APPLIED STATISTICS

(DSSTT31/DBSTT31)

(BSC, BA STATISTICS -III)

ACHARYA NAGARJUNA UNIVERSITY

CENTRE FOR DISTANCE EDUCATION

NAGARJUNA NAGAR,

GUNTUR

ANDHRA PRADESH

CONTENTS

CONTENTS

LESSON - 1

BASIC PRINCIPLES OF SAMPLING & TYPES OF SAMPLING

LEARNING OBJECTIVES

- Upon completion of this lesson, you should be able to :
- Understand the meaning of the sampling attitude.
- Definitions and utility of sampling in statistics.
- Demonstrate the concepts of parameter, statistics, sampling distribution of statistics and standard error.
- Studying various types of sampling.

LESSON OUTILINE

1.1 INTRODUCTION

Sampling in statistics is as common and important as salt is in food. The origin of sampling is as old as our civilisation. Sampling is a procedure of making decisions by studying a few items regarding the characteristics of items in a universe. The use of sampling in day-to-day life has been illustrated with certain examples. In homes, ladies take out one or two rice grains (any other food item) from the cooking pan and test so that they are able to decide wheher the food in the pan is fully cooked or not. In medical sciences, a few drops of the blood are taken. Tested microscopically or chemically to know whether the blood contains some abnormalities or not. Whatever is observed

in the few drops is true for the blood of the whole body. As regards the use of sampling techniques in older days, it was assumed that it did not make much difference how the sample was selected. Now a days, sampling methods are extensively used in socio-economic surveys to the living conditions, cost of living index, and etc. in a class of people. In biological studies, experiments are conducted on some units (persons, animals, or plants) and inferences are drawn about the breed or variety to which the units belong. In the industries, sampling procedures are predominantly used for quality control. As a consequence, scientific sampling techniques have been employed by statisticians and research workers from time to time and their utility has been proved. With the aid of properly selected sample, estimates and conclusions may be drawn about the characteristics of items comprising a population or universe.

1.2 THE SAMPLING ATTITUDE :

The purpose of sampling : The broad aim of statistics is to describe and summarize mass phenomena like births, deaths, and income, and other interrelationships. However, it is often necessary or practicable to base such description on a fraction of the total aggregate, and sometimes an exceedingly small one at that. Such an experiment may be, and usually is quite satisfactory. It is astonishing have effective a well-selected fragment can be : a small snippet from a bolt of cloth; a few drops of blood from the patient's total supply; a few thousand survey of votes, by which we describe the political intentions of millions of voters. Such procedures are standard practice in every day social and economic life, as well as in the branches of scientific activity. In instances of this kind, when the data are partial rather than complete, when they are used to characterise the entire set, we call the fragment a sample, and the total aggregate a universe, or population. We name a specified value of the universe, such as the mean, a parameter, and its counterpart in the sample we term a statistic. The objective of sampling is, therefore, to draw an inference about the parameter, which is unknown, from the sample statistic which is observed. This process of generalizing in a prescribed manner from sample to universe has come to be known as statistical inference.

1.3 WHY SAMPLING IS NECESSARY

VARIOUS FORCES RESPONSIBLE FOR THE WIDE USE OF SAMPLING ARE DISCUSSED BELOW :

(i) **The physical imposibility of checking all the items in the population :** Some of the populations are infinite, therefore, all members of such populations cannot be examined and some of the populations, even though finite, are so vast and inaccessible that it is not possible to study all the items. In such situations, sampling is the only way to get information about such populations. For example to estimate the yield of wheat per acre in Maharastra, neither it possible nor is it advisable to study all wheat fields to make per acre-yield estimates. Similarly, the population of fish, birds, mosquitoes and the like are large and constantly moving. In such cases, sampling is employed to estimate the total number or to study their main characteristics.

(ii) **The destructive nature of goods :** In some fields, sampling is the only possible method of studying the nature of items constituting a population.

For example, some industrial goods are destroyed when they are tested for quality. To measure the life of motor cycle tyre or an electric bulb, it must be used until it is

worthless. If all the tyres and electric bulbs are tested, none would be available for sale. In such situations, sampling is the only way of studying the characteristics of items in a population. For such products, a small part of the output is examined to assess the quality of the product.

(iii) Adequacy of Sampling Results : With the advancement of sampling techniques, sampling results are as accurate as the results of complete enumeration. Even if funds are available for counting all members of a population, it is not necessary to do so because it is doubtful that complete enumeration will increase the accuracy. In census studies, error component is always large because it is not possible to provide proper training to a large army of investigators and it is difficult to supervise their work.

For a sample study, a small team of investigators is needed. They can be trained in the art of data collection. Even better qualified persons can be recruited. As compared to census study, the error component may be much smaller in a sample study because of trained investigators.

Thus a small number of items included in the sample may be examined with greater accuracy as compared to the study of all items in a universe.

(iv) Cost of Study in the entire population : The study of all members in a population entails too much expenditure. Therefore sample surveys are being substituted to an increasing extent for studying the characteristics of members in population inplace of census surveys. The use of samples has reduced the cost of data collection to a point where the businessman or a research worker can afford.

(v) Saving in time : The study of all members in population is merely time waste and laborious. In many cases, it becomes necessary to process the results more quickly than would be possible if all the members in a population are studied. For example, when the Government of India gives dearness allowance to its employees to compensate the rise in prices, it takes a sample of whole sale stores scattered throughout the country for collecting data on prices to compute whole sale price index because a complete counting may take too much time that the results may be of no use by the time they become available.

1.4 DEFINITIONS

1.4.1 The population and the Sample :

Population : In a statistical investigation the interest is usually in the assessment of the general magnitude and study of variation with respect to one or more characteristics relating to individuals belonging to a group. This group of individuals under study is called population or universe. Thus in statistics, population is an aggregate of objects animate or inanimate under study. The population may be finite or infinite.

A population which is consisting of countable and limited number of items, is called finite population such as, the number of books in a college library, the number of workers in a factory, the number of students in a college or the number of state agricultural forms in India.

****** A population consists of a collection of individual units. Which may be persons or experimental outcomes, whose characteristics are to be studied.

An infinite population may be defined as one which is comprised of limit less number of items or the compositions of items such that they cannot be counted. Examples of such populations are throws of a coin or dice or opinions of consumers about certain commodity.

The definition of the population throws light on two properties namely :

(i) Population is a body of members about which information is needed.

(ii) Every member of a population has a certain specified attribute or attributes.

Sample : A finite subset of individuals in a population is called sample and the number of individuals in a sample is called the sample size.

By sample we can determine the characteristics of population, instead of enumerating entire population. So, the sample characteristics are utilised to approximately determine or estimate the population. For example on estimating the sample of a particular stuff we arrive at a decision of purchasing or rejecting that stuff. The error involved in such approximations is known as sampling error and is inherent and unavoidable in any and every sampling scheme.

Sampling is quite often used in our day-to-day practical life. For example, if we selected five students from a class of forty students, five selected students constitute a sample. We select fifty bulbs to test the life of bulbs, from a lot of bulbs manufactured by a factory in a week. Fifty selected bulbs constitute a sample.

1.4.2 Parameter & Statistic :

Parameter : Any population constant is called a parameter. For example, population mean(μ) and population variance $\left(\sigma^2\right)$ are parameters. To obtain a parameter value observations are taken on each and every unit of the population and a value of a constant pertaining to a characterstic is calculated from those observations. This constant value is termed as parameter. Such constant measure (s) of a population characterises that population.

1.4.3 Statistics :

Any sample constant is called a static and does not involve any unknown parameters.

In practice parameter values are estimates based on the sample values generally used. Thus statistic which may be regarded as an estimate of the parameter, obtained from the sample, is a function of sample values only. Example. sample mean $(\bar{x})$ sample variance $\left(s^2\right)$ etc. Also, sample mean $(\bar{x})$ is used to estimate unknown population mean (μ).

** A sample is a portion of the population that is studied to learn about the characteristics of the population**.

Remark : A statistic $t = t(x_1, x_2, x_3, \cdots, x_n)$, a function of the sample values.

$x_1, x_2, \cdots, x_n$ is an unbiased estimate of population parameter θ. If $E(t) = \theta$, in otherwords, if E(Statistic) = parameter, then the statistic is said to be an unbiased estimate of the parameter.

1.4.4 Sampling Distribution : The number of possible samples of size n that can be drawn from a finite population of size N is N_{c_n} (if N is large or infinite then we can draw a large number of such samples). For each of these samples we can compute a statistic, say 't'.... eg., mean, variance, etc., which will vary from sample to sample. The aggregate of various values of the statistic under consideration, so obtained, may be grouped into a frequency distribution which is known as sampling distribution of the statistic. Then, we have the sampling distribution of the sample mean, ($\bar{x}$), sample variance (s^2) etc.,

1.4.5 Standard Error : The standard deviation of the sampling distribution of a statistic is known as its standard error. The standard errors (S.E.) of some of the well - known statistics are given below. Where n is the sample size, σ^2 the population variance, P the population propotion and $Q = 1 - P$.

S.No.	Statistic	Standard Error
1.	$\bar{x}$	$\sigma/\sqrt{n}$
2.	Observed Sample proportion ' p '	$\sqrt{PQ/n}$
3.	Sample standard deviation 'S'	$\sqrt{\sigma^2/2n}$
4.	S^2	$\sigma^2\sqrt{2/n}$
5.	Quartiles	1.32263 $\sigma/\sqrt{n}$
6.	Median	1.25331 $\sigma/\sqrt{n}$
7.	Sample Corelation 'r'	$(1-P^2)/\sqrt{n}$ P being population correlation coefficient
8.	Coefficient of variation	$\frac{V}{\sqrt{2n}}\sqrt{1+\frac{2V^2}{104}} = \frac{V}{\sqrt{2n}}$

Remarks : S.E. plays a very important role in large sample theory and forms the basis of the testing of hypothesis. If t is any statistic then for large sample s.

$$E = \frac{t - E(t)}{\sqrt{V(t)}} = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$$

1.5 PRINCIPLES OF SAMPLE SURVEY

The theory of sampling is based on the following important principles :

(i) Principle of statistical regularity : This principle has its origin in the mathematical theory of probability. According to King "the law of statistical regularity lays down that a moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristics of the large group". This principle stresses the desirability and importance of selecting the sample at random such that each and every unit in the population has an equal chance of being selected in the sample.

An immediate derivation from the principle of statistical regularity in the principle of inertia of large numbers which states that "other things being equal as sample size increases, the results tend to be more reliable and accurate". This is because in, dealing with large numbers with variations in the component parts tend to balance each other and consequently the variation in the aggregate result is likely to be insignificant. For example, in a coin tossing experiment, the result will be 50% heads and 50% tails provided the experiment is performed a fairly large number of times.

(ii) Principle of Validity : By the validity of a sample design we mean that it should enable us to obtain valid tests and estimates with the technique of probability sampling.

(iii) Principle of Optimisation : This principle impresses upon obtaining optimum results in terms of efficiency and cost of design with the resources at our disposal. The reciprocal of sampling variance of the estimate provides a measure of its efficiency while a measure of the cost of the design is provided by the total expences incurred in terms of money and man hour.

The principle of optimisation consists in

(a) achieving a given level of efficiency at minimum cost and

(b) obtaining maximum possible efficiency with given level of cost.

1.6 ERRORS IN SAMPLING SURVEY

In any survey two types of errors are likely to occur.

(i) Sampling errors (ii) non-sampling errors.

(i) Sampling errors : The errors which are introduced due to errors in selection of a sample or the discrepancies between population parameters and estimates which are derived from a random sample. This discrepancy generally decreases as sample size increases. In the begining the decrease in error with increase in sample size is quite rapid but gradually the decrease in error becomes negligible with increase in sample size. In this way we can

minimize the error or keep the error as small as we please and at the same time minimizing the cost of the survey. To determine optimum sample size, a tolerable amount of error is prefixed and smallest sample size is determined to help keep this error with in tolerable limit.

(ii) Non-Sampling Error : An error in sample estimates cannot be attributed to sampling fluctuations. It is experienced that the studies based on complete enumeration do not yield similar results in repeated enumerations. Such a discrepancy occurs due to many errors which are termed as non-sampling errors.

Some better ways of minimizing the sampling errors are the choice of an appropriate sampling scheme, selecting a sample of optimum size and the use of standard techniques of estimation. Where as non-sampling errors can be minimized through superior management of survey or investigation, employing skilled personnel and by using modern computational aids.

1.7 ADVANTAGES OF SAMPLING OVER COMPLETE CENSUS :

The main advantages or merits of sampling technique over the complete enumeration survey may be outlined as follows :

(i) Less Time : There is considerable saving in time and labour since only a part of the population has to be examined. The sampling results can be obtained more rapidly and the data can be analysed much faster since relatively fewer data have to be collected and processed.

(ii) Reduced cost of the survey : Sampling usually results in reduction in cost in terms of money and in man hours. Although the amount of labour and the expenses involved in collecting information are generally greater per unit of sample than the complete enumeration, the total cost of the sample survey is expected to be much smaller than that a complete census. Since in most of the cases our resources are limited interms of money and the time with in which the results of the survey should be obtained, it is usually imperative to resort to sampling rather than complete enumeration.

(iii) Greater Accuracy of Results : The results of a sample survey are usually much more reliable than those obtained from complete census due to the following reasons:

a) It is always possible to determine the extent of the sampling error : and

b) The non-sampling error due to a number of factors such as training of field workers, measuring and recording observations, location of units, biases due to interviewers, etc., are more likely to be of serious nature in complete census than in a sample survey. In a sample survey non-sampling errors can be controlled more effectively by employing more qualified and better trained personnel, better supervision and equipment for processing and analysis of relatively limited data. Effective control of non-sampling errors more than compensates the errors in the estimates due to sampling. As such more sophisticated statistical techniques can be employed to obtain relatively more reliable results.

(iv) Greater Scope : Sample survey has generally greater scope as compared with complete census. The complete enumeration is impracticable, rather inconceivable if

the survey requires a highly trained personnel and more sophisticated equipment for the collection and analysis of the data. Since sample survey saves time and money, it is possible to have a thorough and intensive enquiry because a more detailed information can be obtained from small group of respondents.

(v) If the population is too large, as for example, trees in jungle, we are left with no way but to resort to sampling.

(vi) If the population is hypothetical, as for example in coin-tossing problem where the process may continue indefinitely sampling method is the only scientific method of estimating the parameters of the universe.

1.8 LIMITATIONS OF SAMPLING

Sampling theory has its own limitations and problems which may be briefly outlined as follows :

(i) Proper care should be taken in the planning and execution of sample survey, otherwise the results obtained might be inaccurate and misleading.

(ii) Sampling theory requires the services of trained and qualified personnel and sophisticated equipment for its planning, execution and analysis. In the absence of these, the results of the sample survey are not trust worthy.

(iii) However, if the information is required about each and every unit of the universe, there is no way but to resort to complete enumeration. Moreover, if time and money are not important factors and if the universe is not too large, a complete census may be better than any sampling method.

1.9 TYPES OF SAMPLING

The method of selecting a sample is of fundamental importance in the theory of sampling and usually depends upon the nature of the data and type of enquiry. The procedure of selecting a sample may be classified under the following three heads.

(i) Purposive or judgement sampling

(ii) Probability sampling

(iii) Mixed Sampling

(i) Purposive Sampling : Purposive sampling is one in which the sample units are selected with definite purpose in view and the choice of the sampling units depends entirely on the discretion and judgement of the investigator. For example, we want to give the picture that the standard of living has increased in the city of New Delhi.

This sampling method is seldom used and cannot be recommended for general use since it is often biased due to the element of subjectiveness on the part of the investigator. How ever, if the investigator is experienced and skilled and this sampling is carefully applied, the judgement samples may yield valuable results.

(ii) Probability Sampling : Probability sampling is the scientific method of selecting samples according to some laws of chance in which each unit in the population has some definite

pre-assigned probability of being selected in the sample. The different types of probability sampling are :

(a) Where each unit has an equal chance of being selected.

(b) Sampling units have different probabilities of being selected.

(c) Probability of selection of a unit is proportional to the sample size.

(iii) Mixed Sampling : If the samples are selected partially according to some laws of chance and partly according to fixed sample rule (no assignment of probabilities) they are termed as mixed samples and the technique of selecting such samples is known as mixed sampling.

Some methods of sampling are

(i) Simple random sampling

(ii) Stratified Random sampling

(iii) Systematic sampling

In the following sections, the above methods are explained in detail.

1.10 SUMMARY

Sampling is a procedure of making decisions by studying a few items regarding the characteristics of population. This practice of making decisions about the characteristics of terms in a universe on the basis of a few items is very old even though scientific approach to this problem is of recent origin. Daily life is full of such examples; a wheat merchant, who examines a handful of wheat from a cart-load, in order to ascertain the quality, is making use of sampling. Similarly, in a sweet shop when we asses the quality and taste of sweets by taking a piece of sweet from the sweet tray and then decide to buy it or not, we are making use of sampling to form a conclusion about the whole.

In the present era, sampling is necessary and is employed extensively in business, quality control, agricultural research, medical science, public opinion polls and in many other fields. Sampling techniques have been developed and have been employed by statisticians and research workers. From time to time their utility has been proved. With the aid of properly selected sample estimates conclusions may be drawn about the characteristics of items comprising a population.

QUESTIONS TO STUDY

1. What is sampling ? Why sampling is necessary ?
2. When is sampling useful and why ? What essential characteristics must a sample possess to serve its purpose?
3. Explain Sampling & Non-sampling errors.
4. What are the main sources of error in sample surveys designed to estimate the yield of wheat in India ?
5. What precautions should a statisticians take to reliable estimates of yield of crop from sample of survey?

6. Discuss briefly the basic principles of a sample survey?
7. What are the main steps involved in a sample survey? Discuss them briefly.
8. What are the different sources of errors in a sample survey?
9. Describe briefly how these errors can be controlled.
10. State briefly the advantages of sampling over complete enumeration.

1.11 REFERENCES

1. Basic Statistics - B.C. Agarwal
2. Statistical Researing in Sociology - John H. Muller & Karl. F.SC Huessler
3. Fundamentals of Applied Statistics - S.C. Gupta & V.K. Kapoor
4. Methods of Statistical Analysis - P.S. Grewal

LESSON - 2

SIMPLE RANDOM SAMPLING

LEARNING OBJECTIVES

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

* Understand the meaning of simple random sampling technique and methods of selecting a simple random sample.
* Simple random sample with replacement and simple random sample without replacement
* Notations and Terminology
* Studying the theorems based on simple random sample without replacement.
* Discuss the merits and the limitations of simple random sample.

LESSON OUTILINE

2.1 INTRODUCTION

Simple Random sampling is a technique of drawing a sample in which every item in a population has an equal chance of being included in a sample. Suppose, we want to select a sample of size 3 from 6 numbers 1, 2, 3, 4, 5, & 6. Here there are 20 possible samples and any one of these can be selected. These 20 samples are :

134, 126, 124, 125, 134, 135, 145, 123, 146, 156, 243, 235, 236, 245, 246, 256, 345, 356, 346, 456.

Here each these 20 samples has equal probability of selection and the choice of any one of these samples depends purely on chance. Therefore, this method is also called chance selection method. The principle underlying simple random sampling is that the personal factor is eliminated in the choice of items to be included in the sample.

In this method an equal probability of selection is assigned to each unit of the population at the first draw. It also implies an equal probability of selecting any unit from available units at subsequent draws.

2.2 SIMPLE RANDOM SAMPLING WITH REPLACEMENT

In this case, a sample is drawn from a population with a known probability and the sample is returned to the population before the next draw is made. Thus, in this method at each draw, the population size remains the same and under this sampling plan, a sample has more than one chance of being selected. For example, a card is randomly drawn from a pack of cards, placed back in the pack after not zing its face value, before the next card is drawn or from an urn containing balls of different colours a ball is drawn, its colour is noted and kept back in the urn before another ball is drawn. Such a samplig method is known as sampling with replacement. There are N^n possible samples of size n from a population of N units in case of sampling with replacement.

Simple Random sampling without Replacement : In this selection procedure, if a sample from a population of size N is selected, it is not returned to the population. Thus, for any subsequent draws, the population size is reduced by one. Obviously, at the time of first draw, the population size is N and the probability of a unit being selected randomly is $\frac{1}{N}$; for the second unit to be randomly selected, the population size is $N-1$ and the proabbility of drawing any one of the remaining sampling unit is $\frac{1}{N-1}$, similarly at the third draw, the probability of selection is $\frac{1}{N-2}$ and so on. The selection procedures with replacement and without replacement become equivalent. Whereas in the case of small populations, the values $\frac{1}{N}, \frac{1}{N-1}, \frac{1}{N-2} \cdots\cdots\cdots \frac{1}{N-n+1}$ differ considerably. The sampling from small and large populations is generally expressed as sampling from finite and infinite population respectively. There are $\binom{N}{n}$ possible samples, in case of sampling without replacement. Let E_r be the event that any specified unit is selected at the r^{th} draw. Then $P(E_r) =$ Prob. {that the specified unit is not selected in any one of the previous $(r-1)$ draws and then selected at the r^{th} draw} $P(E_r) = \prod_{i=1}^{r-1} P$ {It is not selected at i^{th} draw} $\times$ P {It is selected at r^{th} draw given that it is not selected at the previous $(r-1)$ draws}

(By multiplication theorem of probability, since draws are independent)

$$\therefore P(E_r) = \prod_{i=1}^{r-1}\left[1 - \frac{1}{N-(i-1)}\right] \times \frac{1}{N-(r-1)}$$

$$= \prod_{i=1}^{r-1} \frac{N-i}{N-(i-1)} \times \frac{1}{N-r+1}$$

$$= \frac{N-1}{N} \times \frac{N-2}{N-1} \times \frac{N-3}{N-2} \times \cdots \times \frac{N-r+1}{N-r+2} \times \frac{1}{N-r+1}$$

$$= \frac{1}{N}$$

$$P(E_r) = \frac{1}{N} = P(E_1)$$

This leads to a very important property of simple random sampling without replacement (SRSWOR), viz.,

"The probability of drawing a specified unit of the population at any given draw is equal to the probability of its being selected at the first draw".

Since a specified unit can be included in the sample of size n in n mutually exclusive ways, viz., it can be selected in the sample at the r^{th} draw $(r = 1, 2, 3, \cdots, n)$ and since

$$P(E_r) = \frac{1}{N}, \quad r = 1, 2, 3, \cdots, n$$

By addition theorem of probability, we get

The probability that a specified unit is included in the sample

$$= \sum_{r=1}^{n} \frac{1}{N} = \frac{n}{N}$$

Remark : Sample Random sampling can also be defined equivalently as follows :

Let us suppose that a sample of size n is drawn from a population of size N. There are $\binom{N}{n}$ possible samples. Simple Random Sampling (S.R.S.) is the technique of selecting the sample in such a way that each of the $\binom{N}{n}$ possible samples has equal chance or probability $P = \frac{1}{\binom{N}{n}}$ of being selected :

In SRS

Probability of selecting any unit at the first draw = $\frac{1}{N}$

Probability of selecting any unit of the remaining $(N-1)$ units in the second draw $=\frac{1}{N-1}$ and so on.

The probability of selecting any unit of the remaining $N-(i-1)$ units at the i^{th} draw

$$=\frac{1}{N-(i-1)},\ (i=3,4,\cdots,n)$$

Since all the draws are independent, by compound probability theorem, the probability of selecting a sample of size n in a fixed specified order, is

$$\frac{1}{N(N-1)(N-2)\cdots\cdots(N-n+1)}$$

Since this probability is independent of the order of the sample and since there are $n!$ permutations of the sampled units, by addition theorem of probability, the required probability of obtaining a sample of size n (in any order) is

$$p=\frac{n!}{N(N-1)(N-2)\cdots\cdots(N-n+1)}=\frac{1}{\left(\frac{N}{n}\right)}$$ as required.

2.3 METHODS OF SELECTING A SIMPLE RANDOM SAMPLE

(a) Lottery Method : In this method, all the items in a population are represented by numbers. These numbers are written on small chits of paper or on cards which are homogeneous in all aspects. These numbered chits of paper must be identical in size, colour and shape. After numbering, these chits are put into a drum which may be rotated either by hand or by machine to mix them thoroughly, then a required number of slips may be drawn, blind folded which constitutes a sample.

(b) Random Numbers Table : The battery method of selecting a random sample is some what cumbersome because randomness in the drum is not as simple as it appears to be because slips may stick together or to the sides. When the size of a population is large, it takes too much time to number the individual numbers of a population. To avoid these short comings, we can make use of table of Random numbers for selecting a random sample. For this purpose, three types of random number tables are available such as :

(i) Tippett's table of random numbers.

(ii) Fisher and Yate's table of random numbers.

(iii) Kendall and Babington Smith's table of random numbers

Among these tables, Tippett's table of ranodm numbers is most popular and is widely employed for selecting a random sample.

The most practical and inexpensive method of selecting a random sample consists in the use of 'Random Numbers Tables', which have been so constructed that each of the digits 0, 1, 2, ..., 9 appear with approximately the same frequency and independently of each other. If we have to select a sample from a population of size $N(\leq 99)$ then the numbers can be combined two by two taken to give pairs from 00 to 99. Similarly for $N \leq 999$ or $N \leq 9999$ and so on.

(i) Tippet's Random Numbers Tables : Tippet number tables consist of 10,400 four digits numbers giving in all 10,400 $\times$ 4 i.e. 41,600 digits selected at random from British census reports.

(ii) Fisher and Yates Tables : Comprises 15000 digits arranged in twos. Fisher and yates obtained these tables by drawing numbers at random from the 10th to 19th digits of A.S. Thomson's 20 figure logarithmic tables.

(iii) Kendall and Babington Smith's, random tables Consists of 1,00,000 digits grouped into 25,000 sets of 4 digited random numbers.

2.4 NOTATIONS AND TERMINOLOGY

Let us consider a finite population of N units and Y be the character under consideration. The capital letters are used to deserible the characteristics of the populations where as small letters refers to sample observations.

Let $Y_i\,(i = 1, 2, \cdots\cdots, N)$ be the value of the character for the i^{th} unit in the population and the corresponding small letters $y_i, (i = 1, 2, \cdots, n)$ denote the value of the character for i^{th} unit selected in the sample. Then we define

Population Mean $= \bar{Y}_N = \dfrac{1}{N}\sum_{i=1}^{N} Y_i$ ..... (i)

Sample mean = $\bar{y}_n = \dfrac{1}{n}\sum_{i=1}^{n} y_i$ ..... (ii)

Sample mean $\bar{y}_n$ may also be written alternatively as

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} a_i Y_i \quad \text{..... (iia)}$$

where

$$a_i = \begin{cases} 1. \text{ if } i^{\text{th}} \text{ unit is included in the sample} \\ 0. \text{ if } i^{\text{th}} \text{ unit is not included in the sample} \end{cases}$$

$S^2 =$ Mean square for the population.

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}_N\right)^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} Y_i^2 - N\bar{Y}_N^2\right] \text{ ..... (iv)}$$

$S^2 =$ Mean square for the sample

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}_n\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}_n^2\right] \text{ ---------- (v)}$$

2.5 THEOREM

2.5.1 Theorem : In simple random sampling without replacement (SRSWOR), the sample mean is an unbiased estimate of the population Mean, i.e. $E\left[\bar{y}_n\right] = \bar{Y}_N$ ..... (vi)

Proof : $E\left[\bar{y}_n\right] = E\left[\frac{1}{n}\sum_{i=1}^{N} a_i Y_i\right]$ [from (iia)]

$$= \frac{1}{n}\sum_{i=1}^{n} E\left(a_i\right) Y_i$$

Since a_i takes only two values 0 and 1.

$$E\left(ai\right) = 1 \cdot p\left(a_i = 1\right) + 0 \cdot p\left(ai = 0\right)$$

$= 1 \cdot p$ [i^{th} unit is included in a sample of size n] + $0 \cdot p$ [i^{th} unit is not included in a sample of size n]

$$= 1 \cdot \frac{n}{N} + 0\left(1 - \frac{n}{N}\right) \qquad \left[\because P\left(E_i\right) = \frac{n}{N}\right]$$

$$= \frac{n}{N}$$

Hence

$$E(\bar{y}_n) = \frac{1}{\cancel{n}} \sum_{i=1}^{N} \frac{\cancel{n}}{N} \cdot Y_i = \frac{1}{N} \sum_{i=1}^{N} Y_i = \bar{Y}_N .$$

$$\therefore E(\bar{y}_n) = \bar{Y}_N \text{ as desired.}$$

Theorem 2.5.2 : In SRSWOR the sample mean square is an unbiased estimate of the population mean square, i.e.,

$$E(s^2) = S^2 \text{ ---------- (vii)}$$

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}_n^2\right]$$

Proof : $$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i{}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i{}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i{}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i{}^2 + \sum_{i \neq j=1}^{n} y_i\, y_j\right)\right]$$

$$= \frac{1}{n-1}\left(1 - \frac{1}{n}\right)\sum_{i=1}^{n} y_i{}^2 - \frac{1}{n(n-1)} \sum_{i \neq j=1}^{n} y_i\, y_j$$

$$= \frac{1}{n}\sum_{i=1}^{n} y_i{}^2 - \frac{1}{n(n-1)} \sum_{i \neq j=1}^{n} y_i\, y_j \text{ ..... (viii)}$$

$$\therefore E(s^2) = \frac{1}{n} E\left[\sum_{i=1}^{n} y_i{}^2\right] - \frac{1}{n(n-1)} E\left[\sum_{i \neq j=1}^{n} y_i\, y_j\right] \text{ ..... (viii a)}$$

We have

$$E\left[\sum_{i=1}^{n} y_i{}^2\right] = E\left[\sum_{i=1}^{N} a_i\, Y_i^2\right], \text{ where } a_i \text{ is defined (in iii)}$$

$$= \sum_{i=1}^{N} E(a_i) Y_i^2 = \frac{n}{N} \cdot \sum_{i=1}^{N} Y_i^2 \text{ ..... (ix)}$$

and $E\left[\sum_{i \neq j=1}^{n} y_i y_j\right] = E\left[\sum_{i \neq j=1}^{N} a_i a_j Y_i Y_j\right] = \sum_{i \neq j=1}^{N} E(a_i a_j) Y_i Y_j$ ----------(*)

where a_i, a_j are defined in (iii)

Now, $E[a_i, a_j] = 1 \cdot p[a_i a_j = 1] + 0 \cdot p[a_i a_j = 0]$

$$= p[a_i = 1 \cap a_j = 1] = p[a_i = 1] \cdot p\left[a_j = \frac{1}{a_i = 1}\right]$$

$$= \frac{n(n-1)}{N(N-1)} \text{ ---------- (x)}$$

because $p[a_i = 1] = p$ [i^{th} unit is included in the sample of size n] $= \frac{n}{N}$

and $p\left[a_j = 1 / a_i = 1\right] = p$ [j^{th} unit is included in the sample given that i^{th} unit is included in the sample]

$$= \frac{n-1}{N-1}$$

Substituting in $(*)$, we get

$$E\left[\sum_{i \neq j=1}^{n} y_i y_j\right] = \sum_{i \neq j=1}^{N} \frac{n(n-1)}{N(N-1)} Y_i Y_j \text{ ---------- (xi)}$$

Substituting from (ix) and (xi) in (viii a), we get

$$E(s^2) = \frac{1}{N} \sum_{i=1}^{N} y_i^2 - \frac{1}{N(N-1)} \sum_{i \neq j=1}^{N} y_i y_j$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{N} y_i^2 - N \bar{y} N^2\right] \text{ (using viii)}$$

$$= S^2$$

$$\therefore E\left(s^2\right) = S^2 \text{ as desired.}$$

Remarks : From the above two theorem we see that unbiased estimates of the population mean and population mean squares are proved by sample and sample mean squares respectively.

Theorem 2.5.3 : In SRSWOR, the variance of the sample mean is given by

$$Var\left(\bar{y}_n\right) = \frac{S^2}{n} \cdot \frac{N-n}{N} \text{ ---------- (xii)}$$

Proof :

$$Var\left(\bar{y}_n\right) = E\left[\bar{y}_n{}^2\right] - \left[E\left(\bar{y}_n\right)\right]^2 \text{ ---------- (xiii)}$$

$$= E\left[\bar{y}_n{}^2\right] - \bar{Y}_N^2 \text{ (on using } v i \text{)}$$

$$\text{Now, } E\left[\bar{y}_n\right]^2 = E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right]^2 = \frac{1}{n^2} E\left[\sum_{i=1}^{n} y_i{}^2 + \sum_{i \neq j=1}^{n} y_i\, y_j\right]$$

$$= \frac{1}{n^2}\left[E\left(\sum_{i=1}^{n} Y_i{}^2\right) + E\left(\sum_{i \neq j=1}^{n} Y_i\, Y_j\right)\right] \text{-------- (xiv)}$$

From (ix), we have

$$E\left[\sum_{i=1}^{n} y_i{}^2\right] = \frac{n}{N}\left[\sum_{i=1}^{N} Y_i^2\right]$$

But $\sum_{i=1}^{N}\left(Y_i - \bar{Y}_N\right)^2 = \sum_{i=1}^{N} Y_i^2 - N\bar{Y}_N{}^2$

$$\Rightarrow \sum_{i=1}^{N} Y_i^2 = \sum_{i=1}^{N}\left(Y_i - \bar{Y}_N\right)^2 + N\bar{Y}_N{}^2 = (N-1)S^2 + N\bar{Y}_N{}^2 \text{ ---------- } (**)$$

$$\therefore E\left(\sum_{i=1}^{n} y_i{}^2\right) = n\left[\frac{N-1}{N} \cdot S^2 + \bar{Y}_N{}^2\right] \text{---------- (xiv a)}$$

Also from (xi), we get

$$E\left[\sum_{i\neq j=1}^{n} y_i\, y_j\right] = \frac{n(n-1)}{N(N-1)} \sum_{i\neq j=1}^{N} Y_i\, Y_j$$

$$= \frac{n(n-1)}{N(N-1)}\left[(\Sigma Y_i)^2 - \sum_{i=1}^{n} Y_i^2\right]$$

$$= \frac{n(n-1)}{N(N-1)}\left[N^2\bar{Y}_N^2 - (N-1)S^2 - N\bar{Y}_N^2\right] \text{ (using * *)}$$

$$= \frac{n(n-1)}{N(N-1)}\left[N(N-1)\bar{Y}_N{}^2 - (N-1)S^2\right]$$

$$= n(n-1)\left[\bar{Y}_N{}^2 - \frac{S^2}{N}\right] \text{ ---------- (xiv b)}$$

Substituting from (xiv a) and (xiv b) in (xiv), we get

$$E\left[\bar{y}_n{}^2\right] = \frac{1}{n}\left[\left(1-\frac{1}{n}\right)S^2 + \bar{Y}_N{}^2\right] + \left(1-\frac{1}{n}\right)\left(\bar{Y}_N{}^2 - \frac{S^2}{N}\right)$$

$$= \bar{Y}_N{}^2 + \left(\frac{1}{n} - \frac{1}{N}\right)S^2 \text{ ---------- (XV)}$$

Substituting in (Xiii), we get

$$E\left[\bar{y}_n{}^2\right] = \frac{1}{n}\left[\left(1-\frac{1}{n}\right)S^2 + \bar{Y}_N{}^2\right] + \left(1-\frac{1}{n}\right)\left[\bar{Y}_N{}^2 - \frac{S^2}{N}\right]$$

$$= \bar{Y}_N{}^2 + \left(\frac{1}{n} - \frac{1}{N}\right)S^2 \text{ ---------- (xv)}$$

substituting in (xiii), we get

$$Var\left(\bar{y}_n\right) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2 = \frac{N-n}{N}\cdot\frac{S^2}{n}$$

Remark 1 : $f = \frac{n}{N}$ is called the sampling fraction and consequently, we get

$$V(\bar{y}_n) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} = (1-f)\frac{S^2}{n} \text{ ---------- (xvi)}$$

The factor $(1-f)$ is called the finite population correction (f.p.c.). If the population size N is very large or if n is small compared with N then $f = \frac{n}{N} \to 0$ and consequently $f \cdot p \cdot c \to 1$.

2 : The standard error (S.E.) of the sampling distribution of $\bar{y}_n$ is given by

$$S.E(\bar{y}_n) = \sqrt{\frac{N-n}{N}} \cdot \frac{S}{\sqrt{n}} \text{ ---------- (XVII)}$$

Usually S is not known and we replace S^2 by its unbiased estimate s^2 and get

$$Est.(S.E\,\bar{y}_n) = \sqrt{\frac{N-n}{N}} \cdot \frac{S}{\sqrt{n}} = \sqrt{(1-f)}\,\frac{s}{\sqrt{n}} \text{ ---------- (x vii a)}$$

3 : If we consider simple random sampling with replacement (SRSWR) from a population with variance σ^2, then

$$Var(\bar{y}_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(y_i)$$

the covarience terms vanish since in SRSWR all the draws are independent and consequently $y_i\,(i = 1,2,\cdots\cdots,n)$ are iid with the seme variance σ^2

$$V(\bar{y}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

But, $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_N\right)^2 \Rightarrow N\sigma^2 = (N-1)S^2$

$$\therefore V(\bar{y}_n) = \frac{N-1}{nN}S^2 \text{ ---------- (xviii)}$$

Comparing this with the expression in (xii) we see that variance of the sample mean is more in SRSWR as compared with its variance in the case of SRWOR. In other words SRSWOR provides a more efficient estimate of $\bar{Y}_N$ relative to SRSWR.

: : Problems : :

2.5.3 A population of finite units has the observations 7, 6, 8, 4, 10

Random samples of units are drawn such that no unit is repeated in a sample and the order of selection does not matter, out of the 5 units, there can be 10 samples each consisting of 3 units. We can show sample mean is an unbiased estimate of population mean, s^2 is an inbiased estimate of S^2.

Proof : Possible samples, their means and variance calculated in the following table.

Sample Nos.	Observations	$\bar{x}$	s^2
1.	(7, 6, 8)	7	1
2.	(7, 6, 4)	17/3	7/3
3.	(7, 6, 10)	23/3	13/3
4.	(7, 8, 4)	19/3	13/3
5.	(7, 8, 10)	13/3	7/3
6.	(7, 4, 10)	7/3	9
7.	(6, 8, 4)	7	4
8.	(6, 8, 10)	6	4
9.	(6, 4, 10)	20/3	28/3
10.	(8, 4, 10)	22/3	28/3
		---	---
		$\Sigma\bar{x} = \frac{210}{3}$	$\Sigma S^2 = \frac{150}{3} = 50$

Mean of $\bar{x} = \frac{1}{10}\Sigma\bar{x} = \frac{1}{10} \times \frac{210}{3} = 7$

Similarly, mean of $s^2 = \frac{1}{10}\Sigma S^2 = \frac{1}{10} \times 50 = 5$

Population mean $\mu = \frac{35}{5} = 7$

$$S^2 = \frac{1}{5-1}\left\{(7-7)^2 + (6-7)^2 + (8-7)^2 + (4-7)^2 + (10-7)^2\right\}$$

$$= \frac{20}{4} = 5$$

From the above calculation we verified the statement that the sample mean is an unbiased estimate of population mean also sample variance $\left(s^2\right)$ is an unbiased estimate of S^2.

i.e. $E(\bar{y}_n) = \bar{Y}_N$ and $E(s)^2 = S^2$

Example 2.5.4 : Consider a population of 6 units with values 1, 2, 3, 4, 5, 6. Write down all possible samples of 2 (without replacement) from this population and verify that sample mean is an unbiased estimate of the population mean.

Also calculate its sampling variance and verify that

(i) It agrees with the formula for the variance of the sample mean, and

(ii) This variance is less than the variance obtained from sampling with replacement.

Solution :

Y	Y^2
1	1
2	4
3	9
4	16
5	25
6	36
Total 21	91

We have

$$\bar{Y} = \frac{1}{N}\Sigma Y = \frac{21}{6} = 3.5$$

$$S^2 = \frac{1}{N-1}\Sigma\left(Y - \bar{Y}\right)^2 = \frac{1}{N-1}\left[\Sigma Y^2 - N\bar{Y}^2\right]$$

$$= \frac{1}{5}\left[91 - 5 \times 12.25\right] = 3.5$$

$$\sigma^2 = \frac{N-1}{N}S^2 = \frac{5}{6} \times 3.5 = 2.917$$

The total number of samples of size $n = 2$ from a population of $N = 6$ units in $N_{C_n} = 15$.

Sample No	Sample Value (y)	Sample mean $(\bar{y})$	$(\bar{y}-\bar{Y})$	$(\bar{y}-\bar{Y})^2$
1	(1, 2)	1.5	- 2.0	4.00
2	(1, 3)	2.0	- 1.5	2.25
3	(1, 4)	2.5	- 1.0	1.00
4	(1, 5)	3.0	- 0.5	0.25
5	(1, 6)	3.5	0	0.00
6	(2, 3)	2.5	- 0.1	0.01
7	(2, 4)	3.0	- 0.5	0.25
8	(2, 5)	3.5	0	0.25
9	(2, 6)	4.0	0.5	0.00
10	(3, 4)	3.5	0	0.25
11	(3, 5)	4.0	1.0	0.25
12	(3, 6)	4.5	1.0	1.00
13	(4, 5)	4.5	1.5	
14	(4, 6)	5.0		2.25
15	(5, 6)	5.5	2.0	4.00
Total		**52.5**	**0**	**17.50**

$$E(\bar{y}) = \frac{\sum_{i=1} \bar{y}_i}{\binom{N}{n}} = \frac{\Sigma \bar{y}_i}{n} = \frac{52.5}{15} = 3.5 = \bar{Y}$$

which implies that sample mean $\bar{y}$ is an unbiased estimate of population mean $\bar{Y}$.

$$Var(\bar{y}) = \frac{1}{15}\sum_{i=1}^{15}(Y_i - \bar{Y})^2 = \frac{17.50}{15} = 1.167 \text{ --------- } (*)$$

Verification : (i) in SRSWOR, the variance of the sample mean is given by the formula.

$$\frac{N-n}{N_n} S^2 = \frac{6-2}{6\times 2} \times 3.5 = \frac{3.5}{3} = 1.167$$

which is same as the value in $(*)$

(ii) we have

$\text{var}(\bar{y})$ in SRSWOR = 1.167 [as in (i) above]

and $\text{var}(\bar{y})$ in SRSWOR = $\frac{\sigma^2}{n} = \frac{2.917}{2} = 1.458$

Hence,

$\text{var}(\bar{y})$ in SRSWR > $\text{var}(\bar{y})$ in SRSWOR.

2.6 MERITS OF SIMPLE RANDOM SAMPLE

(i) The personnel bias of the investigator cannot do any harm to the selection of items because the selection of items entirely depends upon chance.

(ii) Since the sample units are selected at random giving each unit an equal chance of being selected, the element of subjectivity or personal bias is completely eliminated. As such a simple random sample is more representative of the population as compared to the purposive sampling.

(iii) The investigator can assess the accuracy of his results because the sampling error is inversely proportional to square root of the number of items in the sample.

(iv) The statistician can ascertain the efficiancy of the estimates of the parameter by considering the sampling distribution of the stastistics (estimates), eg., $\bar{y}_n$ as an estimate of $\bar{Y}_n$ becomes more efficient as sample size n increases.

2.7 LIMITATIONS OF SIMPLE RANDOM SAMPLING

(i) The selection of a sample in simple random sampling requires an up-to-date frame. Frequently, it is near impossible to identify the units in the population before the sample is drawn and this restricts the use of simple random sampling.

(ii) For selecting a random sample, all the units of the population are to be numbered. In case of large population, the procedure of numbering is costly and time. Consuming in many cases, it is often impossible to number each unit of a population.

(iii) A simple random sample may result in selection of sampling units which are widely spread geographically and in such a case the cost of collecting the data may be much in terms of time and money.

(iv) In case of heterogeneous population, the random sample will fail to show the time characteristics of a population because some of the groups may not represented at all in a sample.

(v) A simple random sample might give most likely non-random results. For example, if we draw a random sample of size 13 cards from pack, we may get all cards from the same suit. However, the probability such happening is extremely small.

(vi) Simple random sampling usually requires larger sample size as compared to stratified random sampling.

2.8 SUMMARY

Simple random sampling is a method of selecting n units out of the N such that each one of the distinct samples has an equal chance of being selected. The simple random sample is drawn unit by unit. The units of the population are numbered 1 to N. A series of Random numbers between 1 to N is then drawn either by means of a table of random numbers or by means of a computer program that produces such a table. At any draw the process used must give an equal chance of drawing any numbers not already drawn in the population. The units that bear these n numbers constitute the sample.

When a number that has been drawn is removed from the population for all subsequent draws, this method is called simple random sampling without replacement. Random sampling with replacement is entirely feasible. Since no matter of what had already happened, all units of the population are given an equal chance of selection. The formulas for the variances and estimated variances of estimates made from the sample are often simpler when sampling is with replacement than it is without replacement. For this reason sampling with replacement is some times used in more complex sampling plans, although at first sight there seems little point in having the same unit two or more times in the sample.

2.9 EXERCISES

1) Define random sampling. Explain various methods of obtaining random sampling.

2) What are random sampling numbers ? Outline the different random number series and explain how these are used to select a simple random sample.

3) What is a simple random sample ? Mention the various methods of drawing a random sample.

4) Define S.R.S.(i) with replacement (ii) without replacement from a finite population. Derive the unbiased estimates of the population mean and its variance based on the above two methods.

5) How does sampling with out replacement differ from that with replacement? Which of these gives a lower value of the standard deviation of the sample mean ? Explain by considering samples of size two from a population consisting of five numbers 2, 3, 6, 8, 11.

6) Define random sampling. Why is it preferred to other methods of sample selection? How will you estimate the mean and the sampling error from a random sampling of a finite population.

7) Prove that in simple random sampling without replacement sample mean square is an unbiased estimate of the population mean square.

8) In selecting 3 unit simple random sampling without replacement from a population having 6 units with the values 1, 5, 8, 12, 15 and 19 show that the sample mean is an unbiased estimator of the population mean.

9) Show that in SRSWOR, the probability of selecting a specified unit of the population at any given draw is equal to the probability of selecting it at the first draw.

10) In SRSWOR, find the probability that

(i) i^{th} unit in the population is included in the sample.

(ii) i^{th} and j^{th} unit $(i \neq j)$ in the population are included in the sample.

2.10 REFERENCES

1. Basic Statistics - B.L. Agarwal
2. Methods of Statistical Analysis - P.S. Grewal
3. Fundamentals of Applied Statistics- S.C. Gupta & V.K. Kapoor
4. Sampling Techniques - Cochran

LESSON - 3

STRATIFIED RANDOM SAMPLING

LEARNING OBJECTIVES

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

* Understand the meaning and technique of Stratified Random Sampling
* Study the advantages of Stratified Random Sampling and its limitations.
* Notations and Terminology
* Estimation of population mean and its variance
* Studying the allocation of sample size
* Comparison of Stratified Random Sampling with Simple Random Sampling without Stratification.

LESSON OUTILINE

3.1 INTRODUCTION

In Simple Random Sampling, the population is treated as homogeneous and the desired number of items to be included in the sample are selected at random. But most of the populations are homogeneous, therefore, Simple Random Sampling fails to show the true characteristics of such population. In such cases, it is recommended to use stratified random sampling for selecting a sample. In Stratified Random Sampling, the whole population is divided into a number of homogeneous groups or strata on the basis of certain characteristic(s) of the sampling units. For example a city is divided into groups according to wards. If the sampling unit is a household, then the households may be divided into homogeneous groups according to the income per family. In

the opinion survey, persons may be divided into homogeneous groups according to their qualifications, age, sex, size of family etc. Each homogeneous group is known as stratum. From each stratum a random sample of required size is selected.

Generally stratification means division into parts. Auxiliary information related to the character under study may be used to divide the population with various groups such that (a) units with in each group are as homogeneous as possible and (b) the group means are as widely different as possible. Thus a population consisting of N sampling units is divided into K relatively homogeneous mutually disjoint (non-overlapping) subgroups, termed as strata, of sizes $N_1, N_2, \cdots\cdots, N_K$ such that $N = \sum_{i=1}^{K} N_i$. If a simple random sample (generally with out replacement) or size n_i, $(i = 1, 2, 3, \cdots, K)$ is drawn from each of the stratum respectively such that $n = \sum_{i=1}^{K} n_i$, the sample is termed as Stratified Random Sample of size n and the technique of drawing such a sample is called Stratified Random Sampling.

Remarks

1 : In Stratified Random Sampling the two important criteria are viz.,

(i) proper classification of the population in various strata, and

(ii) a suitable sample size from each stratum.

2 : The criterion which enables us to classify various sampling units into different strata is termed as 'stratifying factor' (S.F.). This S.F. is called effective if it divides the given population into different stratra which are homogeneous with in themselves and units in different strata are as unlike as possible.

3 : If the study relates to a simple character, it may be easy to choose a variate w.r.t. which the units of the population can be grouped to give homogeneous strata. But if one is dealing with multi-character study one faces the difficulty of choosing an appropriate way of stratification. In such a situation, many times, intuition, judgement of subject matter specialists can all be used effectively in setting up strata. If judgement is excercised in determining the strata, sample result will still be unbiased provided the sampling with in each stratum is carried out by random process. However, if the judgement is good, the sampling variance may be reduced. In many fields of highly varied distributions, stratification is an exceedingly valuable tool.

3.2 ADVANTAGES OF STRATIFIED RANDOM SAMPLING

(i) If the admissible error is given, we need a small sample resulting in reduction of the expenditure.

(ii) In the case where the cost of the survey is fixed, there is reduction in error.

(iii) It helps in determining the estimates for various groups, zones etc., i.e. for each stratum, separately.

(iv) Stratified sampling is convenient from the point of view of organisation, Linguistic zones are are often formed with each zone serving as a stratum.

(v) Sometimes various sampling schemes may be used to draw samples from different strata. But this creates problems in getting the overall estimates.

(vi) In an unstratified random sample some strata may be over-represented, others may be under represented while some may be excluded altogether. Stratified sampling ensures any desired representation in the sample of the various strata in the population. It over-rules the possibility of any essential group of the population being completely excluded in the sample. Stratified sampling thus provide more representative groups of sections of the population and is frequently regarded as the most efficient system of sampling.

(vii) Stratified sampling provides estimates with increased precision. Moreover, stratified sampling enables us to obtain the results of known precision for each of the stratum.

3.3 LIMITATIONS

(i) There is no guarantee that the precision of a stratified random sample would be greater than that of a simple random sample of equal size.

(ii) It is a tedious job to stratify a population into homogeneous groups. In many cases, information needed to set up the groups may not be available.

(iii) The procedure of setting up classes entaitils too much of time.

3.4 NOTATIONS AND TERMINOLOGY

Let K be the number of strata.

N_i = The number of sampling units in the i^{th} stratum, $(i = 1, 2, \cdots\cdots, K)$

$N = \sum_{i=1}^{K} N_i$, total number of sampling units in the population

n_i = The number of sampling units selected with SRSWOR from the i^{th} stratum.

$n = \sum_{i=1}^{K} n_i$, total sample size from all the strata.

Let $Y_{ij}, \left(j = 1, 2, \cdots, N_i, i = 1, 2, \cdots, K\right)$ be the value of the j^{th} unit in the i^{th} stratum.

$\bar{Y}_{Ni} =$ population mean of i^{th} stratum.

$$= \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij}$$

$\bar{Y}_N =$ population mean = $\frac{1}{N}\sum_i\sum_j Y_{ij} = \frac{1}{N}\sum_i^K N_i\ \bar{Y}_{Ni}$

$$= \sum_{i=1}^{K} P_i\ \bar{Y}_{Ni}$$

where $p_i = \frac{N_i}{N}$ is called the weight of the i^{th} stratum.

$S_i^2 =$ population mean square of the i^{th} stratum.

$$= \frac{1}{N_i - 1}\sum_{j=1}^{N_i}\left(Y_{ij} - \bar{Y}_{Ni}\right)^2,\ i = 1, 2, \cdots, K \quad \text{---------- (i)}$$

$Y_{ij} =$ value of j^{th} sampled unit i^{th} stratum.

$\bar{y}_{ni} =$ mean of sample selected from i^{th} stratum.

$$= \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{n_i}\right)^2;\ \left(i = 1, 2, \cdots, K\right) \text{---------- (ii)}$$

We shall consider the following two estimates of the population mean $\bar{Y}_N$, which are

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{K} n_i\ \bar{y}_{n_i} \quad \text{---------- (iii)}$$

$$\bar{y}_{ni} = \frac{1}{N}\sum_{i=1}^{K} N_i\ \bar{Y}_{ni} = \sum_{i=1}^{K} P_i\ \bar{y}_{ni} \quad \text{---------- (iv)}$$

the latter being weighted mean of the strata sample means, weights being equal to strata sizes.

These two estimates of the population mean are identified if $\frac{n_i}{n} = \frac{N_i}{N}$

$$\Rightarrow \frac{n_i}{N_i} = \frac{n}{N} \text{ (constant)}$$

$$\Rightarrow n_i = C\ N_i \text{ where } C = \frac{n}{N}$$

$$\Rightarrow n_i \propto N_i \quad \text{---------- (v)}$$

3.5 ESTIMATE OF POPULATION MEAN AND ITS VARIANCE

Theorem 3.5.1 : $\bar{y}_{st}$ is an unbiased estimate of the population mean $\bar{Y}_N$ i.e.,

$$E(\bar{y}_{st}) = \bar{Y}_N \text{ ---------- (vi)}$$

Proof : Since the sample in each of the stratum is a simple random sample, we have

$$\therefore E(\bar{y}_{st}) = \frac{1}{N}\sum_{i=1}^{K} N_i \cdot E(\bar{y}_{ni}) = \frac{1}{N}\sum_{i=1}^{K} N_i \bar{Y}_{N_i} = \bar{Y}_N$$

Theorem 3.5.2 : $\text{var}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{K} N_i (N_i - n_i)\frac{S_i^2}{n_i}$

$$= \sum_{i=1}^{K} P_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^2 \text{ ---------- (vi)}$$

Proof : Since the sample in each stratum is simple random simple without replacement, we have

$$V(\bar{y}_{ni}) = \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^2 \text{ ---------- } (*)$$

where S_i^2 is as defined in (xii)

$$\text{var}(\bar{y}_{st}) = \text{var}\left(\sum_{i=1}^{K} p_i \bar{y}_{ni}\right) = \sum_{i=1}^{K} p_i^2 \cdot \text{var}(\bar{y}_{ni})$$

the covariance terms vanish since the samples from different strata are independent.

$$\therefore \text{var}(\bar{y}_{st}) = \sum_{i=1}^{K} p_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) s_i^2 \text{ from } (*)$$

$$= \frac{1}{N^2}\sum_{i=1}^{K} N_i (N_i - n_i)\frac{S_i^2}{n_i}$$

Thus we see that $\text{var}(\bar{y}_{st})$ depends on S_i^2, the heterogeneous with in the strata. Thus, if S_i^2, are small, i.e. strata are homogeneous then stratified sampling provides estimates with greater precision.

REMARK

1. In general S_i^2 are not known, since a simple random sample is drawn from each stratum, $E\left(S_i^2\right) = S_i^2;\ i = 1, 2, \cdots, K$.

   Accordingly an unbiased estimate of the $V\left(\bar{y}_{st}\right)$ is given by

$$E_{st}\left[V\left(\bar{y}_{st}\right)\right] = \sum_{i=1}^{K}\left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 \cdot s_i^2$$

$$= \frac{1}{N^2}\sum_{i=1}^{K} N_i\left(N_i - n_i\right)\frac{S_i^2}{n_i} \quad \text{---------- (vii)}$$

2. With stratified random sampling, there is in general no single finite population correlation factor since $\left(N_i - n_i\right)/N_i$ may be different for different values of $i = 1, 2, \cdots, K$. In case all the f.p.c. are negligible then from (vii) and (viii), we get

$$V\left(\bar{y}_{st}\right) = \sum_{i=1}^{K}\frac{N_i^2}{N^2}\cdot\frac{S_i^2}{n_i} \quad \text{---------- (viii)}$$

and $EstVar\left(\bar{y}_{st}\right) = \sum_{i=1}^{K}\frac{N_i^2}{N^2}\cdot\frac{S_i^2}{n_i}$ ---------- (viii a)

3.6 METHODS OF OBTAINING STRATIFIED RANDOM SAMPLE

The main problem in stratified random sampling is to determine the size of various sub-samples which are to be drawn from different sub-groups. The allocation of sample size from each sub-group is provided under the following heads:

(a) Proportional allocation (b) Optimum allocation.

(a) Proportional Allocation : As it indicates, proportional allocation means that we select a small sample from a small stratum and a large sample from a large stratum. The sample size of each stratum is fixed in such a way that for all the strata the sample size allocation n_i is called allocation proportional if the sample fraction is constant for each stratum, i.e.,

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \cdots\cdots = \frac{n_K}{N_K} = \frac{\Sigma n_i}{\Sigma N_i} = \frac{n}{N} = C \text{ (Constant)}$$

$$\Rightarrow \frac{n_i}{N_i} = C = \frac{n}{N} \Rightarrow n_i \propto N_i, \left(i = 1, 2, \cdots, K\right) \text{ --------- (ix)}$$

Thus, in proportional allocation each stratum is represented according to its size.

In proportional allocation, the $V(\bar{y}_{st})$, is given by

$$\text{var}(\bar{y}_{st})_{\text{prop}} = \frac{1}{N^2}\sum_{i=1}^{K} N_i(N_i - n_i)\frac{S_i^2}{n_i}$$

$$= \sum_{i=1}^{K} \frac{N_i}{N^2}\left[\frac{N_i}{n_i} - 1\right] S_i^2$$

$$= \sum_{i=1}^{K} \frac{P_i}{N}\left(\frac{N}{n} - 1\right) S_i^2 \qquad \left(\because \frac{N_i}{n_i} = \frac{N}{n}\right)$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{K} P_i S_i^2 \quad \text{---------- (ix a)}$$

(b) Optimum Allocation : The principle of optimum allocation is the most reliable method of selecting a sample from a heterogeneous population. Here each sub-group is represented in a sample according to its size and variability. The principle of optimum allocation should be employed for selecting a sample in case the following conditions are satisfied :

(a) It is known that stratum standard deviations differ widely from each other.

(b) The stratum standard deviations are determined with precision.

As discussed, variance of estimated mean depends on n_i that is arbitrarily fixed. The allocation problem is two fold :

(a) Minimise the variance (i.e. maximise the precision) of the estimate

(i) for fixed sample size n and (ii) for fixed cost.

(b) Minimise the total cost for fixed desired precision.

The allocation of n_i's to various strata in accordance with the above principles is known as optimum allocation. In optimum allocation n_i's are obtained that satisfy the following criteria :

(1) $\text{var}(\bar{y}_{st})$ is minimum for fixed n

(2) $\text{var}(\bar{y}_{st})$ is minimum for fixed total cost C (say)

(3) Total cost C is minimum for fixed value of $\text{var}(\bar{y}_{st}) = V$ (say)

Cost function : In any sample survey, value of information on the experimental units must always be balanced against the cost of obtaining it. In stratified sampling it may cost more to obtain information about a sample in one stratum than in another. For example, interviewing people in rural areas is going to be more costly because of travel expenses than interviewing people in urban area. Thus, in its simplest form the cost function 'C' in stratified sampling may be given by the linear model

$$C = a + \sum_{i=1}^{K} C_i\, n_i \quad \text{---------- (x)}$$

where 'a' is the overhead cost and C_i is the cost per unit in the i^{th} stratum.

Theorem 3.6.1 : $\text{var}(\bar{y}_{st})$ is minimum for fixed total size of the sample (n) if $n_i \propto N_i\, S_i$.

Proof : Here the problem is to minimise

$$\text{var}(\bar{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{K} N_i (N_i - n_i) \frac{S_i^2}{n_i} \quad \text{---------- (xi)}$$

subject to the condition

$$\sum_{i=1}^{K} n_i = n \text{ (fixed) ---------- (xia)}$$

This is equivalent to minimising

$$\phi = \text{var}(\bar{y}_{st}) + \lambda \left(\sum_{i=1}^{K} n_i - n \right)$$

$$= \frac{1}{N^2} \sum_{i=1}^{K} N_i (N_i - n_i) \frac{S_i^2}{n_i} + \lambda \left(\sum_{i=1}^{K} n_i - n \right)$$

Unconditionally for variation in n_i, λ being unknown Lagrange's Multiplier. For extremum, we should take

$$\frac{\partial \phi}{\partial n_i} = -\frac{N_i^2 s_i^2}{N^2 n_i^2} + \lambda = 0$$

$$\Rightarrow n_i = \frac{N_i S_i}{N\sqrt{\lambda}} \quad \text{---------- (xii)}$$

Also $\dfrac{\partial^2 \phi}{\partial n_i^2} = \dfrac{2 N_i^2 S_i^2}{N^3 n_i^3} > 0$

which implies that n_i's given by (xii) provided a minimum of ϕ

To determine λ, we sum (xi) over i to K to get

$$\sqrt{\lambda} = \frac{\Sigma N_i S_i}{n N}$$

substituting in (xi), we finally get

$$n_i = \frac{n N_i S_i}{\sum_{i=1}^{K} N_i S_i} \quad \text{---------- (xiii)}$$

Thus, in optimum allocation for a fixed total sample size, we have

$$n_i \propto N_i S_i \quad \text{-------- (xiii a)}$$

This is known as Neyman's formula for optimum allocation. This suggests that greater the value of $N_i S_i$ for a given stratum, greater is the number of sampling units to be selected from the stratum in order to obtain the most precise estimate of the population mean.

Substituting the value of n_i from (xiii) in (vi), we get

$$\text{var}\left(\bar{y}_{st}\right)_{\text{opt}} = \frac{1}{N^2} \sum_{i=1}^{K} N_i \left(\frac{N_i}{n_i} - 1 \right) S_i^2$$

$$= \frac{1}{N^2} \sum_{i=1}^{K} N_i \left[\frac{\sum_{i=1}^{K} N_i S_i}{n S_i} - 1 \right] S_i^2$$

$$= \frac{1}{N^2} \left[\frac{1}{n} \left(\sum_{i=1}^{K} N_i S_i \right)^2 - \sum_{i=1}^{K} N_i S_i^2 \right] \quad \text{---------- (xiv)}$$

$$= \frac{1}{n} \left(\sum_{i=1}^{K} P_i S_i \right)^2 - \frac{1}{N} \sum_{i=1}^{K} P_i S_i^2 \quad \text{---------- (xiv a)}$$

Theorem 3.6.2 : In Stratified Random Sampling with given cost function of the form

$$C = a + \sum_{i=1}^{K} C_i n_i,$$

$\text{var}\left(\bar{y}_{st}\right)$ is minimum

if $n_i \propto \frac{N_i S_i}{\sqrt{C_i}}$

Proof : Here, we have to minimise $\text{var}(\bar{y}_{st})$ subject to the condition

$$\sum_{i=1}^{K} C_i n_i = C - a$$

Equivalently, we have to minimise.

$$\phi = \text{var}(\bar{y}_{st}) + \lambda\left[\sum_{i=1}^{K} C_i n_i - C + a\right]$$

$$= \frac{1}{N^2}\sum_{i=1}^{K} N_i (N_i - n_i)\frac{S_i^2}{n_i} + \lambda\left[\sum_{i=1}^{K} C_i n_i - C + a\right]$$

unconditionally for variations in n_i, λ being Lagranges multiplier.

Differentiating ϕ w.r.t. to n_i and equating to 0, we get for extremum

$$\frac{\partial\phi}{\partial n_i} = -\frac{N_i^2 S_i^2}{N^2 n_i^2} + \lambda C_i = 0$$

$$\Rightarrow n_i = \frac{N_i S_i}{N\sqrt{C_i}\sqrt{\lambda i}} \text{ ---------- (xv)}$$

Summing both sides over i from 1 to K, we get

$$\sqrt{\lambda} = \frac{\sum_{i=1}^{K}\left[N_i S_i / \sqrt{C i}\right]}{n N}$$

substituting in (xiv), we get

$$ni = \frac{n Ni Si/\sqrt{Ci}}{\sum_{i=1}^{K}\left[Ni Si/\sqrt{Ci}\right]} \text{ ---------- (xvi)}$$

Thus, in optimum allocation for a fixed cost

$$n_i \propto \frac{N_i S_i}{\sqrt{Ci}} \text{ ---------- (xvi a)}$$

This leads to the following important conclusions :

A larger sample would be required from a stratum if

(i) Stratum size (Ni) is large

(ii) Stratum variability (Si) is large

(iii) Sample cost per unit is low in the stratum

From (xv), we observe that 'n_i' is given interms of n. The value of n depends upon whether the sample is selected so as to meet a specified total cost C or a given specified $\text{var}(\bar{y}_{st})$.

Case 1 : For a fixed C, substituting (xvi) in (x) the optimum sample size n is given by

$$C = a + \Sigma \frac{nCi\, NiSi/\sqrt{Ci}}{\sum_{i=1}^{K}\left(Ni\, Si/\sqrt{Ci}\right)}$$

$$n = \frac{C - a\sum_{i=1}^{K}\left(Ni\, Si/\sqrt{Ci}\right)}{\sum_{i=1}^{K}\left(Ni\, Si\sqrt{Ci}\right)} \text{ ---------- (xvii)}$$

Allocation of n_i's to the various strata in accordance with formula (xvi) is known as optimum allocation. In particular $C_i = C_0 \forall$, $i = 1,2,\cdots\cdots K,$ the allocation is termed as Neyman's allocation.

In this case, we get

$$C - a = C_0 \sum_{i=1}^{K} n_i = C_0 n$$

and consequently the optimum allocation for fixed cost reduces to the optimum allocation for fixed sample size n. Accordingly from (xvi), the optimum values of n_i are given by

$$n_i = \frac{N_i\, S_i}{\sum_{i=1}^{K} N_i\, S_i}$$

which is Neyman's formula for optimum allocation, already obtained in (xiii).

Further, if we take $S_1 = S_2 = \cdots = S_K$, then optimum values of $n_i, (i = 1,2,\cdots,K)$ are given by the formula.

$$n_i = \frac{n}{N} \cdot N_i \Rightarrow \frac{n_i}{N_i} = \frac{n}{N}$$

which is Bowley's formula of proportional allocation already given in (ix)

Case ii : Suppose that n is to be selected so as to meet a given specified $\text{var}(\bar{y}_{st}) = V$ (say) at minimum cost.

Here we have

$$\text{var}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{K} N_i (N_i - n_i)\frac{S_i^2}{n_i} = V \text{ (say)}$$

$$\Rightarrow \frac{1}{N^2}\left[\sum_{i=1}^{K}\frac{N_i^2 Si^2}{n_i} - \sum_{i=1}^{K} N_i S_i^2\right] = V_0$$

$$\Rightarrow \frac{1}{N^2}\left[\sum_{i=1}^{K}\frac{N_i^2 S_i^2}{n_i} = V_0 + \frac{1}{N^2}\sum_{i=1}^{K} N_i S_i^2 \cdots\cdots\right] \text{ --------- (xviii)}$$

Substituting the optimum values in (xvi) in (xviii), we get the formula for minimum sample size required for estimation of the population mean with fixed variance V under optimum allocation as

$$\frac{1}{N^2}\Sigma\left[N_i^2\, S_i^2 \cdot \frac{\sum_{i=1}^{K}\left(N_i S_i/\sqrt{C_i}\right)}{n\, N_i S_i/\sqrt{C_i}}\right] = V_0 + \frac{1}{N^2}\sum_{i=1}^{K} N_i S_i^2$$

$$\Rightarrow n = \frac{\sum_{i=1}^{K}\left(Ni\,Si/\sqrt{Ci}\right)\sum_{i=1}^{K}\left(Ni\,Si/\sqrt{Ci}\right)}{N^2 V0 + \sum_{i=1}^{K} Ni\,Si^2} \text{ ---------- (xxi)}$$

Taking $Ci = C_0 \forall i = 1, 2, \cdots\cdots, n$ we obtain an expression for the optimum sample size for fixed variance V, under Neyman's allocation as

$$n = \frac{\left(\sum_{i=1}^{K} Ni\,Si\right)^2}{N^2 V_0 + \sum_{i=1}^{K} Ni\,Si^2} \text{ ---------- (xx)}$$

3.7 COMPARISON OF STRATIFIED RANDOM SAMPLING WITH SIMPLE RANDOM SAMPLING

Proportional Allocation Vs Simple Random Sampling : The variance of the estimate of population mean in stratified random sampling with proportional allocation and in unstratified simple random sampling is given respectively by

$$\text{var}\left(\bar{y}_{st}\right)_P = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{K} P_i\, Si^2 \quad \text{---------- (xxi)}$$

and $$\text{var}\left(\bar{y}_n\right)_R = \left(\frac{1}{n} - \frac{1}{N}\right)S^2 \quad \text{---------- (xxia)}$$

Where $S^2 = \frac{1}{N-1}\sum_{i=1}^{K}\sum_{j=1}^{N_i}\left(Y_{ij} - \bar{Y}_N\right)^2$

inorder to compare (xxi) and (xxia), we shall first express S^2 interms of Si^2, we have

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{K}\sum_{j=1}^{N_i}\left(Y_{ij} - Y_{Ni} + \bar{Y}_{Ni} - \bar{Y}_N\right)^2$$

$$+2\sum_i\left[\left(\bar{Y}_{Ni} - \bar{Y}_N\right)\sum_{j=1}^{N_i}\left(Y_{ij} - \bar{Y}_{Ni}\right)\right]$$

$$= \sum_{i=1}^{K}\left(N_i - 1\right)Si^2 + \sum_{i=1}^{K} Ni\left(\bar{Y}_{Ni} - \bar{Y}_N\right)^2$$

The product terms vanish since $\sum_{j=1}^{Ni}\left(Y_{ij} - \bar{Y}_N\right)^2$, being the algebric sum of the deviation from mean, is zero.

If we assume that Ni and consequently N are sufficiently large so that we can take $Ni - 1 \simeq Ni$ and $N - 1 \simeq N$, then we get

$$NS^2 \simeq \sum_{i=1}^{K} Ni\, Si^2 + \sum_{i=1}^{K} Ni\left(\bar{Y}_{Ni} - \bar{Y}_N\right)^2$$

$$\Rightarrow S^2 \simeq \sum_{i=1}^{K} P_i\, S_i^{\,2} + \sum_{i=1}^{K} P_i\left(\bar{Y}_{Ni} - \bar{Y}_N\right)^2 \quad \text{---------- (xxi b)}$$

substituting in (xxia), we get

$$\text{var}\left(\bar{Y}_n\right)_R \simeq \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{K} P_i\, S_i^{\,2} + \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{K} P_i\left(\bar{Y}_{Ni} - \bar{Y}_N\right)^2$$

$$\simeq \text{var}\left(Y_{st}\right)_P + \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{K} P_i\left(\bar{Y}_{Ni} - \bar{Y}_N\right)^2 \quad \text{---------- (xxii)}$$

$$\Rightarrow \text{var}\left(\bar{y}_n\right)_R \geq \text{var}\left(\bar{y}_{st}\right)P$$

which leads to an important conclusion : "Greater the difference in the stratum means, greater is the gain in precision of stratified sampling with proportional allocation over unstratified simple random sampling".

Neyman's Allocation Vs. Proportion allocation :

Writing $\text{var}\left(\bar{y}_{st}\right)_P$ and $\text{var}\left(\bar{y}_{st}\right)_N$ for the variance of the estimate of population mean in stratified sampling with proportional allocation and Neyman's optimum allocation respectively, we get from (ixa) and (xiv a)

$$\text{var}\left(\bar{y}_{st}\right)_p - \text{var}\left(\bar{y}_{st}\right)_N = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{K} P_i S_i^2 - \left(\frac{1}{n}\left(\Sigma P_i S_i\right)^2 - \frac{1}{N}\Sigma P_i {S_i}^2\right)$$

$$= \frac{1}{n}\left[\sum_{i=1}^{K} P_i \cdot {S_i}^2 - \left(\sum_{i=1}^{K} P_i S_i\right)^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{K} P_i {S_i}^2 - \left(\sum_{i=1}^{K} P_i S_i\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{K} P_i\left(S_i - \bar{S}\right)^2 \quad \text{---------- (xxiii)}$$

Where, $\bar{S} = \sum_{i=1}^{K} P_i S_i = \frac{1}{N}\sum_{i=1}^{K} N_i S_i$ is weighted mean of the stratum standard deviation, the weights being equal to the stratum sizes.

Since R.H.S. in (xxii) is non-negative, we get

$$\text{var}\left(\bar{y}_{st}\right)_p \geq \text{var}\left(\bar{y}_{st}\right)_N \quad \text{---------- (xxiia)}$$

From (xxii) and (xxiia) we conclude that Neyman's optimum allocation gives better estimates than proportional allocation and greater the difference between the stratum standard deviations, more is the given precision in Neyman's allocation over proportional allocation.

Neyman's Allocation Vs Simple Random Sampling :

Substituting for $\text{var}\left(\bar{y}_{st}\right)_p$ from (xxiii) in (xxii), we get

$$\text{var}\left(\bar{Y}_n\right)_R = \text{var}\left(\bar{Y}_{st}\right)_N + \frac{1}{n}\sum_{i=1}^{K} P_i\left(S_i - \bar{S}\right)^2$$

$$+\left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{K} P_i\left(\bar{Y}_{N_i} - \bar{Y}_N\right)^2 \text{ ---------- (xxiv)}$$

$$\Rightarrow \text{var}\left(\bar{y}_n\right)R \geq \text{var}\left(\bar{y}_{st}\right)N \text{ ---------- (xxiv)}$$

We observe from (xxiv) that as we change from unstratified simple random sampling with Neyman's allocation, the gain in precision of the estimates results from two components, viz.,

(i) The elimination of the difference between stratum means

and (ii) The estimation of the difference among stratum standard deviations.

Remark : From (xxiiia) and (xxiva), we get $\text{var}\left(\bar{y}_n\right)_k \geq \text{var}\left(\bar{y}_{st}\right)_p \geq \text{var}\left(\bar{y}_{st}\right)_N$ --------- (xxv)

Example 3.7.1 : Sample of 30 students is to be drawn from a population consisting of 300 students belonging to two colleges A and B. The means and s.d of their Marks are given below.

	Total Number of students (N_i)	Mean $(\bar{Y}_{Ni})$	Standard deviation (σ_i)
College A	200	30	10
College B	100	60	40

How would you draw the sample using proportional allocation techinque ? Hence obtain the variance of estimate of the population mean and compare its efficiency with simple random sampling without replacement.

Solution : If we regard the colleges A and B as two different strata then the problem is to draw a stratified random sample of size 30 using the technique of proportional allocation. In proportional allocation, we have

$$n_i = \frac{n}{N} \cdot N_i = \frac{30}{300} N_i$$

$$n_1 = \frac{1}{10} \times 200 = 20$$

$$n_2 = \frac{1}{10} \times 100 = 10$$

Thus, the required sample sizes for the colleges A and B are 20 and 10 respectively, for

obtaining $\text{var}(\bar{y}_{st})_p$ and $\text{var}(\bar{y}_n)_R$ we make the calculations as shown below in the table.

College	Ni	$\bar{Y}_{Ni}$	σ_i	σ_i^2	$s_i^2 = \frac{N}{N-1}\sigma_i^2$	$Ni\bar{Y}_{Ni}$	$Ni\bar{Y}_{Ni}^2$	NSi^2	$Ni\sigma_i^2$
A	200	30	10	100	100.3344	6000	180000	20066.88	20000
B	100	60	40	1600	1605.3511	6000	180000	16053.11	16000
Total		90		1700	1705.6855	12000	540000	180601.99	180000

Here $N = 300$ and $n = 30$

$$\bar{Y}_N = \frac{1}{N}\sum_{i=1}^{2} Ni\bar{Y}_{N_i} = \frac{1}{300}\times 12000 = 40$$

$$S^2 = \frac{1}{N-1}\sum\sum\left(Y_{ij} - \bar{Y}_N\right)^2$$

$$= \frac{1}{N-1}\left[\sum_i N_i\,\sigma_i^2 + \sum_i Ni\left(\bar{Y}_{Ni} - \bar{Y}_N\right)^2\right]$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{2} Ni\sigma_i^2 + \sum_{i=1}^{2} Ni\bar{Y}_{Ni}^2 - N\bar{Y}N^2\right]$$

$$= \frac{1}{299}\left[1,80,000 + 5,40,000 - 4,80,000\right] = \frac{2,40,000}{299}$$

In SRSWOR

$$\text{var}(\bar{y}_n)_k = \frac{N-n}{N_n}S^2 = \frac{300-30}{300\times 30} * \frac{2,40,000}{299}$$

$$= \frac{7,200}{299} = 24.08$$

$$\text{var}(\bar{y}_{st})_P = \frac{N-n}{n\cdot N^2}\sum_{i=1}^{2} NiSi^2 = \frac{300-30}{30\times(300)^2}\times 1,80,601.99$$

$$= \frac{270\times 1,80,601.99}{27,00,000} = 18.0602$$

$\therefore$ Gain in efficiency due to stratified random sampling (proportional allocation) over SRSWOR is given by

$$\frac{\text{var}(\bar{y}_n) - \text{var}(\bar{y}_{st})_p}{\text{var}(\bar{y}_{st})\,p} = \frac{24.08 - 18.06}{18.06} = 0.333$$

$\therefore$ Effeciency = 33.3%

3.8 SUMMARY

Random sampling is not always the best method of assessing the population. For example, in estimating the average income of the inhabitants of a city, there is a likelihood that more rich people or more poor pepole may be dominating a sample. For this purpose, it would be better first to divide the city with different strata, say according to localities, slums, middle class socities and bunglaw areas, business localities etc., and then to select individuals at random from each of the localities. This would ensure that all sections of the society are represented in the sample. The above sampling technique is known as "stratified sampling". The size of each group of strata should be proportionate to the relative importance in the population of the stratum represented by the group.

