x, $b_{xy} = \frac{Cov(x,y)}{V(y)} = r_{xy}\frac{s_x}{s_y}$ = regression coefficient of x on y, r_{xy} = correlation coefficient between x and y, $v(x) = s_x^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$, $v(y) = s_y^2 = \frac{1}{n}\sum_i (y_i - \bar{y})^2$.

Derivation of the regression lines

To get the regression line of y on x the best fitted line through n given points we are to find a and b in such a manner that the error sum of squares, that is,

$\sum_i (y_i - Y_i)^2 = \sum_i (y_i - a - bx_i)^2$ gets minimised or least. This is the principle of least squares. Thus we are to minimise $\sum_{i=1} (y_i - a - bx_i)^2$.

This can be obtained by differentiating the error sum of squares $\sum_i (y_i - a - bx_i)^2$ w. r. t. a and b and equating them to zero. The normal equations will then be $\sum_i (y_i - a - bx_i) = 0$ and $\sum_i x_i (y_i - a - bx_i) = 0$.

That is, $\sum y_i = na + b\sum x_i$ and $\sum_i x_i y_i = a\sum x_i + b\sum x_i^2$.

From these two normal equations a and b will be obtained in the following way.

Let the actual relationship between y and x be given by y = a + bx + e ... (1)

where y is the dependent variable, x is the independent variable and e is the error term which takes the influence of the omitted variables into account.

However, we do not get the actual relation. What we get is the estimated relationship given by

Y = a + bx ... (2)

where Y is the estimated value of y for a given value of x.

From (1) and (2) we see that y = Y + e

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \text{ and } \hat{b} = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - (\sum x_i)^2} = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\,\bar{y}}{\frac{1}{n}\sum x_i^2 - \bar{x}^2}.$$

That is, $\hat{b} = \frac{Cov(x,y)}{Var(x)} = b_{yx}$ = the regression coefficient of y on x = the slope of the regression line of y on x and it can be simplified as $b_{yx} = r\frac{s_y}{s_x}$, s_x and s_y being the standard deviations of x and y respectively. So the estimated relation Y = a + bx can be simplified to the following equation $Y = \bar{y} + r\frac{s_y}{s_x}(x - \bar{x})$ or, $Y - \bar{y} = r\frac{s_y}{s_x}(x - \bar{x})$ and is called the regression line of y on x. Here Y gives the predicted or estimated value of y for given values of x.

In a similar way considering y as the independent variable and x as the dependent variable we get the regression line of x on y as

$X = \bar{x} + r\frac{s_x}{s_y}(y - \bar{y})$ where $r\frac{s_x}{s_y} = b_{xy}$ = regression coefficient of x on y = the slope of the regression line of x on y.

2.4.2 Properties of regression equations and regression coefficients :

1. Regression coefficients depend only on scale but not on origin.

Proof : Consider $u_i = \frac{x_i - a}{b}, v_i = \frac{y_i - c}{d}$ for i = 1, 2, ..., n where a and c are origins and b and d are scales. Then $x_i = a + bu_i$ for i = 1, 2, ... n and we get $\bar{x} = a + b\bar{u}$, and $y_i = c + dv_i$ for i = 1, 2, ..., n and we get $\bar{y} = c + d\bar{v}$,

Hence, $Cov(x,y) = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{bd}{n}\sum_i (u_i - \bar{u})(v_i - \bar{v}) = bd\,Cov(u,v)$.

Again, $V(x) = \frac{1}{n}\sum_i (x_i - \bar{x})^2 = \frac{b^2}{n}\sum_i (u_i - \bar{u})^2 = b^2\,Var(u)$

and $V(y) = \frac{1}{n}\sum_i (y_i - \bar{y})^2 = \frac{d^2}{n}\sum_i (v_i - \bar{v})^2 = d^2\,Var(v)$.

Then $b_{yx} = \frac{Cov(x,y)}{V(x)} = \frac{bdCov(u,v)}{b^2V(u)} = \frac{d}{b}b_{vu}$

and $b_{xy} = \frac{Cov(x,y)}{V(y)} = \frac{bdCov(u,v)}{d^2V(v)} = \frac{b}{d}b_{uv}$.

This proves the result.

2. b_{xy}, b_{yx} and r_{xy} are of same sign.

Proof : $r_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}}, b_{xy} = \frac{Cov(x,y)}{Var(y)}$ and $b_{yx} = \frac{Cov(x,y)}{Var(x)}$.

As Var (x) and Var (y) are always positive but Cov (x, y) may be positive or negative. So sign depends on the sign of Cov (x,y).

Hence r_{xy}, b_{xy} and b_{yx} are of same sign.

Both the regression coefficients will always have the same sign. That is, if b_{yx} is positive (negative), b_{xy} will also be positive (negative).

3. Product of two regression coefficients is square of the correlation coefficient.

Proof : $b_{yx} \cdot b_{xy} = \frac{Cov(x,y)}{s_x^2} \cdot \frac{Cov(x,y)}{s_y^2} = \left(\frac{Cov(x,y)}{s_x s_y}\right)^2 = r_{xy}^2$

Thus the coefficient of correlation is the geometric mean of the two regression coefficients, that is,

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

However, the sign of r_{xy} is determined by the common sign of both the regression coefficients.

4. The arithmetic mean of the absolute values of the two regression coefficients cannot be less than the absolute value of the coefficient of correlation. That means,

$$\frac{|b_{yx}| + |b_{xy}|}{2} \geq |r_{xy}|.$$

Proof : We may write $(s_y - s_x)^2 \geq 0$ as it is a perfect square.

That is, $s_y^2 + s_x^2 - 2s_y s_x \geq 0$

That is, $\frac{s_y}{s_x} + \frac{s_x}{s_y} \geq 2$

That is, $|r_{xy}|\frac{s_y}{s_x} + |r_{xy}|\frac{s_x}{s_y} \geq 2|r_{xy}|$

That is, $|b_{yx}| + |b_{xy}| \geq 2|r_{xy}|$

This is, $\frac{|b_{yx}| + |b_{xy}|}{2} \geq |r_{xy}|$

Hence proved.

5. The absolute value of the coefficient of correlation is the ratio between the standard deviation of the estimated values of y, that is, Y and that of the actual values of y.

Proof : $Var(Y) = \frac{1}{n}\sum\left\{\left(Y_i - \overline{Y}\right)^2\right\}$

$= \frac{1}{n}\sum\left\{\left(Y_i - \overline{y}\right)\right\}^2 \qquad \because \overline{Y} = \overline{y}$

$$= \frac{1}{n}\sum\left[r\frac{s_y}{s_x}(x_i - \bar{x})\right]^2$$

$$= r^2 \frac{s_y^2}{s_x^2} s_x^2$$

$$= r^2 s_y^2$$

That is, $r^2 = \frac{s_Y^2}{s_y^2}$.

That is, $|r| = \frac{s_Y}{s_y}$ since standard deviations, s_y and s_Y by definition, cannot be negative.

Hence proved.

Again, $r^2 = \frac{s^2Y}{s^2y} = \frac{\sum(Y_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$.

$$= \frac{\text{Enplained variation}}{\text{Total variation}}.$$

The quantity, r^2, is called the coefficient of determination and may itself be taken as a measure of the usefulness of the linear regression equation as a predicting formula.

6. The mean of the observed values of y is equal to the mean of the corresponding estimated or predicted values, that is, $\bar{y} = \bar{Y}$.

Proof : The regression line of y on x is given by

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

That is, $Y_i = \bar{y} + b_{yx}(x_i - \bar{x})$

That is, $\sum_{i=1}^{n} Y_i = n\bar{y} + b_{yx}\sum(x_i - \bar{x})$

$$= n\bar{y} + b_{yx.0}$$

$$\because \sum_i (x_i - \bar{x}) = 0$$

That is,

$$= n\bar{y}$$

$$\therefore \bar{Y} = \bar{y}.$$

Naturally, the mean of the error term, that is, $\bar{e}$ is zero.

7. The two regression lines will pass through the point $(\bar{x}, \bar{y})$.

Proof : The regression equation of y on x is given by

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

Again, the regression equation of x on y is given by $X_i - \bar{x} = b_{xy}(y_i - \bar{y})$.

Thus $y - \bar{y} = \frac{1}{b_{xy}}(x - \bar{x})$

Hence, $b_{yx}(x - \bar{x}) = \frac{1}{b_{xy}}(x - \bar{x})$

That is, $(x - \bar{x})\left[b_{yx} - \frac{1}{b_{xy}}\right] = 0$.

That is, $x = \bar{x}$ so long as the two regression lines are distinct, that is, $b_{yx} \neq \frac{1}{b_{xy}}$.

Similarly, $y = \bar{y}$.

Hence proved.

8. The two regression lines will coincide if there is perfect correlation between the two variables.

Proof : The regression equation of y on x is given by

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

The regression equation of x on y is given by

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

Hence, $y_i - \bar{y} = \frac{1}{b_{xy}}(x_i - \bar{x})$

Thus, the two regression lines are distinct so long as $b_{yx} \neq \frac{1}{b_{xy}}$.

But $b_{yx} = r\frac{s_y}{s_x}$ and $\frac{1}{b_{xy}} = \frac{s_y}{rs_x}$.

We see that if $r = \pm 1$, $b_{yx} = \frac{1}{b_{xy}}$.

Hence when there is perfect correlation between the two variables, the two regression lines will coincide. In such a case all points in the scatter diagram will lie exactly on a straight line.

9. For a set of n pairs of observations (x_1, y_1), (x_2, y_2), (x_3, y_3), $\ldots(x_n, y_n)$ prove that

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - Y_i)^2 + \sum_i (Y_i - \bar{y})^2$$

where Y_i represents the values of y for given values of $x = x_1, x_2, x_3, \ldots x_n$ obtained from the regression equation of y on x.

Proof : We may write

$$y_i - \bar{y} = (y_i - Y_i) + (Y_i - \bar{y})$$

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - Y_i)^2 + \sum_i (Y_i - \bar{y})^2 + 2\sum_i (y_i - Y_i)(Y_i - \bar{y})$$

$$= \sum (y_i - Y_i)^2 + \sum (Y_i - \bar{y})^2 + 0$$

Now, $\sum (y_i - Y_i)(Y_i - \bar{y})$

$$= \sum \left[(y_i - \bar{y}) - b_{yx}(x_i - \bar{x})\right]\left[b_{yx}(x_i - \bar{x})\right]$$

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

$$= b_{yx}\left[n \operatorname{cov}(x, y) - b_{yx} n \operatorname{var}(x)\right]$$

$$= b_{yx}\left[n \operatorname{cov}(x, y) - \frac{\operatorname{cov}(x, y)}{s^2 x} . n s_x^2\right]$$

$$= b_{yx}.0 = 0.$$

Thus,

$$\sum (y_i - \bar{y})^2 = \sum_i (y_i - Y_i)^2 + \sum_i (Y_i - \bar{y})^2$$

$$= \sum (Y_i - \bar{y})^2 + \sum (y_i - Y_i)^2$$

That is, total variation = Explained variation + Unexplained variation.

In symbols,

Total sum of squares of y = sum of squares explained by the linear regression equation of y on x + sum of squares unexplained by the linear regression equation of y on x.

That is, total SS of y = SS explained by the linear regression equation of y on x + SS unexplained by the linear regression equation of y on x.

Again, unexplained variation in y = total variation in y – explained variation in y

$$= s^2y - s^2Y$$

$$= s^2y - r^2s^2y \qquad \because r^2 = \frac{s^2Y}{s^2y}$$

$$= \left(1 - r^2\right)s^2y.$$

So for knowing the unexplained variation in y the coefficient of correlation play a vital role.

ANGLES BETWEEN TWO REGRESSION LINES

We know that the regression equation of y on x is given by

$$Y_i - \bar{y} = r\frac{s_y}{s_x}\left(x_i - \bar{x}\right)$$

That is, $Y_i = r\frac{s_y}{s_x}\left(x_i - \bar{x}\right) + \bar{y}$ ... (1)

Also, the regression equation of x on y is given by

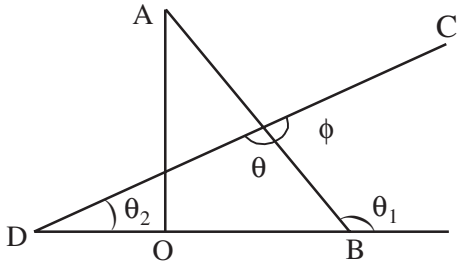
$$X_i - \bar{x} = r\frac{s_x}{s_y}\left(y_i - \bar{y}\right).$$

That is, $y_i - \bar{y} = \left(X_i - \bar{x}\right)\frac{s_y}{rs_x}$

That is, $y_i = \bar{y} + \frac{1}{r}\frac{s_y}{s_x}.X_i - \bar{x}.\frac{1}{r}\frac{s_y}{s_x}$ ... (2)

Let us represent the regression equations in the following diagram where the line AB represents the regression equation of x on y and the line CD represents the regression equation of y on x.

The lines AB and CD have made angles θ_1 and θ_2 with the positive directtion of the X axis. The internal and external angles between the lines AB and CD have respectively been θ and ϕ.



From (1) we see that $\tan\theta_2 = r\frac{s_y}{s_x}$

and from (2)

we see that $\tan\theta_1 = \frac{1}{r}\frac{s_y}{s_x}$.

Now, $\theta = \theta_1 - \theta_2$ so that $\tan\theta = \tan\left(\theta_1 - \theta_2\right) = \frac{\tan\theta_1 - \tan\theta_2}{1 + \tan\theta_1\tan\theta_2}$

$$= \frac{\frac{1}{r}\frac{s_y}{s_x} - r\frac{s_y}{s_x}}{1 + \frac{1}{r}\frac{s_y}{s_x}.r\frac{s_y}{s_x}} = \frac{\frac{s_y - r^2 s_y}{rs_x}}{\frac{s_x^2 + s_y^2}{s_x^2}} = \frac{\frac{s_y(1-r^2)}{rs_x}}{\frac{s_x^2 + s_y^2}{s_x^2}}$$

$$= \frac{1-r^2}{r}.\frac{s_y}{s_x}.\frac{s_x^2}{s_x^2 + s_y^2} = \frac{1-r^2}{r}\frac{s_x s_y}{s_x^2 + s_y^2}$$

That is, $\tan\theta = \frac{1-r^2}{r}\frac{s_x s_y}{s_x^2 + s_y^2}$

Hence $\theta = \tan^{-1}\frac{1-r^2}{r}\frac{s_x s_y}{s_x^2 + s_y^2}$ ... (A)

Again, $\phi = \Pi - \theta$ so that $\tan\phi = \tan(\Pi - \theta) = -\tan\theta$.

That is, $\phi = -\theta = \tan^{-1} -\frac{1-r^2}{r}\frac{s_x s_y}{s_x^2 + s_y^2}$ ... (B)

Combining (A) and (B) we may write that the angles between the two regression lines will be

$$\tan^{-1}\left\{\pm\frac{1-r^2}{r}\frac{s_x s_y}{s_x^2 + s_y^2}\right\}.$$

Let us now make two observations on this angle :

Observation 1 : Let r be ± 1 so that tan θ = 0 = tan 0°. That means, θ = 0°. This implies that when r ± 1, that is, when the two variables are perfectly correlated, the angle between the two regression lines will be 0°, implying also that the two regression lines will coincide.

Observation 2 : Let r be 0 so that tan θ = ∝ = tan 90°. That is, θ = 90°. In this case the angle between the two regression lines is 90°, implying that the two regression lines are perpendicular to each other. In this case Y = $\bar{y}$ and X = $\bar{x}$ will be the equations of the two regression lines.

Combining results of these two observations we see that the greater the angle between the two regression lines, the poorer is the relation between the two variables and smaller the angle between the two regression lines, the stronger is the relation between the two variables.

2.5 Rank Correlation

The product moment correlation coefficient requires measurements in two characters for a group of individuals. Sometimes the measurements on the two characters cannot be available because they are not measurable or even if they are measurable at all, they can not be measured due to lack of cost, labour and time. Sometimes an alternative simple procedure is required for quick appreciation.

Let a group of individuals be arranged in order of efficiency according to their possession of the characters when the characters are not directly measured. Such an ordered arrangement for each of the characters is called ranking and the ordinal numbers of an individual in the ordered arrangements for the characters are called ranks of the individuals in the characters. Specifically, a rank K indicates that (K – 1) individuals have that character in a higher degree than the individual getting rank K viz, best student in a competition gets rank 1, next best gets rank 2 and so on.

If ranks of individuals are available for each of two characters, the association between these ranks of individuals for the two characters can be measured by rank correlation and the measure is called rank correlation coefficient. In other words, rank correlation coefficient is the product moment correlation coefficient between the ranks of individuals in two characters. If two or more individuals are allotted the same rank then it is called a case of tie.

2.5.1 Spearman's rank correlation coefficient

a) First consider the case where there are no ties. Suppose n individuals are ranked according to two characters A and B. Let x_i and y_i be the ranks of the i-th individual for the two characters, i = 1, 2, ..., n. Let $d_i = x_i - y_i$ for i = 1, 2, ..., n. Then

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i = 1+2+\ldots+n = \frac{n(n+1)}{2} \text{ and}$$

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i^2 = 1^2+2^2+\ldots+n^2 = \frac{n(n+1)(2n+1)}{6}.$$

Therefore, $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2$

$$= \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 \left[\text{since } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{n+1}{2}\right]$$

$$= \frac{(n+1)}{12}[2(2n+1)-3(n+1)] = \frac{n^2-1}{12}.$$

Similarly, $\bar{y} = \frac{n+1}{2}$ and $s_y^2 = \frac{n^2-1}{12}$.

Now, $\frac{1}{n}\sum_{i=1}^{n} d_i^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 = \frac{1}{n}\sum_i\left[(x_i - \bar{x}) - (y_i - \bar{y})\right]^2 = s_x^2 + s_y^2 - 2\,\text{Cov}(x,y)$

i.e., covariance between ranks of two characters is

$$\text{Cov}(x,y) = \frac{1}{2}\left[s_x^2 + s_y^2 - \tfrac{1}{n}\sum_{i=1}^{n} d_i^2\right]$$

$$= \frac{1}{2}\left[\frac{n^2-1}{12} + \frac{n^2-1}{12} - \tfrac{1}{n}\sum_{i=1}^{n} d_i^2\right] = \frac{1}{2}\left[\frac{n^2-1}{6} - \tfrac{1}{2}\sum_{i=1}^{n} d_i^2\right]$$

$$= \frac{n^2-1}{12}\left[1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}\right].$$

The correlation coefficient between ranks of two characters is

$$r_R = \frac{\text{Cov}(x,y)}{s_x s_y} = \frac{\frac{n^2-1}{12}\left[1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i^2\right]}{\sqrt{\frac{n^2-1}{12} \times \frac{n^2-1}{12}}} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i^2 .$$

Here r_R is called Spearman's rank correlation coefficient.

This rank correlation coefficient lies between –1 and +1.

As $\sum_{i=1}^{n} d_i^2 \geq 0$, $n > 1$, $\frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i^2 \geq 0$. So $r_R \leq 1$.

Again, $\sum_{i=1}^{n} d_i^2$ will be maximum when $y_i = (n + 1) - x_i$ for i = 1, 2, ..., n i.e., if one character has ranks 1, 2, 3, ..., n then the other character would have ranks n, n – 1, ..., 2, 1.

So, $\sum_{i=1}^{n} d_i^2 \leq \sum_{i=1}^{n}(2x_i - (n+1))^2 = 4\sum_{i=1}^{n}\left(x_i - \frac{n+1}{2}\right)^2 = \frac{4n}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

i.e., $\sum_{i=1}^{n} d_i^2 \leq 4ns_x^2 = 4n\frac{(n^2-1)}{12} = \frac{n(n^2-1)}{3}$

Then $r_R = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i^2 \geq 1 - \frac{6}{n(n^2-1)} \cdot \frac{n(n^2-1)}{3} = 1 - 2 = -1.$

i.e., $r_R \geq -1$.

Two extreme cases are : $r_R = 1$ when $x_i = y_i$ and $r_R = -1$ when $y_i = (n + 1) - x_i$ for i = 1, 2, ..., n. In the first case the ranks of n individuals concord (or are in perfect agreement) in two characters i.e., relation between ranks of individuals in two characters is positive and perfectly linear. In the second case the ranks of n individuals discord (or are in perfect disagreement) in two characters i.e., relation between ranks of individuals in two characters is negative and perfectly linear.

(b) **Case of ties**

Suppose the same rank is allotted to k individuals and these k individual follow r other individuals in the ranking. If there is no tie in these (k + r) ranks these k individuals would have had ranks (r + 1), (r + 2), ..., (r + k). But in case of ties each of these k individuals would be alloted the rank

$$\frac{(r+1)+(r+2)+....+(r+k)}{k} = r + \frac{k+1}{2}$$

The tie does not affect the mean of ranks, i.e., $\bar{x} = \bar{y} = \frac{n+1}{2}$.

But $\sum x^2$ (in tie case) $-\sum x^2$ (in no tie case)

$$= \left[1^2 + 2^2 + \ldots + r^2 + k\left(r + \frac{k+1}{2}\right)^2 + (r+k+1)^2 + \ldots + n^2\right]$$

$$-\left[1^2 + 2^2 + \ldots + r^2 + (r+1)^2 + (r+2)^2 + \ldots + (r+k)^2 + (r+k+1)^2 + \ldots + n^2\right]$$

$$= k\left(r + \frac{k+1}{2}\right)^2 - \left[(r+1)^2 + (r+2)^2 + \ldots + (r+k)^2\right]$$

$$= k\left(r^2 + r(k+1) + \frac{(k+1)^2}{4}\right) - \left[kr^2 + 2r(1+2+\ldots+k) + (1^2 + 2^2 + \ldots + k^2)\right]$$

$$= kr^2 + rk(k+1) + \frac{k(k+1)^2}{4} - kr^2 - rk(k+1) - \frac{k(k+1)(2k+1)}{6}$$

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

So $\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right)$ (in tie case) $= \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right)$ (in no tie case) $- \frac{k^3 - k}{12n}$

i.e., s_x^2 (in tie case) $= \frac{n^2 - 1}{12} - \frac{k^3 - k}{12n}$.

Now suppose that in the ranking with respect to first character there are s ties of lengths $k_1, k_2, ..., k_s$ and in the ranking with respect to the second character there are t ties of lengths $l_1, l_2, ..., l_t$. The variances of the ranks would be

$$s_x^2 = \frac{n^2-1}{12} - T_x \text{ where } T_x = \frac{1}{12n}\sum_{i=1}^{s}\left(k_i^3 - k_i\right)$$

and $s_y^2 = \frac{n^2-1}{12} - T_y$ where $T_y = \frac{1}{12n}\sum_{i=1}^{t}\left(l_i^3 - l_i\right)$.

Similarly, since $2\,\mathrm{Cov}(x,y) = s_x^2 + s_y^2 - \frac{1}{n}\sum_{i=1}^{n} d_i^2$,

in the tie case the covariance is

$$\mathrm{Cov}(x,y) = \frac{n^2-1}{12} - \frac{1}{2}(T_x + T_y) - \frac{1}{2n}\sum_i d_i^2,$$

so that Spearman's rank correlation coefficient in the tie case is

$$r_R = \frac{\frac{n^2-1}{12} - \frac{T_x+T_y}{2} - \frac{1}{2n}\sum_i d_i^2}{\sqrt{\left(\frac{n^2-1}{12} - T_x\right)\left(\frac{n^2-1}{12} - T_y\right)}}.$$

In case of perfect agreement (or cordance) between two series of ranks $T_x = T_y, \sum_i d_i^2 = 0.$

Hence $$r_R = \frac{\frac{n^2-1}{12} - T_x}{\frac{n^2-1}{12} - T_x} = 1.$$

Also in case of perfect disagreement (or discordance) between two series of ranks,

$y_i = (n+1) - x_i$, $d_i = x_i - y_i = 2x_i - (n+1)$ for i = 1, 2, ..., n,

$$s_x^2 = s_y^2 = \frac{n^2-1}{12} - T_x, T_x = T_y$$

and $$\frac{1}{n}\sum d_i^2 = \frac{1}{n}\sum_i (2x_i - (n+1))^2 = 4s_x^2 = \frac{n^2-1}{3} - 4T_x \text{ as } \bar{x} = \frac{n+1}{2}.$$

so that $$r_R = \frac{\frac{n^2-1}{12} - T_x - \left(\frac{n^2-1}{6} - 2T_x\right)}{\frac{n^2-1}{12} - T_x} = \frac{-\left(\frac{n^2-1}{12} - T_x\right)}{\frac{n^2-1}{12} - T_x} = -1$$

Thus, $-1 \le r_R \le 1$ in tie case of ranks.

2.6 Multiple Regression

Multiple regression is obtained when the effect of two or more variables (called independent variables) are considered on another variable (called dependent variable). Multiple regression estimates the dependent variable for given values of independent variables. For example, the yield of crop depends on rainfall, soil and temperature. The regression of yield of crop on rainfall, soil and temperature is considered to estimate the yield of crop.

For the sake of simplicity, consider the linear effect of two independent variables x_2 and x_3 on dependent variable x_1. Let this relation be $x_1 = a + bx_2 + cx_3$ where a, b and c are constants whose values can be obtained by using least squares principle i.e., by minimising $\sum_{i=1}^{n}(x_{1i} - a - bx_{2i} - cx_{3i})^2$ with respect to a, b and c for n given points (x_{1i}, x_{2i}, x_{3i}), i = 1, 2, ..., n. So a, b and c can be obtained from normal equations, obtained by differentiating $\sum_{i=1}^{n}(x_{1i} - a - bx_{2i} - cx_{3i})^2$ with respect to a, b and c and then equating each to zero. The normal equations are :

$$\sum_{i=1}^{n}(x_{1i} - a - bx_{2i} - cx_{3i}) = 0, \sum_{i=1}^{n} x_{2i}(x_{1i} - a - bx_{2i} - cx_{3i}) = 0 \text{ and}$$

$$\sum_{i=1}^{n} x_{3i}(x_{1i} - a - bx_{2i} - cx_{3i}) = 0$$

i.e., $$a = \bar{x}_1 - b\bar{x}_2 - c\bar{x}_3 \qquad (1)$$

and $$\sum_{i}(x_{2i} - \bar{x}_2)\{(x_{1i} - \bar{x}_1) - b(x_{2i} - \bar{x}_2) - c(x_{3i} - \bar{x}_3)\} = 0. \qquad (2)$$

$$\sum_{i}(x_{3i} - \bar{x}_3)\{(x_{1i} - \bar{x}_1) - b(x_{2i} - \bar{x}_2) - c(x_{3i} - \bar{x}_3)\} = 0. \qquad (3)$$

where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ji}, j = 1, 2, 3.$

Dividing (2) and (3) by n, the normal equations to solve for the values of b and c would be

$$r_{12}s_1s_2 = bs_2^2 + cr_{23}s_2s_3 \quad \ldots\ldots\ldots \quad (4)$$

$$r_{13}s_1s_3 = br_{23}s_2s_3 + cs_3^2 \quad \ldots\ldots\ldots \quad (5)$$

where $s_j^2 = \sum_{i=1}^{n}(x_{ji} - \bar{x}_j)^2 / n$ for j = 1, 2, 3, r_{ij} = the correlation coefficient between x_i and x_j for i, j = 1, 2, 3.

Using (1) the equation for the regression line can be written as

$$(x_1 - \bar{x}_1) = b(x_2 - \bar{x}_2) + c(x_3 - \bar{x}_3). \quad \ldots\ldots\ldots \quad (6)$$

Then eliminating b and c from (6), (4) and (5) the multiple regression equation can be written as

$$\begin{vmatrix} x_1 - \bar{x}_1 & x_2 - \bar{x}_2 & x_3 - \bar{x}_3 \\ r_{12}s_1s_2 & s_2^2 & r_{23}s_2s_3 \\ r_{13}s_1s_3 & r_{23}s_2s_3 & s_3^2 \end{vmatrix} = 0 \text{ or } \begin{vmatrix} \frac{x_1 - \bar{x}_1}{s_1} & \frac{x_2 - \bar{x}_2}{s_2} & \frac{x_3 - \bar{x}_3}{s_3} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix} = 0$$

i.e., $$\left(1 - r_{23}^2\right)\frac{x_1 - \bar{x}_1}{s_1} - (r_{12} - r_{13}r_{23})\frac{x_2 - \bar{x}_2}{s_2} - (r_{13} - r_{12}r_{23})\frac{x_3 - \bar{x}_3}{s_3} = 0$$

or $$x_1 - \bar{x}_1 = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \cdot \frac{s_1}{s_2}(x_2 - \bar{x}_2) + \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \cdot \frac{s_1}{s_3}(x_3 - \bar{x}_3)$$

Thus the multiple regression equation of x_1 on x_2 and x_3 can be written as

$$x_1 = \bar{x}_1 + b_{12.3}(x_2 - \bar{x}_2) + b_{13.2}(x_3 - \bar{x}_3) \quad \ldots\ldots\ldots \quad (7)$$

where $b_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \cdot \frac{s_1}{s_2}$ = partial regression coefficient of x_1 on x_2 eliminating the effect of x_3 from both, $b_{13.2} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \cdot \frac{s_1}{s_3}$ = partial regression coefficient of x_1 on x_3 eliminating the effect of x_2 from both. Comparing (7) with the multiple regression equation we get $\hat{b} = b_{12.3}$, $\hat{c} = b_{13.2}$, $\hat{a} = \bar{x}_1 - \hat{b}\bar{x}_2 - \hat{c}\bar{x}_3$

x_1 in the L.H.S. of (7) is the fitted value of x_1 obtained by least squares principle. In order to differentiate it from the observed x_1 we will denote it by $x_{1.23}$.

