U.P. Rajarshi Tandon Open University, Prayagraj

UGSTAT – 104
Applied Statistics

Block: 1 Index Numbers

Unit – 1 : Index Numbers: General Theory

Unit – 2 : Index Numbers: Important Formulae

Unit – 3 : Consumer Price Index Number

Block: 2 Time Series Analysis

Unit – 4 : Time Series

Unit – 5 : Determination of Trend

Unit – 6 : Determination of Seasonal Indices

Block: 3 Demography

Unit – 7 : Sources of Demographic Data

Unit – 8 : Measures of Mortality

Unit – 9 : Measures of Fertility

Unit – 10 : Life Tables

Unit – 11 : Measures of Reproductivity

Block: 4 Statistical Quality Control

Unit – 12 : Introduction to Statistical Quality Control

Unit – 13 : Control Charts for Variables

Unit – 14 : Control Charts for Attributes

Unit – 15 : Principles of Acceptance Sampling

Course Design Committee

Dr. Ashutosh Gupta **Chairman**
Director, School of Sciences, U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anup Chaturvedi **Member**
Department of Statistics, University of Allahabad, Prayagraj

Prof. S. Lalitha **Member**
Department of Statistics, University of Allahabad, Prayagraj

Prof. Himanshu Pandey **Member**
Department of Statistics, D. D. U. Gorakhpur University, Gorakhpur.

Dr. Shruti **Member-Secretary**
School of Sciences, U.P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Dr. Shubhra Katara **Writer** (Block 1)
Department of Statistics, Bareily College, Bareily
Prof. Ramji Tiwari **Writer** (Block 2)
Department of Statistics, University of Jammu, Jammu
Prof. B. P. Singh **Writer** (Block 2)
Department of Statistics, Banaras Hindu University, Varanasi
Dr. Alok Kumar **Writer** (Block 2 & 3)
Department of Statistics, Banaras Hindu University, Varanasi
Prof. S. K. Pandey **Writer** (Block 4)
Department of Statistics, Lucknow University, Lucknow
Dr. Shruti **Writer** (Block 4)
School of Sciences, U.P. Rajarshi Tandon Open University, Prayagraj
Prof. S. K. Pandey **Reviewer** (Block 1)
Department of Statistics, Lucknow University, Lucknow
Dr. D. C. Singh **Reviewer** (Block 2)
Department of Statistics, Harish Chandra P. G. College, Varanasi
Prof. R. C. Yadava **Reviewer** (Block 3)
Department of Statistics, Banaras Hindu University, Varanasi
Prof. V. P. Ojha **Reviewer** (Block 4)
Department of Statistics and Mathematics, D. D. U.,Gorakhpur University, Gorakhpur
Prof. B. P. Singh **Editor** (Block 1)
, Department of Statistics, Banaras Hindu University, Varanasi
Prof. Umesh Singh **Editor** (Block 2 & 4)
Department of Statistics, Banaras Hindu University, Varanasi
Prof. V. P. Ojha **Editor** (Block 3)
Department of Statistics and Mathematics, D. D. U.Gorakhpur University, Gorakhpur
Dr. Shruti **Course/ SLM Coordinator**
School of Sciences, U. P. Rajarshi Tandon Open University, Prayagraj

UGSTAT – 104 APPLIED STATISTICS

First Edition: *March 2008* (Published with the support of the Distance Education Council, New Delhi)
Second Edition: ***July 2021***
ISBN :

Printed By : K.C. Printing & Allied Works, Panchwati, Mathura - 281003

Blocks & Units Introduction

The present SLM on ***Applied Statistics*** consists of fifteen units with four blocks.

The ***Block - 1 – Index Number,*** is the first block, which is divided into three units, deals with theory of Index Numbers. Index numbers are devices designed to measure the relative change in the level of a phenomenon with respect to time, geographical location or other characteristics such as income, profession, social and state affairs of a country as these represent the economic changes and are extensively used for policy making. So a knowledge of theory of index numbers is very important.

Unit – 1- Index Numbers: General Theory, deals with the introduction to the theory of index numbers and various steps in the construction of index numbers.

Unit – 2 - Index Numbers: Important Formulae, describes important methods and formulae for obtaining index number. This unit also deals with the criterion of a good index number.

Unit – 3 - Non-Linear Programming Problem provides the major applications of index numbers which are consumer price index numbers and index of industrial production.

The ***Block - 2 – Time Series Analysis*** is the second block with three units. the time series is a sequence of data values of some variables corresponding to successive point of time. In this block the analysis of time series data is discussed.

Unit – 4 – Time Series, deals with difference component of time series like trend, seasonal components, cyclical variation and random components.

Unit – 5 – Determination of Trend, deals with the various methods of determination of trend.

Unit – 6 – Determination of Seasonal Indices, dealt with various methods of determination of seasonal component of the time series data.

The ***Block - 3 – Demography,*** has five units. Broadly speaking demography is concerned with quantitative study of human population. It focuses its attention mainly on size of population, composition of population and territorial distribution of the population and changes there in and factors responsible for such changes such as fertility, mortality, migration and special mobility. Demography is essentially a science which heavily utilizes data for its study.

Unit – 7 – Source of Demographic Data, describes various sources of data from which relevant data are available for various demographic studies.

Unit – 8– Measures of Mortality and ***Unit – 9 – Measures of Fertility*** are concerned with describing various measures of mortality and fertility which are mainly useful for comparing levels of fertility and mortality of different populations.

In ***Unit – 10 – Life Tables***, gives a detailed description of life table. Life tables are powerful tools in studying longevity of persons of different ages and are useful in many practical situations.

Unit – 11 – Measures of Reproductivity, describes two important measures of reproduction viz. gross reproduction rate and net reproduction rate which are helpful in knowing the quantitative estimate of reproduction of females in their whole reproductive period.

The ***Block - 4 – Statistical Quality Control*** deals with the theory of statistical quality control, ***has*** four units.

Unit – 12 - Introduction to Statistical Quality Control is introductory and gives the concept of control charts, control limits and sub groupings.

Unit – 13 - Control Charts for Variables describes control charts for variables.

Unit – 14 - Control Charts for Attributes discuss control charts for attributes.

Unit – 15 - Principles of Acceptance Sampling, Presents the principles of acceptance sampling along with single sampling plan, its OC, ASN, AQL, LTPD and AOQL functions.

At the end of every block/unit the summary, self assessment questions and further readings are given.

U.P. Rajarshi Tandon Open
University, Prayagraj

UGSTAT – 104
Applied Statistics

Block: 1   Index Numbers

Course Design Committee

Dr. Ashutosh Gupta **Chairman**
Director, School of Sciences, U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anup Chaturvedi **Member**
Department of Statistics, University of Allahabad, Prayagraj

Prof. S. Lalitha **Member**
Department of Statistics, University of Allahabad, Prayagraj

Prof. Himanshu Pandey **Member**
Department of Statistics, D. D. U. Gorakhpur University, Gorakhpur.

Dr. Shruti **Member-Secretary**
School of Sciences, U.P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Dr. Shubhra Katara **Writer** (Block 1)
Department of Statistics, Bareily College, Bareily

Prof. S. K. Pandey **Reviewer** (Block 1)
Department of Statistics, Lucknow University, Lucknow

Prof. B. P. Singh **Editor** (Block 1)
Department of Statistics, Banaras Hindu University, Varanasi

Dr. Shruti **Course/ SLM Coordinator**
School of Sciences, U. P. Rajarshi Tandon Open University, Prayagraj

UGSTAT – 104 **APPLIED STATISTICS**

©UPRTOU
First Edition: *March 2008* (Published with the support of the Distance Education Council, New Delhi)
Second Edition: ***July 2021***
ISBN :

Printed By : K.C. Printing & Allied Works, Panchwati, Mathura - 281003

Block & Units Introduction

The ***Block - 1 – Index Number,*** is the first block, which is divided into three units, deals with theory of Index Numbers. Index numbers are devices designed to measure the relative change in the level of a phenomenon with respect to time, geographical location or other characteristics such as income, profession, social and state affairs of a country as these represent the economic changes and are extensively used for policy making. So a knowledge of theory of index numbers is very important.

Unit – 1- Index Numbers: General Theory, deals with the introduction to the theory of index numbers and various steps in the construction of index numbers.

Unit – 2 - Index Numbers: Important Formulae, describes important methods and formulae for obtaining index number. This unit also deals with the criterion of a good index number.