If the sampling is alone according to the rules of probability, the errors, that are likely to creep in, can be estimated and here lies the importance of random sampling. It is in this method that the rules of probability are applicable so that the statistics of the sample may be read to estimate the parameters of the populations. The results of the two samples may be compared to the test whether they have been drawn from the universe and whether a Hypothesis is to be rejected on the basis of results of a sample. The sampling errors can always be reduced by increasing the sample size; it means corresponding increase in cost and labour.

EXERCISE

1. Describe the procedure of stratified sampling.
2. Describe the advantages of stratified random sampling with illustrations. What are the various methods of allocating a sample in stratified sampling.
3. Work out Neyman's optimum allocation principle of units stratified random sampling.
4. What is stratified sampling. Explain the relative merits and demerits of stratified random sampling.
5. What is the estimator for population mean in case of stratified sampling ?
6. Prove that $\text{var}(\bar{y}_{st})_{opt} < \text{var}(\bar{y}_{st})_{prop} < \text{var}(\bar{y})_k$
7. $\text{var}(\bar{y}_{st})$ is minimum for fixed total size of the sample (n) if $ni \propto NiSi$
8. Explain optimum allocation & proportional allocation.

9. Prove that $\text{var}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{K} Ni(Ni - ni)\frac{Si^2}{ni}$

10. Give essential characteristics of stratified random sampling.

11. Discuss the conditions under which stratified random sampling is more suitable than simple random sampling.

12. A population consists of 12 units which are divided into three strata. The first stratum consists of observed values 1, 3, 4 the second stratum has value 7, 9, 12, 10 and third stratum has values 18, 21, 20, 23, 24. From each of the first and second stratum draw a sample of two units and from the third stratum draw a sample of three units. On the basis of sampled observations, find the mean of the stratified sample and calculate its variance $\text{var}(\bar{y}_{st})$.

3.9 REFERENCES

1. Basic statistics - B.L. Agarwal
2. Fundamentals of Applied Statistics - S.C. Gupta & V.K. Kapoor
3. Sampling Techniques - Cochran

LESSON - 4

SYSTEMATIC SAMPLING

LEARNING OBJECTIVES

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

- Understand the meaning of systematic sampling technique and method of a selecting a systematic sampling
- Merits and demerits of systematic sampling
- Notation and Terminology
- Variance of the estimated mean of systematic sampling
- Comparision with other methods

LESSON OUTLINE

4.1 INTRODUCTION

When the population units occur in a deck or sequence or line and a sample of size n is to be drawn, the population is divided in to n sequential groups and one unit is drawn from each group situated at equal distance. Like that, often we have to obtain the information from cards or registers in a serial order. Sometimes, we might need the sample of trees from a forest or houses in a city. In such a case, a sampling plan known as systematic sampling often works better than the simple random sampling. In order to draw a systematic sample of size n divide the population into n equal parts. Supposing each part to consist of K units, draw a random number from 1 to K. Let the selected number be j. Then select $j, j+k, (j+2k), \cdots (j+(n-1)k)$ th units. This constitutes a systematic sample of size n. For example serially numbered from 1 to 100 we want to draw a

sample of 5 units. Draw a random number from 1 to 20 since $K = 20$ and let the selected number be 7 i.e. $j = 7$. Then select units serial numbers 27, 47, 67, and 87. These units constitute a systematic sample of 5 units.

4.2 METHOD OF SELECTION OF SYSTEMATIC SAMPLING

Systematic sampling is a commonly employed technique if the complete and up-to-date list of the sampling units is available. This consists in selecting only the first unit at random, and the rest being automatically selected according to some predetermined pattern involving regular spacing of units. Let suppose that N sampling units are serially numbered from 1 to N in some order and a sample of size n is to be drawn such that

$$N = nK \Rightarrow K = \frac{N}{n} \quad \text{---------- (xxvi)}$$

where K, usually called the sampling interval, is an integer.

Systematic sampling consists in drawing a random number, say $i \leq K$ and selecting corresponding to this number every K^{th} unit subsequently. This systematic of size n will consist of the units.

$$i, i+K, i+2K, \cdots\cdots i+(n-1)K$$

The random number 'i' is called the random start and its value determines, as a matter of fact, the whole sample.

4.3 ADVANTAGES AND DISADVANTAGES OF SYSTEMATIC SAMPLING

ADVANTAGES :

(i) The method of selection is very simple and is not very expensive.

(ii) The sample is ever the distributed over the whole population and hence all contiguous parts of the population are represented in the sample.

(iii) Systematic sampling is operationally more convenient than simple random sampling or stratified random sampling. Time and work involved is also relatively much less.

(iv) Systematic sampling may be more efficient than simple random sampling provided the frame is arranged wholly at random.

(v) It has an advantage over other sampling plans because of its organising control of the field work.

DISADVANTAGES :

(i) The main disadvantage of systematic sampling is that samples are not in general random samples since the requirement for comparison of systematic sample with that of simple random sampling is rarely fulfilled.

(ii) If N is not a multiple of n, then

(a) the actual sample size is different from that required, and

(b) sample mean is not an unbiased estimate of the population mean. However, these disadvantages can be overcome by adopting a technique known as circular systematic sampling (C.S.S.)

(iii) If the variation in units in periodic, the units at regular intervals are correlated. In this situation the sample becomes highly top-sided and hence the estimates are biased.

(iv) No single reliable formula is available for estimating the standard error of sample mean. A formula is good enough if the population of the type it has been could be expected otherwise not. This is a great draw-back of systematic sampling.

(v) Systematic sampling may yield highly biased estimates if there are periodic features associated with the sampling interval i.e., if the frame has a periodic feature and K is equal to or a multiple of the period.

4.4 NOTATIONS AND TERMINOLOGY

Let y_{ij} denote the observation on the j^{th} unit of the i^{th} sample. $\left(i = 1, 2, \cdots, K;\ j = 1, 2, \cdots, n\right)$.

$\bar{y}_{i.} =$ Mean of the systematic sample

$$= \frac{1}{n}\sum_{j=1}^{n} y_{ij},\ \left(i = 1, 2, \cdots, K\right) \quad \text{---------- (xxvi)}$$

$\bar{Y}_{..} =$ The population mean.

$$= \frac{1}{nK}\sum_{i=1}^{K}\sum_{j=1}^{n} y_{ij} = \frac{1}{K}\sum_{i=1}^{n}\bar{Y}_{ij} \quad \text{---------- (xxvii)}$$

$S^2 =$ population mean square

$$= \frac{1}{N-1}\sum_{i=1}^{K}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}_{..}\right)^2 = \frac{1}{nK-1}\sum_{i=1}^{K}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}_{..}\right)^2$$

The K possible systematic samples.

Random Start	Sample composition units in the same	Probability	Mean
1	$1+K, 1+2K, \cdots\cdots, 1+jK, 1+(n-1)k$	$1/K$	$\bar{Y}_{1\cdot}$
2	$2+K, 2+2K, \cdots\cdots, 2+jK, 2+(n-1)k$	$1/K$	$\bar{Y}_{2\cdot}$
...	.....	.....	.....
...	.....	.....	.....
...	.....	.....	.....
i	$i+K, i+2K, \cdots\cdots, i+jK, i+(n-1)k$	i/K	$\bar{Y}_{i\cdot}$
...	.....	.....	.....
...	.....	.....	.....
K	$2K \cdots\cdots K+jK, \cdots\cdots\cdots, nK$	$1/K$	$\bar{Y}_{K\cdot}$

Thus K rows of the table give K - systematic samples. The columns of the above table are also sometimes referred to as n strata.

$$E\left[\bar{y}_{i\cdot}\right] = \frac{1}{K}\sum_{i=1}^{K}\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot} \text{ --------- (xxix)}$$

Thus, if $N = nK,$ the sample mean provides an unbiased estimate of the population mean

$\bar{Y}_{sys}$ = Mean of systematic sample = $\bar{y}_{i\cdot}$

$$\therefore \text{var}\left(\bar{y}_{sys}\right) = \frac{1}{K}\sum_{i=1}^{K}\left(\bar{Y}_i - \bar{Y}_{\cdot\cdot}\right)^2 \text{ ---------- (xxx)}$$

4.5 VARIANCE OF THE ESTIMATED MEAN

Theorem 4.5.1 :

$$\text{var}\left(\bar{y}_{sys}\right) = \frac{N-1}{N}S^2 - \frac{(n-1)K}{N}S^2_{wsy} \text{ ---------- (xxxi)}$$

where

$$S^2_{wsy} = \frac{1}{K(n-1)}\sum_{i=1}^{K}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2 \text{ ---------- (xxxi a)}$$

is the mean square among units which lie with in the same systematic sample.

Proof. : We have

$$(N-1)S^2 = \sum_{i=1}^{K}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_{..}\right)^2 = \sum_i\sum_j\left(y_{ij}-\bar{y}_i+\bar{y}_i-\bar{y}_{..}\right)^2$$

The co-variance terms vanishes, since

$$\sum_i\sum_j\left(y_{ij}-\bar{y}_{i.}\right)\left(\bar{y}_{i.}\right)\left(\bar{y}_{i.}-\bar{y}_{..}\right) = \sum_{i=K}^{K}\left[\left(\bar{y}_{i.}-\bar{y}_{..}\right)\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_{i.}\right)\right] = 0$$

$$\therefore (N-1)S^2 = \sum_i\sum_j\left(y_{ij}-\bar{y}_{i.}\right)^2 + n\sum_{i=1}^{K}\left(\bar{y}_i-\bar{y}_{..}\right)^2$$

$$= K(n-1)S^2 wsy + nK\ \mathrm{var}\left(\bar{y}_{sys}\right) \text{ from (xxx) and (xxxi)}$$

$$\Rightarrow \mathrm{var}\left(\bar{y}_{sys}\right) = \frac{N-1}{N}S^2 - \frac{K(n-1)}{N}S^2 wsy$$

Corollary : Systematic sampling (vs) simple random sampling

In case of SRSWOR,

$$\mathrm{var}\left(\bar{y}_{i.}\right) = \frac{N-n}{Nn}.S^2$$

where S^2 is defined in (xxviii)

$$\mathrm{var}\left(\bar{y}_n\right) - \mathrm{var}\left(\bar{y}_{sys}\right) = \left(\frac{N-n}{n} - N + 1\right)\frac{S^2}{N} + \frac{K(n-1)}{N}S^2{}_{wsy}$$

$$= \frac{K(n-1)}{N}.S^2{}_{wsy} - \frac{n-1}{n}\cdot S^2$$

$$= \frac{n-1}{n}\left(S^2{}_{wsy} - S^2\right) \qquad \because N = nK \text{ --------- (xxxii)}$$

The systematic sampling gives more precise estimate of the population mean as compared with SRSWOR if and only if

$$\mathrm{var}\left(\bar{y}_n\right) - \mathrm{var}\left(\bar{y}_{sys}\right) > 0$$

$$S_{sys}^2 > S^2 \text{ ---------- (xxxiii)}$$

This leads to the following important conclusion : "A systematic sample is more precise than a simple random sample without replacement if the mean square with in the systematic sample is larger than the population mean square". In otherwords, systematic sampling will yield better results only if the units with in the same sample are heterogeneous.

Theorem 4.5.2 :

$$\text{var}\left(\bar{y}_{sys}\right) = \frac{nK-1}{nK} \cdot \frac{S^2}{n}\left\{1+(n-1)\rho\right\} \quad \text{---------- (xxxvi)}$$

where ρ is the intra class correlation coefficient between the units of the same systematic sample and is given by

$$\rho = \frac{\sum_{i=1}^{K}\sum_{j\neq j'=1}^{n}\left(y_{ij}-\bar{y}_{..}\right)\left(\bar{y}_{ij'}-\bar{y}_{..}\right)}{nK(n-1)\,\sigma^2} \quad \text{---------- (xxx vii)}$$

$$= \frac{\sum_{i=1}^{K}\sum_{j\neq j'=1}^{n}\left(y_{ij}-\bar{y}_{..}\right)\left(\bar{y}_{ij'}-\bar{y}_{..}\right)}{(n-1)(m-1)S^2} \quad \text{---------- (xxx vii a)}$$

[since $N\sigma^2 = (N-1)S^2 \Rightarrow nK\sigma^2 = (nK-1)S^2$]

Proof : We have

$$\text{var}\left(\bar{y}_{sys}\right) = \frac{1}{K}\sum_{i=1}^{K}\left(\bar{y}_{i\cdot}-\bar{y}_{..}\right)^2 = \frac{1}{K}\sum_{i=1}^{K}\left(\frac{1}{n}\sum_{j=1}^{n}y_{ij}-\bar{y}_{..}\right)^2$$

$$\Rightarrow n^2K\ \text{var}\left(\bar{y}_{sys}\right) = \sum_{i=1}^{K}\left[\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_{..}\right)\right]^2 \quad \text{---------- (xxx vii b)}$$

$$= \sum_{i=1}^{K}\left[\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_{..}\right)^2 + \sum_{j\neq j'=1}^{n}\left(y_{ij}-\bar{y}_{..}\right)\left(y_{ij'}-\bar{y}_{..}\right)\right]$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{y}\left(y_{ij}-\bar{y}_{..}\right)^2 + \sum_{i=1}^{K}\sum_{j\neq j'=1}^{n}\left(y_{ij}-\bar{y}_{..}\right)\left(y_{ij'}-\bar{y}_{..}\right)$$

$$= (nK-1)S^2 + (n-1)(nK-1)S^2\rho \qquad \text{[using (xxvii) and (xxxvii a)]}$$

$$= (nK-1)S^2\left[1+(n-1)\rho\right]$$

$$\Rightarrow \text{var}\left(\bar{y}_{sys}\right) = \frac{n_{K-1}}{n_K}\cdot\frac{S^2}{n}\left[1+(n-1)\rho\right]$$

Thus, we see that a positive intraclass correlation between the units of the same sample inflates the variability of the estimates. Due to the multiplier $(n-1)$, the increase is quite significant even for small +ve values of ρ.

4.6 IF THE POPULATION CONSISTS OF A LINEAR TREND, THEN PROVE

$$Var\left(\bar{y}_{st}\right) \le \text{var}\left(\bar{y}_{sys}\right) \le \text{var}\left(\bar{y}_n\right)_R$$

Sol : Let us suppose that the population has the linear trend given by the model

$$Y_i = i : (i = 1, 2, \cdots\cdots N)$$

then $\sum_{i=1}^{N} Y_i = \sum_{i=1}^{N} i^2 = \frac{N(N+1)}{2}$

$$\sum_{i=1}^{N} Y_i^2 = \sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}$$

$$\sum_{i=1}^{N} Y_i^2 = \sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}$$

$$\bar{Y}_N = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{N+1}{2}$$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}_N\right)^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} Y_i^2 - N\bar{Y}_N^2\right]$$

$$= \frac{1}{N-1}\left[\frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{4}\right]$$

$$= \frac{N(N+1)}{12} \text{ (as simplification)}$$

$$\text{var}\left(\bar{y}_n\right)_R = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 = \frac{N-n}{Nn} S^2$$

$$= \frac{n(K-1)}{n^2 K} - \frac{nK(nK+1)}{12} = \frac{(K-1)(nK+1)}{12} \quad \text{---------- (1)}$$

Here we take the usual notation $N = nK$, since comparison has to made with systematic sampling also

$$\text{var}\left(\bar{y}_{st}\right) = \frac{K-1}{n^2 K} \sum_{j=1}^{n} S_i^2$$

$$\therefore \text{var}\left(\bar{y}_{st}\right) = \sum_{j=1}^{n} \left(1 - \frac{1}{K}\right) \frac{1}{n^2} S_i^2 = \frac{K-1}{n^2 K} \sum_{j=1}^{n} S_i^2$$

We have $S^2 = \frac{N(N+1)}{12}$

for population of N units. Since j^{th} stratum consists of K units, we have

$$S_i^2 = \frac{K(K+1)}{12}$$

$$\therefore \text{var}\left(\bar{y}_{st}\right) = \frac{K-1}{n^2 K} \cdot \frac{nK(K+1)}{12} = \frac{K^2-1}{12} \quad \text{---------- (2)}$$

For finding out $\text{var}\left(\bar{y}_{sys}\right)$, we have

$\bar{y}_{i\bullet} =$ mean of the values of i^{th} sample

$$= \frac{1}{n} \sum_{j=1}^{n} Y_{ij}$$

$$= \frac{1}{n}\left[i + (i+K) + (i+2K) + (i+3K) + \cdots + (n-1)K \right]$$

$\because$ of linear model [$*$]

$$= \frac{1}{n}\left[ni + \{1 + 2 + \cdots + n-1\} K \right]$$

$$= i + \frac{n-1}{2}K \quad \text{---------- (3)}$$

Also $\bar{Y}_{..} = \bar{Y}_N = \frac{N+1}{2} = \frac{nK+1}{2} \cdots\cdots$ (4)

$$\Rightarrow \bar{Y}_{i.} - \bar{Y}_{..} = i - \frac{K+1}{2}$$

$$\therefore \text{var}\left(\bar{y}_{sys}\right) = \frac{1}{K}\sum_{i=1}^{K}\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2 = \frac{1}{K}\sum_{i=1}^{K}\left(i - \frac{K+1}{2}\right)^2$$

$$= \frac{1}{K}\sum_{i=1}^{K}\left[i^2 + \left(\frac{K+1}{2}\right)^2 - 2\left(\frac{K+1}{2}\right)i\right]$$

$$= \frac{1}{K}\cdot\sum_{i=1}^{n} i^2 + \left(\frac{K+1}{2}\right)^2 - \frac{(K+1)}{K}\sum_{i=1}^{K} i$$

$$= \frac{(K+1)(2K+1)}{6} + \frac{(K+1)^2}{4} - \frac{(K+1)^2}{2}$$

$$= \frac{K^2-1}{12} \text{ (on simplification) ---------- (5)}$$

From (1), (2) and (5), we get

$$\text{var}\left(\bar{y}_{st}\right) : \text{var}\left(\bar{y}_{sys}\right) : \text{var}\left(\bar{y}_n\right) :: \frac{K+1}{n} : K+1 : nK+1$$

$$\cong \frac{1}{n} : 1 : n \text{ (approx)}$$

$$\Rightarrow \text{var}\left(\bar{y}_{st}\right) \le \text{var}\left(\bar{y}_{sys}\right) \le \text{var}\left(\bar{y}_n\right)$$

Thus, if the population is of a linear trend then stratified random sampling is most effective (with systamatic sampling as the next best) in eliminating the effect of linear trend.

Example 4.6.1 : The data below for a small artificial population which exhibits a fairly steady rising trend. Each column represents a systematic sample and the rows are the strata. Compare the precision of systematic sampling. Random sampling and stratified sampling.

Data for 10 systematic samples with $n = 4,\ K = 10,\ N = nK = 40$

Systematic Sample Number

Strata	1	2	3	4	5	6	7	8	9	10
I	0	1	1	2	5	4	7	7	8	6
II	6	8	9	10	13	12	15	16	16	17
III	18	19	20	20	24	23	25	28	29	27
IV	26	30	31	31	33	32	35	37	38	38

Solution : We have

$$\text{var}\left(\bar{y}_{sys}\right) = \frac{1}{K}\sum_{i=1}^{K}\left(\bar{y}_i - \bar{y}_{..}\right)^2 = \frac{1}{K}\left[\sum_{i=1}^{K}\bar{y}_{i.}^{\,2} - K\bar{y}_{..}^{\,2}\right]$$

$$= \frac{1}{n^2 K}\left[\sum_{i=1}^{K}\left(n\bar{y}_{i.}\right)^2 - n^2 K\,\bar{y}_{..}^{\,2}\right] \text{ ---------- } (*)$$

$$\text{var}\left(\bar{y}_{st}\right)^2 = \frac{1}{nK}\cdot S_{wst}^2 \text{ ---------- } (**)$$

Where $S_{wst}^2 = \frac{1}{n(K-1)}\sum_i\sum_j\left(y_{ij} - \bar{y}_{.j}\right)^2$

is the within (stratum) mean sum of squares

$$\text{var}\left(\bar{y}_n\right)_k = \left(\frac{1}{n} - \frac{1}{N}\right)S^2 \text{ ---------- } (***)$$

where $S^2 = \frac{1}{N-1}\sum_i\sum_j\left(y_{ij} - \bar{y}_{..}\right)^2$ is the total mean sum of squares. The calculation can be best worked out by the analysis of variance technique as explained below

$N = 40,\ n = 4,\ K = 10$

Systematic Sample Numbers

Strata	1	2	3	4	5	6	7	8	9	10	Strata total $T_{.j}$
I	0	1	1	2	5	4	7	7	8	6	41
II	6	8	9	10	13	12	15	16	16	17	122
III	18	19	20	20	24	23	25	28	29	27	233
IV	26	30	31	31	33	32	35	37	38	38	331
Sample total $n\bar{y}_{i.}$	50	58	61	63	75	71	82	88	91	88	$G = \Sigma T_{.j}$ = 727

Total $S \cdot S = \Sigma\Sigma Y_{ij}^2$ - correction factor

$$= 18,527 - \frac{5,28,529}{40} = 5,313.775$$

Between strata $S.S = \sum_{j=1}^{K} \frac{T_{.j}^2}{K} - C.F.$ ------ C.F.

$$= \frac{1,80,415}{10} - 1,3213.225 = 4,828.275$$

with in strata S.S. = Total S.S. - Between strata S.S.

=5313.775 - 4,828.275 = 485.5

$$S_w^2 st = \frac{\text{with in strata . } S.S.}{n(K-1)} = \frac{485.5}{36} = 13.486$$

$$S^2 = \frac{\text{Total } S.S.}{N-1} = \frac{5,313.775}{39} = 136.251$$

ANOVA TABLE

Source of variation	d.f.	S.S.	M.S.S.
Between strata	4 - 1 = 3	4,828.275	
Within strata	39 - 3 = 36	485.50	$\frac{485.5}{36} = 13.486 = S_{wst}^2$
Total	40 - 1 = 39	5,313.775	$\frac{5,313.775}{39} = 136.851 = S^2$

From $(*)$ $\text{var}(\bar{y}_{sys}) = \frac{1}{160}\left[547/3 - 160 \times (18.17)^2\right]$

$$= \frac{1}{160}\left[54713 - 160 \times 330.1489\right]$$

$$= \frac{1}{160}\left[54713 - 52823.824\right] = 11.807$$

From $(**)$ and $(***)$, we get respecitvely

$$\text{var}(\bar{y}_{st}) = \left(\frac{1}{4} - \frac{1}{40}\right)13.486 = 3.034$$

$$\text{var}(\bar{y}_n)R = \left(\frac{1}{4} - \frac{1}{40}\right)136.251 = 30.656$$

Remark : It may be observed that

$$\text{var}(\bar{y}_{st}) < \text{var}(\bar{y}_{sys}) < \text{var}(\bar{y}_n)R$$

We should expect this results, since the data exhibits a fairly steady rising trend.

EXERCISE

1. Define systematic sampling. Discuss its advantages and disadvantages.
2. Explain systematic sampling and state the circumstances when it is optimum.
3. What is systematic sampling ? Give illustrations of situations where such sampling is useful. How will you measure sampling error of a systematic sample mean.
4. In systematic sampling, positive connection between units in the sample inflates the variance with in the systematic sampling and is larger than the population variance of mean of a systematic sampling.
5. Prove that if the population consists of a linear trend, then prove that

   $$\text{var}(\bar{y}_{st}) \leq \text{var}(\bar{y}_{sys}) \leq \text{var}(\bar{y}_n)R$$

6. The data below are for a small artificial population which exhibits a fairly steady rising trend. Each column represents a systematic sample and the rows the strata. Compare the rows or the strata. Compare the precision of systematic sampling, random sampling and stratified sampling.

$$n = 4,\ K = 7,\ N = n\ K = 28$$

Strata	1	2	3	4	5	6	7
I	1	0	3	2	8	6	7
II	5	8	4	6	2	5	8
III	18	19	20	20	24	23	28
IV	26	30	31	31	32	37	38

4.7 SUMMARY

Systematic samples are convenient to draw and to execute. In most of the studies reported in this chapter, both on artificial and on natural populations, they compared favourably in precision with stratified random samples. Their disadvantages are that they may give. Precision when unsuspected periodicity is present and that no trustworthy method for estimating $V\left(\bar{y}_{sy}\right)$ from the sample data is known.

In the light of these results systematic sampling can safely be recommended in the following situations.

1. Where the ordering of the population is essentially random or contains at most a mild stratification. Here systematic sampling is used for convenience, with no expectation of any gain in precision. Sample estimates of error that are reasonably unbiased are available.
2. Where a stratification numurous strata is employed and an independent systematic sample is drawn from each stratum, the effects of hidden periodicities tend to cancel out in these situations and an estimate of error that is known to be an over estimate can be obtained. Alternatively, we can use half the number of strata and draw two systematic samples, with independent random starts, from each stratum. This method gives an unbiased estimate of error.
3. For subsampling cluster units. In this case unbiased estimate or almost unbiased estimate of the sampling error can be obtained in most practical situations. This method gives an unbiased estimate error.
4. For sampling populations with variation of a continuous type, provided that an estimate of the sampling error is not regularly required, if a series of surveys of this type are made, an occassional check on the sampling error may be sufficient.

4.8 REFERENCES :

Fundamentals of Applied Statistics	-	S.C. Gupta & V.K. Kapoor
Basic Statistics	-	B.L. Agarwal
Methods of Mathematical Analysis	-	P.S. Grewal
Sampling Techniques	-	Cochran

LESSON - 5

ANALYSIS OF VARIANCE (ANOVA)

LEARNING OBJECTIVES

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

* Have a clear idea of the theory and the practical utility about the concepts of 'Analysis of variance' for one-way and two-way classification.

LESSON OUTLINE

5.1 INTRODUCTION

The analysis of variance frequently referred to by the contraction ANOVA is a statistical technique specially designed to test whether the means of more than two quantitative populations are equal.

The analysis of variance technique developed by R.A. Fisher in 1920's, is capable of fruitful application to a diversity of practical problems. Basically, it consists of classifying and cross classifying statistical results and testing whether the means of a specified classification differ significantly. In this way it is determined whether the given classification is important in affecting the results. For example, the output of a given process might be cross classified by machines and operators (each operator having worked on each machine). From this cross-classification, it could be determined whether the mean qualities of the outputs of the various machines differed significantly. Also it could independently be determined whether the mean qualities of the outputs of the various machines differed significantly. Such a study would determine, for example, whether uniformity in

quality of outputs could be increased by standardizing the procedures for the operators (say, through special training) and similarly whether it could be increased by standardizing the machines (say, through resetting). Analysis of variance thus enables us to analyse the total variation of our data into components which may be attributed to various "sources" or "causes" of variation.

In the chapter on "Sampling and Tests of significance", the t-test of the difference of means was discussed. This test is an adequate procedure for testing the null hypothesis when we have means of only two samples to consider. However, in a situation, where we have three or more samples to consider at a time an alternative procedure is needed for testing the hypothesis that all samples could likely be drawn from the same population. For example, five fertilizers are applied to four plots each of wheat and we are given the yield of wheat on each of these plots. We may be interested in finding out whether the effects of these fertilizers on the yields are significantly different or, in other words, whether the samples have come from the same universe. The answer to this problem is provided by the technique of analysis of variance.

The analysis of variance originated is agrarian research and its language is thus loaded with such agricultural terms as blocks (referring to land) and treatments (referring to populations or samples) which are differentiated in terms of varieties of seed, fertilizers, or cultivation methods.

The word treatment in analysis of variance is used to refer to any factor in the experiment that is controlled at different levels or values. The treatments can be different point of sale displays, assembly line techniques, sales training programmes or, in short, any controlled factor deliberately applied to the elementary units observed in the experiment.

Today, procedure of this analysis finds application in nearly every type of experimental design in natural sciences as well as social sciences. In fact it has come to acquire a place of great prominence in statistical analysis. This is because of the fact that the analysis of variance is amazingly versatile : it can be readily adopted to furnish, with broad limits, a proper evaluation of data obtained from a large body of experiments which involve several continuous random variables. It can give us answers as to whether different sample data classified in terms of a single variable are meaningful. It can also provide us with meaningful comparisions of sample data which are classified according to two or more variables.

The reader should keep in mind that the analysis of variance test discussed here is not intended to serve the ultimate purpose of testing for the significance of the difference between two sample variances : rather its purpose is to test for the significance of the differences among sample means. They do this via the mechanism of the F-Test for testing for the significance of the difference between two variances, but the test is so designed that the variances being compared are different only if the means under consideration are not homogeneous. In this way, significant values of F indicate that the means are significantly different from one another.

5.3 ASSUMPTIONS IN ANALYSIS OF VARIANCE

The asumptions in analysis of variance are the same as discussed earlier while talking of F test i.e.,

1) Normality
2) Homogenity and
3) Independence of error

It may be noted that theoretically speaking, whenever any of these assumptions are not met, the analysis of variance technique cannot be employed to yield valid inferences. It is indeed fortunate that many economic and business experiments do conform, atleast approximately to these premises. In some cases in experimental work, departure from these assumptions also exists. In such situations the analysis of variance can still be applied after a transformation of the data.

In practice it has been observed that one or more of these assumptions can be "bent" without appreciable loss in the adequacy of the F-test. The research strives to meet the assumptions of the F-test, but one usually finds that if the data are reasonably close to meeting the assumptions, conclusions based on the F test are not markedly affected. If the underlying distributions are bimodal or very skewed the F - test results may not be valid.

Conspicuously greater the variance around the sample means, the samples must be relatively speaking, widely dispersed around the grand mean, very likely not representing random samples from the same population. However, if the sample means are very narrowly dispersed around the grand mean, compared with dispersions around their sample means, the samples are likely to be random samples from a common population.

5.4 TECHNIQUE OF ANALYSIS OF VARIANCE

For the sake of clarity the technique of analysis of variance has been discussed separately for (a) one way classification and (b) two way classification.

5.5 ONE WAY CLASSIFICATION

Let us suppose that N observations $x_{ij}\left(i=1,2,\cdots\cdots k;\ j=1,2,\cdots\cdots n\right)$ of a random variable X are grouped on some basis, into k classes of sizes $n_1, n_2, \cdots\cdots n_k$ respectively, $\left(N=\sum_{i=1}^{k} n_i\right)$ as exhibited below

Observations	Means	Total
$x_{11}\ x_{12}\cdots\cdots x_1 n_1$	$\bar{x}_{1.}$	$T_1 .$
$x_{21}\ x_{22}\cdots\cdots x_2 n_2$	$\bar{x}_{2.}$	$T_2 .$
--	----------	-- --
--	----------	-- --
--	----------	-- --
$x_{i1}\ x_{i2}\cdots\cdots\cdots x_i n_i$	$\bar{x}_i$	$T_{i.}$
--	----------	-- --
--	----------	-- --
$x_{k1}\ x_{k2}\cdots\cdots x_k n_k$	$\bar{x}_{k.}$	$T_{k.}$
		G

The total variation in the observation x_{ij} can be split into the following two components :

(i) The variation between the classes or the variation due to different bases of classification, commonly known as treatments.

(ii) The variation within the classes, i.e., the inherent variation of the random variable within the observations of a class.

The first type of variation is due to assignable causes which can be detected and controlled by human endeavour and the second type of variation is due to chance causes which are beyond the control of human hand.

The main object of analysis of variance technique is to examine if there is significant difference between the class means in view of the inherent variability within the separate class.

In particular, let us consider the effect of 'k' different rations on the yield in milk of N cows (of the same breed and stock) divided into k classes of sizes $n_1, n_2, \cdots, n_K$ respectively, $N = \sum_{i=1}^{k} n_i$.

Here the sources of variation are

(i) Effect of the ration (treatment) $t_i ; i = 1, 2, \cdots, k$

(ii) Error ε produced by numerous causes of such magnitude that they are not detected and identified with the knowledge that we have and they together produce a variation of random nature obeying Gaussian (Normal) Law of errors.

Mathematical Model : In this case the linear mathematical model will be

$$x_{ij} = \mu_i + \varepsilon_{ij}$$

$$= \mu + (\mu_i - \mu) + \varepsilon_{ij}$$

$$= \mu + \alpha_i + \varepsilon_{ij}; \text{ where } (i = 1, 2, \cdots k;\ j = 1, 2, \cdots n_i) \text{ ---------- (1)}$$

(i) x_{ij} is the yield from the j^{th} row, $(j = 1, 2, \cdots n_i)$ fed on the i^{th} ration $(i = 1, 2, \cdots, k)$

(ii) μ is the general mean effect given by

$$\mu = \sum_{i=1}^{k} n_i \mu_i / N \text{ ---------- (2)}$$

where μ_i is the fixed effect due to the i^{th} ration, i.e., if there were no treatment differences and no chance causes then the yield of each cow will be μ,

(iii) α_i is the effect of the i^{th} ration given by

$$\alpha_i = \mu_i - \mu \qquad (i = 1, 2, \cdots, k) \text{ ---------- (3)}$$

i.e., the i^{th} ration increases or decreases the yield by an amount α_i . On using (2) we get

$$\sum_{i=1}^{k} n_i x_i = \sum_i ni(\mu_i - \mu) = \sum_i n_i\mu_i - \mu\Sigma n_i$$

$$= N \cdot \mu - \mu \cdot N$$

$$= 0 \text{ ---------- (3a)}$$

(iv) ε_{ij} is the error effect due to chance.

ASSUMPTIONS IN THE MODEL

(i) All the obeservations x_{ij} are independent.

(ii) Different effects are additive in nature.

(iii) ε_{ij} are i.i.d. $N\left(O, \sigma_e^2\right)$

Under the third assumption, the model (1) becomes :

$$E\left(x_{ij}\right) = \mu_1 = \mu + \alpha_i; \begin{pmatrix} i = 1,2,\cdots\cdots,k \\ j = 1,2,\cdots\cdots,n_i \end{pmatrix} \text{ ---------- (3b)}$$

NULL HYPOTHESIS :

We want to test the equality of the population means, i.e., the homogenity of different rations. Hence null hypothesis is given by

$$H_0 : \mu_1 = \mu_2 = \cdots\cdots = \mu_k = \mu \text{ ---------- (4)}$$

which from (3) reduces to

$$H_0 : \alpha_1 = \alpha_2 = \cdots\cdots = \alpha_k = 0 \text{ ---------- (4a)}$$

STATISTICAL ANALYSIS OF THE MODEL

Let us write

$\bar{x}_{i.}$ = mean of the i^{th} class $= \sum_{j=1}^{ni} x_{ij} \Big/ n_i$; $\left(i = 1,2,\cdots\cdots,k\right)$

and $\bar{x}_{..}$ = overall mean = $\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{N}\sum_{i=1}^{k} n_i \bar{x}_i$.

The parameters μ and α_i in (1) are estimated by the principle of least squares on minimising the error (residual) sum of squares given by

$$E = \sum_i \sum_j \varepsilon_{ij}^2 = \sum_i \sum_j \left(x_{ij} - \mu - \alpha_i \right)^2$$

The normal equations for estimating μ and α_i are

$$\frac{\partial E}{\partial \mu} = -2 \sum_i \sum_j \left(x_{ij} - \mu - \alpha_i \right) = 0 \text{ ---------- } (*)$$

and $$\frac{\partial E}{\partial \alpha_i} = -2 \sum_{j=1}^{n_i} \left(x_{ij} - \mu - \alpha_i \right) = 0 \text{ ---------- } (**)$$

From $(*)$, we get

$$\sum_i \sum_j x_{ij} - N\mu - \Sigma n_i \alpha_i = 0$$

$$\Rightarrow \hat{\mu} = \frac{1}{N} \Sigma\Sigma x_{ij} = \bar{x}_{..} \text{ --------- (5)}$$

$$\left[\because \sum_{i=1}^{k} n_i \alpha_i = 0, \text{ using } (3a) \right]$$

From $(**)$, we get

$$\sum_j x_{ij} - n_i \hat{\mu} - n_i \widehat{\alpha_i} = 0$$

$$\Rightarrow \hat{\alpha}_i = \frac{1}{n_i} \sum_j x_{ij} - \hat{\mu} = \bar{x}_{i.} - \hat{\mu}$$

$$\Rightarrow \hat{\alpha}_i = \bar{x}_{i.} - \bar{x}_{..} \text{ ---------- (5a)}$$

Hence substituting in (1) the model becomes

$$x_{ij} = \bar{x}_{..} + \left(\bar{x}_i - \bar{x}_{..} \right) + \left(x_{ij} - \bar{x}_{..} \right)$$

We introduce the error component $\in_{ij}$ so that both the sides are equal. This is the deviation within the class which is due to randomisation. Transposing $\bar{x}_{..}$ to the left, squaring both sides and summing over i and j we get

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_{..} \right)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_{i.} + \bar{x}_{i.} - \bar{x}_{..} \right)^2$$

$$= \sum_i \sum_j \left(x_{ij} - \bar{x}_{i.}\right)^2 + \sum_i n_i \left(\bar{x}_{i.} - \bar{x}_{..}\right)^2 + 2\left[\sum_i \left\{\left(\bar{x}_{i.} - \bar{x}_{..}\right)\sum_j \left(x_{ij} - \bar{x}_{i.}\right)\right\}\right]$$

But $\sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_{i.}\right) = 0,$ since the algebraic sum of the deviations of the rations from their mean is zero.

$$\therefore \sum_i \sum_j \left(x_{ij} - \bar{x}_{i.}\right)^2 = \sum_i \sum_j \left(x_{ij} - \bar{x}_{i.}\right)^2 + \sum_i n_i \left(\bar{x}_{i.} - \bar{x}_{..}\right)^2 \text{ --------- (6)}$$

$S_T^2 = \sum_i \sum_j \left(x_{ij} - \bar{x}_{..}\right)^2$ is known as total sum of squares $\left(T \cdot S \cdot S\right)$

$S_E^2 = \sum_i \sum_j \left(x_{ij} - \bar{x}_{i.}\right)^2$ is called within sum of squares or error sum of squares (S.S.E.), and

$S_t^2 = \sum_i n_i \left(\bar{x}_{i.} - \bar{x}_{..}\right)^2$ is called S.S due to treatments (S.S.T.).

Then

Total S.S. = S.S.E. + S.S.T --------- (6a)

Degrees of freedom for various S.S. S_T^2, the total S.S. which is computed from the N quantities of the form $\left(x_{ij} - \bar{x}_{..}\right)$ will carry $\left(N-1\right)$ degrees of freedom (.d.f.), one d.f. being lost because of the linear constraint.

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_{..}\right) = 0$$

Similarly the treatment sum of squares $\sum_{i=1}^{K} n_i \left(\bar{x}_i - \bar{x}_{..}\right)^2$ will have $\left(k-1\right)$ d.f. since $\sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{x}_{..}\right) = 0$ and the error S.S. $S_E^2 = \sum_i \sum_j \left(x_{ij} - \bar{x}_{i.}\right)^2$ will have $\left(N-k\right)$ d.f. Since it is based on N quantities which are subject to k linear constraints.

$$\sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_{i.}\right) = 0;\ i = 1, 2, \cdots\cdots k.$$

Hence we see that the d.f. for various S.S. are also additive, since

$$N-1=(N-k)+(k-1) \quad \text{---------- (6b)}$$

MEAN SUM OF SQUARES (MSS) DEFINITION : The sum of squares divided by its degrees of freedom gives the corresponding variance or the mean sum of squares (M.S.S.). Thus

$\frac{S_t^2}{(k-1)}=\frac{S\cdot S\cdot T}{(k-1)}=s_t^2$ (say) is the M.S.S. due to treatments; and

$\frac{S_E^2}{(N-k)}=\frac{S\cdot S\cdot E}{(N-k)}=s_E^2$ (say) is the M.S.S. due to error.

Under H_0, otherwise

$$E(s_t^2)>E(s_E^2)$$

Hence the test statistic for H_0 is provided by the variance ratio.

$$F=\frac{s_t^2}{s_E^2} \quad \text{-------- (7)}$$

Thus, if an observed value of F obtained from (7) is greater than the tabulated value of F for $(k-1,\ N-k)$ d.f. at specified level of significance, (usually 5% or 1%), then H_0 is refuted at that level otherwise it may be retained.

The above statistical analysis is very elegantly presented in the following table, known as the analysis of variance (ANOVA) table.

ANOVA TABLE FOR ONE WAY CLASSIFIED DATA

Sources of Variation	Sum of Squares	d.f.	Mean sum of squares	Variance Ratio
Treatment (Ration)	S_t^2	$k-1$	$s_t^2=\frac{S_t^2}{(k-1)}$	$\frac{s_t^2}{s_E^2}=F_{k-1,\,N-k}$
Error	S_E^2	$N-k$	$s_E^2=\frac{S_E^2}{N-k}$	
Total	S_T^2	$N-1$		

5.6 TWO-WAY CLASSIFICATION : (with one observation per cell)

Let us consider the case when there are two factors which may affect the variate values x_{ij} e.g., the yield of milk may be affected by differences in treatments, i.e. rations as well as the differences in variety, i.e, breed and stock of the cows. Let us now suppose that the N cows are divided into 'h' different groups or classes according to their breed and stock, each group containing k cows and then let us consider the effect of k treatments (i.e., rations given at random to cows in each group) on the yield of milk.

Let the suffix i refer to the treatments (rations) and suffix j refer to the varieties (breed of the cow). Then the yields of milk $\left(x_{ij}\right)\left(i=1,2,\cdots,k,\ j=1,2,\cdots\cdots,h\right)$ of $N=hxk$ cows finish the data for the comparison of the treatments. The yields may be expressed as variate values in the following kxh two way table.

			Means	Total
	x_{11}	$x_{12}\cdots\cdots x_{1j}\cdots\cdots x_{1h}$	$\bar{x}_{1.}$	$T_1.$
	x_{21}	$x_{22}\cdots\cdots x_{2j}\cdots\cdots x_{2h}$	$\bar{x}_{2.}$	$T_2.$
	--	----------	--	--
	--	----------	--	--
	x_{i1}	$x_{i2}\cdots\cdots\cdots x_{ij}\cdots\cdots x_{ih}$	$\bar{x}_{i}$	$T_{i.}$
	--	----------	--	--
	--	----------	--	--
	x_{k1}	$x_{k2}\cdots\cdots x_{k}n_{k}$	$\bar{x}_{k.}$	$T_{k.}$
Means	$\bar{x}_{.1.}$	$\bar{x}_{.2}\cdots\cdots\bar{x}_{.j}\cdots\bar{x}_{.h}$	$\bar{x}_{..}$	
Totals	$T_{.1}$	$T_{.2}\cdots\cdots T_{.j}\cdots\cdots T_{.h}$	$\rightarrow$	$\downarrow$ G

Mathematical Model : Let x_{ij} be the yield from the cow of j^{th} variety fed on the i^{th} ration $\left(i=1,2,\cdots,k,\ j=1,2,\cdots,h\right)$

Let us suppose that $x_{ij}\left(i=1,2,\cdots k;\ j=1,2,\cdots h\right)$ are independent, normally distributed as $N\left(\mu_{ij},\sigma_{e^2}\right)$.

The linear mathematical model becomes

$$E\left(x_{ij}\right)=\mu_{ij}$$

$$x_{ij}=\mu_{ij}+\in_{ij} \quad \text{or}$$

where $\in_{ij}$ are i.i.d. $N\left(0,\sigma_{e^2}\right)$

Obviously, $\sum_{i=1}^{K}\alpha_i=0$

(iii) The effect $\beta_j, \left(j=1,2,\cdots,h\right)$ due to the j^{th} variety (breed of cow) given by

$$\beta_j=\mu_{.j}-\mu$$

where $\mu_{.j}=\frac{1}{k}\sum_{i=1}^{k}\mu_{ij}, \ \left(j=1,2,\cdots\cdots,h\right)$

Obviously, $\sum_{j=1}^{h}\beta_j=0$

(iv) The interaction effect γ_{ij} when the i^{th} level of first factor (rations) and j^{th} level of second factor (breed of cow) occur simultaneously and given by

$$\gamma_{ij}=\mu_{ij}-\mu_{i.}-\mu_{.j}+\mu$$

where $\sum_{j}\gamma_{ij}=0; \quad \forall i=1,2,\cdots\cdots k$

$$\sum_{i}\gamma_{ij}=0; \ \forall \ \ j=1,2,\cdots\cdots h$$

Thus we have

$$\mu_{ij}=\mu+\left(\mu_{i.}-\mu\right)+\left(\mu_{.j}-\mu\right)+\left(\mu_{ij}-\mu_{i.}-\mu_{.j}+\mu\right)$$

and consequently the model $x_{ij}=\mu_{ij}+\in_{ij}$ becomes

$$x_{ij}=\mu+\alpha_i+\beta_j+\gamma_{ij}+\in_{ij}$$

where $\in_{ij}$ is the error effect due to chance and

$$\sum_{i=1}^{k} \alpha_i = 0 = \sum_{j=1}^{h} \beta_j$$

and

$$\sum_{i} \gamma_{ij} = \sum_{j} \gamma_{ij} = 0$$

for all i for all j

As there is only one observation in each cell, the observation corresponding to the i^{th} level of ration and j^{th} level of breed of cow is only one i.e., x_{ij}. But we cannot estimate by one value alone. Hence in this case (one observation per cell), the interaction effect $\gamma_{ij} = 0$ and the model $x_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \in_{ij}$ reduces to

$$x_{ij} = \mu + \alpha_i + \beta_j + \in_{ij}$$

STATISTICAL ANALYSIS OF THE MODEL

Let us write $\bar{x}_{i.}$ = mean yield of i^{th} ration.

$$= \frac{1}{n} \sum_{j=1}^{h} x_{ij}, \ (i = 1, 2, \cdots\cdots, k)$$

$\bar{x}_{.j}$ = Mean yield of the j^{th} variety

$$= \frac{1}{k} \sum_{i=1}^{k} x_{ij}, \ (j = 1, 2, \cdots\cdots, h)$$

$\bar{x}_{..}$ = overall mean = $\frac{1}{hk} \sum_{i} \sum_{j} x_{ij}$

$$= \frac{1}{h} \sum_{j} \left(\frac{1}{k} \sum_{i} x_{ij} \right) = \frac{1}{h} \sum_{j} \bar{x}_{.j}$$

$$= \frac{1}{k} \sum_{i} \left(\frac{1}{h} \sum_{j} x_{ij} \right) = \frac{1}{k} \sum_{i} \bar{x}_{i.}$$

The least square estimates of the parameters μ, α_i, and β_i are obtained on minimising the error sum of squares.

$$E = \sum_{i=1}^{k}\sum_{j=1}^{h} \in_{ij^2} = \sum_i \sum_j \left(x_{ij} - \mu - \alpha_i - \beta_j\right)^2$$

The normal equations for estimating μ, α_i and β_j are respectively.

$$\frac{dE}{d\mu} = 0 = -2\sum_i \sum_j \left(x_{ij} - \mu - \alpha_i - \beta_j\right)$$

$$\frac{dE}{d\alpha_i} = 0 = -2\sum_j \left(x_{ij} - \mu - \alpha_i - \beta_j\right)$$

$$\frac{dE}{d\beta_j} = 0 = -2\sum_i \left(x_{ij} - \mu - \alpha_i - \beta_j\right)$$

Since $\sum_i \alpha_i = 0 = \sum_j \beta_j$, we get from the above equations

$$\hat{\mu} = \frac{1}{hk}\sum_i \sum_j x_{ij} = \bar{x}_{..}$$

$$\widehat{\alpha_i} = \frac{1}{h}\sum_j x_{ij} - \hat{\mu} = \bar{x}_{i.} - \bar{x}_{..}$$

$$\widehat{\beta_j} = \frac{1}{k}\sum_i x_{ij} - \hat{\mu} = \bar{x}_{.j} - \bar{x}_{..}$$

Thus the linear model x_{ij} becomes

$x_{ij} = \bar{x}_{..} + \left(\bar{x}_{i.} - \bar{x}_{..}\right) + \left(\bar{x}_{.j} - \bar{x}_{..}\right) + \left(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..}\right)$ the error term $\in_{ij}$ being so chosen that both sides are equal.

Transposing $\bar{x}_{..}$ to the left side, squaring and summing both sides over i from 1 to k and j from 1 to h we get

$$\sum_i \sum_j \left(x_{ij} - \bar{x}_{..}\right)^2 = \sum_i \sum_j \left[\left(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..}\right) + \left(\bar{x}_{i.} - \bar{x}_{..}\right) + \left(x_{.j} - \bar{x}_{..}\right)\right]^2$$

$$= \sum_i \sum_j \left(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..}\right)^2 + \sum_i \sum_j \left(\bar{x}_{i} - \bar{x}_{..}\right)^2 + \sum_i \sum_j \left(\bar{x}_{.j} - \bar{x}_{..}\right)^2$$

$$+2\sum_i\sum_j\left(\bar{x}_{i.}-\bar{x}_{..}\right)\left(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..}\right)$$

$$+2\sum_i\sum_j\left(\bar{x}_{.j}-\bar{x}_{..}\right)\left(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..}\right)$$

$$+2\sum_i\sum_j\left(\bar{x}_{i.}-\bar{x}_{..}\right)\left(\bar{x}_{.j}-\bar{x}_{..}\right)$$

Now $\sum_i\sum_j\left(\bar{x}_{i.}-\bar{x}_{..}\right)\left(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..}\right)$

$$=\sum_i\left[\left(\bar{x}_{i.}-\bar{x}_{..}\right)\sum_j\left(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..}\right)\right]$$

$$=\sum_j\left[\left(\bar{x}_{i.}-\bar{x}_{..}\right)\left\{\sum_j\left(x_{ij}-\bar{x}_{i.}\right)-\sum_j\left(\bar{x}_{.j}-\bar{x}_{..}\right)\right\}\right]=0,$$

since algebric sum of deviations of a set of observations about their mean is zero.

Similarly it can be easily seen that other product items also vanish.

Hence,

$$\sum_i\sum_j\left(x_{ij}-\bar{x}_{..}\right)^2=h\sum_i\left[\bar{x}_{i.}-\bar{x}_{..}\right]^2+k\sum_j\left(\bar{x}_{.j}-\bar{x}_{..}\right)^2+\sum_i\sum_j\left(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..}\right)^2$$

or $S_{T^2}=S_t^2+S_V^2+S_E^2$

where $S_T^2=\sum_i\sum_j\left(x_{ij}-\bar{x}\right)^2$ is the total S.S.

$S_t^2=h\sum_i\left(\bar{x}_{i.}-\bar{x}_{..}\right)^2$ is the S.S. due to treatments

$S_v^2=k\sum_j\left(\bar{x}_{.j}-\bar{x}_{..}\right)^2$ is the S.S. due to varieties

and $S_E^2=\sum_i\sum_j\left(\bar{x}_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x}_{..}\right)^2$ is the error or residual S.S.

NULL HYPOTHESIS :

We set up all the null hypothesis that the treatments as well as varieties are homogeneous. In other words, the null hypothesis for treatments and varieties are respectively.

$$H_t : \mu_{1.} = \mu_{2.} = \cdots\cdots = \mu_{k.} = \mu$$

$$H_r : \mu_{.1} = \mu_{.2} = \cdots\cdots = \mu_{h} = \mu$$

or from the equivalents

$$H_t : \alpha_1 = \alpha_2 = \cdots\cdots = \alpha_k = 0$$

$$H_v : \beta_1 = \beta_2 = \cdots\cdots = \beta_h = 0$$

Degrees of freedom for various S.S. : The total S.S., $S_{\in^2}$ being computed from $N = hk$ quantities $\left(x_{ij} - \bar{x}_{..}\right)$, which are subject to one linear constraint.

$$\sum_i \sum_j \left(x_{ij} - \bar{x}_{..}\right) = 0, \text{ will carry } (N-1)\text{d.f.?}$$

Similarly S_{t^2} will be based on $(k-1)\text{d.f.}$, since

$$\sum_i \left(\bar{x}_{1.} - \bar{x}_{..}\right) = 0 \text{ and}$$

S_{v^2} will have $(h-1)\text{d.f.}$, since $\sum_j \left(\bar{x}_{.j} - \bar{x}_{..}\right) = 0$ and S_{e^2} will carry

$$(N-1) - (k-1) - (h-1) = (h-1)(k-1)\text{d.f.} \qquad (\because N = hk)$$

Thus the partitioning of d.f. is as follows :

$$(hk-1) = (k-1) + (h-1) + (h-1)(k-1)$$

which implies that the d.f. are additive.

Test Statistic : In order to obtain appropriate test statistic to test the hypothesis H_t and H_v we need the expectations of the various mean S.S. due to each of independent factors.

Using the same notations for the mean S.S. as in the case of one way classified data, we get

Mean S.S. due to treatments = $\dfrac{S_t^2}{k-1} = s_t^2$ (say)

Mean S.S. due to varieties = $\frac{S_{v^2}}{h-1} = s_v^2$, (say)

Error mean S.S. $= \frac{S_E^2}{(h-1)(k-1)} = s_E^2$, (say)

Since $\sum_i \alpha_i{}^2 \geq 0$ and $\sum_j \beta_j{}^2 \geq 0$ we get

$$E(S_{t^2}) = E(S_{E^2}), \text{ under } H_t$$

otherwise $E(S_{t^2}) > E(S_{E^2})$

and $E(S_{V^2}) = E(S_{E^2})$, under H_V

otherwise $E(S_{V^2}) > E(S_{E^2})$

Since various S.S as well as their respective d.f. are additive and since under the null hypothesis H_t and H_v, each of $S_{r^2}, S_{t^2}, S_{v^2}$ and S_{t^2} provides an unbiased estimate of $\sigma_{\in^2}$, under the assumption of normality of parent population, we get by Cochran's theorem

$$\frac{S_t^2}{\sigma_{\in^2}}, \frac{S_v^2}{\sigma_{\in^2}} \text{ and } \frac{S_E^2}{\sigma_{\in^2}}$$

are mutually independent x^2 varieties with $(k-1), (h-1)$ and $(h-1)(k-1)$ d.f. respectively. Hence under H_t and H_v respectively, we get.

$$F_t = \frac{S_{t^2}}{\sigma_{e^2}(k-1)} \div \frac{S_{E^2}}{\sigma_{e^2}(h-1)(k-1)}$$

$$= \frac{S_t^2}{S_E^2} \text{ conforms to } F_{(k-1),(h-1)(k-1)}$$

and

$$F_v = \frac{S_v^2}{\sigma_{e^2}(h-1)} \div \frac{S_E^2}{\sigma_{e^2}(h-1)(k-1)}$$

$F = \frac{S_v^2}{S_E^2}$, conforms to $F_{(h-1),(h-1)(k-1)}$

By comparing these values with the tabulated value of F for respective d.f. and at certain level of significance, the null hypothesis of the homogeneity of various treatments and various varieties may be rejected or accepted.

5.7 ONE WAY CLASSIFICATION

Problem 1 : The following data gives the figures of production of rice of three varities A, B, C of rice shown in 12 plots.

A	14	13	16	17
B	15	18	12	19
C	20	17	11	8

Is there a significant difference between varieties ?

Solution : Let us take the null hypothesis that the varieties don't differ significantly. Carrying out analysis of variance.

Variety A	Variety B	Variety C
X_1	X_2	X_3
14	15	20
13	18	17
16	12	11
17	19	8
$\Sigma X_1 = 60$	$\Sigma X_2 = 64$	$\Sigma X_3 = 56$
$\bar{X}_1 = 15$	$\bar{X}_2 = 16$	$\bar{X}_3 = 14$

$$\bar{X} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{3} = \frac{15+16+14}{3} = 15$$

Variance between samples

X_1	X_2	X_3
$(\bar{X}_1 - \bar{\bar{X}})^2$	$(\bar{X}_2 - \bar{\bar{X}})^2$	$(\bar{X}_3 - \bar{\bar{X}})^2$
1	1	36
4	4	9
1	16	9
4	9	36

$\Sigma(X_1 - \bar{X}_1)^2 = 10$ $\Sigma(X_2 - \bar{X}_2)^2 = 30$ $\Sigma(X_3 - \bar{X}_3)^2 = 90$

Total sum of squares with in samples = 10 + 30 + 90 = 130

Mean square within samples = $\frac{130}{9} = 14.44$

Source of variation	Sum of squares	Degrees of Freedom	Mean square
Between samples	8	2	4
within sample	130	9	14.44
Total	138	11	

$$F = \frac{4}{14.44} = 0.277$$

For $V_1 = 2$ and $V_2 = 9$, $F_{0.05} = 4.26$. The calculated value of F is less than the table value. Our hypothesis holds true. Hence there is no significant difference between varieties.

Problem - 2 : Four salesman were posted in different areas by a company. The number of units of commodity 'X' sold by them are as follows :

A	20	23	28	29
B	25	32	30	21
C	23	28	35	18
D	15	21	19	25

Is there a significant difference in the performance of these salesmen ?

Solution : Let us take the hypothesis that there is no significant difference in the performance of the four salesmen.

Salesmen A	B	C	D
X_1	X_2	X_3	X_4
20	25	23	15
23	32	28	21
28	30	35	19
29	21	18	25
$\Sigma X_1 = 100$	$\Sigma X_2 = 108$	$\Sigma X_3 = 104$	$\Sigma X_4 = 80$
$\bar{X}_1 = 25$	$\bar{X}_2 = 27$	$\bar{X}_3 = 26$	$\bar{X}_4 = 20$

$$\bar{\bar{X}} = \frac{25 + 27 + 26 + 20}{4} = \frac{98}{4} = 24.5$$

Variance between samples.

X_1	X_2	X_3	X_4
$(\bar{X}_1 - \bar{\bar{X}})^2$	$(\bar{X}_2 - \bar{\bar{X}})^2$	$(\bar{X}_3 - \bar{\bar{X}})^2$	$(\bar{X}_4 - \bar{\bar{X}})^2$
0.25	6.25	2.25	20.25
0.25	6.25	2.25	20.25
0.25	6.25	2.25	20.25
0.25	6.25	2.25	20.25
1.00	25.00	9.00	81.00

Sum of squares between samples = 1 + 25 + 9 + 81 = 116

v = 4 - 1 = 3

Mean square between samples $= \frac{116}{3} = 38.67$

Variance within samples

X_1	X_2	X_3	X_4
$(\bar{X}_1 - \bar{X})^2$	$(\bar{X}_2 - \bar{X})^2$	$(\bar{X}_3 - \bar{X})^2$	$(\bar{X}_4 - \bar{X})^2$
25	4	2	25
4	25	4	1
9	9	81	8
16	36	64	25
54	74	158	52

Total sum of squares within samples

= 54 + 74 + 158 + 52 = 338

Mean square within samples = $\frac{338}{12} = 28.17$

Source of variation	Sum of squares	Degrees of freedom	Mean Square
Between samples	116	3	38.67
Within samples	338	12	28.17
Total	454	15	

$$F = \frac{38.67}{28.17} = 1.37$$

For $V_1 = 3$ and $V_2 = 12$, $F_{0.05} = 3.24$

The calculated values of F is less than the table value. Hence there is no significant difference in the performance of salesmen.

Problem - 3 : The following table gives the wheat yield in bushels per acre of 4 varieties grown in 5 blocks. Test whether the mean yields of these varieties differ significantly.

	Varieties			
Block	A	B	C	D
1	16	14	19	20
2	20	23	11	27
3	26	24	23	16
4	12	29	17	21
5	21	25	20	13

Given for $V_1 = 3$, $V_2 = 16$, $F_{0.05} = 3.24$

Solution : Let us take the hypothesis that there is no significant difference in the mean yield as obtained by these four varieties.

X_1	X_2	X_3	X_4
16	14	19	20
20	23	11	27
26	24	23	19
12	29	17	21
21	25	20	13
$\Sigma X_1 = 95$	115	90	100
$\bar{X}_1 = 19$	$\bar{X}_2 = 23$	$\bar{X}_3 = 18$	$\bar{X}_4 = 20$

$$\bar{X} = \frac{19 + 23 + 18 + 20}{4} = 20$$

Sum of squares between samples

X_1	X_2	X_3	X_4
$(\bar{X}_1 - \bar{X})^2$	$(\bar{X}_2 - \bar{X})^2$	$(\bar{X}_3 - \bar{X})^2$	$(\bar{X}_4 - \bar{X})^2$
1	9	4	0
1	9	4	0
1	9	4	0
1	9	4	0
1	9	4	0
5	45	20	0

Sum of squares between samples = 5 + 45 + 20 + 0 = 70

$$v = 4 - 1 = 3$$

Mean sum of squares $= \frac{70}{3} = 23.33$

Sum of squares within samples

X_1	X_2	X_3	X_4
$(\bar{X}_1 - \bar{\bar{X}})^2$	$(\bar{X}_2 - \bar{\bar{X}})^2$	$(\bar{X}_3 - \bar{\bar{X}})^2$	$(\bar{X}_4 - \bar{\bar{X}})^2$
9	81	1	0
1	0	49	49
49	1	25	1
49	36	1	1
4	4	4	49
112	122	80	100

Sum of squares within samples = 112 + 122 + 80 + 100 = 414

Mean sum of squares $= \frac{414}{16} = 25.875$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Between samples	70	3	23.33
Within samples	414	16	25.875
Total	484	19	

$$F = \frac{23.33}{25.875} = 0.902$$

For $v_1 = 3$ and $v_2 = 16$, $F_{0.05} = 3.24$

The calculated value of F is less than the table value. Hence the mean yields of varieties don't differ significantly.

5.8 TWO WAY CLASSIFICATION

Problem 1 : A company manufacturing lipsticks appointed four salesmen and observed their sales in 3 different areas 'X', 'Y' and 'Z'. The number of lipstiks sold is given below.

	Salesmen			
Area	A	B	C	D
X	22	27	38	45
Y	28	32	40	38
Z	25	40	36	22

Carry out the analysis of variance to test whether there is a significant difference in (i) the slaes of four salesmen (ii), in the sales carried out in different areas.

Solution : Let us take the hypothesis that (i) there is no significant difference in sales carried out by the four salesmen and (ii) there is no significant difference in the sales of different areas.

	Salesmen				
Area	A	B	C	D	Total
X	22	23	38	45	132
Y	28	32	40	38	138
Z	25	40	36	22	123
	75	99	114	105	393

Correction factor $= \frac{T^2}{N} = \frac{(393)^2}{12} = \frac{154449}{12} = 12870.75$

Sum of squares between salesmen

$$= \frac{(75)^2 + (99)^2 + (114)^2 + (105)^2}{3} - 12870.75$$

$$= \frac{5625 + 9802 + 12996 + 11025}{3} - 12870.76$$

$$= \frac{39447}{3} - 12870.75$$

$$= 278.25$$

Sum of squares between areas

$$= \frac{(132)^2 + (138)^2 + (123)^2}{4} - 12870.75$$

$$= 12899.25 - 12870.75$$

$$= 28.5$$

Total sum of squares

$$= \Big[(22)^2 + (28)^2 + (25)^2 + (27)^2 + (32)^2 + (40)^2 + (38)^2 + (40)^2$$

$$+ (36)^2 + (45)^2 + (38)^2 + (22)^2 \Big] - 12870.75$$

= [484 + 784 + 625 + 729 + 1024 + 1600 + 1444 + 1600 + 1296 + 2025 + 1444 + 484] - 12870.75

$$= 13539 - 12870.75 = 668.25$$

Source of variation	Sum of Squares	Degrees of Freedom	Mean Square
Between columns (salesmen)	278.25	3	92.75
Between Row (area)	28.50	2	14.25
Residual (or) error	361.50	6	60.25
Total	668.25	11	

$$F_1 = \frac{92.75}{60.25} = 1.54$$

For $v_1 = 3$ and $v_2 = 6$, $F_{0.05} = 4.76$

The calculated value of F is less than the table value. Hence the sales of four different salesmen do not differ significantly.

$$F_2 = \frac{14.25}{60.25} = 0.236$$

For $v_1 = 2$ and $v_2 = 6, F = 5.14$

The calculated value of F is less than the table value. Hence the sales in different areas don't differ significantly.

Problem 2 : The following data gives the number of units produced by 4 different workers using 3 different machines.

	Machine Type		
Workers	A	B	C
1	20	28	26
2	22	30	32
3	26	32	10
4	32	24	24

Test whether (a) the four workers differ with respect to mean productivity (b) whether the mean productivity is the same for different machine types.

Solution : Let us take the hypothesis.

a) The four workers, donot differ with respect to mean productivity (b) the mean productivity is the same for different machines.

	Machine Type			
Workers	A	B	C	Total
1	20	28	26	74
2	22	30	32	84
3	26	32	18	76
4	32	20	24	76
	100	110	100	310

Correction factor = $\frac{T^2}{N} = \frac{(310)^2}{12} = \frac{96100}{12} = 8008.33$

Sum of squares between machine type

$$= \frac{(100)^2 + (110)^2 + (100)^2}{4} - \frac{T^2}{N}$$

$$= \frac{10000 + 12100 + 10000}{4} - 8008.33$$

$$= 8028 - 8008.33 = 19.67$$

Total sum of squares

$$= \Big[(20)^2 + (22)^2 + (26)^2 + (32)^2 + (21)^2 + (30)^2 + (32)^2$$

$$+ (20)^2 + (26)^2 + (32)^2 + (18)^2 + (24)^2 \Big] - \frac{T^2}{N}$$

$$= [400 + 484 + 676 + 1024 + 784 + 900 + 1024 + 400 + 676 + 1024 + 344 + 567] - 8008.33$$

$= 8292 - 8008.33 = 283.67$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares
Between columns (machines)	16.67	2	8.34
Between rows (workers)	19.67	3	6.56
Residual or error	247.33	6	41.22
Total	283.67	11	

$$F_1 = \frac{8.34}{41.22} = 0.202$$

For $v_1 = 2$ and $v_2 = 6$, $F_{0.05} = 5.14$

Since the calculated value of F is less than the table value, our hypothesis is true. Hence there is no significant difference between the machine type.

$$F_2 = \frac{6.56}{41.22} = 0.159$$

For $v_1 = 3$ and $v_2 = 6$, $F_{0.05} = 4.76$

Since the calculated value of F is less than the table value, our hypothesis is true. Hence there is no significant difference in the mean productivity of workers.

5.9 EXERCISE

Example 1 : The following table gives the retail prices of a commodity in some ships selected in three big cities.

City	Price per unit (Rs.)			
A	15	18	23	24
B	23	19	18	24
C	19	16	22	31

carryout the analysis of variance to find out whether there is a significant difference in the prices charged in these cities.

Example 2 : Four different manufacturing processes were tried at three different stations and average measurement of a quality characteristic of the product by these processes obtained, as in the following table. Perform the analysis of variance of the data and list for the difference between the processes.

	Process			
Station	A	B	C	D
I	7	14	11	11
II	15	16	13	10
III	8	15	10	12

Example 3 : Following tables gives the results of an experiment conducted to study the effects of four fertilisers on the yield. Analyse the data and state your conclusions.

	Yield in Kgs.		
Fertiliser	1	2	3
A	3	2	3
B	2	5	2
C	2	3	3
D	4	2	4

Example 4 : Four experiments were conducted to detrmine the moisture content of samples of a powder, each mean taking a sample from each of six consignments. Their assesments are given below :

Observers	1	2	3	4	5	6
A	9	10	9	10	11	11
B	12	11	9	11	11	10
C	11	10	10	12	11	10
D	12	13	11	14	12	10

Analyse the data and discuss the significance of difference between consignments and observers.

Example 5 : 9 pigeons of 3 kinds were trained for postal services. The figures below give the number of correct landings out of 20 flights. Do you suggest that type I is superior ?

Type of piegeons			
I	11	10	9
II	10	8	6
III	12	8	7

The table value of F at 5% level of significance for $v_1 = 2$ and $v_2 = 6$ is 5.14.

5.10 QUESTIONS TO STUDY

1. Define ANOVA and explain the assumptions of ANOVA.
2. Explain ANOVA for one-way calssification.
3. Explain ANOVA for two-ways calssification.

5.11 REFERENCES

1. Fundamentals of Applied Statistics - S.C. Gupta & V.K. Kapoor
2. Methods of Statistical Analysis - P.S. Grewal

LESSON - 6

PRINCIPLES OF EXPERIMENTAL DESIGN

LEARNING OBJECTIVES

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

* Understand the application of completely Randomised Design

LESSON OUT LINE

6.1 INTRODUCTION

The first great stimulus to the development of the theory and practice of experimental design came from agricultural research R.A.. Fisher realised that current practices in field plot trials failed to produce unambiguous conclusions. This led him from about 1923 onward to examine the principles underlying scientific experimentation and to evolve new techniques of design. Not only was it necessary to devise procedures that would permit the drawing of valid inferences from experimental results, but these inferences had to be freed as far possible from the obscuring effect of the variability inherent in the material and the nature of the observations. Not only was randomization needed in order to remove bias, but also for making valid estimates of standard errors. The labour of performing experiments and the number of questions requiring investigation were so great as to make it imperative for techniques that should use most effectively the materials and effort employed and should give results of high precision. To Fisher goes most of the credit for stating and solving these problems and creating a new branch of science from which experimentation in many fields of research has since benefited. Although this science of experimental design is today used widely in biology and else where, the standard nomenclature retains evidence of its agricultural origin.

Definition : Accordingly by the design of experiment may be defined as the logical construction of the experiment in which the degree of uncertainty with which the inference is drawn may be well defined.

Terminology in experimental designs (important terms and definitions)

Experiment : An experiment is a device or a means of getting an answer to the problem under consideration and can be classified into two categories as follows :

(a) Absolute, and (b) Comparative

Absolute experiments consist in determining the absolute value of some characteristics like (i) obtaining the average intelligence quotient (I.Q.) of a group of people, (ii) finding the correlation coefficient between two variables in a bivariate distribution, etc., On the other hand, comparative experiments are designed to compare the effect of two or more objects on some popular characteristics., e.g., comparision of different manures of fertilizers, different kinds of varieties of a crop, different cultivation processes, different pieces of land in a field experiment, or diets or medicines in a dietary or medical experiment respectively etc.

Treatments : Various objects of comparision in a comparative experiment, are termed as treatments, e.g. in field experimentation different fertilizers or different varieties of crop or different methods of cultivation are the treatments.

Experimental unit : The smallest division of the experimental material to which we apply the treatments and on which we make observations on the variable under study, is termed as experimental unit, e.g., in field experiments the plot of 'land' is the experimental unit. In other experiments unit may be a patient in a hospital, a lump of dough, a group of pigs in a pen or a batch of seeds.

Blocks : In agricultural experiments, most of the times we divide the whole experimental unit(field) into relatively homogeneous subgroups or strata. These strata, which are more uniform amongst themselves than the field as a whole, are known as blocks. yield. This measurement of the variable under study on different experimental units (e.g., plots, in field experiments) are termed as yields.

Experimental Error : Let us suppose that a large homogeneous field is divided into different plots. If the yields from some of the treatments are more than those of the others, the experimenter is faced with the problem of deciding if the observed differences are really due to treatment effects or are they due to chance (uncontrolled) factors. In field experimentation, it is a common experience that the fertility gradient of the soil does not follow any systematic pattern but behaves in an erratic fashion. Experience tells us that even if the same treatment is used on all plots the yield would still vary due to the differences in soil fertility. Such variation from plot to plot, which is due to random (or chance or non assignable) factors beyond human control, is spoken of as experimental error. It may be pointed out that the term 'error' used here is not synonymous with 'mistake' but is a technical term which includes all types of extraneous variations due to.

(i) the inherent variability in the experimental material to which treatments are applied,

(ii) the lack of uniformity in the methodology of conducting the experimental or in other words failure to standardise the experimental technique and,

(iii) lack of representativeness of the sample to the population under study.

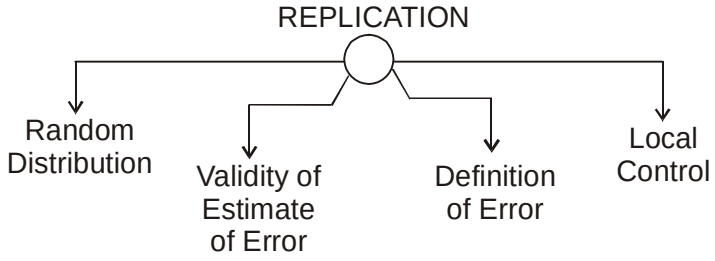
6.3 PRINCIPLES OF AN EXPERIMENTAL DESIGN

Prof. Ronald A. Fisher pioneered the study of experimental designs with his classical book, the design of experiments. According to him, the basic principles of the design of experiments are:

(i) Replication (ii) Randomisation, and (iii) Local control

The figure below due to Fisher, illustrates the function of the various principles.

Replication : As pointed out earlier, replication mean 'the repetition' of the treatment under investigation'. An experimenter resorts to replication in order to average out the influence of the chance factors on different experimental units. Thus, the repetition of treatments results in a more reliable estimate than is possible with a single observation. The figuring are the chief advantages of replication :



The most important purpose of replication is to provide an estimate of the experimental error without which we cannot (i) test the significance of the difference between any two treatments, or (ii) determine the length of the confidence interval. The estimate of the experimental error is obtained by considering the differences in the plots receiving the same treatment in different replications and these is no other alternative of obtaining this estimate.

It is desirable to have as much uniformity or homogeneity as possible within each replication but it is not important to have a great deal of uniformity between replications.

The adequate number of replications for various treatments in an experiment depends upon the knowledge of the variability of the experimental material, e.g. fertility of soil in field experimentation, which is rarely known and as such cannot be suggested in advance. A general rule is to get as many replications which will provide at least 12 degrees of freedom for the error. This follows from the fact that the values of F- statistic do not decrease rapidly beyond $V_2 = 12$. Usually one should not use less than 4 replications.

Randomisation : As discussed earlier, by replication the experimenter tries to average out as far as possible the effects due to uncontrolled factors. This brings him to the question of allocation of treatments to experimental units so that each treatment gets an equal chance of showing its worth. In the absense of the prior knowledge of the variability of the experimental material, this objective is achieved through "randomisation" a process of assigning the treatments to various experimental units in a purely chance manner. The following are the main objectives of randomisation.

(i) The validity of the statistical tests of significances, e.g. the t-test for testing the significance of the difference of two means or the 'Analysis of variance' F-test for testing the homogeneity of several means, depends on the fact that the statistical data under consideration obeys some statistical distribution. Randomisation provides a logical basis for that and makes it possible to draw rigorous inductive inference by the use of statistical theories based on probability theory. This assumption of randomness is necessary since S.E. $(\bar{x}) = \frac{\sigma}{\sqrt{n}}$, for random sampling only. Randomising the treatments over the experimental results by unanticipated influences such as rise in ambient temperature, drift in calibration of instruments and equipment, fertility of the soil or other systematic changes.

The purpose of randomness is to assure that the sources of variation, not controlled in the experiment, operate randomly so that the average effect on any group of units is zero. In other words, randomisation eliminates bias in any form. It equalises even factors of variation over which we have no control.

LOCAL CONTROL : If the experimental material, say field for agricultural experimentation, is heterogeneous and different treatments are allocated to various units (plots) at random over the entire field, the soil heterogenity will also enter the uncontrolled factors and thus increase the experimental error. It is desirable to reduce the experimental error as far as practicable without unduly increasing the number of replications or without interfering with the statistical requirement of randomness, so that even smaller differences between treatments can be detected as significant. In addition to the principles of replication and randomisation discussed earlier, the experimental error can further be reduced by making use of the fact that neighbouring areas in a field are relatively more homogeneous than those that are widely spread out. In order to separate the soil fertility effects from the experimental error, the whole experimental area (field) is divided into homogeneous groups (blocks) row-wise or column-wise [one-way] elimination of fertility gradient, i.e., Randomised block design) or both (elimination of fertility gradient in two perpendicular directions i.e., latin square design). According to the fertility gradient of the soil such that the variation within each block is minimum and between the blocks is maximum. The treatments are then allocated at random within each block. The process of reducing the experimental error by dividing the relatively heterogeneous experimental area (field) into homogeneous blocks (due to physical contiguity as far as field experiments are concerned) is known as local control.

In the following sequences we shall discuss the designing and the analysis of some of the important designs of experiments.

6.4 COMPLETELY RANDOMISED DESIGN (C.R.D.)

The completely randomised design is the simplest of all the designs, based on principles of randomisation and replication. In this design treatments are allocated at random to the experimental units over the entire experimental material. Let us suppose that we have v treatments, the i^{th} treatment being replicated r_i times, $i = 1, 2, \cdots, v$. Then the whole experimental material is divided into $n = \Sigma r_i$ experimental units and the treatments are distributed completely at random over the units subject to the condition that the i^{th} treatment occurs r_i, times, Randomisation assures that extraneous factors do not continually influence one treatment. In particular case if

$$r_i = r \ \forall \ i = 1, 2, \cdots, v$$

i.e. if each treatment is repeated on equal no. of times, r, then $h = rv$ and randomisation gives every group of r units an equal chance of receiving the treatments. In general, equal no of replications for each treatment should be made except in particular cases when some treatments are of greater interest than others or when practical limitations dictate otherwise,

Advantages :

(i) C.R.D. results in the maximum use of the experimental units since all the experimental material can be used.

(ii) The design is very flexible. Any number of treatments can be used and different treatments can be used unequal number of times without unduly complicating the statistical analysis in most of the cases.

(iii) The statistical analysis remains simple if some or all the observations for any treatment are rejected or lost or missing for some purely random accidental reasons. We merely carry out the standard analysis on the avilable data. Moreover the loss of information due to missing data is smaller in comparision with any other design.

(iv) It provides the maximum number of degree of freedom for the estimation of the error variance, which increases the sensitivity or the precision of the experiment for small experiments, i.e. for experiments with small number of treatments.

Disadvantages :

(i) In certain circumstances, the design suffers from the disadvantage of being inherently less informative than other more sophisticated layouts. This usually happens if the experimental material is not homogeneous. Since randomisation is not restricted in any direction to ensure that the units receiving one treatment are similar to those receiving the other treatments, the whole variations among the experimental units are included in the residual variance. This makes the design less efficient and results in less sensitivity in detecting significant effects. As such C.R.D. is seldom used in field experimentation, where due to the fertility gradient of the soil the whole experimental material, viz field is not homogeneous and it is better to use more efficient designs like randomised block design (R.B.D) or Latin Square Design (L.S.D.), etc. discussed later.