2.6.1 Multiple correlation

This is the association between dependent variable (x_1) and the linear regressand of it on other independent variables ($x_{1.23}$) and it is measured by linear correlation between x_1 and $x_{1.23}$. The multiple correlation coefficient of x_1 on x_2 and x_3 may be defined to be the ordinary correlation coefficient between x_1 as observed and x_1 as estimated on the basis of its linear regression on x_2 and x_3. Thus the simple correlation

between x_1 and $x_{1.23}$, that is, between x_1 as such and x_1 as predicted by x_2 and x_3 is called the multiple correlation efficient of x_1 on x_2 and x_3. This is denoted by the symbol, $r_{1.23}$. That is,

$$r_{1.23} = \frac{\text{Cov}(x_1, x_{1.23})}{\sqrt{\text{Var}(x_1)}\sqrt{\text{Var}(x_{1.23})}}.$$

The residual part in x_1 that is not explained by the regression line, is represented by $e_{1.23} = x_1 - x_{1.23}$

$$x_1 = x_{1.23} + e_{1.23}. \text{ i.e.}, e_{1.23} = (x_1 - \bar{x}_1) - b_{12.3}(x_2 - \bar{x}_2) - b_{13.2}(x_3 - \bar{x}_3).$$

Again, Cov $(x_1, x_{1.23})$ = Cov$(x_{1.23} + e_{1.23}, x_{1.23})$ = Var$(x_{1.23})$ + Cov$(e_{1.23}, x_{1.23})$

As $\bar{x}_{1.23} = \frac{1}{n}\sum_i \left[\bar{x}_1 + b_{12.3}(x_{2i} - \bar{x}_2) + b_{13.2}(x_{3i} - \bar{x}_3)\right]$

$= \bar{x}_1 + b_{12.3} \cdot \frac{1}{n}\sum_i (x_{2i} - \bar{x}_2) + b_{13.2} \cdot \frac{1}{n}\sum_i (x_{3i} - \bar{x}_3) = \bar{x}_1.$

$$\text{Cov}(x_{1.23}, e_{1.23}) = \frac{1}{n}\sum_i \left[(\bar{x}_1 - \bar{x}_1) + b_{12.3}(x_{2i} - \bar{x}_2) + b_{13.2}(x_{3i} - \bar{x}_3)\right]$$

$$\times \left[x_1 - \bar{x}_1 - b_{12.3}(x_{2i} - \bar{x}_2) - b_{13.2}(x_{3i} - \bar{x}_3)\right]$$

= 0 [from normal equations (1)]

So $\text{Cov}(x_1, x_{1.23}) = \text{Var}(x_{1.23}) \text{ and } \text{var}(x_1) = s_1^2.$ (8)

$$\therefore \quad V(x_{1.23}) = \text{Cov}(x_1, x_{1.23}) = \frac{1}{n}\sum_i (x_{1i} - \bar{x}_1)\left[\bar{x}_1 + b_{12.3}(x_{2i} - \bar{x}_2) + b_{13.2}(x_{3i} - \bar{x}_3) - \bar{x}_1\right]$$

$$= \frac{1}{n}\sum_i (x_{1i} - \bar{x})(x_{2i} - \bar{x}_2). \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \cdot \frac{s_1}{s_2} + \frac{1}{n}\sum_i (x_i - \bar{x}_1)(x_{3i} - \bar{x}_3) \cdot \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \cdot \frac{s_1}{s_3}$$

$$= r_{12}s_1s_2 \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \cdot \frac{s_1}{s_2} + r_{13}s_1s_3 \cdot \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \cdot \frac{s_1}{s_3} = \left(\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}\right)s_1^2 \quad \ldots\ldots (9)$$

$$\therefore \quad \text{correlation } (x_1, x_{1.23}) = r_{1.23} = \frac{\text{Cov}(x_1, x_{1.23})}{\sqrt{\text{Var}(x_1)\text{Var}(x_{1.23})}} = \sqrt{\frac{\text{Var}(x_{1.23})}{s_1^2}}$$

$$= \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} = \left[\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}\right]^{\frac{1}{2}}, \text{ from (8) and (9).}$$

Similarly, $r_{2.13}$ is the multiple correlation coefficient of x_2 on x_1 and x_3. It represents the degree of association between x_2 and the joint influence of x_1 and x_3 on x_2. It can be computed by using the formula $r_{2.13} = \left[\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}\right]^{\frac{1}{2}}$.

Also, $r_{3.12}$ is the multiple correlation coefficient of x_3 on x_1 and x_2. It shows the degree of relationship between x_3 and the joint effect of x_1 and x_2 on x_3. It can be ascertained by using the formula $r_{3.12} = \left[\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}\right]^{\frac{1}{2}}$.

Remarks : $r_{1.23}$ is non-negative, i.e., $0 \le r_{1.23} \le 1$. Similarly, $0 \le r_{2.13} \le 1$ and $0 \le r_{3.12} \le 1$.

The multiple correlation coefficient, being essentially a simple correlation coefficient, must lie between –1 and +1. But in this case $Cov(x_1, x_{1.23})$ is itself $Var(x_{1.23})$ and since var $(x_{1.23})$ can not be negative, so $Cov(x_1, x_{1.23})$ can not also be negative. This means that the multiple correlation must lie between 0 and +1, that is, $0 \le r_{1.23} \le +1$.

If $r_{12} = r_{13} = 0$, then $r_{1.23} = 0$.

Since $r_{1.23}^2 = \frac{Var(x_{1.23})}{Var(x_1)}$ = proportion of total variance (i.e. total variation) of x_1 that is explained by the multiple regression equation. When $r_{1.23} = 1$, i.e, $Var(x_{1.23}) = Var(x_1)$, then $x_1 = x_{1.23}$ for each i = 1, 2, ..., n, indicating perfect prediction. Again, if $r_{1.23} = 0$ then $V(x_{1.23}) = 0$ which implies $x_{1.23} = \bar{x}$. In this case x_2 and x_3 do not play any role in predicting the dependent variable and this equation fails completely as a predicting formula for x_1. So multiple correlation coefficient $r_{1.23}$ may also be regarded as a measure of the efficiency of multiple regression equation as a formula for predicting x_1 when x_2 and x_3 are given. $r_{1.23}^2$ is also called the coefficient of determination.

2.6.2 Partial Correlation

This is the correlation between two variables when effects of the third variable are eliminated from both. So let us give the definition of the partial correlation coefficient.

So for a multivariate data on these variables the correlation coefficient between variables x_1 and x_2 eliminating the effect of the third variable x_3 from both, is called the partial correlation coefficient between x_1 and x_2 eliminating the effect of x_3 from both x_2 and x_3 and it is denoted by $r_{12.3}$.

So $r_{12.3}$ = correlation coefficient between $e_{1.3}$ and $e_{2.3}$ where

$x_1 = \bar{x}_1 + r_{13}\frac{s_1}{s_3}(x_3 - \bar{x}_3) + \; e_{1.3} = x_{1.3} + e_{1.3}$ and

$x_2 = \bar{x}_2 + r_{23}\frac{s_2}{s_3}(x_3 - \bar{x}_3) + e_{2.3} = x_{2.3} + e_{2.3}$.

$e_{1.3}$ and $e_{2.3}$ are called the residuals in the linear regression of x_1 on x_3 and x_2 on x_3 respectively.

$$\mathrm{Cov}(e_{1.3}, e_{2.3}) = \tfrac{1}{n}\sum_{i=1}^{n}\left[x_{1i} - \bar{x}_1 - r_{13}\frac{s_1}{s_3}(x_{3i} - \bar{x}_3)\right]\left[x_{2i} - \bar{x}_2 - r_{23}\frac{s_2}{s_3}(x_{3i} - \bar{x}_3)\right]$$

since $\bar{e}_{1.3} = \bar{e}_{2.3} = 0$

$$\text{i.e., } \mathrm{Cov}(e_{1.3}, e_{2.3}) = \tfrac{1}{n}\sum_i (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) - r_{23}\frac{s_2}{s_3}\cdot\tfrac{1}{n}\sum_i (x_{1i} - \bar{x}_1)(x_{3i} - \bar{x}_3)$$

$$-r_{13}\frac{s_1}{s_3}\cdot\tfrac{1}{n}\sum_i (x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3) + r_{13}r_{23}\frac{s_1 s_2}{s_3^2}\cdot\tfrac{1}{n}\sum_i (x_{3i} - x_3)^2$$

$$= \; r_{12}\, s_1\, s_2 - r_{23}\, r_{13}\, s_1\, s_2 - r_{13}\, r_{23}\, s_1\, s_2 + r_{13}\, r_{23}\, s_1\, s_2$$

$$= \; (r_{12} - r_{13}\, r_{23}) s_1 s_2.$$

$$\mathrm{Var}(e_{1.3}) = \frac{1}{n}\sum_i \left\{(x_{1i} - \bar{x}_1) - r_{13}\frac{s_1}{s_3}(x_{3i} - \bar{x}_3)\right\}^2$$

$$= \frac{1}{n}\sum_i \left\{(x_{1i} - \bar{x}_1)^2 + r_{13}^2\frac{s_1^2}{s_3^2}(x_{3i} - \bar{x}_3)^2 - 2r_{13}\frac{s_1}{s_3}(x_{1i} - \bar{x}_1)(x_{3i} - \bar{x}_3)\right\}$$

$$= s_1^2 + r_{13}^2\cdot\frac{s_1^2}{s_3^2}\cdot s_3^2 - 2r_{13}\cdot\frac{s_1}{s_3}\cdot r_{13}s_1 s_3$$

$$= s_1^2\left(1 - r_{13}^2\right).$$

Similarly, $\mathrm{Var}(e_{2.3}) = s_2^2\left(1 - r_{23}^2\right)$.

So $$r_{12.3} = \frac{\mathrm{Cov}(e_{1.3}, e_{2.3})}{\sqrt{\mathrm{Var}(e_{1.3})\mathrm{Var}(e_{2.3})}} = \frac{(r_{12} - r_{13}r_{23})s_1 s_2}{\sqrt{s_1^2\left(1 - r_{13}^2\right)s_2^2\left(1 - r_{23}^2\right)}} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}}$$

There are two other partial correlation coefficients, namely

$$r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{32}^2\right)}} \text{ and } r_{23.1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{\left(1 - r_{21}^2\right)\left(1 - r_{31}^2\right)}}.$$

Note : It can be shown that $r_{12.3}$ lies between ± 1. i.e., $-1 \le r_{12.3} \le 1$ and

$1 - r_{1.23}^2 = \left(1 - r_{13}^2\right)\left(1 - r_{12.3}^2\right)$.

2.7 Worked out examples

Example 1. Obtain the correlation coefficient between two variables x and y. Hence obtain the two regression lines and estimate y and x for x = 22 and y = 25 respectively from the following data :

x :	15	20	25	30	35	40	50	55
y :	22	24	26	28	30	36	38	40

Solution :

x	y	$u = \frac{x-30}{6}$	$v = \frac{y-30}{2}$	u^2	v^2	uv
15	22	–3	–4	9	16	12
20	24	–2	–3	4	9	6
25	26	–1	–2	1	4	2
30	28	0	–1	0	1	0
35	30	1	0	1	0	0
40	36	2	3	4	9	6
50	38	4	4	16	16	16
55	40	5	5	25	25	25
Total		6	2	60	80	67

Now n = 8, $\Sigma u = 6$, $\Sigma v = 2$, $\Sigma u^2 = 60$, $\Sigma v^2 = 80$, $\Sigma uv = 67$, $x = 30 + 5u$, $y = 30 + 2v$, $\bar{u} = \frac{6}{8}$, $\bar{v} = \frac{2}{8}$. So $\bar{x} = 30 + 5u = 30 + 5 \times \frac{6}{8} = 30 + 3.75 = 33.75, \bar{y} = 30 + 2 \times \bar{v} = 30 + 2 \times \frac{2}{8}$ = 30 + .5 = 30.5.

$$\text{Var}(u) = \frac{1}{n}\sum u^2 - \bar{u}^2 = \frac{1}{8} \times 60 - \left(\frac{6}{8}\right)^2 = \frac{15}{2} - \frac{9}{16} = \frac{120 - 9}{16} = \frac{111}{16}.$$

$$\text{Var}(v) = \frac{1}{n}\sum v^2 - \bar{v}^2 = \frac{1}{8} \times 80 - \left(\frac{2}{8}\right)^2 = 10 - \frac{1}{16} = \frac{159}{16}.$$

$$\text{Var}(x) = 5^2 \times \text{Var}(u) = \frac{111}{16} \times 25, \text{Var}(y) = 2^2 \times \text{Var}(v) = 4 \times \frac{159}{16} = \frac{159}{4}$$

$$\text{Cov}(u, v) = \frac{1}{n}\sum uv - \bar{u}\,\bar{v} = \frac{1}{8} \times 67 - \frac{6}{8} \times \frac{2}{8} = \frac{67}{8} - \frac{3}{16} = \frac{134 - 3}{16} = \frac{131}{16}$$

$s_x = s.d.(x) = \sqrt{Var(x)} = \frac{\sqrt{111} \times 5}{4} = \frac{10.5357 \times 5}{4} = 13 \cdot 17.$

$s_y = s.d.(y) = \sqrt{\frac{159}{4}} = \sqrt{39 \cdot 75} = 6 \cdot 305$

So correlation coefficient between u and v = $r_{uv} = \frac{Cov(u, v)}{\sqrt{Var(u)Var(v)}} = \frac{\frac{131}{16}}{\sqrt{\frac{111}{16} \times \frac{159}{16}}}$

$= \frac{131}{\sqrt{17649}} = \frac{131}{132 \cdot 85} = 0 \cdot 9861,\ r_{xy} = r_{uv} = 0 \cdot 9861,$

$b_{yx} = r_{xy} \frac{s_y}{s_x} = \cdot 9861 \times \frac{6 \cdot 305}{13 \cdot 17} = \frac{6 \cdot 2174}{13 \cdot 17} = 0 \cdot 472,$

$b_{xy} = r_{xy} \cdot \frac{s_x}{s_y} = \cdot 9861 \times \frac{13 \cdot 17}{6 \cdot 305} = \frac{12 \cdot 9869}{6 \cdot 305} = 2 \cdot 06.$

So the regression line of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$

i.e., y – 30.5 = 0.472 (x – 33.75) i.e., y = 0.472x + 30.5 – 15.93

i.e., y = 0.472x + 14.57

Estimated value of y for given x = 22 is y = 0.472 × 22 + 14.57 = 10.38 + 14.57

= 24.95

Regression line of x on y is $x - \bar{x} = b_{xy}(y - \bar{y})$ i.e., $x - 33.75 = 2.06(y - 30.5)$

or x = 33.75 – 62.83 + 2.06y or x = –29.08 + 2.06y

Estimated value of x for given y = 25 is x = –29.08 + 2.06 × 25 = –29.08 + 51.5

= 22.42

Example 2. Let the two regression lines be y = x – 5 and 16x = 9y + 94. Find the means of x and y, the ratio of variances and the correlation coefficient between x and y.

Solution : As two regression lines pass through the means of the variables we may write $\bar{y} = \bar{x} - 5$ and $16\bar{x} = 9\bar{y} + 94$ where $\bar{x}$ and $\bar{y}$ are means of variables x and y.

$16\bar{x} = 9(\bar{x} - 5) + 94 = 9\bar{x} + (94 - 45)$

or $7\bar{x} = 49$ or $\bar{x} = 7$

Then $\bar{y} = \bar{x} - 5 = 7 - 5 = 2.$

So the means are $\bar{x} = 7$ and $\bar{y} = 2$.

Let y = x – 5 be the regression line of y on x and

16x = 9y + 94 be the regression line of x on y.

Then the regression coefficients, $b_{yx} = 1$ and $b_{xy} = \frac{9}{16}$ so that

$r_{xy}^2 = b_{yx} \times b_{xy} = 1 \times \frac{9}{16} = \frac{9}{16} < 1$. So the assumption is correct. Hence $r_{xy} = \frac{3}{4} = 0.75$, i.e., the correlation coefficient between x and y is 0.75 since r_{xy}, b_{xy}, b_{yx} have same sign.

As $\frac{b_{xy}}{b_{yx}} = \frac{r_{xy} \cdot \frac{\sigma_x}{\sigma_y}}{r_{xy} \cdot \frac{\sigma_y}{\sigma_x}} = \frac{\sigma_x^2}{\sigma_y^2}$, so $\frac{\sigma_x^2}{\sigma_y^2} = \frac{\frac{9}{16}}{1} = \frac{9}{16}$.

Example 3. Calculate Spearman's rank correlation coefficients between two sets of ranks given below in two cases by supervisors of 10 workers working under them in order of efficiency and then comment.

(i)

Workers	1	2	3	4	5	6	7	8	9	10
Supervisor I	5	6	1	2	3	8	9	4	7	10
Supervisor II	6	5	1	3	2	9	7	4	10	8

(ii)

Workers	1	2	3	4	5	6	7	8	9	10
Supervisor I	5	6	1	2	3	$8\frac{1}{2}$	$8\frac{1}{2}$	4	7	10
Supervisor II	$5\frac{1}{2}$	$5\frac{1}{2}$	2	2	2	9	8	4	10	7

Solution : Let x_i, y_i be the ranks of the i-th worker given by supervisors I and II respectively. Let $d_i = x_i - y_i$, for i = 1, 2, 3, 4, ..., 10. There are no ties in ranks of workers in (i).

(i)

i	1	2	3	4	5	6	7	8	9	10	Total
x_i	5	6	1	2	3	8	9	4	7	10	
y_i	6	5	1	3	2	9	7	4	10	8	
d_i	–1	1	0	–1	1	–1	2	0	–3	2	
d_i^2	1	1	0	1	1	1	4	0	9	4	22

Here n = 10.

So Rank correlation coefficient $r_R = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 22}{10(100-1)} = 1 - \frac{132}{990}$

$= \frac{858}{990} = \frac{26}{30} = 0 \cdot 87.$

Though the supervisors do not perfectly agree in their ranking of the workers but they nearly agree in their judgements of ranking the workers.

(ii)

i	1	2	3	4	5	6	7	8	9	10	Total
x_i	5	6	1	2	3	$8\frac{1}{2}$	$8\frac{1}{2}$	4	7	10	
y_i	$5\frac{1}{2}$	$5\frac{1}{2}$	2	2	2	9	8	4	10	7	
d_i	$-\frac{1}{2}$	$\frac{1}{2}$	–1	0	1	$-\frac{1}{2}$	$\frac{1}{2}$	0	–3	3	
d_i^2	$\frac{1}{4}$	$\frac{1}{4}$	1	0	1	$\frac{1}{4}$	$\frac{1}{4}$	0	9	9	21

For ranking by supervisor I there is one tie case of length 2. Then

$$T_x = \frac{1}{10}\left(\frac{2(2^2-1)}{12}\right) = \frac{6}{120} = .05$$

For ranking by supervisor II there are two tie cases of lengths 2 and 3.

Then $T_y = \frac{1}{10}\left[\frac{2(2^2-1)}{12} + \frac{3(3^2-1)}{12}\right] = \frac{1}{10}\left[\frac{6}{12} + 2\right] = \frac{5}{20} = 0.25.$

Here n = 10.

So $r_R = \dfrac{\frac{n^2-1}{12} - \frac{T_x + T_y}{2} - \frac{1}{2n}\sum_{i=2}^{n} d_i^2}{\sqrt{\left(\frac{n^2-1}{12} - T_x\right)\left(\frac{n^2-1}{12} - T_y\right)}} = \dfrac{\frac{99}{12} - \frac{.05+.25}{2} - \frac{1}{20}\times 21}{\sqrt{\left(\frac{99}{12} - .05\right)\left(\frac{99}{12} - .25\right)}}$

$$= \frac{8.25 - .15 - 1.05}{\sqrt{(8.25 - .05)(8.25 - .25)}} = \frac{7.05}{\sqrt{8.2 \times 8}} = \frac{7.05}{\sqrt{65.6}} = \frac{7.05}{8.1} = 0.87$$

As 0.87 is near 1 the supervisors nearly agree in their judgements in ranking the workers.

Example 4. In a three variate multiple correlation analysis the following results were found :

$$\bar{x}_1 = 60,\ \bar{x}_2 = 70,\ \bar{x}_3 = 100,$$

$$s_1 = 3,\ s_2 = 4,\ s_3 = 5,$$

$r_{12} = 0.7,\ r_{13} = 0.6,\ r_{23} = 0.4.$

where for the i-th variable x_i, mean is $\bar{x}_i$ and standard deviation is s_i for i = 1, 2, 3, and r_{ij} is the correlation coefficient between x_i and x_j for i, j = 1, 2, 3.

Find the regression line of x_1 on x_2 and x_3, multiple correlation coefficient $r_{1.23}$ and partial correlation coefficient $r_{12.3}$. Also estimate x_1 when $x_2 = 80$ and $x_3 = 120$.

Solution :

Multiple regression line is given by $\begin{vmatrix} \frac{x_1 - \bar{x}_1}{s_1} & \frac{x_2 - \bar{x}_2}{s_2} & \frac{x_3 - \bar{x}_3}{s_3} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix} = 0$

i.e. $\begin{vmatrix} \frac{x_1 - 60}{3} & \frac{x_2 - 70}{4} & \frac{x_3 - 100}{5} \\ 0.7 & 1 & 0.4 \\ 0.6 & 0.4 & 1 \end{vmatrix} = 0$

or $\left(\frac{x_1 - 60}{3}\right) \times .84 - \left(\frac{x_2 - 70}{4}\right) \times .46 - \left(\frac{x_3 - 100}{5}\right) \times .32 = 0$

or $.28\ (x_1 - 60) = .115\ (x_2 - 70) + .064\ (x_3 - 100)$ is the multiple regression equation of x_1 on x_2 and x_3.

Estimate of x_1 for $x_2 = 80$ and $x_3 = 120$ is obtained as

$.28(x_1 - 60) = .115(80 - 70) + .064(120 - 100)$

$= 1.15 + 1.28 = 2.43$

Therefore, $x_1 = 60 + \frac{2.43}{.28} = 60 + 8.68 = 68.68$

This is the estimated value of x_1 for $x_2 = 80$ and $x_3 = 120$.

Also $r_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} = \sqrt{\frac{(0.7)^2 + (0.6)^2 - 2 \times 0.7 \times 0.6 \times 0.4}{1 - (0.4)^2}}$

$= \sqrt{\frac{.49 + .36 - .336}{1 - .16}} = \sqrt{\frac{.514}{.84}} = \sqrt{.6119} = .7822$

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}} = \frac{0.7 - 0.6 \times 0.4}{\sqrt{\left[1 - (0.6)^2\right]\left[1 - (0.4)^2\right]}} = \frac{0.7 - 0.24}{\sqrt{.64 \times .84}}$$

$$= \frac{0.46}{\sqrt{0.5376}} = \frac{0.46}{0.7332} = 0.6274$$

2.8 Summary

This chapter consists of correlation (linear correlation), its measures and the properties, regression lines, properties of regression coefficients, angle between two regression lines, coefficient of determination along with explained and unexplained variations, rank correlation, multiple regression equation, multiple correlation coefficient, partial correlation coefficient and all the formulae and procedures supported by worked out examples.

2.9 Exercises

1. Define the correlation coefficient. State and prove its important properties.
2. Define the term 'regression'. What are regression lines and regression coefficients? State and prove the important properties of a regression coefficient.
3. Explain what is meant by bivariate frequency distribution. Define marginal frequency distribution.
4. What is scatter diagram? Indicate by means of suitable scatter diagram different values of correlation coefficient that may exist between the variables in bivariate data.
5. Deduce the expression of regression coefficients and regression lines by the method of least squares.
6. Prove that correlation coefficient lies between –1 and + 1.
7. What is rank correlation and how it is measured? Why is rank correlation used?
8. Derive the Spearman's rank correlation coefficient r_R separately in case of ties and in case no tie and prove that r_R lies between –1 and +1 in both the cases :
9. Define regression coefficient and correlation coefficient and prove that the arithmetic mean of regression coefficients is greater than the correlation coefficient when the latter is positive.
10. Determine the angle between the regression lines.

11. Define coefficient of determination and explained variation in dependent variable.

12. Marks of 5 students in Mathematics and Statistics are given below :

Mathematics :	38	48	43	40	41
Statistics :	31	38	43	33	35

Determine the product moment correlation coefficient between marks in Mathematics and marks in Statistics. Determine the regression lines. When marks of a student in Mathematics is 42, estimate his marks in Statistics.

13. Two regression lines are $x + 2y = 5$ and $2x + 3y = 8$. $s_x^2 = 12$. Determine the values of $\bar{x}$, $\bar{y}$, s_y^2 and r.

14. The ranking of 8 individuals at the start and on the completion of a course of training are as follows :

Individuals :	A	B	C	D	E	F	G	H
Rank before :	5	2	8	1	4	6	3	7
Rank after :	4	5	7	3	2	8	1	6

Calculate Spearman's rank correlation coefficient.

15. Ten competitors in a musical contest were ranked by two judges A and B in the following manner. Calculate Spearman's rank correlation coefficient.

Serial no. of a candidate :	1	2	3	4	5	6	7	8	9	10
Rank by Judge A :	10	6	5	1	3	2	4	8	8	8
Rank by Judge B :	5	5	8	3	7	10	1.5	1.5	5	9

16. Obtain the multiple regression equation of variable x_1 on variables x_2 and x_3 in terms of means, standard deviations and correlations of the variables.

17. Prove that $1 - r_{1.23}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)$

Use this relation to show that the multiple correlation coefficient is numerically greater than or equal to any of the total or partial correlation coefficients of x_1 with other variables.

18. In a trivariate distribution

means	:	$\bar{x}_1 = 28.02,$	$\bar{x}_2 = 4.91,$	$\bar{x}_3 = 594$
s.d.'s	:	$s_1 = 4.4,$	$s_2 = 1.1,$	$s_3 = 0.80$
Correlation coefficients	:	$r_{12} = 0.80,$	$r_{23} = -0.56,$	$r_{13} = -0.40$

Find the correlation coefficients $r_{12.3}$, $r_{1.23}$.

Find the multiple regression line of x_1 on x_2 and x_3 and estimate the value of x_1 when $x_2 = 6.0$ and $x_3 = 650$.

19. You are given the following in formation :

	x	y
Arithmetic mean	36	85
Standard deviation	11	8

Correlation coefficient between x and y = 0.66

(i) Find the two regression lines and (ii) Estimate the value of x when y = 75.

20. Given the variance of x = 9 and the regression equations

$8x - 10y + 66 = 0$ and $40x - 18y = 214$.

Find the following :

(i) Average values of x and y,

(ii) Correlation coefficient between x and y

and (iii) Standard deviation of y.

21. The two lines of regression are given to be $5x + 7y - 22 = 0$ and $6x + 2y - 20 = 0$. Identify the regression equations. Find the mean values of x and y. If the variance of y is 15, find the standard deviations of x.

22. The following information were obtained from two variables x and y.

$\bar{x} = 20$, $\bar{y} = 15$, $\sigma x = 4$, $\sigma y = 3$, $r_{xy} = 0.7$.

Obtain the two regression equations and find the most likely value of y when x = 24.

23. Prove that the coeffient of correction is the geometric mean of two regression coefficients. What is the sign of the coefficients of correlation?

24. The coefficient of correlation between income and consumption is found to be 0.9. Interpret this result as clearly as you can.

25. With variables x_1, x_2 and x_3 a scholar has found that $r_{12} = 0.9057$, $r_{13} = 0.3287$ and $r_{23} = 0.8139$. Examine, explaining the basis, if the results can be accepted as free from computational mistakes.

26. Distinguish between the correlation approach and the regression approach to the analysis of bivariate data.

27. Define multiple correlation coefficient and partial correlation coefficient and indicate how they differ from the simple correlation coefficient.

28. Deduce from the first principle the relation between $r_{12.3}$ on the one hand and r_{12}, r_{13} and r_{23} on the other. Show that r_{12} may not be zero, even when $r_{12.3}$ is. Give your interpretation of the result.

29. Without computation find the coefficient of correlation between x and y in the following two cases :

case (a)		case (b)	
x	y	x	y
50	30	110	90
70	45	100	95

Substantiate your answer with proper reasons.

30. Comment and clarify the following statement :

"The degree of linear dependence shown by a value of r = 0.8 is 4 times as strong as that shown by a value of r = 0.4." State and prove the proposition on the basis of which you make your comment.

31. Let x and y be two independent variables with standard deviations S_x and S_y respectively. Show that the coefficient of correlation between x and x ± y is

$$\frac{S_x}{\sqrt{S_x^2 + S_y^2}}.$$

32. Find the coefficient of correlation between x and y from the relation ax + by + c = 0. Is the result surprising to you? If not, why not?

33. Establish the inequality

$$r_{12}^2 + r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23} \leq 1.$$

Why is this inequality important?

34. Obtain the multiple regression equation of x_1 on x_2 and x_3 in terms of the means, the standard deviations and the intercorrelations of the variables.

35. Define rank correlation. Write down Spearman's formula for rank correlation coefficient, r_R. What are the limits of r_R? Interpret the case when r_R takes the minimum value.

36. In a certain investigation, the following values were obtained : $r_{12} = 0.6$, $r_{13} = 0.4$ and $r_{23} = 0.7$. Are these values consistent?

37. Show that the partial correlation coefficient, $r_{12.3}$, is the geometric mean of the two partial regression coefficients and the sign of $r_{12.3}$ is the same as that of $b_{12.3}$ and $b_{21.3}$.

38. If $r_{12} = 0.80$, $r_{13} = 0.40$ and $r_{23} = -0.56$, find the values of $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$.

39. For the variables x and y the equations of two regression lines are $4x - 5y + 33 = 0$ and $20x - 9y = 107$. Identify the regression line of y on x and that of x on y. Find out the co-ordinates of the point of intersection of the two regression lines. Can you find out the coefficient of correlation between x and y?

40. The following sums have been obtained from 100 observation pairs :

    $\Sigma x = 12500$, $\Sigma y = 8000$, $\Sigma x^2 = 1585000$, $\Sigma y^2 = 648100$ and $\Sigma xy = 1007425$.

    (a) Find the regression equation of y on x and estimate the value of y when $x = 130$.

    (b) Compute the coefficient of correlation between x and y.

41. Find the angles between two regression lines. Hence establish the relation between the coefficient of correlation and the angle between the two regression lines.

42. Find the regression equation of y on x and that of x on y from the following information. Also find the value of (i) y when $x = 4$ and (ii) x when $y = 3$.

43. If two variables are independent their coefficient of correlation is zero. Is the converse true? Illustrate your answer with the help of an example.

44. Suppose the coefficient of correlation between income and consumption is 0.9. How can you interpret this result? Explain your answer as clearly as you can.

45. If the coefficient of correlation between x and y is 0.5, what will be the coefficient of corelation between 5x and –3y? Give reasons your answer.

46. In the context of the two variable linear regression analysis establish the relation

    $S^2_{y.x} = S^2_y(1 - r^2)$, all the symbols have usual meaning.

    Hence interpret the cases (a) $r = 0$ and (b) $r = \pm 1$.

47. Ten competitors in a musical contest were ranked by Judges X, Y and Z in the following order :

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

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

48. Prove that $b_{12.3} \times b_{23.1} \times b_{31.2} = r_{12.3} \times r_{23.1} \times r_{31.2}$

Solution : $b_{12.3} \times b_{23.1} \times b_{31.2}$

$$= \frac{\sigma_1}{\sigma_2}\left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right) \times \frac{\sigma_2}{\sigma_3}\left(\frac{r_{23} - r_{12}.r_{13}}{1 - r_{13}^2}\right) \times \frac{\sigma_3}{\sigma_1}\left(\frac{r_{13} - r_{23} \times r_{12}}{1 - r_{12}^2}\right)$$

$$= \frac{r_{12} - r_{13}.r_{23}}{\sqrt{\left(1 - r_{23}^2\right)\left(1 - r_{23}^2\right)}} \times \frac{r_{23} - r_{12}.r_{13}}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{13}^2\right)}} \times \frac{r_{13} - r_{23}.r_{12}}{\sqrt{\left(1 - r_{23}^2\right)\left(1 - r_{12}^2\right)}}$$

$= b_{12.3} \times b_{23.1} \times b_{31.2}$

Hence proved.

49. For a Invariate distribution the following information have been provided :

Means :	$\bar{x}_1 = 50$	$\bar{x}_2 = 60$	$\bar{x}_3 = 90$
S.D.'s :	$\sigma_1 = 2$	$\sigma_2 = 5$	$\sigma_3 = 8$
Coefficients of correlation :	$r_{12} = 0.3$	$r_{13} = 0.5$	$r_{23} = 0.6$

Compute (i) partial regression coefficients and (ii) partial correlation coefficients.

50. Prove that $r_{12.3}^2 = b_{12.3} \times b_{21.3}$

2.10 Suggested Readings

1. Kenney, J. F. and Keeping, E. S. *Mathematics of Statistics Part I and II*, Van Norstrand 1954 and Affiliated East West Press.
2. Goon, A. M., Gupta, M. K. and Dasgupta, B. *Fundamentals of Statistics,* Vol. I, World Press Pvt. Ltd. 2002, Kolkata.
3. Chaudhuri, S. B. *Elementary Statistics,* Vol. I, Shraddha Prakashan, 1986, Kolkata.
4. Yule, G. V. and Kendall, M. G. *Introduction to the Theory of Statistics,* Charles Griffin, 1953.
5. Rao, C. R. *Advanced Statistical Methods in Biometric Research,* John Wiley 1952.