Unit – 3 - Non-Linear Programming Problem provides the major applications of index numbers which are consumer price index numbers and index of industrial production.

At the end of every block/unit the summary, self assessment questions and further readings are given.

Unit-1 Index Numbers: General Theory

Structure

1.1 Introduction

Index numbers are devices for measuring changes in the magnitude of a group of related variables, over a period of time. These changes may have to do with the prices of the commodities, the physical quantity, changes of goods produced marketed or consumed, or such concepts as intelligence beauty or efficiency.

The comparisons may be between periods of time between places, between like categories such as persons, schools or objects. Thus we may have index numbers comparing the cost of living at different times or in different countries or localities, the physical volume of production in different factories or the efficiency of different school systems.

An index number is aggregate measure of the relative change in a collection of presumably related items. For example a price index is an attempt to consider a particular market basket of commodities and services, the prices for each item in the market basket, and to put all that information together in such a way that the general change in price for the total (or aggregate) market basket can be represented by one number.

1.2 Objectives

After going through this unit you must be able to understand:

- What is an index number
- Necessity of an index number
- Computation of index number
- Prime Relatives, Quality and Volume Relatives
- Link and Chain Relatives
- Problems Involves in Computation of Index Numbers

1.3 Definition of an Index Number

Index numbers are statistical devices designed to measure the relative change in the level of a phenomenon with respect to time, geographical location or other characteristics, such as incomes level of product etc. such a study is of great importance for understanding economy or for the making of government policies and for fixing the wages etc. As a matter of fact the index number is an economic barometer as it gives a measure of the economic pressure on the consumers, directly or indirectly.

According to **John I. Griffin** "An index number is quantity which by reference to a base period, shows by its variation the changes in the magnitude over a period of time."

Index Numbers are the number which express the value of a variable at any time (current period) as a percentage of the value of that variable at some reference period or base period.

Edgeworth gave the definition of index number as- "Index Number shows by its variation the change in a magnitude which is not susceptible of either accurate measurement in itself or of direct variation in practice".

1.4 Construction of an Index Number

The methods of index number construction deal with the techniques of combining time series that can't be added because they are not in comparable units.

eg. the annual production of wheat can be measured by simply totaling the output of the individual products or the total amount moving through the markets. Even though there may be different grades of the product, as is generally true for commodities, all types of wheat are nearly enough alike to make the total production a significant amount. In general, items of the same kind, when classified according to differences of kind may be added if all measurements are expressed in the same units.

Many situations arise however when it is desired to measure the composite changes in the production of a number of commodities that are not expressed in the same units and that cannot therefore be added in order to secure a total. Changes in the production of manufactured goods cannot be measured by totaling the production of all kinds of goods since there is no common physical unit of measurement that can be used for all. When the production of the different items cannot be added it would also be improper to average the prices of the items. Accordingly, when it is desired to measure the change in the average level of prices, it is necessary to use the methods of index number construction.

The basic device used in all methods of index number construction is to average the relative change in either quantities or prices, since relative are comparable and can be added even though the data from which they are derived cannot themselves be added, Tons of cotton and Tons of wheat cannot properly be added; but if wheat production was 110% of the previous year's production and cotton production was 106% it is valid to average these two percentages and to say that the volume of these commodities produced was 108% of the previous year. This assumes they are of equal importance, since each is given the same weight. But if cotton production is six times as importance as wheat production, the percentages should be weighted 6 and 1. The average relative secured by this process is referred to as an index number.

Another distinguishing characteristic of all index number is that they are expressed relative to some specified place, time or period called the base. When data can be added and the single series reduced to a fixed base series of relatives, the relatives are called index numbers by some statistical, others reserve the term 'index number' exclusively for an average of relatives derived from series that cannot properly be added. Since many relatives based on a single series are widely used as measures of business conditions, it seems simple to use the term index number to describe both simple relatives and average of relatives.

1.5 Price Relatives

One of the one of the simplest examples of an index number is a price relative which is the ratio of the price of a single commodity in a given period to its price in another period called the base period of reference period. Here we assume prices to be constant for any one period. If they are not an appropriate average for the same can be taken to make this assumption valid.

If p_0 and p_i denote the commodity prices during the base period and given period respectively, then by definition

$$P_{oi} = \text{Price relative} = p_i/p_0$$

and is generally expressed as percentage by multiplying by 100. i.e. price relative = = p_i/p_0 x100 More generally if p_i and p_j are prices of a commodity during periods i and j respectively, the price relative in period j with respect to period I is defined as p_i/p_j and is denoted by p_{ij}.

Properties of Price Relatives

If p_i, p_j , p_k denotes prices in periods i, j, k respectively, the following properties exist for the associated price relatives.

1) Identity Property $P_{ij} = 1$

   It is also evident as $P_{ij} = p_i/p_0=1$ or 100%

   This states that the price related for a given period with respect to the same period is 1 or 100%.

2) Time Reversal Property $P_{ij} . P_{ij} = 1$ or $P_{ij} = 1/P_{ji}$

   As $P_{ij} . P_{ij} =(p_i /p_j).(p_j /p_i) =1$

   This states that if two periods are interchanged the corresponding price relatives are reciprocals of each other. So price relatives follow time reversal property.

3) Cyclical or Circular Property According to this $P_{ij}.P_{jk}.P_{ki} = 1$

   As $P_{ij}.P_{jk}.P_{ki} =(p_i /p_j).(p_j /p_k).(p_k /p_i) =1$

   Modified Cyclical or Circular Property $P_{ij} . P_{jk}=P_{ik}$

1.6 Quantity or Volume Relatives

Instead of comparing prices of a commodity we may be interested in comparing quantities or volumes of the commodity, such as quantity or volume of production consumption, exports, etc. In such cases we speak of quantity relatives or volume relatives. For simplicity as in the case of prices we assume that quantities are constant for any period. If they are not an appropriate average for the period can be taken to make this assumption.

If q_0 denotes the quantity or volume of a commodity produced, consumed exported etc. during a base period while q_i denotes the corresponding quantity produced consumed etc. during a given period we define.

Quantity or volume relative = q_i/q_0

Which is generally expressed as a percentage.

As in the case of price relatives we use the notation q_{ij} to denote the quantity relative in period j with respect i. The same remarks and properties to price relatives are applicable to quantity relatives.

1.7 Value Relatives

If p is the price of a commodity during a period and q is the quantity or volume produced, sold, etc. during the period then pq is called the total value. Thus if 1000 items are sold at 30 Rs. Each the total value is Rs 30 x 1000= Rs. 30,000.

If p_o and q_o denote the price and quantity of a commodity during a base period while p_i and q_i denote the corresponding price and quantity during a given period, the total values during these period are given by v_o and v_i respectively and we define

Value relative $= v_i/v_o = p_i q_i/p_o q_o = (p_i/p_o)(q_i/q_o)$

= price relative x quantity relative

The same remarks notation and properties pertaining to price and quantity relatives can be applied to value relatives.

In particular if p_{ij} q_{ij} and v_{ij} denotes the price quantity and value relatives of period j with respect to period i then we have

$$V_{ij} = p_{ij} - q_{ij}$$

1.8 Link and Chain Relatives

It may often be desirable to express an index number not as a percentage of the original base (fixe or reference) but as a percentage of the proceeding period.

Such an index might employ any of the formulae utilizing weights pertaining to either both of the years or months being compared. Frequently these separate percentages are chained back to the original base by a process of successive multiplication. Such an index is known as a chain Index.

Chain base method or chain relative method consist in calculating a series of index number for each year with the preceding year as base, viz, P_{01}, P_{12}, P_{23},….., where P_{ij} represents the price index with i as base year and j as given year. These are known as link indexes (relatives). The basic index number is obtained by the successive multiplication of the link indexes (relatives).

P_{01} = first link

$P_{12} = P_{01}$ x P_{12}

$P_{03} = P_{01}$x P_{12} xP_{23}

Utility of Chain Indexes

Over a number of years, various changes take place so that commodities shift considerably in their relative importance. Old commodities disappeared form scene are succeeded by new commodities models, styles. Grades of a commodity become obsolete and cease to be manufactured with new models, styles or grades taking their place; marketing centers shift, so that a price quotation at the new center must replace that at the old. Again, it rarely happens that comparable data are available for long periods; thus, the chain index is an extremely valuable tool.