6.5 APPLICATIONS

(i) Completely Randomised Design : Is most useful in laboratory technique and methodological studies, eg, in physics, chemistry or cookery, in chemical and biological experiments, in some green house studies, etc. where either the experimental material is homogeneous or the instrinsic variability between units can be reduced.

(ii) C.R.D. is recommended in situations where an appreciable fraction of units is likely to be destroyed or fail to respond to statistical analysis of C.R.D.

Statistical analysis of a C.R.D. is analogous to the ANOVA for a one-way classified data, the linear model [assuming various effects to be additive] becomes

$$Y_{ij} = \mu + \tau_i + \in_{ij} \begin{pmatrix} i = 1, 2, \cdots v \\ j = 1, 2 \cdots r_i \end{pmatrix}$$

where Y_{ij} is the yield or response from the j^{th} unit receiving the i^{th} treatment, μ is the general mean effect due to chance such that $\in_{ij}$ an identically and independently distributed $(i.i.d) \cdot N\left(0, \sigma_e^{\,2}\right)$. Then $n = \sum_{i=1}^{r} r_i$ is the total number of experimental units. If we write

$\sum_i \sum_j y_{ij} = y_{..} = G =$ Grand total of all the n observations

$\sum_{j=1}^{r_i} y_{ij} = y_i = T_i =$ Total response of the units receiving i^{th} treatment.

Then, as in ANOVA (one-way classified data)

$$\sum_{i=1}^{v} \sum_{j=1}^{r_i} \left(Y_{ij} - Y_{..}\right)^2 = \sum_i \sum_j \left(Y_{ij} - Y_{i.}\right)^2 + \sum_{i=1}^{v} r_i \left(\bar{Y}_{ij} - \bar{Y}_{.}\right)^2$$

i.e. $T.S.S. = S.S.E. + S.S.T.$

where T.S.S., S.S.T., and S.S.E. are the total sum of squares due to treatments (between treatments S.S) and sum of squares due to error (i.e. within treatment S.S.) given respectively.

$$T.S.S = \sum_i \sum_j \left(Y_{ij} - \bar{Y}_{..}\right)^2$$

$$S.S.T = \sum_i r_i \left(\bar{Y}_i - \bar{Y}\right)^2 = {S_T}^2 \text{ (say)}$$

$$S.S.E = \sum_i \sum_j \left(Y_{ij} - \bar{Y}_i\right)^2 = {S_E}^2 \text{ (say)}$$

ANOVA TABLE FOR C.R.D.

Source of variation	d.f.	S.S.	M.S.S.	Variance ratio
Treatments	$v-1$	${S_T}^2$	${s_T}^2 = \frac{{S_T}^2}{(v-1)}$	
Error	$n-v$	${S_E}^2$	${s_E}^2 = \frac{{S_E}^2}{(n-v)}$	$F_T = \frac{{s_T}^2}{{s_E}^2}$
Total	$n-1$	${S_T}^2 + {S_E}^2$		

under the null hypothesis, $H_0 : \tau_1 = \tau_2 = \cdots\cdots = \tau_v$ against the alternative that all τ's are not equal, the statistic,

$$F_T = \frac{S_T^2}{S_E^2} \sim F_{(v-1,\, n-v)}$$

i.e. F_T Follows F(Central) distribution with $(v-1,\, n-v)$ d.f.

6.6 PROBLEMS

1) Carry out the ANOVA for the following CRD

A	8	8	6	8	9
B	10	8	10	...	...
C	12	10	10	9	...

Sol : We set up null hypothesis that the all treatments effects are homogeneous.

Treatments						T_i	T_i^2 / v_i
A	8	8	6	8	9	29	304.2
B	10	8	10	...	...	28	261.33
C	12	10	10	9	....	41	420.25
			Total			108	985.7833

Here $N = \sum_{i=1}^{3} r_i = 12$; Grand Total = 108

$C.F. = \frac{(108)^2}{72} = 972$; Raw sum of squares = 998 = $\Sigma\Sigma x_{ij}^2$

T.S.S. = $S_{T^2} = R.S.S. - C.F. = 26$

Sum of squares due to treatments = $S_{t^2} = \Sigma \frac{T_i^2}{v} - C.F.$

$= 13.7833$

Error sum of squares = S.S.E.

= $S_{E^2} = S_{T^2} - S_{t^2}$

= 12.2167.

ANOVA Table for C.R.D.

Source of variation	Sum of squares	d.f.	Mean sum of squares	Variance Ratio
Treatments	13.7833	2	$S_t^2 = 6.8917$	$F_t = 5.0771 \sim F_{5\%}(2,8) = 4.26$
Error	12.2167	9	$S_E^2 = 1.3574$	
Total	26	11		

From the ANOVA table we have

$\text{cal.}F_t = 5.0771 > (\text{tab})$ critical value = 4.26. Hence we reject the null hypothesis. Hence we conclude that treatment effects are not homogeneous.

2) Carry out the ANOVA using R.B.D.

Blocks	Treatments			
	I	II	III	IV
A	8	8	6	8
B	10	8	9	10
C	12	10	10	9

Sol : We set up null hypothesis H_0 that the treatment as well as blocks are homogeneous.

Blocks	Treatments					
	I	II	III	IV	T_i	T_{i^2}
A	8	8	6	8	30	900
B	10	8	9	10	37	1369
C	12	10	10	9	41	1681
T_i	30	26	25	27	108	3950
T_{i^2}	900	676	625	729	2930	

$N = t \cdot v = 12,\ C_1 = 108$

$C.F. = 972$

$R.S.S. = \Sigma\Sigma x_{i_j}^2 = 998$

T.S.S. = $S_T^2 = \text{R.S.S.} - \text{C.F.} = 26$

$S_T^2 = \text{S.S.T.} = \frac{1}{v}\Sigma T_{i^2} - \text{C.F.}$

$= 4.6667$

$S_E^2 = \text{SSB} = \frac{1}{t}\Sigma T_j^2 - \text{C.F.} = 15.5$

$\text{S.S.E.} = S_E^2 = S_T^2 - S_t^2 - S_B^2 = 5.8333$

ANOVA table for R.B.D.

Source of variation	Sum of Squares	d.f.	Mean sum of squares	Variance Ratio
Treatments	4.6667	3	1.5556	$F_t = 1.6001 \sim F_{5\%}(3,6) = 4.76$
Blocks	15.5	2	7.75	$F_0 = 7.9716 \sim F_{5\%}(2,6) = 5.14$
Error	5.8333	6	0.9722	
Total	26	11	.....	.....

From the above table of ANOVA, comparing the calculated value with tabulated value, we have

for treatments : $F_t = 1.6001 < 4.76 \Rightarrow$ we accept H_0 and hence treatments are homogeneous.

For blocks : $F_B = 7.91716 > 5.14 \Rightarrow$ we reject null hypothesis and hence blocks are not homogeneous.

3) Carryout the ANOVA for the following Latin Square Design

A	B	C	D
12	20	16	10
D	A	B	C
18	14	11	14
B	C	D	A
12	15	19	13
C	B	A	D
16	11	15	20

Sol :

					R_i	R_i^2	Treatments	T_K	T_K^2
	A	D	C	B	58	3364	A	54	12916
	12	20	16	10					
	D	A	B	C	57	3249	B	44	1936
	18	14	11	14					
	B	C	D	A	59	3481	C	61	3721
	12	15	19	13					
	C	B	A	D	62	3849	D	77	5929
	16	11	15	20					
C_j	58	60	61	57	236	13938	Total	236	14502
C_j^2	3364	3600	3721	3249	13934				

We setup the null hypothesis

for rows : $H_{OR} =$ All row effects are homogeneous

for columns : $H_{OC} =$ All column effects are homogeneous.

for treatments : $H_{Ot} =$ All treatment effects are homogeneous.

Here $N = m^2 = 16.4 = 236$ C.F.$=3481$

R.S.S. = 3638

T.S.S. = $S_T^2 = \text{R.S.S.} - \text{C.F.} = 157;$

$$\text{S.S.R.} = S_R^2 = \frac{1}{m}\Sigma R_i^2 - \text{C.F.} = 3.5$$

$$\text{S.S.C.} = S_C^2 = \frac{1}{m}\cdot\Sigma C_j^2 - \text{C.F.} = 2.5;$$

$$\text{S.S.T.} = S_t^2 = \frac{1}{m}\cdot\Sigma T_k^2 - \text{C.F.} = 144.5$$

$$\text{S.S.E.} = S_E^2 = S_T^2 - S_R^2 - S_C^2 - S_t^2 = 6.5$$

ANOVA TABLE

Source of variation	Sum of Squares	d.f.	Mean sum of squares	Variance ratio
Rows	3.5	3	1.1667	$F_r = 1.077 \sim F_{5\%}(3,6)$
Columns	2.5	3	0.8333	$F_C = 1.3 \sim F_{5\%}(3,6)$ $= 4.76$
Treatments	144.5	3	48.1667	$F_t = 44.4629 \sim F_{5\%}(3,6)$ $= 4.76$
Error	6.5	6	1.08333	
Total	157	15	.....	.....

From the above ANOVA table comparing calculated value with critical values, we have, for rows : $F_R < F_{5\%}(3,6) = 4.76$, thus we accept H_{0R} and conclude that row effects are homogeneous.

for columns : $F_C = 1.3 < F_{5\%}(3,6) = 4.76$, thus we accept H_{0C} and conclude that column effects are homogeneous.

for treat : $F_T = 44.4629 > F_{5\%}(3,6) = 4.76$ thus we reject H_{0t} and conclude that treatments effects are not homogeneous.

EXAMPLES :

1) An experiment was carried out on wheat with three treatments in four randomised blocks. The plan and yield per plot in Kgm. are given below.

BLOCKS			
I	II	III	IV
A	C	A	B
8	10	6	10
C	B	B	A
12	8	9	8
B	A	C	C
10	8	10	9

Analyse the data and state the conclusions.

2) A varietal trial was conducted at a Research station. The design adopted for the same was five randomised blocks of 6 plots each. The yields in lb. per plot of $\frac{1}{20}^{th}$ of an acre) obtained from the experiment are as under.

Blocks	Varieties					
	v_1	v_2	v_3	v_4	v_5	v_6
I	30	23	34	25	20	13
II	39	22	28	25	28	32
III	56	43	43	31	49	17
IV	38	45	36	35	32	20
V	44	51	23	58	40	30

Analyse the design and comment on your findings.

3) Set up the analysis of variance for the following results of a latin square design.

A	C	B	D
12	19	10	8
C	B	D	A
18	12	6	7
B	D	A	C
22	10	5	21
D	A	C	B
12	7	27	17

4) The following are the results of the latin square experiment on the effects of five manual treatments A, B, C, D, E on the yield of sugarcane. Test whether the treatments are equally effective, and if they are not so, compare the treatments A and B.

B	A	E	D	C
405	525	463	441	481
C	D	B	A	E
325	445	429	513	493

E	B	A	C	D
471	492	472	381	410
A	C	D	E	B
552	431	425	572	451
D	E	C	B	A
430	469	432	467	460

6.7 QUESTIONS TO STUDY

1. Explain the principles of experimental design.
2. Explain completely Randomised Desion.
3. Explain the statistical Analysis of C.R.D.

6.8 REFERENCES

1. Fundamentals of Applied Statistics - S.C. Gupta & V.K. Kapoor
2. Methods of Statistical Analysis - P.S. Grewal.

LESSON - 7

LAYOUT AND ANALYSIS OF RANDOMISED BLOCK DESIGN (RBD) & LATIN SQUARE DESIGN (LSD)

LEARNING OBJECTIVES

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

* to clear comprehension of the theory and the practical utility about the concepts of Randomised block design and Latin Square Design.

LESSON OUT LINE

7.1 INTRODUCTION

The CRD is seldom used if the experimental units are not alike. For in that case the variation among the units will vitiate the test of significance of the treatment effects. The simplest design which enables us to take care of the variability among the units is the R.B.D. This is also the simplest design using all the three principles enunciated by Fisher.

Suppose we want to compare the effects of 't' treatments each treatment being replicated an equal number of times, say r times. Then we need $n = rt$ experimental units, and these units are not perhaps homogeneous. The RBD consists of two steps. The first step is to divide the units into r more or less homogeneous groups. In each group or block, we take as many units as there

are treatments. Thus, the number of blocks is the same as the common replication number (r). The same technique should be applied to the units of a block. Variation in technique, if any, should be made between the blocks. In agricultural field experiments, some times a fertility gradient is present. In such a situation, it is advisable to place the block across the gradient in order to get homogeneous material for a block and to obtain major differences between blocks. Familiarity with the nature of the experimental units is necessary for an effective blocking of the material.

The second step is to assign the treatments at random to the units of a block. This randomisation has to be done afresh for each block. This is the difference of an RBD from a CRD. In an RBD randomisation is restricted within a homogeneous block.

With this design each treatment will have the same number of replications. If we want additional replications for some treatments; each of these may be applied to more than one unit in a block.

7.3 LAYOUT OF R.B.D.

Let us obtain the layout of an RBD with 5 treatments, each replicated 3 times. So we need 15 units, which are to be grouped into 3 blocks of 5 plots each. We conveniently number the treatments and also the units in a block. Then, following any method of drawing a random sample (as used for the layout of a CRD), we get a random permutation of the digits from 1 to 5, say 4, 3, 1, 5, 2 for the units of block I. Then we apply treatment number 1 to unit 4, treatment number 2 to unit 3 and so on, finally treatment number 5 to unit 2, of block I. We find another random permutation for block II, and so on for the other block.

7.4 STATISTICAL ANALYSIS OF RBD FOR ONE OBSERVATION PER EXPERIMENTAL UNIT

If in an RBD a single observation is made on each of the experimental units, then its analysis is analogous to ANOVA for a two way classified data with one observation per cell and linear model, assuming various effects to be additive, becomes :

$$Y_{ij} = \mu + \tau_i + b_i + \Sigma_{ij} : \begin{pmatrix} i = 1, 2, \cdots t \\ j = 1, 2, \cdots v \end{pmatrix}$$

where Y_{ij} is the response or the yield of experimental unit from i^{th} treatment and j^{th} block; μ is the general mean effect; τ_i is the effect due to i^{th} treatment; b_j is the effect due to j^{th} block or replicate and Σ_{ij} is the error effect due to random component assumed to be independently normally distributed with mean zero and variance σ_{e^2}, i.e., $\in_{ij}$ are i.i.d. $N\left(0.\sigma_{e^2}\right)$.

If we write

$\sum_i \sum_j Y_{ij} = y_{..} = C_1 =$ grand total of all the $t \times r$ observations.

$\sum_j Y_{ij} = Y_{i.} = T_i =$ total for i^{th} treatment.

$\sum_i Y_{ij} = Y_{\cdot j} = B_j =$ Total for j^{th} block;

then, we get

$$\sum_i \sum_j \left(Y_{ij} - \bar{Y}_{..}\right)^2 = \sum_i \sum_j \left[\left(\bar{Y}_{i.} - \bar{Y}_{..}\right) + \left(\bar{Y}_{\cdot j} - \bar{Y}_{..}\right) + \left(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{\cdot j} + \bar{Y}_{..}\right)^2\right]$$

$$= r\sum_i \left[\left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2 + t\sum_j \left(\bar{Y}_{\cdot j} - \bar{Y}_{..}\right)^2 + \sum_i \sum_j \left(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{\cdot j} + \bar{Y}_{..}\right)^2\right]$$

where the product terms vanish since the algebraic sum of deviations from mean is zero. Thus we have

$$T.S.S. = S.S.T. + S.S.B + S.S.E.$$

Where T.S.S. is the total sum of squares and S.S.T., S.S.B. and S.S.E. are the sum of squares due to treatments, blocks and error respectively, given by

$$T.S.S. = \sum_i \sum_j \left(Y_{ij} - \bar{Y}_{..}\right)^2, \quad S.S.T. = S_{T^2} = r\sum_i \left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2$$

$$S.S.B. = S_{B^2} = t\sum_j \left(\bar{Y}_{\cdot j} - \bar{Y}_{..}\right)^2, \; S.S.E. = S_{E^2} = T.S.S. - S.S.T - S.S.B$$

Hence the total sum of squares is partitioned or split into three sum of squares whose degrees of freedom add to the degrees of freedom of T.S.S. Therefore, by Cochran's theorem each of these S.S. divided by $\sigma_{\in^2}$ is independently distributed as x^2 variate.

ANOVA TABLE FOR R.B.D.

Source of Variation	d.f.	S.S.	M.S.S.	Variance Ratio
Treatments	$t-1$	S_T^2	$s_{T^2} = \frac{S_T^2}{(t-1)}$	$F_T = \frac{s_T^2}{S_E^2}$
Blocks or Replicates	$r-1$	S_E^2	$s_{B^2} = \frac{S_B^2}{(r-1)}$	$F_B = \frac{s_B^2}{S_E^2}$
Error or Residual	$(t-1)(r-1)$	S_E^2	$s_{E^2} = \frac{S_E^2}{(t-1)(r-1)}$	

M.S.S. of treatments and replicates (or blocks) are tested for significance aganist error mean S.S.

Under the null hypothesis $H_t : T_1 = T_2 = \cdots\cdots = T_t$ against the alternative that all $T's$ are not equal;

$$F_T = \frac{s_T^2}{s_E^2} \sim F\left[(t-1), (t-1)(r-1)\right]$$

i.e. F_T follows F (central) district, with $(t-1), (t-1)(r-1)$ d.f. Thus if f_T is greater than tabulated F for $(t-1),(t-1)(r-1)$ d.f. at certain level of significance, usually 5% then we reject the null hypothesis H_t and conclude that the treatments differ significantly. If F_t is less than tabulated value then F_T is not significant and we conclude that the data do not provide any evidence against the null hypothesis which may be accepted.

Similarly under the null hypothesis $H_b : b_1 = b_2 = \cdots = b_r$, against the alternative that the $b's$ are not all equal,

$$F_B = \frac{s_B^2}{s_E^2} \sim F\left[(r-1), (r-1)(t-1)\right]$$

7.5 ADVANTAGES AND DISADVANTAGES OF R.B.D.

The RBD has many advantages over other designs. It is quite flexible. It is applicable to a moderate number of treatments. If extra replication is necessary for some treatments, these may be applied to more than one unit (but to some number of units) per block. Since variability among replicates can be eliminated from experimental error, it is not necessary to use continuous blocks. It also enables us to use different techniques to different blocks, though the technique should be the same within a block. The analysis is straight forward and remains so if due to accident data on an entire block or treatment be missing. If data from individual units be missing, then we can use yates, missing plot technique (vide section) to estimate the values and perform the test. By grouping the units, we obtain greater precision than is obtainable with the C.R.D.

This is the most popular design with experimenter in view of its simplicity, flexibility and validity. No other design has been used so frequently as the RBD. If satisfactory results can be obtained with this design, then we shall not use other complicated designs.

The chief disadvantage is that if the blocks are not internally homogeneous, then a large error term will result. As usually occurs in field experiments, with an increase in the number of treatments, the block size increases and so one has a lesser control over error, for the block will include material of a more heterogeneous nature. In such cases, special types of incomplete block designs are used to reduce the block size.

7.6 ESTIMATION OF MISSING VALUE IN R.B.D.

Let the observations $y_{ij} = x$ in the j^{th} block and receiving the i^{th} treatment be missing.

	Treatment		Tot
	1	$2 \cdots\cdots i \cdots\cdots t$	
1	Y_{11}	$Y_{21} \cdots\cdots Y_{i1} \cdots\cdots Y_{t1}$	$Y_{\cdot 1}$
2	Y_{12}	$Y_{22} \cdots\cdots Y_{i2} \cdots\cdots Y_{t2}$	$Y_{\cdot 2}$
--	---	----------	--
--	---	----------	--
--	---	----------	--
j	Y_{1j}	$Y_{2j} \cdots\cdots x \cdots\cdots Y_{ij}$	$Y_{ij}' + x$
--	---	----------	---
--	---	----------	---
--	---	----------	---
r	Y_{1r}	$Y_{2r} \cdots\cdots Y_{ir} \cdots\cdots Y_{tr}$	---
Tot	Y_1	$Y_2 \cdots\cdots \left(Y_i' + x\right) \cdots\cdots Y_r$	$Y_{\cdot\cdot}' + x$

where $Y_{i\cdot'}$ is total of known observations getting i^{th} treatment,

$Y_{\cdot j'}$ is total of known observations in j^{th} block, and

$Y_{\cdot\cdot'}$ is total of all the known observations

$$T.S.S. = \Sigma\Sigma Y_{ij^2} - C.F.$$

$$= x^2 + \text{const w.r.t } x - C.F.$$

$$S.S.T. = \frac{1}{r}\left[\left(y_{i\cdot}' + x\right)^2 + \text{constant w.r.t. } x\right] - C.F.$$

$$S.S.B. = \frac{1}{t}\left[\left(Y_{i\cdot}' + x\right)^2 + \text{constant w.r.t.} x\right] - C.F.$$

where $C.F. = \dfrac{(Y'_{..} + x)^2}{rt}$

$E =$ Residual S.S.

$= T.S.S. - S.S.B - S.S.T.$

$= x^2 - \dfrac{1}{t}\left(Y_{..}' + x\right)^2 - \dfrac{1}{r}\left(Y'_{r.} + x\right)^2$

$E =$ Residual S.S.

$=$ T.S.S. $-$ S.S.B. $-$ S.S.T.

$= x^2 - \dfrac{1}{t}\left(Y_{.j}' + x\right)^2 - \dfrac{1}{r}\left(Y'_{i.} + x\right)^2 + \dfrac{\left(Y_{..}' + x\right)^2}{rt}$ + constant terms independent of x.

We shall choose x such that E is minimum. For E to be minimum for variations in x, we must have

$$\frac{\partial E}{\partial x} = 0 = 2x - \frac{2}{t}\left(Y'_{.j} + x\right) - \frac{2}{r}\left(Y'_{i.} + x\right) + \frac{2}{rt}\left(Y'_{..} + x\right)$$

$$x\left[1 - \frac{1}{t} - \frac{1}{r} + \frac{1}{rt}\right] = \frac{1}{t}Y'_{.j} + \frac{1}{r}Y'_{i} - \frac{1}{rt}Y'_{..}$$

$$x = \frac{ry'_{.j} + ty'_{i.} - y'_{..}}{(r-1)(t-1)}$$

Problem : In the table given below are the yields of 6 varieties in a 4 replicate experiment for which one value is missing. Estimate, the missing value and analyse the data.

Treatments

	1	2	3	4	5	6	Block Total
1	18.5	15.7	16.2	14.1	13.0	13.6	91.1
2	11.7	--	12.9	14.4	16.9	12.5	68.4
3	15.4	16.6	15.5	20.3	18.4	21.5	107.8
4	16.5	18.6	12.7	15.7	16.5	18.0	98.0
Treatment Totals T_i	62.1	50.9	57.3	64.5	64.8	65.7	365.3

Solution : Estimation of missing value.

In the notations of estimation of missing value in R.B.D. we have

$$Y_{.j}' = 68.4,\ Y_{i.}' = 50.9 \text{ and } Y_{..}' = 365.3,\ r = 4 \text{ and } t = 6$$

Hence on using, the missing value x (say) is estimated by

$$x = \frac{ry_{.j}' + ty_{i.}' - Y_{..}'}{(r-1)(t-1)}$$

$$= \frac{4 \times 68.4 + 6 \times 50.9 - 365.3}{(4-1)(6-1)} = \frac{213.7}{15}$$

= 14.25

Analysis of the design. $(H_0) H_t : \tau_1 = \tau_2 = \cdots\cdots = \tau_6$

$$H_b : b_1 = b_2 = \cdots\cdots b_4$$

i.e. treatments as well as replicates are homogeneous.

Substituting the value of x in the given table, we have

Treatment totals :	T_1	T_2	T_3	T_4	T_5	T_6
	62.1	65.15	57.3	64.5	64.8	65.7

Block Totals :	B_1	B_2	B_3	B_4
	91.1	82.65	107.8	98.0

$$G = \sum_i \sum_j Y_{ij} = 379.55,\ N = rt = 4 \times 6 = 24$$

$$C.F. = \frac{G^2}{N} = \frac{(379.55)^2}{24} = 6002.42$$

$$R.S.S. = \sum_{i=1}^{6} \sum_{j=1}^{4} Y_{ij}^2 = 6151.94$$

Total $S.S. = 6151.94 - 6002.42 = 149.52$

$S.S.$ due to treatments $= \frac{1}{4} \sum_{i=1}^{6} T_{i^2} - C.F.$

$$= \frac{24060.0025}{4} - 6002.42$$

= 12.58

S.S. due to Replicates (Blocks) $(S.S.B.)$

$$= \frac{1}{6}\sum_{j=1}^{4} B_j^2 - C.F.$$

$$= \frac{36355.07}{6} - 6002.42 = 56.76$$

Error $S.S. = T.S.S. - S.S.T. - S.S.B. = 80.18$

Using equation, the bias or adjustment factor for treatment $S.S.$ is given by

$$\frac{\left(Y_{.j}' + ty_{i.}' - Y_{..}'\right)^2}{t(t-1)(r-1)^2} = \frac{(68.4 + 6\times 50.9 - 365.3)^2}{6\times 5\times 9} = 0.267$$

$\therefore$ Adjusted value of treatment $S.S. = 12.580 - 0.267 = 12.313$

ANOVA TABLE

Source of Variation	d.f.	S.S.	M.S.S.	Variance Ratio F
Treatments	5	12.313	2.06	$F_t = 2.33$
Blocks	3	56.760	18.92	$F_b = 3.30$
Error	14	80.45	5.72	
Total	22			

Tabulated value of $F_{0.05}(5.14) = 2.96$

and $F_{0.05}(3.14) = 3.34$

Thus we see that neither F_t nor F_b are significant and consequently H_t and H_b may be retained, i.e., we may regard the treatments as well as blocks to be homogeneous.

Example : An experiment was carried out to determine the effect of claying the ground on the field of barley grains; amount of clay used was as follows :

A : No Clay; B = Clay at 100 per acre;

C : Clay at 200 per acre; D : Clay at 300 per acre

The yields were in plots of 8 meters by 8 meters and layout were

Column / Row	I	II	III	IV	Row Total
I	D	B	C	A	83.1
	29.1	18.9	29.4	5.7	
II	C	A	D	B	66.9
	16.4	10.2	21.2	19.1	
III	A	D	B	C	105.2
	5.4	38.8	24.0	37.0	
IV	B	C	A	D	105.0
	24.9	41.7	9.5	28.9	
Column Tot.	75.8	109.6	84.1	90.7	360.2

(a) Perform the ANOVA and calculate the critical difference for the treatment mean yields.

(b) Compare the efficiency of the above latin square design over

(i) R.B.D. and (ii) C.R.D.

(c) Yield under 'A' in the first column was missing estimate the missing value and carry out the ANOVA.

7.8 LATIN SQUARE DESIGN (L.S.D.)

The principle of 'local control' was used in the RBD by grouping the units in one way, i.e. according to blocks. The grouping can be carried one step forward and we can group the units in two ways, each way corresponding to a source of variation among the units, and get the L.S.D. This design is used with advantage in agricultural field experiments where the fertility contours are not always known. Then the LSD eliminates the initial variability among the units in two orthogonal. directions. The LSD has also been used successfully in industry and in the laboratory.

In this design, the number of treatments equals the common replication number per treatment. So letting in stand for the number of treatments as well as the number of replications for each treatment, the total number of experimental units needed for this design is mxm. These m^2 units are arranged in 'm' rows (one source of variation) and 'm' columns (second source of variation). Then the 'm' treatments are alloted to these 'm^2' units at random, subjec to the condition that each treatment occurs once and only once in each row and in each column.

The arrangement of units and allocation of treatments to units make the m rows similar to m complete block of an RBD (the same is true also of the m columns).

The LSD is actually an incomplete three way layout, where all the three factors, rows, columns and treatments, are at the same number of levels (m). For a complete three way layout

with each factor at m levels, we need m^3 experimental units. But in the LSD we take observations on only m^2 of these m^3 units according to the plan stated above.

As an example, let us consider a 4×4 latin square for comparing four varieties of a crop. We take a rectangular field divided into $4\times 4=16$ plots, arranged in four rows and four columns. We represent the varieties by A, B, C, and D. Then the following is a particular 4×4 latin square.

	Columns			
	D	C	B	A
Rows	C	B	A	D
	B	A	D	C
	A	D	C	B

7.9 LAYOUT OF L.S.D.

In connection with the random choice of a latin square, we first define the following :

The totality of L.S.D's obtained from a single LSD by permuting the rows, columns and letters (treatments) is called a transformation set. An $m\times m$ latin square with the m letters A, B, C, ... in the natural order occuring in the first row and in the first column is called a standard square (square in the canonical form). Thus the standard square corresponding to the square cited above is

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

From a standard $m\times m$ latin square, we may obtain $m!(m-1)!$ different LSDs by permuting all the 'm' columns and the $(m-1)$ rows except the first row. Hence there are in all $m!(m-1)!$ different LSDs with the same standard square. Thus the total number of different LSDs in a transformation set is $m!(m-1)!$ times the number of standard LSDs in the set.

As in all other designs, the necessity of randomisation applies to the LSD also. In order to give all $m\times m$ LSDs equal probability of being selected, we select with equal probability one standard square from all the standard $m\times m$ LSDs and then randomise the columns and rows, excluding the first row. More detailed instructions and tables of standard LSDs are given in the introduction to tables XV and XVI of Fisher and Yates. Statistical Tables for Biological, Agricultural and Medical Research.

Two $m\times m$ latin squares are said to be orthogonal if, when these are super imposed, every one of the m^2 pairs of numbers occurs once and once only. A set of $m\times m$ latin squares is called orthogonal if every pair of them is orthogonal.

8.0 ADVANTAGES OF LATIN SQUARE DESIGN

(1) With two way grouping or stratisfication L.S.D. controls more of the variation than C.R.D. or R.B.D.

The two way elimination of variations as a result of cross grouping often results in small error mean sum of squares. Thus in field experimentation if the fertility gradient is in two directions at right angles to each other (i.e. if there is a diagonal trend in fertility) or in one unknown direction than L.S.D. is likely to be more efficient than R.B.D. Infact L.S.D. can be used with advantage in those cases where the variation in experimental material is from two orthogonal sources. As regards the applications of L.S.D., Professor Fisher says, "If experimenters were only concerned with the comparison of four to eight treatments or varieties, it (LSD) would be not merely the principal but almost the universal design employed".

(2) L.S.D. is an incomplete 3-way layout. Its advantage over the complete 3-way layout is that instead of m^3 experimental units only m^2 units are needed. Thus a 4×4 L.S.D. results in saving of 64 - 16 = 48 observations over a complete 3-way layout.

(3) The statistical analysis is simple though slightly complicated than for R.B.D. Even with missing data the analysis remains relatively simple.

(4) More than one factor can be investigated simultaneously and with fewer trials than more complicated designs.

DISADVANTAGES OF L.S.D.

(1) The fundamental assumption that there is no interaction between different factors (i.e., the factors act independently) may not be true in general.

(2) Unlike R.B.D., in L.S.D. the number of treatments is restricted to the number of replications and this limits its field of application. L.S.D. is suitable for the number of treatments between 5 and 10 and for more than 10 to 12 treatments the design is seldom used since in that case the square becomes too large and does not remain homogeneous.

(3) Incase of missing plots, when several units are missing the statistical analysis becomes quite complex. If one or two blocks in a field are attacked by some diseases or pest then in R.B.D. we can easily omit the data for these blocks without complicating the analysis at all whereas a much more complicated analysis is necessitated in L.S. experiment under similar conditions.

(4) In the field layout, R.B.D. is much easy to manage than L.S.D. since the former can be performed equally well on a square or rectangular field or a field of any shape where as for the latter approximately a square field is necessary.

Statistical Analysis of m x m L.S.D. for one Observation per Experimental Unit

Let $Y_{ijk}\,(i, j, k = 1, 2, \cdots\cdots m)$ denote the response from the unit (plot, in field experimentation) in the i^{th} row, j^{th} column and receiving the k^{th} treatment. The triplet (i, j, k) assumes only m^2 of the possible m^3 values of an L.S. selected by the experiment. If S represents the set of m^2 values

then symbolically $(i, j, k) \in S$. If a single observation is made per experimental unit then the linear additive model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma T_k + \in_{ijk}, (i, j, k) \in S$$

where μ is the constant mean effect; α_i, β_j and T_k are the effects due to the i^{th} row, j^{th} column and k^{th} treatment respectively and $\in_{ijk}$ is error effect due to random component assumed to be normally distributed with mean zero and variance $\sigma_{\in^2}$, i.e. $\Sigma_{ijk} \sim N\left(0, \sigma_{\in^2}\right)$. If we write

$C_1 = Y_{...} =$ Total of all the m^2 observations.

$R_i = Y_{i..} =$ Total of the m observations in the i^{th} row.

$C_j = Y_{.j.} =$ Total of the m observations in the j^{th} column.

$T_k = Y_{..k} =$ Total of the m observations from k^{th} treatment.

then heuristically, we have

$$\sum_{i,j,k\in S}\left(Y_{ijk} - \bar{Y}_{...}\right)^2 = \sum_{i,j,k\in S}\left[\left(\bar{Y}_{i..} - \bar{Y}_{...}\right) + \left(\bar{Y}_{.j.} - \bar{Y}_{...}\right) + \left(\bar{Y}_{..k} - \bar{Y}_{...}\right) + \left(Y_{ijk} - \bar{Y}_{i..} - \bar{Y}_{.j.} - \bar{Y}_{..k} + 2\bar{y}_{...}\right)\right]^2$$

$$= m\sum_i\left(\bar{Y}_{i..} - \bar{Y}_{...}\right)^2 + m\sum_j\left(\bar{Y}_{.j.} - \bar{Y}_{...}\right)^2 + m\sum_k\left(\bar{Y}_{..k} - \bar{Y}_{...}\right)^2 + \sum_{i,j,k\in S}\left(Y_{ijk} - \bar{Y}_{i..} - \bar{Y}_{.j.} - \bar{Y}_{..k} + 2\bar{y}_{...}\right)^2$$

the product terms vanish, since the algebraic sum of deviations from mean is zero.

$$\therefore T.S.S. = S.S.R. + S.S.C. + S.S.T. + S.S.E.$$

Where $T.S.S.$ is the total sum of squares and $S.S.R.$, $S.S.C.$, $S.S.T.$, and $S.S.E.$ represent sum of squares due to rows, columns, treatments and error respectively, given by

$$T.S.S. = \sum_{i,j,k\in S}\left(Y_{ijk} - \bar{Y}_{...}\right)^2; S.S.R. = S_{R^2} = m\sum_i\left(\bar{Y}_{i..} - \bar{Y}_{...}\right)^2$$

$$S.S.C. = S_{C^2} = m\sum_j\left(\bar{Y}_{.j.} - \bar{Y}_{...}\right)^2; S.S.T. = S_{T^2} = m\sum_k\left(\bar{Y}_{..k} - \bar{Y}_{...}\right)^2 \text{ and}$$

$$S.S.E. = S_{E^2} = T.S.S. - S.S.R. - S.S.C. - S.S.T.$$

ANOVA TABLE FOR m x m L.S.D.

Soure of Variation	d.f.	S.S.	M.S.S.	Variance Ratio 'f'
Rows	$m-1$	S_R^2	$s_R^2 = \frac{S_R^2}{m-1}$	$F_R = \frac{s_{R^2}}{s_{E^2}}$
Columns	$m-1$	S_C^2	$s_C^2 = \frac{S_C^2}{m-1}$	$F_C = \frac{s_{C^2}}{s_{E^2}}$
Treatments	$m-1$	S_T^2	$s_T^2 = \frac{S_T^2}{m-1}$	$F_T = \frac{s_{T^2}}{s_{E^2}}$
Error	$(m-1)(m-2)$	S_E^2	$s_E^2 = \frac{S_E^2}{(m-1)(m-2)}$	
Total	m^2-1			

Let us set up the null hypothesis

for row effects, $H_\alpha = \alpha_1 = \alpha_2 = \cdots\cdots = \alpha_m = 0$;

for column effects, $H_\beta : \beta_1 = \beta_2 = \cdots\cdots = \beta_m = 0$

and for treatment effects, $H_T : T_1 = T_2 = \cdots\cdots = T_m = 0$

The variance ratios F_R, F_C and F_T follow (central) F distribution with $(m-1), (m-1)(m-2)$ d.f. under the null hypothesis. H_α H_β and H_T respectively.

QUESTIONS TO STUDY

1. Explain R.B.D.
2. Explain Statistical Analysis of L.S.D.
3. Explain layout of R.B.D.
4. Explain missing value technique in R.B.D. with example

REFERENCE

1. Fundamentals of Applied Statistics - S.C Gupta & V.K. Kapoor
2. Methods of Statistical Analysis - P.S. Grewal.

LESSON - 8

BASIC PRINCIPLES OF CONTROL CHARTS AND VARIABLE CONTROL CHARTS

OBJECT OF LESSON

After studying this lesson, the student is expected to have a

* Comprehension of the theory and practical utility about the concepts of SQC and control charts for variables and their applications.

STRUCTURE OF THE LESSON : This consists of - sections as detailed below.

8.1 INTRODUCTION

By statistical quality control we mean the various statistical methods used for the maintenance of quality in a continuous flow of manufactured products. In any manufacturing process, it is not possible to produce goods of exactly the same quality i.e., variation is inevitable. Certain small variation is natural to the process and being due to chance causes, cannot be prevented. This variation, therefore, is called allowable variation. Sometimes superimposed on this there will be variation which occurs when the process goes wrong, the causes of this variation being assignable. Such variation, therefore is called preventable variation. The main purpose of SQC is to devise statistical methods for separating allowable variation from preventable variation, so that we may take appropriate steps as quickly as possible whenever assignable causes are operating in the process. In otherwords, an attempt is made to weed out systematic causes of variation as soon as they occur, so that actual variation may be supposed to be due to the inevitable random causes alone.

In the above type of problem, our aim is to control the manufacturing process so that the proportion of defective items is not excessively large. This is known as process control. In another type of problem, we like to ensure that lots of manufactured goods do not contain excessively large proportions of defective items. This is known as product control. The two are distinct problems because, even when the process is in control, so that the proportion of defective articles for the entire output over a long period will not be large, an individual lot of items may not be of satisfactory quality. Process control is achieved mainly through the technique of control charts, pioneered by W.A. Shewart, where as product control is achieved through sampling inspection pioneered by Dodge and Romig.

ADVANTAGES OF STATISTICAL QUALITY CONTROL

1. It provides a basis for coordination between designers, production personnel and inspection personnel.
2. The detection and elimination of the assignable causes of variation (a) helps in solving many production problems (b) improves the product quality and (c) reduces the spoilage and re-work.
3. It provides continuous inspection to the product at various stages of the manufacturing process. It makes the inspection procedure economical on the basis of samples than what is done by any amount of final inspection.
4. A statistically controlled process is predictable. Here the standard and the number of defectives can be predicted for the product of the furture.
5. The process control eliminates the need of 100% inspection of the finished goods and consequently introduces savings in inspection by a suitable system of sampling.
6. It provides an information which can help the management in taking appropriate steps in either changing the process or making improvement in the machinary or alter the design specifications. The knowledge of the capabilities of the production process helps the management to consider various possible alternatives for the design keeping in view of the market demand and consumer's behaviour.
7. The process control reduces waste of time and material to absolute minimum by giving an early warning about the occurence of defects. Also the rejections by consumers can sufficiently be reduced.
8. Statistical techniques find their applications not only in the sphere of production, but also in other areas like, comparison of budgetary figures with actual performances, advertising packaging, scraps, and spoilage, recoveries etc.

So quality control means better quality with less cost of production, and hence it can better be called profit control by improving the process, or by improving the inspection methods, or by changing the design, or by changing the user's idea of what he wants.

PROCESS AND PRODUCT CONTROL

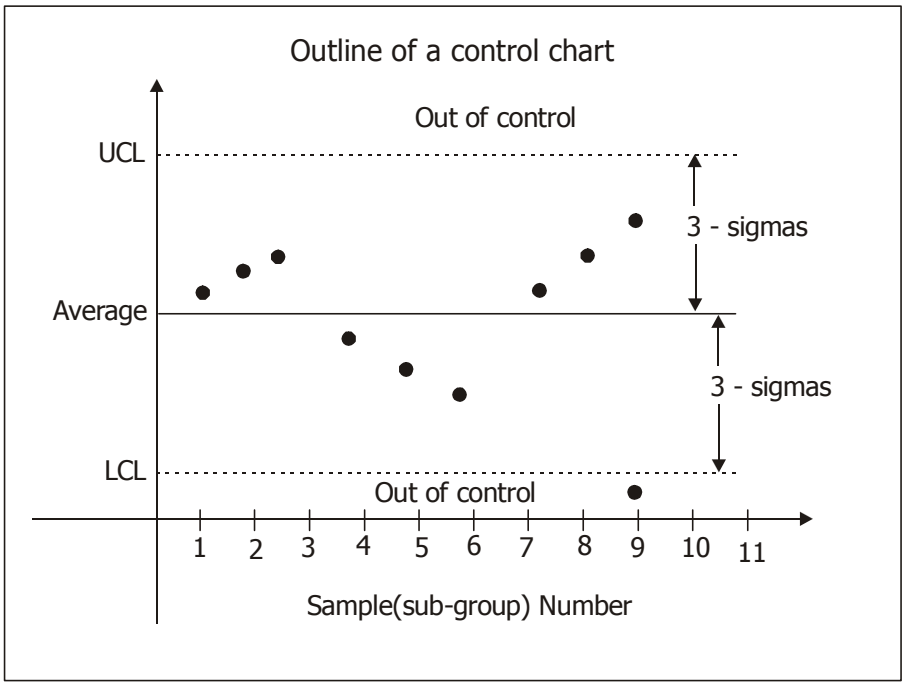
The main objective in any production process is to control and maintain the quality of the manufactured product so that it conforms to specified quality standards. In otherwords, we want to ensure that the proportion of defective items in the manufactured product is not too large. This is called 'process control' and is achieved through the technique of control charts pioneered by W.A. Shewart.

On the otherhand, by product control we mean controlling the quality of the product by critical examination at strategic points and this is achieved through 'sampling inspection plans' pioneered by Dodge and Romig. Product control aims at guaranteeing a certain quality level to the consumer regardless of what quality level is being maintained by the producer. In otherwords, it attempts to ensure that the product marketed by sale department does not contain a large number of defective items.

CONTROL CHARTS

The most common working statistical tools in quality control are the Shewart's control charts. A young industrial statistician and physician W.A. Shewart, of the Bell Telephone laborataries introduced this technique in 1924. Control charts are the running record graphs of the performance of some quality characteristics to examine the presence of assignable causes. While preparing a control chart on the graph, the statistic corresponding to the characteristics under study is taken on y-axis (as ordinates) and the subgroup or sample number is taken on x-axis (as absciss) and the points are plotted as dots.

The control charts consists of three horizontal lines based on 3σ-limits. The upper limit is called Upper Control Limit (UCL), the lower limit is called Lower Control Limit (LCL) and the middle line corresponding to process average is called Central Line (CL). Thus a control chart looks like the one given below.



If all the sample points fall between UCL and LCL we say that the process is under control. Any sample point going outside UCL and LCL (i.e., 3σ control limits) is an indication of the lack of statistical control. i.e., presence of some assignable causes of variation which must be traced, identified and eliminated.

In the control chart, upper control limit (UCL) and lower control limit (LCL) are usually plotted as dotted lines and central line (CL) is plotted as a bold (dark) line. If 't' is the underlying statistic then these values depend on the sampling distribution of t and are given by

$$\text{UCL} = \text{E}(t) + 3\text{S.E.}(t)$$

$$\text{LCL} = \text{E}(t) - 3\,\text{S.E}(t)$$

$$C.L. = E(t)$$

where 't' is the statistic under consideration.

8.2 STATISTICAL BASIS OF CONTROL LIMITS

Consider the statistic $t = t(x_1, x_2, \cdots, x_n)$, a function of the sample observations $x_1, x_2, \cdots, x_n$. Let

$$E(t) = \mu_i \text{ and } var(t) = \sigma_i^2$$

If the statistic 't' is normally distributed, then from the fundamental area property of the normal distribution, we have

$$P[\mu_t - 3\sigma_t < t < \mu_t + 3\sigma_t] = 0.9973$$

$$\Rightarrow P[|t - \mu_t| < 3\sigma_t] = 0.9973$$

$$\Rightarrow P[|t - \mu_t| > 3\sigma_t] = 0.0027$$

In otherwords, the probability that a random value of t goes outside the $3\sigma-$limits i.e., $\mu_t \pm 3\sigma_t$ is 0.0027, which is very small. Hence, if t is normally distributed, the limits of variation should be between $\mu_t + 3\sigma_t$ and $\mu_t - 3\sigma_t$, which are termed respectively the upper control limit (UCL) and lower control limit (LCL). If, for the i^{th} sample, the observed t_i lies between the upper and lower control limits, this variation is attributed to chance that is, the process is statistically under control. If when any observed t_i falls outside the control limits, it is an indication that some assignable cause has crept in, which must be identified and eliminated.

If the assumption regarding normality of the statistic t does not hold, then the above argument does not remain strictly valid. For non-normal population that is, if the sampling distribution of statistic t is not normal we apply Chebychev's Inequality in Probability theory which states that for any constant $K > 0$.

$$P\{|t - E(t)| < K\} \geq 1 - \frac{var(t)}{K^2}$$

$$\Rightarrow P\{|t - \mu_t| < 3\sigma_t\} \geq 1 - \frac{\sigma_t^2}{9\sigma_t^2} = \frac{8}{9}$$

If σ_t is not known and is estimated from the sample data and consequently Chebychev's inequality does not hold if σ_t is not known. Moreover, according to the central limit theorem in probability, the statistics of observations drawn from non-normal populations will exhibit nearly normal behaviour.

8.3 CONTROL CHARTS FOR VARIABLES

The charts used for characteristics on which the actual measurements in numerical forms are possible to be made, are called control charts for variables. Many quality characteristics of a product are measurable and can be expressed in specific units of measurement such as diameter of a screw, tensile strength of steel pipe, specific resistance of a wire, life of an electric bulb, etc. Such variables are of continuous type and are regarded to follow normal probability law. In this category we have three charts.

(i) Mean chart or $\bar{X}$ - Chart

(ii) Range chart or R-Chart

ADVANTAGES :

1. When quality specifications are mentioned in terms of variates, then $\bar{X}$ and R charts are used to know the

   (a) Basic variability of production characteristics.

   (b) Consistent record of the performance.

   (c) Average level of the production characteristics.

Control Chart for Mean or $\bar{X}$ - Chart

Definition : When sample means $(\bar{x})$ are plotted against subgroup number on a control chart, then it is called $\bar{x}$ - chart or mean chart.

Collection of data and computation of statistics : A series of samples often called subgroups is drawn at suitable intervals during the production. The mean, range, s.d. values of each subgroup are computed.

Sub-sample	Values	Mean	Range	S.d.
	$1\ 2 \cdots\cdots n$	$\bar{x}$	R	S
1	$X_{11}\ X_{12} \cdots\cdots X_{1n}$	$\bar{X}_1 = \sum_{j=1}^{n} \frac{X_{1j}}{n}$	R_1	S_1
2	$X_{21}\ X_{22} \cdots\cdots X_{2n}$	$\bar{X}_2 = \sum_{j=1}^{n} \frac{X_{2j}}{n}$	R_2	S_2
.	.....	.....	...	....
.	.....	.....	...	....
.	.....	.....	...	....
K	$X_{K1}\ X_{K2} \cdots\cdots X_{Kn}$	$\bar{X}_K = \sum_{j=1}^{n} \frac{X_{Kj}}{n}$	R_K	S_K

That is, let X_{ij}, $j=1,2,\cdots,n$ be the measurements on the i^{th} sample $(i=1,2,\cdots,K)$. The mean $\bar{X}_i$, the range R_i and the standard deviation s_i for the i^{th} sample given by

$$\left.\begin{aligned}\bar{X}_i &= \frac{1}{n}\sum_j X_{ij}\\ R_i &= \text{Max } X_{ij} - \text{Min } X_{ij}\\ s_i^2 &= \frac{1}{n}\sum_j \left(X_{ij}-\bar{X}_i\right)^2\end{aligned}\right\} \quad (i=1,2,\cdots\cdots,K)$$

then find $\bar{\bar{X}}$, $\bar{R}$ and $\bar{s}$, the averages of sample means, sample ranges and sample standard deviations, respectively as follows :

$$\bar{\bar{X}} = \frac{1}{K}\sum_i \bar{X}_i$$

$$\bar{R} = \frac{1}{K}\sum_i R_i$$

$$\bar{s} = \frac{1}{K}\sum_i s_i$$

STATISTICAL BASIS :

If 'μ' is the process mean and 'σ' is the process standard deviation. We also know that the sample mean $\bar{x}$ will have the normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, n being the size of each subgroup i.e., the number of observations in each subgroup is the same. Then 99.73% confidence limits (or 3σ limits) for sample mean are

$$E(\bar{x}) \pm 3S\cdot E\cdot(\bar{x})$$

i.e., $$\mu \pm 3\cdot\frac{\sigma}{\sqrt{n}} \qquad \left(\because E(\bar{x})=\mu,\ S\cdot E(\bar{X})=\frac{\sigma}{\sqrt{n}}\right)$$

Also from the sampling distribution of range we know that

$$E(R) = d_2\cdot\sigma$$

Where d_2 is a constant depending on the sample size. Thus an estimate of σ can be obtained from $\bar{R}$ by the relation

$$\bar{R} = d_2 \cdot \sigma \Rightarrow \hat{\sigma} = \frac{\bar{R}}{d_2}$$

Also $\bar{\bar{X}}$ gives an unbiased estimate of the population mean μ, since

$$E\left(\bar{\bar{X}}\right) = \frac{1}{K}\sum_{i=1}^{K} E\left(\bar{X}_i\right) = \mu$$

CONTROL LIMITS :

Case 1 : Standards are known or given.

If the process mean μ and the process standard deviation σ are known then the 3σ-limits for $\bar{X}-$ chart are given by

$$E\left(\bar{X}\right) \pm 3 \cdot S \cdot E\left(\bar{X}\right) = \mu \pm \frac{3\sigma}{\sqrt{n}}$$

$$= \mu \pm A\sigma \left(\because A = \frac{3}{\sqrt{n}} \right)$$

If μ' and σ' are known or the specified values of μ and σ respectively, then

Upper Control Limit $\left(UCL_{\bar{x}}\right) = \mu' + A\sigma'$

Central Line $\left(CL_{\bar{x}}\right) = \mu'$

Lower Control Limit $\left(LCL_{\bar{x}}\right) = \mu' - A\sigma'$

Case 2 : Standards are not known or given.

If both μ and σ are not known then these are estimated by the observed data. Here

$\hat{\mu} = \bar{\bar{x}} = \frac{1}{K}\sum_{i=1}^{K} \bar{x}_i$ mean of all means.

$\hat{\sigma} = \frac{\bar{R}}{d_2} = \frac{1}{d_2}\left[\frac{1}{K}\sum_{i=1}^{K} R_i\right]$, Mean of the sample ranges.

or $\hat{\sigma} = \frac{\bar{s}}{c_2} = \frac{1}{c_2}\left[\frac{1}{K}\sum_{i=1}^{K} s_i\right]$, mean of the sample standard deviations.

i.e., when $\sigma-$ chart is paired with $\bar{X}-$ chart, where d_2 and c_2 are factors (constants) based on n and

$$E(\bar{X})=\mu,\ E(\bar{R})=d_2\sigma,\ E(\bar{s})=c_2\sigma$$

(a) Then the control limits for $\bar{X}-$ chart when the control limits are obtained interms of $\bar{R}$,

i.e., estimate of σ obtained from the relation

$$\bar{R}=d_2\hat{\sigma}\Rightarrow\hat{\sigma}=\frac{\bar{R}}{d_2}\text{ then}$$

$3\sigma-$ control limits for $\bar{X}-$ chart are

$$\hat{\mu}\pm3\frac{\hat{\sigma}}{\sqrt{n}}$$

Where $\hat{\mu}=\bar{\bar{x}},\ \hat{\sigma}=\frac{\bar{R}}{d_2}$

$$\therefore \text{UCL}:\bar{\bar{X}}+3\cdot\frac{\bar{R}}{d_2}\cdot\frac{1}{\sqrt{n}}\text{ or }\bar{\bar{X}}+A_2\bar{R}$$

$$\text{CL}:\bar{\bar{X}}$$

$$\text{LCL}:\bar{\bar{X}}-3\cdot\frac{\bar{R}}{d_2}\cdot\frac{1}{\sqrt{n}}\text{ or }\bar{\bar{X}}-A_2\bar{R}\quad\left(\because A_2=\frac{3}{d_2\sqrt{n}}\right)$$

Since d_2 is a constant depending on n, $A_2=\frac{3}{d_2\sqrt{n}}$ also depends only on n and its values have been computed and tabulated for different values of n from 2 to 25 and are given in Table - 1.

(b) If the control limits are to be obtained in terms of $\bar{s}$ then an estimate of σ can be obtained from the relation

$$\bar{s}=c_2\sigma\Rightarrow\hat{\sigma}=\frac{\bar{s}}{c_2}$$

where $c_2 = \sqrt{\frac{2}{n}} \cdot \frac{\left(\frac{n-2}{2}\right)!}{\left(\frac{n-3}{2}\right)!}$

then $3\sigma -$ control limits for $\bar{X}$ - chart are

$$\hat{\mu} \pm \frac{3\hat{\sigma}}{\sqrt{n}}$$

where $\hat{\mu} = \bar{\bar{x}}, \hat{\sigma} = \frac{\bar{s}}{c_2}$

$$\therefore UCL: \bar{\bar{X}} + 3 \cdot \frac{\bar{s}}{c_2} \cdot \frac{1}{\sqrt{n}} \text{ or } \bar{\bar{X}} + A_1 \bar{s}$$

$$CL: \bar{\bar{X}}$$

$$LCL: \bar{\bar{X}} - 3 \cdot \frac{\bar{s}}{c_2} \cdot \frac{1}{\sqrt{n}} \text{ or } \bar{\bar{X}} - A_1 \bar{s}$$

The factor $A_1 = \frac{3}{\sqrt{n}\, c_2}$ has been tabulated for different values of n from 2 to 25.

The values of $\bar{x}_1, \bar{x}_2, \cdots\cdots, \bar{x}_k$ are plotted against the subgroup number $1, 2, \cdots\cdots, K$ on the graphs after drawing the three horizontal lines (UCL, CL, LCL). If one or more points fall outside the control limits, it will indicate the presence of assignable causes, and we conclude that the process is out of control. Otherwise the process is said to be in control.

CONTROL CHART FOR RANGE OR R-CHART :

Definition : When Sample Ranges (R) are plotted against subgroup number on a control chart, then it is called R-Chart or Range chart.

Collection of data and Computation of Statistic : The data is collected in the same form as for $\bar{X} -$ chart and the range (R) of each subgroup is computed.

Statistical Basis : Let μ be the process average and σ be the S.D. It is supposed that range has a normal distribution with mean $\bar{R}$ and standard deviation σ. Then 3σ limits are given by $E(R) \pm 3\sigma_R$. $E(R)$ is estimated by $\bar{R}$ and σ is estimated as $\frac{\bar{R}}{d_2}$.

Control Limits :

Case 1 : When standards are given

when σ is known, then $\bar{R} = d_2\sigma$ and $\sigma_R = d_3\sigma$, where d_2, d_3 are constants depending on n, the subgroup size. The control limits are

$\text{UCL}: d_2\sigma + 3d_3\sigma$ or $(d_2 + 3d_3)\sigma$ or $D_2\sigma$

$\text{CL}: d_2\sigma$

$\text{LCL}: d_2\sigma - 3d_3\sigma$ or $(d_2 - 3d_3)\sigma$ or $D_1\sigma$

Case 2 : When standards are not known.

When σ is not known then it is estimated with the help of observed data. Here

$$\hat{\sigma} = \frac{\bar{R}}{d_2} = \frac{1}{d_2}\left[\frac{1}{K}\sum_{i=1}^{K} R_i\right], \text{ where } E(\bar{R}) = d_2\sigma$$

thus the control limits are

$$\text{UCL}: d_2 \cdot \frac{\bar{R}}{d_2} + 3 \cdot d_3 \frac{\bar{R}}{d_2} \text{ or } \left(1 + \frac{3d_3}{d_2}\right)\bar{R} = D_4\bar{R}$$

$$\text{CL}: d_2 \frac{\bar{R}}{d_2} = \bar{R}$$

$$\text{LCL}: d_2 \cdot \frac{\bar{R}}{d_2} - 3d_3 \frac{\bar{R}}{d_3} = \left(1 - \frac{3d_3}{d_2}\right)\bar{R} = D_3\bar{R}$$

where D_1, D_2, D_3, D_4 are tabulated for different values of n from 2 to 25 in the following table-1.

The values of $R_1, R_2, \cdots, R_K$ are plotted against the subgroup number $1, 2, \cdots, K$ on the graph after drawing three horizontal lines (U.C.L., C.L., L.C.L.).

If one or more points fall outside the control limits, It will indicate the presence of assignable causes and we conclude that the process is out of control.

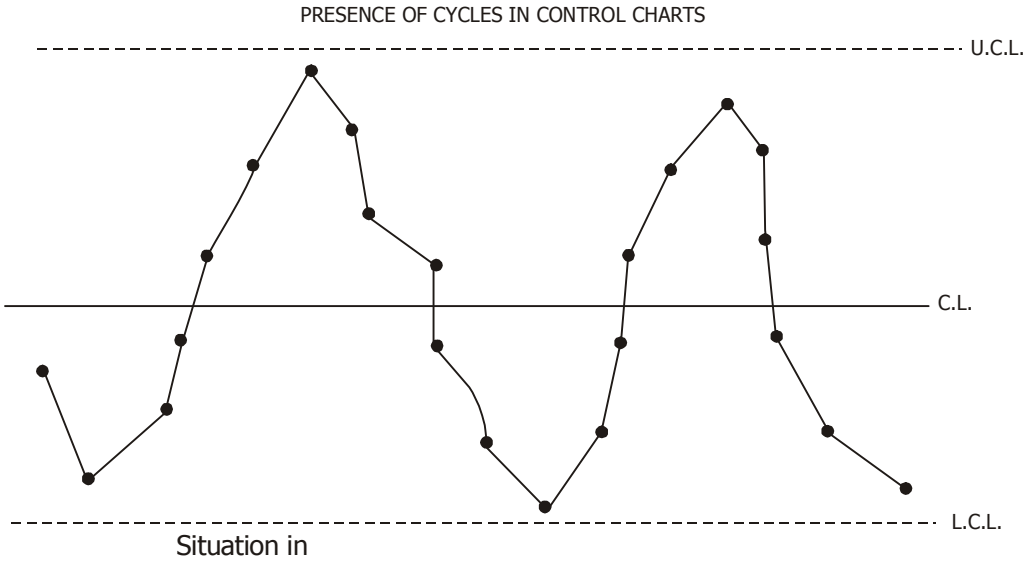
Table - I

FACTORS USEFUL IN THE CONSTRUCTION OF CONTROL CHARTS

Sample size	Mean Chart Factors for control limits			Range Chart Factors for control line	Range Chart Factors for control limits			
n	A	A_1	A_2	d_2	D_1	D_2	D_3	D_4
2	2.121	3.760	1.880	1.128	0	3.686	0	3.267
3	1.732	2.394	1.023	1.693	0	4.358	0	2.575
4	1.500	1.880	0.729	2.059	0	4.698	0	2.282
5	1.342	1.596	0.577	2.326	0	4.918	0	2.115
6	1.225	1.410	0.483	2.534	0	5.078	0	2.004
7	1.134	1.277	0.419	2.704	0.205	5.203	0.076	1.924
8	1.061	1.175	0.373	2.847	0.387	5.307	0.136	1.864
9	1.000	1.094	0.337	2.970	0.546	5.394	0.184	1.816
10	0.949	1.028	0.308	3.078	0.687	5.469	0.223	1.777
11	0.905	0.973	0.285	3.173	0.812	5.534	0.256	1.744
12	0.866	0.925	0.266	3.258	0.924	5.592	0.284	1.716
13	0.832	0.884	0.249	3.336	1.026	5.646	0.308	1.692
14	0.802	0.848	0.235	3.407	1.121	5.693	0.329	1.671
15	0.775	0.816	0.223	3.472	1.207	5.737	0.348	1.652
16	0.750	0.788	0.212	3.532	1.285	5.779	0.364	1.636
17	0.728	0.762	0.203	3.588	1.359	5.817	0.379	1.621
18	0.707	0.738	0.194	3.640	1.426	5.854	0.392	1.608
19	0.688	0.717	0.187	3.689	1.490	5.888	0.404	1.596
20	0.671	0.697	0.180	3.735	1.548	5.922	0.414	1.586
21	0.655	0.679	0.173	3.778	1.666	5.950	0.425	1.575
22	0.640	0.662	0.167	3.819	1.659	5.979	0.434	1.566
23	0.626	0.647	0.162	3.858	1.710	6.006	0.443	1.557
24	0.612	0.632	0.157	3.895	1.759	6.031	0.452	1.548
25	0.600	0.619	0.153	3.931	1.804	6.058	0.459	1.541

INTERPRETATION OF $\bar{X}$ AND R CHARTS

In order to judge if a process is in control, $\bar{X}$ and R charts should be examined together and the process should be deemed in statistical control if both the charts show a state of control. Situations exist where R-chart is in a state of control but $\bar{X}$- chart is not. Here we summarise below, in a tabular form, such different situations and the interpretation in each case.



S.No.	R-Chart	$\bar{X}$ – chart	Interpretation
1	In control	Points beyond limits only on one side	Level of process has shifted
2	In control	Points beyond limits on both sides	Level of process is changing in erratic manner-frequent adjustments.
3	Out of control	Out of control on both sides	Variability has increased
4	Out of control	Out of control on one side	Both level and variabilty have changed
5	In control	Run 7 or more points on one side of central line	Shift in process level
6	In control	Trend of 7 or more points. No point outside control limits	Process level is gradually changing

7	Runs of 7 or more points above central line	----------	Variability has increased
8	Points too close to the central line	----------	Systematic differences within sub-groups
9	----------	Points too close to the central line	Systematic differences within sub-groups.

8.4 WORKEDOUT EXAMPLES

Example 1 : Construct the $\bar{X}$ and R - charts for the following data given the diameter of screws in cms.

Sample No.	I	II	III	IV	V	Mean($\bar{X}$)	Range(R)
1	0.831	0.829	0.836	0.840	0.826	0.8324	0.014
2	0.834	0.826	0.831	0.831	0.831	0.8306	0.008
3	0.836	0.826	0.831	0.822	0.816	0.8262	0.020
4	0.833	0.831	0.835	0.831	0.823	0.8316	0.004
5	0.830	0.831	0.831	0.833	0.820	0.8290	0.013
6	0.829	0.828	0.828	0.832	0.841	0.8316	0.013
7	0.835	0.833	0.829	0.830	0.841	0.8290	0.012
8	0.818	0.838	0.835	0.834	0.830	0.8316	0.020
9	0.841	0.831	0.831	0.833	0.832	0.8336	0.010
10	0.832	0.828	0.836	0.832	0.825	0.8331	0.011
11	0.831	0.838	0.844	0.827	0.826	0.8336	0.018
12	0.831	0.826	0.828	0.832	0.827	0.8306	0.006
13	0.838	0.822	0.835	0.830	0.830	0.8332	0.016
14	0.815	0.832	0.831	0.831	0.838	0.8288	0.023
15	0.831	0.833	0.831	0.834	0.832	0.8334	0.003
16	0.830	0.819	0.819	0.844	0.832	0.8270	0.025
17	0.826	0.839	0.842	0.835	0.830	0.8338	0.016

18	0.813	0.833	0.819	0.834	0.836	0.8270	0.023
19	0.832	0.831	0.825	0.831	0.850	0.8338	0.025
20	0.831	0.838	0.833	0.831	0.833	0.8332	0.007
21	0.823	0.830	0.832	0.835	0.835	0.8310	0.012
22	0.835	0.829	0.834	0.826	0.828	0.8304	0.009
23	0.933	0.836	0.831	0.832	0.832	0.8328	0.005
24	0.826	0.835	0.842	0.832	0.831	0.8332	0.016
25	0.833	0.823	0.816	0.831	0.838	0.8282	0.022
26	0.829	0.830	0.830	0.833	0.831	0.8306	0.004
27	0.850	0.834	0.827	0.831	0.835	0.8354	0.023
28	0.835	0.846	0.829	0.833	0.822	0.8330	0.024
29	0.831	0.832	0.834	0.826	0.833	0.8312	0.008
	Total					**24.111**	**0.406**

$$\bar{\bar{x}} = \frac{1}{K}\sum_{i=1}^{K}\bar{x}_i = \frac{24.111}{29} = 0.8314$$

$$\bar{R} = \frac{1}{K}\sum_{i=1}^{k} R_i = \frac{0.406}{29} = 0.014 \qquad (\because K = 29)$$

$\bar{X}$ – Charts

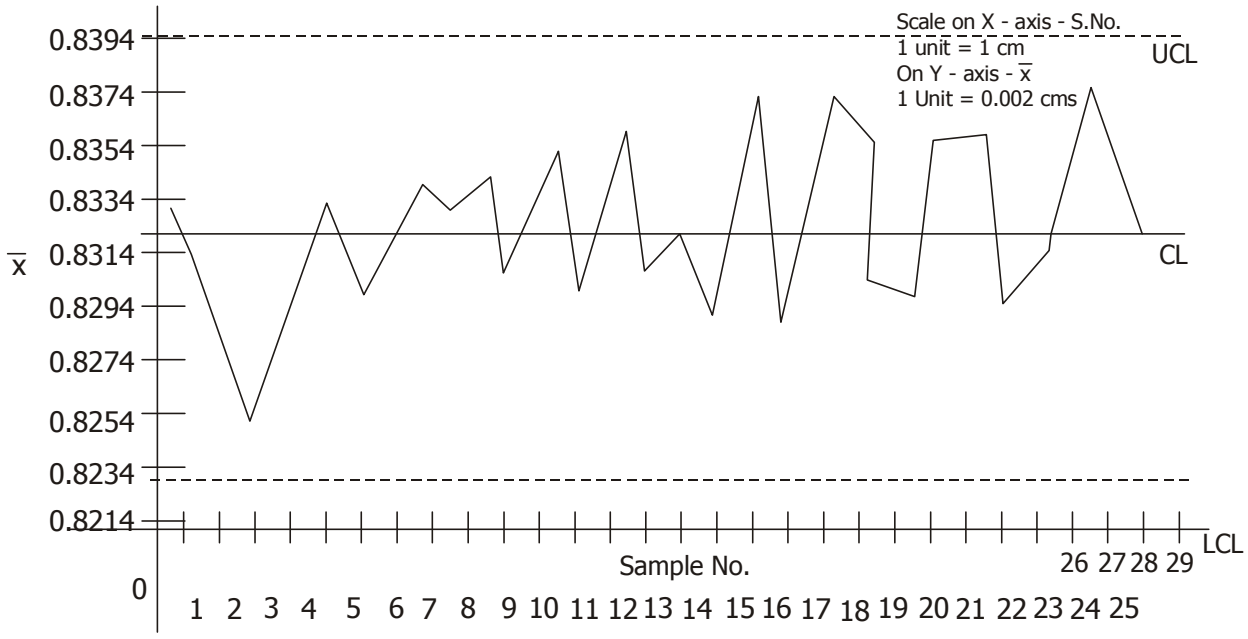
$$UCL: \bar{X} + A_2\bar{R}$$

$$= 0.8314 + 0.577 \times 0.014 = 0.8314 + 0.0081 = 0.8395 \quad (\because A_2 = 0.577 \text{ are taken from Table I})$$

Central line = $\bar{X} = 0.8314$

$$LCL: \bar{X} - A_2\bar{R}$$

$$= 0.8314 - 0.577 \times 0.014 = 0.8314 - 0.0081 = 0.8233$$



($\bar{x}$ Chart)

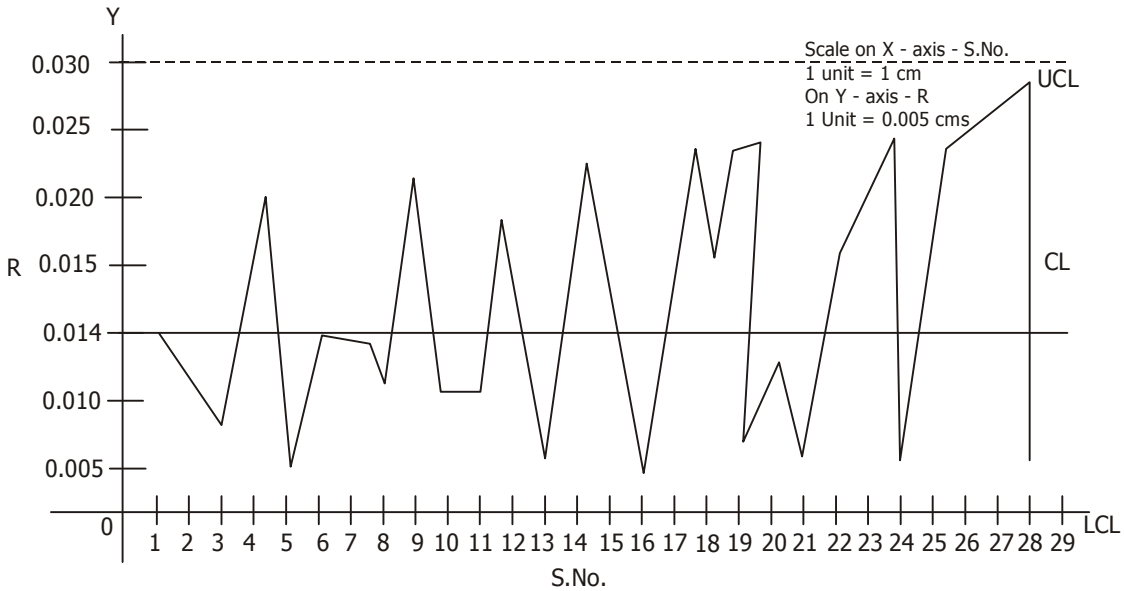
R-Chart :

$$UCL = D_4\bar{R} = 2.115 \times 0.014 = 0.296$$

Central line $CL = \bar{R} = 0.014$

$$LCL = D_3\bar{R} = 0 \times 0.014 = 0$$

$(\because D_4 = 0.2115,\ D_3 = 0$ are taken from Table I.$)$



For $\bar{X}-$ chart and R- chart there is no lack of control in process as no sample point lies outside the control limits.

Example 2 : The following are the mean and ranges of 20 samples of size 5 each. The data pertain to the overall length of a fragmentation bomb base manufactured during the war by the American store camp.

Group No	Mean	Range	Group No	Mean	Range
1	0.8372	0.010	11	0.8380	0.006
2	0.8324	0.009	12	0.8322	0.003
3	0.8318	0.008	13	0.8356	0.013
4	0.8344	0.004	14	0.8322	0.005
5	0.8346	0.005	15	0.8404	0.008
6	0.8332	0.011	16	0.8372	0.011
7	0.8340	0.009	17	0.8282	0.006
8	0.8344	0.003	18	0.8346	0.006
9	0.8308	0.002	19	0.8360	0.004
10	0.8350	0.006	20	0.8374	0.006

(a) From these data, obtain the control limits for $\bar{X}$ and R charts to control the length of bomb bases to be produced in the future.

(b) the above samples were taken every 15 minutes during production. The production rate was 350 units/hr and the tolerances were 0.820 and 0.840 inches.

On the assumption that the lengths of the bomb bases are normally distributed, what percentage of the bomb base would you estimate to have length outside the tolerance limits when the process is under control at the levels indicated by the above data?

Solution :

(a) $$\bar{\bar{X}} = \frac{1}{20}\sum_{i=1}^{20}\bar{X}_i = \frac{16.6796}{20} = 0.83398$$

$$\bar{R} = \frac{1}{20}\sum_{i=1}^{20}R_i = \frac{0.133}{20} = 0.00665$$

From Table - I for $n = 5, A_2 = 0.58, D_3 = 0$ and $D_4 = 2.12$

$$UCL = \bar{\bar{X}} + A_2\bar{R} = 0.83398 + 0.58 \times 0.00665 = 0.837837$$

$$CL = 0.83398$$

$$\text{LCL} = \bar{\bar{X}} - A_2\bar{R} = 0.83398 - 0.58 \times 0.00665 = 0.830123$$

We see that values of $\bar{X}$ corresponding to the sub-groups 11 and 17 are outside the control limits. So excluding these values, we get

$$\bar{\bar{X}}' = \frac{15.0134}{18} = 0.834077, \ \bar{R}' = \frac{0.121}{18} = 0.00672$$

So, the new control limits of $\bar{X}$ chart are

$$\text{UCL}: \bar{\bar{X}}' + A_2\bar{R}' = 0.834077 + 0.58 \times 0.00672 = 0.837975$$

$$\text{CL}: \bar{\bar{X}}' = 0.834077$$

$$\text{LCL}: \bar{\bar{X}}' - A_2\bar{R}' = 0.834077 - 0.58 \times 0.00672 = 0.831179$$

which may be regarded as the final limits of $\bar{X}$, since no value except the rejected ones is outside limits.

$$\text{UCL}_R = D_4\bar{R} = 2.12 \times 0.00665 = 0.014088$$

$$\text{CL}_R = 0.00665$$

$$\text{LCL}_R = 0$$

From the given values of R we see that the values corresponding to all the sub-groups fall inside the control limits, which may be taken as the final control limits.

(b) We are given that

Upper Tolerance Limit $(\text{U.T.L.}) = 0.840''$

Lower Tolerance Limit $(\text{L.T.L}) = 0.820''$

Assuming that the process is in control, the random variable $\bar{X}$ (the length of the bomb bases) is normally distributed with mean and s.d. given by

$$\hat{\mu} = \bar{\bar{X}}' = 0.834077 \text{ and } \hat{\sigma} = \frac{\bar{R}'}{d_2} = \frac{0.00672}{2.326} = 0.002889$$

Hence the proportion 'p' of defectives when the process is in control given by

$$P = P\{X \text{ lies outside tolerance limits}\}$$

$$= 1 - P\{0.82 < X < 0.84\}$$

$$= 1 - P\{-4.7982 < Z < 4.999\}, \text{ when } Z = \frac{X - \hat{\mu}}{\hat{\sigma}}$$

$= 0.0201825$ (from Biometrika Tables)

Hence the percent fraction defective when the process is in control is

$$100 \times 0.0201825 = 2.01825 \cdot$$

Example 3 :

(a) Show that P_n, the probability of the mean of a random sample of size n exceeding $UCL = \mu' + \frac{3\sigma'}{\sqrt{n}}$ when the population mean has shifted to $\mu' + K\sigma'$ is $G\left(3 - K\sqrt{n}\right)$, where

$$G(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-\frac{1}{2}u^2} du$$

(b) If the r^{th} sample mean is the first to exceed UCL, show that $E(r) = \frac{1}{P_n}$

Solution :

(a) If $\bar{X}$ is the mean of a sample of size n from a population with mean μ' and standard deviation σ', then $\bar{X}$ is normally distributed if population is normal and $\bar{X}$ is asymptotically normally distributed (by central limit theorem) with mean μ' and S.D. $\frac{\sigma'}{\sqrt{n}}$. After the shift of the mean to $\mu' + K\sigma'$ has taken place $\bar{X} \sim N\left(\mu' + K\sigma', \frac{\sigma'^2}{n}\right)$

$$p(\bar{x}) = \frac{\sqrt{n}}{\sqrt{2\pi}\,\sigma'} \exp\left[\frac{-n}{2\sigma'^2}(\bar{x} - \mu' - K\sigma')^2\right], \quad -\infty < \bar{x}' < \infty$$

$$p_n = p\{\bar{x} > UCL\} = \int_{UCL}^{\infty} p(\bar{x})\, d\bar{x}$$

$$= \frac{\sqrt{n}}{\sqrt{2\pi}\sigma'} \int_{\mu' + \frac{3\sigma'}{\sqrt{n}}}^{\infty} \exp\left[-\frac{n}{2\sigma'^2}(\bar{x} - \mu' - k\sigma')^2\right] d\bar{x}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\left(3 - K\sqrt{n}\right)}^{\infty} \exp\left(-\frac{1}{2}t^2\right) dt, \quad \left(\because t = \frac{\bar{x} - \mu' - K\sigma'}{\frac{\sigma'}{\sqrt{n}}}\right)$$

$= G\left(3 - K\sqrt{n}\right)$ [By definition of $G(\cdot)$]

(b) If $\bar{X}_i$ is the mean of the i^{th} sample of size n then

$$p_n = p\left\{\bar{X}_i > UCL\right\}$$

$$q_n = p\left\{\bar{X}_i \le UCL\right\} = 1 - P_n$$

If r^{th} sample mean is the first to exceed the UCL, the preceeding $(r-1)$ sample means must be $\le UCL$. Thus if Y is the random variable such that $Y = r(1, 2, \cdots)$ implies that the r^{th} sample mean is the first to exceed UCL then Y follows the geometric distribution with probability function

$$f(r) = p(Y = r) = q_n^{r-1} \cdot p_n,\ r = 1, 2, \cdots$$

$$\therefore E(Y) = \sum_{r=1}^{\infty} r f(r) = \sum r\, q_n^{r-1} \cdot p_n = p_n\left[1 + 2q_n + 3q_n^2 + \cdots\right]$$

Let $S = 1 + 2q_n + 3q_n^2 + 4q_n^3 + \cdots$

$$q_n \cdot S = q_n + 2q_n^2 + 3q_n^3 + \cdots$$

$$\Rightarrow S = \frac{1}{(1-q_n)^2} = \frac{1}{(1-q_n)}$$

$$\therefore E(Y) = \frac{p_n}{(1-q_n)^2} = \frac{1}{p_n}$$

Example 4 : Show that the probability that at least one of the two points $\bar{X}$ and R goes outside the control limits is

$$1 - \left[G\left(\sqrt{n}\,T - 3\rho\right) - G\left(\sqrt{n}T + 3\rho\right)\right]\left[P\left(\frac{R}{\sigma} \le D_2\rho\right) - P\left(\frac{R}{\sigma} \le D_1\rho\right)\right]$$

where $\rho = \frac{\sigma'}{\sigma}$, $T = \frac{(\mu' - \mu)}{\sigma}$ and $G(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-\frac{1}{2}t^2}\, dt$

assuming that the control charts are based on μ' as population mean and σ' as population standard deviation, where the actual values of these parameters are μ and σ respectively.