Unit 3 ❑ Interpolation

Structure

3.0 Objectives

In any computational problem all kinds of numbers, rational or irrational, real or complex, are considered. Our objective will be to get the result with greatest possible accuracy. Numerical data for solving problems are not usually exact. They are just approximations, correct to desired number of figures. We can minimise the computational errors as much as we want to and the methods are given in this chapter.

3.1 Introduction

Existing analytical methods cannot solve all mathematical problems or even if they can solve, for some of them the solutions are so complex that the relevant numerical information from such solution cannot be derived easily. In such cases some numerical methods can be used to get approximate solution of such problems. They

are approximate in the sense that they are correct to certain degree of accuracy. Here solutions are approximate due to approximation of data and approximation of the method of solution. Some numerical methods of interpolation and extrapolation along with the concepts of significant figures being rounded off in numbers are discussed here.

3.2 Significant figures, rounding off numbers and errors

Digits 1, 2, 3, ..., 9 are always significant figures. 0's are sometimes considered as significant figure but not always. In the number 0.021034000, approximated from 0.0210340002, first two zeroes are not significant figures but all other digits are significant figures. Here last three zeroes are significant as they give positions of the number of digits considered, because the number is considered upto 9 places after the decimal point.

The procedure of retaining certain number of digits counting from left in the exact number T and dropping the other numbers on the right causing least possible error for approximating T is called the rounding off the number T. A number is considered correct to p places of decimal when it is rounded off to p places after the decimal point. Also a number correct to p significant figures according as the number is rounded off to p figures starting from the first significant figure.

The rules for rounding off numbers either to p places after decimal or to p significant figures are as follows :

(i) Discard all the digits after p places of decimal point or after p significant figures in a non-decimal number and in second case multiply the number retained, having p significant digits, by 10^k where k = number of discarded digits.

(ii) If the first discarded digit is less than 5, the last of the retained digits remains unchanged.

(iii) If the first of the discarded digits is more than 5, add 1 to the last retained digit.

(iv) If the first of the discarded digit is 5 following atleast one non-zero digit add 1 to the last retained digit.

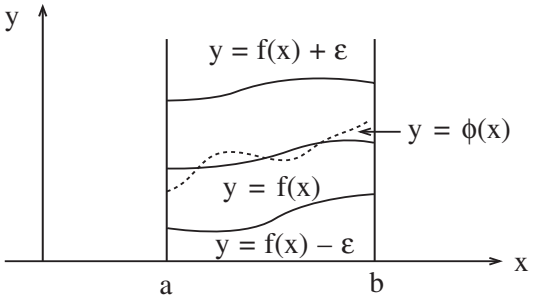
(v) When the only discarded digit is 5, the last of the retained digits will be unaltered if it is an even number and will be increased by 1 if it is an odd number.

For example, the numbers 1.6344, 11.364, 1.6354, 1.635, 1.645 rounded off to 3 significant figures 1.63, 11.4, 1.64, 1.64, 1.64 respectively. Also the numbers 16344, 11364, 16354, 1635, 1645 and 1604 are rounded off to 3 significant figures as 163×10^2, 114×10^2, 164×10^2, 164×10, 164×10 and 160×10 respectively.

If T is the exact value and A is the approximate value of it, then absolute error of A is $|T - A|$, relative error of A is $\frac{|T - A|}{T}$ and percentage error of A is 100 x relative error.

3.3 Interpolation and extrapolation

In y = f(x), y is considered as a dependent variable or entry and x as an independent variable or argument. Interpolation (extrapolation) is some procedure of obtaining the value of the entry y for some given value of argument x lying in (outside) the interval of two extreme values of a given finite set of values of x for which the values of y are available. Here f(x) is approximated by a polynomial of suitable degree using Weirstrass theorem, "Every function f(x), which is continuous in an interval [a, b], can be represented in that interval to any desired degree of accuracy by a polynomial $\phi(x)$ such that $|f(x) - \phi(x)| < \varepsilon$ for every value of x in that interval, where ε is any pre-assigned small positive quantity".



Geometrically, this theorem means that if f(x) is continuous in [a, b] it is possible to find a polynomial $y = \phi(x)$ whose graph lies in the region bounded by the graphs $y = f(x) + \varepsilon$ and $y = f(x) - \varepsilon$ for all values of x lying between a and b, however small positive number ε may be.

For example, the following table gives the values of $\log_e^x$ corresponding to certain values of x.

x	:	45	46	47	48	49
$\log_e^x$	:	3.80666	3.82864	3.85015	3.87120	3.89182

Here the values of $\log_e x$ are rounded off to 5 significant digits after decimal. To get the values of $\log_e 45.4$, $\log_e 47.6$ and $\log_e 48.8$ we use interpolation formulae and to get the value of $\log_e 44.8$ and $\log_e 49.3$ we use extrapolation formula. For extrapolation the values of arguments should not be far off from the extreme values of arguments.

Suppose corresponding to (n + 1) given values $x_0, x_1, x_2, ..., x_n$ of argument x such that $x_0 < x_1 < x_2 < ... < x_n$, $y_0, y_1, y_2, ..., y_n$ are the respective known values of entry y satisfying y = f(x) where the functional form f( ) is either unknown or of complicated nature. As (n + 1) points $(x_0, y_0), (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ are given, we can approximate f(x) by a polynomial $\phi(x)$ of degree at most n,

$$\phi(f) = a_0 + a_1x + a_2x^2 + ... + a_nx^n$$

which passes through the given (n + 1) points. The (n + 1) constants $a_0, a_1, a_2, ... a_n$ can be determined from the (n + 1) relations.

$$y_0 = \phi(x_0),\ y_1 = \phi(x_1),\ y_2 = \phi(x_2),\ ....\ y_n = \phi(x_{n)}$$

where $\phi(x_i) = a_0 + a_1x_i + a_2x_i^2 + ... + a_nx_i^n$ for

$i = 0, 1, 2, ..., n.$ (2)

Also $\phi(x)$ can be written in any of the form, each of degree *n*

(i) $\phi(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1) + ... + b_n(x - x_0)(x - x_1)....(x - x_{n-1})$, (3)

(ii) $\phi(x) = c_0 + c_1(x - x_n) + c_2(x - x_n)(x - x_{n-1}) + ... + c_n(x - x_n)(x - x_{n-1})...(x - x_1)$,(4)

(iii) $\phi(x) = d_0(x - x_1)(x - x_2)...(x - x_n) + d_1(x - x_0)(x - x_2)(x - x_3)...(x - x_n) + ...$

$$+ d_n(x - x_0)(x - x_1) ... (x - x_{n-1}) \quad (5)$$

where $b_0, b_1, b_2, ..., b_n$ in (3), $c_0, c_1, c_2, ..., c_n$ in (4) and $d_0, d_1, d_2, ..., d_n$ in (5) can be determined from $y_i = \phi(x_i)$, i = 0, 1, 2, ..., n.

3.3.1 Use of operators Δ and E

When the values $x_0, x_1, ..., x_n$ of argument x are equidistant and are of interval length h (say), i.e., $x_1 - x_0 = x_2 - x_1 = ... = x_n - x_{n-1} = h$ satisfying $x_0 < x_1 < x_2 < ... < x_n$ then for values $y_0, y_1, ..., y_n$ of entry y, where $y_i = f(x_i)$, i = 0, 1, 2, ..., n, define the first differences (or first order finite differences) as

$$\Delta f(x_i) = f(x_i + h) - f(x_i) = f(x_{i+1}) - f(x_i) = Ef(x_i) - f(x_i)$$

$$= (E - 1)\ f(x_i) \quad \text{for } i = 0, 1, ..., n–1,$$

i.e., $\Delta y_i = (E - 1)y_i$ for i = 0, 1, ..., n–1

and the second differences (or second order finite differences)

$$\Delta^2 f(x_i) = \Delta(\Delta f(x_i)) = \Delta(f(x_i - h) - f(x_i)) = \Delta f(x_i + h) - \Delta f(x_i)$$
$$= [f(x_i + 2h) - f(x_i + h)] - [f(x_i + h) - f(x_i)]$$
$$= f(x_i + 2h) - 2f(x_i + h) + f(x_i)$$
$$= E^2 f(x_i) - 2Ef(x_i) + f(x_i)$$
$$= (E^2 - 2E + 1)f(x_i)$$
$$= (E - 1)^2 f(x_i),$$

i.e., $\Delta^2 y_i = (E - 1)^2 y_i = y_{i+2} - 2y_{i+1} + y_i$ for i = 0, 1, ..., n–2.

Proceeding this way

$$\Delta^3 f(x_i) = f(x_i + 3h) - 3f(x_i + 2h) + 3f(x_i + h) - f(x_i)$$
$$= E^3 f(x_i) - 3E^2 f(x_i) + 3E\, f(x_i) - f(x_i)$$
$$= (E^3 - 3E^2 + 3E - 1)f(x_i) = (E - 1)^3 f(x_i)$$

i.e., $\Delta^3 y_i = y_{i+3} - 3y_{i+2} + 3y_{i+1} - y_i$
$= (E - 1)^3 y_i$ for i = 0, 1, 2, ..., n–3.

Proceeding this way one can get

$\Delta^r y_i = (E - 1)^r y_i$ for i = 0, 1, 2, ..., n–r.

So $\Delta^r \equiv (E - 1)^r$ for r = 1, 2, 3, ...

Here Δ and E are operators and are not quantities

Also $\Delta^r f(x_0) = (E-1)^r f(x_0) = \left[E^r - \binom{r}{1}E^{r-1} + \cdots + (-1)^r\right] f(x_0) = f(x_r) - \binom{r}{1} f(x_{r-1})$

+ .... + $(-1)^r f(x_0)$ where in $\Delta^r f(x_0)$, Δ is operated r times on $f(x_0)$ i.e.,

$\Delta^r f(x_0) = (\Delta\ \Delta\ \Delta$ ... r times) $f(x_0) = \Delta^{r-1}(\Delta f(x_0)) = \Delta^{r-2}\left(\Delta^2 f(x_0)\right)$ etc.

Hence $\Delta^r y_0 = y_r - \binom{r}{1} y_{r-1} + \binom{r}{2} y_{r-2} - \ldots\ldots + (-1)^n_{y_0}$, r = 1, 2, ... n (11)

So laws of algebra satisfy binomial expansion of operators. Generally, if h is a positive constant, called interval of differencing, then

$$\Delta f(x) = f(x + h) - f(x) = Ef(x) - f(x) = (E - 1)f(x)$$

is the first difference of f(x), also sometimes called first forward (or descending) difference of f(x),

Similarly, $\Delta^2 f(x) = (E - 1)^2 f(x) = (E^2 - 2E + 1)f(x) = E^2 f(x) - 2Ef(x) + f(x)$
$= f(x + 2h) - 2f(x + h) + f(x)$ is the second difference of f(x)

Thus the r-th order difference of f(x) is

$$\Delta^r f(x) = (E-1)^r f(x) = \left[E^r - \binom{r}{1}E^{r-1} + \binom{r}{2}E^{r-2} - \ldots + (-1)^r\right] f(x)$$

$$= E^r f(x) - \binom{r}{1} E^{r-1} f(x) + \binom{r}{2} E^{r-2} f(x) - \ldots + (-1)^r f(x)$$

$$= f(r + rh) - \binom{r}{1} f(x + (r-1)h) + \binom{r}{2} f(x + (x-2)h) - \ldots + (-1)^r f(x) \qquad (12)$$

for r = 1, 2, ...

If (x_0, y_0), (x_1, y_1), ..., (x_4, y_4) are the 5 given points where $x_0 < x_1 < x_2 < x_3 < x_4$, construct the finite difference table (also called diagonal difference table) as follows :

x	y_0	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
x_0	y_0				
		Δy_0			
x_1	y_1		$\Delta^2 y_0$		
		Δy_1		$\Delta^3 y_0$	
x_2	y_2		$\Delta^2 y_1$		$\Delta^4 y_0$
		Δy_2		$\Delta^3 y_1$	
x_3	y_3		$\Delta^2 y_2$		
		Δy_3			
x_4	y_4				

where first differences $\Delta y_i = y_{i+1} - y_i$ are placed in the column of Δy and in the middle of the rows corresponding to position of y_i and y_{i+1} for i = 0, 1, 2, 3, second differences $\Delta^2 y_i = \Delta y_{i+1} - \Delta y_i$ are placed in the column of $\Delta^2 y$ and in the middle of the rows corresponding to Δy_i and Δy_{i+1} for i = 0, 1, 2, and the procedure is repeated for third differences $\Delta^3 y_i = \Delta^2 y_{i+1} - \Delta^2 y_i$ for i = 0, 1 and for fourth difference $\Delta^4 y_0 = \Delta^3 y_1 - \Delta^3 y_0$. So if 5 values y_0, y_1, y_2, y_3, y_4 of ordinate y are given then for equidistant values x_0, x_1, x_2, x_3, x_4 satisfying $x_0 < x_1 < x_2 < x_3 < x_4$, 4 values Δy_0, Δy_1 Δy_2, Δy_3 of Δy, 3 values Δy_0, $\Delta^2 y_1$, $\Delta^2 y_2$ of $\Delta^2 y$, 2 values $\Delta^3 y_0$, $\Delta^3 y_1$ of $\Delta^3 y$ and one value $\Delta^4 y_0$ of $\Delta^4 y$ can be obtained, as received in finite difference table above. Then this fixed single value $\Delta^4 y_0$ of $\Delta^4 y$ is assumed to be constant and as a result all higher order finite differences $\Delta^r y$ are zeroes for r > 4.

3.3.2 Rules of finite differences

1. $\Delta[f(x) \pm g(x)] = \Delta f(x) \pm \Delta g(x)$ where f(x) and g(x) are two functions of argument x.
2. $\Delta[cf(x)] = c\Delta f(x)$ where c is constant and f(x) is a function of x.
3. $\Delta c = 0$ where c is constant.

4(a) The first order difference of a polynomial f(x) of degree n is a polynomial of degree n – 1.

Proof : Let $f(x) = a_0x^n + a_1x^{n-1} + ... + a_{n-1}x + a_n$

where a_0, a_1, ..., a_n are constants. Consider h = increment of x. Then

$$\Delta f(x) = f(x + h) - f(x)$$

$$= \left[a_0(x+h)^n + a_1(x+h)^{n-1} + \ldots + a_{n-1}(x+h) + a_n\right] - \left[a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n\right]$$

$$= a_0\left[(x+h)^n - x^n\right] + a_1\left[(x+h)^{n-1} - x^{n-1}\right] + \ldots + a_{n-1}\left[(x+h) - x\right] + a_n[1-1]$$

$$= a_0\left(nhx^{n-1} + \frac{n(n-1)}{2!}h^2x^{n-2} + \ldots + h^n\right) + a_1\left[(n-1)hx^{n-2} + \ldots + h^{n-1}\right] + \ldots + ha_{n-1}$$

$$= nha_0x^{n-1} + \left[\frac{n(n-1)}{2}h^2a_0 + (n-1)ha_1\right]x^{n-2} + \ldots + \left[a_0h^n + a_1h^{n-1} + \ldots + a_{n-1}h\right]$$

$$= b_0x^{n-1} + b_1x^{n-2} + \ldots + b_{n-1}, \qquad (13)$$

where $b_0 = nha_0, b_1 = (n-1)h\left[a_1 + \frac{nha_0}{2}\right], \ldots, b_{n-1} = \left[a_0h^n + a_1h^{n-1} + \ldots + a_{n-1}h\right]$

From (13) it is found that $\Delta f(x)$ is a polynomial of degree n – 1 with coefficient of x^{n-1} as nha_0.

4(b) If k is a positive integer such that $k \leq n$, the k-th order finite difference of a polynomial of degree n, is a polynomial of degree n – k with coefficient of x^{n-k} as $n(n-1) \ldots (n-k+1)\, h^k\, a_0$, where a_0 is coefficient of x^n in the polynomial and h is the interval of differencing.

Proof : $\Delta^2 f(x) = \Delta(\Delta f(x) = \Delta\left(b_0x^{n-1} + b_1x^{n-2} + \ldots + b_{n-1}\right)$,

where $f(x) = a_0x^n + a_1x^{n-1} + ... + a_n$.

So $\Delta^2 f(x) = b_0\left[(x+h)^{n-1} - x^{n-1}\right] + b_1\left[(x+h)^{n-2} - x^{x-2}\right] + \ldots$

$$+ b_{n-2}[(x+h) - x] + b_{n-1}[1-1]$$

$$= b_0(n-1)hx^{n-2} + \left[\frac{(n-1)(n-2)}{2}h^2b_0 + (n-2)hb_1\right]x^{n-3} + \ldots.$$

$$+ \left[b_0h^{n-1} + b_1h^{n-2} + \ldots + b_{n-2}h\right]$$

$$= c_0x^{n-2} + c_1x^{n-3} + \ldots + c_{n-2}(\text{say}),$$

This is a polynomial of degree n – 2 with coefficient of x^{n-2} as $c_0 = b_0\ (n-1)h = a_0\ n\ (n-1)h^2$ obtained from (13).

Proceeding this way $\Delta^k f(x)$ can be seen as a polynomial of degree n – k with coefficient of x^{n-k} as $a_0\ n\ (n-1) \ldots (n-k+1)\ h^k$.

4(c) n-th order difference of a polynomial of degree *n* is constant, and is equal to is $n!h^na_0$.

Proof. This result can be proved by n repetitions of the operation in property 4(a) on $f(x) = a_0x^n + a_1x^{n-1} + ... + a_n$.

4(d)$\Delta^k f(x) = 0$ for $k > n$ when f(x) is a polynomial of degree *n*.

Proof : As f(x) is a polynomial of degree *n* with coefficient of x^n as a_0 and interval of differencing *h*, $\Delta^n f(x) = a_0 n! h^n$ by property 4(c). Then $\Delta^{n+1} f(x) = 0$ since $a_0 n! h^n$ is constant and all higher order differences of f(x) is zero.

3.3.3 Newton's Forward Interpolation Formula

Let $y_0, y_1, y_2, ..., y_n$ be the (n + 1) tabulated values of the function y = f(x) corresponding to the equispaced arguments $x_0, x_1, x_2, ..., x_n$ with interval of length *h* where $x_0 < x_1 < x_2 < ... < x_n$, $x_i = x_0 + ih$ and $y_i = f(x_i)$, i = 0, 1, 2, ..., n. To find y corresponding to an intermediate value of x lying near the beginning of the tabulated values of argument Newton's forward interpolation formula can be used, as

$$y = y_0 + u\Delta y_0 + \frac{u(u-1)}{2!}\Delta^2 y_0 + ... + \frac{u(u-1)(u-2)...(u-n+1)}{n!}\Delta^n y_0$$

where $u = \dfrac{x - x_0}{h}$

Proof : (n + 1) points $(x_0, y_0), (x_1, y_1), ..., (x_n, y_n)$ are given for equidistant arguments of interval length h i.e. $x_1 - x_0 = x_2 - x_1 = ... x_n - x_{n-1} = h$ where $y_i = f(x_i)$, i = 0, 1, 2, ..., n. y = f(x) can be assumed to be a polynomial $y = \phi(x)$ of degree n (when $\Delta^n y_0 \neq 0$) passing through the (n + 1) given points, i.e., $\phi(x_i) = y_i$, i = 0, 1, 2, ..., n, as $\Delta^n y_0$ is a single value and assumed to be constant. Let

$$\phi(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + ... + a_n(x - x_0)(x - x_1)...(x - x_{n-1}) \quad (14)$$

and $a_0, a_1, ..., a_n$ are to be determined from $y_i = \phi(x_i)$, i = 0, 1, 2, ..., n

For $x = x_0$, $y_0 = a_0$

For $x = x_1$, $y_1 = a_0 + a_1.h$ or $a_1 = \dfrac{y_1 - y_0}{h} = \dfrac{\Delta y_0}{h}$.

For $x = x_2$, $y_2 = a_0 + a_1.2h + a_2.2h.h$ i.e., $y_2 = y_0 + 2(y_1 - y_0) + 2h^2a_2$

or $2!h^2.a_2 = y_2 - 2y_1 + y_0 = \Delta^2 y_0$ or $a_2 = \dfrac{\Delta^2 y_0}{2!h^2}$.

For $x = x_3$, $y_3 = a_0 + a_1.3h + a_2 . 3h . 2h + a_3 . 3h . 2h . h$

or $3!h^3a_3 = y_3 - y_0 - 3\Delta y_0 - 3\Delta^2 y_2 = y_3 - y_0 - 3(y_1 - y_0) - 3(y_2 - 2y_1 + y_0)$
$= y_3 - 3y_2 + 3y_1 - y_0 = \Delta^3 y_0$

or $a_3 = \dfrac{\Delta^3 y_0}{3!h^3}$

Proceeding this way get $a_n = \frac{\Delta^n y_0}{n! h^n}$ for $x = x_n$.

Thus $y = \phi(x)$ takes the form as in equation (14) for a given value x of argument lying within outside the interval (x_0, x_n) but having a value close to x_0.

$$\text{i.e., } y = y_0 + (x - x_0)\frac{\Delta y_0}{h} + \frac{(x - x_0)(x - x_1)}{2! h^2}\Delta^2 y_0 + \frac{(x - x_0)(x - x_1)(x - x_2)}{3! h}\Delta^3 y_0$$

$$+ \cdots + \frac{(x - x_0)(x - x_1)\cdots(x - x_{n-1})}{2! h^n}\Delta^n y_0 .$$

If $u = \frac{x - x_0}{h}$, $\frac{x - x_i}{h} = \frac{(x - x_0) - (x_i - x_0)}{h} = u - i$ for i = 1, 2, ..., n – 1

and the above formula can be written as

$$y = y_0 + u\Delta y_0 + \frac{u(u-1)}{2!}\Delta^2 y_0 + \cdots + \frac{u(u-1)\ldots(u-n+1)}{n!}\Delta^n y_0 .$$

This formula is named as forward interpolation formula because it contains tabulated values of the entries starting from y_0 since $\Delta y_0 = y_1 - y_0$, $\Delta^2 y_0 = y_2 - 2y_1 + y_0$ etc. For this reason this formula is used mainly for interpolating the value of y near the beginning of a set of tabular values lying within a close distance forward from y_0 and for extrapolating the value of y lying within a close distance backward from y_0.

3.3.4 Newton's Backward Interpolation Formula

Let $y_0, y_1, \ldots, y_n$ be the (n + 1) tabulated values of the function y = f(x) corresponding to the equidistant values $x_0, x_1, \ldots, x_n$ of argument *x* i.e., $x_i = x_0 + ih$ with interval length h, $x_0 < x_1 < \ldots < x_n$ and $f(x_i) = y_i$ for i = 0, 1, 2, ..., n. To find the value of *y* corresponding to a value of *x* between x_0 and x_n lying near x_n we use Newton's backward interpolation formula

$$y = y_n + v\Delta y_{n-1} + \frac{v(v+1)}{2!}\Delta^2 y_{n-2} + \ldots + \frac{v(v+1)\ldots(v+n-1)}{n!}\Delta^n y_0$$

where $v = \frac{x - x_n}{h}$

Proof. As (n + 1) points (x_i, y_i) are given for equidistant values $x_0, x_1, \ldots, x_n$ of argument *x* of interval length *h*, where $x_0 < x_1 < \ldots < x_n$ and $y_i = f(x_i)$, i = 0, 1, 2, ..., n, represent y = f(x) by a polynomial $y = \phi(x)$ of degree *n* (i.e. $\Delta^n y_0 \neq 0$) passing through (n + 1) given points,

$\phi(x) = b_0 + b_1(x - x_n) + b_2 (x - x_n)(x - x_{n-1}) + \ldots + b_n(x - x_n)(x - x_{n-1})\ldots(x - x_1)$ (15)

where $b_0, b_1, b_2 ..., b_n$ in (15) can be determined from

$y_n = \phi(x_n), y_{n-1} = \phi(x_{n-1}), ..., y_0 = \phi(x_0)$.

As $x_i = x_0 + ih$, $i = 0, 1, 2, ..., n$, $x_r - x_k = (x_0 + rh) - (x_0 + kh) = (r - k)\,h$ for $r, k = 0, 1, 2, ..., n$.

$y_n = \phi(x_n) = b_0$. [Substituting $x = x_n$ in (15)]

$y_{n-1} = \phi(x_{n-1}) = b_0 + b_1(-h)$ or $b_1 h = y_n - y_{n-1} = \Delta y_{n-1}$ or, $b_1 = \dfrac{\Delta y_{n-1}}{h}$. [Substituting $x = x_{n-1}$ in (15)]

Similarly, $y_{n-2} = \phi(x_{n-2}) = b_0 + b_1(-2h) + b_2(-2h)(-h)$

or $2!h^2 b_2 = y_{n-2} - y_n + 2\Delta y_{n-1} = y_{n-2} - y_n + 2(y_n - y_{n-1})$

$= y_n - 2y_{n-1} + y_{n-2} = \Delta^2 y_{n-2}$

so $b_2 = \dfrac{\Delta^n y_{n-2}}{2!h^2}$.

Proceeding this way and noting the symmetry at each stage the value of b_n is derived where

$$b_n = \frac{\Delta^n y_0}{2!h^n}.$$

Thus from (15)

$$y = y_n + (x - x_n)\frac{\Delta y_{n-1}}{h} + (x - x_n)(x - x_{n-1})\frac{\Delta^2 y_{n-2}}{2!h^2} + \ldots.$$

$$+ (x - x_n)(x - x_{n-1})\ldots(x - x_1)\frac{\Delta^n y_0}{n!h^n}$$

i.e., $$y = y_n + v\Delta y_{n-1} + \frac{v(v+1)}{2!}\Delta^2 y_{n-2} + \ldots + \frac{v(v+1)\ldots(v+n-1)}{n!}\Delta^n y_0,$$

where $$\frac{x - x_r}{h} = \frac{x - x_n}{h} + \frac{x_n - x_r}{h} = v + (n - r) \text{ for } r = 0, 1, 2, \ldots, n.$$

This is Newton's backward interpolation formula.

It is named as backward interpolation formula because it contains the tabulated values of entry y starting from y_n and moving backward to the left and none forward to y_n since $\Delta y_{n-1} = y_n - y_{n-1}$, $\Delta^2 y_{n-2} = y_n - 2y_{n-1} + y_{n-2}$ etc. For this reason this formula is used mainly for interpolating the value of y for given x near x_n in the interval (x_0, x_n) and for extrapolating the value of y corresponding to the value of x lying at a short distance forward from x_n.

The Newton's forward and backward formulae can be used to get y when the arguments are equidistant and the value of x lies near the beginning and the end of the ordered set of tabulated values of argument (x). To get a value of y for given value of argument (x) anywhere within or close to the two extreme values where the tabulated values of arguments are not necessarily equidistant when they are arranged in increasing order, we use Lagrange's interpolation formula.

3.3.5 Lagrange's Interpolation Formula

Suppose (x_0, y_0), (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) are $(n + 1)$ pairs of values where $y = f(x)$ and $x_0 < x_1 < ... < x_n$. Lagrange's interpolation formula uses a polynomial of degree n through these $(n + 1)$ points to interpolate the value of entry y for given x lying within or outside but near the extreme values of x. It is given by

$$y = \frac{(x-x_1)(x-x_2)\dots(x-x_n)}{(x_0-x_1)(x_0-x_2)\dots(x_0-x_n)}y_0 + \frac{(x-x_0)(x-x_2)(x-x_3)\dots(x-x_n)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)\dots(x_1-x_n)}y_1 + \dots.$$

$$+ \frac{(x-x_0)(x-x_1)\dots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\dots(x_n-x_{n-1})}y_n.$$

Proof : Through $(n + 1)$ given points (x_i, y_i), $i = 0, 1, 2, ..., n$ a polynomial of degree n, $y = \phi(x)$ can be assumed for the original function $y = f(x)$. where

$$\phi(x) = c_0(x - x_1)(x - x_2)...(x - x_n) + c_1(x - x_0)(x - x_2)(x - x_3)...(x - x_n) + ...$$
$$+ c_n(x - x_0)(x - x_1)...(x - x_{n-1}) \quad (16)$$

and $c_0, c_1, ..., c_n$ can be determined from $y_i = \phi(x_i)$, $i = 0, 1, 2, ..., n$.

When $x = x_0$, from (16), $y_0 = \phi(x_0) = c_0(x_0 - x_1)(x_0 - x_2)....(x_0 - x_n)$ i.e.,

$$c_0 = \frac{y_0}{(x_0-x_1)(x_0-x_2)\dots.(x_0-x_n)}.$$

When $x = x_1$, from (16), $y_1 = \phi(x_1) = c_1(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)....(x_1 - x_n)$ i.e.,

$$c_1 = \frac{y_1}{(x_1-x_0)(x_1-x_2)(x_1-x_3)\dots.(x_1-x_n)}.$$

Proceeding this way finally when $x = x_n$, from (16),

$y_n = \phi(x_n) = c_n(x_n - x_0)(x_n - x_1)....(x_n - x_{n-1})$ i.e.,

$$c_n = \frac{y_n}{(x_n-x_0)(x_n-x_1)\cdots(x_n-x_{n-1})}.$$

Then putting these values of $c_0, c_1, ..., c_n$ in (16), we get

$$y = \frac{(x-x_1)(x-x_2)\dots(x-x_n)}{(x_0-x_1)(x_0-x_2)\dots(x_0-x_n)}y_0 + \frac{(x-x_0)(x-x_2)(x-x_3)\dots(x-x_n)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)\dots(x_1-x_n)}y_1 + \dots.$$

$$+\frac{(x-x_0)(x-x_1)\ldots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\ldots(x_n-x_{n-1})}y_n.$$

This is Lagranges's interpolation formula.

This formula is used even when the arguments are not equidistant and when arguments lie anywhere within or close to two extreme values.

3.3.6 Central Difference Formulae

When entry y is to be interpolated for given value of argument x lying in the central part of a table, the values obtained by the central difference formulae will converge more rapidly to the actual value than by the other formulae discussed before. Thus central difference formula is more useful for interpolating near the middle of an ordered series of values of the argument. The two most important and useful central difference formulae are Stirling's and Bessel's interpolation formulae and their statements are given below without proof.