(1) Commodities may readily be dropped, if they are no longer relevant;
(2) New commodities may be introduced and
(3) Weights may be changed.

Thus account may readily be taken of basic changes in production, distribution and consumption habits with the use of chain indexes.

This disadvantage of the chain index is that, while the percentage of previous year figures gives accurate comparisons of year to year changes, the long range comparisons of the 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 businessman, the percentages of the preceding year provide a flexible and useful tool.

1.9 Problems Involved in Computation of Price Index Number

The problems which the statistical encounters in index-number construction are:-

1) Definition of the purpose for which the index is being compiled.
2) Selection of data for inclusion in index
3) Selection of sources of data
4) Collection of data
5) Selection of base
6) Method of combining data
7) System of weighting.

Not all of the seven problems listed above are of equal importance, nor are they always independent of one another. Thus a simple system of weighting would require a different list of commodities for a price index than would a method that employs a separate weighting system for each subgroup of an index. Likewise, the weighting system to be used depends upon the method of combining the data. Let us discuss all the detail.

1) Definition of the Purpose for which the Index is being Compiled

Before gathering data and making calculations, it is important to know what we are trying to measure, and also how we intend to use our measures. An index number properly designed for the purpose in hand is a most useful and powerful tool. If not properly compiled and constructed, it can be dangerous one. If we wish to know changes in the cost of constructing private dwelling, we should not gather prices of heavy structure steel. Similarly, if we wish to measure the changes in family clothing costs, we should not gather prices of cotton by the bale. To measure the course of retail trade, we should use a sample of department store sales and not data from wholesalers and factories.

2) Selection of Data for Inclusion in Index

Although the method of combining the variables is of considerable importance in constructing index numbers, it is insignificant when compared with the problem of selecting the data that are the raw materials of the index. Too much emphasis cannot be put upon this point. The data must be accurate and homogeneous; and the sample should be a good representative of the whole. A sample cannot be expected to be representative unless adequate number of items are included. To state the idea in other language: a sufficiently large sample of relevant items mush be selected to obtain reliable index numbers.

As noted before, the commodities to be chosen for a price index, and the type of quotation to be selected, depend on what is being measured. A wholesale price index requires wholesale prices. An index of prices paid by consumers necessitates not only retail prices of food, but rents, gas and electric rates, clothing prices, transportation, medical care, and so forth, applying to the class of persons for whom the cost of living is to be ascertained.

3) Selection of Sources of Data

When selecting the sources of data for index numbers, we may rely on regularly published quotations or obtain periodic special reports from the merchant, producers, exporters, or others, who posses the basic information needed. Under either circumstance, we must make sure that the data pertains strictly to the thing being measured. Thus if retail food price changes are being measured, quotations should be from super markets, chain stores, independent stores and any other importance outlets. These different sources should not be mixed indiscriminately, but should be appropriately weighted when combined. Neither should first of the month quotations, middle-of-the-month quotations and end-of-the-month quotations ordinarily be combined in one index.

4) Collection of Data

Most index number are based on samples rather than on population data. This, together with the fact the proper choice of data is of great importance for the construction of index numbers, requires the following considerations to be borne in mind:

i) Accuracy: Statistical data that appear in precise printed form are not necessarily accurate. In this connection, it may be remembered that by and large, the primary source data are relatively more accurate than the secondary source data. It is therefore the responsibility of the statistician to ascertain how the data are collected and to select his source with discrimination.

ii) Comparability over Space and Time: When data are to be drawn from two or more sources, the reliability of each source must be considered and, in-addition, the user must be sure that from the different sources are comparable

5) Selection of Base

Regardless of the formula employed for weighting and combining the data, it customary (although not necessary) combining to select some period of time as 100 percent with which compare the other index numbers. A month is ordinarily too short a period to use of accident or since any one month is likely to be unusual on account of accidental or seasonal influences. A year is sometimes used. However, it is often true that no one year is sufficient "normal"- to be a good basis of comparison. Business and prices are always advancing or receding with the business cycle.

For example comparisons of wage increase with cost of living increases depend heavily on the period selected as the base for the index. It is to be expected that, for a base year in labour management wage negotiations, labour union leaders would select a year when employees were relatively well paid. Representatives of management would select the year before the major wage concession. It is surprising that both sides would prepare their own indexes and that they would arrive at widely varying conclusions. The selection of the base year can have a significant impact on the sympathy of Government and general public toward the demands or positions of the two parties to the negotiations.

For these various reasons, the base periods for most indexes have been changed from time to time. As a general rule and with notable exceptions, the base is no more than about ten to fifteen years prior to the current data. It is quite common for the base to be a single year, but many indexes have a three to five year period average as their base as an attempt to get a more normal period. Major changes in an index, such as the updating of the base, are accompanied by a recalculation of old index numbers to express them relative to the new base and often by changes in the commodities and the quantity weights. Index numbers covering a period of many year are at best only reasonably good estimates or relative prices.

6) Method of Combining Data

There are two method of constructing index numbers"

(1) By computing aggregate values (2) By averaging relatives. Each of these basic procedures may be further described as simple (un weighted) or specifically weighted. Although only the specifically weighted procedures are logically defensible, the simple un weighted

procedures are usually explained because they are sometimes used, and their development leads naturally to the problems and effects of various weighting systems.

The aggregative method obtains the result directly, and produces a result that has a simple and meaning; the method employing relatives is more roundabout and its meaning is more technical. Nevertheless, there are situations in which the aggregative method is not applicable and recourse must be then be had to the averaging of relatives.

7) Systems of Weighting

In selecting the weights for an index, attention must be paid to the use that will be made of the result. If a price index is to measure the changes in the prices paid by consumers, it is necessary for the weights applied to the various commodities to reflect the importance of the individual items to consumers. Likewise, the weights for an index of farm prices should reflect the importance of the individual items to consumers. Likewise, the weights for an index of farm prices should reflect the importance of the various commodities in the income of farmers. Frequently the commodities included in an index and the weights assigned to items do not give the information wanted for the solution of a particular problem. If weighted price indices are to be constructed and if the quantity of each commodity marketed changed from year to year in the same proportion, it would make no difference to what period the weights referred for the results would be identical. In fact however the relative importance of the different commodities is constantly changing, and this is due in part to the change in the relative prices of the different commodities which in turn result from changes in supply and demand. Therein lies a great source of difficulty for which there is no completely satisfactory solution. The answer depends in part on, what the analyst thinks a price index is supposed to do.

One view is that such an index number measure s the changing cost of a constant aggregate of goods. Another view concerns itself not with the goods level of analysis, but with the satisfactions level; an index number, according to this view, should measure the changing cost of aggregates of goods yielding the same utility or satisfactions at two period or two places. Thus suppose we are compare the cost of living of two groups of similar persons at two periods (or places), these groups having at the two periods (or places) the same tastes and capacity for enjoyment, as well as an income that will purchase, and does purchase for the same amount of satisfaction. The commodities of course, will be different but if the expenditures were Rs. 500 the first year and Rs. 550 the second year we may conclude that the cost of living has gone up by 10 percent. It goes without saying that no one has accurately made a measurement of this kind. Although it seems feasible to measure only the varying value of a fixed, aggregate of goods, yet the analyst should select a list of goods that will avoid the certainty of bias in a known direction with respect to the cost of obtaining equal satisfactions at different times.

1.10 Exercises

1) What are index numbers, write a note on necessity of index numbers?

2) What are importance points, which should be kept in mind while obtaining index numbers?

3) Define price index, quantity index and value index.

1.11 Key Words

Index Numbers: An index number is the relative change in the value of some characteristic over a period of time.

Price Relative: Ratio of the price of a single commodity in a given period to its price in another period (base period) is called a price relative.

Quantity Relative: Ratio of the quantity of a single commodity in a given period to its quantity in another period (base period) is called a quantity relative.

Link Relative: When we obtain relatives for a given period of the basis of pervious period, it is called link relative.

1.12 Summary

Index numbers are measures to get the pulse of an economy as they are indicators of inflationary (increasing) or deflationary (decreasing) trends of any economic criteria. We can obtain indicates for prices, wholesales prices, trade, export, import, agriculture production, employment generations or whatever one can think of. The collection of accurate and appropriate data is the most important thing and after that selection of base period is to be done carefully. It should be kept in mind that the two periods are similar for the factors going to effect index number. Choice of formula is also an important criteria, about which we will study in next unit.