Solution : Probability (p_1) that all the points on the $R-$ charts lie within control limits is given by

$$p_1 = P\left[D_1\sigma' \le R \le D_2\sigma'\right] = P\left(R \le D_2\sigma'\right) - P\left(R \le D_1\sigma'\right)$$

$$p = P\left(\frac{R}{\sigma} \le D_2\frac{\sigma'}{\sigma}\right) - P\left(\frac{R}{\sigma} \le D_1\frac{\sigma'}{\sigma}\right),\ \sigma > 0$$

$$= P\left(\frac{R}{\sigma} \le D_2\rho\right) - P\left(\frac{R}{\sigma} \le D_1\rho\right) \text{---------- (1)}$$

Probability (p_2) that all the points on $\bar{X}$-chart lie within the control limits is

$$P_2 = P\left[LCL_{\bar{X}} \le \bar{X} \le UCL_{\bar{X}}\right] = P\left[\mu' - 3\frac{\sigma'}{\sqrt{n}} \le \bar{X} \le \mu' + \frac{3\sigma'}{\sqrt{n}}\right]$$

where $\bar{X}$ is $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. This distribution is exactly normal if population is normal and is asymptotically normal otherwise. Therefore

$$p_2 = P\left[\frac{\mu' - \frac{3\sigma'}{\sqrt{n}} - \mu}{\frac{\sigma}{\sqrt{n}}} \le Z \le \frac{\mu' + \frac{3\sigma'}{\sqrt{n}} - \mu}{\frac{\sigma}{\sqrt{n}}}\right]$$

where $Z = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

$$\Rightarrow p_2 = P\left[\frac{\sqrt{n}(\mu' - \mu)}{\sigma} - \frac{3\sigma'}{\sigma} \le Z \le \frac{\sqrt{n}(\mu' - \mu)}{\sigma} + \frac{3\sigma'}{\sigma}\right]$$

$$= P\left[\sqrt{n}T - 3\rho \le Z \le \sqrt{n}T + 3\rho\right],\ \text{where } T = \frac{\mu' - \mu}{\sigma},\ \rho = \frac{\sigma'}{\sigma}$$

$$= \int_{\sqrt{n}T-3\rho}^{\sqrt{n}T+3\rho} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz$$

$$= \int_{\sqrt{n}T-3\rho}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz - \int_{\sqrt{n}T+3\rho}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz$$

$$= G\left(\sqrt{n}T - 3\rho\right) - G\left(\sqrt{n}T + 3\rho\right) \text{ ---------- (2)}$$

($\because$ by definition of the function $G(\cdot)$)

Assuming normality of the distribution so that $\bar{X}$ and R are independently distributed, the probability (p) that none of the two points $\bar{X}$ and R goes outside the control limits, is given by the compound probability theorem as

$$p = p_1 \cdot p_2$$

$$= \left[p\left(\frac{R}{\sigma} \le D_2\rho\right) - p\left(\frac{R}{\sigma} \le D_1\rho\right)\right]\left[G\left(\sqrt{n}\,T - 3\rho\right) - G\left(\sqrt{n}T + 3\rho\right)\right]$$

$\therefore p$ [atleast one of the two points $\bar{X}$ and R goes outside the control limits] = $1 - p$ as desired.

Example 5 : p_n is the probability of the mean of a sample of size n falling outside the control limits. Show that

(a) The probability that atmost x samples are to be taken for r points to go out of control is

$$1 - \sum_{s=0}^{r=1}\left[\binom{x}{s} p_n^s (1-p_n)^{x-s}\right]$$

(b) The probability that exactly x samples are to be taken for r points to go out of control is

$$\left(\frac{p_n}{1-p_n}\right)^r \cdot \binom{x-1}{r-1}(1-p_n)^x,\ x \ge r$$

Solution : Let Y be the random variable which represents the number of points (sample means) falling outside the control limits in x samples. Then Y is a binomial variate with parameters (x, p_n) and probability function

$$g(s) = p(Y = s) = \binom{x}{s} p_n^s (1-p_n)^{x-s},\ (s = 0,1,2,\cdots\cdots,x)$$

Hence the probability 'p' that in x samples the number of points going out of the control limits greater than or equal to r is given by

$$p = p(Y \ge r) = 1 - p(Y < r)$$

$$= 1 - \sum_{s=0}^{r-1}\binom{x}{s} p_n^s (1-p_n)^{n-s} \text{ ---------- (1)}$$

Now 'p' implies that the r^{th} point goes outside the control limits either at or before the x^{th} sample. In otherwords, for the r^{th} point to go outside the control limit the sample size required is $\leq x$. Hence the R.H.S. of (1) gives the required probability that at most x samples are to be taken for at least r points to go out of control.

(b) The event 'E' that exactly x samples are required for r points to go out of control limits, happens if the r^{th} point goes out of the control limits at the x^{th} sample and the probabilty of this is p_n, and also in the remaining $(x-1)$ samples exactly $(r-1)$ points go out of the control limits, the probability of which is

$$\binom{x-1}{r-1} p_n^{r-1} (1-p_n)^{x-r}$$

Hence by the compound probability theorem, the required probability is given by

$$P(E) = p_n \binom{x-1}{r-1} p_n^{r-1} (1-p_n)^{x-r}$$

$$= \left(\frac{p_n}{1-p_n}\right)^r \cdot \binom{x-1}{r-1} (1-p_n)^x ; \; x \geq r$$

8.5 EXERCISE

1. Give clearly the meaning of each word of the term 'Statistical Quality Control'. Distinguish between 'process' and product control. Does process control also ensure product control necessarily?
2. What do you understand by statistical quality control ? Discuss briefly its need and utility in industry. Discuss the causes of variation in quality.
3. What is meant by process control in industrial statistics ?
4. What is control chart ? Explain the basic principles underlying the control charts. Discuss the role of control charts in manufacturing processes.
5. Explain the justification for using the three sigma limits in the control charts irrespective of the actual probability distribution of the quality characteristic.
6. Explain clearly the basis and working of control charts for mean and range. State the basis and assumptions on which $\bar{X}$ and R charts are developed.
7. Explain how to estimate σ from the mean range of samples of constant size drawn during a continuous production process. What are the other methods of estimating σ ?
8. In order to determine whether or not a process producing bronze castings is in control, 20 subgroups of size 6 are taken. The quality characteristic of interest is the weight of the castings and it is found that $\bar{\bar{X}}$ is 3.126 gm and $\bar{R} = 0.009 gm$.

(i) Estimate the standard deviation of the weight of castings.

(ii) Assuming that the process is in control, find upper and lower control limits for the sub-group means, sub-group ranges.

(iii) Using (i) above, within what limits would you expect 99.73 percent of all individual measurements to fall ?

9) Prepare an $\bar{X} - R$ chart using the following results obtained from sample of size 5 each :

Sample Number :	1	2	3	4	5
Average :	2.5	2.6	2.7	2.7	2.4
Range :	0.2	0.2	0.3	0.4	0.3

10) 20 random samples of 5 units drawn from each lot of wire gave the mean diameters and range as given below. Draw the $\bar{X} - R$ charts and write a comments on your findings.

S.No.	$\bar{X}$	R	S.No.	$\bar{X}$	R
1	61.2	10	11	60.8	13
2	60.8	9	12	65.2	10
3	61.8	5	13	67.8	10
4	62.2	9	14	66.0	9
5	59.2	11	15	64.0	6
6	62.0	14	16	64.8	8
7	61.6	12	17	69.6	13
8	66.0	6	18	59.4	11
9	62.0	15	19	61.0	15
10	66.0	16	20	61.8	13

LESSON - 9

CONTROL CHARTS FOR ATTRIBUTES

OBJECT OF THE LESSON

* After studying this lesson the student is expected to have a clear comprehension of the theory and the practical utility about the concepts of control charts for attributes and their applications.

STRUCTURE OF THE LESSON : This consists of sections as detailed below :

9.1 INTRODUCTION

The variable control charts ($\bar{X}$, R charts) are excellent means for controlling quality, but they do have limitations. One obvious limitation is that they cannot be used for quality characteristics which are attributes. Another limitation concerns the fact that there are many variables in a manufacturing unit. Even a small manufacturing unit could have as many as 10,000 quality characteristics. As one chart is needed for each characteristic, 10,000 such charts would be required. Clearly, this would be too expensive and quite impractical. As against this, control chart for attributes can provide overall quality information at a fraction of the cost and minimize this limitation.

9.2 CONTROL CHART FOR ATTRIBUTES

As an alternative to $\bar{X}$ and R charts, we have the control charts for attributes which can be used for quality characteristics.

(i) Which can be observed only as attributes by classifying an item as defective or non-defective i.e., confirming to specifications or not and

(ii) Which are actually observed as attributes even though they could be measured as variables. ex :, go and no-go gauge test results.

There are two control charts for attributes :

(a) Control chart for fraction defective (p-chart) or the number of defectives (np or d chart)

(b) Control chart for the number of defects per unit (c chart).

A control chart in the hand of statistical quality control engineers is called the engineer's stethoscope. The view point is also much the same as that of a doctor for a patient. He does not look up the patient's temperatures and pulse-rate in a book of specifications but instead measures them to find out how the patient is "running". Similarly, the control chart is the engineer's "stethoscope" for the process, whereby he learns of its conditions.

p-CHART :

Definiton : When we are merely concerned whether a product is "defective" or not i.e., when the quality characteristic is taken as an attribute then the chart employed is called p-chart, i.e., the chart for fraction defective. In fact, here p is the fraction defective. It stands for the fraction of the total number of products in a sample which do not meet the specified requirement and are taken as defectives. If d is the number of defectives found in a sample of size n, then $p = \frac{d}{n}$.

STATISTICAL BASIS :

We should agree with it that the population of the production in a manufacturing process from which samples are taken can safely be taken as an infinite one, so it can be assumed that $d = np$ is distributed binomially with $E(d) = E(np)$ and $V(d) = nPQ = V(np)$ where P is the population fraction defective and $Q = (1 - P)$, so $V(p) = \frac{nPQ}{n^2} = \frac{PQ}{n}$.

CONTROL LIMITS :

Case (i) : Standards are given

Suppose the population fraction defective i.e., the values of p is given as p'. Then 3σ limits for p - chart would be

$$UCL = p' + 3\sqrt{\frac{p'(1-p')}{n}} \quad \text{or} \quad p' + \frac{3}{\sqrt{n}}\sqrt{p'(1-p')}$$

$$\text{or} \quad p' + A\sqrt{p'(1-p')}$$

Central line $CL = p'$

$$LCL = p' - 3\sqrt{\frac{p'(1-p')}{n}} \quad \text{or} \quad p' - \frac{3}{\sqrt{n}}\sqrt{p'(1-p')}$$

$$\text{or} \quad p' - A \cdot \sqrt{p'(1-p')}$$

Case (ii) : When standards are not given

When the population parameter P is unknown, it is estimated by the statistic $\bar{p}$ with the help of the observed data.

Control Chart For Number Of Defectives (d-chart) np-Chart.

If instead of p, the sample proportion defective, we use d, the number of defectives in the sample, then the $3\sigma -$ control limits for d-chart are given by

$$E(d) \pm 3 \cdot S.E(d)$$

$$= np \pm 3\sqrt{np(1-p)}$$

where $E(d) = np \qquad var(d) = nPQ = nP(1-P)$

Case i : Standards are given

If P' is the given value of P then

$$UCL_d = nP' + 3\sqrt{nP' + (1-P')}$$

$$LCL_d = nP' - 3\sqrt{nP'(1-P')}$$

$$CL_d = nP'$$

Case ii : Standards are not given

using $\bar{P}$ as an estimate of P as $\bar{P} = \frac{\Sigma d_i}{\Sigma n_i}$ we get

$$UCL_d = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}$$

$$LCL_d = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}$$

$$CL_d = n\bar{p}$$

Since p cannot be negative, if LCL as given by above formulae comes out to be negative then it is taken to be zero.

(i) p and d charts for Fixed Sample Size :

If the sample size remains constant for each sample i.e., if $n_1 = n_2 = \cdots\cdots = n_K = n,$ then using an estimate of the population proportion p is given by

$$\hat{p} = \bar{p} = \frac{\sum_{i=1}^{K} np_i}{\sum_{i} n} = \frac{n\sum_{i=1}^{K} p_i}{nK} = \frac{1}{K}\sum_{i=1}^{K} p_i$$

$$\bar{p} = \sum_{i=1}^{K} d_i \Big/ \sum_{i=1}^{K} n_i = \frac{\text{total number of defectives of the samples}}{\text{total number of pieces inspected}}$$

$\bar{p}$ is an estimate of p, if the process is in control. But we are not sure that the process may be in control. So we make the hypothesis, that the process is in control and apply the test for examining this hypothesis. Thus, $\bar{p}$ as defined above, may be taken as an unbiased estimate of the unknown parameter p.

Thus, for p - chart

$$UCL = \bar{p} + 3\sqrt{\bar{p}(1-\bar{p})/n}$$

$$CL = \bar{p}$$

$$LCL = \bar{p} - 3\sqrt{\bar{p}(1-\bar{p})/n}$$

PROCEDURE :

(i) Preliminary data for the construction of the limits may be obtained from the past records. Some 20 to 25 samples may be sufficient to examine the quality and to evaluate or to estimate the value of the standard fraction defective for future limits.

(ii) Limits are calculated and plotted on the graph.

(iii) Then individual fraction defectives for each sample is calculated and plotted.

(iv) If a point falls outside the limits it would indicate deterioration in quality.

(v) If the lower control limit comes out to be negative it is taken as zero.

In this case, the same set of control limits can be used for all the samples inspected and it is immaterial if one uses p-chart or d-chart.

(ii) p and d charts for variable sample size :

Method I : If the number of items inspected (n) in each sample varies, for p - chart separate control limits have to be computed for each sample while the centrol line is invariant whereas for d-chart control limits as well as the central line have to be computed for each sample. This type of limits are known as variable control limits. In such a situation p-chart is relatively simple and is preferred to d -chart.

Method II : If n varies separate control limits are calculated for each sample. Since $S.E(p) = \sqrt{\frac{PQ}{n}}$, it should be noted that smaller the sample size wider the control band and vice versa. If the sample size does not vary appreciably then a single set of control limits based on the average sample size $\left(\sum_{i=1}^{K} \frac{n_i}{k}\right)$ can be used. For practical purposes, this holds good for situations in which the largest sample size does not exceed the smallest sample size by more than 20% of the smallest sample size.

Alternatively, for all sample sizes two sets of limits, one based on the largest sample size and the other based on the smallest sample size can be used. The largest sample size gives the smallest control band. Which is called inner band and the smallest sample gives the largest control band which is called outer band. Points falling within the inner band indicate the process in control while points lying outside the outer band are indicative of the presence of assignable causes of variation which must be searched and verified. For other points, action should be based on the exact control limits.

Method III : Another procedure is to standardise the variate, i.e, instead of plotting p or d on the control chart we plot the corresponding standardised values

$$Z = \frac{p - \bar{p}}{\sqrt{\frac{p'Q'}{n}}} \text{ or } \frac{p - \bar{p}}{\sqrt{\bar{p}(1-\bar{p})/n}}$$

According as p is given or not, the symbols have their usual meaning. This stabilises our variables and the resulting chart is called stabilised p-chart or d-chart. In this case the control limits as well as the central line for p and d charts are invariant with n being given by

$$UCL = 3, \quad CL = 0, \quad LCL = -3$$

Conclusions from p - chart :

1. From the p - chart a process is judged to be in statistical control in the same way as done for $\bar{X}$ and R charts. If all the sample points fall within the control limits without exhibiting any specific pattern, the process is said to be in control. In such a case, the observed variations in the fraction defective are attributed to the stable pattern of chance causes and the average fraction defective $\bar{p}$ is taken as the standard fraction defective p.
2. Points outside the UCL are termed as high spots. These suggest deterioration in the quality and should be regularly reported to the production engineers. The reasons for such deterioration could possibly be known and removed if the details of conditions under which data were collected were known and of particular interest and importance is, if there was any change of inspection or inspection standards.
3. Points below LCL are called low spots. Such points represent a situation showing improvement in the product quality. However, before taking this improvement for granted, it should be investigated if there was any slackness in inspection or not.

4. When a number of points falls outside control limits, a revised estimate of P should be obtained by eliminating all the points that fall above UCL. The standard fraction defective p should be revised periodically in this way.

Similarly the conclusions for the control chart for number of defects (d-chart) are the same as that for p-chart.

CONTROL CHART FOR THE NUMBER OF DEFECTS (c - CHART) :

There are many situations in industry where after classifying an item or product as defective one, it is further examined for the number of defects contained in it. For example, a defective photography film may further be examined for the number of surface defects in it, or the number of blemishes may be counted on a defective 100 square yard cloth piece etc. Thus every defective unit contain one or more of the defects. In fact, a defective item is that which does not confirm to one or more of the specifications where as a defect is an instance of the articles deficiency towards specifications. The number of defects is denoted by c.

Definition : A control chart in which the points plotted are the number of defects (c) per each unit is called a c - chart.

Statistical Basis : In a manufacturing process there are numerous opportunities for defects to occur but practically the occurence of defects is quite casual. So the probability (p) for a defect to occur in any one spot of the product is negligible as compared with n which is the area of opportunity here. So for this the experience with variation in the number of defects (c) per unit indicates that the distribution of the variable follows very closely the form of the poisson distribution.

The arithmetic mean and the variance of the poisson variate is the same. So, if the variable that is number of defects per unit 'c' follows the poisson distribution with parameter λ, then $E(c) = \lambda$ and $V(c) = \lambda$.

Collection of data : The number of defects for the c-chart are usually counted in either fixed time, length, area, a single unit or for group of units.

CONTROL LIMITS :

Case 1 : When the standards are known. If the standard value of λ is c', then $3\sigma -$ limits of c - chart would be

$$UCL = c' + 3\sqrt{c'}$$

$$CL = c'$$

$$LCL = c' - 3\sqrt{c'}$$

Case 2 : When the standards are not known. i.e., when the parameter λ is not given, then it is estimated as

$\bar{c} = \sum_{i=1}^{K} \frac{c_i}{K}$ = number of defects per unit.

Here c_i is the number of defects on the i^{th} piece of the sample.

Thus $UCL = \bar{c} + 3\sqrt{\bar{c}}$

$CL = \bar{c}$

$LCL = \bar{c} - 3\sqrt{\bar{c}}$

Application of c - chart

Since c - chart depends on the assumption of poisson distribution, it is used in situations where the chance of occurring a defect at any point is small. The main applications of c-chart are

(i) It is used when a count of defects per item is considered and they are to be eliminated following the 100% inspection. For example, in an aircraft factory, may be the number of defects or of missing rivets in an air wing in the final inspection.

(ii) It can also be used to periodic samples where it is not necessary to remove all the defects. But in order to get a good quality the number is kept low. For example, for controlling the number of defects in a belt of cloth.

(iii) It is also applied to sampling acceptance procedures based on defects per unit.

(iv) Besides, c-chart is also used in the non-manufacturing process like

(a) In accident statistics, where c is the number of accidents at any place at a given time.

(b) In a chemical laboratory, where c may be the number of surface defects in coating or electroplating a sheet of metal.

(c) In studying the effect of an epidemic disease, where c may be the number of deaths per day caused by the disease in a particular region.

Note :

1. c - chart is not necessarily used for a single type of defect but may also be used for all types of defects taken together.
2. If LCL of c-chart comes out to be negative it is considered as zero.

9.3 CHARTS FOR VARIABLES VERSUS THE CHARTS FOR ATTRIBUTES

	Variables		Attributes
1.	Separate charts are constructed for each quality characteristic	1.	A single chart is sufficient for several characteristics.
2.	Sample size may be small, even as small as 4 or 5.	2.	Sample size should be large.
3.	Subgroup sizes should be equal in general.	3.	Even if the subgroup size is not constant from time to time, the charts may be used with varying limits.
4.	The data is collected purposively for the use for charts.	4.	Data collected even for other purpose, may be used.

5.	Cost of collecting and inspecting the data is high.	5.	The cost is not that much.
6.	More care is needed in computations.	6.	Less care is needed.
7.	They are more sensitive in revealing the presence of assignable causes.	7.	They are less sensitive.
8.	They improve the quality more effectively.	8.	They are less effective in improving the quality.
9.	A variable chart can rarely be used for an attribute chart.	9.	An attribute chart can be used in place of a variable chart.
10.	The $\bar{X}$ and R chart do not behave in the manner as corresponding p-chart does	10.	A p-chart is comparable. To summarise the general quality of the output and to provide a record of the quality history.
11.	The sampling acceptance plans will be cheaper in general.	11.	The sampling acceptance plan based on attribute charts will be costlier.

9.4 USES OF DIFFERENT CONTROL CHARTS

The control charts for $\bar{X}$ and R are the extensions of the charts for p and np. Together with the chart for c, they make the total base for the chart procedure meant for controlling the quality.

The $\bar{X}$ and R control charts are charts for variables i.e., for the characteristics that can be measured and expressed in numbers and not to say of those characteristics which are observed as attributes only. Furthermore, with the existing techniques, an $\bar{X}$ and R chart can be used only for one measurable characteristic at a time. For example, if a data consists of 10000 measurable characteristics, each characteristic will need separate $\bar{X}$ and R charts. However, it would be impossible to have 10000 charts, and so only the most important and troublesome characteristics would be plotted. As an alternative to $\bar{X}$ and R charts control charts based on the fraction defective (i.e., p - chart) are used. They are applied to the characteristics, that are actually observed as attributes, even though they might have been measured as variables. The cost of obtaining attribute data is usually less than that for obtaining variable data. The cost of computing and charting them may also be less. Moreover one p - chart can be applied to any number of characteristics simultaneously.

Basically, the p - chart has the same objective as an $\bar{X}$ and R chart. It also discloses the presence of assignable causes of variation. However p - chart is not so sensitive as that of $\bar{X}$ and R charts.

The p-chart can be used whenever data are or can be expressed as percentage counts. If the production consists of separate pieces, each of which can be classified as good or bad, even then the p or np-chart can be used.

The chart for c is used for counts per unit such as the number of chips in a ceramic product, the number of scratches on the table - ware, the number of crabs in aircraft etc. Its special field of usefulness is in cases of studying per article and they are the basis of quality measurement. It is necessary that the unit of inspection be kept uniform. This unit may be one coffee pot, one case of

shoes (12 pairs) or whichever is convenient. When p is small, np and c charts are practically identical and some what a simpler formulae for the c - chart can be used, provided n the sample size remains practically constant. There is no practical method for the adjustment of c or np - charts in varying sample size, since such adjustments would change the position of the central line as well as of the control limits. The p-chart is readily adjusted for varying sample size also since here the adjustments affect the control limits only.

Besides their usual applications, the charts for $\bar{X}$ and R are also applicable in a sugar mill to control the amount of lime used in purifiers or for the losses in molasses. The weights of paper and the breaking strength of yarn, the diameters of wires and filaments, the weight of coal per unit of power and the specific gravity of sulphuric acid can all be adopted to the $\bar{X}$ charts. There are conditions, in which no range chart is needed. For the most part such instances will be those in which $\bar{X}$ itself is a measure of dispersion. On the otherhand there may be many simple conditions that would justify keeping the chart for ranges (R) without one for $\bar{X}$ or Primarily the $\bar{X}$-chart controls the machine setting or some such similar adjustments. The R - chart controls raw materials, operator's skill, alignment between machines, mechanical conditions, and other possible causes of non uniformity. Whenever these causes are of chief interest the R-chart alone may serve the purpose of quality control.

9.5 NATURAL TOLERANCE LIMITS AND SPECIFICATION LIMITS

A process in statistical control implies that the control charts for both the mean and range show complete homogeneity and in such a case a measure of the variation of the individual products is given by the standard deviation (σ), estimated by $\frac{\bar{R}}{d_2}$ from control data. If μ and σ are the process average and process standard deviation respectively then the limits $\mu \pm 3\sigma$, are called the Natural Tolerance Limits. The probability of an observation lying outside these limits is 0.0027. The width '6σ' which is the inherent variability of the process is given a special name Natural Tolerance.

If μ and σ are not known then $\pm 3\hat{\sigma}$ are the estimates of the natural tolerance limits where

$$\hat{\mu} = \bar{X} \text{ and } \hat{\sigma} = \frac{\bar{R}}{d_2}$$

It might happen that even though the process is in statistical control as exhibited by control charts, the customer may not be satisfied with the product. This happens when the process does not conform to specification limits for that item. These specification limits are generally given in terms of upper and lower tolerance limits. A decision whether a process needs adjustment or not can be made at the point by comparing natural tolerance limits and specification limits.

Comparison : Let $\bar{X}_{max}$ and $\bar{X}_{min}$ denote the Upper Specification Limit (USL) and Lower Specification Limit (LSL) respective by for some quality characteristic. When both these limits are specified, a comparison of these with the "natural tolerance limits" may result in one of the following three situations.

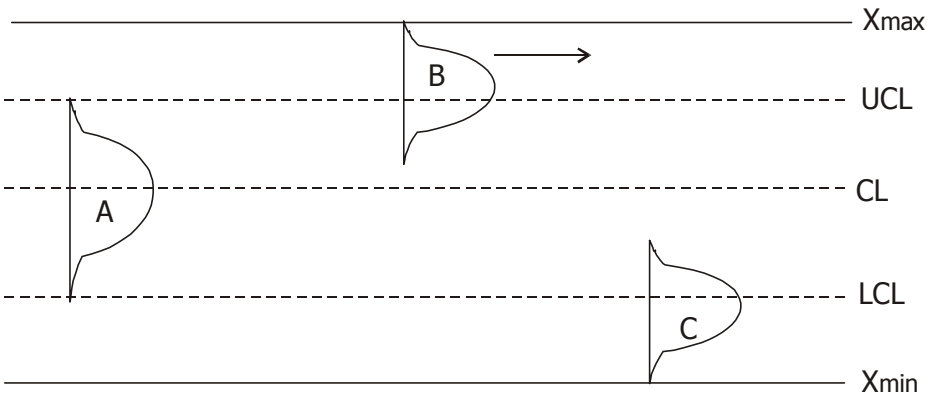
(a) Natural tolerance is considerably smaller than specified tolerance i.e.,

$$\bar{X}_{max} - \bar{X}_{min} > 6\sigma$$

Interpretations :

(i) In such a case almost all the manufactured items will conform to specifications as long as the process is in statistically control and is appropriately centred as in positions A, B, or C as shown in the figure below.

Natural Tolerance Smaller than Specified Tolerance



If the process is operating under one of these conditions, $\bar{X}$ may be permitted to go out of control, provided it does not go too far; In otherwords, the distribution of $\bar{X}$ may be allowed to fluctuate between positions B and C. This will save the time and money for frequent machine setting and delays due to looking for assignable causes of variation which will not be responsible for unsatisfactory product.

(ii) In such a situation since, even considerable shifts in the level of working may not result in the items falling outside specification limits, the time interval between taking successive samples for control chart inspection can be appreciably increased.

(iii) The larger the ratio $(X_{max} - X_{min})$ to the natural tolerance 6σ, the greater is the likelihood of getting good product without assistance from any control chart. This will imply that the process is too good for the product and it may be economical to examine if relaxations in the conditions of production; ex : less costly experiment or processing or material, could be allowed. It may also be worth while to "squeeze" the specification limits, to produce a product superior to the one originally intended.

(b) Specification Limits coincide with tolerance limits i.e.,

$$X_{max} - X_{min} \simeq 6\sigma$$

This is an ideal situation and in this case a process in statistical control obviously implies that the product is meeting the specifications. Here, careful centering of the process is all the more important and if no item is to be rejected then the process has to be centred exactly at the specification mean. Any departure from this centering would result in time of the product going outside the specification limits. As soon as a control chart detects such departure, immediate remedial action should be taken to maintain the centering of the process.

(c) Natural tolerance is greater than specified tolerance

i.e., $X_{max} - X_{min} < 6\sigma$

(i) If the natural tolerances are not included within the specification limits then even with the process in control and the process average perfectly centered at the specification mean, the production of an appreciable quantity of defective articles is inevitable. Here a slight shift in the process average will increase the percent defective. In such a situation, a readjustment of the process is advisable with respect to either the process average or process dispersion or both.

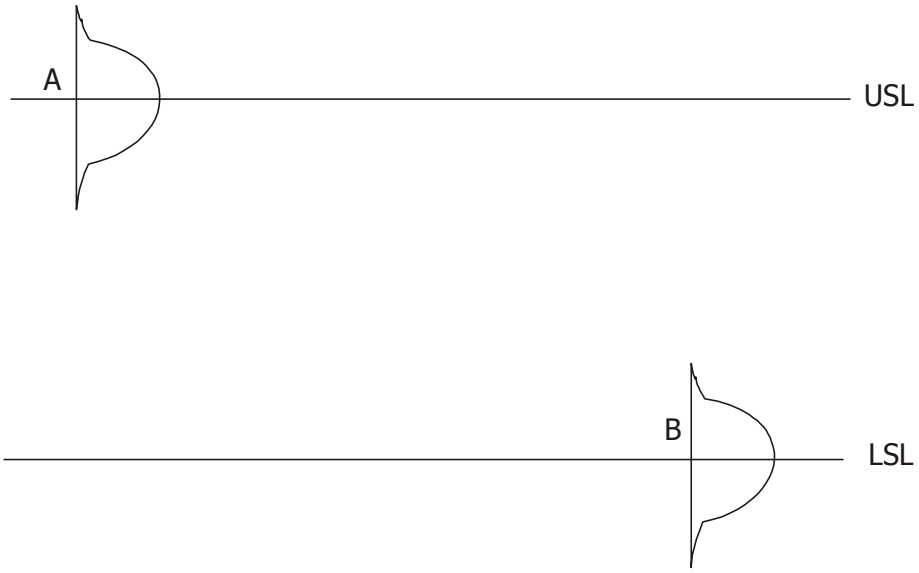
(ii) Also it would be worthwhile to investigate the possibility of relaxing the specified tolerances to the extent of natural tolerances.

If 100% inspection is possible, then defective articles may be sorted out and eliminated, but if 100% inspection is not possible then there is no chance of getting the product all of which will conform to specifications and the only alternative in this case is to relax the specification limits to tolerance limits.

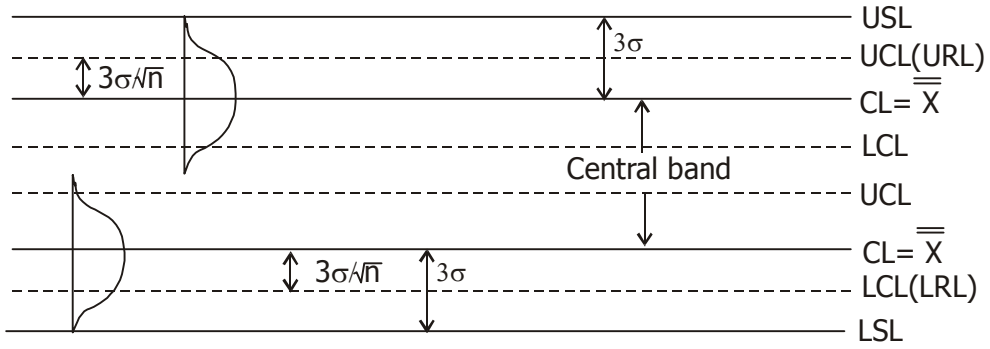
9.6 MODIFIED CONTROL LIMITS

If the specification limits be outside the natural tolerances i.e., $X_{max} - X_{min} > 6\sigma$, then the modified control limits which exhibit the relationship between the specification limits and the $\bar{X}$ values in $\bar{X}$ chart may be used to permit shifts in process levels within permissible limits.

As already pointed out, in such a situation shifts in the values of $\bar{X}$ may be allowed provided it does not go too far. There arises the question : "what are the limits within which $\bar{X}$ - values may be allowed to vary such that the product meets the specification" ? Since if the process is centered at A and B as shown in the figure some of the items will naturally lie outside the specification limits.



Let us have a look at the following figure which shows the statistically controlled universe in its highest and lowest acceptable positions.



The natural tolerances i.e., process dispersion is 6σ. If the universe is at the highest acceptable position, then the process average will be at a distance 3σ below USL and similarly when the universe is at its lowest acceptable position, the process average is at a distance 3σ above the LSL. Thus in this case, instead of fixed central line at $\bar{X}$, we have a central band so that as long as $\bar{X}$ lies in this central band, the product will conform to specifications. The upper and lower edges of the central band are given respectively by

$$USL - 3\sigma,\ LSL + 3\sigma$$

For a sub-group of size n, as it is clear from the figure the highest and lowest satisfactory values of UCL and LCL known as Upper Rejection Limit (URL) and Lowest Rejection Limit (LRL) respectively are given by

$$URL_{\bar{x}} = USL - 3\sigma' + \frac{3\sigma'}{\sqrt{n}}$$

$$LRL_{\bar{x}} = LSL + 3\sigma' - \frac{3\sigma'}{\sqrt{n}}$$

This rejection limits, when used in place of control limits is called modified control limits.

9.7 WORKEDOUT EXAMPLES

Example 1 : The following data gives the defective items in a sample of size 50 presented for inspection. Construct the p - chart.

Sample No.	Number Inspected	Defectives Found	Fraction defectives	Sample No.	Number Inspected	Defectives Found	Fraction defectives
1	50	2	0.04	21	50	1	0.02
2	50	1	0.02	22	50	1	0.02
3	50	2	0.04	23	50	4	0.08
4	50	0	0	24	50	2	0.04
5	50	2	0.04	25	50	2	0.04
6	50	3	0.06	26	50	4	0.08
7	50	4	0.08	27	50	1	0.02
8	50	2	0.04	28	50	3	0.06
9	50	0	0.00	29	50	3	0.06
10	50	3	0.06	30	50	2	0.04
11	50	0	0.00	31	50	3	0.06
12	50	1	0.02	32	50	6	0.12
13	50	2	0.04	33	50	2	0.04
14	50	2	0.04	34	50	3	0.06
15	50	3	0.06	35	50	2	0.04
16	50	5	0.10	36	50	3	0.06
17	50	1	0.02	37	50	1	0.02
18	50	2	0.04	38	50	0	0
19	50	3	0.06	39	50	2	0.04
20	50	1	0.02	40	50	0	0

Solution : The average fraction defective, $\bar{p} = 0.042$

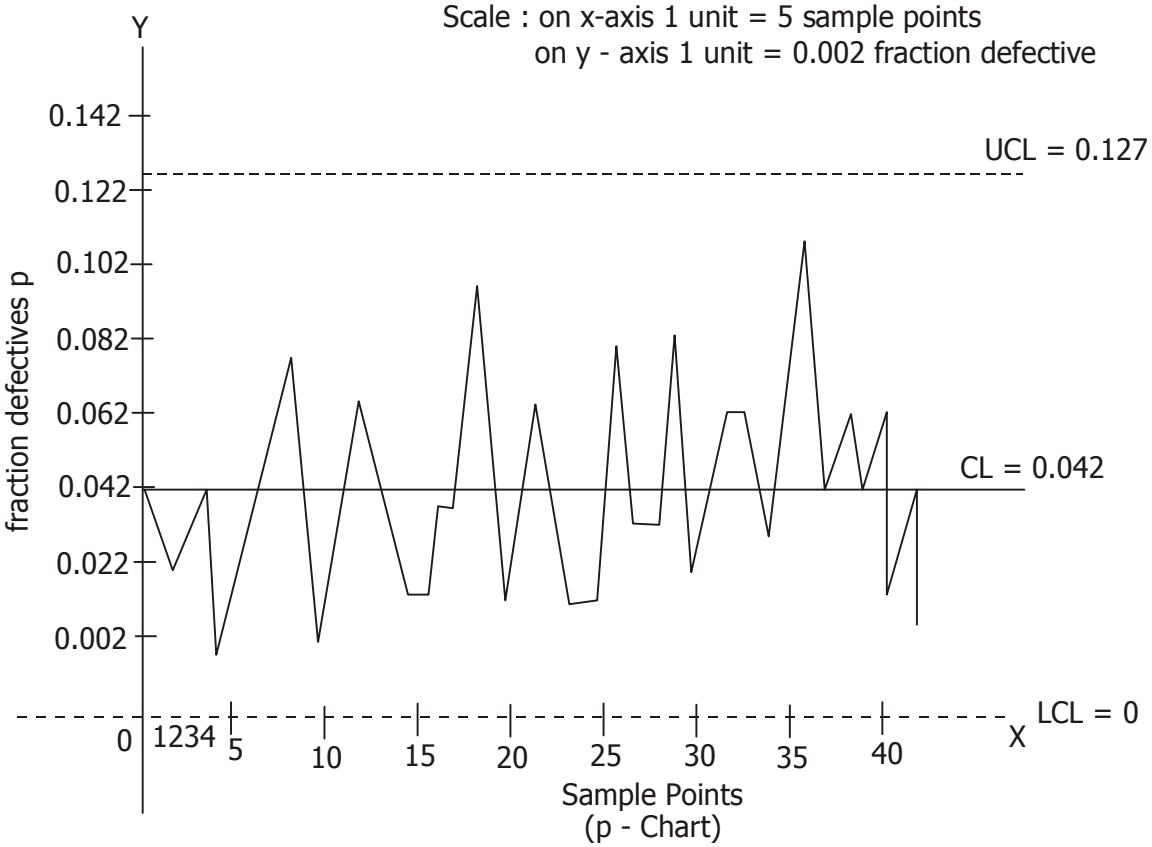
Thus, $UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$ or $\bar{p} + 3\sqrt{\frac{\bar{p}\,\bar{q}}{n}}$

$$= 0.042 + 3\sqrt{\frac{0.042(1-0.042)}{50}} = 0.127$$

$$CL = \bar{p} = 0.042$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}\,\bar{q}}{n}}$$

$$= 0.042 - 3\sqrt{\frac{0.042(1-0.042)}{50}} = -0.043 \text{ = 0}$$



From the graph, we conclude that the process seems to be within control, since no point is out of the control limits.

Example - 2 : The following are the number of defects found in 1000 items of some industry goods inspected every day in a month.

Day	Number of defects	Day	Number of defects
1	1	16	20
2	1	17	1
3	3	18	6
4	7	19	12
5	8	20	4
6	1	21	5
7	2	22	1
8	6	23	8
9	1	24	7
10	1	25	9
11	10	26	2
12	5	27	3
13	0	28	14
14	19	29	6
15	16	30	8

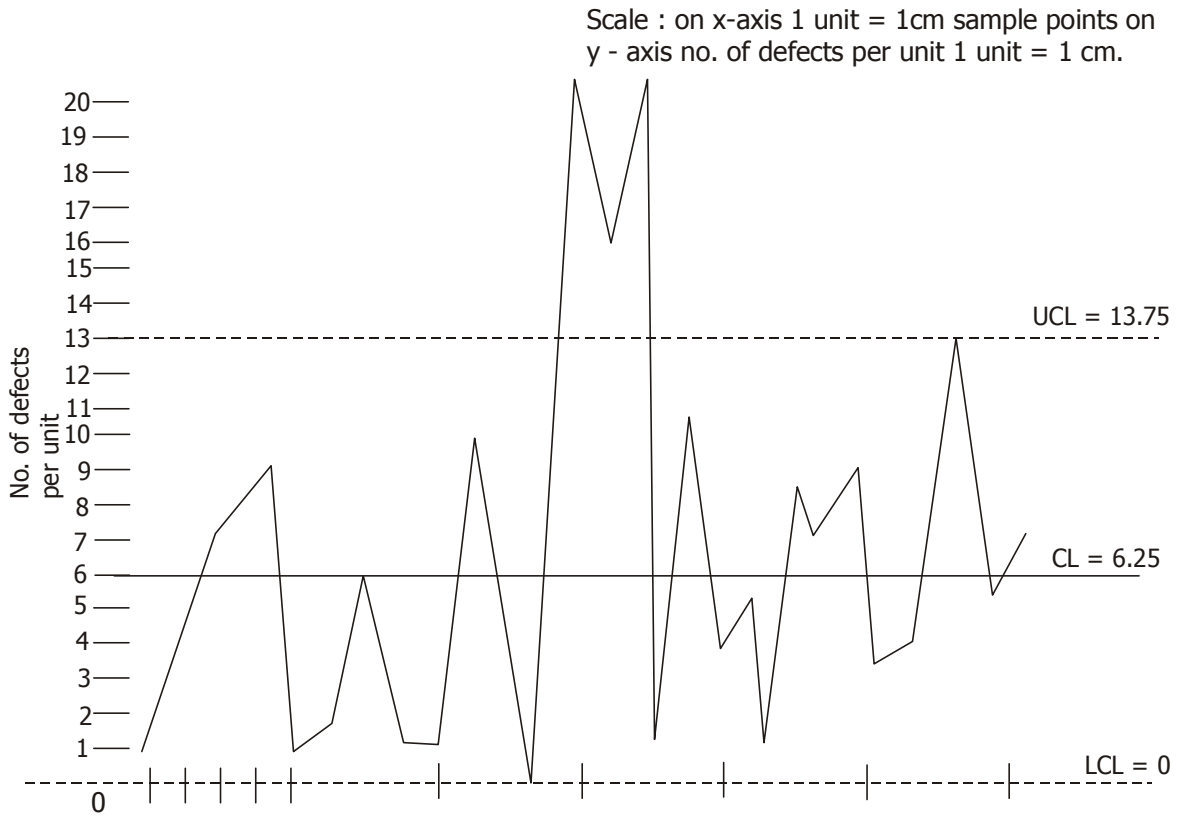
Solution : If c_i denotes the number of defects in the i^{th} group, then we have the average number of defects $\bar{c}$, given by

$$\bar{c} = \frac{187}{30} = 6.23$$

the control limits are

$$\text{UCL} = \bar{c} + 3\sqrt{\bar{c}} = 6.23 + 3\sqrt{6.23} = 13.73$$

$$\text{CL} = \bar{c} = 6.23,\ \text{LCL} = 6.23 - 3\sqrt{6.23} = -1.27 = 0$$



From the graph we find that on 14^{th}, 15^{th}, 16^{th} and 28^{th} day, there were out side the control limits showing the excess of defectives over specified numbers.

Example 3 : Draw a control chart for the following data and state your conclusions.

Sample No. : (each of 100 items)	1	2	3	4	5	6	7	8	9	10	Total
No. of defectives :	12	10	6	8	9	9	7	10	11	8	90

Solution : We have

$$\bar{p} = \frac{12+10+6+8+9+9+7+10+11+8}{10\times 100} = \frac{90}{1000} = 0.09$$

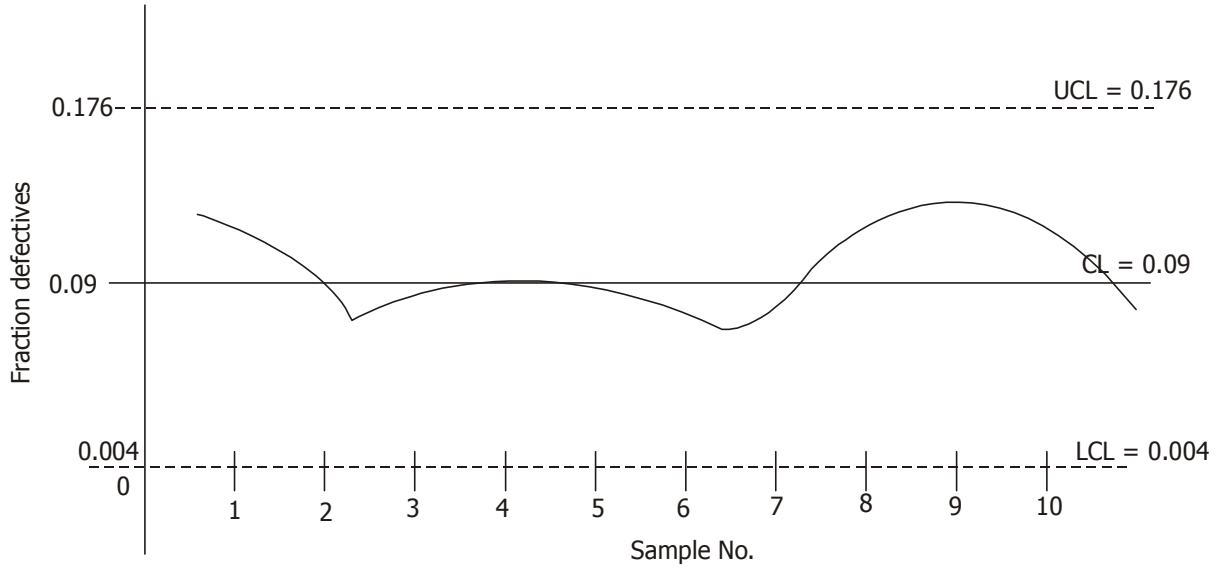
$$\bar{q} = 1 - \bar{p} = 0.91$$

Therefore control limits for the fraction defective p are

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}\,\bar{q}}{n}} = 0.09 + 3\sqrt{\frac{(0.09)(0.91)}{100}} = 0.09 + 0.086 = 0.176$$

$$CL = \bar{p} = 0.09$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}\,\bar{q}}{n}} = 0.09 - 3\sqrt{\frac{(0.09)(0.91)}{100}} = 0.09 - 0.086 = 0.004$$



Since none of the points fall outside the control limits, therefore, the process is in a state of statistical control.

Example 4 : The following data refer to visual defects found by inspection of 10 samples of size 100. Use them to obtain upper and lower control limits for percentage defective in samples of 100. Represent the first ten sample results in the chart you prepare to show the central line and control limits.

Sample No.	1	2	3	4	5	6	7	8	9	10	Total
No. of defectives	2	1	1	3	2	3	4	2	2	0	20

$$p = \frac{20}{10 \times 100} = 0.02$$

$$q = 1 - p = 1 - 0.02 = 0.98$$

$$np = 100 \times 0.02 = 2$$

$$\sqrt{npq} = \sqrt{100 \times 0.02 \times 0.98} = \sqrt{1.96} = 1.4$$

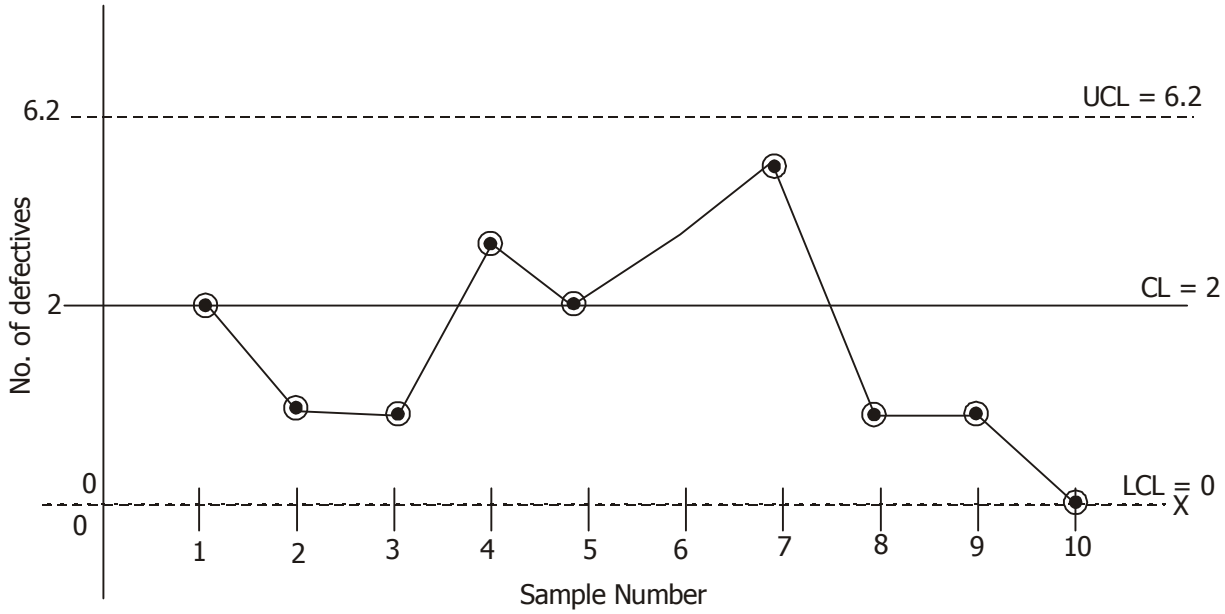
3σ control limits are

$$UCL = np + 3\sqrt{npq} = 2 + 3 \times 1.4 = 6.2$$

$$CL = np = 2$$

$$LCL = np - 3\sqrt{npq} = 2 - 3 \times 1.4 = -2.2 = 0$$

Since there cannot be negative defectives.



Since all sample points are within control limits, the process is under control.

Example 5 : A medium size company carries out the anodizing of various aluminium components. The components are inspected to locate defects in them. The observations from this inspections are given below. Plot a C-chart and draw the conclusion.

Components inspected :	1	2	3	4	5	6	7	8	9	10	Total
No. of defects found during the inspection	2	5	0	5	5	7	2	3	1	7	37 Total No. of defects

Solution : The $3\sigma -$ limits for the construction of C-chart are

$$\left.\begin{aligned} UCL &= \bar{c} + 3\sqrt{\bar{c}} \\ CL &= \bar{c} \\ LCL &= \bar{c} - 3\sqrt{\bar{c}} \end{aligned}\right\} \text{ ---------- (1)}$$

where $\bar{c} = \dfrac{\text{Number of defects in all samples}}{\text{Total number of samples}} = \dfrac{37}{10} = 3.7$

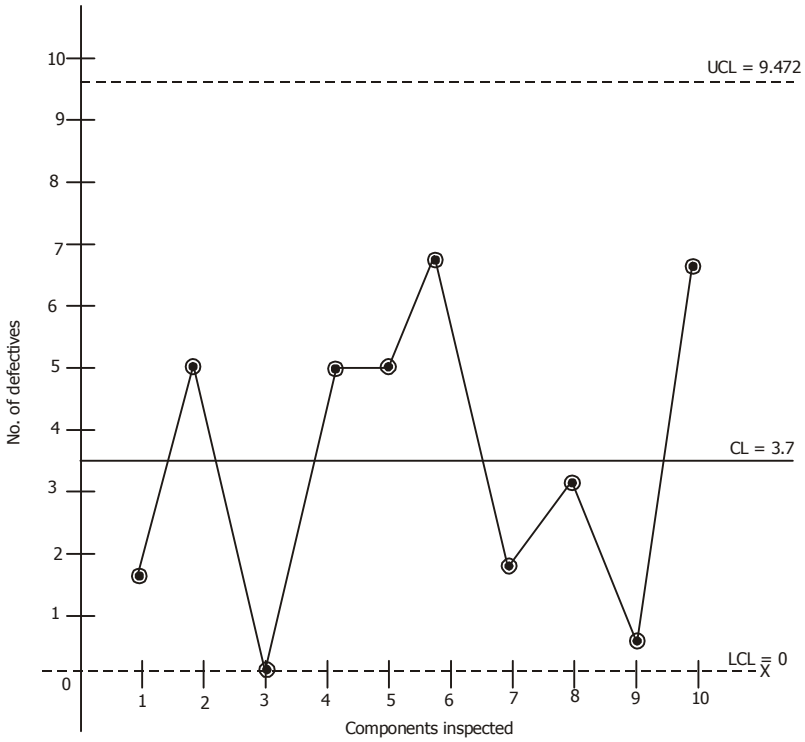
Substituting the value of $\bar{c} = 3.7$ in (1) we get

$$\text{UCL} = 3.7 + 3\sqrt{3.7} = 9.472$$

$$\text{CL} = 3.7$$

$\text{LCL} = 3.7 - 3\sqrt{3.7} = -2.072$ as the limit is negative.

Let us take it as zero.



Since all the points fall within the $3\sigma -$ limits, we conclude that the process is under control.

Example 6 : The following data give the number of defectives in 10 independent samples of varying sizes from a production process.

Sample No	Sample Size	No. of defectives
1	2000	425
2	1500	430
3	1400	216
4	1350	341
5	1250	225
6	1760	322
7	1875	280
8	1955	306
9	3125	337
10	1575	305

Draw the control chart for fraction defective and comment on it.

Solution : Since we have variable sample size, we can draw the control chart for fraction defective in the following three ways.

Method 1 :

Variable Control Limits : In this case we calculate 3σ limits for each sample separately by using the formula.

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}\,\bar{q}}{n_i}} \text{ and } LCL = \bar{p} - 3\sqrt{\frac{\bar{p}\,\bar{q}}{n_i}}$$

where $\bar{p} = \frac{\Sigma d_i}{\Sigma n_i} = \frac{3187}{17790} = 0.1791$

and $d_i =$ No. of defectives in the i^{th} sample.

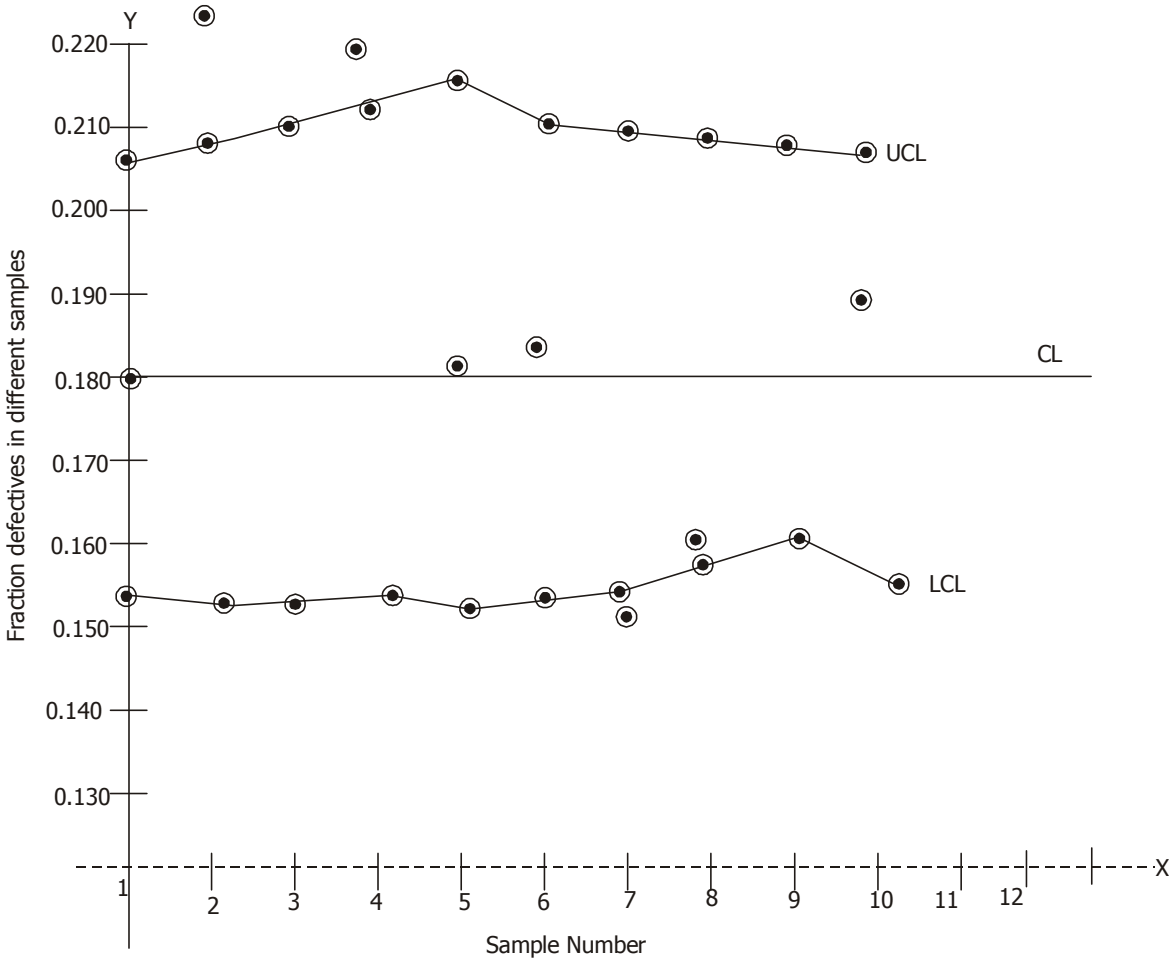
$n_i =$ Sample size of the i^{th} sample

$\therefore \bar{q} = 1 - \bar{p} = 1 - 0.1791 = 0.8209$

$\therefore \bar{p}\,\bar{q} = 0.1791 \times 0.8209 = 0.1470231$

COMPUTATIONS FOR p - CHART (VARIABLE CONTROL LIMITS)

n	d	$p = \frac{d}{n}$	$\frac{\bar{p}\,\bar{q}}{n}$	$\sqrt{\frac{\bar{p}\bar{q}}{n}}$	$3 \cdot \sqrt{\frac{\bar{p}\bar{q}}{n}}$	UCL	LCL
2000	425	0.2125	0.000073	0.008573	0.025719	0.2048	0.1534
1500	430	0.2867	0.000098	0.009899	0.029698	0.2088	0.1494
1400	216	0.1543	0.000105	0.010247	0.030741	0.2098	0.1484
1350	341	0.2526	0.000109	0.010440	0.031321	0.2104	0.1478
1250	225	0.1800	0.000118	0.010863	0.032588	0.2117	0.1465
1760	322	0.1829	0.000084	0.009138	0.027413	0.2065	0.1517
1875	280	0.1495	0.000078	0.008854	0.026562	0.2057	0.1525
1955	306	0.1565	0.000075	0.008672	0.026015	0.2051	0.1531
3125	337	0.1078	0.000047	0.006856	0.020567	0.1997	0.1585
1575	305	0.1937	0.000093	0.009659	0.028977	0.2080	0.1502
17790	3187						

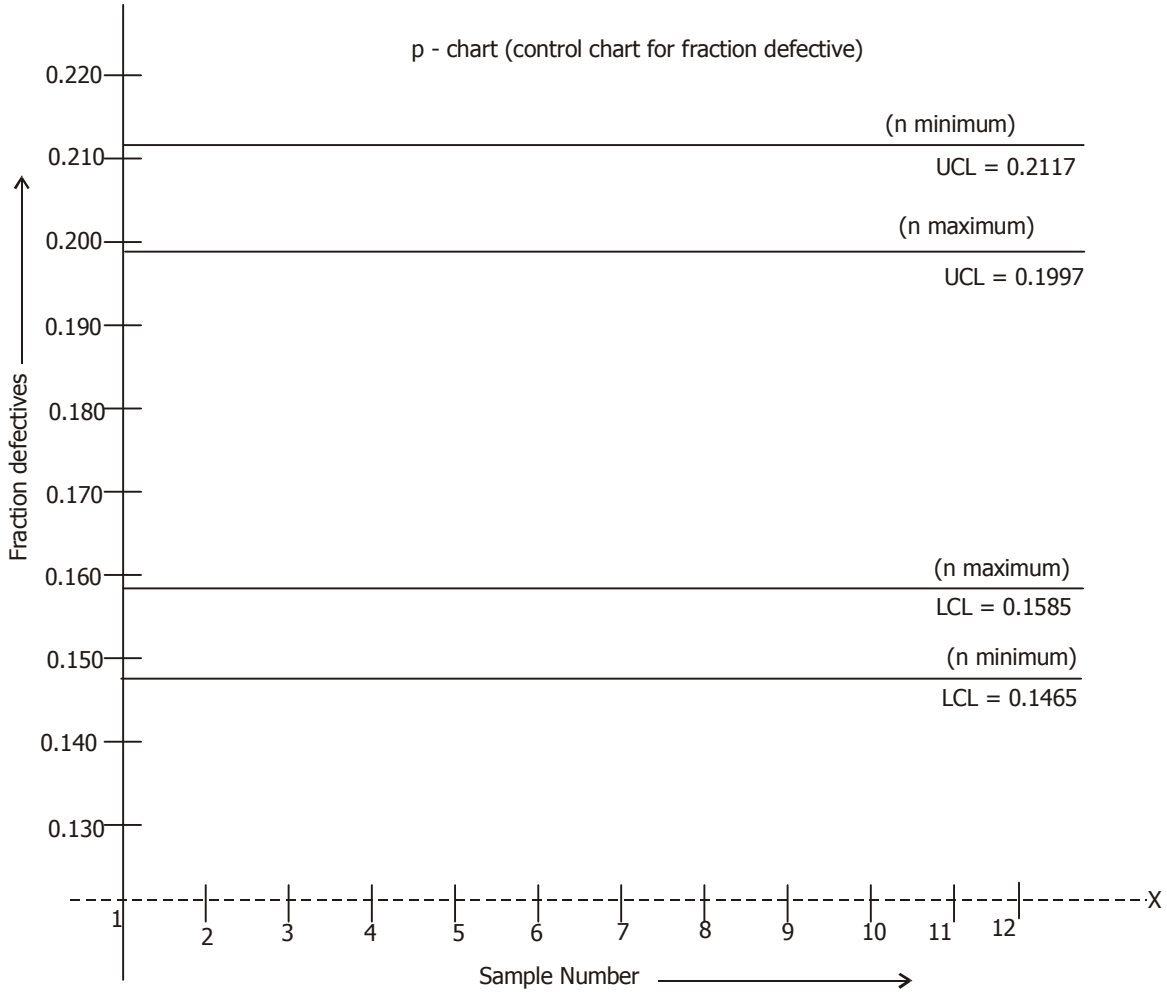


From the chart it is obvious that a number of sample points corresponding to sample numbers 1, 2, 4, 7 and 9 are outside the respective control limits. Hence the process is not in a state of statistical control. This suggests the presence of some assignable causes of variations, which should be detected and eliminated.

Method 2 : Here we construct two sets of control limits, one based on the maximum sample size $n = 3125$ (corresponding to 9^{th} sample) and other based on the minimum sample size $n = 1250$ (corresponding to 5^{th} sample). From the previous Table we note that the corresponding sets of control limits are

For $n = 3125;\ UCL = 0.1997$ and $LCL = 0.1585$

$n = 1250,\ UCL = 0.2117$ and $LCL = 0.1465$



Sample points corresponding to sample number 1, 2, 4 and 9 lie outside the outer band (based on minimum sample size). The process is out of statistical control.

Method 3 : Here we standardise the statistic p by the formula

$$Z_i = \frac{p_i - E(p_i)}{S.E(p_i)} = \frac{p_i - \bar{p}}{\sqrt{\frac{\bar{p}\,\bar{q}}{n}}} \qquad (1)$$

where $\bar{p}$ is computed in (1) and plot the Z - values against the corresponding sample number. Since n is large, $Z_i \sim N(0,1)$ and hence

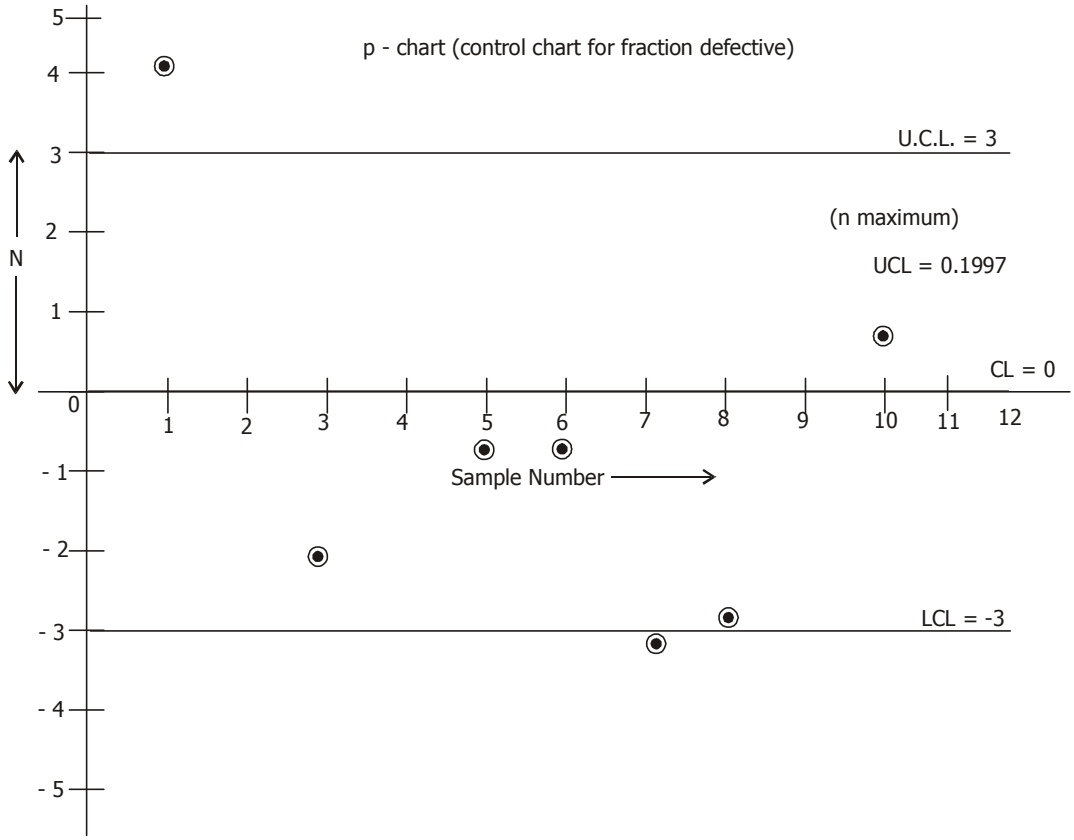
$$UCL_2 = 3; \qquad LCL_2 = -3; \qquad CL_2 = 0$$

COMPUTATION OF Z - VALUES

p	$p - \bar{p}$	$\sqrt{\frac{\bar{p}\,\bar{q}}{n}}$	$Z = \frac{p - \bar{p}}{\sqrt{\frac{\bar{p}\,\bar{q}}{n}}}$
0.2126	0.0334	0.0086	3.8841
0.2867	0.1076	0.0099	10.8686
0.1543	-0.0248	0.0102	-2.4313
0.2526	0.0735	0.0104	5.25
0.1800	0.0009	0.0109	0.0826
0.1829	0.0038	0.0091	0.4176
0.1495	-0.0296	0.0089	-3.2558
0.1565	-0.0326	0.0087	-2.5977
0.1078	-0.0713	0.0069	-10.3333
0.1937	0.0146	0.0097	1.5052

Since a number of sample values Z except the sample numbers 3, 5, 6, 8 and 10 lie outside the limits ± 3, the process is out of statistical control.

Control Chart for Fraction Defectives (Based on Z-values)



9.8 EXERCISE

1. Explain the main control charts for attributes and obtain their control limits. Discuss the advantages and disadvantages of control charts of variables and control charts of attributes.
2. Explain the statistical basis and construction of p and np charts. How is the choice between p and np charts made ?
3. Distinguish between defect and defective. Give some examples of defects for which the C - chart is applicable. How do you calculate control limits for a c - chart ?
4. What is meant by specification limits and control limits. Does a process in statistical control ensure, that all the products will be within specifications ? Justify your statement by means of an example.
5. Explain the relation between the control limits, specification limits and rejection limits. Also explain how the rejection limits are constructed.
6. The number of defects in 20 pieces of cloth each of 100 metres length is given below.

   1, 3, 3, 1, 6, 4, 3, 7, 10, 2, 2, 6, 4, 3, 2, 7, 1, 5, 6, 4

   Draw the appropriate chart and say whether the process can be considered to be in control.

7. Explain briefly the terms (i) specification limits (ii) Tolerance limits (iii) Modified control limits and (iv) control limits as used in statistical quality control.

8. Draw a suitable control chart for the following data pertaining to the number of coloured threads (considered as defects) in 15 pieces of cloth in a certain make of synthetic fibre and state your conclusions.

   7, 12, 3, 20, 21, 5, 4, 3, 10, 8, 0, 9, 6, 7, 20

9. In the final inspection of cars manufactured in a factory each car is checked for minor defects. These defects do not influence the acceptance or rejection of the cars. Rather they require a certain amount of additional labour to satisfy the consumer. A brief study was made to see if the number of minor defects was relatively constant and under control. The results of first twelve units checked are given below.

Unit No. (i)	1	2	3	4	5	6	7	8	9	10	11	12
No. of defects (C_i)	4	3	7	4	5	5	4	5	7	8	6	7

Discuss the control chart for this process stating the underlying assumptions and give your conclusions.

LESSON - 10

VITAL STATISTICS AND DEATH RATES

LEARNING OBJECTIVES

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

* Understand the meaning of vital statistics, Importance of vital events to the mankind.
* Study the Importance of vital statistics
* Estimating vital statistics
* Estimation of Total Population
* Studying the Mortality Rates

LESSON OUTILINE

10.1 INTRODUCTION

Vital Statistics are estimated as a means of preserving evidence of personal status in order to establish the individual's identity. The statistical data related to vital events like birth, death, marriage, divorce, adoption, and legal separation etc., for the proper determination of the various rights and obligations which are connected to life insurance, social security, employment or settlement of properties and inheritance. They enable many social investigations in enterprises undertaken by Government on national and Local basis and these statistics provide invaluable material for the guidance of public health authorities. With the help of vital statistics we study the health conditions of people in a country over a period of time as well as examine the efficiency of public health programmes.

According to Arthur Newsholme, "vital statistics forms perhaps the most important branch of statistics as it deals with mankind in the aggregate. It is the science of numbers applied to the life history of communities and nations".

The statistics of births, deaths and marriages depend upon the compulsory registration of the following information.

Births	:	When and where born, sex, name and parents, profession of the parents; age of mother and father number of children born, number of living children and date of marriage.
Deaths	:	When and where died, name and sex of the diseased, Age, profession, cause of death, married or unmarried, number of children, age of surviving wife/ husband.
Marriages	:	When married, names of husbands, and wives, ages marital conditions, professions etc.;

10.2 MEANING OF VITAL STATISTICS

Vital Statistics deals with registration of births and deaths which influence the future growth of population. Apart from that information vital statistics supply a variety of other information such as marriage, divorce, adoption, legitimation, recognition, sickness, annulment and separation etc..

Arthur News Holme Writes, "that branch of Biometry which deals with data and laws of human mortality morbidity, and demography.

Prof. B. Benjamin defines "vital statistics are conventionally numerical records of marriage, births, sickness and deaths by which the health and growth of a community may be studied".

U.N. Statistical office has elaborately defined vital statistics in these words, "vital statistics has come to mean statistics of live-births, deaths, fatal deaths (still birth), marriage, divorce, adoption, legitimation, recognition, annulment, and legal separation".

The definitions clearly show that vital statistics, in its broader sense, refers to all types of population statistics.

10.3 IMPORTANCE OF VITAL STATISTICS

Vital Statistics is important in almost all the spheres of human activity. The following are the important applications of vital statistics :

1. IMPORTANCE FOR THE GOVERNMENT :

Vital statistics are indispensable for the government in the formulation of new health schemes and assesment of old health schemes. For example, programmes for eradicating infectious diseases can be undertaken only when government has up-to-date information of the number of effected persons; number of persons died from a particular disease. Similarly, other programmes such as accident prevention programmes, crime prevention programmes etc; can be launched only when the Government has full information on the causes of events. Government can adopt certain measures to control the rapid growth in population by framing new population policy.

2. IMPORTANCE TO HEALTH-AUTHORITIES :

Vital statistics is useful to Health Authorities to improve the efficiency of their working. If an infectious disease spreads in an area which causes too many deaths (in that particular area) then health anthorities can launch a disease eradication programme on a war-footing to save the people from the attack of the disease. Vital statistics is useful to public-health services like Hospitals, family welfare centres, mother and children care centres etc... centres established on the basis of information obtained through vital registration. Not only this, public health programmes of post natal care, public safety, accident prevention and crime eradication programmes are formulated on the basis of mortality and morbidity.

3. IMPORTANCE OF THE STUDY OF POPULATION TRENDS :

Vital statistics reflect the change in the pattern of the population of any region, community or country interms of the number of births, deaths and marriages. The size of the population of different regions by births and death rates helps us to form an idea about the population trend of regions or countries and the general standard of living of different regions.

Various types of birth rates help us to form an idea if the population of the region has a tendency to increase, decrease or remain constant, over a period of time, for instance, the net reproductive rate gives the rate of change of population. Vital rates can also be calculated through vital statistics. This clearly helps in determining the growth rate of population with respect to religion, caste, region etc., It helps in knowing the rate at which the present population is being established.

4. IMPORTANCE FOR DEMOGRAPHERS AND MEDICAL RESEARCH :

Vital statistics are indispensable for demographers and medical research. It helps to the demographers to predict the future behaviour of the population. On the basis of these arrangements predictions, necessary arranger can be made to supply food and other necessities of life to people in advance. Where as in medical research to study the effectiveness of a new drug to control births and deaths. Thus vital statistics helps in demographic and medical and pharmaceutical research by providing necessary guidance through mortality and mutuality behaviours of the community.

5. INTER-REGION COMPARISON :

Vital statistics are indispensable to compare the health conditions of people in different regions of the country as well as in the same region at different times. They may also be used to compare the behaviour of population in different regions as well as in the same region.

6. INTERNATIONAL COMPARISON :

Vital statistics are used in comparing the health conditions of people not only at the National level but also at the International level. With the aid of the vital statistics, we can also analyse which country is high in rank compared to others in birth and death rates and why ? Then measures can be suggested to control the birth and death rates and improve the health conditions of the people. For example, the Governments of developing countries are very keen on bringing the birth rates equal to that of developed countries to control the rapid rise in population. Therefore, most of the Governments have adopted family planning programmes to bring the birth rate down.

7. IMPORTANCE FOR THE INDIVIDUAL :

Vital statistics is useful to the individuals as well. For example, Birth record is very much essential for entering a service or to become a voter or for the computation of retirement date. Similarly, death records of persons in a family are used by individuals to become the legal owners of property or to get the benefits from the Government in case the deceased was a Government employee, whether he was insured or not.

8. IMPORTANCE IN PUBLIC ADMINISTRATION :

Statistics in general and vital statistics in particular form the fundamental basis in Public Administration. It helps the administration in planning and evaluation of various plans - economic, educational, health, social securities, defence etc., This is also helpful in rationing of food, construction of voters list, division of constituencies, recruitment in defence and other services, and other allied programmes.

9. IMPORTANCE TO THE ACTURIAL SCIENCE :

The whole of acturial sciences including life insurance is based on the mortality tables. The vital records concerning all possible factors contributing to the deaths in various age group help us in all life insurance schemes.

10.4 METHODS OF OBTAINING VITAL STATISTICS

Vital statistics data are usually obtained by the following methods :

1. Registration Method :

The most important source of obtaining vital statistics data is the registration method which contains the registration of vital events such as births, marriages, deaths etc., A large number of countries have adopted this system which is legally binding on each individual. In many countries of the world, laws have been enacted for the compulsary registration of every vital event with proper authorities together with such information as the age of the mother, religion, social status etc., Registration of births provide information regarding the place of birth, sex, age and religion of the parents, legitimacy, number of previous issues and their sexes, father's occupation and birth place of the parents. Similarly death registration furnishes information on the place of death, sex, marital status, number of issues, birth place, occupation and cause of death. For example the death of a person is to be reported by family members to the concerned authorities, who issue death certificates.

The legal responsibility of giving the information in respect of births and deaths to the registering authority lies with the following persons :

(i) Head of the Household, and in the absence of any such person, the eldest male present therein during the said period.

(ii) In case of births and deaths in a hospital, health centre, nursing home or other such institutions the medical officer incharge or any person authorised by him on his behalf.

(iii) In respect of births and deaths in a Jail, the Jailor incharge.

(iv) In respect of birth and deaths in a choultry, hostel, dharmasala, boarding house, lodging, shop, the person incharge there of

(v) In any other place, such person as may be prescribed.

The registration office issues a certificate of registration for birth or death. This certificate is of great use. The birth certificate is a legal document and helps a person at the time of entering a school, taking a job, or take an insurance policy and to prodcure a ration card. Similarly death certificate helps in getting insurance money and in disputes of inheritance to land and property.

The registration method is easy in operation and very effective. Yet it suffers from the lacuma that a large number of births and deaths are not reported to the registration office as the law has not been enforced effectively.

2. POPULATION CENSUS METHOD :

A census presents a comprehensive profile of a country's population. Census operations are conducted in all most all countries of the world. Population census is conducted at regular intervals of time, usually every ten years. In a census, the enumeration of every individual of all habitational areas is carried out at a specific time. In India, the 1981 census record was updated by enumerators from March 1 to March 5, 1981. Census consists of complete enumeration of the population of the particular area under study and collecting the information from individuals regarding age, sex, marital status, occupation, religion and other economic and social characteristics. But this information is available for the census year only which is the main drawback. Hence, the census method fails to provide the data suitable for vital statistics.

3. AD HOC SURVEYS :

When ever there is no provision for such a registering of births and deaths, or in such situations an ad hoc survey may be under taken to compile data on births and deaths. Ad hoc survey may also be under taken in areas where there is a proper arrangements of registration of vital events to have a check an accuracy.

In addition to these methods, data on vital statistics can also be obtained from hospitals. They maintain proper records of births and deaths which take place in the hospitals.