Stirling's Formula

When (2n + 1) pairs of values (x_{-n}, y_{-n}), $\left(x_{-(n-1)}, y_{-(n-1)}\right)$, ..., (x_{-1}, y_{-1}), (x_0, y_0), (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) of argument x and entry y are given for equidistant values of argument x having interval length h i.e., $x_i = x_0 + i\,h$, $x_{-i} = x_0 - i\,h$, $y_i = f(x_i)$ and $y_{-i} = f(x_{-i})$ for i = 0, 1, 2, ..., n, then y = f(x) is determined for given value of x in $x_0 - h < x < x_0 + h$ by Stirling's interpolation formula, which is

$$y = y_0 + u\frac{\Delta y_0 + \Delta y_{-1}}{2} + \frac{u^2}{21}\Delta^2 y_{-1} + \frac{u\left(u^2-1^2\right)}{3!}\cdot\frac{\Delta^3 y_{-1} + \Delta^3 y_{-2}}{2} + \frac{u^2(u^2-1^2)}{4!}\Delta^4 y_{-2}$$

$$+\ldots+\frac{u^2\left(u^2-1^2\right)\left(u^2-2^2\right)\ldots\left(u^2-(n-1)^2\right)}{(2n)!}\Delta^{2n}y_{-n}$$

where $\quad u = \dfrac{x-x_0}{h}.$

Bessel's Formula

When (2n + 2) pairs of values (x_{-n}, y_{-n}), $(x_{-(n-1)}, y_{-(n-1)})$ ...., (x_{-1}, y_{-1}), (x_0, y_0), (x_1, y_1), ..., (x_n, y_n), $(x_{(n+1)}, y_{(n+1)})$ of argument x and entry y are given for equidistant values of arguments x having interval length h i.e., $x_i = x_0 + ih$, $x_{-i} = x_0 - ih$, $y_i = f(x_i)$, $y_{-i} = f(x_{-i})$ for i = 0, 1, 2, ..., n and $x_{(n+1)} = x_0 + (n+1)h$, $y_{(n+1)} = f(x_{(n+1)})$ then y = f(x) is determined for given value of x in $x_0 - h < x < x_0 + h$ by Bessel's formula, which is

$$y = \frac{y_1 + y_0}{2} + v\Delta y_0 + \frac{\left(v^2 - \left(\frac{1}{2}\right)^2\right)}{2!} \cdot \frac{\Delta^2 y_0 + \Delta^2 y_{-1}}{2} + \frac{v\left(v^2 - \left(\frac{1}{2}\right)^2\right)}{3!} \Delta^3 y_{-1}$$

$$+ \frac{\left(v^2 - \frac{1}{4}\right)\left(v^2 - \frac{9}{4}\right)}{4!} \cdot \frac{\Delta^4 y_{-1} + \Delta^4 y_{-2}}{2} + \ldots + \frac{v\left(v^2 - \left(\frac{1}{2}\right)^2\right)\left(v^2 - \left(\frac{3}{2}\right)^2\right)\ldots\left(v^2 - \left(\frac{2n-1}{2}\right)^2\right)}{(2n+1)!} \Delta^{2n+1} y_{-n}$$

where $v = \frac{x - x_0}{h} - \frac{1}{2} = u - \frac{1}{2}$.

Central difference interpolation formulae are suitable for interpolating near the central part of an ordered tabulated set of values of argument x. More specifically Stirling's formula will give better result when $-0.25 \le u \le 0.25$ and Bessel's formula will give better result when $0.25 \le u \le 0.75$ i.e., $-0.25 \le v \le 0.25$. Stirling's formula will give more accurate result than Bessel's formula to interpolate near the beginning of the central interval and second formula will give more accurate result than the first to interpolate near the middle of the central interval.

3.4 Worked out examples

Example 1. Round off the following numbers correct to four significant figures and find the absolute error, relative error and percentage error in (i) 20478, (ii) 0.018355

Solution : (i) The number 20478 correct to four significant figures is 2048×10 since the digit after fourth significant figure is $8 > 5$ and so 1 is added to the fourth significant digit 7 and fifth place is replaced by zero. As the number is written correct to 4 significant figures it is written as 2048×10.

Then Absolute error = $|20478 - 2048 \times 10| = 2$,

Relative error = $\frac{2}{20478} = 9.7666 \times 10^{-5}$.

Percentage error = Relative error $\times$ 100 = 9.7666×10^{-3}

= .00977%

(ii) The number correct to 4 significant figures is 0.01836 since after the fourth significant figure the only digit remaining is 5 and the digit before the last digit is also 5, an odd number and so the last significant digit becomes the next even number i.e., 6.

So Absolute error = $|0.018355 - 0.01836| = 0.000005$.

Relative error $\frac{0.000005}{0.018355} = 2.724 \times 10^{-4}$,

Percentage error = Relative error × 100 = $2.724 \times 10^{-2} = 0.02724\%$.

Example 2. Find $\Delta^2\left(e^{bx}\right)$ where b is a constant and h is the increment of x.

Solution :

$$\Delta(e^{bx}) = e^{b(x+h)} - e^{bx} = e^{bx}\left(e^{bh} - 1\right)$$

$$\Delta^2\left(e^{bx}\right) = \Delta\left(\Delta e^{bx}\right) = \Delta\left[e^{bx}\left(e^{bh} - 1\right)\right]$$

$$= e^{b(x+h)}\left(e^{bh} - 1\right) - e^{bx}\left(e^{bh} - 1\right)$$

$$= e^{bx}\left(e^{bh} - 1\right)^2.$$

Example 3. (a) Estimate f(3) from the following table

x	:	2	3	4	5	6
f(x)	:	7	–	13	23	37

(b) If f(5) is also missing in the above table how do you estimate the missing values?

Solution : (a) Since only four values of f(x) are given assume f(x) to be a polynomial of degree 3 in *x*. So then 3rd and 4th order finite differences are constant and zero respectively.

Now $\Delta^4 f(x) = 0$. In particular $\Delta^4 f(2) = 0$. Consider $\Delta^4 f(2)$ because it contains all the given f(x)'s including missing value f(3), as $\Delta^4 f(2) = (E - 1)^4 f(2)$

$= (E^4 - 4E^3 + 6E^2 - 4E + 1)f(2) = f(6) - 4f(5) + 6f(4) - 4f(3) + f(2)$

Hence $\Delta^4 f(2) = 0$ gives $f(6) - 4f(5) + 6f(4) - 4f(3) + f(2) = 0$

or $37 - 4 \times 23 + 6 \times 13 - 4f(3) + 7 = 0$

or $37 - 92 + 78 + 7 = 4f(3)$

or $f(3) = \frac{122 - 92}{4} = \frac{30}{4} = 7.5$

So the estimated value of f(3) is 7.5.

(b) When f(3) and f(5) are missing 3 values of ordinates i.e., f(2), f(4) and f(6) would be given. Then f(x) can be assumed to be a polynomial of degree 2 in x. So then 2nd and 3rd order finite differences of f(x) are constant and zero respectively. i.e., $\Delta^3 f(x) = 0$.

In particular $\Delta^3 f(2) = 0$ and $\Delta^3 f(3) = 0$, since two unknown values will appear in both the equations.

Now $\Delta^3 f(2) = (E-1)^3 f(2) = (E^3 - 3E^2 + 3E - 1)f(2) = f(5) - 3f(4) + 3f(3) - f(2)$.

Similarly $\Delta^3 f(3) = f(6) - 3f(5) + 3f(4) - f(3)$. So the two equations are

$f(5) - 3f(4) + 3f(3) - f(2) = 0$ and $f(6) - 3f(5) + 3f(4) - f(3) = 0$,

or $f(5) - 3 \times 13 + 3f(3) - 7 = 0$ and $37 - 3f(5) + 3 \times 13 - f(3) = 0$.

or $f(5) + 3f(3) = 46$ and $3f(5) + f(3) = 76$.

They are solved to give $f(3) = \frac{31}{4}$ and $f(5) = \frac{91}{4}$.

i.e. the estimated values of f(3) and f(5) are 7.75 and 22.75 respectively.

Example 4. Use the method of separation of symbols to prove the following identities :

(i) $\Delta u_x - \frac{1}{2}\Delta^2 u_x + \frac{1.3}{2.4}\Delta^3 u_x - \frac{1.3.5}{2.4.6}\Delta^4 u_x + \ldots = u_{x+\frac{1}{2}} - u_{x-\frac{1}{2}}$

(ii) $u_{2n} - 2\binom{n}{1}u_{2n-1} + 4\binom{n}{2}u_{2n-2} - \ldots + (-2)^n u_n = (-1)^n (c - 2an)$.

$u_x = ax^2 + bx + c$ with interval of differencing unity and a, b, c are constants

Solution : (i) L.H.S. $\left[\Delta - \frac{1}{2}\Delta^2 + \frac{1.3}{2.4}\Delta^3 - \frac{1.3.5}{2.4.6}\Delta^4 + \ldots.\right]u_x$

$$= \Delta\left[1 + \left(-\frac{1}{2}\right)\Delta + \frac{\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)}{2!}\Delta^2 + \frac{\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)\left(-\frac{5}{2}\right)}{3!}\Delta^3 + \ldots\right]u_x$$

$$= \Delta(1+\Delta)^{-\frac{1}{2}} u_x = \Delta E^{-\frac{1}{2}} u_x.$$

R.H.S. $= E^{\frac{1}{2}}u_x - E^{-\frac{1}{2}}u_x = \left(E^{\frac{1}{2}} - E^{-\frac{1}{2}}\right)u_x = (E-1)E^{-\frac{1}{2}}u_x = \Delta E^{-\frac{1}{2}}u_x$.

So L.H.S. = R.H.S.

(ii) L.H.S. $= u_{2n} - 2\binom{n}{1}u_{2n-1} + 4\binom{n}{2}u_{2n-2} - \ldots. + (-2)^n u_n$

$$= E^n u_n + (-2)\binom{n}{1}E^{n-1}u_n + (-2)^2\binom{n}{2}E^{n-2}u_n + \ldots. + (-2)^n u_n$$

$$\left[E^n + \binom{n}{1}E^{n-1}(-2) + \binom{n}{2}E^{n-2}(-2)^2 + \ldots. + \binom{n}{n}(-2)^n\right]u_n$$

$= (E - 2)^n u_n$

$= (\Delta - 1)^n u_n$ since $E \equiv 1 + \Delta$

$= (-1)^n (1 - \Delta)^n u_n$

$$= (-1)^n \left[1 - \binom{n}{1}\Delta + \binom{n}{2}\Delta^2 - \binom{n}{3}\Delta^3 + \ldots + (-\Delta)^n\right] u_n$$

$$= (-1)^n \left[u_n - \binom{n}{1}\Delta u_n + \binom{n}{2}\Delta^2 u_n - \ldots + (-1)^n \Delta^n u_n\right].$$

Now $u_n = an^2 + bn + c$.

So $\Delta u_n = u_{n+1} - u_n = \left[\left(a(n+1)^2 + b(n+1) + c\right) - \left(an^2 + bn + c\right)\right]$

i.e., $\Delta u_n = 2an + a + b$, $\Delta^2 u_n = \Delta u_{n+1} - \Delta u_n = [2a(n + 1) + (a + b)] - [2an + a + b]$

i.e., $\Delta^2 u_n = 2a$ and $\Delta^r u_n = 0$ for $r > 2$.

So L.H.S. $= (-1)^n \left[an^2 + bn + c - n(2an + a + b) + \frac{n(n-1)}{2} \cdot 2a\right]$

$= (-1)^n [an^2 - 2an^2 + an^2 + bn - an - bn - an + c]$

$= (-1)^n [c - 2an]$ = R.H.S.

$\therefore$ L.H.S. = R.H.S.

Example 5. The values of f(x) given below are :

x	:	0	1	2	3	4	5
f(x)	:	1	5	31	121	341	781

Find the values of f(0.5), f(4.5), f(2.5) by using Newton's forward, Newton's backward and Lagrange's formulae respectively. Also determine the value of f(2.2) and f(2.7) by using Stirling's and Bessel's interpolation formulae.

Solution : The finite difference table is

x	f(x)=y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	$\Delta^5 y$
0	1					
		4				
1	5		22			
		26		42		
2	31		64		24	
		90		66		0
3	121		130		24	
		220		90		
4	341		220			
		440				
5	781					

From finite difference table 4th order finite differences $\Delta^4 f(0)$, $\Delta^4 f(1)$ are both 24. So 4th order difference is assumed to be constant i.e., all higher order differences are zero.

Newton's forward interpolation formula :

$$y = y_0 + u\Delta y_0 + \frac{u(u-1)}{2!}\Delta^2 y_0 + \frac{u(u-1)(u-2)}{3!}\Delta^3 y_0 + \frac{u(u^2-1)(u-2)(u-3)}{4!}\Delta^4 y_0$$

where $u = \frac{x - x_0}{h} = \frac{0.5 - 0}{1}$ where $x = 0.5$, $x_0 = 0$. $y_0 = 1$, $\Delta y_0 = 4$,

$\Delta^2 y_0 = 22$, $\Delta^3 y_0 = 42$, $\Delta^4 y_0 = 24$, $\Delta^r y_0 = 0$ for $r \geq 5$. Here $y_i = f(x_i)$, $i = 0, 1, ..., 5$.

$$\therefore\ f(0.5) = 1 + \frac{1}{2} \times 4 + \frac{\frac{1}{2}\left(-\frac{1}{2}\right)}{2!} \times 22 + \frac{\frac{1}{2}\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)}{3!} \times 42 + \frac{\frac{1}{2}\left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)\left(-\frac{5}{2}\right)}{4!} \times 24$$

$$= 1 + 2 - \frac{22}{8} + \frac{21}{8} - \frac{15}{16} = 3 - \frac{1}{8} - \frac{15}{16} = 3 - \frac{17}{16} = \frac{31}{16} = 1.9375$$

So $f(0.5) = 1.9375$.

Newton's backward interpolation formula :

$$y = y_n + v\Delta y_{n-1} + \frac{v(v+1)}{2!}\Delta^2 y_{n-2} + \frac{v(v+1)(v+2)}{3!}\Delta^3 y_{n-3} + \frac{v(v+1)(v+2)(v+3)}{4!}\Delta^4 y_{n-4}$$

$$y = y_5 + v\Delta y_4 + \frac{v(v+1)}{2!}\Delta^2 y_3 + \frac{v(v+1)(v+2)}{3!}\Delta^3 y_2 + \frac{v(v+1)(v+2)(v+3)}{4!}\Delta^4 y_1$$

where $v = \frac{x - x_n}{h} = \frac{x - x_5}{1} = 4.5 - 5 = -0.5 = -\frac{1}{2}$, $y_5 = f(5) = 781$

$\Delta y_4 = 440$, $\Delta^3 y_3 = 220$, $\Delta^3 y_2 = 90$, $\Delta^4 y_1 = 24$, $\Delta^5 y_0 = 0$ and assume $\Delta^4 y_1$ is constant and it is 24. Here $f(x_i) = y_i$, $i = 0, 1, ..., 5$

So $$y = 781 + \left(-\frac{1}{2}\right) \times 440 + \frac{\left(-\frac{1}{2}\right)\left(\frac{1}{2}\right)}{2!} \times 220 + \frac{\left(-\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{3}{2}\right)}{3!} \times 90$$

$$+ \frac{\left(-\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{3}{2}\right)\left(\frac{5}{2}\right)}{4!} \times 24$$

$$= 781 - 220 - \frac{55}{2} - \frac{45}{8} - \frac{15}{16} = 561 - 27.5 - 5.625 - 0.9375$$

$$= 533.5 - 6.5625 = 526.9375$$

So $f(4.5) = 526.9375$.

Lagrange's interpolation formula :

$$y = \frac{(x-x_1)(x-x_2)\ldots(x-x_5)}{(x_0-x_1)(x_0-x_2)\ldots(x-x_5)}y_0 + \frac{(x-x_0)(x-x_2)(x-x_3)(x-x_4)(x-x_5)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)(x_1-x_4)(x_1-x_5)}y_1$$

$$+\ldots+\frac{(x-x_0)(x-x_1)\ldots(x-x_4)}{(x_5-x_0)(x_5-x_1)\ldots(x_5-x_4)}y_5$$

where $x_0 = 0$, $y_0 = 1$, $x_1 = 1$, $y_1 = 5$, $x_2 = 2$, $y_2 = 31$, $x_3 = 3$, $y_3 = 121$, $x_4 = 4$, $y_4 = 341$, $x_5 = 5$ $y_5 = 781$ and $y_i = f(x_i)$, i = 0, 1, 2, ..., 5. x = 2.5.

$$y = \frac{(1.5)(0.5)(-0.5)(-1.5)(-2.5)}{(-)(-2)(-3)(-4)(-5)}\times 1 + \frac{(2.5)(.5)(-.5)(-1.5)(-2.5)}{(1)(-1)(-2)(-3)(-4)}\times 5$$

$$+\frac{(2.5)(1.5)(-.5)(-1.5)(-2.5)}{(2)(1)(-1)(-2)(-3)}\times 31 + \frac{(2.5)(1.5)(.5)(-1.5)(-2.5)}{(3)(2)(1)(-1)(-2)}\times 121$$

$$+\frac{(2.5)(1.5)(.5)(-.5)(-2.5)}{(4)(3)(2)(1)(-1)}\times 341 + \frac{(2.5)(1.5)(.5)(-.5)(-1.5)}{(5)(4)(3)(2)(1)}\times 781$$

$$= \frac{1.40625}{120} - \frac{2.34375}{24}\times 5 + \frac{7.03125}{12}\times 31 + \frac{7.03125}{12}\times 121$$

$$-\frac{2.34375}{24}\times 341 + \frac{1.40625}{120}\times 781$$

$$= 0.01172 - 0.48828 + 18.16406 + 70.89844 - 33.30078 + 9.15234$$

$$= 98.22656 - 33.78906 = 64.4375$$

So f(2.5) = 64.4375.

Stirling's interpolation formula :

$$y = y_0 + u\cdot\frac{\Delta y_0 + \Delta y_{-1}}{2} + \frac{u^2}{2!}\cdot\Delta^2 y_{-1} + \frac{u(u^2-1)}{3!}\cdot\frac{\Delta^3 y_{-1} + \Delta^3 y_{-2}}{2} + \frac{u^2(u^2-1)}{4!}\Delta^4 y_{-2}$$

where $u = \frac{x-x_0}{h} = \frac{2.2-2}{1} = 0.2$, $x_0 = 2$, $y_0 = 31$, $x_{-1} = 1$, $y_{-1} = 5$, $x_{-2} = 0$, $y_{-2} = 1$, $x_1 = 3$, $y_1 = 121$, $x_2 = 4$, $y_2 = 341$, $x_3 = 5$, $y_3 = 781$,

$\Delta y_0 = 90$, $\Delta y_{-1} = 26$, $\Delta^2 y_{-1} = 64$, $\Delta^3 y_{-1} = 66$, $\Delta^3 y_{-2} = 42$, $\Delta^4 y_{-2} = 24$, $\Delta^r{}_y = 0$ for $r \geq 5$.

So $$y = 31 + 0.2\times\frac{90+26}{2} + \frac{(0.2)^2}{2}\times 64 + \frac{0.2(0.04-1)}{6}\times\frac{66+42}{2} + \frac{0.04(0.04-1)}{24}\times 24$$

$$= 31 + 0.2 \times 58 + .02 \times 64 - .032 \times 54 - .04 \times .96$$

$$= 31 + 11.6 + 1.28 - 1.728 - .0384$$

$$= 43.88 - 1.7664 = 42.1136$$

So f(2.2) = 42.1136.

Bassel's interpolation formula :

$$y = \frac{y_1 + y_0}{2} + v\Delta y_0 + \frac{\left(v^2 - \frac{1}{4}\right)}{2!} \cdot \frac{\Delta^2 y_0 + \Delta^2 y_{-1}}{2} + \frac{v\left(v^2 - \frac{1}{4}\right)}{3!}\Delta^3 y_{-1}$$

$$+ \frac{\left(v^2 - \frac{1}{4}\right)\left(v^2 - \frac{9}{4}\right)}{4!} \cdot \frac{\Delta^4 y_{-1} + \Delta^4 y_{-2}}{2}$$

where $v = \frac{x - x_n}{h} - \frac{1}{2} = \frac{2.7 - 2}{1} - \frac{1}{2} = .2,\ x_0 = 2,\ y_0 = 31,\ x_{-1} = 1,\ y_{-1} = 5,\ x_{-2} = 0,\ y_{-2} = 1,$ $x_1 = 3,\ y_1 = 121,\ x_2 = 4,\ y_2 = 341,\ x_3 = 5,\ y_3 = 781,\ \Delta y_0 = 90,\ \Delta^2 y_0 = 130,\ \Delta^2 y_{-1} = 64,\ \Delta^3 y_{-1} = 66,\ \Delta^4 y_{-1} = 24,\ \Delta^4 y_{-2} = 24,\ \Delta^r y = 0$ for $r \geq 5$.

So $y = \frac{121 + 31}{2} + 0.2 \times 90 + \frac{(0.04 - 0.25)}{2} \times \frac{130 + 64}{2} + \frac{0.2(0.04 - 0.25)}{6} \times 66$

$$+ \frac{(0.04 - 0.25)(0.04 - 2.25)}{24} \times \frac{24 + 24}{2}$$

$= 76 + 18 - 0.105 \times 97 - .042 \times 11 + 0.21 \times 2.21$

$= 94 - 10.185 - .462 + .4641 = 94.4641 - 10.647 = 83.8171$

So $f(2.7) = 83.8171$.

Example 6. If $f(0) = 2$, $f(2) = 8$ and $f(3) = 20$, find $f(x)$ and $f(4)$.

Solution : As 3 ordinates are given so we can get second order difference which is considered to be constant. Then $f(x)$ is a polynomial atmost of degree 2. Let $f(x)$ be a polynomial of degree 2 and it is

$$f(x) = a + bx + cx\ (x - 2)$$

Then $2 = f(0) = a$,

$8 = f(2) = a + 2b$

$20 = f(3) = a + 3b + 3c$

So $a = 2,\ b = \frac{8 - a}{2} = \frac{8 - 2}{2} = 3, c = \frac{20 - a - 3b}{3} = \frac{20 - 2 - 9}{3} = 3$

Thus $f(x) = 2 + 3x + 3x\ (x - 2) = 2 - 3x + 3x^2$

and $f(4) = 2 - 3 \times 4 + 3 \times 4^2 = 2 - 12 + 48 = 38$.

3.5 Summary

This chapter includes the significant figures, the rounding off numbers and the errors involved there, interpolation and extrapolation, Weirstrass theorem, Δ and E operators and their uses, rules of finite differences, statements and proof of Newton's

forward and backward interpolation formulae, when arguments are equidistant and of Lagrange's interpolation formula when the arguments may or may not be equidistant, their uses, statements of central difference formulae : Stirling's and Bessel's interpolation formulae and worked out examples.

3.6 Exercises

1. Round off the following numbers correct to four significant figures and find absolute error, relative error and percentage error in each case.
   (i) 3.28236, (ii) 0.0012342, (iii) 16.355, (iv) 16.345, (v) 23455, (vi) 93278, (vii) 363045.
2. Find (i) $\Delta 2^x$, (ii) $\Delta \log x$, (iii) $\Delta^2 (3e^{2x})$, (iv) $\Delta^2 x\,(x-1)\,(x-2)$ when the interval of differencing is unity.
3. Evaluate $\left(\frac{\Delta}{E}\right)^2 e^x$ and $\frac{\Delta^2 e^x}{E^2 e^x}$ where interval of differencing is unity.
4. If $\Delta x = 1$ and $f(x) = x\,(x+1)\,(x+2)\,(x+3)$ show that $\Delta^2 f(x) = 12\,(x+2)\,(x+3)$

5(a) If f(3) = 2, f(4) = 5, f(5) = 16, determine f(x) as a second degree polynomial in x.

(b) Find y as a second degree polynomial in *x* passing through 3 points (0, 0), (1, 3), (2, 8) of rectangular cartesian coordinates.

(c) A third degree polynomial passes through the points (0, –1), (1, 1), (2, 1) and (3, –2). Find the polynomial.

6(a) Find f(5) from the following data f(3) = 8.71940, f(7) = 9.08914, f(9) = 9.19433

(b) Estimate the missing terms in the following table on making suitable assumptions (to be stated by you) on the function u_x

x :	2.0	2.1	2.2	2.3	2.4	2.5	2.6
u_x :	0.135	–	0.111	0.100	–	0.082	0.074

7. Construct a finite difference table and express y as a function of x.

x :	1	2	3	4
y :	4	15	40	85

   (You may use Newton's forward interpolation formula).

8. Use the method of separation of symbols to prove the following identities :
   (a) $u_0 + \binom{n}{1}u_1 x + \binom{n}{2}u_2 x^2 + \binom{n}{3}u_3 x^3 + \ldots.$
   $= (1+x)^n u_0 + \binom{n}{1}(1+x)^{n-1} x\Delta u_0 + \binom{n}{2}(1+x)^{n-2} x^2 \Delta^2 u_0 + \ldots.$
   (b) $u_1 = u_0 + \Delta u_{-1} + \Delta^2 u_{-2} + \Delta^3 u_{-3} + \ldots..$
   (c) $\Delta^n u_{x-n} = u_x - \binom{n}{1}u_{x-1} + \binom{n}{2}u_{x-2} - \binom{n}{3}u_{x-3} + \ldots.$

9. Using a suitable interpolation formula, find f(3.5) and f(2.5) from the following table :

x :	3	4	5	6
f(x) :	4	13	26	43

10. Using suitable interpolation formula find f(24) and f(26) from the following table :

x :	5	10	15	20	25
f(x) :	1.0	1.6	3.8	8.2	15.4

11. The following table gives the normal weight of a baby during the first 5 months of life :

Age (in month) :	1	2	3	4	5
Weight (in lbs) :	5	7	8	10	12

(a) Estimate the weight of the baby at the age of 2.5 months by Lagrange's method.

(b) Estimate the weights of the babies at ages 3.2 months and 3.5 months by Stirling's and Bessel's interpolation formulae.

12. What do you mean by the term 'interpolation'? Distinguish between forward, backward and central interpolation formulae.

13. Define the operators Δ and E and establish the relation between them.

14. Obtain, by using a suitable interpolation formula, the missing observation in the following table :

Year	1881	1891	1901	1911	1921	1931
Population (million)	3.9	5.3	?	9.6	12.9	17.1

15. You are given the following information :

x :	654	658	659	661
y :	2.8156	2.8182	2.8189	2.8202

By applying Lagrange's formula interpolate the value of y when x = 656.

3.7 Suggested Readings

1. Scarborough, J. B. *Numerical Mathematical Analysis*, Oxford University Press 1958 and Oxford Book Co. (Indian Ed.) 1946.
2. Hilderbrand, F. B. *Introduction to Numerical Analysis,* McGraw Hill, 1956 and Tata McGraw Hill, 1974.
3. Goon, A. M.; Gupta, M. K. and Dasgupta, B. *The Fundamentals of Statistics* Vol. 1, The World Press Private Limited, 2002, Kolkata.
4. Freeman, H. *Finite Differences for Actuarial Students,* Cambridge University Press, 1962.

Unit 4 ❑ Theory of Attributes

Structure

4.0 Objectives

In many investigations it is required to collect data for a set of individuals on more than one qualitative characters simultaneously. For example, blindness and deafness can be investigated simultaneously in a community. Then we should observe whether there is any relationship between the two qualitative characters or attributes. Theory of attributes need some different statistical treatment from that of the variables. The theory of sampling of the attributes is, in many respect, simpler than that of variables and in this chapter we shall confine ourselves to it.

4.1 Introduction

Data of qualitative characters are called attributes. For example, religion, literacy, intelligence etc. are attributes, which cannot be measured numerically. Suppose the population is divided into two classes according to the presence and absence of a single attribute. The positive class, which denotes the presence of an attribute, is generally written in capital roman letter like A, B, C, D etc. and negative class, which denotes the absence of the attribute, is generally written in Greek letters like α, β, γ, δ etc. If the attribute, blindness is studied, the population will be divided into two main classes viz, blind (in whom blindness is present) and unblind (in whom blindness is absent). If two attributes blindness A and deafness B are considered together then the population will be divided into a number of classes, which are : blinds A, unblinds α, deafs B, non-deafs β, blinds and deafs AB, blinds and non-deafs Aβ, unblinds and

deafs αB, unblinds and non-deafs $\alpha\beta$, total population. When two attributes A and B are studied simultaneously the total number of classes are $3^2 = 9$ and they are denoted as A, B, α, β, AB, $A\beta$, αB, $\alpha\beta$ and the class including all the N observations. Here A, B, AB and the class having total frequency N are positive classes, α, β, $\alpha\beta$ are negative classes and αB, $A\beta$ are pairs of contrary classes. The corresponding class frequencies are expressed by f_A, f_α, f_B, f_β, f_{AB}, $f_{A\beta}$, $f_{\alpha B,}$ $f_{\alpha\beta}$, N. Thus when two attributes have two categories each say A, α and B, β according to the presence or absence of the attributes A and B then the data can be shown in the form of a bivariate frequency distribution table as shown below. This two way table for attributes A and B is called 2×2 contingency table.

2×2 contingency table of blinds and deafs.

		Blinds		
		A = blinds	a = unblinds	Total
Deafs	B = deafs	f_{AB}	$f_{\alpha B}$	f_B
	b = nondeafs	$f_{A\beta}$	$f_{\alpha\beta}$	f_β
	Total	f_A	f_α	N

So in the above 2×2 contingency table

$f_{AB} + f_{\alpha B} = f_B$, $f_{A\beta} + f_{\alpha\beta} = f_\beta$, $f_{AB} + f_{A\beta} = f_A$, $f_{\alpha B} + f_{\alpha\beta} = f_\alpha$,

$f_A + f_\alpha = N$, $f_B + f_\beta = N$.

If two attributes A and B have *k* and *l* categories A_1, A_2, ..., A_k and B_1, B_2, ..., B_1 respectively then the data can be shown in the form of a frequency table with *k* rows and *l* columns as in a bivariate frequency table above. This two-way table for attributes A and B is called k × l contingency table.

4.2 Types of association

The attributes A and B are said to have no association i.e. they are independent when the presence or absence of one does not affect the presence of the other attribute i.e., when B is independent of A i.e. $\frac{f_{AB}}{f_A} = \frac{f_{\alpha B}}{f_\alpha} = \frac{f_{AB} + f_{\alpha B}}{f_A + f_\alpha} = \frac{f_B}{N}$ i.e., $f_{AB} = \frac{f_A \times f_B}{N}, f_{\alpha B} = \frac{f_\alpha \times f_B}{N}$, and when A is independent of B i.e., $\frac{f_{AB}}{f_B} = \frac{f_{A\beta}}{f_\beta} =$

$\frac{f_{AB}+f_{A\beta}}{f_B+f_\beta}=\frac{f_A}{N}$ i.e., $f_{AB}=\frac{f_A\times f_B}{N}, f_{A\beta}=\frac{f_A\times f_\beta}{N}$. Also $f_{\alpha\beta}=\frac{f_\alpha\times f\beta}{N}$ because $f_{\alpha\beta}=f_\alpha-f_{\alpha B}=f_\alpha-\frac{f_\alpha\times f_B}{N}=\frac{f_\alpha(N-f_B)}{N}=\frac{f_\alpha\times f_\beta}{N}$.

If $f_{AB}>\frac{f_A\times f_B}{N}$ then the attributes A and B are said to be positively associated or simply associated.

If $f_{AB}<\frac{f_A\times f_B}{N}$ then the attributes A and B are said to be negatively associated or disassociated.

Thus we can say if $f_{AB}=\frac{f_A\times f_B}{N}$, the attributes A and B are independent but if $f_{AB}\neq\frac{f_A\times f_B}{N}$ they are associated.

This frequency method can determine the nature of association i.e., whether the attributes are positively associated or negatively associated or independent but not the degree of association of the attributes. Yule's coefficient of association given below determines not only the nature of association but also degree of association.

We now define perfect association in two alternative ways : (i) complete association, (ii) absolute association (either positive or negative).

Complete association (positive) between attributes A and B occur if either all A's are B's or all B's are A's i.e., if either $f_{A\beta}=0$ or $f_{\alpha B}=0$. On the other hand, complete negative association between attributes A and B occur if either no A's are B's or no α's and β's i.e. if either $f_{AB}=0$ or $f_{\alpha\beta}=0$.

Absolute positive association between two attributes A and B occur if all A's are B's and all B's are A's i.e., if both $f_{A\beta}=0$ and $f_{\alpha B}=0$. On the other hand, absolute negative association between attributes A and B occur if no A's are B's and no α's are β's i.e., if $f_{AB}=0$ and $f_{\alpha\beta}=0$.