1.13 Further Readings

- Allen, R.G.D.; Index Numbers of Theory and Practice, MacMillan, 1975.
- Croxton, F.E. and Cowden, D.J.; Applied General Statistical, Prentice Hall, 1967.
- Greenwald, W.I., Statistics for Economics, C.E. Merit Books; 1963.
- Mills, F.C.; Statistical Methods, Henry Holt, 1955.

Unit-2 Index Numbers: Important Formulae

Structure

2.1 Introduction

As described earlier index numbers (or indicates) are those measure which indicate about relative changes in the value of a variable over a period of time the most popular index number being the price index number. From the initial discussion, it is apparent that the index number is the ratio of two quantities with reference to two time periods. In index number is obtained for any given period generally termed as current period in comparison with some references period known as base period. For example if one has to speak about price index for cotton prices in year 1995 with the reference to 1980 as reference year 1980 will be termed as base year while 1995 will be known as current year.

Notation:

Before the discussion of construction of index number, let us explain the notation and terminology used.

I_{0j} $= Index\ number\ for\ the\ current\ year\ j\ with\ reference\ to\ base\ year\ 0$

$I_{01} = Index\ number\ for\ the\ current\ year\ 1\ with\ references\ to\ base\ year\ 0$

$P_{01} = Price\ Index\ number\ for\ the\ current\ period\ 1\ with\ references\ to\ base\ year$

p_{ij} $= Price\ of\ jth\ commodity\ in\ ith\ year$

p_{0j} $= Price\ of\ jth\ commodity\ in\ ith\ year\ '0'$

p_{1j} $= Price\ of\ jth\ commodity\ in\ ith\ year\ '1'$

q_{1j} $= Quantity\ consumed\ of\ jth\ commodity\ in\ ith\ \ year$

q_{0j} $= Quantity\ consumed\ of\ jth\ commodity\ in\ ith\ \ '0'\ base\ year$

q_{1j} $= Quantity\ consumed\ of\ jth\ commodity\ in\ current\ year$

2.2 Objectives

In the previous unit, you have studied about what an index number is and why do one need it. After reading this unit you must be able to understand:

- How to calculate index numbers
- Unweighted and weighted index numbers.
- Important formula of index numbers
- Time reversal factor reversal test.
- Fisher's ideal index number.

2.3 Calculation of Index Numbers

Broadly the calculation of price index can be divided into two subgroups, namely

(a) Simple (unweighted) Aggregate method.

(b) Weighted Aggregates method.

Now, let us discuss both of these one by one.

(a) Simple Aggregate Method: This method consists of expressing aggregate of prices in any year as a percentage of their aggregate in base year.

Thus price index for the ith year as compared to base year ('0') in given as

$$P_{oi} = \left(\frac{\sum p_{ij}}{\sum p_{oj}}\right) \times 100$$

and

Quantity index-

$$Q_{oi} = \left(\frac{\sum q_{ij}}{\sum q_{oj}}\right) \times 100$$

This formula is very simple for the purpose of calculation and so provides a quick measure of index number when one has to obtain Index number for similar type of articles, e.g. crop items like wheat, rice, bajra, gram etc. But if prices of commodities under study have different units i.e. per kg, per meter, per ton, this formula is not of any use different units cannot be summed us directly.

Thus merits and demerits of this formula are-

Merits:

1. It is an easy and quick method and so come in handy for a quick overview of the situation.

Demerits:

1. It does not take into account the fact that articles or commodities whose prices are to be added have different importance.

Thus prices of commodities which are not much important will affect the index number.

2. As said earlier, this formula cannot be used when commodities involved have different units for price or for quantities

During to these demerits, this formula is not of much use for practical purposes.

(b) Weighted Aggregate Method: In this method appropriate weights are assigned to different commodities to make them comparable and thus

$$P_{oi} = \left(\frac{\sum p_{ij} w_j}{\sum p_{oj} w_j}\right) \times 100$$

and

$$Q_{oi} = \left(\frac{\sum q_{ij}\, w_j}{\sum q_{oj}\, w_j}\right) \times 100$$

Where W_j is the weight assigned to jth commodity.

Thus in this method the commodities of higher importance are given weight and vice versa. In this way each commodity selected for obtaining the index number influence it according to its weight (i.e. importance, literally). Or in other words, allotment of weights enables the commodities of greater importance to have more impact on index number.

Now, time to time various renowned statisticians and research workers have suggested different methods for allotment of weight (W_j) to a large number of choices to select from. Some important formulae, which are treated as standard ones are given below.

2.4 Important Formulae

Various methods have been suggested by different workers time to time. These methods give different formula on the basis of choice of different weights. Though the basic character of an index number does not changes and it gives the relative change yet choice of weight changes the utility and purpose of index number and vice versa.

2.5 Laspeyre's Formulae

French economist Laspeyere in 1871 suggested that quantities of commodities consumed in base year can be taken as weights for the purpose of calculating index numbers.

That is

$$P_{oi}^{La} = \frac{\sum p_{ij}\, q_{oj}}{\sum p_{oj}\, q_{oj}}$$

Where notation bear usual meaning. The upper suffix 'La' is added to distinguish this formula as 'Laspeyere's' Index number.

2.6 Paasche's Formulae

By taking year quantities as weights, we get Paasche's formula. This formula is suggested by German statistician Paasche in 1874 and so it is named after him and while writing this formula, as in earlier case we add upper suffice 'Pa'. Thus

$$P_{oj}^{La} = \frac{\sum p_{ij}\, q_{ij}}{\sum p_{oj}\, q_{ij}}$$

2.7 Edgeworth Marshall's Formulae

The statistician Marshall and Edgeworth suggested that to obtain a better index number we should use simple average of base year and current year quantities as the weights.

$$W_{ij} = \frac{q_{oj} + q_{ij}}{2}$$

i.e.

Hence, Marshall Edgeworth Price index no. is-

$$P_{oi}^{ME} = \frac{\sum \frac{q_{oj} + q_{ij}}{2}}{\sum p_{oj} \frac{q_{oj} + q_{ij}}{2}}$$

$$= \frac{\sum p_{ij} \frac{(q_{oj} + q_{ij})}{2}}{\sum p_{oj} \frac{(q_{oj} + q_{ij})}{2}}$$

2.8 Fisher's Formulae

Fisher's index number is the geometric mean of the Lasepeyer's and Paasche's formula. Mathematically-

$$P_{oj} = [P_{oi}^{La} \times P_{oi}^{Pa}]^{\frac{1}{2}}$$

$$= \left[\frac{\sum p_{ij}\, q_{oj}}{\sum p_{oj}\, q_{oj}} \times \frac{\sum p_{ij}\, q_{1j}}{\sum p_{oj}\, q_{1j}}\right]^{1/2} \times 100$$

This formula holds the view that neither base year nor current year quantities are fully appropriate to use as weights, so a geometric mean of Laspeyere's and Passche's formula can be more appropriately used as weight for obtain index number. We will further observe that this formula satisfies several tests and so it is taken as the best of all formulae suggested above. In fact it is known as *"**Fisher's Ideal Index Number**".*

There are several other index numbers also, given my other research workers. Some known names are Dorbish – Bowley's index number and Walsch's index number. But only four formulae mentioned above are most commonly used in general Practice and so these are established standard formulae for calculating Price Indices.

2.9 Other Indices

On the basis of requirement of economic workers various formulae can be developed. For example we can have an index of global warming, an index for depletion of ozone layer, an index for change in literacy rates in different parts of the country etc. The thing is that the sky is the limit when we go for applications of index numbers.On the basis of different roles these indicates play and on the basis of different factors these are based on these indices are known by different names. Some of the importance formulae are listed below.

2.10 Quantity Index Number

We have concentrated ourselves on price index numbers. By interchanging the price (p_{ij}) and quantities (q_{ij}) in the above mentioned formulae we get the corresponding formulae for the calculation of quantity index numbers which reflect the change in the volume of quantity or production. Thus for example,

$$Q_{oj}^{La} = \left(\frac{\sum q_{ij} p_{oj}}{\sum q_{oj} p_{oj}}\right) \times 100$$

$$Q_{oi}^{Pa} = \left(\frac{\sum q_{ij} p_{ij}}{\sum q_{oj} p_{ij}}\right) \times 100$$

$$Q_{oi}^{ME} = \left(\frac{\sum q_{ij} \left(p_{ij} + p_{oj}\right)}{\sum q_{oj} \left(p_{ij} + p_{oj}\right)}\right) \times 100$$

Quantity index number study the changes in the volume of goods produced (manufactured), consumed or distributed like the indices of agriculture production, industrial production, imports and exports etc. They are extremely helpful in studying the level of physical output in an economy.