4. ANALYTICAL METHODS :

When the Ad-hoc surveys are not possible to conduct surveys to asses the population at any period in between two census years. The population estimates at a given time can be obtained by Mathematical Methods with out Ad hoc surveys. The estimates are based on the assumption that the population grows at a constant rate during the inter censal years. For example, given the birth and death rates, population growth may be estimated for an intercensal period. The number of births or deaths at a given period of time in a population sub-group, can also be estimated. Analytical methods which make use of the available data with out requiring any fresh survey are given below :

Estimation of population in a given intercensal year can be done by the formula.

$$\hat{P}_i = P_0 + \frac{n}{N}(P_1 - P_0)$$

where $\hat{P}_i =$ Estimated population at some interensal year t

$P_0 =$ Population in the previous census.

$P_1 =$ Population in the succeeding census.

$N =$ Number of years between the census.

$n =$ Number of years between the given year and previous census year.

Example : The population of Hyderabad city, according to census 1971 and 1981 are 7, 45, 876, and 12, 50, 532 respectively. The population estimates for the year 1975 is

$$P_1 = 12,\ 50,\ 532,\ P_0 = 7,45,876,\ N = 10,\ n = 4$$

$$\text{Thus } \widehat{P_i} = P_0 + \frac{n}{N}(P_1 - P_0) = 7,45,876 + \frac{4}{10}(12,50,532, 7,45,876)$$

$$= 9,47,738$$

10.5 ESTIMATION OF TOTAL POPULATION

The population may be defined as the total number of people (persons) living in a country at any given point of time. Symbolically, the population (P_t) of a country at that time t may be expressed as :

$$P_t = P_0 + B_i - D_t + I_t - E_t$$

where

$P_t =$ Total population at a time - t

$P_0 =$ Total population at a given point of time taken as base

$B_t =$ Total number of births at time - t

$D_t =$ Total number of deaths at time - t

$I_t =$ Total number of Immigrants at time - t

$E_t =$ Total number of emigrants at time - t

Among the four variables, the first two variables namely births and deaths are most important determinants of the change in population. Therefore, we study in detail two variables to examine the behaviour of population over a period of time. The number of births depends upon fertility and the number of deaths upon mortality rates. Thus, to study the size of the population, we have to study in detail fertility and mortality rates in the population.

10.6 MEASUREMENT OF MORTALITY RATES

For the study of mortality conditions of a population following rates of mortality are frequently used :

1. Crude Death Rate
2. Specific Death Rate
3. Age Specific Death Rate
4. Standardised Death Rate
5. Infant Mortality Rate

10.6.1 CRUDE DEATH RATE

This is the simplest measure of all the measures of mortality. The crude death rate refers to the number of deaths per thousand in the population of any geographic location or community during a given period of time. The crude death rate is calculated with by the expression of the form:

$$\text{CDR} = \frac{\text{Total number of deaths in a year}}{\text{Annual Mean population}} \times 1000$$

or

$$\text{CDR} = \frac{\Sigma D}{\Sigma P} \times 100$$

where $\Sigma D =$ Total Number of deaths; $\Sigma P =$ Annual Mean population.

MERITS :

1. It is easy to calculate and simple to understand
2. The main factor which is responsible for wide use of crude death rate is to study the health conditions.
3. Since the entire population of the region is exposed to the risk death CDR is a probability that emphasizes the possibility that a person belonging to the given population may die in the said (given) period.

DEMERITS :

1. In crude death rate, ignores the age and sex completely, remains a draw-back.
2. Children, in the lower age group and those who belong to older generation are exposed to higher risk of mortality when compared to younger generation.
3. Death rate differs for females, irrespective of the age group when compared males.
4. C.D.R. is not suitable for comparing the mortality in two places or in the same place for two periods.

LIMITATIONS :

It is exposed to the following limitations :

(i) There is no possibility for inter regional comparison because of variations in age and sex of two or more populations.

(ii) A population having older age group persons above 60 is certain to have a high crude death rate compared to a population having large proportion of persons in the age group of 20 to 50 years.

(iii) The crude death rate may be inaccurate because most of the deaths go unrecorded. For example, deaths of newly born children are rarely registered.

ILLUSTRATION 1 : From the following data, compute the death rate for populations A and B.

Age-group	Population A			Population B		
Years	Population	Deaths	Death Rate	Population	Deaths	Death Rate
0 - 15	20000	640	34	12000	500	41.7
15 - 30	12000	240	20	30000	600	20.0
30 - 45	50000	1200	24	60000	900	15.0
45 - 60	30000	500	16.7	15000	660	44.0
Above 60	8000	300	37.5	3000	220	73.3
Total	120000	2880		120000	2880	

Population A

$$C \cdot D \cdot R \cdot = \frac{\Sigma D_A}{\Sigma P_A} \times 1000 = \frac{2880}{120000} \times 1000 = 24$$

Population B

$$C \cdot D \cdot R \cdot = \frac{\Sigma D_B}{\Sigma P_B} \times 1000 = \frac{2880}{120000} \times 1000 = 24$$

The crude death rate is identical in both populations. i.e. 24 per thousand. But there are glaring differences in death rates of various age groups.

10.6.2 SPECIFIC DEATH RATE :

The death rate for any specific part of the population like age, sex, occupation etc., is generally known as specfic death rate. It is more useful than C.D.R.. For instance, people interested in infant or child welfare work would like to know the mortality condition in the age groups below 1 year, 1 - 4 years, 5 - 9 years etc., Similarly, those interested in material health may be interested in knowing about the number of deaths that occured among women of child bearing age in the total population.

Death rate computed for a particular specified section of the population is termed as specific death rate (S.D.R.). S.D.R. for a given Geographical region during a given period is defined as

$$\text{S.D.R.} = \frac{\text{Total number of deaths in the specified section of the population}}{\text{Total population of the specified section in the same period}} \times 1000$$

10.6.3 AGE SPECIFIC DEATH RATE (AGE S.D.R.) :

Study of death rate specific to age group, is known as Age specific death rate.

The age specific death rate (ASDR)

$$D_x = \frac{D}{P_x} \times 1000$$

where $D_x =$ Age specific death rate per thousand

$D =$ Deaths of the specific age group

$P_x =$ Population of the specific age group

Remark : One can calculate the crude death rate with the help of Age specific death rate by using the following formula :

$$\text{C.D.R.} = \frac{\Sigma\left(P_X \cdot D_X\right)}{\Sigma P_X}$$

MERITS :

1. The death rates specific to age and sex overcome the draw back of C.D.R.
2. Specific death rates provide the more appropriate measures of the relative mortality situations in the regions.
3. The specific death rate for age and sex is one of the most important and widely applicable type of death rate.
4. It supplies one of the essential components required for the computation of net reproduction rate and construction of life table.

DEMERITS :

1. For overall comparison of mortality conditions prevailing in two different regions, specific death rates are not much useful.
2. Specific death rates completely ignore the factors like social, occupational and topographical. So, they are called differential mortality. In order to eliminate such spurious effects, standard death rates are computed.

10.6.4 STANDARDISED DEATH RATE :

The crude death rate is not suitable for inter regional comparison because of age and sex-composition differentials. So the comparison of crude death rates would lead to misleading and fallacious conclusions. To overcome this problem, the death rate is standardised. The processing of standardisation is nothing but making the population similar for comparing mortality conditions. Standardisation is also known as "connected rate", or 'adjusted rate' or 'age adjusted death rate'.

The standardisation can be : (i) direct or (ii) indirect.

(i) Direct Method : According to this method the standard death rate is computed by applying mortality rates of each age-group of two populations or regions to some standard populations. Thus if P_x^S is the number of persons in the age group x to $x+1$ in the standard population, then the standardised death rates for the regions A and B are given as :

$$(STDR)_A = \frac{\sum_x m_x^a \; P_x^s}{\sum_x P_x^s};$$

where $m_x^a =$ mortality rate of the persons at region A.

$$(STDR)_B = \frac{\sum_x m_x^b \; P_x^s}{\sum_x P_x^s};$$

$m_x^b =$ mortality rate of the persons at region B.

These age adjusted death rates for regions A and B respectively are nothing but the C.D.R. observed in the standard population if it were subjected to Age S.D.R. of the regions A and B.

(ii) Indirect method : This method may be used when we are supplied standard death rates for each age-group of the population. These rates enable us to compute the number of deaths that would have occured in each age-group on the assumption that standard death rates apply to populations understudy.

The expected number of deaths can be totalled and can be employed to compute Index rates of two populations. Symbolically

$$\text{Index Rate} = \frac{\text{Total no. of expected deaths}}{\text{Total population}} \times 1000$$

The index rate varies according to age composition of the population. If population has a large number of infants and old persons compared to the second population then its index rate would be high compared to the second population in which a larger number of young persons are present. In case the index rates vary, the age composition of two populations is not identical. In this situation Index rates are adjusted to eliminate the difference in age-composition. The adjustment factor often termed as standardising factor denoted by "C" and is defined as

$$C = \frac{\text{Standard mortality rate for all ages}}{\text{Index rate}}$$

Lastly crude death rate multiplied by C to compute standardised death rate,

Symbolically,

$$STDR = C \cdot D \cdot R \times C$$

$$= C \cdot D \cdot R \cdot \times \text{Standardising factor}$$

Illustration 2 : Compute crude and standardised death rates of the populations of town A and B. regarding A as standard population, from following data.

Age Group	Town A		Town B	
(years)	Population	Deaths	Population	Deaths
Below 10	30000	720	80000	2000
10 - 45	40000	800	104000	2080
Above 45	20000	560	16000	480
Total	90000	2080	200000	4560

Determine which of the Towns A or B is more healthy ?

Solution : Computation of crude and standardised death rates.

Age Group	Town A			Town B			
(years)	Population	Deaths	Death rate per 1000	Population	Deaths	Death rate per 1000	
	P_x^a	D_x^a	m_x^a	P_x^b	D_x^b	m_x^b	$P_x^a\ m_x^b$
Below 10	30000	720	24	80000	2000	25	750000
10 - 45	40000	800	20	104000	2080	20	800000
Above 45	20000	560	28	16000	480	30	600000
Total	90000	2080		200000	4560		2150000

Crude Death Rates :

$$\text{C.D.R. for population A} = \frac{\sum_x D_x^a}{\sum_x P_x^a} \times 1000$$

$$= \frac{2080}{90000} \times 1000 = 23.11$$

$$\text{C.D.R. for population B} = \frac{\sum_x D_x^b}{\sum_x P_x^b} \times 1000$$

$$= \frac{4560}{200000} \times 1000 = 22.8$$

Standard Death Rates :

Since population A is taken as standard population,

$$\text{STDR for A = C.D.R. for A} = 23.11$$

$$\text{STDR for B} = \frac{\sum_x P_x^{\,a}\ m_x^{\,b}}{\sum_x P_x^{\,a}}$$

$$= \frac{2150000}{90000} = 23.89$$

We may, therefore, conclude that death rate in population B is greater than that in population A. Hence Town A is healthier than Town B.

ILLUSTRATION 3 : Calculate the standardised death rate (with respect to age) using following data :

Age Specific :	0 - 2	2 - 5	5 - 15	15 - 40	40 - 60	60 above
death rates :	375	210	85	41	93	195
Standardised :	15	10	12	38	15	10
Population ('000)						

Solution : Computation of standardised Death Rate.

Age	**Specific Death Rate** (D_X)	**Standardised population** (P_X)	$P_X \cdot D_X$
0 - 2	375	15000	5625000
2 - 5	210	10000	2100000
5 - 15	85	12000	1020000
15 - 40	41	38000	1558000
40 - 60	93	15000	1395000
60 above	195	10000	1950000
Total		$\Sigma P_x = 1,00,000$	$\Sigma P_x \cdot D_x = 13,648,000$

$$\text{Standardised Death Rate} = \frac{\Sigma P_x \times D_x}{\sum_x P_x}$$

$$= \frac{13,648,000}{1,00,000} = 136.48$$

ILLUSTRATION 4 : Estimate the standardised death rates for the following two countries :

Age group (years)	Death Rate per 1000		Standardised population (in Lakhs)
	Country A	Country B	
Below 5	20.00	5.00	100
5 - 14	1.00	0.5	200
15 - 24	1.40	1.00	190
25 - 34	2.00	1.00	180
35 - 44	3.30	2.00	120
45 - 54	7.00	5.00	100
55 - 64	15.00	12.00	70
65 - 74	40.00	35.00	30
75 above	120.00	110.00	10

Solution : Calculation of standard Death Rates

Age-Group (years)	**Death rate per 1000**		**Standard Population**	$m_x^a \cdot P_x^s$	$m_x^b \cdot P_x^s$
	Country A m_x^a	**Country B** m_x^b	P_x^s		
Below - 5	20.0	5.0	100	2000.	500.0
5 - 14	1.0	0.5	200	200.0	100.0
15 - 24	1.4	1.0	190	266.0	190.0
25 - 34	2.0	1.0	180	360.0	180.0
35 - 44	3.3	2.0	120	396.0	240.0
45 - 54	7.0	5.0	100	700.0	500.0
55 - 64	15.0	12.0	70	1250.0	840.0
65 - 74	40.0	35.0	30	1200.0	1050.0
75 and above	120.0	110.0	10	1200.0	1100.0
Total			1000	7372	4700

Standardised Death Rate for Country A

$$(\text{STDR})_A = \frac{\sum_x m_x^a \; P_x^s}{\sum_x P_x^s} = \frac{7372}{1000} = 7.372$$

Standardised Death Rate for Country B

$$(STDR)_B = \frac{\sum_x m_x^b P_x^s}{\sum_x P_x^s} = \frac{4700}{1000} = 4.7$$

ILLUSTRATION 5 : Find the standardised death rate by direct and indirect methods for the data given below :

Age	Standard Population		Local Population	
	Population in '000	Specific Death Rate	Population in '000	Specific Death Rate
0 - 5	8	50	12	48
5 - 15	10	15	13	14
15 - 50	27	10	15	9
50 and above	5	60	10	59

Solution : Computation of STDR by direct and Indirect methods.

Age	**Standard Population**			**Local Population**				
	P_x^s	m_x^s	$m_x^s \cdot P_x^s$	P_x^l	M_x^l	$P_x^l \cdot M_x^l$	$m_x^a P_x^s$	$m_x^l P_x^a$
0 - 5	8000	50	400000	12000	48	576000	384000	600000
5 - 15	10000	15	150000	13000	14	140000	140000	195000
15 - 50	27000	10	270000	15000	9	135000	243000	150000
50-above	5000	60	300000	10000	59	295000	295000	600000
Total	50000		1120000	50000		1483000	1062000	1545000

Direct Method :

$$(STDR)_A = \frac{\sum_l m_x^l P_x^s}{\sum_x P_x} = \frac{1062000}{50000} = 21.24$$

Indirect Method :

$$(CDR)_L = \frac{\Sigma m_x^L P_x^L}{\sum_x P_x^L} = \frac{1483000}{50000} = 29.66$$

Adjustment Factor =

$$\hat{C} = \frac{\Sigma m_x^{\,s}\, P_x^{\,s}}{\Sigma P_x^{\,s}} \times \frac{\Sigma P_x^{\,l}}{\Sigma m^{s} x\, P_x^{\,l}}$$

$$= \frac{1120000}{50000} \times \frac{50000}{1545000} = 0.7249$$

$$\therefore (\text{STDR})_A = (\text{CDR})_A \times \hat{C}$$

$$= 0.7429 \times 29.66$$

$$= 21.5005$$

10.6.5 INFANT MORTALITY RATE :

It is a kind of size specific rate, where age group is taken as (0 - 1) years. Generally, the child of this age is known as "Infant". Also, the probability of death in this age group is higher. Hence, it is necessary to know the death rate of this age group to formulate a reasonable policy to reduce this. This is calculated using the following formula :

$$\text{Infant Mortality Rate} = \frac{\text{No. of deaths below age one year}}{\text{No.of children born in the same year in the same region}} \times 1000$$

10.7 EXERCISES

1. Define vital statistics. What is the importance of these statistics ?
2. Explain the methods of estimating vital statistics.
3. Explain crude and standardised death rates. In what way standardised death rates superior to crude death rates ? Give briefly the direct and indirect method of finding standardised death rates.
4. Distinguish between crude death rates and specific death rates. Explain the purpose and procedure of standardising death rates.
5. Calculate the crude and standardised death rates from the following data.

Age group (years)	Population	Death rate per 1000	Standard age distribution
0 - 10	400	40	600
10 - 20	1500	4	1000
20 - 60	2400	10	3000
Above 60	700	30	400

6. From the following data compute specific death rate for group each and from these values calculate crude death rates.

Age group :	Below 5	5 - 15	15 - 35	35 - 55	Above 55
Population ('000) :	7	22	14	5	2
No. of deaths :	200	350	160	50	100

7. What is meant by standardised rate ? Calculate the standardised death rate (with respect to age) using the following data.

Age Interval :	0 - 2	2 - 5	5 - 15	40 - 60	Above 60
Specific death rates :	375	210	85	41	93
Standard Population size :	15	10	12	38	15

8. Compute the crude and standardised death rates of the two populations A and B regarding A as standard population, from the following data.

Age group	**A**		**B**	
(Years)	**Population**	**Deaths**	**Population**	**Deaths**
Under 10	20,000	600	12,000	372
10 - 20	12,000	240	30,000	660
20 - 40	50,000	1250	62,000	1,612
40 - 60	30,000	1050	15,000	325
Above 60	10,000	500	3,000	180

9. Calculate C.D.R. and STDR's of local population from the following data. Compare them with the C.D.R. of the standard population.

Age in	**Standard Population**		**Local Population**	
Years	**Population**	**Deaths**	**Population**	**Deaths**
0 - 10	600	18	400	16
10 - 20	1000	5	1500	6
20 - 60	3000	24	2400	24
Above 60	400	20	700	21

10.8 SUMMARY

Vital statistics mean data pertaining to the vital events of a population under references especially with regard to births, deaths, marriages, health, migrations etc., In present times, vital statistics not only consider the number of people but also the quality of human life Vital statistics maintains the records of marriage, divorce, adoption, legitimation, recognition, annulment and legal separation, live births, deaths, foetal deaths, morbidity. It is a part of demography.

Thanks to the rapid advancement of science and technology in respect of health and medical sciences, together with improvement in standard of living, there has been a considerable decline in the mortality rate all over the world. But birth rates are not been declining in the same proportion. This has resulted in the unprecedented growth in world population and is posing a threat to its survival. Consequently, the world society is actively involved in taking necessary steps to maintain the population equilibrium by way of reducing the birth rate.

LESSON - 11

MEASUREMENT OF FERTILITY AND POPULATION PROJECTION

LEARNING OBJECTIVES

On completion of this lesson you should be able to :

- □ Understand the meaning of Fertility rate and Measurement of Fertility. Studying their merits and demerits. We can observe the increase in population due to births in a specified time period.
- □ We can study the growth of population by using the technique of measurement of population.

LESSON OUTILINE

11.1 INTRODUCTION

Fertility rates are the barometers to indicate the increase in population due to births in a specified period, usually a year.

The rates are expressed per thousand women of child bearing age (15 - 50 years). In 1981 census of our country questions relating to fertility were canvassed. The questions are (1) the age at marriage (2) number of surviving children, (3) number of children born last year. This (4) question was canvassed incase of newly married women.

To study the behaviour of the population of a country in a given period of time, we have to count the number of live births. The number of births depend on fertility rate which refers to the actual birth of children in a particular region or country during a given period.

The countries in the world are being divided into two distinct fertility groups high and low. The high fertility developing countries, that account for about two-thirds of the world's population, have birth rates usually above 30 per thousand. The low fertility developed countries have birth rates below 30 per thousand.

11.2 FERTILITY RATES

Fertility is a "vital event" which determines the population, distribution and growth rate. The fertility rate may be defined as, "The number of births to the women of child bearing or reproductive age as against the total population".

In the words of W.G. Barclay, the fundamental motion of fertility is an actual level of performance in a population, based on the numbers of "live births" that occur.

In demography, the word fertility is used in relation to the actual number of children on occurance of births, specially live births". Fertility must be distinguished from fecundity which refers to the capacity to bear children. Infact, fecundity provides an upper limit for fertility. To measure the growth of population, various fertility rates are computed.

GENERAL FACTORS ASSOCIATED WITH LOW FERTILITY :

The list of causes which resulted in the declion of Birth rate in the developed world are given below :

1. Decreasing mortality (especially infant mortality) requiring fewer births for the same number of grown-up children.
2. Industrialisation and the division of labour generate a much more complex social structure and completely new opportunities for social promotion and movement.
3. Growing urbanisation with increasing facilities for communication and exchange and differential penalties of various types for large families.
4. Shift of functions from the family unit to the other units.
5. Increasing participation of women in professional activities.
6. Development of secular, rational attitudes or of a new kind of hedonism.
7. Availability of contraceptives & intraceptives.
8. The possible lessening of biological potentialities etc.

11.3 MEASURES OF FERTILITY RATES

Some of the important rates which are computed only to have an idea about the fertility. The following are the important fertility rates.

1. Crude Birth Rate (C.B.R)
2. General Fertility Rate (G.F.R.)
3. Age specific Fertility Rate (ASFR)
4. General Marital Fertility Rate (GMFR)

5. Age specific marital Fertility Rate (ASMFR)
6. Total marital Fertility Rate (TMFR)
7. Total Fertility Rate (TFR)

11.3.1 Crude Birth (Fertility) Rate (C.B.R.) : It is the easiest of all the measures of fertility. It relates to the number of live births in the total population in a region or country during a given period of time. It indicates the speed or rate at which the population of a country is advancing. It is computed with the following expression.

$$\text{Crude Birth Rate} = \frac{\text{Total number of live-births in a country during a given period}}{\text{Total population of the given region}} \times 1000$$

$$= \frac{\Sigma B^t}{\Sigma P^t} \times 1000$$

where $B^t =$ Total number of live births in the given region or locality during a given period, say t.

$P^t =$ Total population of the given region during the period t.

MERITS :

1. It is simple to understand and easy to calculate.

DEMERITS :

1. The crude birth rate, though simple is unreliable because, it ignores the age and sex distribution of population.
2. C.B.R. is not a probability ratio, since the whole population P^t cannot be regarded as exposed to the risk of producing children. In fact, only the females and only who are of child bearing age group are exposed to the risk of producing children. But risk varies from one group to another, because a woman under 30 is certainly at greater risk as compared to a woman over 40.
3. The child bearing age-groups are not identical in all the countries, it may vary due to climatic conditions. In tropical countries, the period starts, or apparently at an earlier date than in cold weather countries. So C.B.R. does not help us to compare fertility situations in different countries.
4. C.B.R. shows that women in all the ages have the same fertility conditions which is not true since younger women have, in general higher fertility than older women.
5. The C.B.R. is determined by a number of factors such as age and sex distribution of the population, fertiity of the population, sex ratio, marriage rate, migration, family-planning measures and so on. Thus relatively high C.B.R. may be observed in population, with favourable age and sex structure even though fertility is low.

LIMITATIONS :

1. C.B.R. doesn't throw light upon the age-structure and sex-composition of the population. For example, two countries with the same population may have different crude birth rates because of variation in age structure and sex-composition. A country having large proportion of population C.B.R.in the age group 20-45 may have high than the proportion of population of the country in the age group's of 5 - 15 and 45 - 70.
2. Since the total population refers to male and female the crude birth rate is not a reliable and accurate measure of fertility. The crude birth rate is affected by the following factors.
   (i) The age structure of population.
   (ii) The sex composition of the population
   (iii) The marriage age
   (iv) The fertility of the population.

11.3.2 GENERAL FERTILITY RAE (G.F.R.)

General fertility rate gives the rate of births per thousand women of child bearing age (15 - 49) of a country or region in a given year with-out giving any cognizance to any other factor. G.F.R. is obtained as the annual number of births divided by the total number of females in child bearing age multiplied by 1000.

$$\text{G.F.R.} = \frac{\text{Total number of live-births}}{\text{Total number of females in child bearing-age}} \times 1000$$

Thus, general fertility rate may be defined as the number of babies per 1000 women in the reproductive age group.

Remark : Generally the child bearing-age is 15 - 49. Thus births to females out-side this range are very rare. Such births, if any, are recorded separately and are included in the age group below 15 and above 49 respectively.

MERITS :

1. General fertility rate is a probability rate since the denominator of G.F.R. formula consists entirely of female population which is exposed to the child bearing-age.
2. General fertility rate increases the population through live births.

DEMERITS :

1. G.F.R. over looks the age composition of the female population in the child bearing-age. It suffers from the draw back of non-comparability in respect of time and country.
2. The general fertility rate suffers from one main limitation. It ignores fecundity of women which varies according to age-groups.

11.3.3 AGE-SPECIFIC FERTILITY RATE (ASFR) :

The general fertility rate is a better measure of population growth than crude birth rate. G.F.R. gives only a general trend of the fertility rate of child bearing age. The capacity to bear a child is not uniform through out the child bearing age of 15 - 50. Therefore, it is essential to compute the fertility rate for each age-group of child bearing-age. This rate is called Age Specific Fertility Rate and is computed by the formula :

$$\text{A.S.F.R.} = \frac{\text{No. of live births to the women of age group x to (x+n) during a year}}{\text{Average No. of women in the same age group during that year}}$$

The age specific fertility rate is useful because :

(i) It reveals the pattern of the rate of child bearing through out all ages and the extent of differences in fertility and between the age groups.

(ii) It permits the study of fertility in terms of cohorts of women, tracing their fertility as they pass through life.

(iii) It is the basis on which other measures of fertility like general fertility rate, gross reproduction rate are computed.

11.3.4 GENERAL MARITAL FERTILITY RATE (GMFR) :

The births are compared to married women only barring expectation of births to unmarried or widow women. Hence, for family planning, emphasis is laid down to married women only. For this, it appears logical to study the fertility rates among married women only.

General marital fertility rate can be defined as the number of off springs born alive during a period (usually a year) per thousand married women of child bearing age. It can be formulated as:

$$\text{GMFR} = \frac{\text{No. of births during a year to married women}}{\text{Mid-year population of married women during that year of age 15-49 years}} \times 1000$$

G.M.F..R. does not provide classified fertility rates of married women, particularly, in respect of age. Hence, the information through G.M.F.R. is not complete.

11.3.5 AGE SPECIFIC MARITAL FERTILITY RATES (ASMFR) :

It is known that fecundity varies with age. For example, in Indian women fecundity is maximum in the age group 25 - 29 and is little less in the age-group 20 - 24. Also after 29 year of age it goes down. Hence, for the success of family planning schemes, it becomes necessary to study Age specific marital fertility rate.

A.S.M.F.R. is defined as, the number of children born alive during a given period (usually a calender year) per thousand married women of a particular age or age-group. As a formula

$$\text{A.S.M.F.R.} = \frac{\text{No. of live births to married women in the age-group}}{\text{Average population of married women during that year in the same age group}}$$

11.3.6 TOTAL MARITAL FERTILITY RATE (T.M.F.R.) :

Total marital fertility rate gives the total number of live births that would have taken place per thousand married women, had the current schedule of age-specific marital fertility rates been been applicable for entire child bearing period. For the age group x to $x + n$ years, the total marital fertility rate,

$$T.M.F.R. = A.S.M.F.R. \times n$$

Where, n is the interval in years.

11.3.6 TOTAL FERTILITY RATE :

Total fertility rate is a measure which gives approximately the magnitude of complete family size that the total number of children, a women would bear on an average in her life time assuming no mortality and no adoption of family planning measures. Total fertility rate is computed by adding the age specfic fertility rates of various age-groups of child bearing age. This indicates the number of children which a women expects to deliver in the child bearing age i.e. from 15 to 50 years on the assumption that she would be subjected to the given fertility condition and none of these women die before passing the child bearing age. It is computed by

$$T.F.R = \Sigma\left(A.S.F.R\right) \times C$$

Where C stands for the number of years in an age-group.

Illustration 1 : From the following information of a town of 400000 population in 1991 compute (i) C.B.R. (ii) G.F.R. (iii) S.F.R. (iv) T.F.R.

Age-group	Female Population ('000) $\left(P_f^+\right)$	Live - births $\left(B^+\right)$
15 - 19	17	340
20 - 24	18	1980
25 - 29	20	2900
30 - 34	15	1500
35 - 39	12	840
40 - 44	10	400
45 - 49	8	40
Total	100	8000

Solution :

(i) $$\text{Crude Birth Rate (C.B.R.)} = \frac{\text{Total number of live births in the town during a given period}}{\text{Total population of the given region}} \times 1000$$

$$= \frac{\Sigma B^t}{\Sigma P^t} \times 1000$$

$$= \frac{8000}{4000000} \times 1000 = 20$$

C.B.R. = 20 per thousand.

(ii) $$\text{G.F.R.} = \frac{\text{Total number of live-births}}{\text{Total number of females in child bearing-age}} \times 1000$$

$$= \frac{\Sigma B^t}{\sum_{15}^{49} P_f^t} \times 1000 = \frac{8000}{100000} \times 1000 = 80$$

$\therefore$ G.F.R. = 80 per thousand.

(iii) Computation of specific fertility Rate

Age-group	**Female Population** $\left(P_f^t\right)$	**Live-birth** B^t	**Age-specific fertilities** **Rate,** $\frac{B^t}{P_f^t} \times 1000$
15 - 19	17000	340	20
20 - 24	18000	1980	110
25 - 29	20000	2900	145
30 - 34	15000	1500	100
35 - 39	12000	840	70
40 - 44	10000	400	40
45 - 49	8000	40	5
Total	**1,00,000**	**8,000**	**490**

(iii) Age Specific Rates :

$$\text{S.F.R.} = \frac{B_X^t}{f_{P_x}} \times 1000$$

Thus $\text{S.F.R.}(15-19) = \frac{340}{17000} \times 1000 = 20$ per thousand.

Similarly for other age groups, it is given in the last column of the table.

(iv) Total Fertility Rate (T.F.R.)

$$\text{T.F.R.} = \sum_{15}^{49} \text{S.F.R.} \times \text{C}$$

$= 490 \times 5 = 2450$ per thousand.

This presumis that there is no death of any woman (15 - 49) years, then 1000 females will give birth to 2450 babies or 2.45 babies per woman.

Illustration 2 : The female populations and live births with the age of mother, in 1979, of a country, are given below. Calculate general fertility rates and age specific fertility rates.

Age-Group	**Female population**	**Live-births**
15 - 19	1,16,410	10,468
20 - 24	1,13,610	16,983
25 - 29	1,02,930	12,522
30 - 34	93,300	7,083
35 - 39	73,920	3,456
40 - 44	62,700	1,140
Total	5,62,870	51,652

Solution : Calculation of G.F.R. and A.S.F.R.

Age-group (years)	Female populations (P_f)	Live-births $\left(B^t\right)$	Age Specific Fertility Rates (ASFR) $= \frac{{}^nB_x^t}{{}^nP_{f_x}} \times 1000$
15 - 19	1,16,410	10,468	89.92%
20 - 24	1,13,610	16,983	149.48%
25 - 29	1,02,930	12,522	121.65%
30 - 34	93,300	7,083	75.92%
35 - 39	73,920	3,456	46.75%
40 - 44	62,700	1,140	18.18%
Total	5,62,870	51,652	

The General Fertility Rate (G.F.R.)

$$\left(\text{G.F.R.}\right)=\frac{\Sigma B^{t}}{\Sigma P_{f}}\times 1000$$

$$=\frac{51,652}{5,62,870}\times 1000=91.76$$

Age Specific Fertility Rate (ASFR)

$$\text{A.S.F.R.}=\frac{\text{No. of live births to the women of age group x to (x+n) during a year}}{\text{Average No. of women in the same age group during that year}}$$

Thus $\text{A.S.F.R.}\ 15-19=\frac{10,468}{1,16,40}\times 1000=89.92$

Similarly for other age-groups, given in the last column gives the above ASFR.

Illustration 3 : Compute (i) G.F.R. (ii) S.F.R. (iii) Total Fertility Rate.

Age-group child bearing	:	15-19	20-24	25-29	30-34	35-39	40-44	45-49
Females								
Number of women ('000)	:	16.0	16.4	15.8	15.2	14.8	15.0	14.5
Total Births	:	260	2244	1894	1320	916	280	145

Solution : Computation of G.F.R., A.S.F.R & T.F.R.

Age	**Number of women**	**Number of births**	**Age-specific fertility rate**
	$\left(nP_{fx}\right)$ **'000**	nB_{x}^{t}	$=\frac{nB_{x}^{t}}{nPf_{x}}\times 1000$
15 - 19	16.0	260	15
20 - 24	16.4	2244	137
25 - 29	15.8	1894	116
30 - 34	15.2	1320	86
35 - 39	14.8	916	62
40 - 44	15.0	280	19
45 - 49	14.5	145	10
Total	107.7	7059	446

(i) General Fertility Rate (G.F.R.)

$$G.F.R. = \frac{\Sigma^n B_x^t}{\Sigma^n Pf_x} \times 1000 = \frac{7059}{107700} \times 1000 = 65.55$$

$G.F.R. = 65.55$ per thousand.

(iii) Total Fertility Rate :

$$T.F.R. = \sum_{15}^{49} S.F.R \times C = 446 \times 5 = 2230$$

$T.F.R. = 2230$ per thousand.

11.4 MEASUREMENT OF POPULATION GROWTH :

Population growth mainly depends on the sex of the newly born children. The number of births in a population depends upon the number of women in a reproductive age. Thus, fertility rates fail to reflect on the rate of population growth because they do not take into consideration the sex of the newly born children. Therefore, to measure the rates of population growth, it becomes necessary to take into account the female births and their mortality before they reach the child bearing age. The rate of population growth can be measured interms of reproduction rates. There are two indices of reproduction.

1. Gross reproduction rate (GRR)
2. Net reproduction rate (NRR)

1. Gross Reproduction Rate : The rate of population growth is dependent on the number of female births that are actually the future mothers. Thus total fertility rate can be calculated interms of female births restricting the birth in the age specific fertility rates of female births.

The demographic year book published by United Nations in 1954, has defined GRR as, "The gross reproduction rate indicates the average number of daughters who would be born to a group of girls, beginning life together in a population where none died before the upper limit of child bearing age and where the given set of fertility rates was in operation".

The gross reproduction rate is based on the following assumptions :

1. No female children die till they attain the upper limit of child bearing age which is 49 years in India.
2. Female population remains stagnant inspite of migration.
3. The age specific fertility rate currently prevailing will continue to operate throughout the child bearing age (period).

The gross reproductive rate may be computed by the following expression.

$$G.R.R. = \sum_{x=15}^{50} f_x \frac{B_f}{B}$$

where f_x denotes age specific fertility rate.

B_f denotes female births

B denotes total number of births.

Here male children are excluded in the calculation of G.R.F. because there is no upper limit for males producing a child.

MERITS :

G.R.R. posseses two advantages :

(i) The G.R.R. may be used to compare the fertility of two or more regions.

(ii) It may also be used for comparing the fertility in the same region.

LIMITATIONS :

It is exposed to the following limitations :

1. In computation of gross reproduction rate, we have not considered the mortality. All newly born female children do not survive upto child bearing age.
2. It is based on the current rate of fertility which may undergo change with the passage of time.
3. There may be un registered number of births and wrong statements of age at the time of registration.

2. Net Reproduction Rate (N.R.R.) : Though gross reproduction rate gives an idea about the growth of the population it excludes the effect of mortality of the birth rate. But in the case of net reproduction Rate both fertility and mortality are taken into account.

The net repdocution rate indicates the number of daughters that would be born to female population in the child bearing age on the assumption that they are subject to same fertility and mortality rates through out the child bearing age.

Suppose that 10,000 mothers give birth to a certain number of female babies that do not live upto child bearing age i.e., some die during infant stage, and same during the child bearing age. It is also certain that some may not marry, some may become widows in the very first year, some may not be conceive the child through the child bearing age and it is only the balance who pass through fertility period and add to the population growth. Thus, the net reproduction rate represents the rate of replacement of that population.

The net reproduction rate, as defined in the demographic year book, United Nations, 1954 is, "The Net Reproduction rate may be interpreted as the average number of daughters that would be produced by women through out their life-time if they were exposed at each age to the fertility and mortality rates on which the calculation is based".

The Net reproduction rate is computed by the formula :

$$\text{N.R.R.} = \sum_{x=15}^{50} f_x \cdot L_x \frac{B_f}{B}$$

where L_x is the probability of female population surviving upto age x.

F_x stands for age specific fertility rate.

B_f stands for female births.

B stands for total live births.

NRR cannot exceed GRR because it doesn't ignore mortality rate among newly born females.

MERITS :

1. The NRR may be used to compare the fertility of two regions to study the behaviour of population.
2. It takes mortality rate into consideration.

DEMERITS :

1. N.R.R. are based on the constant rates of fertility and mortality over generation which is not true to the real life phenomenon.
2. N.R.R. do not take into account the number of emigrants or immigrants. Many times, the number is also large that affect the reproductive rate.
3. A.S.F.R. are also so constant as they are taken to be for the purpose of N.R.R.

INTERPRETATIONS :

1. Normally N.R.R varies from 0 to 5.
2. N.R.R. cannot exceed G.R.R. i.e. $\text{N.R.R.} \leq \text{G.R.R.}$
3. N.R.R. will be equal to G.R.R. If all the newly born female children survive till their maximum child bearing age.
4. If N.R.R. = 1, the female population will exactly replace itself into new generation and population remains constant.
5. If $\text{N.R.R.} < 1$. This will result into the reduction in the number of mothers and will thus cause reduction in population.
6. If $\text{N.R.R.} > 1$, there will be a number of mothers in the next generation which will tend to increase the population.

The N.R.R. may also be computed by other formula.

$$\text{N.R.R.} = \Sigma \frac{\left(\text{No. of female births} \times \text{Survival Rate}\right)}{1000}$$

$$= \frac{\text{Total number of female children born and survived}}{1000}$$

$$N.R.R. = \Sigma\left(A.S.F.R. \times S.R.\right) \times C$$

where C stands for no. of years in age-group.

ILLUSTRATION 5.1 :

From the data given below, calculate the G.R.R and N.R.R.

Age-group	Number of children born to 1,000 women passing through the age-group	Mortality Rate (per 1000)
16 - 20	150	120
21 - 25	1500	180
26 - 30	2000	150
31 - 35	800	200
36 - 40	500	220
41 - 45	200	230
46 - 50	100	250

Sex Ratio being males : Females 54 : 48.

Solution : Calculation of G.R.R and N.R.R.

Age-group	Number of children born to 1000 women passing through the age group	Number of female children f_{Bx} = 48% of (2)	Survival Rate (S.R.) = 1 - Mortality Rate	Number of female children survived $f_{Bx} \times S.R.$
(1)	(2)	(3)	(4)	(5)
16 - 20	150	$\frac{150 \times 48}{100} = 720$	1 - 0.12 = 0.88	63.36
21 - 25	1500	$\frac{1500 \times 48}{100} = 720$	1 - 0.18 = 0.82	604.80
26 - 30	2000	$\frac{2000 \times 48}{100} = 960$	1 - 0.15 = 0.85	806.00
31 - 35	800	$\frac{800 \times 48}{100} = 384$	1 - 0.20 = 0.80	307.20

36 - 40	500	$\frac{500\times 48}{100}=240$	1 - 0.22 = 0.78	187.20
41 - 45	200	$\frac{200\times 48}{100}=96$	$1-0.23=0.77$	73.92
46 - 50	100	$\frac{100\times 48}{100}=48$	$1-0.25=0.75$	36.00
Total		2520		2078.48

$\therefore$ Gross reproduction Rate per woman (GRR)

$$= \frac{\text{Total number of children born}}{1000}$$

$$\text{G.R.R.} = \frac{2520}{1000} = 2.52$$

Net Reproduction Rate (N.R.R.) per woman

$$= \frac{\text{Total number of female children born and survived to 1000 women}}{1000} = \frac{2078.84}{1000} = 2.08$$

N.R.R. = 2.08

ILLUSTRATION 11.5.2 :

The population and its distribution by sex and number of births in a year 1991 and survival rates are given in the following table. Compute G.R.R. and N.R.R.

Group	Population	Female Births	Survival Rate
15 - 19	11832	60	0.91
20 - 24	10538	132	0.90
25 - 29	9375	127	0.84
30 - 34	7843	81	0.87
35 - 39	7270	56	0.85
40 - 44	6315	15	0.83
45 - 49	5394	3	0.82

Solution : Calculation of G.R.R. & N.R.R.

ILLUSTRATION 11.5.3 : Compute gross reproduction Rate and net reproduction rate from the data given below.

Age-group	Female Population	Female births	Survival Rate
15 - 19	10,000	300	0.90
20 - 24	9,000	630	0.89
25 - 29	8000	480	0.88
30 - 34	7000	280	0.87
40 - 44	5000	35	0.83

Solution : Calculation of GRR & NRR

Age group	Female population	Female births	A.S.F.R per women	Survival rate (S.R.)	$A.S.F.R. \times S.R$
15 - 19	10,000	300	$\frac{300}{10000} = 0.030$	0.90	0.02700
20 - 24	9000	630	$\frac{630}{9000} = 0.070$	0.89	0.06230
25 - 29	8000	480	$\frac{480}{8000} = 0.060$	0.88	0.05280
30 - 34	7000	280	$\frac{280}{7000} = 0.040$	0.87	0.03480
35 - 39	6000	150	$\frac{150}{6000} = 0.025$	0.85	0.02125
40 - 44	5000	35	$\frac{35}{5000} = 0.007$	0.83	0.00581
Total			**0.232**		**0.20396**

$$G.R.R. = \Sigma(A.S.F.R.) \times C$$

$$= 0.232 \times 5 \qquad (\because C = 5)$$

$$G.R.R. = 1.160$$

$$N.R.R. = \Sigma(A.S.F.R. \times S.R.) \times C$$

$$= 0.20396 \times 5 = 1.01980$$

Illustration 11.5.4 : Calculate the gross and net reproductive rates from the data given below :

Age-group	Female Population	Female births	Survival Rate
15 - 19	1399	15133	0.9694
20 - 24	1422	94155	0.9668
25 - 29	1521	102676	0.9632
30 - 34	1726	72490	0.9584
35 - 39	1421	31402	0.9524
40 - 44	1689	10640	0.9524
45 - 49	1667	700	0.9279

Solution : Computation of G.R.R. and N.R.R.

Age group	Female population	Female births	A.S.F.R per women	Survival rate (S.R.)	A.S.F.R. × S.R
	P_f	B_f	$= B_f \div P_f$		S.F.R.×S.R.
15 - 19	1399000	15133	0.01082	0.9694	0.01049
20 - 24	1422000	94155	0.06621	0.9668	0.06401
25 - 29	1521000	102676	0.06751	0.9632	0.06503
30 - 34	175600	72490	0.04128	0.9584	0.03596
35 - 39	1451000	31402	0.02164	0.9519	0.02060
40 - 44	1689000	10640	0.00632	0.9424	0.00594
45 - 49	1667000	700	0.00042	0.9279	0.00039
Total	10965000	327196	0.21418		0.20602

$$\text{G.R.R.} = \sum_{x=15} \frac{B_f}{P_f} \times 5 = 0.21418 \times 5 = 1.0709$$

$$\text{N.R.R.} = \sum_{x=15} \left(\frac{B_f}{P_f} \times 5 \right) \times 5 = 0.20602 \times 5 = 1.0301$$

Hence, population is growing.

THEORETICAL EXERCISES

1. Explain the various measures of Fertility used in vital statistics.
2. Describes the various measures of fertility rates and state their limitations.
3. Explain the concepts of Gross reproduction and net reproduction rates.
4. Differentiate between the following.
   (a) Gross Reproduction Rate and Net Reproduction Rate
   (b) Age Specific Fertility Rate and Total Fertiltiy Rate
5. Distinguish between
   (i) General fertility rate (ii) age-specific fertility rate (iii) Total fertility Rate.

11.6 NUMERICAL EXERCISE

1. From the following figures, calculate the gross and net reproduction rates.

Age-group (years)	No. of female children borned to 1000 women	Percent survival rate rate
15 - 19	220	85
20 - 24	365	80
25 - 29	300	70
30 - 34	188	65
35 - 39	115	60
40 - 44	42	50
45 - 49	5	45

2. The following data pertain to the female population and number of live births in different age groups of reproductive age. Calculate general and specific fertility rates.

Age-group :	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39	40 - 44	Total
Female Population :	20,240	24,565	21,138	18,319	15,661	13,045	1,12958
No. of live births :	2046	6313	5792	4048	2380	1094	24673

3. From the given data below, calculate the gross reproduction rate, assuming that for the given population, the ratio of female babies total birth is 48.8%.

Age-group :	16 - 20	21 - 25	26 - 30	31 - 35	36 - 40	41 - 46	46 - 50
Fertility Rate per : 1000 women	17	173	253	201	157	67	9

4. Calculate (i) the gross reproduction rate (ii) the net reproduction from the data given below

Age (years)	No. of women ('000)	No. of female births in the age group	No. of total births to women in the age group	Survival Rate
15 - 19	16.0	140	260	0.669
20 - 24	16.4	1130	2240	0.967
25 - 29	15.8	980	1894	0.963
30 - 34	15.2	670	1320	0.958
35 - 39	14.8	460	916	0.952
40 - 44	50.0	150	280	0.942
45 - 49	14.5	80	145	0.928

11.7 SUMMARY

Fertility is measured interms of number of births per 1000 women. Measures of fertility are made on the basis of data from the census, vital registrations and sample surveys. Measures of fertility are basically of two-types -period measures and cohort measures. Period measures include birth rate, general fertility rate; child women ratio and age-specific fertility rate. Cohort methods include total fertility rate, gross reproduction rate and net reproduction rate.

Measures of fertility are useful to understand fertility levels in a country and also compare fertility levels between two or more populations.

Thus, the factors which effect fertility are social, religious and economic. These factors are responsible for births high fertility or low fertility in almost all societies, and in developed and under developed countries. In a developing country like India fertility differentials have become increasingly distinct because at present the Indian society is in a transitional stage of declining fertility. Therefore, various social groups exhibit fertility differentials which have become increasingly more pronounced.

11.8 REFERENCE BOOKS

Basic Statistics	:	B.L. Agarwal
Fundamentals of Statistics	:	S.C. Srivasthava, Sangya Srivasthava
Methods of Statistics	:	P.S. Grewal
Applied Statistics	:	S.C. Gupta & V.K. Kapoor
Methods in Bio-statistics	:	B.K. Mahajan

LESSON - 12

LIFE TABLE & ABRIDGED LIFE TABLE

LEARNING OBJECTIVES

Upon the completion of this lesson, you should be able :

* To describe mortality in a population.
* To exhibit the numbers, living and dying at each age on the basis of experience of a cohort (life history).
* To study the probability of dying and living separately at each age.
* To provide the expectation of life at any age in a tabular form.
* To study the Abridged life table to overcome the problems faced in the construction of complete life table.

LESSON OUTILINE

12.1 INTRODUCTION

A life table is composed of a set of values showing how a group of infants born on the same day and living under similar conditions would gradually die out. In simple words, a life-table provides conveniently summaries of the mortality or longevity of any *cohort.

George Berclay expressed life-table as, "The table is a life history of a hypothetical group or cohort or people, as it is diminished gradually by deaths. The record begins at birth of each number

Cohort : Life history of hypothetical group or population.

and continues until all have died". However, life-tables can be constructed for males and females separately, for different segments of the country, for different causes of death which affects the hypothetical population.

Life-table in the real scense came into existence and popularity at the end of 18th century for the purpose of life assurance. This helps in actuaries to calculate the insurance risks and the premium rate. In recent years, life-table technique being increasingly applied to follow up studies of chronic diseases of hospital patients chalking out welfare programmes of different cohorts, etc.,

12.2 ASSUMPTION OF LIFE-TABLE

Certain assumptions made while constructing a life-table are given below :

i) Mortality rates are not the same for all age groups. Hence, life-tables utilise only the age specific mortality rates.

ii) The deaths are distributed uniformly over the period $(x, x+1, x+2)$ for each x (except for the first years).

iii) The life-tables for males and females are constructed separately, because mortality rates differ significantly for males and females.

iv) There is no effect of immigrations and emigration on the cohort. It means that the reduction in number of the initial cohort is merely due to deaths.

v) The deaths recorded for a cohort through any method are correct without any errors and omissions.

vi) The cohort originates from some standard number of births, say 10,000 or 1,00,000 which is called the radix of the table.

12.3 USES OF LIFE-TABLES

Primarily, life table was developed to meet the needs of life assurance offices. A life-table mainly displays the death rate between two consecutive days and expectation of life at any age of a cohort. The information revealed by a life-table has applications in many fields. Today lifetable is widely accepted as an important basic material in demographic and public health studies. Some of the specific uses of life tables are :

(i) Life tables are of maximum utility to actuaries to work out the rate of premium for persons of different age groups.

(ii) Population projections may be constructed by age and sex with the help of life-tables.

(iii) Life-table clearly shows the distribution of people according to sex which helps a great deal in planning of education, employment and work force :

(iv) The life-table helps to assess the accuracy of census figures, death and birth registrations.

(v) The computation of net reproduction rate and true rate of natural increase in population is easily predicted by the use of life tables.

(vi) It helps to assess the impact of family planning programmes on population growth.

(vii) To find expectation of life or longevity of life at birth or any other age. Increase in longevity of life means reduction in mortality. Thus, life table is another source to information compare mortality two places, periods, professions or groups.

(viii) How far, the better medical aid, high standard of living and new scientific inventions have increased the span of life can be evaluated through life tables.

(ix) The demographers may use the life table techniques for other types of demographic data like in the computation of probability of marrying, specific age and sex, on the basis of census data classified by marital status.

However, the accuracy and usefulness of life tables depends mainly upon the accuracy and completness of the registration of deaths and of the enumeration of population at census. Deficiency in registration may be greater than the census enumeration.

12.4 CONSTRUCTION OF LIFE-TABLE

The data used for constructing a life table are census data and death registration data. Life tables are generally constructed for various groups of people, which as experience shows, have different patterns of mortality. Thus the life-tables are constructed for different races, occupational groups and sex. Life tables are constructed on regional basis and other factors according differential mortality.

A life table starts with a convenient cohort size like 1,00,000 or 10,000 known as radix. The record of a life table begins at the birth of each member and continues fill all have died. It is worth pointing out that a life-table diminishes gradually. A life table consists of eight columns as expressed below :

(1) The age in years = x

(2) Persons living at age $x = \ell_x$

(3) Dying between age x and $x+1 = d_x = \ell_x - \ell_{x+1}$

(4) Probability of dying between age x and $x+1 = q_x = \frac{d_x}{\ell_x}$

(5) Probability of surviving between age x and $x + 1 = p_x = 1 - q_x = \frac{\ell_{x+1}}{\ell_x}$.

(6) Living between age x and $x+1 = L_x = \frac{\ell_x + \ell_{x+1}}{2} = \ell_x - \frac{1}{2} d_x$

(7) Living above the $x = T_x = L_x + L_{x+1} + L_{x+2} \cdots\cdots = L_x + T_{x+1}$

(8) Expectation of life at age $x = e_x^0 = \frac{T_x}{\ell_x}$

Each symbol has been explained further to familiarise the current usage :

x – Exact age in years till last birthday.

ℓ_x – The number of survivors at age x out of the initial cohort which is usually taken as 1,00,000.

d_x = The number of deaths in the interval x to $(x+1)$ in the initial cohort. It gives the number of those who could celebrate their xth, birth-day but not the $(x+1)$th birth day.

q_x = It is the Probability of persons dying between the age of x and $x+1$ to the number of persons alive at the age of x, i.e. at the beginning of the interval.

p_x = It gives the probability of persons surviving to the end of the interval, i.e. at age $(x+1)$ years to the number of persons alive at the beginning of the interval. i.e. at age x.

L_x = The number of persons lived between the age x and $(x+1)$ in the hypothetical life table stationary population.

T_x = The number of years lived by the cohort after x years of age; In otherwords Tx denotes the total future years lived by ℓ_x persons of the cohort who have attained age x.

e_x^0 = The average remaining life time. It gives the average number of years a person age x is likely to survive under the existing mortality rate.

Description of a Life Table : A typical life table has generally the following columns.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
x	l_x	d_x	q_x	p_x	L_x	T_x	e_x^0

The following theorems establish the relation between the various quantities defined above.

Theorem 12.4.1 : ${}_np_x = p_x \cdot p_{x+1} \cdot p_{x+2} \cdots p_{x+n-1}$

Proof : We have by definition

$${}_np_x = \frac{\ell_{x+n}}{\ell_x}$$

$$= \frac{\ell_{x+1}}{\ell_x} \cdot \frac{\ell_{x+2}}{\ell_{x+1}} \cdots\cdots \frac{\ell_{x+n}}{\ell_{x+n-1}} = p_x \cdot p_{x+1} \cdot p_{x+2} \cdots\cdots p_{x+n-1}$$

Theorem : 12.4.2

$$_{n}q_{x} = \frac{d_{x+n-1}}{\ell_{x}}$$

Proof : $_{n}q_{x}$ = probability that a person aged x may die in the nth year.

= Probability that a person aged x survives till age $(x+n-1)$ but dies in the age period $(x+n-1, x+n)$.

= $p[\text{A person aged } x+n-1 \text{ dies with in a year}]$ [By compound probability Theorem].

$$= \frac{\ell_{x+n-1}}{\ell_{x}} \times \frac{d_{x+n-1}}{\ell_{x+n-1}} = \frac{d_{x+n-1}}{\ell_{x}}$$

$$_{n}q_{x} = \frac{d_{x+n-1}}{\ell_{x}}$$

Corollary : We have

$$_{n}p_{x} - {}_{n+1}p_{x} = \frac{\ell_{x+n}}{\ell_{x}} - \frac{\ell_{x+n+1}}{\ell_{x}}$$

$$= \frac{\ell_{x+n} - \ell_{x+n+1}}{\ell_{x}} = \frac{d_{x+n}}{\ell_{x}}$$

$$_{n}p_{x} - {}_{n+1}p_{x} = {}_{n+1}q_{x} \quad \left[\because {}_{n}q_{x} = \frac{d_{x+n-1}}{\ell_{x}}\right]$$

Theorem 12.4.3 : If w is the last age at which ℓ_{x} vanishes, i.e., if $\ell_{w} = 0$, then

$$\ell_{w} = \sum_{i=x}^{w-1} d_{i}$$

Proof : $\sum_{i=x}^{w-1} d_{i} = d_{x} + d_{x+1} + d_{x+2} + \cdots\cdots + d_{w-1}$

$$= (\ell_{x} - \ell_{x+1}) + (\ell_{x+1} - \ell_{x+2}) + \cdots\cdots + (\ell_{w-2} - \ell_{w-1}) + (\ell_{w-1} - \ell_{w})$$

$$= \ell_{x} \quad [\because \ell_{w} = 0]$$

Theorem 12.4.4 : $T_x = \frac{1}{2}\ell_x + \ell_{x+1} + \ell_{x+2} + \cdots\cdots$

Proof : By definition

$$T_x = \sum_{t=0}^{\infty} L_{x+t} = \sum_{t=0}^{\infty} \frac{1}{2}\left(\ell_{x+t} + \ell_{x+t+1}\right)$$

$$= \frac{1}{2}\ell_x + \sum_{t=1}^{\infty} \ell_{x+t}$$

$$= \frac{1}{2}\ell_x + \ell_{x+1} + \ell_{x+2} + \ell_{x+3} + \cdots\cdots$$

12.4.5 : $e_x = \left(\sum_{n=1}^{\infty} \frac{\ell_{x+n}}{\ell_x}\right)$

Proof : Since d_{x+i} is the number of persons dying in the age period $(x+i,\ x+i+1)$. i.e. dying in $(i+1)$th year after completion of i years at age x. Thus the total number of years lived by d_{x+i} individuals is given by

$$i \times d_{x+i}\ (i = 0,1,2,\cdots\cdots)$$

Thus $e_x =$ Average number of years survived by individuals of the given age x.

$$= \left(\sum_{i=0}^{\infty} \frac{i\, d_{x+i}}{\ell_x}\right)$$

$$= \frac{1}{\ell_x}\left[d_{x+1} + 2\cdot d_{x+2} + 3\cdot d_{x+3} + \cdots\cdots\right]$$

$$= \frac{1}{\ell_x}\left[\left(\ell_{x+1} - \ell_{x+2}\right) + 2\left(\ell_{x+2} - \ell_{x+3}\right) + 3\left(\ell_{x+3} - \ell_{x+4}\right) + \cdots\cdots\right]$$

$$= \frac{1}{\ell_x}\left[\ell_{x+1} + \ell_{x+2} + \ell_{x+3} + \cdots\cdots\right]$$

$$= \frac{\left(\sum_{n=1}^{\infty} \ell_{x+n}\right)}{\ell_x}$$

Corollary 1 : $e_x^0 = e_x + \frac{1}{2} = \left[\frac{\ell_{x+1} + \ell_{x+2} + \cdots\cdots}{\ell_x}\right] + \frac{1}{2}$

$$= \frac{\left(\frac{1}{2}\ell_x + \ell_{x+1} + \ell_{x+2} + \cdots\cdots\right)}{\ell_x}$$

$$e_x^0 = \frac{T_x}{\ell_x} \qquad \left(\because T_x = \frac{1}{2}\ell_x + \ell_{x+1} + \ell_{x+2} + \cdots\cdots\right)$$

Corollary 2 :

We know

$$e_x = \frac{\ell_{x+1} + \ell_{x+2} + \ell_{x+3} + \cdots\cdots}{\ell_x}$$

$$\Rightarrow \ell_x \cdot e_x = \ell_{x+1} + \ell_{x+2} + \ell_{x+3} + \cdots\cdot$$

$$\therefore \ell_{x+1} \cdot e_{x+1} = \ell_{x+2} + \ell_{x+3} + \ell_{x+4} + \cdots\cdot$$

$$\ell_x \cdot e_x - \ell_{x+1} \cdot e_{x+1} = \ell_{x+1}$$

$$\ell_x e_x = \ell_{x+1}\, e_{x+1} + \ell_{x+1} = \ell_{x+1}(e_{x+1} + 1)$$

$$\frac{\ell_{x+1}}{\ell_x} = \frac{e_x}{e_{x+1} + 1}$$

$$p_x = \frac{e_x}{1 + e_{x+1}}$$

Also, $q_x = 1 - p_x = \frac{1 - (e_x - e_{x+1})}{1 + e_{x+1}}$

12.4.6 : $m_x = \frac{2q_x}{2 - q_x}$

Proof : $m_x = \frac{\text{number of deaths with in age interval x to } (x+1)}{\text{Average } \ell_x \text{ of the cohort in that interval}}$

$$\Rightarrow m_x = \frac{d_x}{L_x} = \frac{d_x}{\ell_x - \frac{1}{2}d_x} \qquad \left[\because L_x = \ell_x - \frac{1}{2}d_x\right]$$

$$= \frac{2\frac{dx}{\ell_x}}{2-\frac{dx}{\ell x}} = \frac{2q_x}{2-q_x} \text{ ---------- (1)}$$

Solving (1) for q_x, we get

$$q_x = \frac{2m_x}{2+m_x}$$

12.5 ILLUSTRATION

12.5.1 Illustration : Given the following table for ℓ_x the number of rabbits living at age x, compute the life-table for rabits :

x :	0	1	2	3	4	5	6
ℓ_x:	100	90	80	75	60	30	0

Solution : The complete life table for the above data is given below :

Age(x)	ℓ_x	$d_x = \ell_x - \ell_{x+1}$	$q_x = \frac{d_x}{\ell_x}$	$L_x = \frac{\ell_x + \ell_x + 1}{2}$	$T_x = \sum_{i=x} L_i$	$e_x^0 = \frac{T_x}{\ell_x}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0	100	10	0.10	95	385	3.85
1	90	10	0.11	85	290	3.22
2	80	5	0.06	77.5	205	2.56
3	75	15	0.20	67.5	127.5	1.70
4	60	30	0.50	45	60	10
5	30	30	1.00	15	15	0.5
6	0	--	--	--	--	--

ILLUSTRATION 12.5.2 :

A life table with two years age returns with certain missing values is presented below.

Age	ℓ_x	d_x	p_x	q_x	L_x	T_x	e_x^0
35	9,345	?	?	?	?	1,63,819	?
36	9,243	149	?	?	?	?	?

Provide missing values in the table.

Solution : The life table has been completed using the relationship of unknown terms with other terms.

$$dx = \ell_x - \ell_{x+1}$$

$$d_{35} = \ell_{35} - \ell_{36} = 9345 - 9243 = 102$$

$$d_{35} = 102$$

$$q_x = \frac{dx}{\ell_x} \Rightarrow q_{35} = \frac{d_{35}}{\ell_{35}} = \frac{102}{9345} = 0.0109$$

$$p_x = 1 - q_x \Rightarrow p_{35} = 1 - q_{35} = 1 - 0.0109 = 0.9891$$

$$p_x = 0.9891$$

$$q_{36} = \frac{d_{36}}{\ell_{36}} = \frac{149}{9243} = 0.0161, \quad p_{36} = 1 - 0.0161 = 0.9839$$

$$L_x = \ell_x - \frac{1}{2}dx \Rightarrow L_{35} = \ell_{35} - \frac{1}{2}d_{35} = 9345 - \frac{1}{2}(102) = 9294$$

$$L_{36} = \ell_{36} - \frac{1}{2}d_{36} = 9243 - \frac{1}{2}(149) = 9168.5$$

$$T_{x+1} = T_x - L_x \Rightarrow T_{36} = T_{35} - L_{35} = 163819 - 9294$$

$$L_{35} = 154525$$

$$e_x^0 = \frac{T_x}{\ell_x} \Rightarrow e_{35}^0 = \frac{T_{35}}{\ell_{35}} = \frac{163819}{9345} = 17.53$$

$$e_{36}^0 = \frac{T_{36}}{\ell_{36}} = \frac{154525}{9243} = 16.72$$

substituting the values of known terms, the completed life table is,

Age(Years)	ℓ_x	d_x	p_x	q_x	L_x	T_x	e_x^0
35	9345	102	0.9891	0.0109	9294	163819	17.53
36	9243	149	0.9839	0.0161	9168.5	154525	16.72

ILLUSTRATION 12.5.3 :

Fill in the blanks which are marked with a query(?) in the following life-table.

Age(Years)	ℓ_x	d_x	q_x	L_x	T_x	e_x^0
30	762227	?	?	?	27296372	?
31	758580	-	-	-	?	?

Solution :

$$d_x = \ell_x - \ell_{x+1} \Rightarrow d_{30} = \ell_{30} - \ell_{31} = 762227 - 758580 = 3647$$

$$q_x = \frac{dx}{\ell_x} \Rightarrow q_{30} = \frac{d_{30}}{\ell_{30}} = \frac{3647}{762227} = 0.00478$$

$$p_x = 1 - q_x \Rightarrow p_{30} = 1 - q_{30} = 1 - 0.00478 = 0.99522$$

$$L_x = \frac{\ell_x + \ell_{x+1}}{2} \Rightarrow L_{30} = \frac{\ell_{30} + \ell_{31}}{2} = \frac{762227 + 758550}{2}$$

$$L_{30} = 760403.5$$

$$T_{x+1} = T_x - L_x \Rightarrow T_{31} = T_{30} - L_{30} = 27296372 - 760403.5$$

$$= 26536328.5$$

$$e_x^0 = \frac{T_x}{\ell_x} \Rightarrow e_{30} = \frac{T_{30}}{\ell_{30}} = \frac{27296732}{762227} = 35.81$$

$$e_{31} = \frac{T_{31}}{\ell_{31}} = \frac{6536328}{2758580} = 34.78$$

Illustration : 12.5.4 : Obtain the missing columns of life table which are marked by question marks

Age(x)	ℓ_x	d_x	q_x	p_x	L_x	T_x	e_x^0
20	693435	2762	?	?	?	35081126	?
21	690673	-	-	-	-	34389077	?

Solution : We shall use various relatioships for computation of missing columns of life-table.

$$q_x = \frac{dx}{\ell_x} \Rightarrow q_{20} = \frac{d_{20}}{\ell_{20}} = \frac{2762}{693435} = 0.00398;\ p_{20} = 1 - 920$$

$$= 1 - 0.00398$$

$$p_{20} = 0.99602$$

$$L_{20} = \ell_{20} - \frac{1}{2}d_{20} = 693435 - \frac{1}{2}2742 = 692054$$

$$e_{20}^{0} = \frac{T_{20}}{\ell_{20}} = \frac{35081126}{693435} = 50.59;\ e_{21} = \frac{T_{21}}{\ell_{21}} = \frac{34389077}{690673} = 49.79$$

Age(Years)	ℓ_x	d_x	q_x	p_x	L_x	T_x	e_x^0
20	693435	2762	0.00398	0.99602	692054	34389072	50.59
21	690673	--	?	?	--	--	--

ILLUSTRATION 12.5.5 :

The table below gives the life table for the rabbits. Find $d_2, p_1, q_2, L_3, e_4^0$

x	0	1	2	3	4	5	6
ℓ_x	100	90	80	75	60	30	0

Solution : $d_x = \ell_x - \ell_{x+1} \Rightarrow d_2 = \ell_2 - \ell_3 = 80 - 75 = 5$

$$q_x = \frac{dx}{\ell_x} = \frac{\ell_x - \ell_{x+1}}{\ell_x}$$

$$q_0 = \frac{\ell_0 - \ell_1}{\ell_0} = \frac{100 - 90}{100} = 0.1$$

$$p_0 = 1 - 90 = 1 - 0.1 = 0.9$$

$$p_1 = 1 - \frac{d_1}{\ell_1} = 1 - \frac{90 - 80}{90} = 0.89$$

$$L_x = \ell_x - \frac{1}{2}d_x \Rightarrow L_3 = \ell_3 - \frac{1}{2}d_3 = \ell_3 - \frac{\ell_3 - \ell_4}{2}$$

$$L_3 = 75 - \left(\frac{75 - 60}{2}\right) = 67.5$$

$$e_x = \frac{2p_x}{1 - p_x} \Rightarrow e_4 = \frac{2p_4}{1 - p_4},\ p_4 = 1 - 94$$

$$p_4 = 1 - \left(\frac{\ell_4 - \ell_5}{\ell_4}\right) = 1 - \left(\frac{60 - 30}{60}\right) = 0.5$$

$$\therefore e_4 = \frac{2p_4}{1-p_4} = \frac{2\times 0.5}{1-0.5} = \frac{1.0}{0.5} = 2$$

$$e_4^0 = e_4 + \frac{1}{2} = 2 + \frac{1}{2} = \frac{5}{2} = 2.5$$

ILLUSTRATION 12.5.6 :

Age in years	ℓ_x	d_x	p_x	q_x	L_x	T_x	e_x^0
7	90000	500	?	?	?	4850000	?
8	?	400	?	?	?	?	?

SOLUTION :

$$\ell_8 - \ell_7 - d_7 = 90,000 - 500 = 89500$$

$$p_7 = \frac{\ell_8}{\ell_7} = \frac{89500}{90000} = 0.9944,\ q_7 = 1 - 0.9944 = 0.0056$$

Also $t_9 = \ell_8 - d_8\ \ 89500 - 400 = 89100$

$$p_8 = \frac{89100}{89500} = 0.9955$$

$$q_8 = 1 - p_8 = 1 - 0.9955 = 0.0045$$

$$L_7 = \frac{\ell_7 + \ell_8}{2} = \frac{90000 + 89500}{2} = 89750$$

$$L_8 = \frac{\ell_8 + \ell_9}{2} = \frac{89500 + 89100}{2} = 98300$$

$$e_7^0 = \frac{T_L}{\ell_7} = \frac{4850000}{90000} = 53.89$$

$$T_8 = T_7 - L_7 = 4850000 - 89750 = 4760250$$

$$\therefore e_8^0 = \frac{T_8}{\ell_8} = \frac{4760250}{859500} = 53.19$$

Age in years	ℓ_x	d_x	p_x	q_x	L_x	T_x	e_x^0
7	90000	500	0.9944	0.0056	89750	4850000	53.89
8	89500	400	0.955	0.0045	98300	4760250	53.19

12.5.7 ILLUSTRATION

Given the returns for the two ages $x = 9$ years and $x + 1 = 10$ years with a few life-table values as $\ell_9 = 75,824$, $\ell_{10} = 75362$, $d_{10} = 418$ and $T_{10} = 49, 53, 195$. Give the complete life-table for two ages of persons :

Solution : Using the relations between different columns of the life-table we get the completed life-table.

$$d_x = \ell_x - \ell_{x+1} \Rightarrow d_9 = \ell_9 - \ell_{10} = 75824 - 75362$$

$$d_9 = 462$$

$$q_x = \frac{d_x}{\ell_x} \Rightarrow q_9 = \frac{d_9}{\ell_9} = \frac{462}{75824} = 0.00609$$

$$q_{10} = \frac{d_{10}}{\ell_{10}} = \frac{418}{75362} = 0.00555$$

$$p_9 = 1 - q_9 = 1 - 0.00609 = 0.99391$$

$$p_{10} = 1 - q_{10} = 1 - 0.00555 = 0.99445$$

$$L_x = \ell_x - \frac{1}{2}dx \Rightarrow L_9 = \ell_9 - \frac{1}{2}d_9 = 75824 - \frac{1}{2} \times 462 = 755963$$

$$L_{10} = \ell_{10} - \frac{1}{2}d_{10} = 75362 - \frac{1}{2} \times 418 = 75253$$

$$T_x = L_x + T_{x+1} \Rightarrow T_9 = L_9 + T_{10} = 75593 + 7953195$$

$$T_9 = 5028788$$

$$e_x^0 = \frac{T_x}{\ell_x} \Rightarrow e_9^0 = \frac{T_9}{\ell_9} = \frac{5028788}{75824} = 66.32$$

$$e_{10}^0 = \frac{T_{10}}{\ell_{10}} = \frac{4953195}{75362} = 65.72$$

12.6 ABRIDGED LIFE-TABLE

A life-table that describes the life expectation for all ages at an interval of one year is called complete life table. Where as an abridged life table consists of all columns of a life-table for ages distant apart usually 5 to 10 years or for age groups usually of width 5 to 10 years. In this way there is tremendous reduction in size of life tables and hence such life-tables are named as Abridged life-tables.

The principal methods used for the construction of abridged life table are

(i) Reed-Merrell method

(ii) Greville's method

(iii) King method

of these, (i) and (ii) are meant for the type of abrigement in (ii) and (iii) is used for type of abridgement in (i).

Remark : The values of q_x are often a smooth function of x and show some irregular variations and as such they should be graduated by Gompertz Makehams law or by some other summation formula. This difficulty can also be over-come by suitable grouping the value of x and obtaining the life table functions for these grouped values of x. This is what exactly is done in abridged life-tables. Thus main objective in the construction of abridged lifetable is to overcome the problems faced in the construction of complete life tables.

12.7 CONSTRUCTION OF ABRIDGED LIFE-TABLE

An abridged life table for an age interval of n years consists of the following columns :

(i) The age advances from x to $(x + n)$

(ii) $\ell_x -$ The number of persons living at the initial age of the interval x to $(x + n)$ out of a cohort ℓ_0.

(iii) $q_x -$ the probability of dying of a person in age group x to $(x + n)$.

$$_nq_x = \frac{\ell_x - \ell_{x+n}}{\ell_x} = 1 - \frac{\ell_{x+n}}{\ell_x}$$

(iv) $nd_x =$ the number of deaths in the interval x to $(x + n)$ i.e.

$$nd_x = \ell_x \times nq_x$$

(v) $_np_x -$ the probability of living of a person between the age x and $(x + n)$ can be given as :

$$_np_x = \frac{\ell_{x+n}}{\ell_x} = 1 - nq_x$$

(vi) ${}_nL_x$ – the number of persons of the life-table stationary population in the interval x to $(x+n)$ years can be obtained as

$$nL_x = \int_0^n \ell_{x+t}\, dt$$

(vii) T_x – The number of persons living after the age x.

(viii) e_x^0 – The expectation of life at the age x and in given as,

$$e_x^0 = \frac{T_x}{\ell_x}$$ complete expectation of life at age x.

(I) REED - MERRELL METHOD :

Construction of abridged life table by Reed-Merrell method is based on central mortality rate. If ${}_np_x^z$ and ${}_nd_x^z$ are the average population size and number of deaths in the age interval x to $(x+n)$ in the calender year z, then

$${}_nm_x^z = \frac{{}^nd_x^z}{{}_np_x^z}$$

and $${}^nq_x^z = \frac{2n\left({}_nm_x^z\right)}{2+n\left({}_nm_x^z\right)}$$

also $np_x = 1 - nq_x$, $\ell_{x+n} = \ell_x \times {}^n p_x$

once we known ${}^nq_x$, it is trivial to construct the abridged life-table.

This method due to L.J. Reed and M. Merrell is based on the following fundamental result which we state in the form of a lemma.

12.7.1 LEMMA :

$${}_nq_x^z = \frac{2n\left({}_nm_x^z\right)}{2+n\left({}_nm_x^z\right)} \quad \text{---------- (1)}$$

where ${}_np_x^z$ and ${}_nd_x^z$ are respectively the average number of persons and the number of deaths between ages x to $(x+n)$ in the calender year z and

$$ {}_{n}m_{x}^{z} = \frac{{}_{n}d_{x}^{z}}{{}_{n}p_{x}^{z}} \text{ ---------- (2)} $$

the control rate of mortality in the calender year z, n being the length of the age group.

Proof : Let the life table is $\ell_x = E_x^z$

Assuming that deaths are uniformly distributed in the interval $(x, x+n)$ or equivalently assuming the linearity of ℓ_{x+t} for $t \in (0,n)$, we get

$$ {}_{n}p_{x}^{z} = \int_{0}^{n} \ell_{x+t}\,dt \simeq \frac{n}{2}\left(\ell_x + \ell_{x+n}\right) $$

$$ = \frac{n}{2}\left[\ell_x + \left(\ell_x - n d_x^z\right)\right] $$

Proof : Let the life table is $\ell_x = E_x^z$

Assuming that deaths are uniformly distributed in the interval $(x, x+n)$ or equivalently assuming the linearity of ℓ_{x+t} for $t \in (0,n)$, we get

$$ {}_{n}p_{x}^{z} = \int_{0}^{n} \ell_{x+t}\,dt \simeq \frac{n}{2}\left(\ell_x + \ell_{x+n}\right) $$

$$ = \frac{n}{2}\left[\ell_x + \left(\ell_n - {}_{n}p_{x}^{z}\right)\right] $$

$$ = n E_x^z - \frac{n}{2}\left(n d_x^z\right) \quad \left[\because \ell_x = E_x^z\right] $$

$$ \Rightarrow E_x^z = \frac{1}{n}\left({}_{n}p_{x}^{z}\right) + \frac{1}{2}\left(n d_x^z\right) $$

By definition, we have

$$ {}_{n}q_{x}^{z} = \frac{n d_x^z}{E_x^z} = \frac{n d_x^z}{\frac{1}{n}\left({}_{n}p_{x}^{z}\right) + \frac{1}{2}\left(n d_x^z\right)} $$

$$= \frac{\frac{nd_x^z}{np_x^z}}{\frac{1}{n} + \frac{1}{2}\left(\frac{nd_x^z}{np_x^z}\right)} = \frac{{}_n m_x^z}{\frac{1}{n} + \frac{1}{2}\left({}_n m_x^z\right)}$$

$$= \frac{2n\left(nm_x^z\right)}{2 + n\left(nm_x^z\right)}$$

a result which is similar to (1)

(II) GREVILLE'S METHOD :

Greville's made use of the age specific central mortality rate ${}_n m_x$ for estimating the death ${}_n q_x$ in the age interval x to $(x+n)$ years. The formula for the estimation of ${}_n q_x$ proponded by Greville is

$$nq_x = \frac{2n\left({}_n m_x\right)}{1 + {}_n m_x\left[n + \frac{n^2}{6}\left(n^{m_x} - \log_e^c\right)\right]}$$

The above formula involves C which is estimated by making assumption that n^{m_x} follows the Compertz (exponential) law,

$$n^{m_x} = BC^x$$

Also following usual notations, the other constituents of the abridged life-table are estimated as follows :

An approximation to ${}_n L_x$ based on numerical quadrature is,

$${}_n L_x = \frac{n}{2}\left(\ell_x + \ell_{x+n}\right) + \frac{n}{24}\left({}_n d_{x+n} - {}_n d_{x-n}\right)$$

If ℓ_x vanishes at the age $w+n$, then

$$nL_x = \frac{\ell_w}{n^m w}$$

and thus,

$$T_x = n^{L_x} + n^{L_{x+n}} + \cdots\cdots + n^{L_w}$$

$$= n^{L_w} + T_{x+n}$$

Obviously, the last column of life-table is

$$e_x^0 = \frac{T_x}{\ell_x}$$

(III) KINGS METHOD :

Central mortality rate can be calculated precisely provided the population p_x^0 and the number of deaths D_x^0 for the central age in the age interval $[x, x+n]$ can be estimated precisely from the known values of population and deaths np_x and np_x respectively in the age group x to $(x + n)$. Under the assumption that p_x^0 and D_x^0 can be estimated by a second degree parabola. Giving estimated then by the formula,

$$p_x^0 = \frac{1}{n}\left({}_n p_x\right) - \frac{1}{24n}\left[1 - \frac{1}{n^2}\right]\Delta^2\left(np_x\right)$$

$$D_x^0 = \frac{1}{n}\left({}_n D_x\right) - \frac{1}{24n}\left[1 - \frac{1}{n^2}\right]\Delta^2\left({}_n Dx\right)$$

Once we know p_x^0 and D_x^0, the control mortality rate,

and $q_x = \frac{2mx}{2 + m_x}$, also

$$m_x = \frac{D_x^0}{p_x^0} \text{ for } x = x_0, x_0 + n, x_0 + 2n, \cdots\cdots$$

$p_x = 1 - qx$, where x_0 is the initial age from where to start. Other constituents of the abridged life-table are given as

$$\ell_{x+n} = \ell_x\left({}_n p_x\right),\ \ell_{x+2n} = \ell_{x+n}\left({}_n p_{x+n}\right)$$

where np_x is computed by relation

$$\log\left({}_n p_x\right) = \sum_{i=1}^{n-1} \log p_{x+i}$$

The value of np_x can be improved by interpretation technique. King denoted the number of years lined by the radix ℓ_x during interval $(x+n)^b T^*_{x\cdot n}$. Under the assumption that deaths are distributed uniformly over the year x to $x+n$,

$T^*_{x\cdot n}$ is given as

$$T^*_{x\cdot n} = L_x + L_{x+1} + \cdots\cdots + L_{x+n-1}$$

$$= \sum_{i=1}^{n-1} L_{x+i} - \frac{1}{2}(\ell_x - \ell_{x+n})$$

Finally ℓ^0_x, the expectation of life at the age x by the persons living in the interval $(x, x+n)$ is obtained by the formula

$$\ell^0_x = \frac{T^*_{x\cdot n}}{\ell_x}$$

The construction of abridged life-table by King's method involves many approximations and lengthy in calculation.