4.3 Measures of Association for the 2 × 2 Contingency Table

A measure of association between two attributes A and B should satisfy the following conditions :

(i) The measure should be independent of total frequency, N.

(ii) The measure should have zero value in case of independence, negative value in case of negative association and positive value in case of positive association.

(iii) The measure should vary between two definite limits –1 and +1. Its values are +1 when there is perfect positive association between the attributes and –1 when there is perfect negative association between them.

In case of 2 × 2 contingency table there are three important measures :

1. Yule's coefficient of association,
2. Yule's coefficient of colligation,
3. Another measure of association.

We describe these measures below one by one.

1. Yule's coefficient of association :

Yule's method is the most popular method of measuring association between two attributes A and B.

$$Q_{AB} = \frac{f_{AB}f_{\alpha\beta} - f_{A\beta}f_{\alpha B}}{f_{AB}f_{\alpha\beta} + f_{A\beta}f_{\alpha B}}$$

This Yule's coefficient of association between two attributes A and B i.e., Q_{AB} lies between –1 and +1.

When $Q_{AB} = +1$ there is perfect positive association between A and B and vice versa. When $Q_{AB} = -1$ there is perfect negative association between A and B and vice versa. When $Q_{AB} = 0$ then attributes A and B are said to be independent and vice versa.

2. Yule's coefficient of colligation :

Another measure of association developed by Yule is the coefficient of colligation and is given by

$$Y_{AB} = \frac{\sqrt{f_{AB}f_{\alpha\beta}} - \sqrt{f_{A\beta}f_{\alpha B}}}{\sqrt{f_{AB}f_{\alpha\beta}} + \sqrt{f_{A\beta}f_{\alpha B}}}.$$

When $Y_{AB} = 0$ the attributes A and B are independent and vice versa. When $Y_{AB} = +1$ there is perfect positive association between two attributes A and B where $f_{A\beta} = 0 = f_{\alpha B}$ and vice versa.

When $Y_{AB} = -1$ there is perfect negative association between A and B where $f_{AB} = f_{\alpha\beta} = 0$ and vice versa.

Relation between Q_{AB} and Y_{AB} :

$$1+Y_{AB}^2 = 1+\frac{f_{AB}f_{\alpha\beta}+f_{A\beta}f_{\alpha B}-2\sqrt{f_{AB}f_{\alpha\beta}f_{A\beta}f_{\alpha B}}}{f_{AB}f_{\alpha\beta}+f_{A\beta}f_{\alpha B}+2\sqrt{f_{AB}f_{\alpha\beta}f_{A\beta}f_{\alpha B}}}$$

$$= \frac{2\left(f_{AB}f_{\alpha\beta}+f_{A\beta}f_{\alpha B}\right)}{\left(\sqrt{f_{AB}f_{\alpha\beta}}+\sqrt{f_{A\beta}f_{\alpha B}}\right)^2}$$

So $$\frac{2Y_{AB}}{1+Y_{AB}^2} = 2\left(\frac{\sqrt{f_{AB}f_{\alpha\beta}}-\sqrt{f_{A\beta}f_{\alpha B}}}{\sqrt{f_{AB}f_{\alpha\beta}}+\sqrt{f_{A\beta}f_{\alpha B}}}\right) \Bigg/ \frac{2\left(f_{AB}f_{\alpha\beta}+f_{A\beta}f_{\alpha B}\right)}{\left(\sqrt{f_{AB}f_{\alpha\beta}}+\sqrt{f_{A\beta}f_{\alpha B}}\right)^2}$$

$$= \frac{f_{AB}f_{\alpha\beta}-f_{A\beta}f_{\alpha B}}{f_{AB}f_{\alpha\beta}+f_{A\beta}f_{\alpha B}} = Q_{AB}.$$

3. Another measure of association :

This measure of association between attributes A and B is

$$V_{AB} = \frac{f_{AB}f_{\alpha\beta}-f_{A\beta}f_{\alpha B}}{\sqrt{f_A f_\alpha f_B f_\beta}}$$

It is also called coefficient of association.

If $V_{AB} = 0$, the attributes A and B are independent and vice versa.

If $V_{AB} = +1$, the attributes A and B have absolute positive association, i.e. $f_{A\beta} = f_{\alpha B} = 0$ and vice versa.

If $V_{AB} = -1$, the attributes A and B have absolute negative association i.e. $f_{AB} = f_{\alpha\beta} = 0$ and vice versa.

So, if $V_{AB} = \pm 1$, the attributes A and B have absolute association between them and vice versa.

4.4 Worked out examples

Example 1. From the following class frequencies find out whether the data are consistent or not.

$f_{AB} = 75$, $f_A = 65$, $f_B = 140$, $N = 200$.

Solution : Putting all the given figures in a 2×2 contigency table, we get

	A	α	Total
B	75		140
β			
Total	65		200

$f_{\alpha B} = f_B - f_{AB} = 140 - 75 = 65$, $f_{A\beta} = f_A - f_{AB} = 65 - 75 = -10$,

$f_\beta = N - f_B = 200 - 140 = 60$, $f_\alpha = N - f_A = 200 - 65 = 135$,

$f_{\alpha\beta} = f_\alpha - f_{\alpha B} = 135 - 65 = 70$

Class frequency $f_{A\beta} = -10$ i.e. negative value. This is not possible. So the data are inconsistent.

Example 2. On the basis of following class frequencies calculate Yule's coefficient of association, Yule's coefficient of colligation and coefficient of absolute association. Determine whether two attributes A and B are independent or not. If dependent which type of association holds between the attributes?

$(AB) = 80$, $(A\beta) = 30$, $(\alpha B) = 120$, $(\alpha\beta) = 20$.

Solution : Class frequencies $f_{AB} = 80$, $f_{A\beta} = 30$, $f_{\alpha B} = 120$, $f_{\alpha\beta} = 20$ are tabulated below

	A	α	Total
B	80	120	200
β	30	20	50
Total	110	140	250

Yule's coefficient of association, $Q_{AB} = \dfrac{f_{AB}f_{\alpha\beta} - f_{A\beta}f_{\alpha B}}{f_{AB}f_{\alpha\beta} + f_{A\beta}f_{\alpha B}}$

$$= \frac{80\times 20 - 30\times 120}{80\times 20 + 30\times 120} = \frac{1600 - 3600}{1600 + 3600} = -\frac{2000}{5200} = -\frac{5}{13} = -\cdot 3846$$

Yule's coefficient of colligation, $Y_{AB} = \dfrac{\sqrt{f_{AB}f_{\alpha\beta}} - \sqrt{f_{A\beta}f_{\alpha B}}}{\sqrt{f_{AB}f_{\alpha\beta}} + \sqrt{f_{A\beta}f_{\alpha B}}}$

$$= \frac{\sqrt{80\times 20} - \sqrt{30\times 120}}{\sqrt{80\times 20} + \sqrt{30\times 120}} = \frac{40 - 60}{40 + 60} = -\frac{20}{100} = -0.2$$

Coefficient of association, $V_{AB} = \dfrac{f_{AB}f_{\alpha\beta} - f_{A\beta}f_{\alpha B}}{\sqrt{f_A f_\alpha f_B f_\beta}}$

$$= \frac{80\times 20 - 30\times 120}{\sqrt{110\times 140 + 200\times 50}} = \frac{1600 - 3600}{1000\sqrt{154}} = -\frac{2000}{1000\sqrt{154}} = -\frac{2}{12\cdot 4097}$$

$= -\ 0.1612.$

In all cases we see the attributes A and B are not independent since $Q_{AB} \neq 0$, $Y_{AB} \neq 0$, $V_{AB} \neq 0$.

The different coefficients show substantial degree of association and attributes are negatively associated since all the coefficients are negative.

4.5 Summary

This chapter consists of notations used for the frequencies in simultaneous study of two attributes, types of association between the attributes, measures of association for the 2 × 2 contingency table, three important measures of association between the attributes and relation between the first two measures and worked out examples.

4.6 Exercises

1. Given n, f_A, f_B and f_{AB}, how would you find the other cell frequencies and marginal frequencies of a 2 × 2 contingency table?
2. Show that the number of individuals who have attribute A but not B or have attribute B but not A is $f_A + f_B - 2f_{AB}$.
3. For the case of two attributes, define independence and association (positive and negative). What are the different measures of association and what are their properties?
4. Establish the relation between Yule's coefficient of association (Q) and Yule's coefficient of colligation (Y) in 2 × 2 case.

   Hence or otherwise show that the coefficient of association is greater in absolute value than the coefficient of colligation except when both are zero or unity in absolute value.
5. Given n = 2500, f_A = 420, f_{AB} = 85 and f_n = 670, prepare a 2 × 2 contingency table and compute Yule's coefficient of association and interpret the result.
6. Given the following information : f_A = 490, f_{AB} = 294, f_α = 570, $f_{\alpha B}$ = 380, determine whether attributes A and B are positively associated, negatively associated or independent.

7. Find whether attributes A and B are independent, positively associated or negatively associated in the cases

   (i) N = 1000, f_A = 470, f_B = 620, f_{AB} = 320

   (ii) f_A = 490, f_{AB} = 294, f_α = 570, $f_{\alpha B}$ = 380.

8. Investigate the association between darkness of eye-colour in father and son from the following frequencies :

   Father with dark eyes and sons with dark eyes : 50

   Father with dark eyes and sons with out dark eyes : 79

   Father with out dark eyes and sons with dark eyes : 89

   Father with out dark eyes and sons with out dark eyes : 782

   Also tabulate the frequencies that would have been observed in case of no heredity and compare with the values above.

9. The following table shows the result of inoculation against cholera.

	Not attacked	Attacked
Inoculated	431	5
Non-inoculated	291	9

   Examine the effect of inoculation in controlling susceptibility to cholera.

10. Given the following in formation : f_A = 490, f_{AB} = 294, f_α = 570, $f_{\alpha\beta}$ = 380, establish the nature of association between the attributes A and B.

11. What do you mean by complete association and absolute association?

12. What is Yule's coefficient of association? Show that Yule's coefficient of association is +1 in the case of perfect positive association and it is –1 in the case of perfect negative association between two attributes.

13. In an experiment on immunization of cattle from tuberculosis the following results were obtained :

	Affected	Unaffected
Inoculated	12	26
Not inoculated	16	6

   By calculating Yule's coefficient of association, examine the effect of vaccine in controlling the disease.

   [**Hints :** the f_{AB} = 12; $f_{A\beta}$ = 16; $f_{\alpha B}$ = 26, $f_{\alpha\beta}$ = 6 and Q_{AB} = –0.705. Thus there is a negative association between tuberculosis and inoculation. Thus vaccine is effective in controlling the disease.]

14. 6000 students appeared in the MBA Entrance Examination and of those 1800 were successful. 1050 had attended a preparatory class and of those 600 came out successful. Estimate the utility of the preparatory class. Also calculate the coefficient of colligation.

15. Eighty eight residents of an Indian city, who had been interviewed during a sample survey, have been classified below according to their smoking and tea drinking habits. Calculate Yule's coefficient of association and comment on its value.

	Smokers	Non-smokers
	(A)	(α)
Drink tea (B)	40	33
Not drinking (β)	3	12

Hints : Positive association between drinking tea and smoking.

4.7 Suggested Readings

1. Yule, G. U. and Kendall, M. G. *An Introduction to the Theory of Statistics,* Charles Griffin 1953.
2. Kendall, M. G. and Stuart, A. *Advanced Theory of Statistics* Vol. II, Charles Griffin 1960.
3. Goon, A. M.; Gupta, M. K. and Dasgupta, B. *Fundamentals of Statistics* Vol. I, The World Press Private Limited, Kolkata, 2002.

Unit 5 ❑ Index Number

Structure

5.0 Objectives

Index numbers are developed to measure the effect of change in prices. They are most widely used as statistical devices and there is hardly any field today where they are not used. Newspaper headlines give the fact that prices are going up and down, that industrial production is rising or falling, that imports or exports are increasing or decreasing in a particular period in comparison to a previous period as disclosed by an index number. Actually, the index numbers are used to feel the pulse of an economy and they are being used as indicators of inflationary or deflationary tendencies. The index numbers are barometers of economic and business activities. If one wants to get an idea about what is happening to an economy he should look at important indices like index number of prices in retail and wholesale markets, index number of industrial production, agricultural production, business activity etc.

5.1 Introduction

Index number is the measurement of relative change in one variable on an average for a group of commodities at one period of time with respect to another. If the variable is price then it is called the price index number and if it is quantity then it is the quantity index number. So price index number is considered as relative change of price on the average for a group of *k* commodities (say) of current period (denoted by 1 or n) with respect to the base period (denoted by O) and it is represented by I_{01} or I_{0n}.

Let p_{0i} and p_{1i} be the prices of the i-th commodity in the base period and in the current period respectively. The change in price of the i-th commodity from base period to current period can be expressed as absolute change $p_{1i} - p_{0i}$ or as relative change $\frac{p_{1i}}{p_{0i}}$, which is also called price relative. Price relative is independent of units of price whereas absolute change is not.

Different persons obtained different price index number formulae considering those price relatives of *k* commodities with different types of weights and using different averages formulae.

5.2 Problems in the construction of a price index number

The following are the various factors which should be considered in the construction of a price index number.

(i) The purpose of constructing the index number should clearly be stated. Selection of items, proper source of data, use of appropriate formula are some of the major aspects of a proper index number formula.

(ii) As we compare price levels of two time periods, the first being base period and the latter being current period, the base periods are to be selected with utmost care following some rules to determine the price index number. The base period should be a normal period showing economic stability i.e., free from booms and depressions arising from catastrophies like wars, floods etc. It should be a period of recent past i.e., not far apart, to get comparable figures. Otherwise tastes, habits, customs of the related people may change with the passage of time. It should not be too short or too long. It should be a period of some economic importance.

(iii) To minimise the time, cost and labour, most representative items are to be selected by judgement according to the purpose and type of index number

instead of considering all the commodities used by the people. Items should be representative of tastes, habits and traditions of the related people. Commodities should be classified into groups having similar type of price fluctuations and commodities selected should be representative of each of these groups. Qualities of selected commodities should not be changed from base period to current period and they should be available at both periods. Number of commodities should neither be too large nor should it be too small.

(iv) The price of a commodity at a particular period of time will vary from one market to another for the same goods and also for different grades of the same goods. Instead of collecting the prices of the commodities from all markets, price quotations should be taken from a number of representative markets for a few important grades of the commodities. For obtaining price quotations proper selections of reporting centre and reporting agencies and frequencies of price quotations are to be considered. To construct cost of living index number retail prices are to be considered and to construct wholesale price index number wholesale prices are to be considered.

(v) As price changes, different commodities are of different importance so different weights will have to be used to find the average change in prices. Let the price index number of current year (n) with respect to base year (0) be I_{0n} and it is generally expressed in percentages by multiplying it by 100. Let p_{ni} and p_{0i} be the prices of current period and base period of the i-th commodity and q_{ni} and q_{0i} be the quantities of the i-th commodity of the current period and base period for i = 1, 2, ..., k. Then different price index number formulae are as follows :

(a) **Simple aggregative method :** $$I_{0n} = \frac{\sum_{i=1}^{k} P_{ni}}{\sum_{i=1}^{k} P_{0i}}. \quad (1)$$

(b) **Weighted aggregative method :** $$I_{0n} = \sum_{i=1}^{k} P_{ni}\, W_i \Big/ \sum_{i=1}^{k} P_{0i}\, W_i \quad (2)$$

where W_i is the weight of the i-th commodity for i = 1, 2, ..., k. If $W_i = q_{0i}$ we get Laspeyres' formula $$I_{0n} = \sum_{i=1}^{k} p_{ni} q_{0i} \Big/ \sum_{i=1}^{k} p_{0i} q_{0i}. \quad (3)$$

If $W_i = q_{ni}$ we get Paasche's formula, $$I_{0n} = \sum_i p_{ni} q_{ni} \Big/ \sum_i p_{0i} q_{ni}. \quad (4)$$

If $W_i = \frac{q_{0i} + q_{ni}}{2}$ we get Marshall-Edgeworth's formula

$$I_{0n} = \sum_i p_{ni}(q_{ni} + q_{0i}) \Big/ \sum_i p_{0i}(q_{ni} + q_{0i}). \quad (5)$$

Fisher's price index number formula is the geometric mean of Laspeyres' and Paasche's formulae. So Fisher's price index number formula is

$$I_{0n} = \sqrt{\frac{\Sigma p_{ni} q_{0i}}{\Sigma p_{0i} q_{0i}} \times \frac{\Sigma p_{ni} q_{ni}}{\Sigma p_{0i} q_{ni}}} \times 100, \quad (6)$$

where summations are taken over number of commodities.

(c) **The average of price relative method :**

Simple arithmetic mean of price relatives, $I_{0n} = \frac{1}{k} \sum_i \frac{p_{ni}}{p_{0i}}$ (7)

Simple geometric mean of price relatives, $I_{0n} = \left(\prod_{i=1}^{k} \frac{p_{ni}}{p_{0i}} \right)^{\frac{1}{k}}$ (8)

Simple harmonic mean of price relatives, $I_{0n} = \frac{k}{\sum_{i=1}^{k} \frac{p_{0i}}{p_{ni}}}$ (9)

(d) **The weighted average of price relative method :**

Let the price relatives have weight W_i for the i-th commodity, i = 1, 2, ..., k.

Weighted A.M. of price relatives, $I_{0n} = \frac{1}{\sum_i w_i} \sum_i \frac{p_{ni}}{p_{0i}} w_i$ (10)

Weighted G.M. of price relatives, $I_{0n} = \left[\prod_{i=1}^{k} \left(\frac{p_{ni}}{p_{0i}} \right)^{w_i} \right]^{\frac{1}{\sum_i w_i}}$ (11)

Weighted H.M. of price relatives, $I_{0n} = \frac{\sum_i w_i}{\sum_i \left(\frac{p_{0i}}{p_{ni}} \right) w_i}$ (12)

(vi) Interpretation of the index number depends on the purpose of index number. For example, the cost of living index number measures the changes in the amount of money required to purchase the same amount of commodities on

the average to maintain same standard of living in current year in comparison to base year. Here retail prices of goods are considered. Similar interpretation holds for wholesale price index number which measures the changes in general wholesale price level of goods on the average in a country from base period to current period.

5.2.1 Errors involved in index numbers

(i) **Formula error.** This arises because of choice of a particular formula to construct an index number. There is no universally accepted price index number formula which measures the change in price exactly.

(ii) **Sampling error.** This arises due to the selection of a sample of goods from the complete list of goods marketed, for construction of index numbers.

(iii) **Homogeneity error.** Due to passage of time old commodities disappear and commodities of new quality appear in the market and it is difficult to maintain the homogeneity between the items selected for base and current year in construction of the index number. This error increases as the gap between the base period and the current period increases.

5.2.2 Uses of index number

Index number guides many business policies and even framing policies. One use of cost of living index number is the decision for increasing or decreasing dearness allowances (DA) of the employees.

Another use is to determine purchasing power of money since price index number gives the amount of money required to purchase a fixed amount of goods in the current year in comparison to the base year. Purchasing power of money is the inverse of index number.

Index number is very useful in calculating the deflation of values. Deflated values are obtained by dividing the values by a price index number. Wages are deflated by dividing them by cost of living index number.

Index numbers are used in studying the general business condition. A company may plan its activities by studying wholesale price index number. The index of industrial production would give the changes in the volume of production.

Index number is used to get the real wages of people

$$\text{where real wage} = \frac{\text{Actual wage}}{\text{Cost of living index number}} \times 100$$

5.2.3 Tests of adequacy of an index number formula

There are a number of tests for judging the adequancy of an index number formula. However, we shall consider only two such tests.

Irving Fisher considered two tests of consistency, which a price index number should satisfy.

1. Time Reversal Test :

An index number formula to be accurate should be time consistent. That is, we should get the same picture of the change in price if the base period (o) and the current period (n) be interchanged. In symbols, this suggests $I_{on} \times I_{no} = 1$,
where I_{0n} is the price index number of current year (n) with respect to base year (0) and I_{n0} is the price index number of current year (0) with respect to base year (n) i.e., I_{n0} is obtained from I_{0n} by interchanging $_0$ and n i.e., by interchanging base year and current year in prices and quantities. If for an index number by $I_{0n} \times I_{n0} \neq 1$ then there exists time bias in that index number.

This test is satisfied by simple aggregative formula (1), weighted aggregative formula (2), Edgeworth-Marshall's formula (5), Fisher's formula (6), simple geometric mean of price relatives (8), weighted geometric mean of price relatives (11), but not by other formulae considered in this chapter above. For example, in the case of Fisher's formula

$$I_{0n} = \sqrt{\frac{\sum_i p_{ni}q_{0i}}{\sum_i p_{0i}q_{0i}} \times \frac{\sum_i p_{ni}q_{ni}}{\sum_i p_{0i}q_{ni}}} \quad \text{and} \quad I_{n0} = \sqrt{\frac{\sum_i p_{0i}q_{ni}}{\sum_i p_{ni}q_{ni}} \times \frac{\sum_i p_{0i}q_{0i}}{\sum_i p_{ni}q_{0i}}}$$

$$\text{and so } I_{0n} \times I_{n0} = \sqrt{\frac{\sum_i p_{ni}q_{0i}}{\sum_i p_{0i}q_{0i}} \times \frac{\Sigma p_{ni}q_{ni}}{\Sigma p_{0i}q_{ni}} \times \frac{\Sigma p_{0i}q_{ni}}{\Sigma p_{ni}q_{ni}} \times \frac{\sum_i p_{0i}q_{0i}}{\sum_i p_{ni}q_{0i}}}$$

$$= \sqrt{1} = 1,$$

and so time reversal test is satisfied by this price index number formula of Fisher.

In case of Laspeyres' price index number formula

$$I_{0n} = \frac{\sum_i p_{ni}q_{0i}}{\sum_i p_{0i}q_{0i}} \text{ and } I_{n0} \text{ is obtained by interchanging 0 and n}$$

$$\text{i.e., } I_{n0} = \frac{\sum_i p_{0i}\, q_{ni}}{\sum_i p_{ni}\, q_{ni}}.$$

Thus $I_{0n} \times I_{n0} = \frac{\sum_i p_{ni}\, q_{0i}}{\sum_i p_{0i}\, q_{0i}} \times \frac{\sum_i p_{0i}\, q_{ni}}{\sum_i p_{ni}\, q_{ni}} \neq 1$ and so time reversal test is not satisfied by this index number of Laspeyre.

Similarly, Paasche's index number also does not satisfy the time reversal test.

However, Marshall-Edgeworth's index number satisfies the time reversal test because for this index number

$$I_{0n} \times I_{n0} = \frac{\sum p_{ni}(q_{0i} + q_{ni})}{\sum p_{0i}(q_{0i} + q_{ni})} \times \frac{\sum p_{0i}(q_{0i} + q_{ni})}{\sum p_{ni}(q_{0i} + q_{ni})}$$
$$= 1.$$

2. **Factor Reversal Test :**

If the price and quantity factors in the price index for number formula (P_{0n}) be inter-changed so as to get the quantity index for number formula (Q_{0n}), the product of these two indices should give the value index (V_{0n}) $= \frac{\sum p_{ni} q_{ni}}{\sum p_{0i} q_{0i}}$.

In symbols, $P_{0n} \times Q_{0n} = V_{0n} = \frac{\sum p_{ni} q_{ni}}{\sum p_{0i} q_{0i}}$.

We will see that Laspeyres' formula does not satisfy factor reversal test since

$$P_{0n} = \frac{\sum_i p_{ni} q_{0i}}{\sum_i p_{0i} q_{0i}}, Q_{0n} = \frac{\sum_i q_{ni} p_{0i}}{\sum_i q_{0i} p_{0i}} \text{ and } V_{0n} = \frac{\sum_i p_{ni} q_{ni}}{\sum_i p_{0i} q_{0i}}$$

But $P_{0n} \times Q_{0n} \neq V_{0n}$.

Similarly, Paasche's formula also does not satisfy the factor reversal test.

Out of formulae (1) to (12) only Fisher's price index number formula satisfies this test because

$$P_{0n} = \sqrt{\frac{\sum p_{ni}\, q_{0i} \times \sum p_{ni}\, q_{ni}}{\sum p_{0i}\, q_{0i} \times \sum p_{0i}\, q_{ni}}},\ Q_{0n} = \sqrt{\frac{\sum q_{ni}\, p_{0i} \times \sum q_{ni}\, p_{ni}}{\sum q_{0i}\, p_{0i} \times \sum q_{0i}\, p_{ni}}}.$$

and $V_{0n} = \frac{\sum p_{ni}\, q_{ni}}{\sum p_{0i}\, q_{0i}}$.

We see that $P_{0n} \times Q_{0n} = V_{0n}$.

However, other index number formulae do not satisfy this test.

5.2.4 Fixed base method

Price index number is constructed on the basis of prices and quantities of two years, base year (0) and current year (n). Prices and quantities of intermediate years are not considered. This method of constructing index numbers is called fixed base method.

5.2.5 Chain index and chain base method

Another method of constructing index number is called chain base method where series of indices are constructed taking the immediately preceding year as the base year and the next following year as the current year and they are multiplied to get the required index number. Chain base index number I'_{0n} is the index number of current year (n) with respect to base year (0). Let the intermediate years be 1, 2, ..., n–1 and the index numbers I_{01}, I_{12}, I_{23}, ..., $I_{(n-1),n}$ are determined as fixed base index numbers where $I_{(i-1),i}$ is the index number of current year (i) with respect to base year (i–1) for i = 1, 2, ..., n. These index numbers I_{01}, I_{12}, I_{23}, ..., $I_{n-1,\,n}$ called link indices, are multiplied to get the chain base index number I'_{0n} which is an index number of current year n with respect to the base year 0 i.e.,

$$I'_{0n} = I_{01} \times I_{12} \times I_{23} \times \ldots \times I_{n-1,n}.$$

By multiplying successive link indices, chain indices are obtained.

$$I'_{02} = I_{01} \times I_{12}$$

$$I'_{03} = I_{01} \times I_{12} \times I_{23}$$

- - - - - - - - - - - - - - - - - -

$$I'_{0n} = I_{01} \times I_{12} \times \ldots \times I_{n-1,n}.$$

These chain indices I'_{0n} are not, in general, equal to fixed base index number I_{0n} unless the formula used, meets the circular test which is

$$I'_{0n} \times I_{n0} = 1$$

i.e., $I_{01} \times I_{12} \times I_{23} \times \ldots \times I_{n-1,n} \times I_{n0} = 1$.

This circular test will be time reversal test when $n = 1$. Simple aggregative index (1), weighted aggregative index (2), simple geometric mean of price relatives (8) and weighted geometric mean of price relatives (11) satisfy this circular test. But other formulae will not satisfy this test.

For example, for the index number formula (2)

$$I_{01} = \frac{\Sigma p_{1i} w_i}{\Sigma p_{0i} w_i},\ I_{12} = \frac{\Sigma p_{2i} w_i}{\Sigma p_{1i} w_i}, \ldots, I_{n-1,n} = \frac{\Sigma p_{ni} w_i}{\Sigma p_{(n-1)i} w_i}$$

and $I_{n0} = \dfrac{\Sigma p_{0i} w_i}{\Sigma p_{ni} w_i}$, where all summations are taken over number of commodities and this formula (2) satisfies circular test, because

$$I_{01} \times I_{12} \times I_{23} \times \ldots \times I_{(n-1),\,n} \times I_{n0} = 1$$

For Fisher's index number,

$$I_{01} = \sqrt{\frac{\sum p_{1i}q_{0i}}{\sum p_{0i}q_{0i}} \times \frac{\sum p_{1i}q_{1i}}{\sum p_{0i}q_{1i}}},\ I_{12} = \sqrt{\frac{\sum p_{2i}q_{1i}}{\sum p_{1i}q_{1i}} \times \frac{\sum p_{2i}q_{2i}}{\sum p_{1i}q_{2i}}},\ \ldots\ldots,$$

$$I_{(n-1),n} = \sqrt{\frac{\sum p_{ni}q_{(n-1)i}}{\sum p_{(n-1)i}q_{(n-1)i}} \times \frac{\sum p_{ni}q_{ni}}{\sum p_{(n-1)i}q_{ni}}} \text{ and } I_{no} = \sqrt{\frac{\sum p_{0i}q_{ni} \times \sum p_{0i}q_{0i}}{\sum p_{ni}q_{ni} \times \sum p_{ni}q_{0i}}}$$

and this formula does not satisfy $I_{01} \times I_{12} \times \ldots \times I_{(n-1),n} \times I_{n0} = 1$ i.e., Fisher's index number formula does not satisfy circular test.

5.2.6 Comparison between fixed base and chain base index numbers

The fixed base index number has been constructed by taking prices and quantities of the base year (0) and the current year (n). Fixed base index number method does not utilise the information on prices and quantities of the intermediate years but the chain base method does. So chain base method is more realistic in nature than the fixed base method.

If new commodities appear and old commodities disappear in the current year in comparison to the base year, chain base is an appropriate one to get I_{0n} instead of fixed base method.

The chain base method takes into account of the dynamics of transition and there link indices enable the comparison between two adjacent time periods but the fixed base method does not take this into account.

The fixed base method is simple and is not laborious but the chain base method is complex and it is laborious.

The chain base method may involve cumulative error but the fixed base method may involve non-cumulative error.

5.2.7 Cost of Living Index Number (CLI)

A cost of living index number is defined as the relative change in the amount of money required for a particular group of people in some geographical area to get equal satisfaction in current year in comparison to base year. Purchasing power of money is defined as inverse of cost of living index number and real wage $= \frac{\text{Actual wage}}{\text{Cost of living index number}} \times 100$.

5.2.8 Construction of a cost of living index number

For the construction of a cost of living index number we first decide on the geographical area and the people for whom the index number will be constructed and define the scope of the index number clearly.

Then we construct a family budget enquiry covering the group of people for whom the index number is to be designed. The enquiry is conducted on a random basis i.e., some families are selected from the total number of families by simple random sample and their family budgets are scrutinised in detail. The objective of conducting family budget enquiry is to determine the amount, on an average, a family spends on different items of consumption. During family budget enquiry the number of commodities consumed are to be taken into account.

The items on which the money income is spent, are classified into certain groups namely, food, clothing, fuel and lighting, house rent, miscellaneous. Each of these groups is further divided into sub-groups. For example, the group 'Food' may be divided into wheat, rice, sugar etc.

Retail prices of items consumed are collected at regular interval of time from important local markets where the people make their purchases. Price quotations are taken at regular interval of time at least once a week.

After quotations have been collected from the retail markets, from where the involved people make their purchases, an average price for each of the items included in the index will be worked out. Such averages are calculated for the base period and then calculated for the current period of the index number. Here the base period should be free from uncertainties like war, flood etc.

While obtaining retail prices it should be noted that (i) it is for fixed list of items and their quality should be fixed, (ii) retail prices are the prices charged from the consumers, (iii) discount, if any, given to consumers should be considered, (iv) in a period of price control if illegal prices are charged openly, they should also be considered.

A separate index number is computed for each group using Laspeyres' formula in the form of weighted average of price relatives and then

$$\text{Group index} = \sum_i w_i \left(\frac{p_{ni}}{p_{0i}} \right) \text{ where } w_i = \frac{p_{0i} q_{0i}}{\sum_i p_{0i} q_{0i}} \times 100$$

i.e. w_i is the percentage of values (or expenditure) for the i-th commodity in the group in relation to total value (or expenditure) in the group as obtained from family budget enquiry.