2.11 Key Index

Value index numbers are given by the aggregate expenditure for any given year expressed as a percentage of the same in the base year.

Thus

$$V_{oi} = \left(\frac{\sum q_{ij}\, p_{ij}}{\sum q_{oj}\, p_{ij}}\right) \times 100$$

However these indices are not as common as price and quantity indices.

2.12 Some Worked Examples

Example 1: The table below relates to the daily pay of the wage earners on a company's pay roll:

	April 1978		April 1983	
	Number	Total Pay (Rs.)	Number	Total Pay (Rs.)
Men aged 21 and over	350	2500	300	4200
Women aged 21 and over	400	1600	1200	8000
Youths and boys	150	450	100	560
Girls	100	250	400	1540
Total	1000	4800	2000	14300

Construct an index of daily earning based on 1978 as base showing the rise of earning for all employees as one figure.

Solution: Regarding number wage earns as quantities and the weekly wages per labourer as prices we are given the figure q_0 and p_0q_0 for 1978 and the values q_1 and p_1q_1 for 1983. The following table can be easily completed.

q_{oj}	p_{oj}	q_{1j}	p_{1j}	$p_{oj}\, q_{oj}$	$p_{oj}\, q_{1j}$	$p_{1j}\, q_{1j}$	$p_{1j}\, q_{oj}$
350	7.14	300	14.00	2,500	2,142	4,200	4,900
400	4.00	1,200	6.67	1,600	4,800	4,800	2,668
150	3.00	100	5.60	450	300	560	840
100	2.50	400	3.85	250	1,000	1,540	385
Total				**4,800**	**8,242**	**14,300**	**8,793**

$$P_{oi}^{La} = \frac{\sum p_{ij}\, q_{oj}}{\sum p_{oj}\, q_{oj}} \times 100 = \frac{8{,}793}{4{,}800} \times 100 = 183$$

$$P_{oi}^{Pa} = \frac{\sum p_{ij}\, q_{ij}}{\sum p_{oj}\, q_{ij}} \times 100 = \frac{14{,}300}{8{,}242} \times 100 = 173.5$$

Hence,

$$P_{oi}^{f} = \{183 \times 173.5\}^{\frac{1}{2}} = 178.3$$

Example 2: Given the data

Commodities		
	A	B
P_o	1	1
q_o	10	5
p_1	2	X
q_1	5	2

Where p and q respectively stand for price and quantity and subscripts stand for time period. Find X, if the ratio between Laspeyre's (L) and Paasche's (P) index numbers is L : P: : 28: 27

Solution:

CALCULATIONS FOR LASPEYRE'S AND PAASCHE'S INDICES

Commodities	p_{oj}	q_{oj}	p_{1j}	q_{1j}	$p_{1j}\,q_{oj}$	$p_{oj}\,q_{0j}$	$p_{1j}\,q_{1j}$	$p_{0j}\,q_{1j}$
A	1	10	2	5	20	10	20	5
B	1	5	X	2	5X	5	2X	2
Total					20+5X	15	10+2X	7

$$P_{oi}^{La} = \frac{\sum p_{ij}\,q_{oj}}{\sum p_{oj}\,q_{oj}} \times 100 = \frac{20+5X}{15} \times 100$$

$$P_{oi}^{Pa} = \frac{\sum p_{ij}\,q_{ij}}{\sum p_{oj}\,q_{ij}} \times 100 = \frac{10+2X}{7} \times 100$$

$$\frac{P_{oi}^{La}}{P_{oi}^{Pa}} = \frac{\frac{20+5X}{15}}{\frac{10+2X}{7}} = \frac{28}{27}$$

$$\frac{20+5X}{15} \times \frac{7}{10+2X} = \frac{28}{27}$$

X= 4 (on simplification)

Example 3.: Calculate a suitable weighted price index from the following data:

Material Required	Unit	Quantity Required	Prices During	
			Base Year (Rs.)	Current Year (Rs.)
Cement	100 Ib	500Ib	5.0	8.0
Timber	c.ft.	2,000 c.ft.	9.5	14.2
Steel Sheet	cwt.	50 cwt.	34.0	4.0
Bricks	Per '000	20,000	12.0	24.0

Solution: Since the weights here are fixed (neither relating to current year nor to base year), we use fixed base method (Kelly's method) for computing the index.

Material Required	Unit	Quantity q	p_o	p_1	$p_o q$	$p_1 q$
Cement	100Ib	500 Ib	5.0	8.0	25	40
Timber	c.ft.	2,000 c.ft.	9.5	14.2	19,000	28,400
Steel Sheet	cwt.	50 cwt.	34.0	4.0	1,700	2,100
Bricks	per'000	20,000	12.0	24.0	240	480
Total					20,965	31,020

$$P_{oi}^{k} = \frac{\sum p_{ij}\, q_j}{\sum p_{oj}\, q_j} \times 100 = \frac{31,020}{20,965} \times 100 = 148$$

Example 4 Construct Quantity Index Numbers taking 1980 as the base:

Commodities	Average Price	Production			
		1980	1981	1982	1983
A	1.00	62	65	66	90
B	1.50	138	120	110	60
C	0.25	500	540	580	800
D	2.25	10	10	10	10

Solution.

CALCULATION FOR QUANTITY INDEX NUMBERS

Commodities	Price P	q_o	pq_o	q_1	pq_1	q_2	pq_2	q_3	pq_3
A	1.00	62	62.0	65	65.0	66	66.0	90	90.0
B	1.50	138	207.0	120	180.0	110	165.0	60	120.0
C	0.25	500	125.0	540	135.0	580	145.0	800	200.0
D	2.25	10	22.5	10	22.5	10	22.5	10	22.5
Total			416.5		402.5		398.5		432.5
			100	$\frac{402.5}{416.5} \times 100 = 96.6$		$\frac{398.5}{416.5} \times 100 = 95.7$		$\frac{432.5}{416.5} \times 100 = 103.8$	

2.13 Criteria of a Good Index Number

A number of mathematical test discussed below have been suggested for comparing various index numbers.

1) **Unit Test:** This test requires the index numbers to be independent of the units in which prices and quantities are quoted. This test is satisfied by all the formulae discussed in 2.2.

2) **Time Reversal Test:** This is one of the two very important test proposed by Irving Fisher as tests f consistency for a good index number. According to this the formula for calculating an 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 notations we have-

$$P_{ij} \times P_{ji} = 1 \;\; or \;\; P_{ij} = 1/P_{ji}$$

For example if we take the Laspeyre's formula

$$P_{oi}^{La} = \frac{\sum p_{ij}\, q_{oj}}{\sum p_{oj}\, q_{oj}} \times 100$$

$$P_{oi}^{Pa} = \frac{\sum p_{ij}\, q_{ij}}{\sum p_{oj}\, q_{ij}} \times 100$$

$$P_{oi}^{La} \times P_{oi}^{Pa} = \frac{\sum p_{ij}\, q_{oj}}{\sum p_{oj}\, q_{oj}} \times \frac{\sum p_{ij}\, q_{ij}}{\sum p_{oj}\, q_{ij}} \neq 1$$

Hence laspeyre's formula does not satisfy Time Reversal Test. Similarly it can be seen that Paasche's formula also does not satisfy this test.

For the Fisher's ideal formula.

$$P_{oi}^{F} = \left[\frac{\sum p_{1j}\, q_{oj}}{\sum p_{oj}\, q_{oj}} \times \frac{\sum p_{ij}\, q_{oj}}{\sum p_{oj}\, q_{ij}}\right]^{1/2}$$

$$P_{io}^{F} = \left[\frac{\sum p_{oj}\, q_{ij}}{\sum p_{ij}\, q_{ij}} \times \frac{\sum p_{oj}\, q_{oj}}{\sum p_{ij}\, q_{oj}}\right]^{1/2}$$

$$P_{oi}^{F} \times P_{io}^{F} = 1$$

Hence Fisher's ideal satisfies Time Reversal Test.

3) **Factor Reversal Test:** This is the second test of consistency suggested by I. Fisher. In his words: " Just as our formula should permit the interchange of two items without giving inconsistent results, so it ought to permit interchanging the prices and quantities without giving inconsistent results –i.e. the two result multiplied together should give the true value ratio, except for a constant of proportionally".