12.8 SUMMARY

A life table, in other words called a mortality table indicates the probability of surviving from age to any subsequent age according to the age specific death rates prevailing at a particular time and place (William Peterson). Strictly speaking, the life-table is not a standardization procedure but rather a method by which a brief account of mortality can be obtained. Hence life tables prevent in component form the age sex specific mortality rates in a given period and place. If we observe a cohort and the experience of this cohort is followed until all the numbers had died, it is possible to give a detailed account of the mortality of this group and construct a "generation" life table. Since this type of life table can be constructed only after the death of all the numbers of the cohort, the utility of such a table is limited. Another type of life table, which has practical significance is called a "time specific life table". This table presents the mortality experience of a population from birth to death and constantly used by demographers, planners, aswell as by insurance companies.

The life table makes it possible to calculate "Expectation of life' which is the average expectation of life for a person of a specific age. The average life expectancy is the expected average number of years a person lives right from his birth. This table provides total number of members and the years to be lived by the entire cohort before the death of the last survivor in the annual computation.

12.9 EXERCISES

1. Define a life table, explain its assumptions.
2. Explain important uses of life-table.
3. Describe the various components of life table. How is the expectation of life at birth determined from life-table?

4. Explain Abridged life table. Enumerate the columns of an abridged life table and describe the steps in their construction.

5. Fill in the blanks of the following table which are marked with questions marks :

Age (x)	ℓ_x	d_x	p_x	q_x	L_x	T_x	e_x^0
20	693435	?	?	?	?	35081126	?
21	690673	--	--	--	--	?	?

6. Fill in the blanks in a portion of life table given below :

Age in years	ℓ_x	d_x	p_x	q_x	L_x	T_x	e_x^0
4	95000	500	?	?	?	4850300	?
5	?	400	?	?	?	?	?

7. Calculate e_x^0 from the following data

Age(x)	dx	Tx
0	1273	--
1	1129	--
2	1090	58510
3	1066	57432
4	1049	56374

9. Fill in the blanks which are maked with a query in the following skeleton life table.

Age (x)	ℓ_x	d_x	p_x	q_x	L_x	T_x	e_x^0	mx
30	762227	3647	?	?	?	27296632	--	--

10. Given that the complete expectation of life at ages 30 and 31 for a particular group are respectively 21.39 and 20.91 years and that the number living at age 30 is 41, 176, find

    (i) The number that attains the age 37 and

    (ii) the number that will die without attaining the age 31.

12.10 REFERENCE BOOKS :

Fundamentals of Statistics : S.C. Srivasthava & Sangya Srivasthava

Applied Statistics : S.C. Gupta & V.K. Kapoor

Basic Statistics : B.L. Agarwal

Methods & Bio-Statistics : B.K. Mahajan

LESSON - 13

COMPONENTS OF TIME SERIES-TREND

OBJECT OF THE LESSON

After studying this lesson the student is expected to have a

* Clear comprehension of the theory and the practical utility about the concepts of components of time series-trend and their applications.

STRUCTURE OF THE LESSON

This consists of sections as detailed below

13.1 INTRODUCTION

Arrangement of statistical data in chronological order i.e., in accordance with occurence of time is known as 'Time series'. Such series have a unique place of importance in the field of economic and business statistics since the series relating to prices, consumption and production of various commodities; money in circulation; bank deposits and bank clearings; sales and profits in a departmental store, agricultural and industrial production, national income and foreign exchange reserves, prices and dividends of shares in a stock exchange market, etc., are all time series spread over a long period of time. A time series depicts the relationship between two variables, one of them being time; for example the population (U_t) of a country in different years (t); Temperature (U_t) of a place on different days (t) etc.,

According to Ya-lun Chou,

"A time series may be defined as a collection of readings belonging to different time periods, of some economic variable or composite of variables".

Mathematically, a time series is defined by the functional relationship

$$U_t = f(t)$$

where U_t is the value of the variable under consideration at time t.

For example (i) the population (U_t) of a country or a place in different years(t) .

(ii) the number of births and deaths (U_t) in different months (t) of the year.

(iii) the sale (U_t) of a departmental store in different months (t) of the year,

(iv) the temperature (U_t) of a place on different days (t) of the week and so on, constitute time series. Thus if the values of a variable at times $t_1, t_2, \cdots\cdots t_n$ are $U_1, U_2, \cdots\cdots, U_n$ respectively, then series

$$t : t_1, t_2, \cdots\cdots, t_n$$

$$U_t : U_1, U_2, \cdots\cdots, U_n$$

Constitute a time series. Thus, a time series invariably gives a bivariate distribution, one of the two variables being time (t) and the other being the value U_t of a variable at different points of time. The values of t may be given yearly, monthly, weekly, daily or even hourly, usually but not always at equal intervals of time.

If the data are segregated by time i.e., days, months, years etc., the value of the variable under consideration changes from time to time. These fluctuations are affected not by a single force but are due to the net effect of multiplicity of forces pulling it up and down and if these forces were in a state of equilibrium the series would remain constant. For example, the retail prices of a particular commodity are influenced by a number of factors viz., the crop yield which further depends on weather conditions, irrigation facilities, fertilizers used, transportation facilities, consumer demand etc.,

13.2 COMPONENTS OF TIME-SERIES

Experience from many examples of time series has revealed the presence of certain characteristic movements or fluctuations in a time series. These characteristic movements of time-series may be classified into different categories called components of time-series. Broadly speaking, the components of time-series are

i) long term movements or fluctuations

ii) short term movements or periodic changes.

iii) erratic or irregular or random fuctuations.

Longterm fluctuations of the time series are called the secular trend or trend. The short term fluctuations are again divided into seasonal fluctuations and cyclical fluctuations. Thus a time-series has the following components in all.

1. Longterm fluctuations or secular trend or simply trend

2. Seasonal fluctuations or variations
3. Cyclical fluctuations or variations and
4. Random fluctuations or variations.

The value of a time series u_t at any time t is regarded as the resultant of the combined impact of above components. In the following section we shall briefly explain them one by one.

1. Trend : By secular trend or simply trend we mean the general tendency of the data to increase or decrease during a long period of time. This is true of most of series of business and economic statistics. For example, an upward tendency would be seen in data pertaining to population, agricultural production, currency in circulation etc., while a downward tendency will be noticed in data of births and deaths, epidemics etc., as a result of advancement in medical sciences, better medical facilities, literacy and higher standard of living. Clearly trend is the general smooth, longterm, average tendency. It is not necessary that the increase or decrease should be in the same direction throughout the given period. It may be possible that different tendencies of increase or decrease or stability are observed in different sections of time. However, the overall tendency may be upward, downward or stable. Such tendencies are the result of the forces which are, more or less constant for a long time or which change very gradually and continuously over a long period of time such as the change in the population, tastes, habits and customs of the people in a society and so on. They operate in an evolutionary manner and do not reflect sudden changes. For example, the effect of population increase over a long period of time on the expansion of various sectors like agriculture, industry, eductaion, textiles, etc., is a continuous but a gradual process. Similarly, the growth or decline in a number of economic time sries is the interaction of forces like advances in production technology, large-scale production, improved marketing, management and business organisation, the discovery of new natural resources and the exhaustion of the existing resources and so on-all of which are gradual processes.

It should not be inferred that all the series must show an upward or downward trend. We might come across certain series whose values fluctuate round a constant reading which does not change with time, for example the series of barometric readings or the temperature of a particular place. If the time series values plotted on graph cluster more or less, round a straight line, then the trend exhibited by the time series is termed as linear as opposed to non linear (curvi-linear). In a straight line trend, the time series values increase or decrease more or less by a constant absolute amount i.e., the rate of growth is constant. Although, in practice, linear trend is commonly used, it is rarely obtained in economic and business data. In an economic and business phenomenon, the rate of growth or decline is not of constant nature throughout but varies considerably in different sectors of time. Usually, in the beginning the growth is slow, then rapid which is further accerlated for quite sometime after which it becomes stationary or stable for some period and finally retards slowly.

The term 'long period of time' is a relative term and cannot be defined exactly. In some cases a period as small as a week may be fairly long while in some cases, a period as long as 2 years may not be long enough. For example, if the data of agricultural production for 24 months shows an increase it won't be termed as secular change over a period of 2 years whereas if the count of bacterial population of a culture every five minutes for a week shows an increase then we would regard it as a secular change.

2. Periodic Changes : It would be observed that in many social and economic situations one

finds that a part from the growth factor in a time series there are forces at work which prevent the smooth flow of the series indicates that in a particular direction and tend to itself over a period of time. These forces do not act continuously but operate in a regular spasmodic manner. The resultant effect of such forces may be classified as

(i) Seasonal variations &

(ii) Cyclic variations

(i) Seasonal variations : These variations in a time series are due to the rhythmic forces which operate in a regular and periodic manner over a span of less than a year i.e., during a period of 12 months and have the same or almost the same pattern year after year. Thus, seasonal variations in a time series will be there if the data are recorded quarterly (every 3 months), monthly, weekly, daily, hourly and so on. Although in each of the above cases, the amplitudes of the seasonal variations are different, all of them have the same period i.e., 1 year. Thus, in a time series data where only annual figures are given, there are no seasonal variations. Most of economic time series are influenced by seasonal fluctuations. For example prices, production and consumption of commodities; sales and profits in a departmental store; bank clearings and bank deposits etc., are all affected by seasonal variations. Also seasonal variations may be attributed to the following two causes.

(a) Those resulting from natural forces : As the name suggests, the various seasons or wheather conditions and climatic changes play an important role in seasonal movements. For example the sale of umbrellas pick up very fast in rainy season, the demand for electric fans goes up in summer season. The sale of ice and ice cream increases very much in summer; the sale of woolens goes up in winter, all being affected by natural forces i.e., weather or seasons. Likewise, the production of certain commodities such as sugar, rice, pulses, eggs etc., depends on seasons similarly the prices of agricultural commodities always go down at the time of harvest and then pick up gradually.

(b) Those resulting from man-made conventions : These variations in a time series within a period of 12 months are due to habits, fashions, customs and conventions of the people in the society. For example sale of jewellary and ornaments goes up during marriages; the sales and profits in departmental stores go up considerably during marriages, and festivals like Diwali, Dusherra, Christmas etc., such variations operate in regular spasmodic manner and recurr year after year.

The main objective of the measurement of seasonal variations is to isolate them from the trend and study their effects. A study of the seasonal patterns is extremely useful to business men, producers, sales-managers etc., in planning future operations and in formulation of policy decisions regarding purchase, production, inventory control personnel requirements, selling and advertising programmes. In the absence of any knowledge of seasonal variations, a seasonal upswing may be mistaken as indicator of better business conditions while a seasonal slump may be mis-interpretated as deteriorating business conditions. Thus, to understand the behaviour of the phenomenon in a time series properly, the time series data must be adjsuted for seasonal variations. This is done by isolating them from trend and other components by dividing the given time series value (U_t) by the seasonal variations (S_t). This technique is called de-seasonalisation of data.

(ii) Cyclic variations : The oscillatory movements in a time series with period of oscillation more

than one year are termed as cyclic fluctuations. One complete period is called a 'cycle'. The cyclic movements in a time series are generally attributed to the so called 'Business cycle', which may also be referred to as the four-phase cycle' composed of prosperity recession, depression and recovery and normally lasts from seven to eleven years. The up and down movements in business depends upon the commulative nature of the economic forces and the interaction between them. Most of the economic and commercial series ex : series relating to prices, production and wages etc., are affected by business cycles. Cyclic fluctuations, though more or less regular are not periodic.

iii) Irregular component : Apart from the regular variations, almost all the series contain another factor called the random or irregular or residual fluctuations which are not accounted by secular trend and seasonal and cyclic variations. These fluctuations are purely random, erratic, unforeseen, unpredictable and are due to numerous non-recurring and irregular circumstances which are beyond the control of human hand but at the same time are a part of our system such as earthquakes, wars, floods, famines, revolutions, epidemics, etc. These isolated or irregular but powerful fluctuations due to floods, revolution, political upheavals, famines etc., are also called episodic fluctuations. In some cases the importance of irregular fluctuations may not be significant while in others these may be very effective and might give rise to cyclic movements.

13.3 ANALYSIS OF TIME SERIES - IMPORTANCE

There are four basic types of variations i.e., trend, seasonal variations, cyclical variations, random variations in a time series which are super imposed and act as periodic in time to give the series its erratic appearance. To discover and measure the effect of these components and to isolate them individually is known as the analysis of time series.

The main problem in the analysis of time series is to identify the components or factors at work and to isolate, study and measure them independently. This is the way to obtain the maximum possible information from the given data arranged in chronological order.

Importance : Time series analysis is the basis for understanding the past behaviour, evaluating the current accomplishments and anticipating the future operations. It is also used for comparing the components of different sets of data. The analysis of time series makes the, time series data more useful for economists, scientists, sociologists and biologists etc.,

According to B.J. Mendel "Quantitative data on the past and present behaviour of an activity in combination with knowledge about the various factors that influence it makes it possible to forecast the activity, future magnitude. Thus, time series data on a given activity are useful in deliberately forcing a comprehensive study of the factors affecting it, learning of its past behaviour, interpreting its current behaviour and forecasting its probable future magnitude". In short following are the advantages or uses of time series analysis.

(i) Knowledge of past behaviour of the factors which are responsible for the variations.

(ii) Projecting past trends into the future.

(iii) Evaluating the present achievements.

(iv) Interpretation of the variations as to how they are related with each other, with net effect of their interaction and also with the similar changes in other time series data.

(v) The segregation and study of the various components is of paramount importance to a business man in the planning of future operations and in the formulation of executive and policy decisions.

13.4 MATHEMATICAL MODELS FOR TIME SERIES

The following are the two models commonly used for the decomposition of a time series into its components.

(i) Decomposition by Additive Hypothesis : According to the additive model, a time series can be expressed as

$$U_t = T_t + S_t + C_t + R_t \qquad (13.4.1)$$

Where U_t is the time series value at time t. T_t represents the trend value S_t, C_t, R_t represent the seasonal, cyclic and random fluctuations at time t. Obviously, the term S_t will not appear in a series of annual data. The additive model (13.4.1) implicitly implies that seasonal forces, in different years, cyclical forces in different cycles and irregular forces in different long time periodic operate with equal absolute effect irrespective of the trend value. As such C_t and S_t will have positive or negative values, according as whether we are in an above normal or below normal phase of the cycle and the total of positive and negative values, for any cycle will be zero. R_t will also have positive or negative values and in the long run ΣR_t will be zero. Occasionally there may be a few isolated occurrences of extreme R_t of episodic nature.

(ii) Decomposition by Multiplicative Hypothesis : On the other hand if we have reasons to assume that the various components in a time series operate proportionately to the general level of the series, the traditional or classical multiplicative model is appropriate. According to the multiplicative model.

$$U_t = T_t \cdot S_t \cdot C_t \cdot R_t \qquad (13.4.2)$$

Where S_t, C_t and R_t instead of assuming positive and negative values are indices fluctuating above or below unity and the geometric means of S_t in a year, C_t in a cycle and R_t in a long-term period are unity. In a time series with both positive and negative values, the multiplicative model (13.4.2) cannot be applied unless the time series is translated by adding a suitable positive value. It may be pointed out that the multiplicative decomposition of a time series is the same as that of the additive decomposition of the logarithmic values of the original time series.

$$\log U_t = \log T_t + \log S_t + \log C_t + \log R_t$$

Limitations of the Hypothesis of Decomposition of a Time-series : Hypothesis of decomposition presupposes that the trend and periodic components are determined by separate forces acting independently so that simple aggregation of the components could constitute the series. But in reality, it is possible that this year's value of the series will depend to some extent on last year's value so that trend and periodic movement will get inextricably mixed up and no meaningful separation of them will be possible. In such a case any variations of this year may affect the whole future course of the series and no meaningful separation of trend and periodic components will be possible.

In addition to the addition and multiplication models discussed above the components in a time series may be combined in a large number of other ways. The different models, defined under different assumptions will yield different results. Some of the mixed models resulting from different combinations of additive and multiplicative models are given below.

$$U_t = T_t C_t + S_t R_t$$

$$U_t = T_t + S_t C_t R_t$$

$$U_t = T_t + S_t + C_t R_t$$

The model (13.4.1) or (13.4.2) can be used to obtain a measure of one or more of the components by elimination, i.e., subtraction or division. For example, if trend component (T_t) is known, then using multiplication model, it can be isolated from the given time series to give

$$S_t \times C_t \times R_t = \frac{U_t}{T_t} = \frac{\text{original values}}{\text{Trend values}}$$

thus for the annual data, for which the seasonal component S_t is not there, we have

$$U_t = T_t \times C_t \times R_t \Rightarrow C_t \times R_t = \frac{U_t}{T_t}$$

13.5 MEASUREMENT OF TREND

Trend can be studied and or measured by the following methods :

(i) Graphic method or Trend by Inspection

(ii) Method of semi-Averages

(iii) Method of curve fitting by principles of least squares and

(iv) Method of moving averages

(i) Graphic Method : A free-hand smooth curve obtained on plotting the values U_t against 't' enables us to form an idea about the general 'trend' of the series. Smoothing of the curves eliminates other components, i.e., regular and irregular fluctuations. This method does not involve any complex mathematical techniques and can be used to describe all types of trend linear or non-inear, thus, simplicity and flexibility are strong points of this mehtod. Its main drawbacks are

1. The method is very subjective i.e., the bias of the person handling the data plays a very important role and as such different trend curves will be obtained by different persons for the same set of data. As such trend by inspection should be attempted only by skilled and experienced statistician and this limits the utility and popularity of the method.
2. It does not enable us to measure trend.

(ii) Method of Semi-Averages : If it is clear from the data that a straight line is appropriate for the trend for which the semi-average method works well. In this method, the series is divided into two

parts. The first part consists of the first half years (periods) and the second part of the remaining in years (periods). In case the series consists of an even number of years say, 2K where K is an integer. The series is divisible into two halves each consisting K years. Here, we find the average of K years of the two parts of series and place these values in the mid-year of each of the half series. It is easy to locate the mid year if K is odd. In this case $\left(\frac{K+1}{2}\right)^{th}$ year will be the mid year for the first half series and $\frac{(3K+1)^{th}}{2}$ year for the second half series. For example, if 2K = 10, then K = 5. In this case the third year will be the mid-year for the first half series and the eighth year for the second half series. But in the case where K is even, none of the year in the half series will be a mid year. Here the middle point of the two mid-years will be the mid-period where the average of half series has to be located. For example, if the data pertain to 12 years i.e., from 1971 to 1982. The half series will consist of 1971 to 1976 and 1977 to 1982. The mid period for the first half will be the middle of 1973 - 74, i.e, 1st July, 1973 and similarly the mid period of second half series will be the mid-point of 1979-80, i.e., 1st July, 1979, enter the average values against these mid-points in the table. Plot the data on graph paper and these two average values at the mid-points. The original data should be joined by dotted lines and the two average values by a smooth straight line. This straight line shows the trend.

Now we consider the case, when the number of years in the series is odd say, $(2K+1)$, where K is an integer. In this case, it is not possible to divide the series into two equal halves. Here, the first K years form the first half series and the last K years form the last half series. The value for $(K+1)$ th year is either included in both the half series or it is excluded completely. The later part is generally implemented. It is advisable to include this value in both the half series if K is even and omit it if K is odd. For instance, if there are 11 years in the series, 2K+1 = 11 and thus, K = 5. Here 6th value in the sereis will have to be either included or excluded. Once the series is divided into halves,the rest of the procedure remains the same as given, in the above respective paragraph.

The semi-average method is superior to the freehand method, as it is not subjective. In this method, everyone will get the same trend line. But it possesses all the demerits of an average, i.e., it will be affected by the extreme values if present in the series. In this situation, it does not depict the true trend line. A semi average method does not ensure the elimination of short-term and cyclic variations. This danger is more if the period for average is small. Hence this method may be used when the data are given for a longer period. All the situations have been dealt with in the three numerical examples in section 13.6.

(iii) Method of Curve Fitting by Principle of Least Squares : The principle of least squares is the most popular and widely used method of fitting mathematical functions to a given set of data. The method yields very correct results. If sufficiently good appraisal of the form of the function to be fitted is obtained either by a scrutiny of the graphical plot of the values over time or by a theoritical understanding of the mechanism of the variable change.

Fitting of Straight Line by Least Square Method :

Let the equation of the straight line be

$$U_t = a + bt \qquad \text{(A)}$$

Principle of least squares consists in minimizing the sum of squares of the deviations between the given values of U_t and their estimates given by (A). In otherwords we have to find a and b such that for given values of U_t corresponding to n different values of t,

$$Z = \sum_t (U_t - a - bt)^2$$

is minimum. For a maxima or minima of z, for variations in a and b, we should have

$$\frac{\partial Z}{\partial a} = 0 = -2\Sigma(U_t - a - bt)$$

$$\frac{\partial Z}{\partial b} = 0 = -2\Sigma t(U_t - a - bt)$$

$$\left.\begin{aligned} \Rightarrow \Sigma U_t &= na + b\Sigma t \\ \Sigma t\, U_t &= a\Sigma t + b\Sigma t^2 \end{aligned}\right\} \quad \text{(B)}$$

which are the normal equations for estimating a and b. The values of ΣU_t, $\Sigma t U_t$, Σt, Σt^2 are obtained from the given data and the equations (B) can now be solved for a and b. With these values of a and b, the line (A) gives the desired trend line. The solution of normal equations (B) provides a minima of Z.

FITTING OF SECOND DEGREE PARABOLA :

Let the equation of the second degree parabola is

$$U_t = a + bt + ct^2$$

Proceeding similarly as in the case of a straight lines the normal equations for estimating a, b, and c are given by

$$\left.\begin{aligned} \Sigma U_t &= na + b\Sigma t + c\Sigma t^2 \\ \Sigma t U_t &= a\Sigma t + b\Sigma t^2 + c\Sigma t^3 \\ \Sigma t^2 U_t &= a\Sigma t^2 + b\Sigma t^3 + c\Sigma t^4 \end{aligned}\right\} \quad (1)$$

FITTING OF EXPONENTIAL CURVE :

Let the equation of the exponential curve is

$$U_t = ab^t$$

$$\Rightarrow \log U_t = \log a + t \log b \quad (2)$$

$$\Rightarrow y = \log U_t,\ A = \log a,\ B = \log b \quad (3)$$

(2) is a strraight line in t and Y and thus the normal equations for estimating A and B are

$$\left.\begin{aligned}\Sigma Y &= nA + B\Sigma t \\ \Sigma tY &= A\Sigma t + B\Sigma t^2\end{aligned}\right\} \quad (4)$$

These equations can be solved for A and B and finally from (3) we get

$$a = \text{anti}\log(A);\ b = \text{anti}\log(B)$$

SECOND DEGREE CURVE FITTED TO LOGARITHMS :

Suppose the trend curve is

$$U_t \Rightarrow a \cdot b^t \cdot c^{t^2} \quad (1)$$

Taking logarithms on both sides, we get

$$\log U_t = \log a + t\log b + t^2 \log c$$

$$\Rightarrow V_t = A + Bt + Ct^2 \quad (2)$$

Where $V_t = \log U_t$; $A = \log a$; $B = \log b$ and $C = \log c$, Now (2) is a second degree parabolic curve where in V_t and t can be fitted by the technique explained earlier. We can finally obtain.

$a = \text{Anti}\log(A);\ b = \text{Anti}\log(B)$ and $c = \text{Anti}\log(C)$. With these values of a, b and c the curve where (1) becomes the best second degree curve fitted to logarithm.

The method of curve fitting by the principle of least squares is used quite often in trend analysis particularly when one is interested in making projections for future times. Obviously, the reliability of the estimated values primarily depend upon the appropriateness of the form of the mathematical function fitted to the given data. If the function is determined on an adhoc basis by the scrutiny of the plotted values, the projections based on it may be valid for the near future . However, if the study of physical mechanism of the variable change forms the basis for the selection of the function, there is very little likelihood that the function will change for sufficiently long period and hence in this case reliable long term projections can be made.

MERITS AND DEMERITS OF TREND FITTING BY THE PRINCIPLE OF LEAST SQUARES

Merits : The method of least squares is the most popular and widely used method of fitting mathematical functions to a given set of observations. It has the following advantages.

1. Because of its mathematical or analytical character, this method completely eliminates the element of subjective judgement or personnel bias on the part of the investigator.
2. Unlike the method of moving averages, this method enables us to compute the trend values for all the given time periods in the series.
3. The trend equation can be used to estimate or predict the values of the variable for any period t in future or even in the intermediate periods of the given series and the forecast values are sufficiently reliable.

4. The curve fitting by the principle of least squares is the only technique which enables us to obtain the rate of growth per annum for yearly data, if linear trend is fitted.

Demerits :

1. This method is quite tedious and time consuming as compared to other methods. It is rather difficult for a non-mathematical person to understand and use this method.
2. The addition of even a single new observation necessitates all calculations to be done afresh.
3. Future predictions or forecasts based on this method are based only on the long term variations i.e., trend and completely ignore the cyclical, seasonal and irregular fluctuations.
4. The most serious limitation of the method is the determination of the type of the trend curve to be fitted. i.e., whether we should fit a linear or a parabolic trend or some other more complicated trend curve.
5. It cannot be used to fit growth curves like modified exponential curve, Gompertz curve and logistic curve to which most of the economic and business time series data conform.

(iv) Moving Average Method : It consists in measurement of trend by smoothing out the fluctuations of the data by means of a moving average. Moving average of a period m is a series of successive averages (arithmetic means) of m terms at a time, starting 1st, 2nd, 3rd term etc. Thus the first average is the mean of the 1st m terms, the 2nd is the mean of the m terms from 2nd to (m+1)th term, the third is the mean of the m terms from 3rd to (m+2)th term and so on.

If m is odd = (2K+1) say, moving average is placed against the mid-value of the time interval it covers, i.e., against t = K+1 and if m is even = 2K it is placed between the two middle values of the time interval it covers, i.e., between t = K and t = K+1. In the latter case the moving average does not coincide with the original time period and an attempt is made to synchronise the moving averages and the original data by centering the moving averages which consists in taking a moving average of period two, of these moving averages and putting the first of these values against $t = K + 1$. The graph obtained on plotting the moving average against time gives trend.

In this method, the main problem which is of paramount importance lies in determining the period of the moving average which will completely eliminate the oscillatory movements affecting the series. It has been established mathematically that if the fluctuations are regular and periodic then the moving average completely eliminates the oscillatory movements provided (a) the period of moving average is exactly equal to the period of oscillation. (b) the trend is linear. Since different cycles vary in amplitude and period in such cases the appropriate period of moving average should be equal to or somewhat greater than the mean period of the cycles in the data. In such cases the moving averages method does not completely wipe out the cyclic movements and hence can't give a nice picture of the general trend.

Moving average method is very flexible in the sense that the addition of a few more figures to the data simply results in some more trend values.

The moving average method has the following drawbacks:

(i) It does not provide trend values for all the terms. ex.: for a moving average of period 2K+1, we have to forego the trend values for the first K and last K terms of the series.

(ii) It cannot be used for forecasting or predicting future trend, which is the main objective of trend analysis.

Properties of the moving average method :

1. Cyclical fluctuations with a uniform period can completely be ironed out by this method. To iron out this type of fluctuations through moving averages means to remove their influence. Thus, the moving averages are used as a tool in all parts of the time series analysis.
2. The moving averages can adopt themselves to changing circumstances i.e., any change in the trend will duly be reflected by them.
3. It can adopt to the addition or subtraction of the values.
4. The method is simple to understand and easy to adopt.
5. The method is not subjective. The degree of the polynomial is to be decided by the statistician.
6. It reduces the irregular fluctuations also if its time period happens to be sufficiently large.

Limitations of the moving average method :

1. Some trend values at the beginning and some at the end remain unestimated and their number increases with the increase in the time period of the moving averages.
2. The choice of the period of moving average is sometimes subjective. Actually there is no hard and fast rule for the choice of the period and things depend mostly on personal judgement.
3. The period and the amplitude of the cyclical fluctuations vary from cycle to cycle. In such cases cycles will not completely be eliminated even if the moving average period is calculated by averaging the period of these cycles.
4. Moving averages are affected by the extreme values.
5. Irregular fluctuations are not removed completely.
6. The method gives good results if the trend is linear or atleast approximately so. If trend is non-linear; say concave upward then a moving average will over estimate the trend values, and if it is convex it will under estimate the trend values.
7. Moving averages cannot be represented readily by the mathematical formulae. Thus, the method is not useful for comparison of the trends and cannot be used to extend the trend line for forecasting purposes.

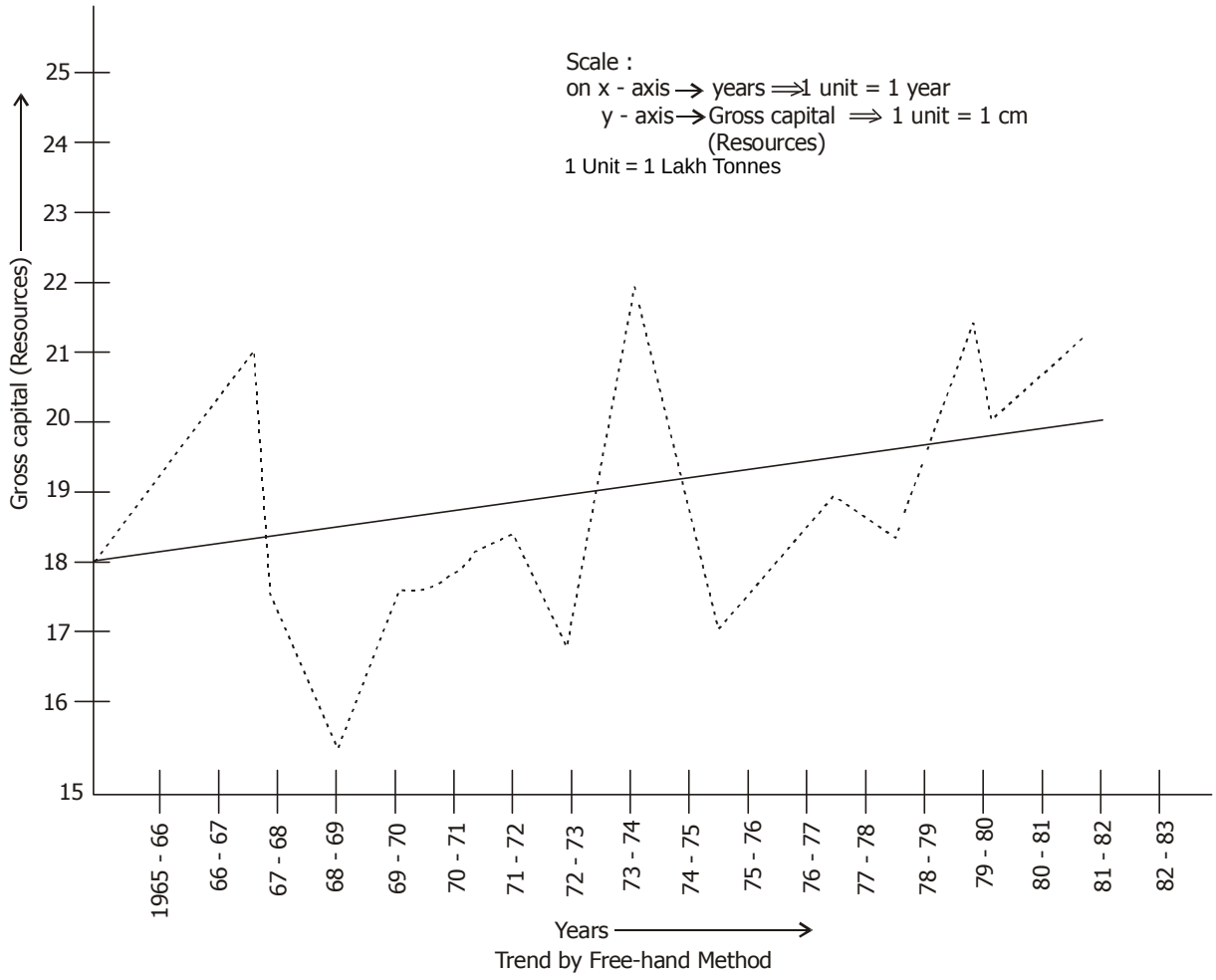
13.6 WORKEDOUT EXAMPLES

Example 1 : The data below gives the rate of gross capital formation from 1965 - 66 to 81 - 82 in Rs. Crores at 1970 - 71 prices.

Years :	1965 - 66	66 - 67	67-68	68-69	69-70	70-71	71-72	72-73	73-74
Gross : Capital formation (Rs. crores)	19.3	20.9	17.8	16.1	17.6	17.8	18.3	17.3	21.4

Years :	74-75	75-76	76-77	77-78	78-79	79-80	80-81	81-82
Gross : Capital formation (Rs. crores)	19.3	18.1	19.5	19.2	22.2	20.9	21.5	21.9

Solution : Here years are taken on X - axis and gross capital formation on Y - axis. The trend line by a free-hand method is drawn.



Example 2 : The figures given below show the export of sugar from India for the years 1970-71 to 1979 - 80.

Years :	1970 - 71	71 - 72	72 - 73	73 - 74	74 - 75	75 - 76	76 - 77	77 - 78
Export : (Lakh Tonnes)	3.9	1.3	1.1	4.4	9.4	9.6	3.4	2.5

Years :	1978 - 79	79 - 80
Export : (Lakh Tonnes)	8.6	2.9

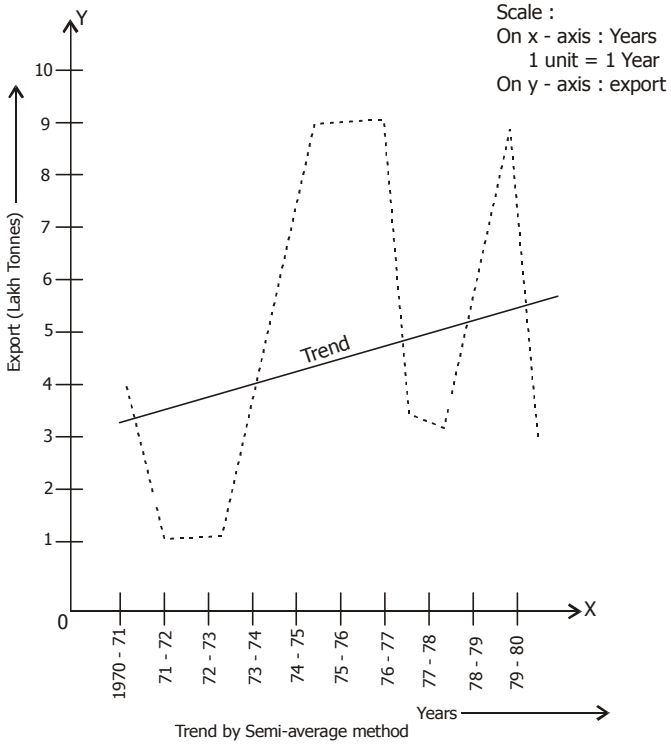
Solution : The trend of the export of sugar from India can be found by the semi average method in the manner given in theory.

Here, 2K = 10, or K = 5, an odd integer. We divide the series into halves consisting of the first five years and the last five years.

The average of the first five years $= \frac{1}{5}(3.9+1.3+1.1+4.4+9.4) = 4.02$

The average of the last five years $= \frac{1}{5}(9.6+3.4+2.5+8.6+2.9) = 5.4$

Plot the data on a graph paper and join them by dotted lines. Plot the average value of the first half series against the year 1972 - 73 and the average of the last half series against 1977 - 78 on the same graph. On joining these two points, we get the trend line as shown in the following graph.



Trend by Semi-average method

Example 3 : The following table gives the export data of a country from 1968 to 1976.

Year :	1968	1969	1970	1971	1972	1973	1974	1975	1976
Export :	720	681	723	666	679	951	1093	1031	1144

(Rs. Crores)

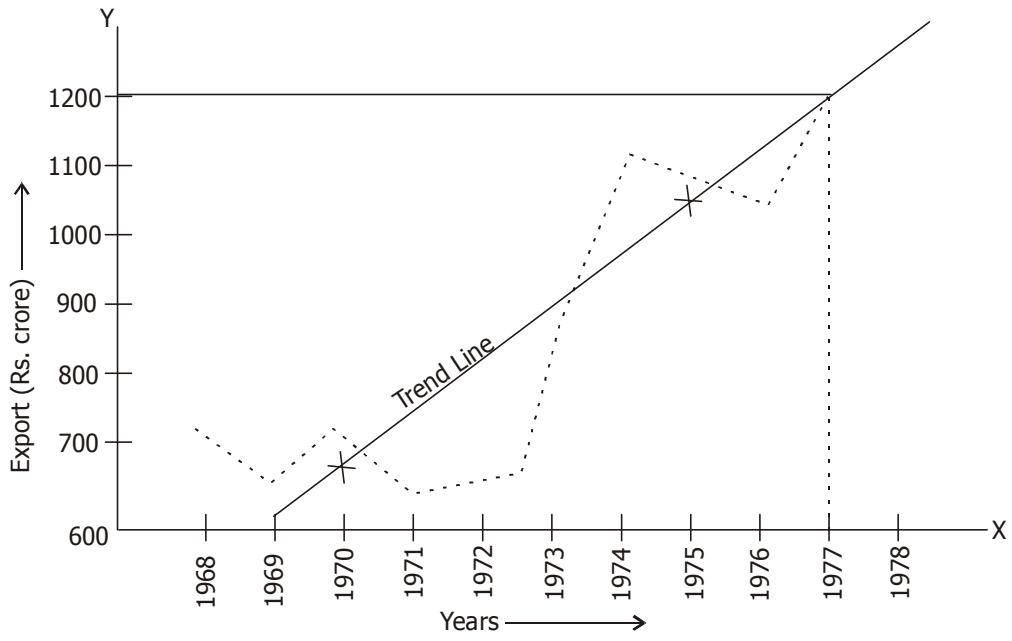
Solution : The given sereis contains data for the nine years, hence $2K+1=9$ or K = 4. We take the first four year's figures in half series and the last four year's data to form the last half of the series. The year 1972 is not included in either of the half series. Since the number of years in the half series is four, we include 1972's data in both the half series. Now the half series consist of five year's data each.

Find the average of each half series separately.

The average of the first half series $=\frac{1}{5}(720+681+723+666+679) = 693.8$

The average of the last half series $=\frac{1}{5}(679+951+1093+1031+1144) = 979.6$

Plot the data on a graph paper and join the plotted points by dotted lines. Plot the average values against the years 1970 and 1974 respectively. Join these average values by a straight line which gives the trend of export as shown in the following graph.



The extent of exports in the year 1977 can be predicted by the trend line. Locate the year 1977 on the X - axis and from this point draw a line parallel to the Y - axis so that it cuts the trend line. Now from the point of intersection P draw a line parallel to the X - axis so that it touches the Y - axis. The reading at this point on the vertical scale, which is the ordinate of point P, gives the predicted value of exports for the year 1977. The estimated value of exports comes out to be Rs. 1190 crores.

Example 4 : The following table gives the number of black and white films produced in India from the year 1970 to 1977.

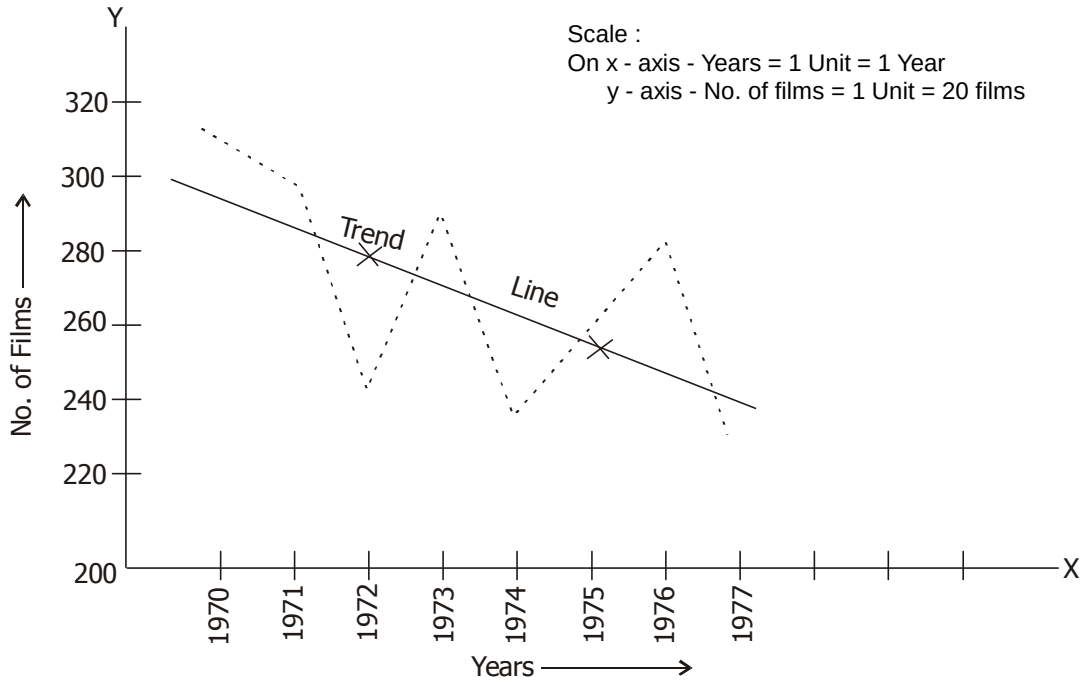
Year :	1970	1971	1972	1973	1974	1975	1976	1977
No. of films :	311	289	236	278	225	243	267	217

Solution : The number of years in the given series is eight therefore, the half series will consist of the first and the last four year's data.

The average of the first half series $= \frac{1}{4}(311 + 289 + 236 + 278) = 278.5$

The average of the last half series $= \frac{1}{4}(225 + 243 + 267 + 217) = 238$

The average of the first half series has to be plotted against the mid-point of 1971-72 and the last half series average has to be plotted against the mid-point of 1975-76. Plot the data and join the points in order, by dotted lines to show the variability of the series. By plotting the average values as stated above and joining these two points by a straight line, we get the trend of production of black and white films as shown in the following graph.



Trend of Production of black and white films by semi-average method.

Example 5 : Fit a straight line trend by the method of least squares to the following data. Assuming that the same rate of change continues, what would be the predicted sales for the year 1985 ?

Year :	1976	1977	1978	1979	1980	1981	1982	1983
Sales :	76	80	130	144	138	120	174	190

(Lakhs Rs.)

Solution : Here $n = 8$, i.e., even. Hence shift the origin to the arithmetic mean of the two middle years. i.e., 1979 & 1980. We define

$$t = \frac{x - \frac{1}{2}(1979 + 1980)}{\frac{1}{2}(\text{interval})} = \frac{x - 1979.5}{\frac{1}{2} \times 1} = 2x - 3959 \qquad (1)$$

where t values are in units of six month (i.e., half year)

Computation of Straight Line Trend

Year (x)	Sales (Lakh Rs.) U_t	t	$t \cdot U_t$	t^2	Trend values $U_t = 131.5 + 7.33t$
1976	76	-7	-532	49	80.19
1977	80	-5	-400	25	94.85
1978	130	-3	-390	9	109.51
1979	144	-1	-144	1	124.17
1980	138	1	138	1	138.83
1981	120	3	360	9	153.49
1982	174	5	870	25	168.15
1983	190	7	1330	49	182.81
	$\Sigma U_t = 1052$	$\Sigma t = 0$	$\Sigma t \cdot U_t = 1232$	$\Sigma t^2 = 168$	

Let the straight line trend equation between U_t and t be

$$U_t = a + bt \qquad (2)$$

Since $\Sigma t = 0$, the normal equation for estimating a and b

$$a = \frac{\Sigma U_t}{n} = \frac{1052}{8} = 131.5$$

$$b = \frac{\Sigma t\, U_t}{\Sigma t^2} = \frac{1232}{168} = 7.33$$

Hence the least square trend line becomes

$$U_t = 131.5 + 7.33t \qquad (3)$$

where $b = 7.33$ units represents half yearly increase in the earnings. The trend values for the years 1976 to 1983 can now be obtained from (3) on putting t = -7, -5, ... 5, 7 respectively as shown in the last column of the table.

Estimate for 1985. When 1985, we get from (1)

$$t = 2x - 3959 = 2 \times 1985 - 3959 = 11$$

Hence the predicted sales for 1985 are

$U_t = 131.5 + 7.33 \times 11 = 212.13$ (Lakhs Rs).

Example 6 : Below are given the figures of production (in thousand quintals) of a sugar factory.

Year :	1973	1975	1976	1977	1978	1979	1982
Production :	77	88	94	85	91	98	90

(i) Fit a straight line by the "least squares method" and tabulate the trend values.

(ii) Eliminate the trend. What components of the time series are thus left over.

(iii) What is the monthly increase in the production of sugar ?

Solution : Here n = 7, i.e., odd. Hence we shift the origin to the year 1977. We define

$$t = \frac{x - 1977}{1} = x - 1977 \qquad (1)$$

Computation of Trend values

Year	Production U_t	t	$t\,U_t$	t^2	Trend values (1000, quintals)	Elimination of trend
1973	77	-4	-308	16	83.32	-6.32
1975	88	-2	-176	4	86.06	+1.94
1976	94	-1	-94	1	87.43	+6.57
1977	85	0	0	0	88.80	-3.80
1978	91	1	91	1	90.17	+0.83

1979	98	2	196	4	91.54	+6.46
1982	90	5	450	25	96.65	-5.65
	623	1	159	51	622.97	

Let the trend equation be $U_t = a + bt$

Normal equations for estimating a and b are

$$\Sigma U_t = na + b\Sigma t$$

$$\Sigma tU_t = a\Sigma t + b\Sigma t^2$$

$\Rightarrow 623 = 7a + b$

$159 = a + 51b$

Solving for a and b, we get $a = 88.80$ and $b = 1.37$

$\therefore$ Trend equation is $U_t = 88.8 + 1.37t$

Substituting the values of t. i.e, -4, -2 etc., successively, we get the required trend values as shown in the table. For example trend value for 1973 is 88.8+1.37(-4) = 83.32.

(ii) Assuming additive model for the time series, the trend values are eliminated by subtracting them from the given values as shown in the table. The resulting values show the short-term fluctuations which change with a period more than one year.

(iii) Yearly increase in the production of sugar, as provided by linear trend.

$U_t = a + bt$ is $'b' = 1.37$

$\therefore$ Monthly increase $= \frac{1.37}{12} = 0.114$ thousand quintals.

Example 7 : Calculate 3 yearly moving average for the following data and plot both original and trend values on the growth.

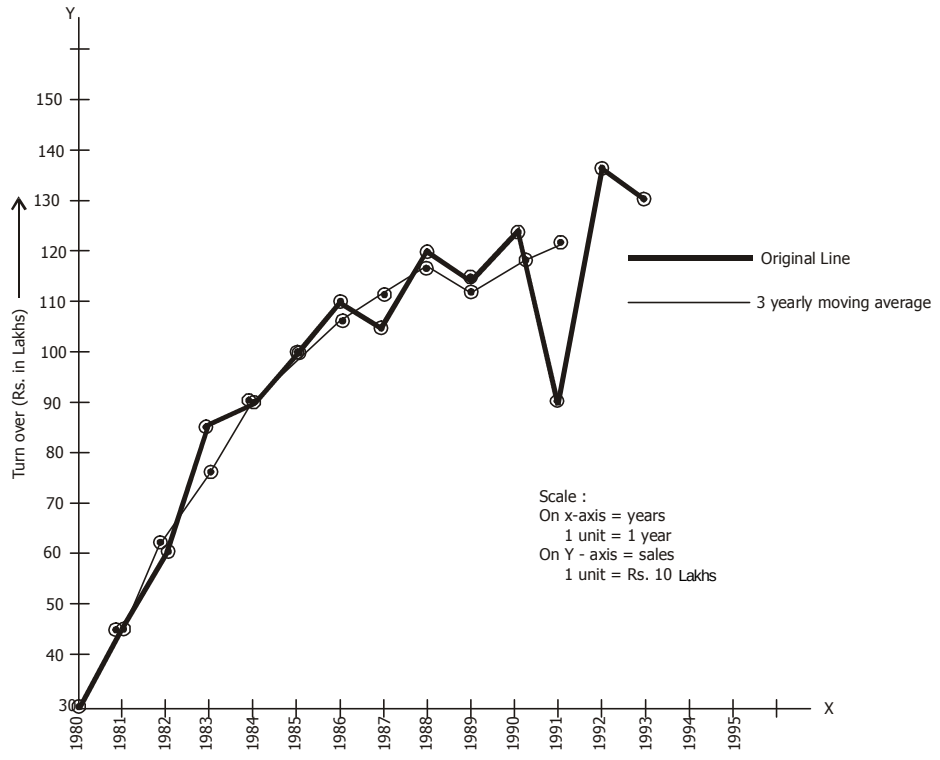
Year :	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Sales in Rupees (in lakhs)	30	45	60	82	90	100	110	105	120	112

Year :	1990	1991	1992	1993
Sales in Rupees(in lakhs)	122	90	135	130

Solution : Calculation of 3 yearly moving averages are as follows.

Year	Turn over Rs. (in Lakhs)	3 yearly Totals	3 yearly Moving average
1980	30	--	--
1981	45	135	45
1982	60	187	62.33
1983	82	232	77.33
1984	90	272	90.66
1985	100	300	100.00
1986	110	315	105.00
1987	105	335	111.67
1988	120	337	112.33
1989	112	354	118.00
1990	122	324	108.00
1991	90	347	115.66
1992	135	355	118.33
1993	130	--	--

Three yearly moving averages can be shown on graph.



Example 8 : The following are the index numbers of retail price of wheat. Plot the figures on a graph and hence find out the periodicity of the given series. Taking the periodicity as the interval of the moving averages. Find out the linear trend and plot on the same graph paper.

Year	Index	Year	Index
1873	100	1883	103
1874	94	1884	91
1875	81	1885	89
1876	78	1886	103
1877	112	1887	121
1878	147	1888	123
1879	158	1889	118
1880	118	1890	117
1881	96	1891	137
1882	101		

Solution : Average periodicity is the average of the business cycles of different periods and is given by

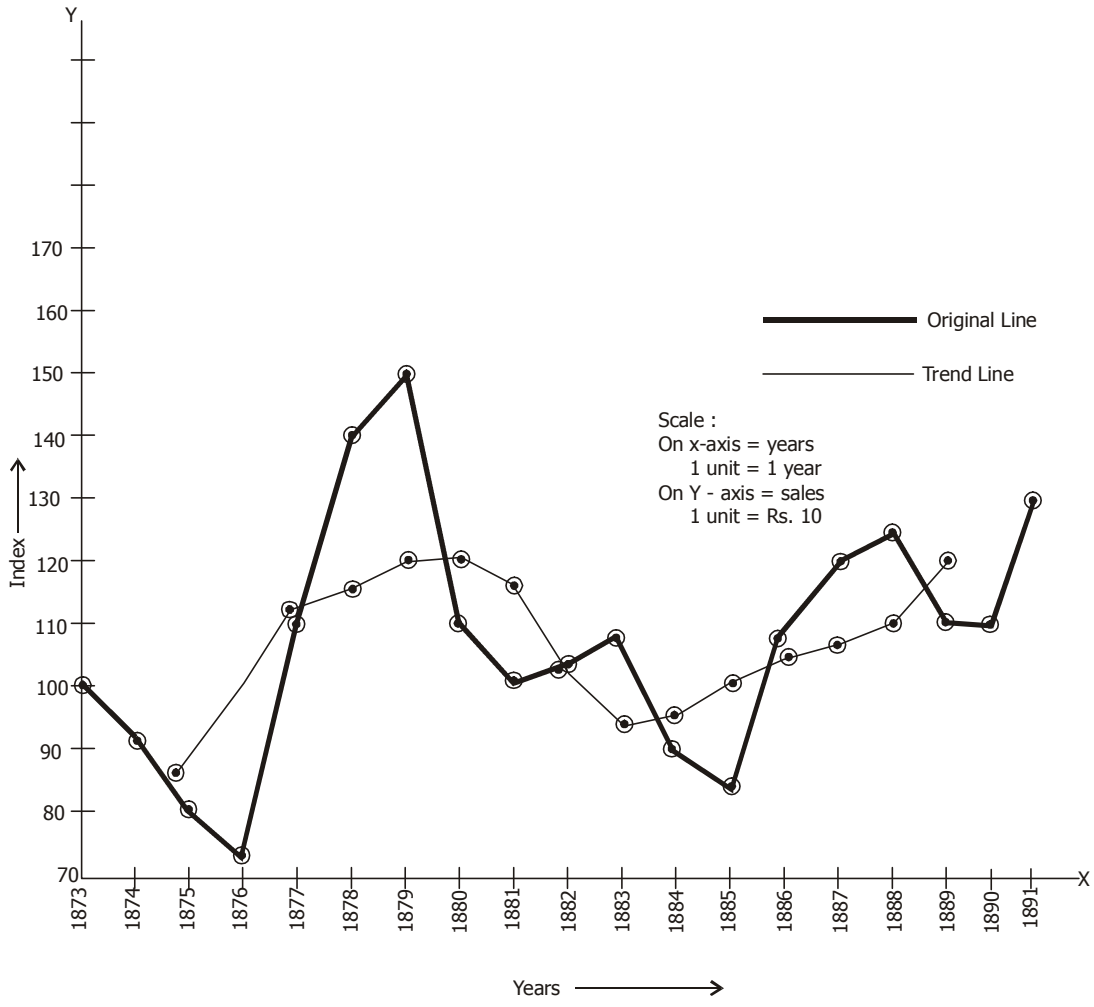
$$\text{Average periodicity} = \frac{7+4+5+3}{4} = \frac{19}{4} = 4.75 \simeq 5$$

Calculation of trend values by a moving average of interval 5.

Year	Index	Moving sum of 5 values	Moving Average
1873	100		
1874	94		
1875	81	445	89.0
1876	78	502	100.4
1877	112	566	113.2
1878	147	603	120.6
1879	158	621	124.2
1880	118	620	124.0
1881	96	576	115.2
1882	101	509	101.8

1883	103	480	96.0
1884	91	487	97.4
1885	89	507	101.4
1886	103	527	105.4
1887	121	554	110.8
1888	123	582	116.4
1889	118	616	123.2
1890	117	--	--
1891	137	--	--

The relevant graph is shown below.



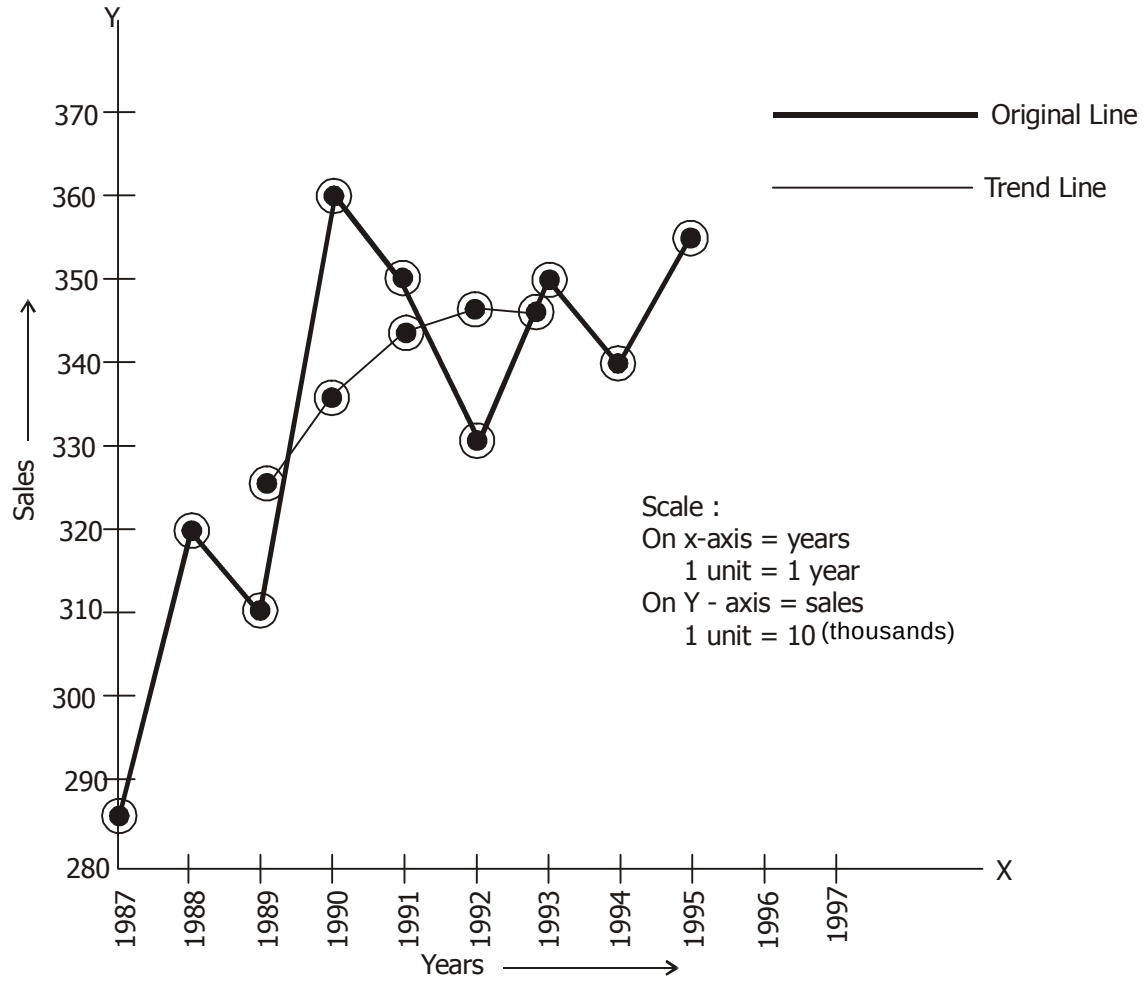
Example 9 : Assuming 4 yearly cycle, calculate the trend values by the method of moving averages and plot the actual data and the trend values on the graph.

Year :	1987	1988	1989	1990	1991	1992	1993	1994	1995
Sales : (1000 Rs)	285	320	310	362	355	336	350	342	359

Solution : since the period of moving average is even i.e., 4, the moving averages will be calculated as follows.

Calculations of Moving Averages Trend :

Year	Sales in Rs. 1000	4 yearly moving total	4 yearly moving average	Moving average centered (Trend)
1987	285	--	--	--
1988	320	1277	--	--
1989	310	1347	319	$\frac{656}{2} = 328$
1990	362	1363	337	$\frac{678}{2} = 339$
1991	355	1403	341	$\frac{692}{2} = 346$
1992	336	1377	351	$\frac{695}{2} = 348$
1993	350	1387	344	$\frac{691}{2} = 346$
1994	342	--	347	--
1995	359	--	--	--



13.7 EXERCISES

1. Define a time series. Mention its important components with illustrations and describe a method of smoothing of time series.
2. Describe the nature of the component of a time series. Explain the additive and multiplicative models of a time series stating clearly the assumptions and discuss their relative merits.
3. Explain clearly what is meant by trend of a time series? Describe the moving average method for determining trend. Explain how the method is related to the method of fitting curves by the principle of least squares.
4. Describe any one of the methods of fitting trend to time series.
5. Explain how the principle of least squares is used to estimate trend in a time series. Below are given the figures of production (in thousand tons) of a sugar factory.

Year (t) :	1975	1976	1977	1978	1979	1980	1981
Production Y_t :	77	88	94	85	91	98	90

(a) Fit a straight line by the method of least squares and obtain the trend values.

(b) What is montly increase in production ?

6. What is moving average? Describe how it is useful in the analysis of economic series data ?

7. What are the components of a time sereis ? Explain briefly, the method of moving averages for determination of trend.

8. Describe in brief the various methods used in estimating the secular trend.

9. Apply the method of semi-averages for determining the trend of the following data.

Year :	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
Sales : (Lakhs Rs.)	18	20	17	9	11	27	14	24	21	19

10. State the two methods of determining secular trend. Apply these methods in finding out the trend values in the following data.

Year :	1963	1964	1965	1966	1967	1968
Index No. :	100	111	101	112	105	116

11. Describe the method of least squares for measuring secular trend in time series. Fit a straight line trend to the following data by the method of least squares.

Year :	1981	1982	1983	1984	1985	1986	1987
Production of : wheat (10000 tons)	10	12	14	15	18	21	26

12. Plot the following data and 4 yearly moving average trends on a graph paper.

Years	Index No.	Years	Index No.	Years	Index No.
1946	154	1954	201	1962	316
1947	167	1955	237	1963	357
1948	206	1956	194	1964	247
1949	202	1957	206	1965	295
1950	171	1958	270	1966	291
1951	154	1959	342	1967	277
1952	172	1960	309	1968	266
1953	198	1961	360	1969	283

13. The data below gives the rate of gross capital formation from 1966 - 1982 in Rs. crores prepare the 5 yearly moving average.

Year :	1966	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82
Gross : capital formation (Rs. crores)	19.3	20.9	17.8	16.1	17.6	17.8	18.3	17.3	21.4	19.3	18.1	19.5	19.2	22.2	20.9	21.5	21.9

LESSON - 14

SEASONAL INDICES - OTHER COMPONENTS

OBJECT OF THE LESSON

After studying this lesson the student is expected to have a :

* Clear comprehension of the theory and the practical utility about the concepts of seasonal Indices - other components of time series and their applications.

STRUCTURE OF THE LESSON

The consists of sections as detailed below.

14.1 Measurement of Seasonal variations or Indices

- a) Method of simple Averages
- b) Ratio to trend Method
- c) Ratio to moving Average method
- d) Link Relative method

14.2 Workedout Examples

14.3 Exercise

14.1 MEASUREMENT OF SEASONAL VARIATIONS OR INDICES

The seasonal variations in a time series data are those short-term variations contained within a year that recur regularly. They are periodic in nature and the periods involved may be days, weeks, months, quarters or the season like summer, rainy, winter etc., Seasonal variations generally refer to the annual repetitive patterns of economic activity. Most economic and business time series pass through a more or less regular oscillation with in each twelve month period as the changing seasons and consumer habits come to bear on the behaviour of producers, distributors and consumers every where. Seasonal variations are studied for the following purposes.

1) To eliminate the effect of seasonal factors or in otherwords to find out the effect of seasonal forces on a time-series data.

2) To deseasonalise or eliminate the seasonal variations from the time series, so that remaining data in time series, is free from the effect of seasonal factors.

3) To study the pattern of seasonal variations themselves.

4) To help short-term forecasting and planning.

Once a time series is deseasonalised it is relatively easy to study the effects of cyclical and irregular variations. Thus, deseasonalisation helps in the process of decomposition of time series into its various components i.e, trend, seasonal variations, cyclical variations and irregular or random variations.

The analysis of seasonal variations requires the elimination of seasonal variations from trend, cyclical and random influences in the time series. However, before using any method, whether additive or multiplicative for measuring seasonal variations one should ensure that the following.

i) It should measure only the seasonal forces in data. It should be free from the influence of trend or cycle that may be present.

ii) It should modify the erratic fluctuations in the data with an acceptable system of averaging.

iii) It should recognize slowly changing seasonal patterns that may be present and modify the index to keep up with these changes.

It should be remembered that when the time series is on an annual basis, seasonal variations do not enter into the analysis. When monthly or otherwise within a year data are collected over a period of years, it is possible to determine the variations in an activity for a periods within a year. Generally, monthly or quarterly data that exhibits the seasonal variations may be of two kinds.

1) An underlying seasonal pattern arising from the existing of recurrent seasonal factors and

2) Irregular or random fluctuations which arise from a wide variety of events that do not occur in any stable pattern.

Following are the commonly used methods of elimination of seasonal variations.

a) Method of simple averages

b) 'Ratio to trend' method

c) 'Ratio to moving average' method

d) 'Link relative' method

A) METHOD OF SIMPLE AVERAGES

This is the simplest of all the methods of measuring seasonality and consists of the following steps.

i) Arrange the data by years and months (or quarters if quarterly data are given).

ii) Compute the average $(\bar{x}_i)(i=1,2,\cdots\cdots,12)$ for the i^{th} month for all the years. [Here i^{th} month, $i=1,2,\cdots\cdots,12$ represents January, February,....., December respectively].

iii) Compute the average $\bar{x}$ of the montly averages

$$\text{i.e., } \bar{x}=\frac{1}{12}\sum_{i=1}^{12}\bar{x}_i$$

iv) Seasonal indices for different months are obtained by expressing monthly averages as percentage of $\bar{x}$.

Seasonal Index for i^{th} month $= \frac{\bar{x}_i}{\bar{x}} \times 100, \ (i = 1, 2, \cdots, 12)$

or $$\text{Seasonal Index} = \frac{\text{Monthly average for the month}}{\text{Average of monthly averages}} \times 100$$

i.e., Seasonal Index for January $= \frac{\text{Monthly Average for the January}}{\text{Average of Monthly Averages}} \times 100$

N.B. 1. If instead of monthly averages, we use monthly totals for all the years, the result remains the same.

2. If we are given quarterly data for different years, then seasonal index for any quarter is obtained as

$$\text{Seasonal index for any quarter} = \frac{\text{Quarterly Average for the quarter}}{\text{Average of Quarterly averages}} \times 100$$

MERITS AND DEMERITS :

Merits :

(a) This method is simple to calculate and easy to understand.

(b) The seasonal indices are used for comparison purposes. Comparing a current year's seasonal variations with the normal pattern, or comparing a current month's production or sales with normal production or sales that should occur in relation to the normal pattern.

DEMERITS :

(a) This method is not a true reflection of the normal seasonal variation because it is obtained from the original data, that is affected not only by seasonal movements but also by the three remaining components of time series. It can not be assumed that these movements have negligible effect and hence, can be ignored. This is a very serious limitation since most of the economic and business time series exhibit definite trends and are affected to a great extent by cycles. Hence, the seasonal indices obtained by this method contain the influence of trend, cyclical and random variations.

(b) Furthermore, the effects of the cycles of the original data are not eliminated by the process of averaging. This depends on duration of cycles and the no. of months for which average is calculated.

(c) Also, this method tries to eliminate the effect of irregular or random variations by averaging the monthly or quarterly values over different years.

Hence, in order to arrive at any meaningful seasonal indices, first of all remove the effect of trend from the data, this is done in 'ratio to trend' and 'ratio to moving average' methods.

B. RATIO TO TREND METHOD

This method is an improvement over the simple averages method and is based on the assumption that seasonal variation for any given month is a constant factor of the trend. The measurement of seasonal variation by this method consists in the following steps :

(i) Obtain the trend values by the least square method by fitting a mathematical curve, straight line or second degree polynomial etc.,

(ii) Express the original data as the percentage of the trend values. Assuming the multiplicative model, these percentages will contain the seaosnal, cyclic and irregular components.

(iii) The cyclic and irregular components are then wiped out by averaging the percentages for different months (quarters) if the data are monthly (quarterly), thus leaving us with indices of seasonal variations. Either arithmatic mean or median can be used for averaging but median is preferred to arithmatic mean since the latter gives undue weigtage to extreme values, which are not primarily due to seasonal swings. If there are few abnormal values, modified mean may be used with advantage.

(iv) Finally these indices, obtained in step (iii) are adjusted to a total 1200 for monthly data or 400 for quarterly data by multiplying them throughout by a constant K given by

$$K = \frac{1200}{\text{Sum of Monthly Indices}} \text{ or } K = \frac{400}{\text{Sum of quarterly Indices}}$$

Merits and Demerits

Merits

(i) Compared to other methods of computing seasonal indices, this method is more rational and logical, as it computes the seasonal variations after eliminating the trend component.

(ii) It has no advantage over the moving average procedure too, because it has a ratio to trend value for each month (or quarter) for which data are available. Thus, the data remains in fact here, which is not possible in moving averages. This is a distinct advantage, specially, when the period of time series is very short.

(iii) It is simple to compute and easy to understand.

Demerits

(i) If the data exhibits pronounced cyclical swings, then seasonal indices based on this method will be more biased than calculated by ratio to moving average method.

(ii) In the time sereis, having fairly large number of observations, the fitting of trend equation by least squares method is not an easy job.

C. RATIO TO MOVING AVERAGE METHOD

As mentioned earlier moving average eliminates periodic movements if the extent of the period of moving average is equal to the period of the oscillatory movements sought to be eliminated. Thus for a monthly data, a 12 month moving average should completely eliminate the seasonal movements if they are of constant pattern and intensity. The method of getting seasonal indices by

moving average involves the following steps :

(i) Calculate the centred 12 month moving average of the data. These moving average values will give estimates of the combined effects of trend and cyclic variations.

(ii) Express the original data (except for 6 months in the beginning and 6 months at the end) as percentages of the centred moving average values. Using multiplicative model, these percentages would then represent the seasonal and irregular components.

(iii) The preliminary seasonal indices are now obtained by eliminating the irregular or random component by averaging these percentages. Either arithmetic mean or median can be used for averaging, but median is preferred to arithmetic mean since the latter gives undue weightage to extreme values which are not primarily due to seasonal swings. If there are few abnormal values, modified mean may be used with advantage.

(iv) The sum of these indices = S (say) will not, in general be 1200. Finally an adjustment is done to make the sum of the indices 1200 by multiplying throughout by a constant factor = 1200/S, i.e.; by expressing the preliminary seasonal indices as the percentage of their arithmetic mean. The resultant gives the desired indices of seasonal variations.

MERITS

(i) Of all the methods of measuring seasonal variations, the ratio to the moving average method is the most satisfactory, flexible and widely used method. These indices do not fluctuate as the indices by the ratio to trend method.

(ii) This method does not completely utilise the data, for example in the case of 12 month moving average seasonal indices cannot be obtained for the first 6 months and for the last 6 months.

Note 1 : The seasonal indices for each month(quarter) of different years are also known as specific seasonal and the average of specific seasonals for each month (quarter) for a number of years are termed as typical seasonals.

2 : If we use additive model of the time series, then the method of moving averages for computing seasonal indices involves the following steps. Here we state the steps for monthly data and these can be modified accordingly for quarterly and other data.

DEMERITS

(i) Obtain 12 month moving average values. These will contain trend and cyclic components, i.e., they will represent $(T + C)$

(ii) Trend eliminated values are obtained on subtracting each moving average values from the given time series values to give

$$U_t - \text{M.A. value} = (T + S + C + I) - (T + C) = S + I$$

(iii) Irregular component is eliminated on averaging these $(S + I)$ values for each month over different years and we get the preliminary indices for each month.

(iv) Sum of the indices should be zero. In case it is not. the preliminary indices in step (iii), are adjusted to a total of zero by subtracting from each of them a constant factor.

$$\text{i.e., } \frac{K}{12}[\text{Sum of monthly seasonal indices}]$$

D. LINK RELATIVE METHOD

This method, also known as Pearson's method is based an averaging the link relatives. Link relative is the value of one season expressed as a percentage of the preceeding season. Here the word 'season' refers to time period, it would mean month for monthly data, quarter for quarterly data etc., Thus for monthly data :

$$\text{Link relative for any month} = \frac{\text{Currently month's figure}}{\text{Previous month's figure}} \times 100$$

The steps involved in this method are :

(i) Translate the original data into link relatives (L.R.) as explained above.

(ii) As in the case of ratio to trend method, average the link relatives for each month (quarter) if the data are monthly (quarterly). Mean or median may be used for averaging.

(iii) Convert the average (mean or median) link relatives into chain relatives on the base of the first season. Chain relative (C.R.) for any season is obtained on multiplying the link relative of that season by the chain relative of the preceeding season and dividing by 100. Thus for monthly data, the chain relative for first season (month) i.e., for January is taken by 100.

$$\text{C.R. for Febrauary} = \frac{\text{L.R. of February} \times \text{C.R. of January}}{100}$$

$$(\because \text{C.R. of Jan} = 100)$$

$$\text{C.R. for March} = \frac{\text{L.R. of March} \times \text{C.R. for February}}{100}$$

$$\text{C.R. for December} = \frac{\text{L.R. of December} \times \text{C.R. for November}}{100}$$

Now, by taking this December value as a base, a new chain relative for January can be obtained as

$$\frac{\text{L.R. of January} \times \text{C.R. for December}}{100}$$

Usually, this will not be 100 due to trend and so we have to adjust the chain relatives for trend.

(iv) This adjustment is done by subtracting a 'correction factor' from each chain relative. If we write

$$d = \frac{1}{12}[\text{second (new) C.R. for January} - 100]$$

Then assuming linear trend, the correction factor for February, March, ... December is $d, 2d, \cdots, 11d$ respectively.

(v) Finally, adjust the corrected chain relatives to total 1200 by expressing the corrected chain relatives as percentages of their arithmetic mean. The resultant gives the adjusted monthly indices of seasonal variations.

MERITS AND DEMERITS

(i) The link relatives averaged together contain both the trend and cyclic components. Although the trend is subsequently eliminated by applying correction, the method is effective only if the growth is constant amount or constant rate.

(ii) Though not so simple as the moving average method, or so readily adaptable as others to the construction of some or more complex types of seasonal movements, the actual computations of the link relative method are much less extensive.

(iii) This method utilises data more completely than moving average method. There is only one less link relative while a 12-month moving average results in cut of six months at each end.

14.2 WORKEDOUT EXAMPLES

Example 1 : The data below give the average quarterly prices of a commodity for four years. What are the seasonal indices for various quarters.

Year	I Quarter	II Quarter	III Quarter	IV Quarter
1977	3.7	4.1	3.3	3.5
1978	3.7	3.9	3.6	3.6
1979	4.0	4.1	3.3	3.1
1980	3.3	4.4	4.0	4.0

Solution : Assuming that the trend is absent in the above data the differences in the averages of various quarters (if there is any) will be due to seasonal changes.

Year	I Quarter	II Quarter	III Quarter	IV Quarter
1977	3.7	4.1	3.3	3.5
1978	3.7	3..9	3.6	3.6
1979	4.0	4.1	3.3	3.1
1980	3.3	4.4	4.0	4.0
Total	14.7	16.5	14.2	14.2
Average	3.675	4.125	3.550	3.550
Seasonal Index	98.7	110.8	95.3	95.3

Average of Quarterly averages $= \frac{3.675 + 4.125 + 3.550 + 3.550}{4}$

$= \frac{14.9}{4} = 3.725$

Seasonal Index for first quarter $= \frac{3.675}{3.725} \times 100 = 98.7$

Seasonal Index for second quarter $= \frac{4.125}{3.725} \times 100 = 110.8$

Seasonal Index for (IIIrd) third quarter $= \frac{3.550}{3.725} \times 100 = 95.3$

Seasonal Index for fourth quarter $= \frac{3.550}{3.725} \times 100 = 95.3$

Example 2 : Use the method of monthly averages to determine the monthly indices for the following data of production of a commodity for the year 1979, 1980, 1981 : Production in lakhs of Tonnes given for 1979, 1980, 1981.

Month	**Jan**	**Feb**	**Mar**	**April**	**May**	**June**	**July**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
1979	12	11	10	14	15	15	16	13	11	10	12	15
1980	15	14	13	16	16	15	17	12	13	12	13	14
1981	16	15	14	16	15	17	16	13	10	10	11	15

Solution : Computation of Seasonal Indices.

Month	Production in Lakhs of Tonnes			Total	3 Yearly	Seasonal	Indices
	1979	1980	1981	Monthly	Average	$\frac{(5)}{41} \times 100$	$\frac{(6)}{13.6} \times 100$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
January	12	15	16	43	14.3	104.88	104.88
February	11	14	15	40	13.3	97.56	97.56
March	10	13	14	37	12.3	90.24	90.24
April	14	16	16	46	15.3	112.20	112.20
May	15	16	15	46	15.3	112.20	112.20
June	15	15	17	47	15.6	114.63	114.63
July	16	17	16	49	16.3	119.51	119.51

August	13	12	13	38	12.6	92.68	92.68
September	11	13	10	34	11.3	82.93	82.93
October	10	12	10	32	10.6	78.05	78.05
November	12	13	11	36	12.0	87.80	87.80
December	15	14	15	44	14.6	107.32	107.32
Total of Monthly Total =				492	163.5	$\frac{1200}{12}=100$	$\frac{1200}{12}=100$
Grand Average =					$\frac{492}{12}=41$		$\frac{163.5}{12}=13.6$

EXPLANATION OF THE TABLE :

(i) Column (5) gives the total for each month for 3 years.

(ii) Column (6) gives the average for each month i.e., column (5) divided by number of years (3)

(iii) Column(8) gives the seasonal index. The average of monthly average which is 13.6 taken equal to 100 and the index for each months average is calculated on that basis

for example, for January $\frac{14.3}{13.6}\times 100 = 104.88$ and so on.

(iv) Column (7) also gives the seasonal index. Here monthly total for 3 years i.e., column (5) is divided by average of monthly total i.e., 41. and multiplied by 100. For example for January seasonal index will be $\frac{43}{41}\times 100 = 104.88$ and so on. Thus the results obtained in (iii) & (iv) are same.

Example 3 : Find the seasonal variation by the ratio-to-trend method from the data given below.