The weighted average of group index numbers gives cost of living index number (C L I) i.e.,

$$CLI = \frac{\sum_j I_j W'_j}{\sum_j W'_j} \text{ where } \sum_j W'_j = 100 \text{ and } W'_j \text{ of the j-th group is the percentage}$$

of total expenditure of an average family on the j-th group as obtained from family budget enquiry.

Cost of living index numbers are generally constructed for a week. The average of the weekly index numbers in a month gives index number for that month. The average of monthly index numbers gives the cost of living index number for the whole year.

5.2.9 Biases in Laspeyres' and Paasche's index numbers

A true index number is a measure of the ratio of the money values required for a particular group of people to get equal satisfaction in two different situations i.e., in current period and base period.

Let q'_{ni} be the quantity which gives the same satisfaction at the current period as q_{0i} did in the base period. Then the true cost of living index number I_1 for the current period in comparison to base period is

$$I_1 = \frac{\sum_i p_{ni} q'_{ni}}{\sum_i p_{0i} q_{0i}}.$$

Let q'_{0i} be the quantity which gives same satisfaction at the base period as q_{ni} does in the current period. Then another true cost of living index number I_2 for the current period in comparison to base period is

$$I_2 = \frac{\sum_i p_{ni} q_{ni}}{\sum_i p_{0i} q'_{0i}}.$$

I_1 and I_2 will be almost identical if the two periods are such that change in real income resulted in almost no change in consumption pattern. That is generally possible when the two periods are close.

As determinations of q'_{ni} and q'_{0i} are practically impossible, the true cost of living index number I_1 and I_2 cannot be determined. Laspeyres' and Paasche's price index numbers L_p and P_p will approximate the true indices I_1 and I_2 respectively where

$$L_p = \frac{\sum_i p_{ni} q_{0i}}{\sum_i p_{0i} q_{0i}},\ P_p = \frac{\sum_i p_{ni} q_{ni}}{\sum_i p_{0i} q_{ni}}.$$

Common experience shows that more expensive items are replaced by relatively cheaper items. As a result

$$\sum_i p_{ni} q_{0i} > \sum_i p_{ni} q'_{ni} \text{ i.e., } \frac{\sum_i p_{ni} q_{0i}}{\sum_i p_{0i} q_{0i}} > \frac{\sum_i p_{ni} q'_{ni}}{\sum_i p_{0i} q_{0i}} \text{ or } L_p > I_1,$$

$$\text{and } \sum_i p_{0i} q_{ni} > \sum_i p_{0i} q'_{0i} \text{ i.e., } \frac{\sum_i p_{ni} q_{ni}}{\sum_i p_{0i} q_{ni}} < \frac{\sum_i p_{ni} q_{ni}}{\sum_i p_{0i} q'_{0i}} \text{ or } P_p > I_2.$$

The above results show that Laspeyre's price index number has an upward bias while Paasche's price index number has a downward bias in constructing the true cost of living index number.

So $L_p - P_p = (L_p - I_1) + (I_2 - P_p) + (I_1 - I_2)$.

As $L_p > I_1$ and $P_p < I_2$, $(L_p - I_1) + (I_2 - P_p) > 0$ provided tastes and preferences of consumers remain unchanged.

If current period is very close to the base period, I_1 and I_2 differ very slightly and then $I_1 = I_2$. Then $L_p - P_p > 0$ or $L_p > P_p$. So Laspeyres' price index number formula has an upward bias than Paasche's price index number formula.

5.3 Worked out examples

Example 1. Compute price index number of the following data by using (a) simple aggregate, (b) weighted aggregate, (c) simple arithmetic mean of price relatives, (d) weighted arithmetic mean of price relatives, (e) simple geometric mean of price relatives, (f) weighted geometric mean of price relatives.

Commodity	:	A	B	C	D
Quantity	:	4 kg	2 kg	3 kg	6 kg
Price in 1986 (Rs)	:	20	30	10	40
Price in 1991 (Rs)	:	25	30	15	45

Solution : Let p_{ni}, p_{0i} be prices in 1991, 1986 and w_i be the quantity of commodity i, for i = A, B, C, D.

(a) Simple aggregative index $= \frac{\Sigma p_{ni}}{\Sigma p_{0i}} \times 100 = \frac{115}{100} \times 100 = 115\%$

(b) Weighted aggregative index $= \frac{\Sigma p_{ni} W_i}{\Sigma p_{0i} W_i} \times 100$

$$= \frac{25 \times 4 + 30 \times 2 + 15 \times 3 + 45 \times 6}{20 \times 4 + 30 \times 2 + 10 \times 3 + 40 \times 6} \times 100$$

$$= \frac{100 + 60 + 45 + 270}{80 + 60 + 30 + 240} \times 100 = \frac{475}{410} \times 100 = 115.85\%$$

Commodity (i)	:	A	B	C	D	Total
Quantity (W_i)	:	4kg	2kg	3kg	6kg	15kg
Price relatives $\left(\frac{p_{ni}}{p_{0i}}\right)$	:	1.25	1	1.5	1.125	4.875
$\log \left(\frac{p_{ni}}{p_{0i}}\right)$	:	0.0969	0	0.1761	0.05115	0.32415
$W_i \times \frac{p_{ni}}{p_{0i}}$	:	5	2	4.5	6.75	18.25
$W_i \log \times \left(\frac{p_{ni}}{p_{0i}}\right)$	:	0.3876	0	0.5283	0.3069	1.2228

(c) Simple arithmetic mean of price relatives $= \frac{1}{\sum w_i} \sum_i w_i \frac{p_{ni}}{p_{0i}} \times 100 = \frac{4.875}{4} \times 100$

= 121.88%

(d) Weighted a.m. of price relatives $= \frac{1}{\sum w_i} \sum_i w_i \frac{p_{ni}}{p_{0i}} \times 100 = \frac{18.25}{15} \times 100$

= 121.67%

(e) Simple g.m. of price relatives $= \left[\text{anti log}\left\{\frac{1}{4}\sum_i \log\left(\frac{p_{ni}}{p_{0i}}\right)\right\}\right] \times 100$

$= \left[\text{anti log}\left(\frac{0.32415}{4}\right)\right] \times 100 = (\text{anti log } 0.0810375) \times 100$

$= 1.2051 \times 100 = 120.51\%.$

(f) Weighted g.m. of price relatives $= \left[\text{anti log}\left\{\frac{1}{\Sigma w_i}\sum_i w_i \log\left(\frac{p_{ni}}{p_{0i}}\right)\right\}\right] \times 100$

$= \left[\text{anti log}\left\{\frac{1.2228}{15}\right\}\right] \times 100 = [\text{anti log } 0.8152] \times 100 = 1.2065 = 120.65\%$

Example 2. Calculate price index number and quantity index number by (i) Laspeyres', (ii) Paasche's, (iii) Fisher's and (iv) Edgeworth-Marshall methods.

Commodity	1989		1991	
	Price (Rs)	Total Value (Rs).	Price	Total value (Rs.)
A	10	100	12	120
B	12	144	14	196
C	14	140	16	192
D	16	192	18	216
E	18	270	20	320

Also show that Fisher's price index number satisfies time reversal test and factor reversal test.

Solution : Let p_{0i} and q_{0i} be the price and quantity of base year 1989 and p_{ni} and q_{ni} be the price and quantity of current year 1991 of the i-th commodity for i = A, B, C, D, E. Thus $p_{0i}\, q_{0i}$ = Total value in 1989 and $p_{ni}\, q_{ni}$ = Total value in 1991 of the i-th commodity for i = A, B, C, D, E.

Commodity (i)	p_{0i}	$p_{0i}\ q_{0i}$	q_{0i} (units)	p_{ni}	$p_{ni}q_{ni}$	q_{ni} (units)	$p_{ni}q_{0i}$	$p_{0i}q_{ni}$
A	10	100	10	12	120	10	120	100
B	12	144	12	14	196	14	168	168
C	14	140	10	16	192	12	160	168
D	16	192	12	18	216	12	216	192
E	18	270	15	20	320	16	300	288
Total		846			1044		964	916

So $\sum_i p_{0i}\, q_{0i} = 846, \sum_i p_{0i} q_{ni} = 916, \sum_i p_{ni} q_{0i} = 964, \sum_i p_{ni} q_{ni} = 1044.$

Price index numbers (P_{on}) and quantity index numbers (Q_{on}) are obtained as follows :

(i) Laspeyres' method : $P_{on} = \dfrac{\sum_i p_{ni} q_{0i}}{\sum_i p_{0i} q_{0i}} \times 100 = \dfrac{964}{846} \times 100 = 113.95\%$

$$Q_{on} = \frac{\sum_i q_{ni}\, p_{0i}}{\sum_i q_{0i}\, p_{0i}} = \frac{916}{846} \times 100 = 108.27\%.$$

(ii) Paasche's method : $P_{0n} = \dfrac{\sum_i p_{ni}\, q_{ni}}{\sum_i p_{0i}\, q_{ni}} \times 100 = \dfrac{1044}{916} \times 100 = 113.97\%.$

$$Q_{0n} = \frac{\sum_i q_{ni}\, p_{ni}}{\sum_i q_{0i}\, p_{ni}} \times 100 = \frac{1044}{964} \times 100 = 108.30\%.$$

(iii) Fisher's method : $P_{0n} = \sqrt{\dfrac{\Sigma p_{ni}\, q_{0i}}{\Sigma p_{0i}\, q_{ni}} \times \dfrac{\Sigma p_{ni}\, q_{ni}}{\Sigma p_{0i}\, q_{ni}}} \times 100 = \sqrt{\dfrac{964}{846} \times \dfrac{1044}{916}} \times 100$

$$= \sqrt{1.1395 \times 1.1397} \times 100 = \sqrt{1.298688} \times 100 = 113.96\%.$$

$$Q_{0n} = \sqrt{\frac{\Sigma q_{ni}\, p_{0i}}{\Sigma q_{0i}\, p_{0i}} \times \frac{\Sigma q_{ni}\, p_{ni}}{\Sigma q_{0i}\, p_{ni}}} \times 100 = \sqrt{\frac{916}{846} \times \frac{1044}{964}} \times 100$$

$$= \sqrt{1.0827 \times 1.0830} \times 100 = \sqrt{1.172564} \times 100 = 108.28\%.$$

(iv) Edgeworth-Marshall method : $P_{0n} = \dfrac{\sum_i p_{ni}(q_{ni} + q_{0i})}{\sum_i p_{0i}(q_{ni} + q_{0i})} \times 100 = \dfrac{1044 + 964}{916 + 846} \times 100$

$$= \frac{2008}{1762} \times 100 = 113.96\%.$$

$$Q_{0n} = \frac{\Sigma q_{ni}(p_{ni} + p_{0i})}{\Sigma q_{0i}(p_{ni} + q_{0i})} = \frac{1044 + 916}{964 + 846} \times 100 = \frac{1960}{1810} \times 100$$

$$= 108.29\%.$$

For time reversal test $P_{0n} \times P_{n0} = 1$. By Fisher's method

$$P_{0n} \times P_{n0} = \sqrt{\frac{\Sigma p_{ni}\, q_{0i}}{\Sigma p_{0i}\, q_{0i}} \times \frac{\Sigma p_{ni}\, q_{ni}}{\Sigma p_{0i}\, q_{ni}}} \times \sqrt{\frac{\Sigma p_{0i}\, q_{ni}}{\Sigma p_{ni}\, q_{ni}} \times \frac{\Sigma p_{0i}\, q_{0i}}{\Sigma p_{ni}\, q_{0i}}}$$

$$= \sqrt{\frac{964}{846} \times \frac{1044}{916}} \times \sqrt{\frac{916}{1044} \times \frac{846}{964}} = \sqrt{\frac{964}{964} \times \frac{1044}{1044} \times \frac{846}{846} \times \frac{916}{916}}$$

$$= \sqrt{1} = 1.$$

So time reversal test is satisfied by Fisher's price index number.

For factor reversal test $P_{0n} \times Q_{0n} = V_{0n}$ where P_{0n}, Q_{0n} and V_{0n} are price index number, quantity index number and value index numbers respectively.

By Fisher's method $P_{0n} = \sqrt{\frac{\Sigma p_{ni} q_{0i}}{\Sigma p_{0i} q_{0i}} \times \frac{\sum p_{ni} q_{ni}}{\sum p_{0i} q_{ni}}} = \sqrt{\frac{964}{846} \times \frac{1044}{916}}$,

$$Q_{0n} = \sqrt{\frac{\Sigma q_{ni} p_{0i}}{\Sigma q_{0i} p_{0i}} \times \frac{\sum q_{ni} p_{ni}}{\sum q_{0i} p_{ni}}} = \sqrt{\frac{916}{846} \times \frac{1044}{964}} \text{ and } V_{0n} = \frac{\Sigma p_{ni} q_{ni}}{\Sigma p_{0i} q_{0i}} = \frac{1044}{846}$$

So $P_{0n} \times Q_{0n} = \sqrt{\frac{964}{846} \times \frac{1044}{916}} \times \sqrt{\frac{916}{846} \times \frac{1044}{964}} = \sqrt{\frac{(1044)^2}{(846)^2}} = \frac{1044}{846} = V_{0n}$

Hence Fisher's price index number satisfies factor reversal test.

Example 3.

The cost of living indices with percentages of total expenditures for 5 different groups are given for middle class people of Kolkata in 2011 with 2001 as base. Determine the cost of living index number in 2011 with 2001 as base :

Group :	Food	Clothing	Light and fuel	House rent	Others
Group Index No. (%)	625	320	250	300	180
Percentage of total expenditure	40	16	15	20	9

Mr. X got a salary of Rs. 2500/- in 2001. Determine how much he will have to receive as salary in 2011 to maintain his same standard of living as in 2001.

Solution :

Suppose the group index is represented by I and percentage of total expenditure is represented by weight W. Then cost living index number of 2011 with 2001 as base

$= \frac{\sum_{j=1}^{5} W_j}{\sum_{j=1}^{5} W_j}$ where summation is taken over different groups.

$\therefore$ the required CLI $= \frac{625 \times 40 + 320 \times 16 + 250 \times 15 + 300 \times 20 + 180 \times 9}{40 + 16 + 15 + 20 + 9}$

$= \frac{25000 + 5120 + 3750 + 6000 + 1620}{100} = \frac{41490}{100} = 414.90\%$

To maintain same standard of living as in 2001, in the year 2011 he will have to receive salary = 2500 × 414.90% = 2500 × 4.1490 = Rs. 10372.50.

5.4 Summary

This chapter contains the definition of index number, problems arising in construction of an index number, errors and uses of index number, tests of consistency in the formula used, fixed base and chain base index number and their comparisons, cost of living index number and its construction and worked out examples.

5.5 Exercises

1. What do you mean by an index number? State the uses of an index number.
2. Discuss the various problems in the construction of an index number.
3. Discuss the importance of use of weights in the construction of an index number.
4. Explain price index number with an example and give its uses.
5. Explain the formula and construction of the following index number of prices : (i) Laspeyres', (ii) Paasche's, (iii) Marshall-Edgeworth's and (iv) Fisher's.
6. What weights are used in (i) Laspeyres', (ii) Paasche's, (iii) Marshall-Edgeworth's price index numbers. Show that the third index number lies between the first two.
7. Explain the Time Reversal Test and Factor Reversal Test of index number and examine whether the following index numbers satisfy these tests :
   Laspeyres' index number, Paasche's index number, Marshall-Edgeworth's index number, Fisher's index number.
8. What is a cost of living index number? What does it measure? How is it constructed?

9. What is the chain base method of construction of index number and how does it differ from the fixed base method?

10. What do you mean by a link index? Discuss the relative merits and demerits of chain base and fixed base index number.

11. Explain what is meant by saying that Laspeyres' formula has an upward bias while Paasche's formula has a downward bias.

12. What is a chain index? What do you mean by circular test? Show that the chain indices are equal to fixed base indices if the formula used satisfies circular test. Discuss the advantages and disadvantages of the former over the latter.

13. Write Laspeyres' and Paasche's price index number formulae. Show that both of them can be expressed as weighted arithmetic mean and weighted harmonic mean of price relatives.

14. Discuss the considerations that should guide us in the choice of a base period and the choice of weights while constructing a price index number.

15. Determine the price index number of 1999 with 1995 as base year from the following data using (i) Laspeyres', (ii) Paasche's, (iii) Marshall-Edgeworth's, (iv) Fisher's methods.

Commodity	Unit	Year 1995		Year 1999	
		Quantity	Price (Rs)	Quantity	Price (Rs)
A	Kg	5	2.00	7	4.50
B	Quintal	7	2.50	10	3.20
C	Dozen	6	3.00	6	4.50
D	Kg	2	1.00	9	1.80

16. Calculate the price index numbers from the following data using (i) simple aggregative formula, (ii) weighted aggregative formula, (iii) simple arithmetic mean of price relatives and (iv) weighted arithmetic mean of price relatives.

Commodity	Base price (Rs.)	Current price (Rs.)	Weight
A	80	110	14
B	10	15	20
C	40	56	35
D	50	95	15
E	12	18	16

17. Calculate the price number for the year 1981 with 1971 as base from the following data by using (i) Laspeyres', (ii) Paasche's, (iii) Marshall-Edgeworth's and (iv) Fisher's formula.

Commodity	Unit	1971		1981	
		Price (Rs)	Money value (Rs)	Quantity consumed	Money value (Rs)
A	Kg	2.40	36.00	14	42.00
B	500 gm	3.60	75.60	21	84.00
C	meter	10.80	64.80	5	70.00
D	tin	2.50	7.50	2	10.00

(Here money value means the total value of the commodity).

18. From the table of group index numbers and group expenditures given below, calculate the cost of living index number.

Group	Food	Clothing	Light and fuel	House rent	Others
Index No.	428	250	220	125	175
Percentage of total expenditure	45	15	8	20	12

19. From the following figures determine the relative importance for the food group, given that the cost of living index number for 1975 with 1970 as base is 175.

Group	Food	Clothing	Light and fuel	House rent	Miscellaneous
% increase in expenditure	65	90	20	70	150
Weight	–	12	18	10	20

20. The price relatives in percentage and weights of a set of commodities are given in the following table.

Commodity	A	B	C	D
Price relative	115	110	125	116
Weight	W_1	W_2	$2W_1$	W_2-2

If the sum of the weights is 30 and the price index number for the set of commodities is 118%, find the numerical values of W_1 and W_2.

21. Distinguish between fixed base and chain base index numbers and point out their merits and demerits.

22. Using the following data, prove that the Time Reversal Test and the Factor Reversal Test are satisfied by Fisher's index number :

	Base year		Current year	
Commodity	Price	Quantity	Price	Quantity
A	6	50	10	56
B	2	100	2	120
C	4	60	6	60
D	10	30	12	60
E	8	40	12	36

23. Write a brief note on errors in index number.
24. Define Fisher's index number. Why is this index number called the ideal index number?
25. The cost of living index for a certain consumer group goes up from 110 in 1991 to 200 in 1996 and the salary of a worker is raised from Rs. 325 in 1991 to Rs. 500 in 1996. Does the worker really gain and if so, how much in real terms?
26. Show that both Laspeyre's and Paasche's price index numbers may be regarded as weighted averages of price relatives. Aslo specificy the weight in each case.
27. Define an index number. Explain clearly the various steps involved in the construction of an index number.
28. If the wages of a group of workmen are increased by 40% and the cost of living rises by 25%, how much greater is their purchasing power than before the change took place?
29. Monthly income of an employee was Rs. 800 per month in 2005. The consumer price index number was 160 in 2005 and it rose to 225 in 2010. Calculate the additional dearness allowances to be paid to the employee if he is to be rightly compensated.
30. If price index of Laspeyre (P_L) is greater than the price index of Paasche (Pp), prove that the quantity index of Laspeyre (Q_L) is greater than the quantity index of Paasche (QP).

Hint : $P_L \times Q_P = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$

Also, $Q_L \times P_P = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$

Thus $P_L \times Q_P = Q_L \times Pp$.

That is, $\frac{P_L}{P_P} = \frac{Q_L}{Q_P}$.

But $P_L > Pp$. Hence, $Q_L > Q_P$ proved.

31. You are given the following information :

Commodities	Base year price (Rs.)	Base year quantities (Kg.)	Current year price (Rs.)	Current year quantities (Kg.)
A	5	50	10	56
B	3	100	4	120
C	4	60	6	60
D	11	30	14	24
E	7	40	10	36

Compute the quantity index numbers by using (i) Laspeyre's formula, (ii) Paasche's formula, (iii) Fisher's formula and (iv) Marshall Edgeworth's formula.

5.6 Suggested Readings

1. Croxton, F. E. and Cowden, D. J., *Applied General Statistics,* Prentice Hall 1967 and Prentice Hall of India, 1969.
2. Mills, F. C., *Statistical Methods,* Henry Holt, 1955.
3. Mudgett, B. D., *Index Numbers*, John Willey, 1951.
4. Goon, A. M., Gupta, M. K. and Dasgupta, B., *Fundamentals of Statistics* Vol. II, The World Press Private Limited, Kolkata, 2002.
5. Gupta, S. P., *Statistical Methods,* Sultan Chand & Sons, 1980.

Unit 6 ❑ Time Series

Structure

6.0 Objectives

Every business man wants to gain from his business and every economist, particularly in the planning sector of a country, wants to plan the economy of that country in such a way that the country can develop in terms of the growth of national income. For this the most important task before the economists and business men is to forecast the results in the future from the past data. For example, a business man is interested in growing his sales by long term planning and try to gain by increasing his production in growth sector and by decreasing the production in decline sector of demand for products which are purchased by the customers. Similarly, an economist is interested in estimating the likely productions in the coming years so that proper planning can be carried out with regard to food supply, jobs of the people etc. The first step to get estimates for the future consists of gathering information from the past. For this statistical data are to be collected and recorded at regular interval of time. These data are called time series data. Here time sequence is of prime importance and the data require special techniques for analysis of the series. Analysis of past data helps to understand the future data.

6.1 Introduction

Time series is the set of values of some statistical variable that depends on time and is generally measured over a uniform set of time points. For example, (i) annual turnover of a factory for 10 successive years and (ii) number of employees for each

quarter of 5 consecutive years are time series data. The time series is mathematically represented as

$$y = f(t)$$

where y is the dependent variable, say, production and t is the independent variable, time.

In time series, cycle is a general pattern which broadly repeat and occur. For example, monthly sales of a business will exhibit some natural 12 monthly cycle. To use time series effectively for the control of current business activities or to plan for further business activities, records are to be kept in regular basis and data have to be organised and analysed.

6.2 Time Series Analysis

It is the evaluation and extraction of the components of a model that breakdown a particular series into understandable and explainable portions and enables (a) trend to be identified, (b) extraneous factors to be eliminated and (c) forecasts to be made. The understanding, description and use of these processes is known as time series analysis.

- **Necessity of time series analysis**

The analysis of time series is of great importance to the economists, business men and different research workers. (a) It helps in understanding past behaviour and in predicting future behaviour. (b) It helps in planning future operations by forecasting i.e., by determining the relationship between time and yield. (c) It helps in evaluating accomplishments by comparing actual performance with expected performance and in judging the cause of variation. (d) It facilitates comparison between different time series data and different conclusions drawn therefrom.

- **Factors for the change in the time series data**

It is a fact that by time series analysis one cannot foretell the course of future actions with 100% accuracy. But this accuracy can be achieved nearly if influence of various forces, which affect these time series, can be known. The forces are climate, tastes, habits, customs and traditions, growth and decline of factors etc. For this some important adjustments in a time series are necessary. These are of three types : (a) Calender variations, (b) Population changes and (c) Price changes.

(a) Calender variations

We all know that the number of days in different calender months of a year varies from month to monty. Besides, if holidays and weekends are considered, variations will be much higher. The adjustment for calender variations is made by dividing each monthly total by the month of days (very often by the number of working days) in the month and this gives the daily average of each month.

(b) Population changes

Very often it becomes necessary to adjust the data for population changes. This is specially applicable when national income (or production) is increasing but percapita income (or consumption) is decreasing due to change in the size of population. Under these situations the original data are divided by the appropriate population totals to reduce them to percapita values which are appropriate for comparison.

(c) Price changes

Adjustment for price changes is made by a process know as deflation. This method needs in dividing the original observations by an appropriate price index number.

- **Standard time series models**

In order to explain movement of time series data models are to be constructed combining different components to form individual data values. To describe the time series data two main models, namely (i) additive model $y = T + S + C + I$ and (ii) multiplicative model $y = T \times S \times C \times I$ can be used where y = time series value, T = the trend component, S = seasonal component, C = cyclical component and I = Irregular component. Here T, S and C are regular components and they can be determined. Some describe C + I in additive model and C × I in multiplicative model as residual component. Evaluation and interpretation of these components is the main aim of time series analysis. Note that although the trend component is same for the two models the value of seasonal and cyclical components will depend on the model being used.

6.2.1 Components of time series

Secular trend (or simply trend, T) is the smooth, regular and long term movement of time series when data are observed for long enough. Trend may be upward trend, downward trend or constant trend. Here frequent changes are not possible though trend may change its direction somewhere within the long period of time. For example, population is increasing with the passage of time. This gives an increasing trend.

Seasonal fluctuation (S) is the periodic movement in a time series where the period is not longer than one year. If the movement recurs at regular interval of time then it is periodic. Sale of woolen garments in a year is an example of seasonal fluctuation.

Cyclical fluctuation (C) represents the oscillatory movement in time series data, the period of oscillation being more than a year. One complete period is called a cycle. Cyclical fluctuations are not necessarily periodic since the lengths of the cycles and the intensities of fluctuations may change from one cycle to another. There are four well defined phases in business cycle, namely prosperity, recession, depression and recovery.

Irregular variations are erratic and random in nature and do not repeat in a definite pattern. They are unpredictable. These variations are due to unforeseen events like flood, earthquake, revolution etc., which do not follow any law.

- **Adjustment in time series data**

As number of working days are different in different months then to make the data comparable for production of a factory in different months data are to be taken per working day on the average (obtained by dividing the monthly data by number of working days in different months). Adjustments should also be done in terms of units.

6.2.2 Measurement of secular trend

The following methods determine the trend.

(i) **Method of drawing free hand curve (Or the method of Inspection)**

Time is plotted along the x-axis and data are put along the y-axis. A line diagram or curve is drawn through the given points so that fitting of the diagram to the data is best. From that fitted diagram we get trend values for different time points. This method is subjective since different persons can draw different curves which appear to them to be the best fitted one through the given points. So the method is quite flexible and does not assume any mathematical law. For determining trend this method is not suitable for forecasting purposes. This method should be used by a specialist with long experience but not by the beginners. This method is simple to understand and gives a rough idea about the matter of trend revealed in the time series.

(ii) Method of moving averages

The moving average period of a time series data gives a new series of arithmetic means, for each of k successive observations of the time series when arranged in chronological order. We start with first k observations and calculate the average. Then we exclude the 1st observation and include the (k + 1)st observation in this set and get the average. The procedure is repeated till we reach last k observations and get the arithmetic mean of them. Each of these means are centred against the time which is the mid-point of the time interval included in the calculation of moving averages. If k, the period of moving average, is odd the moving average values correspond to the tabulated time periods for which the time series is given. If k is an even number, the average values correspond to the mid-value between two centred tabulated time periods and we place those average values in mid-way between corresponding two time periods. Then we calculate the two-item moving averages to get the centred moving average values, which correspond to the tabulated time periods. These

averages corresponding to the particular time periods give the trend values and the effect of averaging gives a smoother value, which has less influence of fluctuations that pulls the figures away from the general trend.

Moving averages with a properly selected period will smooth out cyclical fluctuations and other fluctuations than trend from the series and give estimates of trend. Thus the central problem in this method is to select an appropriate period which will eliminate these fluctuations.

Cycles in economic time series are not strictly periodic (i.e., not of periods of fixed length). If that be so cyclical variation will be completely eliminated when we consider moving averages of data with that period or multiple of that period provided the trend is linear. The period and the amplitude generally vary from cycle to cycle. In such cases the best results may be obtained by using the period as average of the periods of the cycles to get moving averages. This, however, will not completely eliminate the cycles though there is no other way to overcome this in the method of moving averages.

There will be further complications if the trend is non-linear. If the trend curve is concave upward, a moving average will over estimate and if the curve is convex upward it will underestimate the trend values.

As this method does not assume any law of change it cannot be used for forecasting purposes. Besides, by this method a number of trend values cannot be estimated at each end of the series.

Merits : This method is simple. It is a flexible method of measuring trend for the reason that if a few more figures are added to the data the entire calculations will not change. Only a few of them need to be recalculated. If the periods of moving average coincide with the period of cyclical fluctuation in the data, such fluctuations are automatically eliminated. It is effective if the trend of a series is very irregular.

Demerits : All trend values cannot be computed. Since moving average is not represented by a mathematical law, this method cannot be used for forecasting, which is the main object of calculating the trend. Exact period of moving average can not be obtained, instead we get average period from periods of cycles though it is approximate. If the trend is non-linear the moving average either overestimates or underestimates according as the trend is concave upward or it is convex upward.

(iii) Method of fitting mathematical curves

This is the best and most objective method of determining trend. Here appropriate type of trend equation is at first selected and the constants are determined by the

method of least squares. The choice of appropriate polynomial is facilitated by a graphical representation of data and for this apart from arithmetic scale, semi-logarithmic or double logarithmic scale may also be used. By looking at the graph of data we consider approximately whether the general trend equation is linear or quadratic or exponential etc.

For linear trend equation $T_t = a_0 + a_1 t$ fitted through *n* given points (t, y_t), t = 1, 2, ..., n is obtained after deriving values of a_0 and a_1 by minimising $E = \sum_{t=1}^{n} (y_t - a_0 - a_1 t)^2$ i.e., a_0 and a_1 are obtained from normal equations :

$$\sum_t y_t = n a_0 + a_1 \sum_t t \qquad \text{... (1)}$$

and $$\sum_t t y_t = a_0 \sum_t t + a_1 \sum_t t^2 \qquad \text{... (2)}$$

Here normal equations are obtained by differentiating E with respect to a_0 and a_1 and equating them to zero.

The quadratic trend equation $T_t = a_0 + a_1 t + a_2 t^2$ will be best fitted through *n* points (t, y_t), t = 1, 2, ..., n if a_0, a_1 and a_2 are so determined that $\sum_{t=1}^{n} \left(y_t - a_0 - a_1 t - a_2 t^2\right)^2$ is minimum. Here a_0, a_1 and a_2 are determined from the normal equations :

$$\sum_t y_t = n a_0 + a_1 \Sigma t + a_2 \Sigma t^2 \qquad \text{... (3)}$$

$$\sum_t t y_t = a_0 \Sigma t + a_1 \Sigma t^2 + a_2 \Sigma t^3 \qquad \text{... (4)}$$

$$\sum_t t^2 y_t = a_0 \Sigma t^2 + a_1 \Sigma t^3 + a_2 \Sigma t^4 \qquad \text{... (5)}$$

To fit the exponential trend equation $T_t = ab^t$ through the *n* given points (t, y_t), t = 1, 2, ..., n we determine unknowns *a* and *b* so that $\sum_t (\log y_t - A - Bt)^2$ is minimised. The exponential trend equation can be written as

$$\log T_t = \log a + t \log b \text{ i.e., } \log T_t = A + Bt,$$

where A = log *a* and *B* = log *b*.