Symbolically we should have

$$P_{oi} \times Q_{oi} = \frac{\sum V_{ij}}{\sum V_{oj}} = \frac{\sum p_{ij} p_{ij}}{\sum p_{oj} p_{oj}}$$

For example

$$P_{oi}^{F} = \left[\frac{\sum p_{ij}\, q_{oj}}{\sum p_{oj}\, q_{oj}} \times \frac{\sum p_{ij}\, q_{oj}}{\sum p_{oj}\, q_{ij}}\right]^{1/2}$$

and

$$Q_{oi}^{F} = \left[\frac{\sum q_{ij}\, p_{oj}}{\sum q_{oj}\, p_{oj}} \times \frac{\sum p_{ij}\, q_{ij}}{\sum q_{oj}\, p_{ij}}\right]^{1/2}$$

$$P_{oi}^{F} \times Q_{oi}^{F} = \frac{\sum p_{ij}\, q_{ij}}{\sum p_{oj}\, q_{ij}} = V_{oi}$$

Hence Fisher's ideal index satisfies Factor Reversal Test. It may be pointed out that none of the other formulae satisfies the factor reversal test.

So we see that Fisher's index satisfies both time reversal and factor reversal test. Hence it is termed as 'Ideal Index Number'.

2.14 Key Words

Laspeyre's Index Number: An index number obtaining by using the base year quantities as weights.

Paasche's Index Number: An index obtained by using current year quantities as weights.

Fisher's index number: It is the geometric mean of the Laspeyre's and Paasche's index number.

2.15 Exercises

1) The prices per unit the number of united consumed for three commodities A, B, and C for two time-periods are given below:

Commodities	Base Year		Current Year	
Quantity	Price	Quantity		Price
A	1	6	3	5

B	3	5	8	5
C	4	8	10	6

Construct price index number using simple aggregative price index number formula.

2) Prepare price and quantity index numbers for 1983 with 1982 as base year from the following data by using (i) Laspeyre's. (ii) Paasche's (iii) Marshall- Edgworth, and (iv) Fisher's Method.

Year	Article I		Article II		Article III		Article IV	
	Price	Qty.	Price	Qty.	Price	Qty.	Price	Qty.
1982	5.00	5	7.75	6	9.63	4	12.50	9
1983	6.50	4	8.80	10	7.75	6	12.75	9

Also show that Fisher's formula satisfies Factor Reversal and Time Reversal Test.

2.16 Summary

Index numbers of the data related to prices, production, profits imports and exports etc. are indispensable for any organization for efficient planning and formulation of executive decisions. Although the real aim of index numbers in the beginning was to deal with changes in economic patterns, they are now widely used by sociologists, psychologists, health and educational authorities etc.

To real problem in the construction of index numbers is the selection of appropriate weights. Thus it is the main decision which is to be taken by the maker of the index number that how to rationalize the weights.

Availability of the information regarding index number consist in expressing the aggregate of prices in any year (under consideration) as a percentage of their aggregate in base year. But in this method relative importance of the various commodities is neglected and if the group of articles carries different units we cannot use this method.

In Laspeyre's index number base year quantities are used as weight and in Paasche's index number current year quantities are used as weights.

Fisher's index number is the geometric mean of the Laspeyre's and Paasche's index number. It is also termed as ideal index number as it satisfies both time reversal and factor reversal tests.

2.17 Further Readings

- Allen, R.G.D.; Index Numbers of Theory and Practice, MacMillan, 1975.
- Croxton, F.E. and Cowden, D.J.; Applied General Statistical, Prentice Hall, 1967.
- Greenwald, W.I., Statistics for Economics, C.E. Merit Books; 1963.
- Mills, F.C.; Statistical Methods, Henry Holt, 1955.

Unit-3 Consumer Price Index Number

Structure

3.1 Introduction

As we have studied earlier, price index no. is the measure of relative changes in the prices of some commodity over a period of time. In practice these indices are used for various different purposes. One very important use of the theory of index number is in obtaining consumer price index or alternative also called cost of living index number.

It is a well known fact that the prices of the commodities required for day to day living go on increasing e.g. prices of food items like wheat, rice, oil etc. are different in different years. This increase (or decrease, if there is any) in prices of commodities directly hit the purchasing power of consumer. A consumer price index is therefore devised to measure the over all changes in the purchasing power of the consumer.

A consumer price index or cost of living index, is a measure which indicates the relative changes in the prices of a group of items, necessary for the living for a selected group of consumers. In a way it tells us about what should be the increase in the wages of consumer so that they are able to maintain some standard of living in two time periods. For this purpose, the total expenditure of a household are categorized like food, clothing, rent, electricity, entertainment, education, medicines, miscellaneous etc.

Consumer price index no. or cost of living index numbers measures the effect of changes in the prices of the described basket of goods and services on the purchasing power of a particular class of people during current period as compared with some base period. Change in the cost of living of an individual between two periods means the change in his money income, which will be necessary for him to maintain the same standard of living in both periods. There the cost of living index numbers are intended to measure the average charge in the cost of maintaining the same standard in a given year as in the base year.

3.2 Objectives

The purpose of this unit is to make you aware of the most important applications of the index number i.e. Consumer price index after reading this unit will be able to:

- Get an idea of consumer price index number
- Know how to computer consumer price index number
- Know about base shifting, splicing and deflating of index numbers.
- Get an idea of industrial production index.

3.3 Construction of Consumer Price Index Number

Thus the consumer price index numbers, also known as cost of living index numbers are generally intended to represent average change over time in the prices paid by the ultimate consumer of a specified group, goods and services. The need for constructing consumer price indices arises because the general index numbers fail to give an exact idea of effect of the change in the general price level on the cost of living of different classes of people, since a given change in the level of prices aces affects different classes of people in different manners. Different classes of people consumer different types of commodities and even the same types of 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 various widely. Not only this, the consumption habits of the people of the same class differ from place to place. For example the mode of expenses true 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 determining the effect of rise and fall in prices on different classes or consumers living in different areas, 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 full of goods and services in a given period compared to the certain period. To bring out clearly this fact, the sixth international conference of Labor Statisticians recommended that the term ‘cost of living index’ should be replaced in appropriate circumstances by the terms ‘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 actual living costs of the two areas. A higher index for one area than for another with the same period is not indication that living costs are higher in the one than in the other. All it means is that as compared with the base period prices have risen more 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 region and for different classes of population.

3.4 Computation of Consumer Price Index Number

For computing consumer price index number or cost of living index number we need to have prices and quantity consumed for different categories of items in base a year and current year.

Let

$p_{0j} = Price\ of\ jth\ commodity\ in\ year\ '0'\ (base\ year)$

$p_{ij} = Price\ of\ jth\ commodity\ in\ I\ year\ (current\ year)$

$q_{0j} = Quantity\ consumed\ for\ jth\ commodity\ in\ '0'\ year\ (base\ year)$

Now consumer price index number is derived as the weighted average of the price relatives, the weight being the values of the quantities consumed in the base year.

Price relative

$$P_j = \frac{P_{ij}}{p_{oj}} \times 100$$

And weight

$$W_j = p_{oi} \times q_{oj}$$

Then,

Consumer Price Index

$$= \frac{\sum W_j P_j}{\sum W_j}$$

3.5 Some Illustrations

Example 3.1: Construct the cost of living index for the year 1982 (base 1980=100).

Item	Unit	Price (1981)	Price (1983)	Weight
A	Kg./	0.50	0.75	10%
B	Litre	0.60	0.75	25%
C	Dozen	2.00	2.40	20%
D	Kg.	0.80	1.00	40%
E	One pair	8.00	10.00	5%

Solution: The consumer price index is obtained by the method of weighted price relatives.

COMPUTATION OF COST OF LIVING INDEX NUMBER

Item	Price in Rupees		Price Relatives (base 1980)	Weight	PW
	1981 (P_0)	1983 (P_1)	$P = 100 \times \frac{P_1}{P_0}$	W	
A	0.50	0.75	$100 \times \frac{0.75}{0.50} = 150$	10	1,500
B	0.60	0.75	$100 \times \frac{0.75}{0.60} = 125$	25	3,125
C	2.00	2.40	$100 \times \frac{2.40}{2.00} = 120$	20	2,400
D	0.80	1.00	$100 \times \frac{1.00}{0.80} = 125$	40	5,000
E	8.00	10.00	$100 \times \frac{10.00}{8.00} = 125$	5	625
Total				100	12,650

$$\frac{\sum PW}{\sum W} = \frac{12,650}{100} = 126.5$$

Cost of living index= 126.5

Example 3.2: From data given below, calculate the cost of living index number for the current year by the aggregate expenditure method .