Year	I Quarter	II Quarter	III Quarter	IV Quarter
1977	60	80	72	68
1978	68	104	100	88
1979	80	116	108	96
1980	108	152	136	124
1981	160	184	172	164

Solution : For obtaining the seasonal variations by ratio to trend method : first of all the trend for the yearly data by method of least squares is obtained and then it is converted into quarterly data.

Let our trend line equation be $Y = a + bx$

Hence the normal equations are

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

Computation of Trend by Method of Least Squares

Year	Yearly Total	Yearly Average (Y)	Deviation 1979 = X	X^2	XY	Trend Values Y_c
1977	280	70	- 2	4	- 140	64
1978	360	90	- 1	1	- 90	88
1979	400	100	0	0	0	112
1980	520	130	1	1	130	136
1981	680	170	2	4	340	160
				$\Sigma X^2 = 10$	$\Sigma XY = 240$	

$\Sigma Y = 560$ $\Sigma X = 0$

$N = 5$

$\because \Sigma X = 0,\ a = \frac{\Sigma Y}{N} = \frac{560}{5} = 112$

and $b = \frac{\Sigma XY}{\Sigma X^2} = \frac{240}{10} = 24$

Thus, trend equation is $Y_c = 112 + 24X$

Substituting the values of $X = -2, -1, 0, +1, 2,$ trend values for 1977 to 1981 and are given in last column of the above table.

Now yearly increment is = 24

Therefore quarterly increment $= \frac{24}{4} = 6$

Next we determine the quarterly trend values as follows :

For 1977, the trend value for the middle quarter, i.e., half of second quarter and half of third quarter, is 64 and since the quarterly increment is 6. Therefore the trend value for the second quarter of 1977 would be (64 - 3) = 61 and for the third quarter it would be (64 + 3) or 67. The value for the first quarter 1977 would be (61 - 6) or 55 and for the last quarter (67 + 6) or 73. Similarly, trend values for the various quarter of the other years can be calculated. These values are as follows.

Quarterly Trend Values

Year	I Quarter	II Quarter	III Quarter	IV Quarter
1977	55	61	67	73
1978	79	85	91	97
1979	103	109	115	121
1980	127	133	139	145
1981	151	157	163	169

The given values of time series will now be expressed as percentages of corresponding trend values given above. Thus for 1st quarter of 1977, this percentage would be $\frac{60}{55}\times 100$ or 109.09 for the second quarter it would be $\frac{80}{61}\times 100 = 131.15$. Similarly other values can be calculated and these values are given in the following table.

Given quarterly values as percent of Trend values (Trend eliminated values)

Quarterly Trend Values

Year	I Quarter	II Quarter	III Quarter	IV Quarter
1977	109.09	131.15	107.46	93.15
1978	86.08	122.35	109.89	90.72
1979	77.67	106.42	93.91	79.34
1980	85.04	114.29	97.84	85.52
1981	105.96	117.20	105.52	97.04
Total	463.84	591.41	514.62	445.77
Average	92.77	118.28	102.92	89.15
Adujusted Seasonal Index	92.05	117.36	102.12	88.42

The adjustment of seasonal index is done as

Total of averages = $92.77 + 118.28 + 102.92 + 89.15 = 403.12$

This total is more than 400, hence the correction factor (C.F.) would be $\frac{400}{403.12}$. Each average would be multiplied by $\frac{400}{403.12}$. Thus, the adjusted seasonal index for the first quarter of 1977 would be

$$= \frac{92.77 \times 400}{403.12} = 92.05$$

Similarly other adjusted seasonal indices are calculated.

Example 4 : Apply ratio to moving average method to ascertain seasonal indices from the following data.

Year and Month 1981	No. of persons visiting a place of interest	Year and Month 1982	No. of persons visiting a place of interest	Year and Month 1983	No. of persons visiting a place of interest
January	90	January	100	January	110
February	85	February	89	February	93
March	70	March	74	March	78
April	60	April	62	April	66
May	55	May	55	May	58
June	45	June	47	June	40
July	30	July	30	July	35
August	40	August	43	August	45
September	70	September	65	September	72
October	120	October	127	October	130
November	115	November	118	November	118
December	118	December	120	December	124

Solution : Elimination of trend by moving average

Year and Month	No. of persons visiting a place of interest	12-point moving total	12 point M.A. i.e., $\left(\frac{\text{Col.3}}{12}\right)$	12-point M.A. Centred	Ratio to M.A. $\left[\frac{\text{Col.2}}{\text{Col.5}} \times 100\right]$
1981 Jan	90				
Feb	85				
March	70				
April	60				
May	55				
June	45				
July	30	898	74.83		
Aug	40	908	75.67	75.3	39.8
Sep	70	912	76.00	75.8	52.8

	Oct	120	916	76.33	76.2	91.9
	Nov	115	918	76.50	76.4	157.1
	Dec	118	918	76.50	76.5	150.3
1982	Jan	100	920	76.66	76.66	154.0
	Feb	89	920	76.66	76.7	130.4
	March	74	923	76.91	76.8	115.9
	April	62	918	76.50	76.7	96.5
	May	55	925	77.16	76.8	80.7
	June	47	928	77.33	77.2	71.2
	July	30	930	77.50	77.4	60.7
	Aug	43	940	78.33	77.9	38.5
	Sep	65	944	78.66	78.5	54.8
	Oct	127	948	79.00	78.8	82.5
	Nov	118	952	79.33	79.2	160.4
	Dec	120	955	79.58	79.5	148.4
1983	Jan	110	948	79.00	79.3	151.3
	Feb	93	953	79.41	79.2	138.9
	Mar	78	955	79.58	79.5	117.0
	Apr	66	962	80.16	79.9	97.6
	May	58	965	80.41	80.3	82.2
	June	40	965	80.41	80.4	72.1
	July	35	996	80.75	80.6	49.6
	Aug	45				
	Sept	72				
	Oct	130				
	Nov	118				
	Dec	124				

Computation of Adjusted Seasonal Indices

Month	1981	1982	1983	Seasonal Indices (Arithmetic Average)	Adjusted Seasonal Indices (Seasonal Indices x C.F.)
Jan		130.4	138.9	134.7	135.0
Feb		115.9	117.0	116.5	116.7
March		96.5	97.6	97.1	97.3

April		80.7	82.2	81.5	81.7
May		71.2	72.1	71.7	71.8
June		60.7	49.6	55.2	55.3
July	39.8	38.5		39.2	39.9
Aug	52.8	54.8		53.8	53.3
Sep	91.9	82.5		87.1	87.3
Oct	157.1	160.4		158.8	159.1
Nov	150.3	148.4		149.4	149.7
Dec	154.0	151.3		152.7	153.0
	Total			1197.7	1200.1

Here, correction factor (C.F.) for obtaining adjusted seasonal Indices $= \frac{1200}{1197.7} = 1.0019$

Example 5 : Calculate the seasonal indices by the ratio to moving average method from the following data. Use multiplicative as well as additive models and illustrate the difference between them.

Year	Quarter	Y
1972	I	75
	II	60
	III	54
	IV	59
1973	I	86
	II	65
	III	63
	IV	80
1974	I	90
	II	72
	III	66
	IV	85
1975	I	100
	II	78
	III	72
	IV	93

Solution : Assuming multiplicative model of time series, the trend values are eliminated by expressing the given values (U_t) as a percentage of trend values and are given in the last column of the above table.

Year	Quarter	Pirce U_t	4-quarter Moving Total	Sum of two 4-quarter Moving totals	4-quarter Moving Average $(6)=[(5)\div(8)]$	Ratio to Moving Average $(7)=[(3)\div(6)]\times 100$	U_t – M.A.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1972	I	75					
	II	60					
	III	54	248	507	63.375	85.2071	-9.375
	IV	59	259	523	65.375	90.2485	-6.375
1973	I	86	264	537	67.125	128.1192	18.875
	II	65	273	567	70.875	91.7108	-5.875
	III	63	294	592	74.000	85.1351	-11.000
	IV	80	298	603	75.375	106.1360	4.625
1974	I	90	305	613	76.625	117.4551	13.375
	II	72	308	621	77.625	92.7536	-5.625
	III	66	313	636	79.500	83.0189	-13.500
	IV	85	323	652	81.500	104.2945	3.500
1975	I	100	329	664	83.000	120.4819	17.000
	II	78	335	678	84.750	92.0354	-6.750
	III	72	343				
	IV	93					

Computational of Seasonal Indices

Trend Eliminated Values

Year	I Quarter	II Quarter	III Quarter	IV Quarter	Total
1972	--	--	85.2071	90.2485	
1973	128.1192	91.7108	85.1351	106.1360	
1974	117.4551	92.7536	83.0189	104.2945	
1975	120.4819	92.0354	--	--	
Total	366.0562	276.4998	253.3611	300.6790	

Average (A.M.) (S.I.)	122.0187	92.1666	84.4537	100.226	398.865
Adjusted Seasonal Indices	122.3603	92.4246	84.6902	100.5066	$399.985 \simeq 400$

The seasonal indices obtained as averages (A.M.) above are adjusted to a total of 400, by multiplying each of them by a constant factor.

$$K = \frac{400}{\text{Sum of Seasonal Indices}} = \frac{400}{398.865} = 1.0028$$

If we assume additive model of the time series, then the trend eliminated values, (short term and irregular fluctuations) are obtained on subtracting the trend (M.A.) values from the given time series values i.e., by the formulae.

Short term fluctuations = $U_t - (\text{M.A. values}) = S + I,$ and are given in the last column of the first table. These values are then used to obtain the seasonal indices as explained in the following table.

Computation of Seasonal Indices

Trend Eliminated Values

Year	I Quarter	II Quarter	III Quarter	IV Quarter
1972	--	--	-9.375	-6.375
1973	18.875	-5.875	-11.000	4.625
1974	13.375	-5.625	-13.500	3.500
1975	17.000	-6.750	--	--
Total	49.250	-18.250	-33.875	1.750
Average (A.M.)	16.417	-6.083	-11.292	0.583
(S.I.)				
Adjusted Seasonal Indices	16.511	-5.989	-11.198	0.677

Sum of seasonal Indices (S.I.) = 16.417 - 6.083 - 11.292 + 0.583 = -0.375

Since the sum is not zero, these indices are adjusted to a total of zero by subtracting from each of them a constant factor.

$$K = \frac{\text{Sum of Indices}}{4} = -\frac{0.375}{4} = -0.094$$

Adjusted seasonal indices for Ist quarter $=16.417-(-0.094)=16.511$. Similarly we can obtain the adjusted seasonal Indices for the remaining quarters, which are given in the last row of the above table.

Example 6 : The data below gives the average quarterly prices of a commodity for five years. Calculate the seasonal variation indices by the method of link relatives.

Year/Quarter	1979	1980	1981	1982	1983
I	30	35	31	31	34
II	26	28	29	31	36
III	22	22	28	25	26
IV	31	36	32	35	33

Solution : Calculations for Seasonal Indices by the Method of Link Relatives.

Link Relatives

Year	First Quarter	Second Quarter	Third Quarter	Fourth Quarter
1979	--	86.7	84.6	140.9
1980	112.9	80.0	78.6	163.6
1981	86.1	93.5	96.6	114.3
1982	96.9	100.0	80.7	140.0
1983	97.1	105.9	72.2	126.9
A.M.	(393.0)/4=98.25	(446.1)/5=93.22	(442.7)/5=82.54	(685.7)/5=137.14
Chain Relative	100	$\frac{100\times 93.22}{100}=93.22$	$\frac{93.22\times 82.54}{100}=76.95$	$\frac{76.95\times 137.14}{100}=105.4$
Adjusted Chain Relatives	100	92.345	75.200	102.775
Seasonal Indices	108.02	99.75	81.23	111.00

Explanation :

The second chain relative for first quarter $=\frac{105.4\times 98.25}{100}=103.5$

Correction Factor $d=\frac{1}{4}(103.5-100.0)=\frac{3.5}{4}=0.875$

Adjusted chain relatives are obtained by subtracting 0.875, 2 x 0.875, 3 x 0.875 from the chain relatives of second, third and fourth quarter respectively.

Average of adjusted chain relatives $= \frac{100 + 92.345 + 75.200 + 102.725}{4} = 92.58$

Seasonal Variation Index for any quarter $= \frac{\text{Adjusted C.R.}}{92.58} \times 100$

Seasonal Indices have been obtained in the above table.

Example 7 : Calculate the seasonal variation indices by the method of link relatives for the following data.

Quarterly Data for five years

	Year				
Quarter	**1969**	**1970**	**1971**	**1972**	**1973**
I	45	48	49	52	60
II	54	56	63	65	70
III	72	63	70	75	84
IV	60	56	65	72	86

Solution : Computation of seasonal indices by link relative.

Link Relatives

Year	**Ist Quarter**	**IInd Quarter**	**IIIrd Quarter**	**IVth Quarter**	**Total**
1969	--	120.00	133.33	83.33	
1970	80.0	116.67	112.50	88.89	
1971	87.50	128.57	111.11	92.86	
1972	80.00	125.00	115.38	96.00	
1973	85.71	116.67	120.00	78.57	
Total	333.21	606.91	592.32	439.65	
A.Mean	83.303	121.382	118.464	87.93	
Chain relatives	100.00	121.382	143.793	126.431	
Adjusted Chain relatives	100	121.382-1.33 = 120.052	143.793-1.33 = 142.463	126.438-1.33 =125.108	487.623
Seasonal Indices	100 x 0.82 = 82.0	120.052 x 0.82 = 98.2	142.463 x 0.82 = 116.9	125.108 x 0.82 = 102.6	400

The explanations of above table is

$$\text{Link Relatives for a figure} = \frac{\text{the figure}}{\text{Previous Quarter Figure}} \times 100$$

As such L.R. for this first quarter of the second year $\frac{54}{60} \times 100 = 90$ and similarly for others.

Adjustment for chain relatives has been done on the following basis.

(i) Original C.R. for 1st quarter = 100

(ii) C.R. for Ist quarter calculated on the basis of the C.R. of the last quarter

$$= \frac{126.438 + 83.303}{100} = 105.33$$

(iii) Adjustment factor to be subtracted from C.C. for 2nd, 3rd and 4th quarter would be

$$= \frac{105.33 - 100}{4} = 1.33$$

Seasonal indices have been calculated multiplying the chain relatives by the correction factor which is $\frac{400}{487.622} = 0.82$

14.3 EXERCISES

1. Explain what is meant by seasonal fluctuations of a time series. Discuss the different methods for determining seasonal fluctuations of a given time series. Discuss the relative merits and demerits of each of these methods. Also state the conditions of applicability for each of the methods.

2. Discuss the different methods for obtaining measures of seasonal variation. Discuss their relative merits and demerits. Using the ratio to trend method, determine the quarterly seasonal indices.

Production of Coal (In Million of Tonnes)

Years	I	II	III	IV
1	68	60	61	63
2	70	58	56	60
3	68	63	68	67
4	65	59	56	62
5	40	55	51	58

3. Give various methods of determining seasonal indices for a time series. Estimate the influence of trend, seasonal and random variation in the following data.

Year/Quarter	I	II	III	IV
1978	10	27	21	40
1979	11	35	29	57
1980	14	51	33	74
1981	19	57	43	78
1982	22	67	45	101

4. Explain Ratio to moving average method for determining seasonal index. The following percentages of moving average have been obtained for a dairy's ice cream sales. Determine the seasonal index for each quarter.

	Quarter			
Year	Winter	Spring	Summer	Fall
1969	--	--	156	111
1970	49	92	137	109
1971	53	93	148	108
1972	42	91	162	104
1973	51	89	153	110
1974	51	90	151	112
1975	48	88	--	--

5. What do you understand by seasonal variations in a time-series? Give Examples. Explain the link relative method of computing the indices of seasonal variations.

6. Apply the method of link relatives to compute seasonal indices from the following data.

Year	1	2	3	4	5	6	7	8	9	10	11	12
1962	120	115	110	100	90	80	90	100	110	120	130	140
1963	150	140	130	120	110	100	110	120	130	140	150	160
1964	170	160	150	140	130	120	130	140	150	160	170	170

7. The following data give the sales of a company (million in Rs.) during 1966-70. Compute seasonal indices by the method of simple averages.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1966	3	4	9	10	8	6	5	6	7	11	10	14
1967	5	6	7	10	10	6	8	7	8	10	11	15
1968	4	7	8	12	9	8	10	6	9	12	11	17
1969	7	5	8	13	11	7	8	8	7	11	11	17
1970	11	13	13	15	17	13	14	8	9	16	22	27

8. Using the data given below, compute seasonal indices by applying ratio to moving average method.

Cement Production (1000 tons)

	Quarter			
Year	**I**	**II**	**III**	**IV**
1968	83	81	98	114
1969	124	113	115	152
1970	163	162	168	175
1971	191	180	184	167

9. What do you understand by seasonal variation ? Explain with examples. Describe the method of ratio to moving average for finding seasonal indices ?

LESSON - 15

INDEX NUMBERS

(Simple and Weighted)

LEARNING OBJECTIVES

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

- □ Comprehend the theory and the practical utility of the concepts of simple and weighted Index numbers.

LESSON OUTILINE

15.1 INTRODUCTION

Historically, the first index was constructed in 1764 to compare the Italian price index in 1750 with the price level in 1500. Though originally developed for measuring the effect of change in prices, index numbers have today become one of the most widely used statistical devices and there is hardly any field where they are not used. Newspapers headline the fact that prices are going up or down, that industrial production is rising or falling, that imports are increasing or decreasing, that crimes are rising in a particular period compared to the previous period as disclosed by index numbers. They are used to feel the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies. Infact, they are described as 'barometers of economic activity' i.e. if one wants to get an idea as to what is happening to economy, one should look at important indices like the index number of industrial production, agricultural production, business activity, etc.,

Index Number Defined

Index numbers can be classified into the following three broad categories.

1) A measure of change

* It is a numerical value characterizing the change in complex economic phenomenon over a period of time or space ***--- Maslow***

* An index number is a quantity which, by reference to a base period shows by its variations the changes in the magnitude over a period of time. In general, index

numbers are used to measure changes over time in magnitudes which are not capable of direct measurement.

--- John I Raffin.

* An index number is a statistical measure designed to show changes in a variable or a group of related variables with respect to time, geographic location or other characteristics. ***-- Speigel***

* Index number is a single ratio (usually in % or percentage) which measures the combined i.e.; averaged change of several variables between two different times, places or situations. ***-- A.M. Tuttle***

A DEVICE TO MEASURE CHANGE

* Index numbers are devices measuring differences in the magnitude of a group of related variables. ***-- Cosxton and Cowden***

* An index number is a device which shows by its variation the changes in a magnitude which is not capable of accurate measurement in itself or direct valuation in practice.

-- Wheldom

3. A Series Representing The process of Change

* Index numbers are series of numbers by which, changes in the magnitude of phenomenon are measured from time to time or place to place. ***-- Horace Secris***

* A series of index numbers reflects in its trend and fluctuations the movements of some quantity to which it is related ***-- B.L. Bowley***

* An index number is a statistical measure of fluctuations in a variable arranged in the form of series, and using a base period for making comparisons.

-- L.J. Kaplan

It is clear from the above definitions that an index number is a specialized average designed to measure the change in a group of related variables over a period of time. Thus, when we say that the index number of wholesale prices is 112 for January 2000 compared to January 1999, it means that there is a net increase in the prices of wholesale commodities to the extent of 12 per cent during the year.

Index numbers, in effect, relate a variable or variables in a given period to the same variable or variables in another period, called the base period. An index, the simplified name for index numbers, which is computed from a single variable, is called a univariate index, where as an index which is constructed from a group of variables is considered a composite index.

For a proper understanding of the term index number, the following points are worth considering.

(i) Index numbers are specialized averages : As explained in the chapter on measures of central value, an average is a single figure representing a group of figures. However, to obtain an average the items must be comparable; for example, the average weight of men, women and children of a certain locality has no meaning at all. Furthermore, the unit of measurement must be the same for all the items. Thus, an average of the weight expressed in Kg; Lb., etc., has

no meaning. However, this is not so with index numbers. Index numbers are used for the purposes of comparison in situations where two or more series are expressed in different units or the series are composed of different types of items. For example while constructing a consumer price index the various items are divided into broad heads, namely (i) Food, (ii) Clothing (iii) Fuel and lighting (iv) House Rent and (v) Miscellaneous. These items are expressed in different units : Thus, under the head 'food' wheat and rice may be quoted per quintal, ghee per Kg. etc. Similarly, cloth may be measured in terms of metres. An average of all these items expressed in different units is obtained by using the technique of index number.

(ii) Index numbers measure the change in a group of related variables. Since index numbers are essentially averages they descrbe in one single figure the increase or decrease in a group of related variables under study. The group of variables may be the prices of a specified set of commodities, the volume of production in different sectors etc., Thus, if the consumer price index of working class for Delhi has gone up to 113 in February 2000 compared to February 1999 it means that there is a net increase of 13 per cent in the prices of commodities included in the index. Similarly, if the index of industrial production is 118 in 2000 compared to 1999 it means that there is a net increase in industrial production to the extent of 18 per cent. It should be carefully noted that even where an index is showing a net increase, it may include some items which have actually decreased in value and others which have remained constant.

(iii) Index numbers measure the effect of changes over a period of time. Index numbers are most widely used for measuring changes over a period of time. Thus, we can find out the net change in agricultural prices from the beginning of First plan period to the end of the Eighth plan period. i.e, from 1951 to 1996. Similarly, we can compare the agricultural production, industrial production, imports, exports, wages etc., at two different times. However, it should be noted that index numbers not only measure changes over a period of time but also compare economic conditions of different locations, different industries, different cities or different countries. But since the basic problems are essentially the same and most of the important index numbers published by the Government and private research organisations refer to data collected at different times. We shall consider in this chapter index numbers measuring changes relative to time only. However, methods described can be applied to other areas also.

15.2 USES OF INDEX NUMBERS

Index numbers are indispensable tools of economic and business analysis. Their significance can be best appreciated by the following points. (i) They help in framing suitable policies. Many of the economic and business policies are guided by index numbers. For example, while deciding the increase in dearness allowance of the employees, the employers have to depend primarily upon the cost of living index. If wages and salaries are not adjusted in accordance with the cost of living, very often it leads to strikes and lock-outs which in turn cause considerable waste of resources. The index numbers provide some guideposts that one can use in making decisions.

Though index numbers are most widely used in the evaluation of business and economic conditions, there is a large number of other fields also where index numbers are useful. For example, sociologists may speak of population indices, psychologists measure intelligence quotients which are essentially index numbers comparing a person's intelligence score, with that of an average for his or her age; health authorities prepare indices to display changes in the adequacy of hospital facilities and educational research organisations have devised formulae to measure changes in the effectiveness of school systems. (ii) They reveal trends and tendencies. Since index numbers

are most widely used for measuring changes over a period of time, the time series so formed enable us to study the general trend of the phenomenon under study. For example , by examining index number of imports for India for the last 8-10 years we can say that our imports are showing an upward tendency. i.e., they are rising year after year. Similarly, by examining the index numbers of industrial production, business activtiy etc; for the last few years we can conclude about the trend of production and business activity. By examining the trend of the phenomenon under study we can draw very important conclusions as to how much change is taking place due to the effect of seasonality, cyclical forces, irregular forces, etc. Thus index numbers are highly useful in studying the general business conditions. (iii) They are important in forecasting future economic activity. Index numbers are useful not only in studying the past and present workings of our economy, but they are also important in forecasting future economic activity. Index numbers then are often used in time series analysis, the historical study of long-term trend, seasonal variations and business cycle development, so that business leaders may keep pace with changing economic and business conditions and have better information available for decision - making purposes. (iv) Index numbers are very useful in deflating. Index numbes are higly useful in deflating, i.e., they are used to adjust the original data for price changes, or to adjust wages for cost of living changes and thus transform nominal wages into real wages. Moreover, nominal income can be transformed into real income and nominal sales into real sales through appropriate index numbers.

CLASSIFICATION OF INDEX NUMBERS

Index numbers may be classified in terms of what they measure. In economics and business the classifications are 1) Price; 2) quantity; 3) Value; 4) Special Purpose.

Only price and quantity index numbers are discussed in detail. The others will be mentioned, but without detail, of how to construct them since both value and special purpose index numbers do not offer new problems in construction since the method of construction of various types of index numbers can be understood if the technique of constructing price index number is clear. Hence, we shall devote greater attention to the price index numbers.

15.3 PROBLEMS IN THE CONSTRUCTION OF INDEX NO.

Before constructing index numbers a careful thought must be given to the following problems.

1. The Purpose of the Index : At the very outset the purpose of constructing the index must be very clearly decided-what the index is to measure and why? There is no all-purpose index. Every index is of limited and particular use. Thus, a price index that is intended to measure consumer's prices must not include wholesale prices. And if such an index is intended to measure the cost of living of poor families, great care should be taken not to include goods ordinarily used by middle class and upper income groups. Failure to decide clearly the purpose of the index would lead to confusion and wastage of time with no fruitful results. Other problems such as the base year, the number of commodities to be included, the prices of the commodities etc., are to be decided in the light of the purpose for which the index is being constructed.

The problem of the scope of the index, i.e., the field covered by the index, is bound up with the purpose of the index and the data available. The data available, or rather the lack of it, may necessitate the modification of the purpose.

2. Selection of a Base Period : Whenever index numbers are constructed a reference is made to some base period. The base period of an index number (also called the reference period) is the

period against which comparisons are made. It may be a year, a month or a day. The index for base period is always taken as 100. Though the selection of the base period would primarily depend upon the object of the index, the following points need careful consideration of base period.

(i) The base period should be a normal one, i.e., it should be free from abnormalities like wars, earthquakes, famines, booms, depressions, etc. However, at times it is really difficult to select a year which is normal in all respects - a year which is normal in one respect may be abnormal in another. To solve this problem an average of a number of years, 3 or 4 (preferably covering one complete cycle), may be taken as the base. The process of averaging will reduce the effect of extremes. Thus, the average of the period from 2000 to 2002 may be considered normal whereas no individual year in that span can be considered normal.

(ii) The base period should not be too distant in the past. It is desirable to have an index based on a fairly recent period, since comparisons with a familiar set of circumstances are more helpful than comparisons with vaguely remembered conditions. For example, for deciding increase in dearness allowance at present there is no advantage in taking 1970 or 1980 as the base : the comparison should be with the preceding year or the year after which dearness allowance has not been revised.

(iii) Fixed base or chain base : While selecting the base a decision has to be made as to whether the base shall remain fixed or not. i.e., whether we have a fixed base or chain base index. In the fixed base method, the year or the period of years to which all other prices are related is constant for all times. On the otherhand, in the chain base method the prices of a year are linked with those of the preceeding year and not with the fixed year. Naturally the chain base method gives a better picture than what is obtained by the fixed base method. However, much would depend upon the purpose of constructing the index.

3. Selection of Number Items : The items included in an index should be determined by the purpose for which the index is constructed; every item cannot be included while constructing an index number and hence one has to select a sample. For example, while constructing a price index it is impossible to include each and every commoditiy. Hence which it is necessary to decide commodities are to be included. These commodities should be selected in such a manner that they are representative of the tastes, habits and customs of the people for whom the index is meant. Thus, in a consumer price index for the working class, items like scooters, motor cars, refrigerators, consmetics, etc., find no place. A decision must also be made on the number of commodities to be included and their qualities. Here we should note that the larger the number of commodities included, the more representative shall be the index but at the same time the greater shall be the cost and the time taken. The purpose of the index shall help in deciding the number of commodities. Thus, in a general price index a larger number of commodities shall have to be included as compared to a specific purpose index as the index number of the prices of food grains or industrial raw materials.

It is also necessary to decide the grade or quality of the items to be included in the index. Index numbers shall give wrong result if at one time one set of qualitites is included and at another time another set. To avoid confusion about qualities it is desirable that as far as possible, only standardized or graded items are included so that they can be easily identified after a time lapse.

4. Price Quotations : After the commodities have been selected, the next problem is to obtain price, quotations for these commodities. It is a well known fact that prices of many commodities

vary from place to place and even from shop to shop in the same market. It is impracticable to obtain price quotations from all the places where a commodity is dealt with. A selection must be made of representative places and persons. These places should be those which are well known for trading for that particular commodity. After the places from where the price quotations are to be obtained is decided, the next thing is to appoint some person or institution who can supply price quotations as and when required. Great care must be exercised to see that price reporting agency is unbaised. In order to check the inaccuracy of price quotations supplied by an agency quotations are obtained from more than one agency. If there is some reliable journal or magazine supplying price quotations then it may be utilised.

In order to ensure uniformity in the manner in which prices are to be quoted must also be decided. There are two methods of quoting prices. (i) money prices, and (ii) quantity prices. In the former case prices are quoted per unit of commodity, for example sugar Rs. 500 per quintal (100 Kg) and in the latter case prices are quoted per unit of money. Thus sugar may be quoted as 1/5 Kg for one rupee. The former method is free from confusion and is generally adopted while quoting prices.

A decision must also be made as to whether the wholesale prices or retail prices are required. The choice would depend upon the purpose of the index. Thus in a consumer price index the wholesale price shall not be representative at all. If the prices of certain commodities are controlled by the government then it is these controlled prices that should be taken into account and not the black market prices which may be much higher.

5. Choice of an Average : Since index numbers are specialized averages a decision has to be made as to which particular average (i.e. arithmetic mean, median, mode, geometric mean or harmonic mean) should be used for constructing the index. Median, mode and mean are almost never used in the construction of index numbers. Basically a choice has to be made between arithmetic mean and geometric mean. Theoritically speaking, geometric mean is the best average in the construction of index numbers because of the following reasons : (i) in the construction of index numbers, we are concerned with ratios of change; (ii) geometric mean is less susceptible to major variations as a result of violent fluctuations in the values of the individual items; and (iii) index numbers calculated by using this average are reversible and therefore base shifting is easily possible. The geometric mean index always satisfies the time reversal test.

Despite theoritical justification for favouring geometric mean, arithmetic mean is more popularly used while constructing index numbers. This is for the reason that arithmetic mean is much more simple to compute than the geometric mean. However, where ever possible in the interest of greater accuracy, geometric mean should be preferred. It is gratifying to note that with the growing use of electronic computers geometric mean is becoming more popular in constructing index numbers.

6. Selection of Appropriate Weights : The problem of selecting suitable weights is quite important and at the same time quite difficult to decide. The term 'weight' refers to the relative importance of the different items in the construction of the index. All terms are not of equal importance and hence it is necessary to devise some suitable method whereby the varying importance of the different items is taken into account. This is done by allocating weights. Thus, we have broadly two types of indices, unweighted indices and weighted indices. In the former case, no specific weights are assigned whereas in the latter case specific weights are assigned to various items. It may be pointed out here that no index is unweighted in the strict sense of the term as weights implicitly

enter in unweighted indices because we are giving equal importance to all the items and hence weights are unity. It is, therefore, necessary to adopt some suitable method of weighting so that arbitrary and haphazard weights may not affect the results.

There are two methods of assigning weights : (i) implicit, and (ii) explicit. In implicit weighing, a commidity or its variety is included in the index a number of times. Thus, if wheat is to be given in an index twice as much weight as rice, then two varieties of wheat against one of rice may be included in the sereis. On the other hand, in the case of explicit weighting some outward evidence of importance of the various items in the index is given. When the explicit weights are assigned the questions are : (i) By what do we weigh ? and (ii) what type of weight do we use?

(i) In order to bring out the economic importance of the commodities involved the weight can be production figures, consumption figures or distribution figures.

(ii) Weights are of two types : quantity weights and value weights. A quantity weight, symbolised by q, means the amount of commodity produced, distributed, or consumed in some time period. A value weight, on the other hand, combines price with quantity 'produced', 'distributed' or 'consumed' value is in terms of rupees and is symbolised by p x q where p stands for the price and q for the quantity.

Now the question is whether to choose quantity weights or value weights. If the aggregate method is used while constructing index, then quantities are used as weights because price times quantity will always give the same units, namely rupees. On the other hand in averaging price relatives quantity figures cannot be used. It is for this reason that if we multiply percentages by quantities expressed in different units, we get results in different units; for example, percentage times tonnes will give tonnes and percentage times Kg. will give Kg. such figures cannot be used in computation. But if percentages are multiplied by value figures, which are always expressed in rupees, we get answer in rupees only. Hence the statistician will use q as a weight in the method of aggregating actual prices and must use p x q as a weight in the method of averaging price relatives.

Another problem in connection with weights is that of deciding whether the weight shall be fixed or fluctuating. Since the relative importance of the different items does not remain the same for all times it is logical to vary the weight from time to time. Such an index would give better results. However, when fluctuating weights are used one must be very careful in interpreting the index because not only changes in prices but also changes in weights are affecting the index.

7. Selection of An Appropriate Formula : A large number of formulae have been devised for constructing the index. The problem very often is that of selecting the most appropriate formula. The choice of the formula would depend not only on the purpose of the index but also on the data available. Prof. Irving Fisher has suggested that an appropriate index is that which satisfies time reversal test and factor reversal test. Theoretically, Fisher's method is considered as "ideal" for constructing index number. However, from a practical point of view there are certain limitations of this index which shall be discussed later. As such, no one particular formula can be regarded as the best under all circumstances. On the basis of this knowledge of the characteristics of different formulae, a discriminating investigator will choose technical methods adapted to his data and appropriate to his purposes.

None of the above problems is simple to solve in practice and the final index is usually the product of compromise between theoretical standards and the standards attainable with the given

data.

15.4 METHODS OF CONSTRUCTING INDEX NUMBERS

A large number of formulae have been devised for constructing index numbers. Broadly speaking, they can be grouped under two heads;

(a) Unweighted indices; and

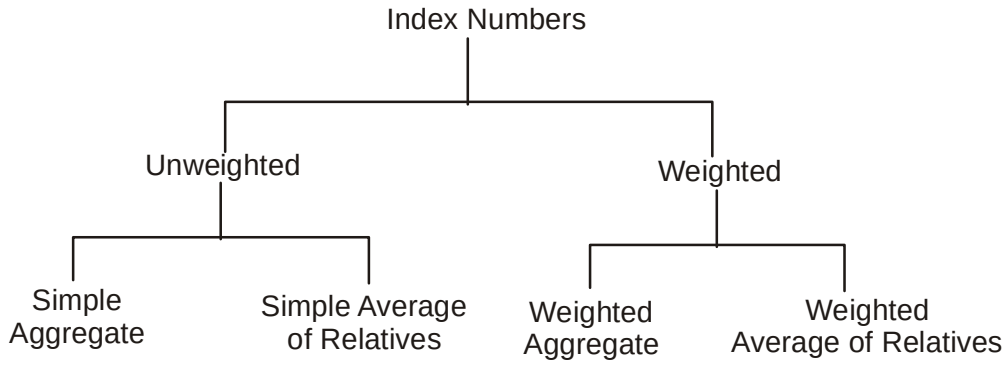
(b) Weighted indices

In the unweighted indices weights are not expressly assigned whereas in the weighted indices weights are assigned to the various items. Each of these types may be further divided under two heads;

(i) Simple aggregate, and

(ii) Simple average of Relatives

The following chart illustrates the various methods :



Unweighted Index Numbers

1. Simple Aggegate Method : This is the simplest method of constructing index numbers. When this method is used to construct a price index the total of current year prices for the various commodities in question is divided by the total of base year prices and the quotient is multiplied by 100. Symbolically :

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

$\Sigma P_1 =$ Total of current year prices for various commodities.

$\Sigma P_0 =$ Total of base year prices for various commodities.

This method of constructing the index is the simplest of all the methods.

The steps required in computation are :

(i) Add the current year prices for various commodities i.e., ΣP_1

(ii) Add the base year prices for the same commodities. i.e., ΣP_0

(iii) Divide ΣP_1 by ΣP_0 and multiply the quotient by 100.

Illustration 1 : From the following data construct an index for 1999 taking 1998 as base :

Commodity	Price in 1998(Rs.)	Price in 1999(Rs.)
A	50	70
B	40	60
C	80	90
D	110	120
E	20	20

Solution : Construction of Price Index

Commodity	Price in 1998	Price in 1999
	P_0	P_1
A	50	70
B	40	60
C	80	90
D	110	120
E	20	20
	$\Sigma P_0 = 300$	$\Sigma P_1 = 360$

$$P_{01} = \frac{\Sigma P_1}{\Sigma P_0} \times 100 = \frac{360}{300} \times 100 = 120$$

This means that as compared to 1998, in 1999 there is a net increase in the prices of commodities included in the index to the extent of 20%

Limitations : There are two main limitations of the simple aggregate index :

* The units used in the price or quantity qutotations can exert a big influence on the value of the index.

* No consideration is given to the relative importance of the commodities.

SIMPLE AVERAGE OF PRICE RELATIVES METHOD :

When this method is used to construct a price index, first of all price relatives are obtained

for the various items included in the index and then average of these relatives is obtained using any one of the measures of central value i.e. arithmetic mean, median, mode, geometric mean or harmonic mean. When arithmetic mean is used for averaging the relatives, the formula for computing the index is

$$P_{01} = \frac{\Sigma\left(\frac{P_1}{P_0} \times 100\right)}{N}$$

Where N refers to the number of items (commodities) whose price relatives are thus averaged.

Although any measure of central value can be used to obtain the overall index, price relatives are generally averaged either by the arithmetic or the geometric mean. When geometric mean is used for averaging the price relatives the formula for obtaining the index becomes

$$\log P_{01} = \frac{\Sigma \log\left(\frac{P_1}{P_0} \times 100\right)}{N} \quad \text{or} \quad \frac{\Sigma \log P}{N}$$

where $P = \frac{P_1}{P_0} \times 100$

$$P_{01} = \text{anti}\log\left[\frac{\Sigma \log\left(\frac{P_1}{P_0} \times 100\right)}{N}\right] = \text{anti}\log\frac{\Sigma \log P}{N}$$

Other measures of central value are not in common use for averaging relatives.

Illustration : From the following data construct an index for 2001 taking 2000 as base by the average of relatives method using (a) arithmetic mean, and (b) geometric mean for averaged relatives.

Commodity	Price in 2000 (Rs.)	Price in 2001 (Rs.)
A	50	70
B	40	60
C	80	90
D	110	120
E	20	20

Solution : (a) Index numbers using arithmetic mean of price relatives.

Commodity	Price in 2000(Rs.)	Price in 2001(Rs.)	Price relatives
	P_0	P_1	$\frac{P_1}{P_0}\times 100$
A	50	70	140.0
B	40	60	150.0
C	80	90	112.5
D	110	120	109.1
E	20	20	100.0

$$\frac{\Sigma P_1}{P_0}\times 100 = 611.6$$

$$P_{01} = \frac{\Sigma \frac{P_1}{P_0}\times 100}{N} = \frac{611.6}{5} = 122.32$$

(b) Index numbers using geometric mean of price relatives.

Commodity	Price in 2000 (P_0)	Price in 2001 (P_1)	Price Relatives P	$\log P$
A	50	70	14.0	2.1461
B	40	60	150.0	2.1761
C	80	90	112.5	2.0512
D	110	120	119.1	2.0378
E	20	20	100.0	2.0000

$$\sum \log p = 10.4112$$

$$P_{01} = \text{Anti}\log\left[\frac{\Sigma \log P}{N}\right] = \text{Anti}\log\left[\frac{10.4112}{5}\right] = 120.9$$

Merits : This method has the following two advantages over the previous method.

* Extreme items do not influence the index. Equal importance is given to all the times.
* The index is not influenced by the units in which prices are quoted or by the absolute level of individual prices. Relatives are pure numbers and are therefore, divorced from the original units. Consequently, index number computed by the relatives method would be the same regardless of the way in which prices are quoted. This simple average of price relatives is said to meet what is called the unit test.

Limitations : Despite these merits this method is not very satisfactory because of two reasons :

* Difficulty is faced with regard to the selection of an appropriate average. The use of the arithmetic mean is considered as questionable sometimes because it has an upward bias. Other averages are almost never used while constructing index no's.
* The relatives are assumed to have equal importance. This is again a kind of concealed weighing system that is highly objectionable since economically some relatives are more important than others.

Weighted Index Numbers :

The so-called un-weighted index numbers discussed above are un-weighted in the true sense of the term. They assign equal importance to all the items included in the index and as such they are in reality weighted, weights being implicit rather than explicit. As discussed earlier in case of unweighted indices it is possible to get different results by changing the implicit weighing for the unweighted index that is far from realistic in most of the cases. Construction of useful index numbers requires a conscious effort to assign to each commodity a weight in accordance with its importance in the total phenomenon that the index is supposed to describe.

* Weighted index are of two types.
* Weighted aggregate indices, and
* Weighted Average of Relatives.

Weighted Aggregative Indices : These indices are of simple aggregate type with the fundamental difference that weights are assigned to the various items included in various methods of assigning weights and consequently a large number of formulae for constructing index numbers have been devised of which some of the more important ones are

* Laspeyres method
* Paasche method
* Dorbish and Bowley's method
* Fisher's ideal method
* Marshall-Edgeworth method, and
* Kelly's method

All these methods are named after the person who suggested them.

(i) Laspeyres Method : The Laspeyers price index is a weighted aggregate price index, where the weights are determined by quantities in the base period. The formula for constructing the index is :

$$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

Steps :

* Multiply the current year prices of various commodities with base year weights and obtain $\Sigma p_1 q_0$.
* Multiply the base year prices of various commodities with base year weights and obtain $\Sigma p_0 q_0$.
* Divide $\Sigma p_1 q_0$ by $\Sigma p_0 q_0$ and multiply the quotient by 100. This gives us the price index.

Laspeyres index attempts to answere the question : What is the change in aggregate value of the base period list of goods when valued at given period prices ? This index is very widely used in practical work. The primary disadvantage of the Laspeyres method is that it does not take into consideration the consumption pattern, the Laspeyers index has an upward bias. When prices increase, there is a tendency to reduce the consumption of higher priced items. Hence, by using base year weights, too much weight will be given to those items which have increased in price similar, when prices decline, to consumers shift their purchases to those items which decline the most. By using base period weights, too little weight is given to those items which decrease most in price, again overstating the index.

(ii) Paasche's Method : The Paasche price index is a weighted aggregate price index in which the weights are determined by quantities in the given year. The formulae for constructing the index is

$$p_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

Steps :

* Multiply the current year prices of various commodities with current year weights and obtain $\Sigma p_0 q_1$.
* Multiply the base year prices of various commodities with current year weights and obtain $\Sigma p_0 q_0$.
* Divide $\Sigma p_1 q_1$ by $\Sigma p_0 q_1$ multiply the quotient by 100.

In general this formula answers the question : What would be the value of the given period list of goods when value d at base-period prices?

The difficulty in computing the Paasche index in practice is that revised weights, or quantities, must be computed each year or for each period, adding to the data collection expenses in the preparation of the index. For this reason, the Paasche index is not used, frequently in practice where the no. of commodities is large.

Comparison of Laspeyres and Paasche methods : Laspeyres index measures change in a "fixed market basket" of goods and services. The same quantities are used in each period. The Paasche index continually updates the quantities to the levels of current consumption. These two

approaches tend do produce opposite extremes in index values computed from the same data. From a practical point of view, Laspayres index is often preferred to Paasche's for the simple reason that in Laspayres index weights (q_0) are the base year quantities and do not change from one period to the next. On the other hand, use of Paasche index requires the continuous use of new quantity weights for each period considered and in most cases of these weights are difficult and expensive to obtain. In most countries index numbers are constructed by using Laspeyers formula.

An interesting property of Laspeyres and Paasche indices is that the former is generally expected to overestimate or to leave an upward bias whereas the latter tends to underestimate. i.e., shows a downward bias. When the prices increase there is usually a reduction in the consumption of those items for which the increase has been the most pronounced and hence, by using base year quantities we will be giving too much weight to the prices that have increased the most and the numerators of the Laspeyres index will be too large when the prices go down. Consumers often shift their preference to those items which have declined the most and, hence by using base year quantities we will be giving too much weight to the prices that have increased the most and the numerators of the Laspeyres index will be too large when the prices go down, consumers often shift their preference to those items which have declained the most and, hence by using base period weights in the numerator of the Laspeyers index we shall not be giving sufficient weight to the prices that have gone down the most and the numerator will again be too large. Similarly because people tend to spend less on goods when their prices are rising the use of the Paasche or current weighing produces an index which tends to underestimate the rise in prices i.e., it has a downward bias. But the above arguments do not imply that Laspeyres index must necessarily be larger than Pasche's.

Unless drastic changes have taken place betwen the base year and the given year, the difference between Laspeyres and Paasche's will generally be small and either could serve as a satisfactory measure in practice. However, the base year weighted Laspeyres type index remains the most popular for reasons of its practicability. The Paasche type index can only be constructed when up-to-date data for the weights are available. Furthermore, the price index of a given year can be compared only with the base year. For ex : let $P_{93} = 100$, $P_{94} = 130$ and $P_{95} = 145$. Then P_{94} and P_{95} are using different weights and cannot be compared with each other. If there indices had been obtained by the Laspyres formula they could be compared because in that case the weights are the same base year weights (q_0). For these reasons, in practice the Paasche formula is usually not used and the Laspeyres type index remains the most popular.

(iii) Dorbish and Bowley's Method : Dorbish and Bowley have suggested simple arithmetic mean of the two indices (Laspeyres and Paasche mentioned above so as to take into account the influence of both the periods, i.e., current as well as base periods. The formula for constructing the index is :

$$P_{01} = \frac{L + P}{2}$$

Where L = Laspeyres index, P = Paasche index

or $$P_{01} = \frac{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}{2} \times 100$$

(iv) Fisher's ideal index : Prof. Irving Fisher has given a number of formule for constructing index and of these he calls the one as the ideal index. The Fisher's ideal index is given by the formula

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100 \quad \text{or} \quad P_{01} = \sqrt{L \times P}$$

It shall be clear from the above formula that Fisher's ideal index is the geometric mean of the Laspeyres and Paasche indices. Thus, in the Fisher's method we average geometrical formula that err in opposite directions.

(i) It is based on the geometric mean which is theoritically considered to be the best average for constructing index numbers.

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

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

(iv) It is free from bias. The two formulae Laspeyre and Paascher's that embody the opposing type and weight biases are i.e., by an averaging process that of itself has no bias. The result is the complete cancellation of biases of the kinds revealed by time reversal and factor reversal tests.

It is not however, a practical index to compute because it is excessively laborious. The data, particularly for the Paasche segment of the index, are not readily available. In practice, statisticians will continue to rely upon simple, although perhaps less exact, index number formulae.

(v) Marshal Edgeworth Method : In this method also both the current year as well as base year prices and quantities are considered. The formula for constructing the index is

$$P_{01} = \frac{\Sigma (q_0 + q_1) p_1}{\Sigma (q_0 + q_1) p_0} \times 100$$

On opening the brackets

$$P_{01} = \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$

It is a simple, readily constructed measure, giving a very close approximation to the result obtained by the ideal formula.

(vi) Kelly's Method : Truman L. Kelly has suggested the following formula for constructing index numbers.

$$P_{01} = \frac{\Sigma p_1 q}{\Sigma p_0 q} \times 100$$

Here weights are the quantities which may refer to some period not necessity the base year or current year. Thus, the average quantity of two or more years may be used as weights. In the Kelly's formula the average of the quantities of two years is used as weights, the formula becomes

$$P_{01} = \frac{\Sigma p_1 q}{\Sigma p_0 q} \times 100; \text{ where } q = \frac{q_0 + q_1}{2}$$

Similarly, the average of the quantities of 3 or more years can be used as weights. This method is known as fixed weight aggregate index and is currently in great favour in the construction of index number series. An important advantage of this formula is that like Laspeyres index it does not demand yearly changes in the weights. Moreover, the base period can change with out necessitating corresponding change in the weights. This is very important because the construction of appropriate quantity weights for a general purpose index usually requires a considerable amount of work. Weights can thus be kept constant until new census (or other survey) data become available to revise the index.

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

1. Laspeyres Method
2. Paasche Method
3. Bowley's Method
4. Fisher's ideal Method
5. Marshall-Edgeworth Method

Commodity	1999		2000	
	Price	Quantity	Price	Quantity
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

Solution : Calculation various indices.

Commodity	1990		2000		p_1q_0	p_0q_0	p_1q_1	p_0q_1
	Price	Qty	Price	Qty				
A	2	8	4	6	12	16	24	12
B	5	10	6	5	60	50	30	25
C	4	14	5	10	17	56	50	40
D	2	19	2	13	38	38	26	60
					Σp_1q_0	Σp_0q_0	Σp_1q_1	Σp_0q_1
					= 200	= 160	= 130	= 103

1. Laspeyres Method :

$$p_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 \text{ , where } \Sigma p_1 q_0 = 200, \Sigma p_0 q_0 = 160$$

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

2. Paasche's Method :

$$p_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100, \text{ where as } \Sigma p_1 q_1 = 130 , \quad \Sigma p_1 q_1 = 103$$

$$p_{01} = \frac{130}{103} \times 100 = 126.21$$

3. Bowley's Method :

$$p_{01} = \frac{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100}{2}$$

$$= \frac{\frac{200}{160} + \frac{130}{103} \times 100}{2}$$

$$= \frac{1.25 + 1.26}{2} \times 100 = 125.6$$

$$p_{01} = \frac{L + P}{2} = \frac{125 + 126.2}{2} = 125.6$$

4. Fisher's Ideal Method :

$$p_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

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

$$= \sqrt{1.578} \times 100 = 125.6$$

5. Marshall-Edgeworth Method :

$$P_{01} = \frac{\Sigma(q_0 + q_1)p_1}{\Sigma(q_0 + q_1)p_0} \times 100$$

$$= \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100 = \frac{200 + 130}{160 + 103} \times 100$$

$$= \frac{330}{263} \times 100 = 125.47$$

Weighted Average of Relatives : In the weighted aggregate method discussed above price relatives were not computed. However, like unweighted relative method it is also possible to compute weighted average of relatives. For the purpose of averaging we may use either the arithmetic mean or the geometric mean. The steps in the computation of the weighted arithmetic mean or relatives index number are as follows :

(i) Express each item of the period for which the index number is being calculated as a percentage of the same item in the base period.

(ii) Multiply the percentages as obtained in step (i) for each item by the weight which has been assigned to that item.

(iii) Add the results obtained from the several multiplications carried out in step (ii).

(iv) Divide the sum obtained in step (ii) by the sum of the weights used. The result is the index number. Symbolically.

$$P_{01} = \frac{\Sigma pv}{\Sigma v}$$ where $p =$ price relative

$v =$ value weights i.e., $p_0 q_0$

Instead of using arithmetic mean the geometric mean may be used for assigning relatives. The weighted geometric mean of relatives is computed in the same manner as the unweighted geometric mean of relatives index number except that weights are introduced by applying them to the logarithms of the relatives when this method is used and the formulae for computing the index is

$$P_{01} = \left[\frac{\Sigma V \cdot \log p}{\Sigma V} \right]$$

where $p = \frac{p_1}{p_0} \times 100$

$V =$ value weight, i.e., $p_0 q_0$ for each item.

Steps :

* Obtain percentage relatives for each item.
* Find the logarithm of each percentage relative found in step (i)
* Multiply the logarithms by the weights assigned.
* Add the results obtained in step (iii)
* Divide the total obtained in step (iv) by the sum of the weights.
* Find the antilogarithm of the quotient obtained in step (v)

This is weighted geometric mean of relatives index number.

The following example, shall illustrate the steps.

Illustration : From the following data compute price index by supplying weighted average of price method using

(a) arithmetic mean and

(b) geometric mean

Commodity	p_0 (Rs)	q_0	P_1 (Rs)
Sugar	3.0	20Kg.	4.0
Flour	1.5	40 Kg.	1.6
Milk	1.0	10Lt.	1.5

(a) Index number using weighed A.M. of price relatives.

Commodity	p_0	q_0	p_1	p_0q_0	$\frac{p_1}{p_0}\times 100$	p_v
Sugar	Rs. 3.0	20Kg.	40	60	$\frac{4}{3}\times 100$	8,000
Flour	Rs. 1.5	40Kg.	1.6	60	$\frac{1.6}{1.5}\times 100$	6,400
Milk	Rs. 1.0	10Lt.	1.5	10	$\frac{1.5}{1.0}\times 100$	1,500

$\Sigma p_v = 15,900$

This means that there has been a 22.3 per cent increase in prices over the base level.

(b) Index number using geometric mean of price, relatives.

Commodity	p_0	q_0	p_1	v	p	log p	v log p
Sugar	Rs. 30	20Kg.	Rs.4.0	60	133.3	2.1249	127.494
Flour	Rs. 1.5	40 Kg.	1.6	60	106.7	2.0282	121.692
Milk	Rs. 1.0	10 Lt.	1.5	10	150.0	2.1761	21.761
							$\Sigma v \log p$
							$= 270.947$

$$p_{01} = \text{Anti log}\left[\frac{\Sigma v \log p}{\Sigma v}\right] = \text{Anti log}\left[\frac{270.947}{130}\right]$$

$$= \text{Anti log}(2.084) = 120.9$$

15.5 QUANTITY OR VOLUME INDEX NUMBERS

Price index numbers measure and permit comparison of the price of certain goods; quantity index numbers on the otherhand; measure the physical volume of production, construction or employment. Though price indices are more widely used, production indices are highly significant as indicators of the level of output in the economy or in parts of it.

In constructing quantity index numbers, the problems confronting the statistician are analogous to those involved in price indices. We measure changes in quantities, and when we weigh we use prices or values as weights. Quantity indices can be obtained easily by changing p to q and q to p in the various formulae discussed above.

Thus, when Laspeyres method is used $Q_{01} = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times 100$

When Paasche's formula is used $Q_{01} = \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1} \times 100$

When Fisher formula is used $Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} \times 100$

These formulae represent the quantity index in which the quantities of the different commodities are weighted by their prices. However, any other suitable weights can be used instead.

Illustration : From the following data compute a quantity index :

Commodity	Quantity		Price in
	1997	1998	1997
A	30	25	30
B	20	30	40
C	10	15	20

Solution : Computation of Quantity Index

Commodity	q_0	q_1	p_0	q_1p_0	q_0p_0
A	30	25	30	750	900
B	20	30	40	1200	800
C	10	15	20	300	200

$\Sigma q_1p_0 = 2{,}250 \qquad \Sigma q_0p_0 = 1900$

$$q_{01} = \frac{\Sigma q_1p_0}{\Sigma q_0p_0} \times 100 = \frac{2250}{1900} \times 100 = 118.42$$

Thus compared to 1997 the quantity index has gone up by 18.42 per cent in 1998.

15.6 VALUE INDEX NUMBERS

The value of a single commodity is the product of its price and quantity. Thus, a value index v is the sum of the values of a given year, divided by the sum of the values of the base year. The formula, therefore is

$$v = \frac{\Sigma p_1q_1}{\Sigma p_0q_0} \times 100, \; v = \text{value index}$$

where $\Sigma p_1q_1 =$ Total value of all commodities in the given period, and

$\Sigma p_0q_0 =$ Total value of all commodities in the base period

Since in most cases the value figures are given the formula can be stated more simply.

$$v = \frac{\Sigma v_1}{\Sigma v_0}$$

in which v stands for value.

In this type of index both price and quantity are variable in the numerator. Weights do not have to be applied, since they are inherent in the value figures. A value index, therefore, is an aggregate of values. It measures the change in actual values between the base and the given period.

The value index is not in wide use, although because of the unsatisfactory nature of price and quantity indices, it has been occassionally suggested that they be replaced by the value index. The temptation, however, must be resisted, since the concepts of price level and quantity level answer questions that cannot be answered by the value level. Furthermore, an aggregate of values may be viewed as the product of a price level and a quantity level. The division of an aggregate of value into its price and quantity factors may be arbitrary. But this arbitrariness need not create any confusion as long as our concepts of the two factors are consistent.

The test of consistency is that the product of the price and quantity indices must produce the value index.

15.7 QUESTIONS TO STUDY

1. Explain problems involved in the construction of Index Numbers & its uses
2. Explain the construction of Index Numbers.
3. Write about volume index numbers.

15.8 REFERENCES

1. Fundamentals of Applied Statistics - S.C. Gupta & V.R. Kapoor.

LESSON - 16

TESTS OF ADEQUACY OF INDEX NUMBER FORMULAE

BASE SHIFTING & SPLICING OF INDEX NUMBER

LEARNING OBJECTIVES

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

* Have a clear comprehension of the theory and the practical utility of the concepts of "Tests of Adequacy of Index Number, Base shifting and Splicing of Index Number".

LESSON OUTILINE

16.1 TEST OF ADEQUACY OF INDEX NUMBER FORMULAE

Several formulae have been suggested for constructing index numbers and the problem is that of selecting the most appropriate one in a given situation. The following tests are suggested for choosing an appropriate index :

* Unit test
* Time Reversal Test
* Factor Reversal Test
* Circular Test

Unit Test : The unit test requires that the formulae for constructing an index should be independent of the units in which, or for which, prices and quantities are quoted. Except for the simple (unweighted) aggregate index all other formulae discussed in this chapter satisfy this test.

Time Reversal Test : Prof. Irving Fisher has made a careful study of the various proposals for computing index numbers and has suggested various tests to be applied to any formula to indicate whether or not it is satisfactory. The two most important of these he calls the time reversal test and the factor reversal test.

Time reversal test is a test to determine whether a given method will work both ways in time, forward and backward. In the words of Fisher, "The test is that the formulae for calculating the index number should be such that it will give the same ratio between one point of comparison and the other, no matter which of the two is taken as base". In other words, when the data for any two years are treated by the same method, but with the bases reversed, the two index numbers secured should be reciprocals of each other so that their product is unity. Symbolically, the following relation should be satisfied.

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

where p_{01} is the index for time "1" on time "0" as base and p_{10} is the index for time "0" on time "1" as base. If the product is not unity, there is said to be a time bias in the method. Thus, if from 1998 to 1999 the price of wheat increased from Rs. 480 to Rs. 560 per quintal the price in 1996 should be $133\frac{1}{3}$ per cent of the price in 1998 and the price in 1996 should be 75 per cent of the price in 1999. One figure is the reciprocal of the other; their product (1.33×0.75) is unity. This is obviously true for each individual price relative and, according to the time reversal test, it should be true for the index number.

The test is not satisfied by Laspeyres method and the Paasche method as can be seen below;

When Laspeyres method is used

$$p_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}$$

$p_{10} = \frac{\Sigma p_0 q_1}{\Sigma p_1 q_1}$; $p_{01} \times p_{10} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \neq 1$ and the test is not satisfied.

When Paasche method is used

$$p_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}; \quad p_{10} = \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}$$

$p_{01} \times p_{10} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0} \neq 1$ and the test is not satisfied.

There are five methods which do satisfy the test.

1. The Fisher's ideal formula.
2. Simple geometric mean of price relatives.
3. Aggregates with fixed weights.
4. The weighted geometric mean of price relatives if we use fixed weights.
5. Marshall-Edgeworth method.

Let us now see how Fisher's ideal formula satisfies the test.

Proof : $P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$

Changing time, i.e., 0 to 1 and 1 to 0.

$$P_{10} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}} = \sqrt{1} = 1$$

Since $P_{01} \times P_{10} = 1$, the Fisher's ideal index satisfies the test.

FACTOR REVERSAL TEST

Another test suggested by Fisher is known as factor reversal test. It holds that the product of a price index and the quantity index should be equal to the corresponding value index. In the words of Fisher, "Just as each formula should permit the interchange of the two times without giving inconsistent results, so it ought to permit interchanging the prices and quantites without giving inconsistent result, i.e., the two results multiplied together should give the true value ratio". In otherwords, the test is that the change in price multiplied by the change in quantity should be equal to the total change in value. The total value of a given commodity in a given year and the prodcut of the quantity and the price per unit (total value = $p \times q$). The ratio of the total value of one year to the total value in the preceeding year is $\frac{p_1 q_1}{p_0 q_0}$. If from one year to the next, both price and quantity could double, the price relative would be 200, the quantity relative 200, and the value relative 400. The total value in the second year would be four times the value in the first year. In otherwords, if p_1 and p_0 represent prices and q_1 and q_0 the quantities in the current year and the base year, respectively, and if p_{01} represents the change in price in the current year and Q_{01} the change in quantity in the current year, then

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

If product is not equal to the value ratio, there is, with reference to this test, an error in one or both of the index numbers.

The factor reversal test is satisfied only by the Fisher's ideal index.

Proof : $P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$

changing p to q and q to p.

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

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

$$= \sqrt{\frac{(\Sigma p_1 q_1)^2}{(\Sigma p_0 q_0)^2}} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

Since $P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$, the factor reversal test is satisfied by the Fisher's ideal index. This means, of course, that the formula serves equally well for constructing indices of quantities as for constructing indices of prices, the quantity index being derived by interchanging p and q in the ideal formula. None of the simple or weighted forms of elementary indices - arithmetic mean, harmonic mean, geometric mean-fulfil the requirements of factor reversal test. It is thus obvious that the strong restrictions imposed by the factor reversal test compel its being ignored in the construction of many highly reputable index numbers.

Some authorities on the subject argue that these are no good logical reasons for claiming that an index number ought to meet these tests. For example, Karmel has pointed out that as far as time reversal test is concerned collections of goods included in p_{01} is different from that included in p_{10} (q_0 as against q_1) and, therefore, one could hardly hope for consistent results.

CIRCULAR TEST

Another test of the adequacy of the index number formula is what is known as 'circular test'. If in the use of index numbers interest attaches not merely to a comparison of two years, but to the measurement of price changes over a period of years, it is frequently desirable to shift the base. A formula is said to meet this test if, for example, the 1995 index with 1990 as the base is 200, and the 1995 index with 1985 as the base must be 400. Clearly, the desirability of this property is that it enables us to adjust us the index values from period to period without referring each time to the original base. A test of this shiftability of base is called circular Test.

This test is just an extension of the time reversal test. The test requires that if an index is constructed, for year a on base year b and for the year b on base year c, we ought to get the same result as if we calculated direct an index for a on base year c without going through b as an intermediary.

Symbolically, if there are three indices p_{01}, p_{12} and p_{20} the circular test will be satisfied if :

$$p_{01} \times p_{12} \times p_{20} = 1$$

The Laspeyres index does not satisfy the test as can be seen from the following :

If the three years are 0, 1, 2 the index by Laspeyres method will be

$$\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_2}{\Sigma p_2 q_2}$$

The product of all these is not equal to 1. Hence the test is not satisfied. Similarly, it can be shown that the Paasche's index and Fisher's index do not satisfy the test. However, the simple aggregate method and the fixed weight aggregate method satisfy the test as can be seen from the following :

When the test is applied to the simple aggregate method, we will get

$$\frac{\Sigma p_1}{\Sigma p_0} \times \frac{\Sigma p_2}{\Sigma p_1} \times \frac{\Sigma p_0}{\Sigma p_2} = 1$$

Similarly when applied to fixed weight aggregate method, we will get

$$\frac{\Sigma p_1 q}{\Sigma p_0 q} \times \frac{\Sigma p_2 q}{\Sigma p_1 q} \times \frac{\Sigma p_0 q}{\Sigma p_2 q} = 1$$

An index which satisfies this test has the advantage of reducing the computations every time a change in the base year has to be made. Such index numbers can be adjusted from year to year without referring each time to the original base.

Illustration : Calculate Fisher's ideal index from the following data and prove that it satisfies both the time reversal and factor reversal tests.

Commodity	2000		2001	
	Price	Expenditure	Price	Expenditure
A	8	80	10	120
B	10	120	12	96
C	5	40	5	50
D	4	56	3	60
E	20	100	25	150

Solution : Calculation of Fisher's Ideal Index

Commodity	2000		2001					
	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	8	10	10	12	100	80	120	96
B	10	12	12	8	144	120	96	80
C	5	8	5	10	40	40	50	50

D	4	14	3	20	42	56	60	80
E	20	5	25	6	125	100	150	120
					Σp_1q_0	Σp_0q_0	Σp_1q_1	Σp_0q_1
					$=451$	$=396$	$=476$	$=426$

$$p_{01} = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100 = \sqrt{\frac{451}{396} \times \frac{476}{426}} \times 100$$

$$= \sqrt{1.2726} \times 100 = 1.128 \times 100 = 112.8$$

Time Reversal Test : Time reversal test is satisfied when $p_{01} \times p_{10} = 1$

$$p_{10} = \sqrt{\frac{\Sigma p_0q_1}{\Sigma p_1q_1} \times \frac{\Sigma p_0q_0}{\Sigma p_1q_0}} = \sqrt{\frac{426}{476} \times \frac{396}{451}}$$

$$p_{01} \times p_{10} = \sqrt{\frac{451}{396} \times \frac{476}{426} \times \frac{426}{476} \times \frac{396}{451}} = \sqrt{1} = 1$$

Hence, time reversal test is satisfied.

Factor Reversal Test : Factor reversal test is satisfied when :

$$P_{01} \times Q_{01} = \frac{\Sigma p_1q_1}{\Sigma p_0q_0}$$

$$Q_{01} = \sqrt{\frac{\Sigma q_1p_0}{\Sigma q_0p_0} \times \frac{\Sigma q_1p_1}{\Sigma q_1p_0}}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{451}{396} \times \frac{476}{426} \times \frac{426}{396} \times \frac{476}{451}} = \frac{476}{396}$$

This is also the value of $\frac{\Sigma p_1q_1}{\Sigma p_0q_0}$. Hence the above data also satisfies the Factor Reversal Test.

16.3 THE CHAIN INDEX NUMBERS

In the fixed base method discussed so far, the base remains the same throughout the series of the index. This method, though convenient, has certain limitations. As time elapses conditions which were once important become less significant and it becomes more difficult to compare accurately present conditions with those of a remote period. New items may have to be included and old ones may have to be deleted in order to make the index more representative. In such cases it may be desirable to use the chain base method. When this method is used the comparisons are not made with a fixed base; rather the base changes from year to year. For

example, for 1999, 1998 will be the base : for 1998, 1997 will be the base, and so on. If however, it is desired to associate these relatives to a common base the results may be chained to obtain chain indices. Thus, in its simplest form, the chain index is one in which the figures for each year (or sub-period there of) are first expressed as percentages of the preceeding year. These percentages are then chained together by successive multiplication to form a chain index.

In constructing a chain index following steps are desirable.

(i) Express the figures for each year as percentages of the preceeding year.

(ii) The result so obtained are called link relatives. Chain together these percentages by successive multiplication to form a chain index of previous year divided by 100. In the form of formula chain index for current year =

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

The link relatives obtained in step (i) facilitate comparision from one year to another, i.e. between closely situated periods in which the q' are a process of chaining binary comparisons facilitate long-term comparisons.

Chain relatives differ from fixed-base relatives. Chain relatives are computed from link relatives where as fixed base relatives are computed directly from the original data. The results obtained by the two different methods should be the same, but they may differ from each other slightly due to rounding off of decimal places. Since the process of computing chain relatives is quite complicated and the results are same as the fixed base relatives obtained from the original data, chain relatives should be used when the original data are not available but the link relatives are

Illustration : From the following data of the wholesale prices of wheat for ten years construct index numbers taking (a) 1998 as base, and (b) by chain base method :

Year	Price of wheat (Rs. per 40 Kg)	Year	Price of wheat (Rs. per 40Kg)
1989	50	1994	78
1990	60	1995	82
1991	62	1996	84
1992	65	1997	88
1993	70	1998	90

Solution : (a) Construction of Index numbers taking 1989 as base

Year	Price of Wheat	Index No. (1989 = 100)	Year	Price of Wheat	Index No 1989 = 100
1989	50	100	1994	78	$\frac{78}{50} \times 100 = 156$

1990	60	$\frac{60}{50}\times 100 = 120$	1995	82	$\frac{82}{50}\times 100 = 164$
1991	62	$\frac{62}{50}\times 100 = 124$	1996	84	$\frac{84}{50}\times 100 = 168$
1992	65	$\frac{65}{50}\times 100 = 130$	1997	88	$\frac{88}{50}\times 100 = 176$
1993	70	$\frac{70}{50}\times 100 = 140$	1998	90	$\frac{90}{50}\times 100 = 180$

This means that from 1989 to 1990 there is a 20 per cent increase; from 1990 to 1991 there is a 24 per cent increase, from the 1991 to 1992 there is a 30 per cent increase.

If we are interested in finding out increase from 1989 to 1990, from 1990 to 1991, from 1991 to 1992, we shall have to compute the chain indices.

(b) Construction of chain indices

Year	Price of wheat	Link Relatives	Chain indices (1989 = 100)
1989	50	100.00	
1990	60	$\frac{60}{50}\times 100 = 120.00$	$\frac{120\times 100}{100} = 120$
1991	62	$\frac{62}{60}\times 100 = 103.33$	$\frac{103.33\times 120}{100} = 124$
1992	65	$\frac{65}{62}\times 100 = 104.84$	$\frac{104.84\times 124}{100} = 130$
1993	70	$\frac{70}{65}\times 100 = 107.69$	$\frac{107.69\times 130}{100} = 140$
1994	78	$\frac{78}{70}\times 100 = 111.43$	$\frac{111.43\times 140}{100} = 156$
1995	82	$\frac{82}{78}\times 100 = 105.13$	$\frac{105.13\times 156}{100} = 164$
1996	84	$\frac{84}{82}\times 100 = 102.44$	$\frac{102.44\times 164}{100} = 168$

1997	88	$\frac{88}{84}\times 100 = 104.76$	$\frac{104.76\times 168}{100} = 176$
1998	90	$\frac{90}{88}\times 100 = 102.27$	$\frac{102.27\times 176}{100} = 180$

16.4 CONVERSION OF CHAIN INDEX TO FIXED INDEX

At times it may be desired to convert the chain base index numbers into fixed base index numbes in such a case the following procedure is followed :

1. For the first year the fixed base index will be taken the same as the chain base index. However, if the index numbers are to be constructed by taking first year as the base in that case the index for the first year is taken 100.
2. For calculating the indices for others years the following formula is used.

$$\text{Current year's F.B.I.} = \text{current year's C.B.I.} \times \frac{\text{previous year's F.B.I.}}{100}$$

F.B.I. = Fixed base index number C.B.I. = Chain Base Index Number

Illustration : From the chain base index numbers given below prepare fixed base index numbers

1991	1992	1993	1994	1995
80	110	120	90	140

Solution : Computation of Fixed Base Index Numbers

Year	Chain base Index	Fixed base index No.
1991	80	80
1992	110	$\frac{110\times 80}{100} = 88.00$
1993	120	$\frac{120\times 88}{100} = 105.60$
1994	90	$\frac{90\times 105.6}{100} = 95.04$
1995	140	$\frac{140\times 95.04}{100} = 133.06$

MERITS OF THE CHAIN BASE METHOD :

The merits of this method are enumerated here :

1. The chain base method has a great significance in practice because in economic and business data, we are more often concerned with making comparison with the previous

period and not with any distant period. The link relatives obtained by chain base method serve this purpose.

2. Chain base method permits the introduction of new commodities and the deletion of old ones without necessitating either the recalculation of entire series or other drastic changes. Because of this flexibility, chain index is used in many types of indices such as the consumer price index and the wholesale price index.

3. Weights can be adjusted as frequently as possible. This flexibility is of great significance in many types of index numbers.

4. Index numbers calculated by the chain base method are free to a greater extent from seasonal variations than those obtained by the other methods.

Limitations of the chain Index : The main limitation of the chain index is that as the percentages of previous year figures give accurate comparisons of year-to-year changes. The long-range comparisons of chained percentages are not strictly valid. However, when the index number user wishes to make year-to-year comparisons, as is so often done by the business man, the percentages of the preceeding year provide a flexible and useful tool.

BASE SHIFTING, SPLICING AND DEFLATING THE INDEX NUMBERS

16.5 BASE SHIFTING

For a variety of reasons, it frequently becomes necessary to change the reference base of an index number series from one time period to another without returning to the original raw data and recomputing the entire series. This change of reference base period is usually referred to as "shifting the base". There are two important reasons for shifting the base.

1. The previous base has become too old and is almost useless for purposes of comparison. By shifting the base it is possible to state the series in terms of a more recent time period.

2. It may be desired to compare several index number series which have been computed on different base periods; particularly if the several series are to be shown on the same graph. It may be desirable for them to have the same base period. This may necessitate a shift in the base period.

When base period is to be changed, one possibility is to recompute all index numbers using the new base period. A simpler approximate method is to divide all index numbers for the various years corresponding to the old base period by the index number corresponding to the new base period, expressing the results as percentages. These results represent the new index numbers, the index number for the new base period being 100 per cent.

Mathematically, speaking, this method is strictly applicable only if the index numbers satisfy the circular test. However, for many types of index numbers the method, fortunately, yields results which in practice are close enough to those which would be obtained theoretically.

Illustration : The following are the index numbers of

Year	Index	Year	Index Prices (1990=100)
1990	100	1995	410
1991	110	1996	400
1992	120	1997	380
1993	200	1998	370
1994	400	1999	340

shift the base from 1990 to 1996 and recast the index numbers.

Solution : Index Numbers With 1990 as Base (1990 = 100)

Year	Index No's (1990 = 100)	Index No's (1996 = 100)	Year	Index No's 1990 = 100	Index No 1996 = 100
1990	100	$\frac{100}{400}\times 100 = 25$	1995	410	$\frac{410}{400}\times 100 = 102.5$
1991	110	$\frac{110}{400}\times 100 = 27.5$	1996	400	= 100
1992	120	$\frac{120}{400}\times 100 = 30.0$	1997	380	$\frac{380}{400}\times 100 = 95.0$
1993	200	$\frac{200}{400}\times 100 = 50.0$	1998	370	$\frac{370}{400}\times 100 = 92.5$
1994	400	$\frac{400}{400}\times 100 = 100$	1999	340	$\frac{340}{400}\times 100 = 85.0$

The new series with 1996 as base is obtained very easily by dividing each entry of the first column by 400. i.e., the values of the index for 1996 and multiplying the ratio by 100. Thus

$$\text{Index number for } 1990 = \frac{\text{Index no. for } 1990}{\text{Index no. for } 1996}\times 100$$

$$= \frac{100}{400}\times 100 = 25$$

$$\text{Index number for } 1991 = \frac{\text{Index no. for } 1991}{\text{Index no. for } 1996}\times 100$$

$$= \frac{110}{400}\times 100 = 27.5$$

16.6 SPLICING OF INDEX NUMBER

Sometimes an index number series is available for a period of time, and then undergoes substanial revision including a shift in the reference period.

For example, the weights of an index number many become out of date and we may construct another index with new weight. Thus, two indices would result. At times, it may be necessary to convert these two indices into a continuous series. The procedure employed for this conversion is called splicing.

The process of splicing is very simple and is akin to that used in shifting the base. It is expressed in the form of a formula as follows.

$$\text{Spliced Index No.} = \frac{\text{Index no. of current year} \times \text{Old index of new base year}}{100}$$

The following example illustrates the procedure. Assume that a price index number series was revised by inclusion of certain new products, exclusion of some old products and change in the definition of some other products. In the following table is shown such an old series on a reference base of 1988 and a revised series on a base of 1990.

Year	Old Price Index (1988 = 100)	Revised Price Index (1990 = 100)	Spliced Price Index (1990 = 100)
1987	96		87.27
1988	100		90.91
1989	105		95.40
1990	110	100	100.00
1991		104	104.00
1992		106	106.00
1993		112	112.00

The splicing of the two series to obtain a continuous series stated on the new base of 1990 is accompanied by dividing each figure in the old index figure for 1990 and multiplying the quotient by 100. Thus for 1987 spliced index is obtained as follows.

$$\frac{96}{110} \times 100 = 87.27$$

We can obtain the same result by dividing each figure in the old series by 1.1. Had it been desired to state on the basis of old series and the old reference base of 1988, each multiplying by 1.1.

It may be noted that there must be an overlapping period for the old and revised series to provide for the splicing or linking of the two series. The period of overlap in the example is 1990.

USE OF INDEX NUMBERS IN DEFLATING

By deflating we mean making allowances for the effect of changing price levels. A rise in prices level means a reduction in the purchasing power of money. To take the case of a single commodity. Suppose the price of wheat rises from Rs. 300 per quintal in 1986 to Rs. 600 per quintal in 1996; it means that in 1996 one can buy only half of wheat if he spends the same amount which he was spending on wheat in 1986. Thus, the value (or purchasing power) of a rupee is simply the reciprocal of an appropriate price index written as a proportion. For example, if prices increase by 60 percent the price index is 1.60 and what a rupee will buy is only $\frac{1}{1.60}$ or $\frac{5}{8}$ of what it used to buy. In other words, the purchasing power of rupee is $\frac{5}{8}$ of what it was or approximately 63 paise. Similarly, if prices increase by 25 percent, the price index is 1.25 (125 per cent), and the purchasing power of the rupee is $\frac{1}{1.25} = 0.80$ or 80 paise.

This is the same as saying that the purchasing power of money is the reciprocal of the price index. The general expression may be given thus :

$$\text{Purchasing power of Money} = \frac{1}{\text{Price index}}$$

It shall be clear from above that since the value of money goes down with rising prices the workers or the salaried people are interested not so much in money wages as in real wages, i.e. not how much they earn but how much their income or wage will buy.

For calculating real wages we can multiply money wages by a quantity measuring the purchasing power of rupee, or better we divide the cash wages by an appropriate price index. This process is referred to as deflating. In principle it appears to be very simple but in practice the main difficulty consists in finding appropriate index to deflate a giving set of values or appropriate deflators. The process of deflating can be expressed in the form of a formula as follows.

$$\text{Real wage} = \frac{\text{Money Wage}}{\text{Price Index}} \times 100$$

$$\text{Real wage or income inde x number} = \frac{\text{Index of money wage}}{\text{Consumer price index}}$$

Illustration : The following table gives the annual income of a worker and the general index numbers of price during 1988 - 1996. Prepare index number to show the changes in the real income of the worker and comment on price increase.

Year	1988	1989	1990	1991	1992	1993	1994	1995	1996
Income price	3600	4200	5000	5500	6000	6400	6800	7200	7500
Index No.	100	120	145	160	250	320	450	530	600

Solution : Index Number showing changes in the real income of the worker.