The normal equations to solve for A and B are :

$$\sum_t \log y_t = nA + B \Sigma t \qquad \text{... (6)}$$

and $$\sum_t t \log y_t = A \Sigma t + B \Sigma t^2 \qquad \text{... (7)}$$

The calculations will be simplified in the above equations if we take suitable origin and scale of t so that $\Sigma t = \Sigma t^3 = 0$. That is, the sum of odd powers of t becomes zero. In all the trend equations origin of t and unit of t should be properly mentioned.

Derivation of monthly or quarterly trend equation from annual data

Sometimes monthly or quarterly trend values are required instead of yearly trend values from the linear trend equation or parabolic trend curve based on data of monthly or quarterly averages.

Consider the parabolic trend equation fitted to annual total for an odd number of years, given as $y = A_0 + A_1t + A_2t^2$, where origin is at the middle-most year and unit of t is 1 year. For an even number of years let the equation be given as $y = B_0 + B_1t + B_2t^2$, where origin is at the middle of the two middle-most years and unit of t is $^1/_2$ year.

Now as monthly average in a year $= \frac{1}{12} \times$ Total yield in that year, the trend equation based on monthly averages in both cases are :

$$y = \frac{A_0}{12} + \frac{A_1}{12}t + \frac{A_2}{12}t^2$$

and $$y = \frac{B_0}{12} + \frac{B_1}{12}t + \frac{B_2}{12}t^2$$

origin and unit of t remain as usual.

Then monthly trend equations would be

$y = \frac{A_0}{12} + \frac{A_1}{12} \cdot \frac{t + \frac{1}{2}}{12} + \frac{A_2}{12} \frac{\left(t + \frac{1}{2}\right)^2}{144}$, origin is at July of the middle-most year and unit of t is 1 month

and $y = \frac{B_0}{12} + \frac{B_1}{12} \cdot \frac{t + \frac{1}{2}}{6} + \frac{B_2}{12} \cdot \left(\frac{t + \frac{1}{2}}{6} \right)^2$, origin is at January of the second middle-most year and unit of t is 1 month.

Also as quarterly average in a year is $\frac{1}{4}$th of total yield in that year, the trend equation based on quarterly averages in both the cases will be

$$y = \frac{A_0}{4} + \frac{A_1}{4}t + \frac{A_2}{4}t^2$$

and $$y = \frac{B_0}{4} + \frac{B_1}{4}t + \frac{B_2}{4}t^2$$

origin and unit of t remain as usual. Then quarterly trend equations would be

$y = \frac{A_0}{4} + \frac{A_1}{4} \cdot \frac{t + \frac{1}{2}}{4} + \frac{A_2}{4} \cdot \left(\frac{t + \frac{1}{2}}{4} \right)^2$, origin is at the 3rd quarter of the middle-most year and unit of t = 1 quarter,

and $y = \frac{B_0}{4} + \frac{B_1}{4} \cdot \frac{t + \frac{1}{2}}{4} + \frac{B_2}{4} \cdot \left(\frac{t + \frac{1}{2}}{2} \right)^2$, origin is at the first quarter of the second middle-most year and unit of t = 1 quarter.

Merits and demerits of mathematical curve fitting

Merits : This is a mathematical method of measuring trend. So there is no possibility of subjectiveness. This method eliminates personal bias.

This method follows some law. The curve obtained by this method is called the curve of best fit since sum of squares of deviations of original value and fitted value of yields is minimum.

This method has the great advantage of forecasting future. Values for given and future time periods can be determined from the trend equation.

The great advantage of fitting the exponential trend is that it enables us to find compound annual growth (CAG) rate of the depending variable, say, national income, population, exports, imports and the like. The CAG is

$r = (b - 1) \times 100\%$

and is widely used in research.

However, the least squares method fails in fitting the parameters involved in modified exponential curve, Gompertz curve and the logistic curve. For fitting these curves which are very often used in research, some other techniques are available.

Demerits : The calculation procedure of this method is tedius than other methods.

Extra data in the time series require fresh calculations.

6.2.3 Measurement of seasonal variation

There are various methods for measuring seasonal variation in time series depending on the elimination of the other components like trend, cyclical and irregular variations. The following methods are popularly used for measuring seasonal variations :

1. Method of monthly (or quarterly) averages.
2. Ratio to moving average method.

3. Ratio to trend method.

4. Method of link relatives.

However, the last three methods are quite popular.

2. Ratio to moving average method

If the period of fluctuations of the data is the same as the period of moving average then moving average will smooth out periodic fluctuations. For quarterly data four quarterly moving averages are taken while in the case of monthly data 12 period moving averages are taken, after centering them properly we get $m = T \times C \times I_1$ where T = trend, C = cyclical variation and I_1 = some part of irregular variation while original data y is expressed as $y = T \times S \times C \times I$ where T and C are defined as above, S = seasonal variation and I = irregular variation in the case of multiplicative model. Now $I = I_1 \times I_2$ where I_1 is defined above and I_2 is other part of irregular component. Then ratio of the data to the moving average is

$$\frac{y}{m} = \frac{T \times S \times C \times I}{T \times C \times I_1} = S \times I_2 .$$

First express these ratios in percentages for each quarter for which moving averages are available. Other data for which moving averages are not obtained, are to be ignored. These percentages are arranged in a table against the quarters (months) of the given number of years in chronological order. Different values for each quarter (month) are then averaged so that irregular fluctuations may be removed.

If G is the total for unadjusted quarterly averages then an adjustment is made by multiplying each quarterly average by the correction factor $\frac{400}{G}$ and then the seasonal indices for the four quarters Q_1, Q_2, Q_3 and Q_4 (or for twelve months twelve seasonal indices) are obtained.

For the additive model the components T, S, C and I of the time series satisfy $y = T + S + C + I$ and moving average value m' gives $m' = T + C + I'$ where $I = I' + I''$, I' and I'' are two parts of irregular component (I) where I' is obtained in moving average value (m) as shown above. Then $y - m' = S + I''$ and these deviations are arranged for quarters of the years. Then for each quarter values are averaged to remove irregular fluctuation. Then grand mean is subtracted from quarterly averages to get the adjusted seasonal indices for different quarters.

For the monthly and daily data, the moving averages are obtained by using periods of 12 months and 7 days respectively. Then the data received either by dividing or by

subtracting the moving average value from the original data of the corresponding period, depending on whether the model is multiplicative or additive, are arranged in the table for each month or day as the case may be. The procedure is repeated to get seasonal indices for 12 months or 7 days as in the method of quarterly averages. Let G_1 and G_2 be the totals of unadjusted monthly averages and that of unadjusted daily averages respectively $\frac{1200}{G_1}$ and $\frac{700}{G_2}$ would be multiplied to or $\frac{G_1}{12}$ and $\frac{G_2}{7}$ would be subtracted from the monthly averages and daily averages respectivly in multiplicative model or additive model.

3. Ratio to trend method

By plotting the data in a graph paper considering time along the horizontal axis and data along the vertical axis, first determine the type of curve T = f(t) of trend equation which fits the data best. Then the constants in f(t) would be determined by fitting the curve y = f(t) to the given data by least squares method. Let the data given be of different quarters in some consecutive years, then obtain quarterly trend equation first and obtain trend values from that equation for the given quarters of the given years. Then obtain ratio of trend values to the yield given and express them in percentages. These percentage values, obtained for each quarter of different years are then arranged and averaged and these averages are adjusted to a total of 400 by multiplying each by a factor $\frac{400}{G}$ where G is the total of averages obtained for different quarters.

Here by averaging the ratio of original data to the trend values we eliminate irregular and cyclical variations.

For additive model the trend values are subtracted from the original data and then averages are taken for each quarter over different years and they are adjusted to total zero by subtracting average of average quarter values so obtained from each average quarter values.

Simplification in calculation can be made by fitting the trend equation to yearly totals (or averages) instead of quarterly values and then obtaining the quarterly trend values by a suitable modification of the trend equation.

4. Method of link relatives

Each quarterly value is to be expressed as a percentage of the previous quarterly value. These percentages are called link relatives. Let l_t be the link relative of time

series data y_t at time period, quarter t. Then $l_t = \frac{y_t}{y_{t-1}} \times 100$ is the link relative for quarter t. Link relative of all the given quarters of the given years in chronological order are obtained by the formula $l_t = 100\ y_t/y_{t-1}$ but link relative of first given quarter cannot be obtained. Then the average of the link relatives for each quarter is calculated. Then convert these averages to chain relatives on the basis of chain relative of first quarter taken as 100. Let A_1, A_2, A_3 and A_4 be the averages of link relatives of the quarters Q_1, Q_2, Q_3 and Q_4 respectively. Then chain relatives C_1, C_2, C_3 and C_4 for four consecutive quarters starting from the first would be obtained in the following way :

$$C_1 = 100, C_2 = \frac{A_2 \times C_1}{100}, C_3 = \frac{A_3 \times C_2}{100}, \text{ and } C_4 = \frac{A_4 \times C_3}{100}.$$

If $\frac{A_1 \times C_4}{100} = 100$, it will be all right, otherwise determine the correction factor $b = \left(\frac{A_1 \times C_4}{100} - 100 \right) \Big/ 4$ because there are four quarters. Then adjusted chain relatives are C_1, $C_2 - b$, $C_3 - 2b$ and $C_4 - 3b$ for first to fourth quarters Q_1, Q_2, Q_3 and Q_4. If the total adjusted chain relatives be G then multiply the adjusted chain relatives by $\frac{400}{G}$ to get seasonal indices for the quarters in case of multiplicative model of time series data. This method is illustrated clearly in the worked out example number 5 later.

6.2.4 Measurement of Cyclical Variations

Business cycles are the most important type of fluctuations in economic data. They are most difficult type of economic fluctuations to measure because successive cycles vary widely in timing, amplitude and pattern and cyclical fluctuations are inextricably mixed up with irregular factors. For this it is impossible to construct meaningful cycle indices or curves similar to those that have been developed for determining trends and seasonal indices. Cyclical fluctuations can be determined by the following residual method.

Residual method

This method is commonly used to determine cyclical fluctuations. Here trend (T_t) and seasonal variation (S_t) are eliminated from the time series data y_t to get $C_t \times I_t$ i.e. product of cyclical and irregular variations considering multiplicative model. In case of additive model, $C_t + I_t$ can be obtained from the time series data y_t by eliminating trend

(T_t) and seasonal variation (S_t). In the first case $C_t \times I_t = \frac{y_t}{T_t \times S_t} = \frac{T_t \times S_t \times C_t \times I_t}{T_t \times S_t}$ and in the second case $C_t + I_t = y_t - T_t - S_t$. Then irregular component (I_t) is removed from the residuals ($C_t \times I_t$) or ($C_t + I_t$) by smoothing using moving averages of a suitable period or by fitting a suitable curve. Though this method is laborious, it gives more or less accurate results.

Use of Cycle Analysis

There is some difference between the forecasting by cycle analysis (i.e., cyclical movement analysis) and that by trend and seasonal variation analysis. The forecasts by cycle analysis must be related to the forecasts of general business conditions. Since the business conditions of a particular industry will be affected in general by the general business conditions, it is a difficult job. Only a partial success has been obtained in forecasting the turning points of business conditions.

6.2.5 Business Forecasting

Successful business activity requires an accurate forecasting of future business conditions upon which decisions regarding production, inventories, cost etc. depend. To eliminate guess work, modern statistical methods are employed as useful tools of forecasting. The methods require a knowledge of the past and present conditions and a proper analysis of time series data, isolating them into four components T_t, S_t, C_t and I_t.

There are three important methods used in business forecasting,

(i) Economic rhythm method

(ii) Specific historical analysis

and (iii) Cyclical sequence method.

(i) Economic rhythm method

The time series is first analysed to study trend, seasonal variation and cyclical variation by the methods given earlier in this chapter. The trend is projected for a number of years ahead by drawing a free hand curve or by drawing a mathematical curve. Then an estimate of future cyclical movement is superimposed on the trend. Superimposition of cyclical component for a number of years into the future is done on the assumption that a period of prosperity with a certain value as the product of duration of prosperity and its intensity will be followed by a period of depression with an approximately equal product of duration of depression and its intensity. Then multiply the product of trend and cyclical factors by an appropriate seasonal index to get the result.

(ii) Specific historical analysis

A specific business event in the past having some similarities with the present situation is chosen and studied. On the basis of this study forecast is made about future business conditions assuming that the same set of conditions hold in the future.

(iii) Cyclical sequence method

A series correlated with the business series which we want to predict with a certain time lag is considered. From the two series trend (T_t) and seasonal component (S_t) are eliminated first and the approximate time lag between the two series is determined graphically or otherwise. The lag regression equation $y_t = a_0 + a_1 x_{t-k}$ may be used for predicting the future conditions. On the basis of past figures x_{t-k} the value of y_t can be predicted.

6.3 Worked out examples

Example 1.

The following table gives the number of workers employed in an industry between 1985 and 1996. Calculate the trend values using the moving average method.

Year	No. of workers	Year	No. of workers	Year	No. of workers
1985	463	1989	513	1993	518
1986	495	1990	494	1994	514
1987	479	1991	493	1995	540
1988	481	1992	528	1996	500

Also determine moving average values with period 4 years.

Solution : First part

Here the peak years are 1986, 1989, 1992 and 1995 where periods between peaks of values are 3, 3, 3 years. Periods between toughs of values are 4 and 3 years since values are lowest in the years 1987, 1991 and 1994. Moving average period is the average of these periods of peaks and toughs i.e. it is $\frac{1}{5}(3+3+3+4+3) = \frac{16}{5} = 3 \cdot 2$ whose nearest integer is 3. So consider the period of moving average as 3 years. First consider 3 yearly moving total and place it in the centre and divide it by 3 to get 3 yearly moving average (or trend) in the next column.

Year	No. of Workers	3-Yearly moving total	3 Yearly moving average = Trend
1985	463	–	–
1986	495	1437	479
1987	479	1455	485
1988	481	1473	491
1989	513	1488	496
1990	494	1500	500
1991	493	1515	505
1992	528	1539	513
1993	518	1560	520
1994	514	1572	524
1995	540	1554	518
1996	500	–	–

Then 3 yearly moving average values or trend values for the years from 1986, 1987, ..., 1995 are 479, 485, ..., 518 have been obtained in the last column.

Second part

For 4-yearly moving average values proceed as below. First obtain 4-yearly moving totals and place them in the centres and then in the next column centered moving totals are obtained to get the moving totals corresponding to different years. Then in the last column divide these centred moving totals by 8 to get moving averages.

Year (1)	No. of Workers (2)	4-Year moving total (3)	2-Yr moving total (4)	4-Years centered moving average (5) = (4) ÷ 8
1985	463	–	–	–
1986	495	1918	–	–
1987	479	1968	3886	485.75
1988	481	1967	3935	491.875
1989	513	1981	3948	493.5
1990	494	2028	4009	501.125
1991	493	2033	4061	507.625
1992	528	2053	4086	510.75
1993	518	2100	4153	519.125
1994	514	2072	4172	521.5
1995	540		–	–
1996	500	–	–	–

The 4 yearly moving averages i.e., trend values are 485.75, 491.875, 493.5, 501.125, 507.625, 510.75, 519.125 and 521.5 for the years 1987, 1998, 1989, 1990 1991, 1992, 1993 and 1994 respectively.

Example 2.

Fit a straight line trend and a parabolic trend using least squares method from the following data.

Year :	1981	1982	1983	1984	1985
Production ('000 tons) :	22	21	23	22	24

Estimate the production of 1990 by using both types of trends. Also obtain quarterly trend equation in both the cases.

Solution

Let $y = a + bt$ be the best fitted straight line to the given data and $y = c + dt + et^2$ be the best fitted parabola to the given data where *y* is production in thousand tons and *t* is time (year) where origin is considered at 1983 and unit of *t* is 1 year. Here *a* and *b* are obtained from normal equations :

$$\left.\begin{aligned} \Sigma y &= na + b\Sigma t \\ \Sigma ty &= a\Sigma t + b\Sigma t^2 \end{aligned}\right\} \quad (1)$$

Also, c, d and e are obtained from normal equations :

$$\left.\begin{aligned} \Sigma y &= nc + d\Sigma t + e\Sigma t^2 \\ \Sigma ty &= c\Sigma t + d\Sigma t^2 + e\Sigma t^3 \\ \Sigma t^2 y &= c\Sigma t^2 + d\Sigma t^3 + e\Sigma t^4 \end{aligned}\right\} \quad (2)$$

In both the cases n = 5.

Year	y = Production (.000 tons)	t = Year–1983	t^2	ty	t^3	t^4	t^2y
1981	22	–2	4	–44	–8	16	88
1982	21	–1	1	–21	–1	1	21
1983	23	0	0	0	0	0	0
1984	22	1	1	22	1	1	22
1985	24	2	4	48	8	16	96
Total	112	0	10	5	0	34	227

From normal equations (1) i.e. $112 = 5a$, $5 = 10b$ we get $a = 22.4$ and $b = 0.5$. So straight line trend is $y = 22.4 + 0.5t$, origin = 1983 and unit of t = 1 year. Then the estimate of production for 1990, for which t = 7, is $y = 22.4 + 0.5 \times 7 = 25.9$ thousand tons.

From normal equations (2) i.e. 112 = 5c + 10e, 5 = 10d and 227 = 10c + 34e we get c = 21.97, d = 0.5 and e = 0.214. Then the trend equation is y = 21.97 + 0.5t + $0.214t^2$, origin 1983 and unit of *t* is 1 year. Then estimate of production of 1990 is y = 21.97 + 0.5 × 7 + 0.214 × 49 = 35.956 thousand tons.

Linear quarterly trend equation is $y = \frac{22.4}{4} + \frac{0.5(t + \frac{1}{2})}{16}$ i.e., *y* = (5.6 + .016) + .03t i.e., y = 5.616 + .03*t*, origin is at 1983 third quarter and unit of *t* = 1 quarter.

Parabolic trend equation is $y = \frac{21.97}{4} + \frac{0.5(t + \frac{1}{2})}{16} + \frac{0.214(t + \frac{1}{2})^2}{64}$

i.e., $y = 5.49 + .03(t + 0.5) + .0033\left(t^2 + t + \frac{1}{4}\right)$

= (5.49 + .015 + .0008) + (.03 + .0033)t + $.0033t^2$

or *y* = 5.5058 + .0333 t + .0033 t^2 origin of *t* is at 1983 third quarter and unit of *t* = 1 quarter.

Example 3.

Fit a straight line trend and a parabolic trend using least squares method from the following data :

Year :	1981	1982	1983	1984	1985	1986
Production ('000 tons) :	22	21	23	22	24	23

Estimate the production of 1990 by using both types of trends.

Solution :

Let *y* = a + bt ... (i)

origin at *t* = 1983.5 and unit of *t* = $^1/_2$ year, be the best fitted trend line.

Also, let y = c + dt + et^2 ... (ii)

origin at *t* = 1983.5, unit of *t* = $^1/_2$ year, be the best fitted parabola.

To get *a* and *b* from the first equation (i) and c, d and e from the second equation (ii) necessary normal equations are :

$\Sigma y = na + b\Sigma t$, $\Sigma ty = a\Sigma t + b\Sigma t^2$ ... ... (1)

and $\Sigma y = nc + d\Sigma t + e\Sigma t^2$, $\Sigma ty = c\Sigma t + d\Sigma t^2 + e\Sigma t^3$,

$\Sigma t^2 y = c\Sigma t^2 + d\Sigma t^3 + e\Sigma t^4$ ... ... (2)

In both cases n = 6.

Year	y = Production ('000 tons)	t = 2 (Year–1983.5)	t^2	t_y	t^3	t^4	t^2y
1981	22	–5	25	–110	–125	625	550
1982	21	–3	9	–63	–27	81	189
1983	23	–1	1	–23	–1	1	23
1984	22	1	1	22	1	1	22
1985	24	3	9	72	27	81	216
1986	23	5	25	115	125	625	575
Total	135	0	70	13	0	1414	1575

From normal equations (1) 135 = 6a, 13 = 70b we get a = 22.5, b = 0.1857. So the trend equation is $y = 22.5 + 0.1857\,t$, origin is at the end of 1983, unit of $t = \frac{1}{2}$ year. The estimate of production in 1990 is $y = 22.5 + 0.1857 \times 13 = 24.914$ thousand tons, since for the year 1990, $t = 2(1999 - 1983.5) = 2 \times 6.5 = 13$.

From normal equations (2), 135 = 6c + 70e, 13 = 70d, 1575 = 70c + 1414e, we get c = 22.5, d = 0.1857, e = 0. So former fitted straight line is the best fitted straight line and not the parabola. This trend line is $y = 22.5 + 0.1857t$ and ultimately the estimate of production in 1990 is 24.914 thousand tons as obtained above.

Example 4

Obtain the seasonal fluctuations of four quarters from the following data applying moving average method and deseasonalise the time series.

		Quarters		
Years	I	II	III	IV
1974	72	68	80	70
1975	76	70	82	74
1976	72	66	84	80
1977	76	74	84	78

Solution

First arrange the data in chronological order and determine the period of moving average by the average of periods of peaks and troughs.

Year	Quarter	Yield (y)	3-year moving total	Trend Value = T = 3 year moving average	$\frac{y}{T} \times 100$
1974	Q_1	72	–	–	–
	Q_2	68	220	73.33	92.73
	Q_3	80	218	72.67	110.09
	Q_4	70	226	75.33	92.92
1975	Q_1	76	216	72	105.56
	Q_2	70	228	76	92.11
	Q_3	82	226	75.33	108.85
	Q_4	74	228	76	97.37
1976	Q_1	72	212	70.67	101.88
	Q_2	66	222	74	89.19
	Q_3	84	230	76.67	109.56
	Q_4	80	240	80	100
1977	Q_1	76	230	76.67	99.13
	Q_2	74	234	78	94.87
	Q_3	84	236	78.67	106.78
	Q_4	78	–	–	–

Period of moving average $= \frac{(2+2+4+4)+(2+2+4+4)}{8} = 3$ Quarters.

Let the time series model be a multiplicative one.

Calculation of seasonal variation

Year \ Quarter	Q_1	Q_2	Q_3	Q_4	Total
1974	–	92.73	110.09	92.92	
1975	105.56	92.11	108.85	97.37	
1976	101.88	89.19	109.56	100	
1977	99.13	94.87	106.78	–	
Total	306.57	368.9	435.28	290.29	
Average :	102.19	92.225	108.82	96.763	399.998
Adjusted Seasonal index	102.19	92.225	108.82	96.763	

Seasonal indices of 4 quarters I, II, III and IV are 102.19, 92.225, 108.82, 96.763 respectively.

Year and quarter	1974				1975			
	Q_1	Q_2	Q_3	Q_4	Q_1	Q_2	Q_3	Q_4
Yield	72	68	80	70	76	70	82	74
Seasonal index (%)	102.19	92.225	108.82	96.763	102.19	92.225	108.82	96.763
Deseasonalised yield	70.46	73.73	73.52	72.34	74.37	75.90	75.35	76.48

Year and quarter	1976				1977			
	Q_1	Q_2	Q_3	Q_4	Q_1	Q_2	Q_3	Q_4
Yield	72	66	84	80	76	74	84	78
Seasonal index (%)	102.19	92.225	108.82	96.763	102.19	92.225	108.82	96.763
Deseasonalised yield	70.46	71.56	77.19	82.68	74.37	80.24	77.19	80.61

Example 5

Calculate seasonal indices by the method of link relatives from the following data.

Year / quarter	1985	1986	1987	1988
Q_1	85	96	100	110
Q_2	70	75	82	88
Q_3	64	73	76	82
Q_4	69	90	95	103

Solution :

L_t = Link relatives of any quarter = $\dfrac{\text{Value of that quarter } (y_t)}{\text{Value of Previous year } (y_{t-1})} \times 100$

Let averages of link relatives of quarters Q_1, Q_2, Q_3 and Q_4 be A_1, A_2, A_3 and A_4 respectively.

C_1 = chain relative for Q_1 = 100

C_2 = chain relative for $Q_2 = \frac{A_2 \times C_1}{100}$,

C_3 = chain relative for $Q_3 = \frac{A_3 \times C_2}{100}$,

C_4 = chain relative for $Q_4 = \frac{A_4 \times C_3}{100}$,

correction factor b = $\left[\frac{(\text{Average of link relative of } Q_1) \times (\text{chain relative of } Q_4)}{100} - 100\right] \Big/ 4$.

Adjusted chain relative of $Q_1 = C_1$

Adjusted chain relative of $Q_2 = C_2 - b$

Adjusted chain relative of $Q_3 = C_3 - 2b$

Adjusted chain relative of $Q_4 = C_4 - 3b$

If total adjusted chain relative for all the quarters is G then correction factor $= \frac{400}{G}$.

Seasonal index for quarter Q_i = (Adjusted chain relative of Q_i) $\times \frac{400}{G}$ for i = 1, 2, 3, 4.

Calculation of seasonal indices by method of link relatives

Quarter / Year	Q_1	Q_2	Q_3	Q_4	Total
1975	–	82.35	91.43	107.81	
1976	139.13	78.12	97.33	123.29	
1977	111.11	82	92.68	125	
1978	115.79	80	93.18	125.61	
Average	122.01 $= A_1$	80.62 $= A_2$	93.66 $= A_3$	120.43 $= A_4$	
Chain relative	$C_1 = 100$	$C_2 = 80.62$	$C_3 = 75.51$	$C_4 = 90.94$	
Adjusted chain relative	$C_1 = 100$	$C_2 - b$ $= 77.88$	$C_3 - 2b$ $= 77.03$	$C_4 - 3b$ 82.72	330.63 = G
Seasonal indices	120.98	94.219	84.722	100.075	

Trend correction $= b = \left(\frac{A_1 \times C_4}{100} - 100 \right) \Big/ 4 = \left(\frac{122.01 \times 90.94}{100} - 100 \right) \Big/ 4$

$= \frac{1}{4}(110.9559 - 100) = \frac{1}{4} \times 10.9559 = 2.74$

Correction factor $= \frac{400}{G} = \frac{400}{330.63} = 1.2098$

So seasonal indices for 4 quarters Q_1, Q_2, Q_3 and Q_4 are 120.98, 94,219, 84.722 and 100.075 respectively.

6.4 Summary

This chapter consists of definition of time series with examples, time series analysis and necessity of it, factors for the change in time series data, standard time series models, Components of time series, Adjustment in time series data, measurement of secular trend by the free hand curve fitting method, moving average method and mathematical curve fitting method including their merits and demerits and also derivation of monthly or quarterly trend equation from annual trend equation, Measurements of Seasonal variation by ratio to moving average method, ratio to trend method and link relative method, measurement of cyclical variation by residual method, business forecasting and worked out examples.

6.5 Exercises

1. What do you mean by time series? Give two examples.
2. What is the necessity of time series analysis?
3. Describe the different components of a time series.
4. Describe the factors responsible for change in time series data.
5. What is trend? Describe the different methods of determining trend.
6. What is seasonal variation? Describe the different methods of determining seasonal variation.
7. What is cyclical fluctuation? How can it be determined?
8. Explain the nature of cyclical variations in a time series. How do seasonal variations differ from them?
9. What is business forecasting? How can it be done?

10. Work out the trend values from the following data :

Year	'63	'64	'65	'66	'67	'68	'69	'70	'71	'72	'73	'74
Production of iron ore in 1000 tonnage	110	125	118	134	121	132	145	155	159	148	162	159

When period of moving average (i) is known to be 4 years (ii) is unknown.

11. Fit (i) a straight line trend and (ii) a second degree polynomial to the following time series data

(i)

Year	1980	1981	1982	1983	1984	1985	1986	1987
Average production per month ('000 tons)	20	22	21	23	22	24	23	25

(ii)

Year	1980	1981	1982	1983	1984	1985	1986
Average production per month ('000 tons)	20	22	21	23	22	24	23

Determine the estimate of average production per month in 1990 in both the time series data.

12. Fit an exponential trend equation to the following data of production of sugar to the following data :

(i)

Year :	1981	1982	1983	1984	1985
Production ('000 tons)	81	89	98	109	120

(ii)

Year	1981	1982	1983	1984	1985	1986
Production ('000 tons)	81	89	98	109	120	133

Estimate the production of sugar in 1990 in both time series data.

13. Calculate the seasonal indices in the case of following quarterly data in certain units using (i) ratio to moving average method, (ii) ratio to trend method, (iii) link relative method :

Total production of paper ('000 tons)

Year \ Quarter	Q_1	Q_2	Q_3	Q_4
1975	37	38	37	40
1976	41	34	25	31
1977	35	37	35	41

14. Trend equation for certain production data is $y = 240 + 36t$ (origin = 1980, unit of t = 1 year, y = annual production in thousand tons). Estimate the trend value for March 1982.

15. Sales of XYZ company rose from Rs. 39,45,000 to Rs. 46,21,000 from 2nd quarter to third quarter of 2008. The seasonal indices for the quarters are 103 and 150 respectively. The owner of the company holds that it is a losing company. Do you think so? If so, explain how.
16. Describe the method of moving average and discuss its role in isolating trend movement in time services.
17. Obtain the three year moving averages for the following data :

Years	2001	2002	2003	2004	2005	2006	2007	2008	2009
Annual Sales (in crore)	36	43	43	34	44	54	34	24	14

18. What do you mean by seasonal variation? When write a few examples the utility of such a study.
19. Which of the components of time series is applicable to the following cases? (a) Fire in a factory, (b) The increase in demand for wool in the winter season, (c) The decline in the mortality rate due to advancement in medical science, (d) The increase in the percentage of literacy in a developing country, (e) A period of depression, (f) Number of books borrowed from National Library increases during the weekend, (g) Average number of eggs per laying hens increases in the month of June, (h) Increased production of pig iron over a long period of time.
20. What is the time sries? What is the need to analyse time series? Enumerate the different methods of finding the secular trend.
21. Deseasonalise the following profit data and interpret the results.

Quarter	I	II	III	IV
Profit (Rs. 00)	25.8	28.2	26.5	38.6
Seasonal index	82.0	98.0	85.0	135.0

22. Fit a trend line to the following data by the free hand method :

Year	2001	2002	2003	2004	2005	2006	2007	2008
Sales of a firm (in million Rs.)	63	65	67	64	68	65	70	68

23. What do you mean by a time series? What are its components? Briefly discuss the objectives of time series anslysis.
24. Fit an exponential trend $y = ab^t$ to the following data by the method of least squares. Hence obtain the annual compound rate of growth of sales of the company

Years	2002	2003	2004	2005	2006	2007	2008
Sales	87	97	113	129	202	195	193

25. Discuss the various methods for measuring seasonal fluctuations from time series data.

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

Year	2009	2010	2011	2012	2013	2014	2015
Sales (000 tonnes)	110	121	116	136	140	157	170

27. XYZ company estimates its average monthly sales in a particular year to be Rs. 20,00,000. The seasonal indices of the sales data are given below :

Month :	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
Seasonal Index :	78	75	100	126	138	121	101	104	99	103	80	75

Assuming that there is no trend, use the above information to draw up a monthly sales budget for the company.

[**Hits :** Seasonal Effect = Seasonal index ÷ 100. Estimated Sales = Seasonal effect × 20,00,000.]

28. Obtain the five year moving average for the following data :

Year :	2000	2001	2002	2003	2004	2005	2006	2007	2008
Annual Sales (in Rs. crore)	36	43	43	34	44	54	34	24	14

29. What do you mean by business forecasting? Indicate its uses.

30. You are given the population figures of India.

Census year (x) :	1911	1921	1931	1941	1951	1961	1971
Population (in crores) :	25.0	25.1	27.9	31.9	36.1	43.9	54.7

Fit an exponential trend $y = ab^x$ to the above data by the method of least squares and find the trend values. Estimate the population in 1981.