Article	Quantity Consumed in base year	Unit	Price (in Rs.) per unit	
			Base Year	Current Year

Rice	5 quintals	Quintals	60	80
Millets	5 quintals	Quintals	40	50
Wheat	1 quintals	Quintals	50	100
Gram	1 quintals	Quintals	30	60
Arhar	½ quintals	Quintals	40	60
Other pulses	2 quintals	Quintals	30	40
Ghee	4 kg	Kg	12.5	20
Gur	2 quintals	Quintals	25	50
Salt	12kg	Kg	40	50
Oil	24kg	Kg	200	250
Clothing	40 metres	Metre	2.5	5
Firewood	10 quintals	Quintals	5	8
Kerosene	1 tin	Tin	40	60
House Rent		-	120	150

Solution

Aggregate Expenditure Method

Article	Quantity Consumed in base year	Unit	Price in Base Year (P_0)	Price in Current Year (P_1)	Aggregate Expenditure in base year ($p_o q_o$)	Aggregate Expenditure in current Year ($p_1 q_o$)
Rice	5 quintals	Quintal	60	80	300	400
Millets	5 quintals	Quintal	40	50	200	250
Wheat	1 quintals	Quintal	50	100	50	100
Gram	1 quintals	Quintal	30	60	30	60
Arhar	½ quintals	Quintal	40	60	20	30
Other pulses	2 quintals	Quintal	30	40	60	80
Ghee	4 kg	Kg	12.5	20	50	80
Gur	2 quintals	Kg	25	50	50	100
Salt	12kg	Kg	40	50	400	625

Oil	24kg	Kg	200	250	4,800	6,000
Clothing	40 metres	Metre	2.5	5	100	200
Firewood	10 quintals	Quintal	5	8	50	80
Kerosene	1 tin	Tin	40	60	40	60
House Rent		-	120	150	120	160
Total					6,370	8,215

Index number for current year

$$= \frac{\sum p_1 q_o}{\sum p_0 q_o} \times \frac{8215}{6370} \times 100 = 128.9$$

3.6 Steps in Construction of Consumer Price Index

The main steps which are required for construction of CPI are described below.

(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 to teachers, we are referring to primary teachers, middle class teachers, etc. or to all the teachers taken together. Along with the class of people it is also necessary to decide the geographical area covered by the index.

(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 consumptions. While conduction such an enquiry, 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 period . The sixth international conference of labour statisticians held in Geneva 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.

(3) Deciding on the items. The items on which the money is spent are classified into certain well- accepted groups. One of the choicest and most frequently used classification is-

a) Food
b) Clothing
c) Fuel and lighting
d) House rent
e) 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 alterations in supply and for which regular and comparable quotations or prices can be obtained.

(4) Obtaining price quotations. The collection of retail prices is very important and at the same time, very tedious and difficult task also. That is 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 or from where they usually make their purchases. Some of the principles recommended to be observed in the collection of retail price data required for purpose 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 to consumers for cash sales.

(c) Discount should be taken into account if it is automatically 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 confirm 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 prices form the most important component of cost of living indices, considerable attention has to be paid to the methods of price collections and to the price collection personnel . Prices are collected usually by special agent for through mailed questionnaire or in some cases through published price lists. The greatest reliance can be placed on the price collection through special agents as they visit the selected retail out-lots and collect the prices from them. However these agents should be properly selected and trained and should be given a manual of instructions as well as manual of specifications of items to be priced. Appropriate methods of price verification should be followed such as 'check pricing' in which price quotations are verified by means of duplicate prices obtained by different agents or purchase checking in which actual purchases of goods are made.

(5) Working on CPL. 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 every month if the index is maintained on monthly basis. The method of averaging the quotations should be should be such as to yield unbiased estimates of averages 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 weighting 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 averages family is decided and these constitute the weights. Percentages 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 the group weight.

3.7 Uses and Limitations of Consumer Price Index

(1) These indices are compiled for different groups or classes of people (such as low income, middle income, clerical, labor class, etc.) and are useful to assess the general price movement of the commodities consumed by them.

(2) Cost of living index number indicate whether the real wages are raising or falling, money wages remaining unchanged. In other words they are used for the calculation of real wages and for determining the change in the purchasing power of the money.

(3) Cost of living indices are used for the regulation of dearness allowance of living or the grant of bonus to the workers so as to enable them to meet national accounts.

(4) These indices are also used for deflation of income and value series in national accounts.

(5) By itself cost of living index number does not throw much light on the inflationary or deflationary trend on the soundness of an economy but in conjunction with other tools such as the indices of wholesale prices, wages, profits, production, employment etc., it serves as an economic indicator for the analysis of price situation.

3.8 Base Shifting of Index Numbers

Base shifting means the changing of the given base period (year) of a series of index numbers and recasting them into a new series based on some recent new base period. This step is quite often necessary under the following situations:

(i) When the base year is too old or too distance from the current period to make meaningful and valid comparisons.

(ii) If we want to compare series of index numbers with different base periods to make quick and valid comparisons both the series must be expressed with a common base period.

Base shifting requires the recompilation of the entire series of the index numbers with the new base. However this is a very difficult and time consuming job. A relatively much simple though approximate method consist in taking the index number of the new base year as 100 and then expressing the given series of index numbers as a percentage of the index number of the time period selected as the new base year. Thus, the series of index numbers, recast with a new base year. Thus, the series of index numbers, recast with a new base is obtained by the formula:

$$\text{Recast Index number of any year} = \frac{Old\ index\ number\ of\ the\ year}{Index\ number\ of\ New\ base\ year} \times 100$$

$$= \frac{100}{Index\ Number\ of\ New\ base\ year} \times Index\ number\ of\ the\ Year$$

In other words, the new series of index numbers is obtained on multiplying the old index numbers with a common factor:

$$= \frac{100}{Index\ Number\ of\ New\ base\ year}$$

Example 3.3: Form the index numbers given below, find out index numbers by shifting base from 1960 to 1979.

Year	1960	1961	1962	------	1979	1980	1981	1982
Index No.	100	76	68	------	50	60	70	75

Solution:

Year	Index No.	Index Number with base 1960
1960	100	$\frac{100}{50} \times 100 = 200$
1961	76	$\frac{76}{50} \times 100 = 152$
1962	68	$\frac{68}{50} \times 100 = 136$
1979	50	$\frac{50}{50} \times 100 = 100$
1980	60	$\frac{60}{50} \times 100 = 120$
1981	70	$\frac{70}{50} \times 100 = 140$
1982	75	$\frac{75}{50} \times 100 = 150$

If arithmetic mean or median is used for averaging the price relatives then the method of base shifting consists in calculating the price relatives for each individual item w.r.t. in new base and then averaging their totals, i.e., the whole of the series is to be reconstructed. How ever, in practice even in these cases the approximate method describe above gives result which are fairly close to those obtained otherwise.

3.9 Splicing Two Index Numbers Series

In order to obtain continuity in the comparison of two or more overlapping series of index numbers, we combine or splice them into a single continuous series. For example, suppose an index number series 'A' with base period 'a' is discontinued in 'period' [b] due to certain reasons and a new series 'B' of index numbers is computed with base period 'b' (nd the same items). In order to compare the series 'B' with 'A' we splice the series from 'a' onwards. The process is very much alike to that of base shifting and is illustrated below:

SPLICING OF TWO INDEX NUMBER SERIES

Year	Series I Base 'a'	Series II Base 'a'	Series II (Base 'a') spliced to series I	Series I spliced to series II (Base 'b')
a	100		100	$\frac{100}{a_k} \times 100$
a+1	a_1		a_1	$\frac{100}{a_k} \times a_1$
a+1	a_2		a_2	$\frac{100}{a_k} \times a_2$
b-1	a_{k-1}		a_{k-1}	$\frac{100}{a_k} \times a_{k-1}$
b	a_k	100	a_k	100
b+1		b_1	$\frac{a_k}{100} \times b_1$	b_1
b+2		b_2	$\frac{a_k}{100} \times b_2$	b_2
b+3		b_3	$\frac{a_k}{100} \times b_3$	b_3

Explanation. When series II is spliced to series I to get a continuous series with base 'a'.