Year	Income	Price Index No.	Real Income	Real Income Index No.
1988	3600	100	$\frac{3600}{100}\times 100 = 3600$	100
1989	4200	120	$\frac{4200}{120}\times 100 = 3500$	97
1990	5000	145	$\frac{5000}{145}\times 100 = 3448.27$	95.78
1991	5500	160	$\frac{5500}{160}\times 100 = 3437.50$	95.48
1992	6000	250	$\frac{6000}{250}\times 100 = 2400$	66.60
1993	6400	320	$\frac{6400}{320}\times 100 = 2000$	55.55
1994	6800	450	$\frac{6800}{450}\times 100 = 1511.11$	41.97
1995	7200	530	$\frac{7200}{530}\times 100 = 1358.49$	37.73
1996	7500	600	$\frac{7500}{600}\times 100 = 1250$	34.72

The method discussed above is frequently used to deflate individual values, value series or value indices. Its special use is in problems dealing with such diversified things as rupee, sales, rupee inventories of manufacturers, whole saler's and retailer's income wages and the like.

16.7 QUESTIONS TO STUDY

1. Explain test of adeqacy of Index numbers (or) ExplainTest of Reversability.

   (or) Show that Fisher's Index number satisfies the time reversal & Factor reveral tests.

2. Explain Chain Index number with an example.
3. Write short notes in Base shifting.
4. Write short notes on Splicing of Index Number

16.8 REFERENCES

1. Fundamentals of Applied Statistics - S.C.Gupta & V.K. Kapoor

LESSON - 17

COST OF LIVING INDEX NUMBER AND WHOLESALE PRICE INDEX NUMBER

OBJECT OF THE LESSON

After studying this lesson the student is expected to have a clear comprehension of the theory and practical utility about the concepts of cost of living and wholesale price index numbers.

LESSON OUTILINE

17.1 CONSUMER PRICE INDEX NUMBERS (OR) COST OF LIVING INDEX NUMBERS

The consumer price index numbes, also known as cost of living index numbers, are generally intended to represent the average change over time in the prices paid by the ultimate consumer. Hence, the consumer price indices are better indicators over the general index numbers, since the general index numbers fail to give an exact idea of the effect of the change in the general price level on the cost of living of different classes of people in different manners.

Different classes of people consume different types of commodities and these commodities are not consumed in the same proportion by different classes of people. For example, the consumption pattern of rich, poor and middle class people varies widely. Not only this, the consumption habits of the people of the same class differ from place to place.

For example, the mode of expenditure of a lower division clerk living in Delhi may differ widely from that of another clerk of the same category living in, say, Mumbai. The consumer price index helps us in determining the effect of rise and fall in prices on different classes of consumers living in different areas. The construction of such an index is of great significance because very often the demand for a higher wage is based on the cost of living index and the wages and salaries in most countries are adjusted in accordance with the consumer price index.

It should be carefully noted that the cost of living index does not measure the actual cost of living nor the fluctuations in the cost of living due to causes other than the change in the price level. Its object is to find out how much the consumers of a particular class have to pay more for a certain basket of goods and services in a given period, compared to the base period. To bring out this fact, clearly, the Sixth International Conference of Labour Statisticians recommended that "the term cost of living index, should be replaced in appropriate circumstances by "the terms or price of living index ' cost of living price index, or consumer price index". At present, the three terms, namely, cost of living index, consumer price index and retail price index are in use in different countries with practically no difference in their connotation.

It should be clearly understood at the very outset that two different indices representing two different geographical areas cannot be used to compare the actual living cost of the two areas. A higher index for one area than for another with the same period is no indication that living costs are higher in one than in the other. All it means is thats compared with the base periods, prices have risen in one area than in another. But actual costs depend not only on the rise in prices as compared with the base period, but also on the actual cost of living for the base period which will vary for different regions and for different classes of population.

17.2 UTILITY OF THE CONSUMER PRICE INDICES

The consumer price indices are of great significance as can be seen from the following :

1. The most common use of these indices is in wage negotiations and wage contracts. Automatic adjustments of wage or dearness allowance component of wages are governed in many countries by such indices.
2. At governmental level, the index numbers are used for wage policy, price policy, rent control, taxation and general economic policies.
3. The index numbers are also used to measure the change in the purchasing power of the currency, real income, etc.,
4. Index numbers are also used for analysing markets for particular kinds of goods and services.

Construction of a consumer price Index : The following are the steps in constructing a consumer price index :

(i) Decision about the class of people for whom the index is meant. It is absolutely essential to decide clearly the class of people for whom the index is meant, i.e., whether it relates to industrial workers, teachers, officers etc., The scope of the index must be clearly defined. For example, when we talk of teachers, whether we are referring to primary teachers, middle level teachers etc., or to all the teachers taken together. Along with the class of people it is also necessary to decide the geographical ares covered by the index. Thus, in the example taken above it is to be decided whether all the teachers living in Delhi are to be included or those living in a particular locality of Delhi, say Chandini Chowk or Karol Bagh, etc.,

(ii) Conducting family budget enquiry - once the scope of the index is clearly defined the next step is to conduct a family budget enquiry covering the population group for whom the index is to be designed. The object of conducting a family budget enquiry is to

determine the amount that an average family of the group included in the index spends on different items of consumption. While conducting such an enquiry, therefore, the quantities of commodities consumed and their prices are taken into account. The consumption pattern can thus be easily ascertained. It is necessary that the family budget enquiry amongst the class of people to whom the index series is applicable should be conducted during the base period. The Sixth International Conference of Labour statisticians held in Geneva, in 1946 suggested that the period of enquiry of the family budgets and the base periods should be identical as far as possible.

The enquiry is conducted on a random basis. By applying lottery method some families are selected from the total number and their family budgets are scrutinized in detail. The items on which the money is spent are classified into certain well-accepted groups, namely :

* Food
* Clothing
* Fuel and lighting
* House Rent
* Miscellaneous

Each of these groups is further divided into sub-groups. For example, the broad group 'food' may be divided into wheat, rice, pulses, sugar etc. The commodities included are those which are generally consumed by people for whom the index is meant. Through family budget enquiry an average budget is prepared which is the standard budget for that class of people. While constructing the index only such commodities should be included as are not subject to wide variations in quality or to wide seasonal alternations in supply and for which regular and comparable quotations of prices can be obtained.

(iii) **Obtaining Price Quotations :** The collection of retail prices is a very important and, at the same time, very tedious and difficult task because such prices may vary from place to place, shop to shop and person to person. Price quotations should be obtained from the localities in which the class of people concerned reside are from where they usually make their purchases. Some of the principles recommended to be observed in the collection of retail price data required for purposes of construction of cost of living indices are described below :

(a) The retail prices should relate to a fixed list of items and for each item, the quality should be fixed by means of suitable specifications.

(b) Retail prices should be those actually charged from consumers.

(c) Discount should be taken into account if it is given to all customers.

(d) In a period of price control or rationing, where illegal prices are charged openly, such prices should be taken into account along with the controlled prices.

The most difficult problem in practice is to follow principle (a) i.e., the problem of keeping the weights assigned and qualities of the basket of goods and services constant with a view to

ensuring that only the effect of price change is measured. To conform to uniform qualities, the accepted method is to draw up detailed descriptions or specifications of the items priced for the use of persons furnishing or collecting the price quotations.

Since the prices form the most important component of cost of living indices, considerable attention has to be paid to the methods of price collection and to the price collection personnel. Prices are collected usually by special agents or through mailed questionnaire or in some cases through published price lists. The greatest reliance can be placed on the price collection through special agents who visit the selected retail outlets and collect the prices. However, these agents should be properly selected and trained and should be given a manual of instructions as well as a manual of specifications of items to be priced. Appropriate methods of price verification should be followed such as 'check pricing' in which the price quotations are verified by means of duplicate prices obtained by different agents or 'purchase checking' where actual purchases of goods are made.

After quotations have been collected from all retail outlets an average price for each of the items included in the index has to be worked out. Such averages are first calculated for the base period of the index and later for every month if the index is maintained on a monthly basis. The method of averaging the quotations should be such as to yield unbiased estimates of average prices as being paid by the group as a whole. This, of course will depend upon the method of selection of retail outlets and also the scope of the index.

In order to convert the prices into index numbers the prices or their relatives must be weighted. The need for weighing arises because the relative importance of various items for different classes of people is not the same. For this reason, the cost of living index is always a weighted index. While conducting the family budget enquiry the amount spent on each commodity by an average family that constitute the weights is decided. Aggregates of expenditure on the different items constitute the individual weights allocated to the corresponding price relative and the percentage expenditure on the five groups constitute 'group weight'.

17.3 METHOD OF CONSTRUCTING COST OF LIVING INDEX NUMBER

After the above mentioned problems are carefully decided the index may be constructed by applying any of the following methods :

1. Aggregate expenditure method or Aggregate method.
2. Family Budget Method or the Method of Weighted Relatives.

1. Aggregate Expenditure Method : When this method is applied the quantities of commodities consumed by the particular group in the base year are estimated which constitue the weights. The prices of commodities for various groups for the current year are multiplied by the quantities consumed in the base year and the aggregate expenditure incurred in buying those commodities is obtained. In a similar manner the prices of the base year are multiplied by the quantities of the base year and aggregate expenditure for the base period is obtained. The aggregate expenditure of the current year is divided by the aggregate expenditure of the base year and the quotient is multiplied by 100. Symbolically,

$$\text{Consumer Price Index} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

This is infact the Laspeyres method discussed earlier. This method is the most popular method for constructing consumer price index.

2. Family Budget Method : When this method is applied the family budgets of a large number of people for whom index is meant are carefully studied and the aggregate expenditure of an average family on various items is estimated. These constitute the weights. The weights are thus the value weights obtained by multiplying the prices by quantities consumed (i.e., $p_0 q_0$). The price relatives for each commodity are obtained and these price relatives are multiplied by the value weights for each item and the product is divided by the sum of the weights. Symbolically,

$$\text{Consumer price Index} = \frac{\Sigma pv}{\Sigma v}$$

where $= \frac{p_1}{p_0} \times 100$ for each item.

v = value weights, i.e. $p_0 q_0$

This method is the same as the weighted average of price relatives method discussed earlier.

It should be noted that the answer obtained by applying aggregate expenditure method and the family budget method shall be the same.

17.4 ILLUSTRATION (PROBLEMS ON C.L.I.)

Construct the consumer price index number for 1996 on the basis of 1995 from the following data using (i) the aggregate expenditure method, and (ii) the family budget method.

Commodity	Quantity Consumed in 1995	Units	Price in 1995		Price in 1996	
			Rs.	Ps.	Rs.	Ps.
A	6 Quintal	Qui	5	75	16	0
B	6 Quintal	Qui	5	0	8	0
C	1 Quintal	Qui	6	0	9	0
D	6 Quintal	Qui	8	0	10	0
E	4 Kg	Kg	2	0	1	50
F	1 Quintal	Qui	20	0	15	0

Solution : Computation of consumer price index number for 1996 (Base 1995 = 100). By the aggregate expenditure method.

Commodity	Quantites Consumed	Unit	Price in 1995	Price in 1996	p_1q_0	p_0q_0
A	6 Qtl.	Qtl.	5.75	6.00	36..00	34.50
B	6 Qtl.	Qtl.	5.00	8.00	48.00	30.00
C	1 Qtl.	Qtl.	6.00	9.00	9.00	6.00
D	6 Qtl.	Qtl.	8.00	10.00	60.00	48.00
E	4 Kg.	Kg.	2.00	1.50	6.00	8.00
F	1 Qtl.	Qtl.	20.00	15.00	15.00	20.00
					Σp_1q_0 = 174	Σp_0q_0 = 146.5

Consumer price index $= \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \frac{174}{146.5} \times 100 = 118.77$

Construction of Consumer Price Index Number

for 1996 (Base 1995 = 100). By the Family Budget Method.

Commodity	Quantities Consumed (q_0)	Unit	Price in 1995 (p_0)	Price in 1996 (p_1)	$\frac{p_1}{p_0} \times 100 (p)$	p_0q_0 (v)	pv
A	6 Qtl.	Qtl.	5.75	6.0	104.34	34.5	3600
B	6Qtl.	Qtl.	5.00	8.0	160.00	30.0	4800
C	1 Qtl.	Qtl.	6.00	9.0	150.00	6.0	900
D	6 Qtl.	Qtl.	8.00	10.0	125.00	48.0	6000
E	Kg	Kg.	2.00	1.5	75.00	8.0	600
F	1 Qtl.	Qtl.	20.00	15.00	75.00	20.0	1500
						ΣV =146.5	ΣPV =17400

Consumer price index $= \frac{\Sigma PV}{\Sigma V} = \frac{17400}{146.5} = 118.77$

Thus, the answer is the same by both the methods. However, the reader should prefer the aggregate expenditure method because it is far more easier to calculate compared to the family budget method.

17.5 INDEX NUMBER OF INDUSTRIAL PRODUCTION

The index number of industrial production is designed to measure increase or decrease in the level of industrial production in a given period compared to some base period. It should be noted that such an index measures changes in the quantum of production and not in values. For constructing such an index it is necessary to obtain the data about the level of industrial output in the base period and the given period. Usually data about production are collected under the following heads :

1. Textile Industries - Cotton, Woollen, Silk etc.,
2. Mining Industries - Iron Ore, Coal, Copper, Petroleum etc.
3. Metallurgical Industries - iron and steel etc.
4. Mechanical Industries - Locomotives, Ships, Aeroplanes etc.,
5. Industries Subject to Excise Duties - Sugar, Tobacco, Matches, etc.,
6. Miscellaneous - Glass, Soap, Chemical, Cement etc.,

The figures of output for the various industries classified above are obtained on a monthly, quarterly or yearly basis. Weights are assigned to various industries on the basis of some criteria such as capital invested turnover, net output, production, etc. usually the weights in the index are based on the values of a net output of different industries. The index of industrial production is obtained by taking the simple arithmetic mean or geometric mean of the relatives. When simple arithmetic mean is used the formula for constructing the index becomes.

$$\text{Index of Industrial production} = \Sigma\left(\frac{q_1}{q_0}\right)\frac{W}{\Sigma W}$$

where q_1 = Quantity produced in the given period.

q_0 = Quantity produced in the base period

W = relative importance of different outputs.

For determining the relative share of an individual output to total output the concept of value added is most commonly used.

17.6 ILLUSTRATION

Construct an index number of business activity in India from the following data

[use (a) Arithmetic mean, and (b) Geometric mean]

Items	Weight	Index
1. Industrial Production	36	250
2. Mineral Production	7	135
3. Internal Trade	24	200

4. Financial activity	20	135
5. Exports and imports	7	325
6. Shipping Activity	6	300

Solution : Construction of Index no. of business activity

S.No.	Items	Weight	Index (T)	1W
1.	Industrial Problem	36	250	9000
2.	Mineral Production	7	135	945
3.	Itneral Production	24	200	4800
4.	Financial Activity	20	135	2700
5.	Exports and Imports	7	325	2275
6.	Shipping Activity	6	300	1800
		$\Sigma W = 100$		$\Sigma 1W = 21520$

Index No. of Business activity = $\frac{\Sigma 1W}{\Sigma W}$

$$= \frac{21520}{100} = 215.2$$

Index of Business Activity using Geometric Mean

Items	Index	$\log \ell$	W	$\log \ell \times W$
Industrial Production	250	2.3979	36	86.3244
Mineral Production	135	2.1303	7	14.9121
Internal trade	200	2.3010	24	55.2240
Financial Activity	135	2.1303	20	42.6060
Exports and Imports	325	2.5119	7	17.5833
Shipping Activity	300	2.4771	100	14.8626

$$\Sigma \log \ell \times W = 231.51$$

Index No. $= A\lg\left(\frac{231.5124}{100}\right) = Al_g\,(2.3151) = 206.5$

The second index is better.

MISCELLANEOUS ILLUSTRATIONS :

1. ILLUSTRATION : Compute the cost of living index number using both the aggregate expenditure method and the family budget method, from the following information.

Commodity	Unit consumption in Base year	Price in base year	Price in current year
Wheat	200	1.00	1.20
Rice	50	3.00	3.50
Pulses	50	4.00	5.00
Ghee	20	20.00	30.00
Sugar	40	2.50	5.00
Oil	50	10.00	15.00
Fuel	60	2.00	2.50
Clothing	40	15.00	18.00

2. Calculate the chain base index No's chained to 1994 from the average price of following three commodities.

Commodity	**1994**	**1995**	**1996**	**1997**	**1998**
Wheat	4	6	8	10	12
Rice	16	20	24	30	36
Sugar	8	10	16	20	24

3. Construct Fisher's ideal index no. for the following data and show how it satisfies the time and Factor Reversal Tests.

	2001		**2002**	
Commodities	**Quantity**	**Price**	**Quantity**	**Price**
M	20	12	30	14
N	13	14	15	20
O	12	10	20	15
P	8	6	10	4
Q	5	8	5	6

4. In 1988 for working class people wheat was selling at an average price of Rs. 120 per 20 Kg. cloth Rs. 20 per metre, house rent Rs. 300 per house and other items Rs. 100 per unit. By 1998 cost of wheat rose by Rs. 180 per 20 Kg, house rent by Rs. 450 and other items doubled in price. The working class cost of living index for the year 1998 with 1988 as base was 160. By how much the cloth rose in price during the period ?

5. Construct the consumer price index no for 1999 on the basis of 1989 from the following data using family budget method.

Items	Price in 1990(Rs.)	Price in 1999(Rs.)	Weights
Food	200	280	30
Rent	100	200	20
Clothing	150	120	20
Fuel and Lighting	50	100	10
Miscellaneous	100	200	20

6. A textile worker in the city of Mumbai earns Rs. 3500 per month. The cost of living index for a particular month is given as 136. Using the following information, find out the amount of money he spent on house rent and clothing.

Group	Expenditure	Group Index
Food	1400	180
Clothing	?	150
House Rent	?	100
Fuel and Lighting	560	110
Misc.	630	80

17.7 LIMITATIONS OF INDEX NUMBERS

Though the Index Numbers are of great significance, the reader must also be aware of their limitations so that he avoids errors of interpretation. The chief limitations of index numbers are

* Since Index numbers are generally based on a sample it is not possible to take into account each and every item in the construction of the index.
* While taking the sample, random sampling is seldom used. This is so because to sample from a population of literally millions of commodities and services, the random procedure could be neither practical nor representative. Typically, indices are constructed from samples deliberately selected. This is likely to introduce errors and every effort must be made to minimise these errors.
* It is often difficult to take into account changes in the quality of products. In a really typical index, qualitites of commodities should remain the same over a period of time because differences in quality would mean differences in prices also. But very often it is not practicable and it makes comparisons over long periods less reliable.
* A large number of methods are designed for constructing index numbers and different methods of computation give different results. Very often the selection of an appropriate formula creates problems and in the interest of comparability. It is necessary to ensure that the same formula is adopted over a period of time for constructing a particular

index. There is no index number method which is most satisfactory from various points of view may be taken logically or practically. Index numbers are averages, and all averages are basically compromised between opposing extremes or forces.

* Just like other statistical tools, index numbers can also be misused in such a manner as to draw the desired conclusions. Choosing a freak year is a favourite trick of those who use statistics to mislead. A dishonest capitalist could choose a record year of profits as base and so 'prove' subsequent profits to be pitifully low. Similarly, in order to prove that the current prices are intolerably high a dishonest trade unionist may choose a year of exceptionally low prices as base.

* Since in the construction of index numbers a large number of factual questions are involved lack of adequate and accurate data often becomes a serious limitation of the index itself. In most cases, one cannot collect the data himself and, therefore, one has to rely on a published source. Ordinarily, we draw upon many sources of data which are geographically dispersed. Problems of comparability and reliability thus multiply and the chances of spurious results are increased. One mistake may "bias" the index such as including the price of one commodity for one time period, or the price of a slightly different commodity for another period, or taking the manufacture's price at one time and whole saler's or retailer's price at another.

17.8 QUESTIONS TO STUDY

1. What is cost of Living Index Number. Explain its construction.
2. What are the limitations of Index Number ?

17.9 REFERENCES

1. Fundamentals of Applied Statistics - S.C. Gupta & V.K. Kapoor

(c) Concurrent List : This list includes such subjects for which ordinances about the collection of data can be formulated by both central & state governments. This list includes labour-welfare, social insurance, trade-union, vital statistics, price control, planning etc.,

Thus, both the central and state governments can collect statistics. The central government provides technical expertise to state governments and establishes coordination between different statistical activities.

Existing Statistical Set-up in India : It has been mentioned earlier that statistical organisation in India is a decentralised system. There are one and more than one statistical units under each ministry for the co-ordination of different statistical units. There is a central statistical organisation under the department of statistics of the central cabinet secretariat.

The statistical organisations working in India (attached to different ministries) can be divided in to the following categories :

18.2 ORGANISATION SPECIALLY SET UP FOR COLLECTION AND COMPILATION OF DATA :

There are certain statistical organisations which have been primarily set-up for the collection and analysis of statistics of different types. Such organisations are most important units of the statistical system of the country. Such organisations are : Directorate General Commercial Intelligence Statistics (D.G.C.I.S.), Labour Bureau, Industrial Statistics Directorate (I.S.D.); National Income Unit (N.I.U.), Registrar General and Census Commissioner (R.G.C.C.). Army Statistical Organisation (A.S.O.), etc., There are also such organisations in the states. Besides them, National Sample Survey Organisation (N.S.S.O.) now a part of C.S.O., and a computer centre (C.C.) have also been set-up for filling the gap for the organisation of the statistical material available in the country.

(ii) Organisations for Processing of data available as by products of Administration:

There are many organisational units at the governmental level which collect data (which is not their primary function) while discharging their administrative functions. The statistics collected by them in the process of administration is of great value in formulating future policies. Boards of Direct and Indirect Taxes : Central Board of Revenue; post and Telegraph Department; Police, Railways and Roadways etc., are such organisations.

(iii) Organisations for control of Production and Distributions :

The units associated with the control of production and distribution of commodities in short supply, come under this category. The statistical units associated with Textile Commissioner, Iron and Steel Controller, Central Power Commission, Controller of Imports and Exports etc., are few such organisations.

(iv) Research Organisations : Research Departments of Reserve Bank of India; Statistical Division of the Indian Council for Agricultural Research and many other agencies which are involved in research works in different fields, fall in this category. These organisations collect data analyze and draw conclusions from them and advise the government for proper policy formulation.

(v) Non-Governmental Organisations : A number of non-governmental organisations like Indian Statistical Institute, Economics and Statistics Department of Tata Research and Training Department of the Indian Merchant Chamber etc., can be put in this category, which are busy in researches in various fields.

(vi) Statistical Organisations in States : In all the states there are directorates of economics and statistics, which take care of the needs of the Statistical Systems of the state.

Statistical Units under Central Ministries : There are 153 statistical units in different central ministries which employ 15,459 personnel. A brief description of some of the central organisations engaged in statistical work under different central ministries is given below :

I. Ministry of Agriculture and Irrigation : There are 44 statistical units employing 1,745 personnel, in the three departments - Agriculture, Co-operation and community Development of this ministry. It has the following major statistical units :

(A) : Directorate of Economics & Statistics :

This directorate was set-up in 1947-48 with the responsibility for collection, compilation and publication of agricultural statistics - covering the fields of agriculture, live-stocks, fisheries and forestry etc., The data are collected by state governments, partly through administrative channels and partly through specially designed random sample surveys. Besides this, directorate also advises the government on the problems related to agriculture and co-ordinates the various agricultural development programmes. Major publications of this ministry are :

Weekly : 1. Bulletin on Agricultural Prices

Monthly : 1. Agricultural Situation in India.

2. Rubber Statistical News.

Annual : 1. Indian Agricultural Statistics Vols. I & II

2. Abstracts of Agricultural Statistics.

3. Estimates of Area & Production of Principal crops in India, Vols I & II

4. Indian Live stock statistics.

5. Indian Forest statistics

6. Agricultural Prices in India

7. Agricultural wages in India.

8. Bulletin on Food Statistics

9. Indian Rubber Statistics.

10. Tea statistics

11. Coffee Statistics

12. Financial Statement of Irrigation works in India.

Two-yearly : 1. Bulletin of commercial crop statistics.

2. Average yield per acre of principle crops in Inida.

Five Yearly : 1. Live stock census of India.

Other 1. Studies in Agricultural Economics

2. Indian Crop Calendar

3. Indian Agricultural Atlas

4. Agricultural Law in Inida.

(B) The Statistical Wing of Indian Council of Agricultural Research (I.C.A.R.) :

This department is known as Insititute of Agricultural Research Statistics - IARS. The main work of this department is to organise random sample surveys, to train the people about the techniques associated with agriculture statistics and to advise the government on agricultural subjects.

Its main publication is "Statistical News, Letter and Abstract". It is a three - monthly publication.

(C) Administrative Intelligence Section of the Department of Community Development : This department co-ordinates and publishes data related to the progress of community development programmes and activities at the all India level. Its annual (or) yearly publications are :

Annual : 1. Annual Appraisal of C.D. Programmes.

2. High lights of C.D. Programmes

3. Panchayat Raj at a Glance.

4. Pocket book of Information.

(D) Directorate of Marketing and Inspection : This organisation conducts surveys from time to time of the marketing of agricultural products.

(E) Other Units / Departments : Besides above departments, there are many other units / institutes which are engaged in collecting and publishing agricultural statistics for the ministry of agriculture. A few of them worth mentioning are :

1. Development of Cooperation

2. Directorate of Sugar and Vanaspati.

3. National Sugar Institute, Kanpur.

4. Indian Sugar Cane Research Inistitute, Lucknow,

5. Central Rice Research Institute, Cuttack,

6. Central Potato Institue, Shimla

7. Indian Diary Research Institute, Karnal

8. Indian Animal Husbandry Research Institute, Izzat Nagar.

9. Central Marine Fisheries Research Station, Maudapam.

10. Forest Research Institute, Dehradun. Its main publications are :

1. Statistical Methods in Forest Products Research
2. Statistical Quality control methods in wood based industries.

II. Ministry of Finance : There are three statistical units in this ministry which give employment to 1019 personnel.

(A) Department of Research and Statistics of Reserve Bank of India : This is a department of Reserve Bank of India, which performs the collection, analysis and publication of financial statistics of the nation. This department has five sub-departments :

(i) Monetary Research : This department conducts surveys for the determination of policies related to credit and finance.

(ii) International Finance : This department takes care of the activities associated with foreign exchange and trade balances in the country.

(iii) Banking Research : This department conducts researches related to banking.

(iv) Rural Economics : This department (conducts researches) takes care of rural finances.

(v) Statistics : This department presents statistical analysis of the facts available to them.

The main publication of this department are :

Monthly :
1. Reserve Bank of India Bulletin

Half yearly :
1. Banking Statistics in India

Annual :
1. Report on Currency and Finance
2. Report of the Central Board of Directors.
3. Trend and Progress of Banking in India.
4. Statistical Tables relating to Banks in India.
5. Statistical Statement Relating to co-operative movement in India.
6. Some Basic Statistics Relating to Indian Economy.
7. Statement of Affairs of Scheduled Banks.
8. Banking and monetary Statistics of India (Adhoc)

This department collects statistics related to the department.

(C) Statistics and Intelligence Branch of the Central Board of Revenue : It collects and publishes on monthly basis the statistics related to central excise and production in its 'monthly statistical branch of central board of direct taxes : its main publication is annual : "All India Income-Tax Report and Returns"

III. Ministry of Commerce : Commerce ministry has eight statistical units which has 385 employees.

(A) Directorate General of Commercial Intelligence and Statistics : This is the oldest statistical units of Governement of India, which was established to collect and publish the data related to Internal and Foregin Trade and Navigation in India. Its main publications are :

Weekly : 1. Indian Trade Journal.

Monthly : 1. Monthly Statistics of Foreign Trade of India by countries & currency areas (MSFTI), vols I & II.

2. Accounts Relating to Coastal Trade & Navigation of India.

3. Accounts Relating to the Island (Rail and River Borne) Trade of India.

Annual : 1. Annual Statement of the Foreign Seaborne Trade of India.

2. Statistics of Maritime Navigation of India.

3. Indian Customs & Central Excise Traffic Vol. I & II.

(B) Office of the Chief Controller of Imports and Exports : Its main publications are :

Weekly : 1. Weekly Bulletin of Industrial Licences and export Licences.

Annual : 1. Annual Bulletin of Statistics of Exports & Imports

2. Annual Administrative Report.

(C) Directorate of Commercial Publicity : This department was primarily under Foreign Trade Ministry till Feb. 1973 and after this it has been transferred to commerce ministry. The main function of this department is to publicise the foreign trade and publish the activities of the department. Its main publications are :

1. Economic and Commercial News
2. Foreign trade of India.
3. Indian Exports
4. Udyog-Vyapar Patrika.

18.3 OTHER UNITS

There are other units and institutions under this ministry, which are engaged on collecting statistics and publishing the related data their departments of some of the few main units are :

1. Textile commissioner
2. Tea Board
3. Coffee Board
4. Coir Board and
5. Handicraft Board

(IV) Ministry of Labour :

There are 4 statistical units, with 633 persons, under this ministry, few of them are as under :

(A) Labour Bureau : Labour Bureau was established in 1946 in Shimla for collecting statistics related to labour, establishing coordination between them, and for construction of cost of living index numbrs and for their publication. The main publications of this department are :

Monthly 1. Indian Labour Journal

Yearly 1. Indian Labour Year Book

2. Indian Labour Statistics

3. Pocket Book of Labour Statistics

Biannual : 1. Trade Unions in India

2. Employment Review

Besides the above publications, annual reports on the following ordinances are published which contain important statistics.

1. Minimum Wages Act
2. Trade Union Act.
3. Factory Act.
4. Labour Compensation Act.
5. Employee's Insurance Act.

(B) Directorate General of Employment & Training : This department publishes statistics related to employment & training. Its main publication is Annual : 1. Hand Book of Training facilities available in the country.

(C) Directorate General of Mines Safety : This department publishes annually the informations (data) related to mines and mine-workers in India.

(D) Directorate General of Factory Advice, Service and Labour Institute : This unit conducts researches about the working conditions and safety, industrial psychology and health conditions etc., It has certain publications also :

(V) Ministry of Industrial Development :

There are seven statistical units employing 182 persons, under this ministry. Its main units are :

(A) Office of Economic Advisor : This unit was established in 1938. Its main function is to prepare weekly wholesale price Index numbers. It publishes a journal "Index numbers of wholesale prices in India - Revised Series" - Today, whole-sale price index numbers based on 1981-80 as the base year 360 commodities are published.

(B) Development Commissioner : Small Scale Industries : This unit publishes reports on monthly, half-yearly and yearly basis related to development of small scale industries.

(C) Others : Besides this, information related to Directorate General of Technical Development, Textile commissioner, Handloom Board, Coir Board etc., are published annually. Some of its main publications are :

Annual : 1. Annual Survey of Industries - Census & Sample Sector

2. Statistics Related to DGTD units.

3. Statistics for Iron & Steel Industries

4. Indian Petrolium and Petro-Chemical Statistics.

Monthly : 1. Monthly Statistics of Production of Selected Industries.

VI. Ministry of Home - Affairs :

There are 30 statistical units under this ministry which employ 2808 personnel. The main unit of this ministry is :

(A) Office of the Registrar General and Census Commissoner of India : In 1948, this unit was established under this ministry on a permanent basis to look after the population census work on a continuing basis. It conducts population census and publishes census reports.

Further registration of vital-statistics, conducting Adhoc Demographic surveys are also performed by this office there are Directorates for Census operations for different states work under this ministry. Its main publications with detailed studies are :

Annual : 1. Vital Statistics of India.

Biannual : 2. Indian Population Bulletin.

VII. Ministry of Planning :

The statistical department of cabinet secretariat was transferred to ministry of planning in 1963. This has 4 statistical units employing 4545 personnel. The statistical units under this ministry are :

(A) Central Statistical Organisation (C.S.O.)

(B) National Sample Surveys Organisation (NSSO)

(C) Indian Statistical Institute (ISI)

(D) Computer Centre

(E) Programme Evaluation Organisation (PEO)

(VIII) Other Ministries :

Besides the above ministries, Ministry of Railways, Education and Social welfare, health and family welfare, Heavy Industries, Defence, Petroleum and Chemical, Irrigation and Energy, Navigation and Transportation, Information and Broad Casting, Tourism and Civil Aviation, Steel and Mines, Construction and Housing etc., do collect statistics related to their ministries and publsh them regularly. Al lthese ministries have their own statistical units.

(IX) Non-Governmental Statistical Organisations :

There are several non-governmental and semi-governmental organisations in the country which are engaged in collection, compilation, research and training related to enriching the statistics in the country, few of them are mentioned here :

1. Indian Statistical Institute, Calcutta.
2. National Council of Applied Economic Research, NewDelhi.
3. Institute of Economic Growth, Delhi.
4. Gokhale Institute of Economics and Politics, Pune.
5. Tata Institute of Social Sciences, Mumbai.
6. Institute of Agricultural Research Statistics.
7. Indian Institute of Foreign Trade, New Delhi.
8. Institute of Labour Research, Mumbai.
9. Institute of Applied Man Power Research, Delhi.
10. Universities of the Country. etc.,,

18.4 AGRICULTURAL STATISTICS

About 72 percent of the people in India are dependent on land for their living. Agriculture and allied activities account for nearly a half of the country's national income. Among the main objects of the Five Year Plans for the economic development of India, production of adequate quantities of food to meet the requirements of the rapidly growing population of India occupies an important place. The increased agricultural production is conditional upon the fuller and more intensive utilisation of existing resources. In this respect the role of comprehensive and reliable statistics for realistic and detailed planning of agricultural development and implementation of plans cannot be minimised. The progress of various plans for agricultural development need be periodically assessed to find out the rate of progress and the weakness with a view to discover suitable correctives. The existing statistical data on land utilisation, yield, live-stock and rural population are inadequate for our needs.

Broadly speaking, all statistics having a bearing on agricultural economy may be termed as agricultural statistics. These may include, among others, statistics of land utilisation, production of crops, live-stock and animal husbandry, hides and skins, forestry, fisheries, mines and minerals, agricultural prices, wages, trade, imports and exports, land revenue, agricultural implements, poultry farming and dairy etc. In the present chapter the term agricultural statistics is used in its narrower sense so as to include statistics of area and yield of principal crops only. Other agricultural statistics are dealt with separately. It is, however, not unusual to find statistics pertaining to these large varieties of topics in publications relating to agricultural statistics. For example, 'Agricultural Statistics' a publication by the Department of Agriculture of USA is a very comprehensive document dealing with all possible data relating to agricultural and rural life in the USA. Similarly, Indian Agricultural Statistics, volume I and II deal with a large variety of statistics relating to agriculture and rural life of India.

Collection of agricultural statistics has been in vogue in India from very early times. Unlike other countries of the world, in India, the responsibility of collecting, these statistics has remained with the government. In ancient days rulers collected such statistics and maintained complete records of land acreage and production so as to determine land revenue which was the principal source of income for the state. Koutilya wrote in his Arthshastra that land records were elaborate and complete even in the days of Ramayan. According to him, during the reign of Raja Ram Chandra

land revenue was based on agricultural produce. The state had no other source of income and land revenue was realised as one-fourth of total agricultural produce. Since taxes were few, or rather non-existent, the collection of agricultural statistics of area and yield become compulsory for the state to be able to assess land revenue. Historical documents of Moghul Period, specially those of the reign of Babar and Akbar, indicate the manner in which agricultural statistics were collected. Akbar wrote in his autobiography, Aina-i-Akbari, details of agricultural administration in the country. In his reign there was an elaborate administrative machinery for the collection of land revenue and maintenance of agricultural records . It was during this period that a ministry for agriculture was established under the charge of Raja Todarmal. The ministry maintained detailed records of the classification of land area under cultivation of cereals, orchards, fallows, forests and land under state occupation.

The value of agricultural statistics was recognised by the British when they took the administration of the country in their hands. They realised the importance of such statistics which helped them in the determination and collection of land revenue. These statistics were also helpful to the administration in other ways, specially when in the latter half of the 19th century famines and droughts became a regular feature. The government collected statistics of agricultural production. However, the British Government in India confined its activities to the field of agricultural statistics for the preparation and issue of crop forecasts only.

The Government of India created a new Department of Agriculture at the centre in 1871 on the recommendations of Lord Mayo who believed that for generations to come, the progress of Indian economy would be dependent on agriculture. As already stated in an earlier chapter, this department was later closed down as a measure of emergency created by the Afghan war. The Royal Commission on Famine recommended to the Government of India for the revival of this Department. As a result of their recommendations, the Central Department of Agriculture was revived in 1879 and agricultural departments were created in other provinces also in which Uttar Pradesh (then North Western Province) was the first to establish such a department "These departments were created to ascertain more systematically and completely, and to render more generally, available statistics of important agricultural and economic facts in order that the government and its officers may always be in possession of an adequate knowledge of the actual condition of the country, its population and its resources".

The system of crop forecasts to which our agricultural statistics were confined till recently dates back to the year 1861 but such forecasts, as would be seen later, were not very useful and effective in the beginning. The Secretary of State for India in England wanted certain improvements in the methods of collecting agricultural statistics. In 1882 he suggested improvements in the forms for the collection of agricultural statistics and wanted the preparation and publication of crop forecasts in India and England before the crops were actually harvested in India. The question of bringing uniformity in collection and publication of agricultural statistics was discussed thoroughly at statistical conference of provincial directors of agriculture at Calcutta in December 1883. The conference agreed for early crop forecasts in regard to crops of considerable commercial importance and favoured, discontinuance of after-harvest reports previously published. The Government of India accepted the recommendations of the statistical conference and issued instructions to provincial governments to the same effect in 1884. It was also decided that crop Forecast Reports from provinces should be compiled and published by the Revenue and Agriculture Department of the Government of India. In the first year the Provincial Governments started with the crop forecasts of wheat alone. Later other commodities were also included in the list. By 1896, about the half a

dozen crops comprised the list including Sugar-cane and ground-nuts.

SCOPE OF AGRICULTURAL STATISTICS :

We have already seen that agricultural statistics include a wide variety of data pertaining to rural and agricultural life of the people. Following the classification given by the Food and Agricultural Organisation (FAO) of the United Nations, Agricultural Statistics can be broadly classified as follows:

1. Basic agricultural statistics pertain to statistics of land holdings and their characteristics. The latter include the size of holdings, form of tenure fragmentation, land utilisation, employment, mechanisation etc. The data under this category throw light on the resources and structure of agriculture in India.
2. Agricultural statistics proper comprising statistics of area and yield, and of livestock and their products.
3. Agricultural statistics in the wider sense refer to statistics of agricultural stocks, trade prices and consumption including that of livestock. The coverage of statistics belonging to this category extends to cost of living of farmers, loans to them, taxes and other levies on farms, labour employed and data pertaining to forests and fishing.

Data compiled under the second and the third categories throw light on the conditions of agriculture in India and help in the laying down of agricultural policies.

The FAO has laid down the following requirements which they consider as basic and essential for all agricultural statistics to satisfy -

(i) Utility (ii) Significance (iii) Reliability (iv) Adequate Coverage

(v) Timeliness (vi) International Comparability

As a matter of fact, these are the requirements which all statistical data must fulfil in order that they may be regarded as satisfactory. Agricultural statistics must be generally useful to the trade, commerce and industry of the country, and to the economic planners and to all others who have occasion to use them. From this point of view it is desirable for them to be comprehensive and detailed. They should be significant to the national economy and must be capable of lending meaning in the total context of an economy. Reliability is an attribute which is self-explanatory. The very purpose of statistics is defeated if they cannot be relied upon. From a geographical point of view they should cover as much of the territory of the country as possible. If the statistics are not available in time much of their utility is impaired. To enable comparisons being made, agricultural statistics must conform to standard concepts and be compiled on internationally accepted proforma for national and international comparability.

In the following pages we shall study some of the major agricultural statistics.

18.5 AREA STATISTICS

Two series of acreage statistics, namely,

(i) The official series and

(ii) The NSSO series are at present available in India.

1. Official Series : The official series relate to statistics of land utilisation giving the area of land put to different uses and the area under different crops. These statistics are available since 1884. The problems of area statistics differ according to the nature of land revenue settlement. From this point of view the entire country has been divided into two broad categories, viz., (i) Temporarily settled areas of UP, Punjab and Madras where land revenue has been temporarily settled and is subject to revision at the time of next settlement; and (ii) Permanently settled areas of Bihar, Bengal, Orissa and eastern part of UP where land revenue has been permanently settled and is fixed unless changed by legislation.

(i) Temporarily Settled Areas : The system of Temporary Settlement or Ryotwari system was introduced in India in the year 1892. The purpose of this type of settlement was to determine land revenues for a fixed period which was subject to change at the time of next settlement. The interval between two settlements, till the abolition of Zamindari, was about thirty years. Under this arrrangement for the collection of land revenue, statistics of land value, cost of cultivation, prices of produce and crop yield were compiled for the determination of land revenue and for making estimates of crop forecasts. It is after the abolition of Zamindari that more exhaustive statistics are being collected under the present arrangement in temporarily settled areas. But even for the period before the abolition of Zamindari, the area statistics in temporarily settled areas were, by and large, correct. "It is generally agreed that the anual figures of areas shown with various crops are on the whole accurate and they compare in this respect very favourably with those published for any other country in the world.

Primary reporting agency : There is an elaborate machinery for the collection of area statistics. It is known as the primary reporting agency whose main function is to collect and maintain agricultural statistics as a part of land records. This agency, which is maintained by the Revenue Departments of State Governments, consists of the village accountant in the last rung of the ladder. These are known by different names in different parts of the country like Lekhpals (formerly patwari) in UP, Karnams in the sourth, Telhati or Talati or Kulkarni in Bombay, Karmchari in Bihar and Patwaris in Punjab.

One Lekhpal is incharge of one village or a group of villages, depending upon the size of the population and land area. He keeps records of individual fields, crops sown and ownership of fields in a special register known as Khasra. He is expected to make field-to-field inspection called partal and enter in his register the name and area of crops. The minimum number of harvest inspections as fixed by the state government is one in MP, Two in Punjab and Bombay and three in UP. But the minimum inspection fixed by the Government is treated as maximum for all practical purposes.

Then, there are revenue inspectors commonly known as Kanoongos who control and supervise the work of Lekhpals. Such inspectors are expected to check 7% entries in the Khasra and make 1% personal inspection of fields to verify the accuracy of entries in the books of Lekhpals Superior to the revenue inspectors are taluka officers, Naib Tehsildars, Tehsildars and Deputy Collectors or Sub-divisional Officers incharge of their respective areas. Such officers supervise and verify the work of their junior officers. Finally there are District Officers, called Collectors, who control the work of revenue collection in the entire district under their charge. The Collector is the supreme officer of the Revenue Department in his district who is sub-ordinate to the Regional Revenue Commissioner and controls and supervises the work of land records and realisation of land revenue.

Sowing of mixed crops :: In temporarily settled areas all fields have been codastrally (i.e., showing the extent, value and owner ship of land for taxation) surveyed, mapped and alloted survey

numbers. Total acerage under each crop is accurately known in these areas. It appears that the method is perfect and satisfactory. In actual practice, however, grave errors crop in and the unreliable nature of these figures is a by-word in the study of agricultural statistics in India. There is no difficulty in ascertaining the accurate area under each crop when the field contains one such crop only, but great difficulty arises in case of mixed farming wherein two or even more than two crops mixed together are sown in the same field.

Irrespective of the reasons that motivate mixed cropping, the different systems as practised in the different parts of the country present considerable difficulties in the accounting of area under each specified crop in terms of a certain net figure. The gross areas of some major crop mixtures which are popularly practised are published by some states in their annual tables of Agricultural statistics and season and crop reports. The allocation of grass area of a crop mixture to its different component crops is done either at source, i.e., at the field level, by the village accountant during the course of his regular crop inspection and the net area of each crop is recorded accordingly in the basic village from Khasra. The village accountant is also allowed to record the whole area of crop mixture treating it as a single crop and in this case total area of component crop is separated at the district level. The assignment of net area to different component crops at the field level is made in proportion to the number of their rows, if sown in separate lines. In case the mixed crop is sown after thoroughly mixing, the net area of component crop is assigned in proportion of seed weights sown for each crop. The component crops occupying a negligible area or area below a certain minimum is ignored and is assigned to major crops. At the district level the apportionment of net areas of component crops of a mixture is done on the basis of the ratio representative of average condition with regard to different crops.

The different states of India can be grouped into three categories with regard to the system followed for the allocation of net areas of component crops of a mixture. These categories are :

(i) States in which allocation is done entirely at the field level, e.g., Assam, Bengal, Andhra Pradesh, Tamil Nadu, Mysore, Orissa, Maharashtra, Gujarat, and Kerala.

(ii) States in which popularly used mixtures are recorded as a single crop by the village accountant and the components are allocated net areas at the district level. Less popular mixtures are apportioned according to net areas at field level, e.g. in Bihar, Punjab, Rajasthan and Jammu and Kashmir.

(iii) States in which the allocation is done entirely at the district level on the basis of the determined fixed ratios representative of average crop conditions, e.g. in UP and MP. The following fixed ratio is popularly used for allocation of components in a mixed crop in some states of India.

Ratio fixed for apportionment of components of the mixed crops

State	Wheat-gram	Wheat-linseed	Gram-linseed	Wheat barley	Gram-barley
U.P...	50 : 50	. .	. .	50 : 50	75 : 25 to 50 : 50
M.P....	93 : 10 to 50 : 50	95 : 5 to 50 : 50	95 : 5 to 20 : 80	. .	. .

Punjab..	50 : 50	. .	. .	50 : 50	50 : 50
Bihar	50 : 50	. .	. .	50 : 50	50 : 50
Rajasthan	70 : 30	. .	. .	66 : 34	50 : 50
	to 39 : 61			to 34 : 66	to 34 : 66

The old system of apportionment of mixed crops should be changed as the technique of mixed cultivation has changed and along with it the ratio of different crops has also changed. The ratio of mixed crops should be tested before state governments decide to fix such ratios for ascertaining land areas under mixed crops. Crop - cutting experiments may be of great help in this direction. In practice, even the fixed ratio of mixed farming is not properly followed by the Lekhpal. He makes some adjustments in important mixed crops like wheat and barley, wheat and gram, gram and barley etc., but in case of minor mixing he generally allocates the entire area to the major crop.

Accuracy of land acreage is doubtful regarding land area sown and land area harvested. The Lekhpal generally records the former and ignores the latter when that alone need be shown in records. Further, such areas where no germination takes place should be excluded from account but that is seldom done. Then, at least, some land area of cultivable fields is occupied by ridges which demacrate the boundaries of two adjoining fields. These ridges are never taken into account and are also included in areas sown which makes the entire land record wrong. These ridges are very common in hilly regions and form not less than 2% of total land acerage on the whole. In addition, land record statistics from many villages are received much later than the scheduled time and are thus excluded or omitted from annual returns. Such omission makes the acerage statistics inaccurate. At times, to make good the deficiency, the Revenue Department makes a broad guess on the basis of old records. If Lekhpals are more alert, this deficiency can be removed.

Although statistics of land acreage are comparatively more accurate in temporarily settled areas, they cannot claim perfect accuracy due to certain defects in the primary reporting agency of the agriculture and revenue department which have been discussed in detail later in this chapter.

(ii) Permanently Settled areas :

The primary reporting agency is conspicuous by its absence in these areas. Since the land revenue is fixed and not subject to change, the state is not very much interested in collecting the information concerning land area and production of crops. Efficient system for the measurement of land area is lacking at most of the places in these areas and different practices are followed in different regions in Bihar, West Bengal and Orissa. Statistics usually available from such areas are regarded as almost worthless'. They are 'mere guesses and are, not infrequently, manifestly absurd guesses'.

In permanently settled areas the police chowkidar or the village mukhia or Headman is entrusted with the task of collecting statistics of land revenue. He also maintains certain statistical records but he is not trained in this respect. Most of the statistical information of these areas is a mere guess work because there is no supervisory staff to check and verify the entries in the registers of the village headman. Statistical records of land acreage and yield of crops maintained

by the village Headman are transferred to the Sub-Divisional officer who modifies the figures by his own experience and forwards them to the District Officer concerned. The District Officer after having modified the figures for his entire district forwards them to the Director General of the state. The Director General modifies statistical figures on the basis of figures submitted to him by various district officers and thus makes available a consolidated figure of land acreage for the whole state.

It is futile to expect agricultural statistics, whether they be of area or of field, to be perfectly accurate. In the very nature of their source, some inaccuracies are bound to creep in and this state of affairs hass come across, in a more or less degree, in all countries of the world. It is possible to verify the accuracy of area statistics by comparing them, (i.e., the official estimates) with the settlement records. In permanently settled areas settlement records are compiled decennially. Often these records are not available for ready reference. But a comparison of the figures available from the two sources shows that the official data are generally under-estimates. This is also the view of the Royal Commission on Agriculture in India. On a rough estimate, it is said that the settlement records are over-estimates by about 15-20 percent. However, this cannot be regarded as conclusive because random sample surveys carried out by the government to check the records of the Patwaris relate to a point of time and the error is not likely to be constant.

Recently there has been a slight change in the existing practice for the collection of agricultural statistics in permanently settled areas, but detailed survey and mapping of land plots is long overdue atmost places in these areas. In 1944 - 45 the Bihar Government established a machinery of primary reporting agency of Karmcharis who have been recording statistics of land area on the basis of complete field inspection. The Government of Bengal conducted plot to plot survey in 1944 - 45 and collected figures of land acreage on the basis of random sampling. The Government also appointed an extensive staff of Agricultural assistants in the province to collect statstics of land acreage and production of crops on a permanent basis. Ad hoc investigations were also carried out by The Bengal Government for this purpose but the system of sampling did not suit all places. In certain inaccessible areas experiment was conducted by the Indian Council of Agricultural Research by aerial photography in 1947-48 with the help of coloured plates for different crops but the scheme was not quite successful. The Government of Orissa did not make any progress in this respect and all schemes of reforms were given up on the ground of financial stringency. As a matter of fact, an agency parallel to the primary reporting agency in temporarily settled areas is needed in permanently settled areas also to improve the situation in respect of agricultural statistics. It is encouraging to note that suitable reporting agencies have been established at most places in permanently settled areas also after the integration of princely states in the Indian Union in 1948.

In permanently settled areas also there is need for proper survey of all land areas for which maps should be prepared separately for various Tehsils and Talukas. Area statistics need be collected on the basis of complete enumeration and sampling should not be resorted to for all places. Aerial photography, as recommended by the ICAR, may be tried at inaccessible places. Although such experiments were not successful in the beginning, but we should not abandon them. In areas not yet surveyed where complete enumeration may not be possible, services of experienced Lekhpals and other land record officers should be utilised. These officials would be helpful in preparing maps and estimates of the area of each field by actual measurement. At places where such facilities may not be available, other alternatives may be resorted to NSSO series :

The Directorate of NSS are collecting data during the regular rounds of surveys regarding land utilisation showing area under different crops. They employ random sampling method for this purpose. The estimates of area under different crops are given in terms of gross area and 'allocated

area'. They define 'gross area' as the area under the crop grown singly (one crop) together with the area under all mixed crops having that crop as one of the components. The 'allocated area' under a crop has been defined as : area under the crop grown singly plus the apportioned area under the crop from all the mixed crops having that particular crop as one of the components. The apportioned area under the crop is allocated to its different components at the plot level by eye estimation on the basis of relative intensity of plants. The estimates are presented for all - India level and for various population zones.

A comparison between the NSSO estimate of acreage under cereals and the official estimates show close agreement at the all - India level. The two estimates differ widely in case of zone-wise and crop-wise figures. The difference between the estimates of two agencies is accounted for by the following :

(i) differences in coverage of crops and seasons to which the figures relate;

(ii) non-comparability of the experience in the field work between the two agencies;

(iii) different classification of area under grain crop and fodder crop;

(iv) differences in the methods in general and particularly in allocation of area under mixed crop; and

(v) possible sampling errors in NSSO series.

Sources of error in area Statistics : Errors in the statistics of area arise from several sources. They may be enumerated as follows :

1. **The Lekhpal**. The Lekhpal being the primary reporting agency, much depends upon how he takes his task. It is unfortunate that this official who is the pivot around whom the revenue administration revolves is regarded as mischievous and corrupt. He is required to make a field-to-field survey in his area. His work is checked and supervised by other administrative officers of the revenue department. The verification of his entries is done in a routine manner and generally they are taken as correct. Lekhpal is not so inefficient as corrupt. His records are mostly wrong. It is common knowledge that he fills in the Khasra or schedules sitting at his house according to his vague impressions, whims, moods and relations with the party concerned. Under these circumstances the entries are bound to be wrong. He rarely conducts any field-to-field survey and shows a remarkable tendency to avoid a change from past figures. His work is checked by the Kanungo, the Naib Tehsildar and the sub-divisional officer in turn.

   As mentioned earlier, the scrutiny of his entries is a formality which is completed in a routine manner at the headquarters of these respective officials and not on the spot. It is surprising that this defect in the primary reporting agency is an age-old defect, but the Government has not been able to remove it. In UP, the Government decided in 1953 to replace the Patwaris by Lekhpals. The Lekhpals have, however, inherited most of the traditional methods of the Patwaris with the result that there is no improvement in the position with regard to agricultural statistics. It is necessary that the remedy be more fundamental. For example, the Lekhpal should be paid slightly more than what he is getting at present. The recruitment should be made from among the educated classes and they should be imparted adequate training before they are actually required to take up the work. There should be complete enumeration at the stage of primary collection

and checking by higher officials should be carried out in fields selected on a random basis. These officials of the revenue department should not be over-burdened with executive duties. Particularly, the Patwari should be assigned revenue work only and should not be asked to do other stray jobs for the Government like assisting the small savings scheme, preparation of the electoral roll, Five year plan propaganda etc., The superior officials should give more detailed instructions and must take to their task of checking and sueprvision more seriously. These remedies are expected to go a long way in improving the situation. It is a matter of satisfaction that the Government of India are alive to the situation and they have introduced a scheme of Central Supervision and Random checking. A check up of the record of the Lekhpals indicates that it involves an underestimation of as much as 15 percent.

2. **Non-survey of certain areas**. Some areas have not been surveyed, mapped and numbered. In estimates such ares are sometimes left out of account altogether.

3. **Sowing of mixed crops**. Wrong returns of area under different crops arise also on account of the practice of sowing mixed crops and it becomes difficult to estimate the area under several crops. In such cases the State Government lays down the ratio in which different crops are to be apportioned. There is a fixed formula referred to earlier according to which area under mixed crops is divided into different crops. It is obvious that a common formula cannot be regarded as satisfactory for all parts of the State and under all circumstances. The State Government should carry out crop-cutting experiments regularly and on that basis decide the ratio of area under different crops. If it is not possible to arrive at a fair basis of division under different crops, the statistics of area published must contain separate records of area under minor mixed crops. Such area should be distributed among various components according to some reasonable basis.

4. **Lack of uniformity in the practice of considering the area sown or area cropped.** It is simple that the area sown is regarded as the area under cultivation. When a crop fails it should be excluded from the figures of area under cultivation. The present practice is that if the piece of land is utilised for sowing other crops it is included in the figures for area under the new crop. If no new crop is sown then the area is allowed to continue under the old crop and only in a few cases it is excluded from it. When a new crop is sown there is no uniformity in the practice that the area is excluded from the old crop. Sometimes it is allowed to continue under the old crop also, with the result that the same piece of land is shown simultaneously under two or more crops. In such cases the commonsense point of view should prevail. If there is no germination at all, or if the crop fails subsequently, the area should be excluded from the forecast at the first opportunity. Later, if the field is utilised for other sowings, it should, naturally, be included under the other crop sown.

5. **Inclusion of ridges.** Figures of area are considerable over estimates on account of inclusion of ridges and bunds in the area under cultivation. Area falling under ridges assumes a large magnitude in Inida on account of vast fragmentation and subdivision of land. Embankments have to be raised to separate pieces of land from those of others. Thus, a considerable proportion of land goes under these embankments but when estimating the area under cultivation no account is taken of them. It is suggested that while measuring the fields, the ridge area should be excluded, or on a uniform basis 2 percent allowance be made for them.

6. Area statistics are also wrong because it is a common practice in the villages to estimate the area according to the quantity of seed down. The area, therefore, depends upon the manner of sprinkling the seed.

However, in order to enhance the reliability of area statistics, the DES has drawn out a scheme-rationalised supervision of work of area enumeration by primary reporting agencies. Training is also imparted by the DES to land record officers.

18.6 QUESTION TO STUDY

1. Explain Historical Development of Statistics before and After independence.
2. Write short notes on Agricultural Statistics and Area Statistics.

LESSON - 19

CENTRAL STATISTICAL ORGANISATION (C.S.O.) NATIONAL SAMPLE SURVEY ORGANISATION (N.S.S.O.) LIVE STOCK & POULTRY STATISTICS

OBJECTIVE OF THE LESSON

After studying this lesson the student is expected to have a clear comprehension of the functions of C.S.O., N.S.S.O., poultry statistics, and the concepts of Forest statistics, Fisheries statistics, Mines and Mineral statistics.

LESSON OUTILINE

19.1 ESTABLISHMENT OF CENTRAL STATISTICAL ORGANISATION (C.S.O)

The years since independence have witnessed a great spurt in statistical activities in India. A central statistical unit was set up in the cabinet secretariat in 1949, which developed into Central Statistical Organisation (C.S.O.) in 1951, with functions relating to laying downs of standards, co-ordination of statistical activities at the centre, states and union territories, supply of National data

(vii) Co-ordination of Population Statistics : This organisation estimates the annual population on the basis of census and vital-statistics. It also makes adjustments in different demographic statistics.

(viii) Compilation of Industrial Statistics : The C.S.O. was given the work of Industrial Statistics wing in 1957. Hence, its industrial statistical wing performs the work related to collection, processing, analysis and publication of Industrial statistics in India.

(ix) Training : This organisation organises short-period statistical training programmes for statistical officers, students and some times foreigners also. This training is organised in Delhi and Calcutta. Some of the training programmes are :

(i) Evening courses (ii) Short term courses for university students (iii) Training courses for Semi or Statistical Officers (iv) Training for foreigners.

(x) Organising Conferences : This also organises seminars and conferences of central and state statistical oficers, technicians and working groups and sends representatives international conferences.

(xi) Inventional works : This organisation has started surveys regarding capital formation in the country. This has published the 'Estimates of gross capital formation in India from 1948 - 49 to 1960 - 61'. This also has estimated the 'Rural Unemployment in the country' and worked on the distribution of man-power, capital, labour and health services in some industries.

(xii) Special works : C.S.O. performs several ad-hoc works related to statistics on the instruction of state and central government.

(xiii) Display : To present and publish the statistical data/materials by diagram and graphs is a vital work of the C.S.O.

(xiv) Publications : C.S.O. works as a central clearing house of statistical intelligence and publishes them from time to time.

These are Periodically published in their reports, abstracts and others.

19.3 C.S.O.'S PUBLICATIONS

The main publications of C.S.O. are as under : Monthly : (1) Monthly Abstract of Statistics (2) Monthly Statistics of Production of selected Industries of India (3) Monthly Bulletin Showing Production of Selected Industries of India and the Index of Industrial Production. (4) Consumer price Index Numbers for Urban Non-Manual Employees.

Quarterly : (1) Statistical News letter

(2) DOCSTAT - It contains the summaries of researches in the field of Economics and Statistics in the world / India.

2. ORGANISATION AND MANAGEMENT :

N.S.S.O. is managed by independent governing council, which has 15 members. Out of these 15 members, 5 are non-governmental educationists, 5 economists from Governmental departments and 5 are directors of working branches. The Director of this council always belongs to Non-Governmental members and Chief-Executive Officer is its member secretary.

Management of N.S.S.O. (Members)

Non-Government	Government	Organisation Staff
1. Chairman/Director	1. Director C.S.O.	1. Four directors of four departments
2. Two Statisticians from I.S.I. State Stai.	2. Two directors of S.S.S.O.	2. Chief-Executive Officer is member secretary.
3. Two Economists from universties or Research organisations.	3. Two statisticians/ economic advisors from central ministries.	

3. THESE ARE FOUR IMPORTANT DEPARTMENTS OF N.S.S.O. :

1. Survey, Design and Research Division;
2. Field - Operation Division - F.O.D. which is the of Ms.S.
    a) Socio-Economic Statistical Field
    b) Industrial Statistics Field
    c) Agricultural Statistics Field
3. Data Processing Division
4. Economic Analysis Division

To minimise the time-lag between the correction of data, its processing and analysis, the tabulation of important statistical items has been done at different manual tabulation centres established in 1972 - 1973.

19.6 FUNCTIONS OF N.S.S.O. :

National Sample Survey Organisation has four major functions :

(a) Collection of data for Ministries : Collection of data on planning, income and other related fields for ministries and planning commission is the major function of N.S.S.O.

(b) Socio-Economic Surveys : N.S.S.O. conducts socio-economic surveys related to subjects like consumption, price, wages, population etc. on random basis.

(x) "Mysore Demographic Survey" under the auspices of Health Division of U.N.O.

(xi) "Rehabilitation survey in Maharastra and West Bengal's urban areas" for the enquiry committee of Rehabilitation ministry.

(xii) "Primary Research for Land-Census" proposed by planning ministry.

VII (C) INDIAN STATISTICAL INSTITUTE (I.S.I) :

The Indian Statistical Institute was registered on April 28, 1932 as a non-profit distributing society under the Societies Registration Act XXI of 1860. Since then it has been carrying out an integral programme of research, training and practical applications of statistics in projects. The institute was declared to be an institution of national importance and given powers to award degrees by an Act of Parliament passed in 1959. This together with the recognition of statistics as a key technology' in the words of late sir Ronald A. Fisher, has resulted in a new approach to the learning of statistics in the institute, which is also expressed in its research programmes.

The International Statistical Education Centre (ISEC) is operated jointly by the International Statistical Institute and the Indian Statistical Institute, under the auspices of the UNESCO and the government of India. The centre functions under a Board of Directors with Dr. C.R. Rao as its first Chairman. The official organ of ISI is 'Sankhya'.

The statistical Quality Control (SQC) Division, set up by the Government of India in 1952 under the administrative control of the Indian Statistical Institute, continued its activities with the object of promoting the application and development of SQC and operational Research (OR) methods, training of technologists, training of trainers, service to industrial establishments and research in SQC operational research and allied methods.

VII (D) COMPUTER CENTRE - ELECTRONIC DATA PROCESSING :

Today, the analysis and processing of huge amount of data is done with the help of computers. The first digital computer was instituted in 1956 in Indian Statistical Institute, Calcutta. From then, there has been a flood of computers in the country.

The application of computers in administrative work was for the first time introduced in 1966, when the Government of India developed a computer centre under statistics department and established ten money well model. New computers were established in Jan. 1982. This computer centre was developed between Jan. 1982 and 1988.

'Computer centre works like an open shop'. Any body who wants to take advantage of this centre, is welcome. The centre provides technical support.

FUNCTIONS OF COMPUTER CENTRE :

The main functions of computer centre are :

(i) To provide electronic-processing of data and its analysis under the statistics department.

(ii) To help perform statistical analysis to various governmental organisations and public enterprises.

Indian Livestock Census (Quinquennial) : This publication contains information concerning number of cattle and poultry. Collected information in 1961 live stock census has been classified according to : (i) bovines (ii) Other live-stock and (iii) Poultry.

Bovinces are classified into cattle and buffaloes, each into males and females above three years of age and below that. Male bovines above three years of age are sub classified into breeding bulls and buffaloes, and other bovindees. Such bovines which are used for work have been further classified as castrated and uncastrated and also bovines over 3 years of age are classified as those in milk in December 1960 and on April 15, 1961; dry; not calved even once; used for work; and others. Young stock of bovinces is classified into : (i) one to three years, and (ii) below one year. These young stocks are again sub-classified each into male and female cattle and buffaloes.

Other live-stocks are classified into sheep, goats, horses and ponies, donkeys and miscellaneous including mules, camels and pigs. Each category is sub-classified into male and female, again classified as up to one year of age, one to three years and those above three years.

Poultry has been classified into fowls, ducks and others. Fowls are sub-classified into hens, cocks and chickens desi and improved. Ducks have three sub-classes as ducks (females), drakes (males) and ducklings (young stock).

The live-stock censuses are being conducted on further improved lines with adequate supervision and sample testing. Census figures are available for different districts of India separately and also for individual states. While drawing any inference from these figures the following must be borne in mind :

(i) Date regarding successive years are not complete and available figures do not provide any idea regarding the trend in number of live stocks over the preceeding years.

(ii) Coverage of the census is not complete. Number of states participating in cattle census has varied from census to census; for example, Orissa and Manipur states did not participate in the census of 1951. Further, at many places lekhpals and patwaris did not count the cattle but reported the figures from the records of the preceeding census, without any alteration.

(iii) Accuracy of available figures is doubtful because patwaris who collected such figures in temporarily settled areas and local agents doing the same service in permanently settled areas are not given adequate training in the task which is highly technical. In order to improve the accuracy of the livestock statistics, the National Income Committee has suggested a system of annual partial census instead of a quinquennial census, to which in a spam of be carried out in the whole area in five years. This would decrease the burden on the primary revenue agency as this activity is spread over five years. Actually, the census should be carried out in one fifth of those villages in which crop cutting experiments are carried out.

(iv) The date to which census relates is not the same for all the states although May 31 was fixed for the purpse. The census due in 1950 was postponed till 1951.

(v) House to house enumeration is expected in every census but it is rarely done and the patwari fills in the schedule from his memory or guess.

(iii) Utilisation of Live-stock Products : Statistics of utilisation of milk and milk products, wool, hide and skin are given separately, specially in respect of wool used for yarn and blanket. Food values of different food stuffs, production calendar for cow and buffalo milk, cost of producton of milk, ghee and butter etc., are also separately given. Figures are also available in respect of working cattle, cultivated areas, co-operative milk societies, 'gau-shalas', cattle breeding and dairy farms, animal slaughter etc. Information concerning import and excise duty and cess on live stock product and price of such products is also published.

(iv) Foreign trade : Figures of import and export of live stock, live stock products, milk and milk products (like butter, cheese, ghee), hides, wools, bones etc., are published in this volume. Imports are classified by source and exports by destination in respect of important live-stock products.

(v) Live-Stock Statistics of Foreign Countries : The volume publishes figures of live-stock statistics in foreign countries also. Such figures include number of cattle, buffaloes, goats, pigs, poultry, horses, ponies, mules and camels in different countries. Most of the foreign figures are based on the 'Year Book of Food and Agricultural Organisation' of the UNO. Statistics of average annual production of milk, butter, cheese and ghee, eggs meat and wool and also figures of foreign trade in live-stock and live stock products of principal foreign countries are given in this publication. This volume also contains one detailed table which provides figures of per capita consumption of dairy product and meat in foreign countries.

While drawing inferences from publications cited above it is interesting to note what Dr. P.V. Sukhatme (FAO), says about these statistics. He observes : "In India live stock statistics are collected from nearly 90 percent of land. From some states annual figures are also obtained. These statistics cannot be called quite correct because, in addition to their availability at five year internal there is inadequate attention on their enumeration at the primary stage. Live-stock product statistics are more or less completely inadequate in the Indian Statistical Organisation". It should be borne in mind that statistics are not comparable from year to year as information is not available for all the states in respect of each census and coverage also did not remain the same through out. In addition, classification of live-stock and poultry is not uniform all over the country and in certain states there is lack of proper agency for the collection of required statistics. Further, some states take annual census while others confine to quinquennial census alone.

Apart from the two volumes mentioned above viz., the Indian Live-Stock Census (Quinquennial) and the Indian Livestock Statistics (Annual), statistics of live-stock products are also available in : (i) Marketing Reports issued occasionally by the DMI, Government of India, (ii) Agricultural Statistics of India (Vol. I and II) (iii) Abstract of Agricultural Statistics in India, and (iv) Statistical Abstract of India Published by the C.S.O.

Need for Improvement : The scope of information compiled under the live-stock census is limited. It relates only to the total number of live-stock and poultry. The method of conducting the census also needs improvement. It is necessary to compile data pertaining to live-stock products e.g. milk, butter, ghee, eggs, meat etc. However, recently efforts are being made to evolve suitable

(c) Miscellaneous : Detailed statistics concerning forest area are available in respect of forests where their economic exploitation has been made possible and also where it was so done previously but suspended for the time being. Statistics are also available for forest areas yet to be expolited when transport facilities develop and the areas become accessible. This latter category is known as "potentially exploitable forests". Area statistics are also available for forests which are open to grazing throughout the year and those which are opened occassionally. Some information is also given relating to survey work undertaken by the Forest Departments of State Governments during the period.

(b) Volume of Standing timber, fire-wood and their increment in exploitable forests. Indian Forest Statistics provides separate figures for different types of available timber along with gross annual increment, natural loss, annual falling and net increment etc. Such figures are available for some states only; hence all figures India on this aspect are not included in the publication.

(c) Out-turn of timber and other minor produce. The publication provides figures separately for round wood, pulp wood and charcoal wood along with other minor forest produce like bomboos, canes, drugs, spices, fibres, fodder grass, resin, lac, rubber, dyeing and tanning material, animal products, vegetable oils and oil seeds.

(d) Employment. Figures regarding number of persons employed on the first day of each month are given. Detailed figures are available for mainly and partially dependent persons. Seperate figures are also available for quantity and value of forest produce consumed by these units.

(e) Revenue and expenditure statistics of Forest Department and Forestry. These are also published in this volume in a summary form.

(f) Foreign trade. This publication also gives figures of quantity and value of import and export of wood and timber along with minor forest produce separately for each year.

Another important publication on the subject is 'A Review of Forest Administration in India' which is a quinuennial volume on forest statistics in India published by the Government of India. This publication deals mainly with fluctuation in forest area under different ownership, comparative increase or decrease in forest output etc. It also gives a report on the progress of survey work undertaken in different states and gives the analysis of the financial results and makes a comparative study with the previous years. An interesting feature of forest statstics is that the two figures concerning forest area published by the Forest Department and Agriculture Department do not tally and often lead to confusion. The difference in area figures is inevitable because of the diversity in the definition of the term 'forest' adopted by the two departments and the purpose for which such statistics are collected by them. Then, there are differences in coverage and in the period to which these statistics relate to in these two different departments of state Governments.

19.9 FISHERIES STATISTICS

Very little statistics concerning fisheries are collected in India and whatever statistics available relate mostly to marine fish and little information on inland fisheries is available. Although fish stands on par with other food items in coastal areas and India on all three, sides is surrounded by sea not much has been done to collect statistics not even in South India. Infact India lacks organised

Detailed statistics regarding fish are available only in respect of Madras, Malabar Coasts and Kerala, while in other states of India statistics available are only in respect of rail and river borne trade with some figures of export and import where available. In port towns on prices data are also collected.

In Madras Fish statistics are collected through officers incharge of curing yards. In addition, some adhoc surveys are also conducted. In other states also, having fish curing yards, detailed statistics of quantity of fish cured and salt consumed are maintained for departmental recors. In India there are more than one hundred fish curing yards on the coast of South India, each under the charge of one sub-inspector or officer. The officer incharge of fish curing yard collects statistics regarding daily landing of fish, its weight, locality of catch, its distance from shore, nets used and the market price of fish. In Madras and other southern states, once in every five years a census of fishermen, fishing-tackle and craft employed by them is taken.

Regarding other states mention may be made of Bombay where statistics are available of quantities of fish brought to Bombay for curing and also of salt consumed in the process. In Punjab, district staff collects fish statistics from Octori and railway authorities and also from contractors. In Bengal, statistics are available regarding quantity and price of imports of fish into Calcutta and Howrah which are provided by fish dealers under Fish Dealers licencing order. Some statistics of export and import of marine fish are available in Baroda. In Kerala, statistics are available concerning export of fish of different varities and quantity cured along with numbers of licensed implements employed by the fishing industry of the state.

For most of the states detailed statistics regarding in land fisheries are not available due to certain inherent difficulties. Firstly, fishing is a perennial industry in which numerous individuals, spread all over the country are engaged, with the result that complete record of catch cannot be had daily. Secondly, fishing practices vary from place to place and no uniformity can be maintained. Lastly, fishing organisation for processings marketing and distribution is not adequate, which renders it difficult to record the volume of fish entering different stages of processing and marketing statistics of fish consumption can be had from octroi offices regarding quantity arriving in urban areas for which regular registers may be maintained. Price statistics may be collected in a similar fashion from important marketing centres. Curing yards may also supply detailed statistics regarding cured fish.

The State Government may also help in improving the situation by the introduction of a simple schedule for collection of fish statistics. Such schedule need contain only simple, and fewer questions. For detailed statistics state Governments may conduct adhoc survey and census, at fixed intervals. In spite of all improvements it would be difficult to record detailed information of fish consumed by fishermen themselves and fish purchased by businessmen on the high seas. In view of the importance of fisheries statistics in the context of planning it is necessary that data on fisheries resources, fish catches, economic intelligence, socio-economic aspect of fish culture etc.,should be collected on a uniform basis throughout the country.

India has nearly 3000 miles of coast line and several million acres of territorial waters scattered all over. Her fishery resources, therefore, are large. These resources are capable of much greater development both in yield and in the industrial opportunities they offer. The present

The Director General of Geological Survey of India (G.S.I.) publishes statistical information concerning mines and minerals in an annual publication called Records issued by the G.S.I. Figures in this publication are based on returns submitted to the office of the Geological Survey by State Governments except in case of mines under Indian Mines Act. for which statistical information is published by the Chief Inspector of Mines, Dhanabad in his 'Annual Report'. In Records minerals are classified into two : (i) those for which trust worthy annual returns are available, and (ii) those for which regular and reliable figures are lacking. Information published provides all India figures and its comparison with the preceding year. It also contains a table which provides information regarding consumption of various minerals in India along with the details of their production, import and export. Some information on minerals of lesser importance is also contained therein, like building construction materials, sand, lime, stone and road metals etc., Figures for petroleum and building materials are not reported by the Chief Inspector of Mines, Dhanabad.

Among other important publications dealing with mineral statistics of India, mention may be made of the Quinquennial Review of Mineral production in India. This volume provides consolidated all. India figures and comments on general tendencies. The 'Review' also provides comparative figures of production and foreign trade of minerals. Historical account of various minerals in India is also given.

Quarterly statistics of production of coal, gold and petroleum gives statistical information concerning production, consumption, export and import of coal, gold and petroleum in India. Quarterly statistics do not deal with unorgainsed mineral industry, especially building materials and state, annual statements in this publication provide figures regarding value of managnese output in India. Mica figures do not include scrap mica. No information is provided on trade of mica because the mica Act does not cover mica dealers and purchasers.

Monthly survey of Business conditions in India, issued by the Chief Inspector of Mines, Dhanabad, provides latest available monthly figures of production of coal, petroleum and kerosene in India.

Statistical information in the Indian Trade Journal (weekly) is avilable concerning manufacturing and import of salt in India.

Indian Coal Statistics, issued by the Department of Commercial Intelligence and statistics, Calcutta, provides comprehensive statistical information concerning production, export, import, consumption price, freight, capital invested and labour employed in Indian coal industry. Separate figures of production are available for individual fields together with the percentage of total production. This publication also gives figures of average export value, freight rate by land and sea and also figures of consumption of 'coal by major consumers in India like Railways, power generating plant etc..

Statistical abstract of India also provides figures of quantity and value of output of each mineral for the 10 preceeding years. It also contains information regarding number of persons employed daily in mineral production.

19.11 QUESTION TO STUDY

1. Explain C.S.O.
2. Explain N.S.S.O.

Prior to 1919 - 20 some of the states followed financial year ending with 31st March but later on uniformity was brought about except in two states referred to above.

(ii) Area sown under a crop in India means area actually sown whether the crop reaches maturity or not except where land area after the failure of germination is sown with other crops. In the latter case the area is included in the area for the other crop. Area statistics are also wrong due to different methods followed in different states for apportioning mixed crops. In temporarily settled areas the patwari apportions the area under mixed crops from his personal inspection and when the mixture is not important generally the entire area is shown under the major crop. States have no uniform rules for this purposes although efforts are being made to bring uniformity in them. In addition to mixed crops, ridges are also responsible for wrong statistics of area. Land area occupied by ridges which is not less than 2% of the total cultivated area is always included in area sown and no allowance is made for that.

(iii) Statistics of area under crops are fairly reliable for temporarily settled areas but not equally reliable for permanently settled areas where village papers are not maintained. Efforts are being made in different states for the recruitment of primary reporting agency and some of the states have already done that.

(iv) Crop statistics for different years are not comparable due to gradual expansion for reporting areas, specially after 1947.

(v) Statistics collected for different crops do not have uniform coverage in respect of area and crop from year to year.

The FAO World Agricultural Census, 1970. This was the third census. Similar censuses were organised by the FAO in 1950 and 1960 also. Most of the free nations of the world are taking part in these censues.

It is recognised that the main task of an agriculturist is to wage a world wide war against hunges and malnutrition. It is also recognised that the role of an agricultural statistician is all the more important in this context.

A four year programme for the training of high ranking administrators and statisticians has been drawn up by the FAO in collaboration with U.S.A. The programme started with the training of 38 statisticians in 1967 in Washington D.C. The training included theoritical institutions, workshop and demonistration and practical field work.

According to the scheme of the census, all nations will collect data on the following items, process and publish it and send to the FAO for discrimination.

(1) Number of agricultural holdings, their main features - size, tenure system, land utilisation i.e. cultivable under permanent crops, permanent pastures, forests etc., use for domestic consumption or for sale.

(2) Area under cultivation of main crops, number of poultry and live stock.

(3) Employment in agriculture and output of work by families and labourers.

(4) Agricultural population - land owners, co-operative cultivators, land the size of the families.

(5) Number of agricultural implements, availability of transport and use and type of fertilizers.

(6) Production of fish and wood for fuel, logging and paper.

(7) Extent to which agriculture is involved with other industries.