6.6 Suggested Readings

1. Croxton, F. E. and Cowden, D. J., *Applied General Statistics,* Prentice-Hall 1976 and Prentice-Hall of India 1969.
2. Mills, F. C., *Statistical Methods,* H. Holt 1955.
3. Goon, A. M., Gupta, M. K. And Dasgupta, B., *Fundamentals of Statistics,* Vol. II, The World Press Private Limited, Kolkata 2002.
4. Kendall, M. G. and Stuart, A., *The Advanced Theory of Statistics, Vol 3,* Charle's Griffin 1966.

Unit 7 ❑ Statistical Quality Control

Structure

7.0 Objectives

As competition is growing it has become necessary for a businessman to keep a continuous watch over the quality of the goods produced. If the consumer is satisfied with quality, price etc, a goodwill of the company will be produced, resulting in an

increase in sales and profits. Otherwise, it will be difficult for the manufacturer to survive in the market.

As needs for maintaining or improving the standard of the quality of goods produced grow with the increase in competition, quality control is becoming essential to the manufacturer. Statistical quality control involves the statistical analysis of the inspection data, which involves sampling and the principle of normal distribution. These techniques were developed by A. Shewhart.

Presently the statistical quality control is used virtually in all kinds of industry. In fact, it has become an integral and permanent part of management control. Consistency in quality standard is more desirable than maintaining absolute standard due to cost, labour and time.

7.1 Introduction

Statistical Quality Control (SQC) is the statistical technique used to maintain uniform quality of products in a continuous flow of manufacturing process. Here quality is a characteristic of the product, which is of interest. In any manufacturing process the products produced, are not of exactly same quality. Certain variations are inevitable and are called allowable variations. Other than these there may be other type of variations, known as preventable or assignable variations. These variations occur due to some defects in the machine or due to bad raw materials etc.

Process control is the procedure in the production process to control and to maintain a satisfactory quality of goods produced by separating allowable variation from preventable variation and taking remedial action to prevent the occurrences of preventable variation in the process. This is usually done by drawing control charts of suitable characteristics, which ensure that the products to be manufactured, conform to a specified quality.

By product control or lot control we mean the sampling inspection plan by which a lot of items, conforming to a specified level of quality is accepted and a lot of non-conforming items is rejected. The decision is made through sampling inspection which helps buyers to dispense with 100% inspection of the lot of items.

The two problems, process control and product control, are distinct, because even when the process is in control in individual lot of items, may not be of satisfactory quality.

Variation in the quality of products in a manufacturing process are attributed to two different types of causes. One type is assignable causes which are relatively large,

occurred due to specific causes like mistake of an inexperienced person, faulty machines etc. The second type is non-assignable or random causes for which variations are inevitable. Then the variations are correspondingly called assignable variation and random variation.

7.2 Types of quality measure

Some quality characteristics are measurable and can be expressed numerically and they are called variables. Some other quality characteristics are not measurable and cannot be expressed numerically and they are called attributes. For example, length of a bolt or diameter of a nut are variables but defective articles or defects of similar objects are attributes.

7.2.1 Rational sub-group and control chart

Divide the sample products into homogeneous sub-groups in which variations of elements within a sub-group may be attributable entirely due to chance causes. Assignable variations, due to variation from one sub-group to another exist. So, for each sub-group there exists chance variations but not assignable variations. These sub-groups are called rational sub-groups.

For each rational sub-group consider T as a statistic for the quality characteristic of the sample products. It is either a variable or an attribute. The values of T are the estimates of some population parameter θ. If T is normally distributed with mean μ_T and variance σ_T^2 then $P(\mu_T - 3\sigma_T \leq T \leq \mu_T + 3\sigma_T) = P(|Z| \leq 3) = 0.9973$ approximately where Z is a normal deviate (i.e., standard normal variable). So we see that almost all values of T lie within 3σ limits i.e. between $\mu_T - 3\sigma_T$ and $\mu_T + 3\sigma_T$ and 0.27% of values only lie outside 3σ limits. If any sample point T lies outside the control limits $\mu_T - 3\sigma_T$ and $\mu_T + 3\sigma_T$ then it is suspected that assignable causes of variation are playing a role in the process. These horizontal lines are drawn in a graph paper to represent Lower Control Limit (LCL) $\mu_T - 3\sigma_T$, Central Line (CL) μ_T and Upper Control Limit (UCL) $\mu_T + 3\sigma_T$. Then the values of T for each sub-group or sample are plotted on the graph paper at equal interval considering sample numbers along horizontal axis and statistic T along vertical axis.

7.2.2 Control chart of variables

When quality characteristics are measurable and can be represented numerically as a continuous variable (x) consider control chart of mean (μ), standard deviation (σ) and range (R). For *k* independent samples or sub-groups, each of size *n* with means μ_1, μ_2,

..., μ_k, standard deviations, σ_1, σ_2, ..., σ_k and ranges R_1, R_2, ..., R_k to examine whether the process is in control, we need to see whether the means μ_1, μ_2, ..., μ_k, the standard deviations (s.d.'s) σ_1, σ_2, ..., σ_k and ranges R_1, R_2, ..., R_k are the same. For manufactured articles subject to random variation only the variable (x) may be supposed to be normally distributed. This is due to central limit theorem and the fact that the value of the variable can be considered as the sum of a large number of independent components each of which contributes a relatively negligible proportion of the total variability. The four types of situations may arise : (a) the process is in control when mean and s.d. (or R) are in control, (b) the mean is out of control but not the s.d. (or R), (c) the s.d. (or R) is out of control but not the mean and (d) the mean and s.d. (or R) are out of control i.e., the process is out of control in (d).

The statistic corresponding to μ and σ are $\bar{x}$ and s. Though R is inferior to *s* from theoretical point of view, R is simpler to compute than *s*. Hence in quality control charts for $\bar{x}$ and R are often used in preference to charts for $\bar{x}$ and *s*.

7.2.3 Control chart for mean

Case I : Standard given

For a sample of size n per sub-group and for a stable system $\bar{x}$ is taken as a statistic to estimate μ for which $E(\bar{x}) = \mu$ and $v(\bar{x}) = \frac{\sigma^2}{n}$ where we assume that observations in each sub-group are independent. As standard value of μ and σ are known as μ' and σ' the control chart for mean will be given by

$$LCL = \mu' - 3\frac{\sigma'}{\sqrt{n}} = \mu' - A\sigma', CL = \mu', UCL = \mu' + 3\frac{\sigma'}{\sqrt{n}} = \mu' + A\sigma'$$

where $A = \frac{3}{\sqrt{n}}$.

Case II : Standards not given

Let there be *k* sub-groups and let the successive sample means be $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$, the successive standard deviations be s_1, s_2, ..., s_k and successive ranges be R_1, R_2, ..., R_k. As μ and σ are not known they are estimated from the samples. Let the means of sample means, standard deviations and ranges be $\bar{\bar{x}} = \frac{1}{k}\sum_{i=1}^{k}\bar{x}_i$, $\bar{s} = \frac{1}{k}\sum_{i=1}^{k}s_i$ and $\bar{R} = \frac{1}{k}\sum_{i=1}^{k}R_i$ and they will give the estimates of μ and *s* from the relations.

$E(\bar{\bar{x}}) = \mu, E(\bar{s}) = c_2\sigma$ and $E(\bar{R}) = d_2\sigma, c_2$ and d_2 being function of n, considering the distribution of the variable x as normal with mean μ and variance σ^2. So then the control chart for mean will be given by

$$LCL = \bar{\bar{x}} - \frac{3\bar{s}}{c_2\sqrt{n}} = \bar{\bar{x}} - A_1\bar{s},\ CL = \bar{\bar{x}} \text{ and } UCL = \bar{\bar{x}} + \frac{3\bar{s}}{c_2\sqrt{n}} = \bar{\bar{x}} + A_1\bar{s}$$

in terms of means and standard deviations of sub-groups, where $A_1 = 3/\left(c_2\sqrt{n}\right)$

and $$LCL = \bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}} = \bar{\bar{x}} - A_2\bar{R},\ CL = \bar{\bar{x}} \text{ and } UCL = \bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}} = \bar{\bar{x}} + A_2\bar{R}$$

in terms of means and ranges of sub-groups, where $A_2 = \frac{3}{d_2\sqrt{n}}$. The values of A, A_1 and A_2 can be obtained from the table of factors useful in construction of control chart given in any standard book of quality control.

7.2.4 Control chart for standard deviation

Case 1 : Standard given

As variable x is normally distributed with mean μ and variance σ^2, $E(s) = c_2\sigma$ $v(s) = \left(\frac{n-1}{n} - c_2^2\right)\sigma^2$ and where c_2 is a function of n. If the standard value of σ is σ' the control chart will be based on

$$LCL = c_2\sigma' - 3\sigma'\sqrt{\frac{n-1}{n} - c_2^2} = B_1\sigma', CL = c_2\sigma',$$

$$UCL = c_2\sigma' - 3\sigma'\sqrt{\frac{n-1}{n} - c_2^2} = B_2\sigma',$$

where $B_1 = c_2 - 3\sqrt{\frac{n-1}{n} - c_2^2}$ and $B_2 = c_2 + 3\sqrt{\frac{n-1}{n} - c_2^2}$.

Case II : Standard not given

When standard value of σ is not given then estimate σ by $\frac{\bar{s}}{c_2}$ where $\bar{s} = \frac{1}{n}\sum_{i=1}^{k} s_i$

where s_i is the standard deviation of x from the i-th sub-group for i = 1, 2, ..., k and the control chart will be based on

$$LCL = \bar{s} - 3\frac{\bar{s}}{c_2}\sqrt{\frac{n-1}{n} - c_2^2} = B_3\bar{s}, CL = \bar{s}, UCL = \bar{s} + 3\frac{\bar{s}}{c_2}\sqrt{\frac{n-1}{n} - c_2^2} = B_4\bar{s}$$

where $B_3 = 1 - \frac{3}{c_2}\sqrt{\frac{n-1}{n} - c_2^2}$ and $B_4 = 1 + \frac{3}{c_2}\sqrt{\frac{n-1}{n} - c_2^2}$

The values of B_1, B_2, B_3 and B_4 will be obtained from a table of factors useful in the construction of control chart from any Quality control's book.

7.2.5 Control chart for range

Case I : Standard given

For a normally distributed variable x we have $E(R) = d_2\sigma$ and var $(R) = D\sigma$ where d_2 and D are functions of n. If standard value of σ is given to be σ' then the chart for R will be based on

$$LCL = d_2\sigma' - 3D\sigma' = D_1\sigma',\ CL = d_2\sigma',\ UCL = d_2\sigma' + 3D\sigma' = D_2\sigma'$$

where $D_1 = d_2 - 3D$ and $D_2 = d_2 + 3D$

Case II : Standard not given

When standard value of σ is not given then it is estimated by $\frac{\overline{R}}{d_2}$ where d_2 is function of n. The chart will be based on

$$LCL = \overline{R} - 3\frac{D}{d_2}\overline{R} = D_3\overline{R},\ CL = \overline{R},\ UCL = \overline{R} + 3\frac{D}{d_2}\overline{R} = D_4\overline{R}$$

where $D_3 = 1 - 3\frac{D}{d_2}$ and $D_4 = 1 + 3\frac{D}{d_2}$.

The values of D_1, D_2, D_3 and D_4 will be obtained from the table of factors in construction of control charts of any books in SQC.

7.2.6 Control chart for attribute

When the quality characteristic is an attribute each item of the sample drawn is recorded as defective or non-defective. Consider P and p as the proportion of defectives in the population and in the sample respectively so that $p = \frac{d}{n}$ where d = number of defectives in the sample, n being the size of the sample.

7.2.7 Control chart for number of defectives

Case 1 : Standard given

Random samples each of size n, have been drawn without replacement from an infinite population and distribution of d is binomial with parameters n and P so that $E(d) = nP$ and $v(d) = nP(1–P)$ if the process is in control. If the standard is given

i.e. given value of P is say P', then control chart of number of defectives is constructed on the basis of

$$LCL = nP' - 3\sqrt{nP'(1-P')},\ CL = nP',\ UCL = nP' + 3\sqrt{nP'(1-P')}.$$

Case II : Standard not given

If the standard i.e. value of P is not given then P is estimated by $\bar{p} = \frac{1}{k}\sum_{i=1}^{k} p_i$ where P_i is the proportion of defectives in the i-th sub-group for i = 1, 2, ..., k, then control chart of number of defectives is constructed on the basis of

$$LCL = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})},\ CL = n\bar{p},\ UCL = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}.$$

Anyhow if LCL is negative then it should be taken as zero.

7.2.8 Control chart for fraction defective

Case I : Standard given

For the fraction defective of the sample i.e., for p, E(p) = P, $v(p) = \frac{P(1-P)}{n}$. When standard value of P is given, say P' then control chart of fraction defective can be constructed on the basis of

$$LCL = \bar{P} - 3\sqrt{\frac{\bar{P}(1-\bar{P})}{n}},\ CL = \bar{P} \text{ and } UCL = \bar{P} + 3\sqrt{\frac{\bar{P}(1-\bar{P})}{n}}.$$

Case II : Standard not given

When standard value of P is not given then P is estimated by $\bar{p} = \frac{1}{k}\sum_{i=1}^{k} p_i$, p_i being the fraction defective in the i-th sub-group, i = 1, 2, ..., k. the control chart of fraction defective is constructed on the basis of

$$LCL = \bar{P} - 3\sqrt{\frac{\bar{P}(1-\bar{P})}{n}},\ CL = \bar{P} \text{ and } UCL = \bar{P} + 3\sqrt{\frac{\bar{P}(1-\bar{P})}{n}}.$$

If LCL is negative, consider it to be zero.

7.2.9 Control chart for number of defects (c-chart)

Every defective article contains one or more defects. In almost all manufactured articles there are numerous opportunities (say n) for defects to occur but the chance (say p) of a defect at a single opportunity is very small (i.e., $n \to \infty$ and $P \to 0$) so that $np = \lambda$ constant. Hence number of defects (c) may be supposed to follow Poisson distribution with parameter λ. Then $E(c) = \lambda$, var (c) = $V(c) = \lambda$.

Case I : Standard given

If standard value of λ is given as λ' then the control chart of the number of defects can be constructed on the basis of

$$\text{LCL} = \lambda' - 3\sqrt{\lambda'}, \text{CL} = \lambda' \text{ and } \text{UCL} = \lambda' + 3\sqrt{\lambda'}.$$

Case II : Standard not given

If standard value of λ is not given then we estimate λ by $\hat{\lambda} = \bar{c} = \frac{1}{k}\sum_{i=1}^{k} c_i$ where c_i is the value of c i.e., number of defects for the i-th sub-group, i = 1, 2, ..., k. The control chart for number of defects is constructed from

$$\text{LCL} = \bar{c} - 3\sqrt{\bar{c}}, \text{CL} = \bar{c} \text{ and } \text{UCL} = \bar{c} + 3\sqrt{\bar{c}}.$$

Anyhow if LCL is negative then consider it to be zero.

7.3 Uses of control charts

Control chart is used to determine whether past operations of a process have been in control or not, and use that information as a basis for action in future. Lack of control with regard to past production is indicated by points falling outside the control limits LCL and UCL. Then assignable causes of variation exist and one may correct the production process so as to avoid future troubles. Control chart is more useful for taking future actions.

7.4 Assumptions used in S. Q. C.

1. The process has only two states : (1) operating state and (2) failure state, and initially it is in operating state.
2. Process can go from state 1 to state 2 but after repair it can come back from state 2 to state 1.
3. Quality characteristic is considered to be normally distributed.
4. Observations, considered, occur independently.

7.5 Advantages and limitations of S.Q.C.

1. This reduces cost of inspection due to inspection of a fraction of output.
2. It reduces work load and as a result efficiency of work increases.

3. This is easy to apply and this control can be operated by a person with little training after the system is established.
4. Through S.Q.C. an objective check is done on the quality of the product to conform to the laid down standards and specifications.
5. Through S.Q.C. it can be known whether the process is in control or not. If it has gone out of control then necessary remedial actions can be taken so that the production of goods of standard below the unit should be stopped to avoid waste of material, time and money.
6. If the producers follow efficient and strict S.Q.C. system then that will increase their goodwill because the users may then rely on their products and may not resort to a thorough check.
7. The quality of the product can be defended before any Government enquiry on the basis of S.Q.C. records.
8. As a byproduct of S.Q.C., good deal of statistical data are made available and these data can be used by the management to evaluate plant and machinery as well as technical staff.

7.6 Specification and Tolerance Limits

The limits for an individual product set forth by the authority for satisfying quality to the customer before the production starts are called specification limits. Limits for an individual product arising in the existing state of production are called the natural tolerance limits or simply, tolerance limits. If μ and σ are the process average and standard deviation respectively, then the limits $\mu \pm 3\sigma$ which include, on the average, 99.73% out of all the items are called natural tolerance limits of the process. When μ and σ are unknown they are estimated by $\bar{\bar{x}}$ and $\frac{\bar{s}}{c_2}$ or $\frac{\bar{R}}{d_2}$ as stated earlier.

If the estimated tolerance limits are well within specification limits then the process is too good. Otherwise a readjustment of the process will be advisable with respect to either process average or process standard deviation or both. Otherwise a revision of specification limits is called for.

7.7 Total Quality Management

It is a system approach to assure quality levels, designed into products throughout the production system. Quality of a manufactured article is the fitness for use, meeting

the requirements of the customers who use the articles. It can also be defined as the totality of characteristics of an entity that is based on its ability to satisfy stated and implied needs. Total quality of products may be defined as the performance superiority of products which satisfy customers' needs in values and activities. Here the process may be refined to provide error-free planned outputs with consistency. Concept of total quality includes reliability, performance, safety aspects, serviceability, aesthetics, durability and perceived package of benefits accrued as seen by the customers.

Total quality management leads to a continuous journey for improvement towards excellence. Implementation of it requires refocussing on entire organisation's efforts, changing the attitudes and priorities. It deals with every aspect of an organisation viz, customers' satisfaction, employee relation, corporate image, community relation, Government affairs, ecofriendliness and corporate profitability and productivity.

Total quality is measured by (i) value price ratio, (ii) value cost ratio and (iii) error-free performance. Here value means the perception of quality of the customer and is measured by the relation between quality and cost. Total quality performance require (1) Customer orientation i.e., customers should be given more time and should be entertained with nice behaviour (2) Human resource excellence or excellence created by cohesive teams i.e. everyone in the organisation has to participate in total quality mission. As the total quality performance is people-driven, this should be achieved by giving training and education according to individual needs to create motivation of work of total quality mission.

(3) Management leadership which is required to understand customers' perceptions of value by designing products, processes and services as a total package.

(4) Movement of the management which changes intents to united success by setting annual total quality improvement objectives and reviewing regularly the progress with clarity and consistency and also by measuring the cost of non-conformance properly and by giving feedbacks to avoid the repetition of similar non-conformance.

For total quality performance the tools are :

(1) *Total quality fitness* which reviews organisation performance within 12 conditions of excellence viz (i) work culture, (ii) quality planning, (iii) improving quality information cycle, (iv) establishing accountability at each

process stage, (v) design for value edge of new products, (vi) diagnosis of value adding process to eliminate non-performing activities, (vii) accurate and precise information flow, (viii) closer relation with the venders, (ix) active individual participation to enhance quality, (x) need-based training, (xi) motivation to remove bias for change and (xiii) customer orientation.

(2) *Total quality improvement process* which identifies the main issues for improvement, defines and develops strategies to tap opportunities, take up improvement projects and implement through combined team work.

(3) *Operation cycle time* which critically examines the business operation and production system to set action plan at priority basis and to reduce cycle time to improve responsiveness to market needs and struggles against problems like unscheduled machine breakdown, inefficient work flows, incomplete customer information etc.

(4) *Quality function development* which is a process of building quality into the products by integrating the needs of the customers with new developments.

(5) *Cause effect diagram* which identifies the avenues of improvement by problem selection, team formation to investigate into the problem, brain storming for diagonistic journey, identification of root causes, recommendation of remedies, evaluation of the suggested remedies and standardisation and implementation

(6) Controlled *cost of poor quality* (i.e., non-conformance), which consists of internal field failure costs, external field failure cost inclusive of servicing, warranty, loss of reputation, etc., while controllable cost of poor quality consists of approval cost and prevention cost.

(7) *PDCA cycle* which means Planning, Do, Check and Act for continuous improvement in performance. Planning process sets improvement target over specified time period for proper customers' satisfaction. Here phase "Do" deals with resolving the problem by providing clues about deficiencies to be eliminated and by moving from symptoms to the root causes. The phase "Check" enables the team to assess the suitability of the suggested remedies in providing solution to the problem being investigated. The phase "Act" standardises the new system for future.

7.8 Worked out examples

Example 1.

A machine delivers packets of a given weight (in kg). Means and ranges of 10 samples of 5 packets each are recorded below :

Sample No. :	1	2	3	4	5	6	7	8	9	10
Mean :	15	16	15	18	17	14	18	15	17	14
Range :	7	7	5	9	8	7	10	4	10	5

Calculate the control limits and draw the control charts of mean weight and range of weight of packets and comment on the state of control. (Given that for n = 5, A_2 = 0.58, D_3 = 0, D_4 = 2.11).

Solution :

Let $\bar{x}_2$ and R_i be the mean weight and range of weights of the i-th sample for i = 1, 2, ..., 10. Standards of mean and range are not given. So $\bar{\bar{x}} = \frac{1}{10}\sum_{i=1}^{10}\bar{x}_i$ and $\bar{R} = \frac{1}{10}\sum_{i=1}^{10}R_i$ i.e.

$$\bar{\bar{x}} = \frac{159}{10} = 15\cdot 9,\ \bar{R} = \frac{72}{10} = 7\cdot 2$$

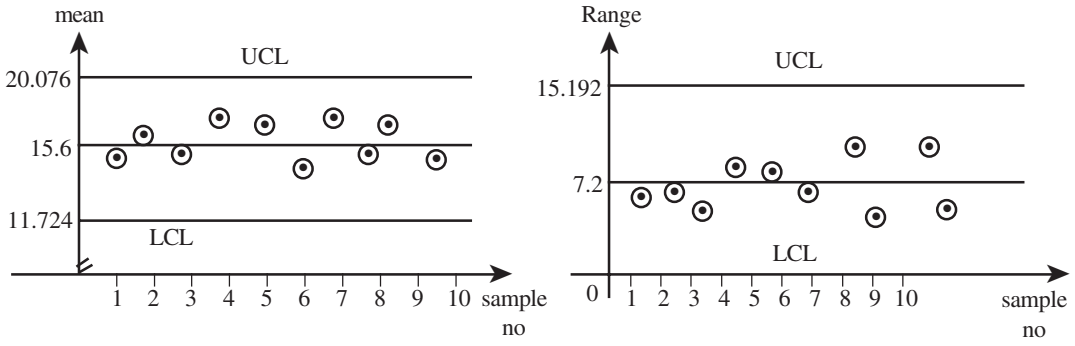
Then for mean chart LCL = $\bar{x} - A_2\bar{R} = 15.9 - 0.58\times 7.2 = 11.724.$

$$\text{CL} = 15.9 \text{ and UCL} = \bar{\bar{x}} + A_2R == 15.9 + 0.58\times 7.2 = 20.076.$$

As all the means lie within LCL and UCL the process is under control.

The limit for range chart LCL = $D_3\bar{R} = 0\times 7.2 = 0, \text{CL} = 7.2,$

$\text{UCL} = D_4\bar{R} = 2.11\times 7.2 = 15.192.$ As all the ranges lie between LCL and UCL i.e., between 0 and 15.192 so the process is under control. In both mean chart and range chart the process is under control. It is better to draw the control charts considering sample numbers along horizontal axis at equal distances and means or ranges along vertical axis i.e., by drawing the lines LCL, CL and UCL and plotting the points of means in the chart just as below :



Mean chart Range chart

From both the charts we see that all the points lie within UCL and LCL. So the process is in control.

Example 2.

The number of defective pens out of 100 of a lot has been observed for 10 lots as 10, 16, 17, 14, 8, 8, 16, 12, 15, 12. Construct a suitable control chart and hence draw a conclusion regarding the state of control.

Solution :

Standard is not given. Then proportion of defectives are 0.10, 0.16, 0.17, 0.14, 0.08, 0.08, 0.16, 0.12, 0.15, 0.12.

Then $\bar{p} = \frac{1}{10}\sum_{i=1}^{10} p_i = \frac{1.28}{10} = 0.128,$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{100}} = \cdot 128 - 3\sqrt{\frac{\cdot 128 \times \cdot 872}{100}} = \cdot 128 - \frac{3 \times 0 \cdot 3341}{10}$$

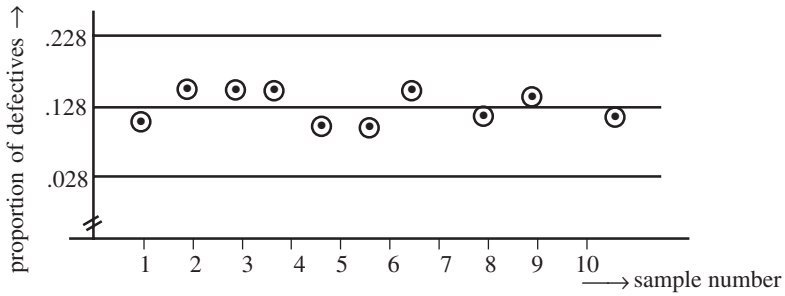
$= .128 - .100 = 0.028$

$CL = \bar{p} = 0.128,$

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{100}} = \cdot 128 + 3\sqrt{\frac{\cdot 128 \times \cdot 872}{100}} = \cdot 128 + \frac{3 \times \cdot 3341}{10}$$

$= .128 + .100 = 0.228.$

As all the values of p lie between LCL and UCL in control chart below of fraction defective, the process is in control.



Control chart for fraction defective

Example 3.

Following are the numbers of defects found in 1000 items of cotton piece goods inspected every day in first 15 days of a certain month.

1, 1, 3, 5, 7, 1, 2, 5, 1, 1, 11, 5, 0, 12, 5

Do these data come from a controlled process?

Solution :

If c_i denotes the number of defects in the i-th sub-group (here the i-th day) we have

$$\bar{c} = \frac{\sum_i c_i}{15} = \frac{60}{15} = 4.$$

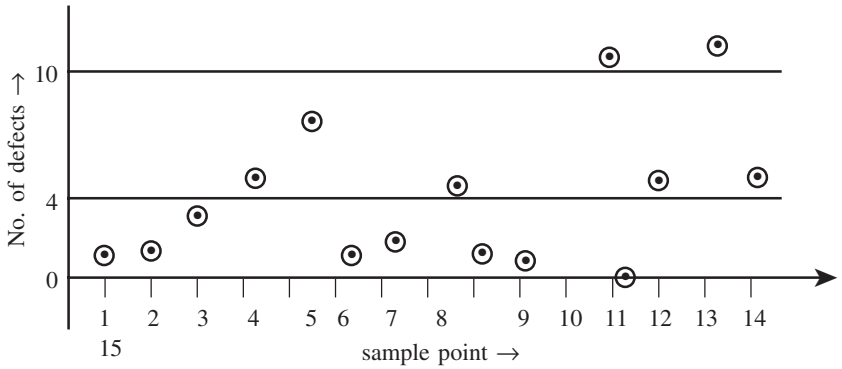
The control limits and the central line are as follows

$$\text{LCL} = \bar{c} - 3\sqrt{\bar{c}} = 4 - 3\sqrt{4} = 4 - 6 = -2.$$

As LCL is negative so we take LCL = 0,

$$\text{Central line} = \bar{c} = 4, \quad \text{UCL} = 4 + 3\sqrt{4} = 10.$$

The process is out of control because two values of number of defects 11 and 12 are more than UCL. We show this by control chart.



There are two points outside the control limits. So the process is out of control.

Example 4.

It has been noticed from past records of a factory using quality control methods that on an average 4 articles produced are defective out of a batch of 100. Find the maximum number of defective articles likely to be encountered in the batch of 400, assuming that the production process is in a state of control.

Solution :

Suppose p' is the fraction of defective in the lot.

Hence $p' = \frac{4}{100} = 0.04$ (given)

The sample size is given to be n = 400.

If d is the number of defectives in a sample of size n, then d = np and the control limits for np chart are (where standard given).

$$UCL = np' + 3\sqrt{np'(1-p')}$$

$$= 400 \times 0.04 + 3\sqrt{400 \times 0.04(1-0.04)}$$

$$= 16 + 3 \times 3.919 = 27.757$$

$$CL = np' = 400 \times 0.04 = 16$$

and $LCL = np' - 3\sqrt{np'(1-p')} = 16 - 3 \times 3.919 = 4.243$

Thus if the production process is in a state of control, the number of defective articles (d) likely to be encountered in the batch of 400 must lie between the control limits of 4.243 and 27.757, that is, between 4 and 28.

7.9 Summary

This chapter includes definition of S.Q.C., process control and product control, types of quality measure, rational sub-groups and control chart, construction of control chart of variable (mean, s.d and range charts) and attributes, (number of defective, proportion of defective and number of defects charts) specification limit and tolerance limits, total quality management and worked out examples.

7.10 Exercises

1. Explain the following terms : Statistical Quality Control, Rational sub-groups, Control charts, Process Control and Product Control.
2. Describe the construction of control charts for mean and range from appropriate data when standards are given and when standards are not given.
3. What are the assumptions used in S. Q. C?
4. Describe the construction of control charts for number defectives and fraction defectives from appropriate data when standard is given and when standard is not given.
5. Describe the construction of control charts for number of defects from appropriate data when standard is given and when standard is not given.
6. Define specification limits and natural tolerance limits.
7. What are the uses of S. Q. C.?
8. What are the advantages and limitations of S. Q. C.?
9. Write notes on total quality management.
10. A machine is set to deliver packets of a given weight in kg. Ten samples of size 5 each were recorded. Mean and range of each sample are given below :

Sample No. :	1	2	3	4	5	6	7	8	9	10
Mean :	14	18	16	15	17	16	18	15	17	14
Range :	7	6	8	4	8	5	10	4	10	5

Draw a mean chart and a range chart of weights of packets and comment on the state of control.

Given that for n = 5, $A_2 = 0.577$, $D_3 = 0$ and $D_4 = 2.115$.

11. The following are the number of defective transistors in 10 lots of 100 transistors each

17, 6, 10, 8, 10, 14, 7, 17, 2 and 15

Construct a suitable control chart and write a brief report on the evidence of the cast.

12. Ten pieces of clothes out of different rolls of equal length contained the following number of defects :

    3, 0, 2, 8, 4, 2, 1, 3, 7 and 1.

    Prepare a control chart and state whether the production process is in a state of control.

7.11 Suggested Readings

1. Grant, E. L., *Statistical Quality Control*, McGraw Hill 1964.
2. Cowden, D. J. *Statistical Methods in Quality Control*, Prentice Hall, 1957 and Asia Publishing House 1960.
3. Montgomery, D. C., *Introduction to Statistical Quality Control*, John Wiley 1985.
4. Gun, A. M., Gupta, M. K. and Dasgupta, B. Fundamentals of Statistics, Vol. 2, The World Press Pvt. Ltd., 2002, Kolkata.
5. Chaudhuri, S. B., *Elementary Statistics,* Vol. II, Shraddha Prakashan, 1986, Kolkata.