100 of II series becomes a_k

b_1 of II series becomes $\frac{a_k}{100} \times b_1$

and b_2 of II series becomes $\frac{a_k}{100} \times b_2$ and so on. Thus multiplying each index of series II with a constant factor $\frac{a_k}{100}$ we get the new series of index numbers spliced to series I (base 'a'). In this case series I is also said to be spliced forward.

If we splice series I to series II to get a new continuous series with base 'b' then,

a_k of first series becomes 100

a_{k-1} of first series becomes $\frac{100}{a_k} \times a_{k-1}$

a_2 of first series becomes $\frac{100}{a_k} \times a_2$

Thus the new series of index numbers with series I Spiced to series II (Base ‘b) obtained on multiplying each index of series I by new constant factor ($100/a_k$). In this case we say that series is spliced backward.

Example 3.4: Given below are two price index series. Spline them on the base 1974 = 100. By what percent did the price of steel rise between 1970 and 1975?

Year	Old price index for steel base (1965=100)	New price index for steel base (1974=100)
1970	141.5	
1971	163.7	
1972	158.2	
1973	156.8	99.8
1974	157.1	100.0
1975		102.3

Solution:

SPLICING OF OLD PRICE INDEX TO NEW PRICE INDEX

Year	Old price index for steel base (1965=100)	New price index for steel base (1974=100)
1970	141.5	$\frac{100}{157.1} \times 141.5 = 99.06$
1971	163.7	$\frac{100}{157.1} \times 163.7 = 104.19$
1972	158.2	$\frac{100}{157.1} \times 158.2 = 100.69$
1973	156.8	$\frac{100}{157.1} \times 156.8 = 99.80$
1974	157.1	100.0
1975		102.3

The percentage increase in the price of steel between 1970 and 1975 is

$$\frac{102.30 - 90.06}{90.06} \times 100 = 0.1359 \times 100 = 13.59$$

Hence the require increase is 13.59%

3.10 Deflating the Index Number

Deflating means “making allowance for the effect of changing price levels”. The increase in the prices of consumer goods for a class of people over a period of years means a reduction in the purchasing power for the class. For example the increase in price of a particular commodity from Rs. x in base year ‘a’ to Rs. 2x in the year ‘b’ implies that in ‘b’ a person can by only half the amount of the commodity with Rs. x which he was spending on it in ‘a’. Thus the purchasing power of a rupee in only 50 paise in ‘b’ as compared to ‘a’.

The idea of “the purchasing power of money” or “a measure of the real income” for a class of people is obtained on deflating the wage series by dividing each item by an approximate price index e.g. the cost of living index. The real wages so obtained my be converted into index number if desirable. More precisely,

$$Real\ wage = \frac{Money\ or\ Nominal\ Wages}{Price\ Index} \times 100$$

The real income is also known as deflated income. This technique is extensively used to deflate value series or value indices, rupee sales inventories, income wages and so on.

The following The real income is also known as deflated income. This technique is extensively used to deflate value series or value indices, rupee sales inventories, income wages and so on.

The following example illustrates the technique of ‘deflating’.

Example 3.5: The following table shows the average wages in Rs. per hour of railroad workers during the year 1947 to 1958. So also are given the consumer Price Indices for these years with 1947 to 1958. So also are given the consumer Price Indices for year with 1947 to 1949 as the base period.

Year	Average wage of workers in Rs. per hour	Consumer price indices (1947-49) as base period
1947	1.19	99.5
1948	1.33	102.8
1949	1.44	101.8
1950	1.57	102.8
1951	1.75	111.0
1952	1.84	113.5

1953	1.89	114.4
1954	1.94	114.8
1955	1.97	114.5
1956	2.13	116.2
1957	2.28	120.2
1958	2.45	123.5

(a) Determine the real wages of the rail road workers during the years 1947-1958 as compared to their wages in 1947.

(b) Use the consumer Price Index to determine the purchasing power of rupee for the various years assuming that in 1947, one rupee was strictly worth rupee one in purchasing power.

Solution:

Year (1)	Average wage of workers in Rs. per hour (2)	Consumer price indices (1947-49) as base period (3)	Consumer price index (base 1947)(4)	Deflated or Real Wages (Base 1947) (5) = $\frac{2}{5}\times 100$	Purchasing power of rupee (base 1947)
1947	1.19	99.5	100	1.19	1.00
1948	1.33	102.8	$\frac{102.8}{95.5}\times 100 = 107.6$	1.24	0.93
1949	1.44	101.8	$\frac{101.8}{95.5}\times 100 = 106.5$	1.35	0.94
1950	1.57	102.8	$\frac{102.8}{95.5}\times 100 = 107.6$	1.46	0.93
1951	1.75	111.0	$\frac{111.0}{95.5}\times 100 = 116.2$	1.51	0.86
1952	1.84	113.5	$\frac{113.5}{95.5}\times 100 = 118.8$	1.55	0.84
1953	1.89	114.4	$\frac{114.4}{95.5}\times 100 = 119.7$	1.58	0.83
1954	1.94	114.8	$\frac{114.8}{95.5}\times 100 = 120.2$	1.61	0.83
1955	1.97	114.5	$\frac{114.5}{95.5}\times 100 = 119.9$	1.64	0.83

1956	2.13	116.2	$\frac{116.2}{95.5} \times 100 = 121.7$	1.75	0.82
1957	2.28	120.2	$\frac{120.2}{95.5} \times 100 = 125.9$	1.81	0.79
1958	2.45	123.5	$\frac{123.5}{95.5} \times 100 = 129.3$	1.89	0.77

The real wage of the workers in any year as compared to 1947 will be the wage in that year multiplied by the purchasing power of one rupee in that year as compared to 1947.

3.11 Index of Industrial Production

The index of industrial production is aimed at reflecting changes (increase or decrease) in the volume of industrial production (i.e., production of non agricultural commodities) in a given period compared to some base period. These indices measure, at regular intervals the general movement in the quantum of industrial production. Such indices are useful for studying:

(i) The progress of general industrialization of a country and

(ii) The effect of tariff on the development of particular industries.

These indices of industrials activity are of importance in the formulation and implementation of industrial plans. For the construction of the indices of industrial production, the data about production of various industries are usually collected under the following heads:

(i) Textile Industries: Cotton, silk, woolen, etc.

(ii) Metallurgical Industries: Iron an steel, etc

(iii) Mining Industries: Coal, pig-iron and Ferro-alloys, petrol Kerosene, copper (virgin metal), etc.

(iv) Mechanical Industries: Locomotives, sewing machines, aero planes, etc.

(v) Industries subject to excise duty: Tea, sugar, cigarettes and tobacco, distilleries and breweries, etc.

(vi) Electricity, gas and steam; Electric lamps, electric fans, electrical apparatus and appliances, etc.

(vii) Miscellaneous: glass, paints ad varnish, paper and paper board, cement chemical etc.

Usually the data (figures of output) are obtained for various industries on monthly basis and the indices of industrial production are obtained as the weighted arithmetic mean (or sometimes geometric mean) of the production (quantity) relatives by the formula:

$$I_{oj} = \frac{\sum Q_j W_j}{\sum W_j}$$

Where Q_j = production relatives = q_{ij} / q_{oj} ,

And W_j is the weight may be assigned to jth them (industry).

The weights may be assigned to various industries on the basis of, say, capital invested, net output, production etc. The concept of 'value added by manufacture' is the most commonly used criterion for determining the weights to be assigned to different industries.

3.12 Exercises

(1) What is meant by consumer price index number?

(2) Explain the method of obtaining consumer price index number.

(3) What are the uses and limitations of a cost of living index number?

Calculate cost of living index number for the following data-

	Group	Weight	Index No.
i	Food	48	152
ii	Fuel and lighting	2	110
iii	Clothing	10	130
iv	House rent	12	100
v	Miscellaneous	15	80

(5) What is meant by base shifting of a series of index numbers?

(6) Write a note on the index of industrial production?

3.13 Summary

A consumer price index number measures the relative change in the amount of money required to produce equivalent satisfaction in two different situations. This index is constructed by comparing the consumer prices for two different time points. Cost of living index or CPI covers food, clothing, fuel, house rent and miscellaneous groups. It has been developed on the assumption that taste and habits of the group under consideration are same in two situations.

3.14 Further Readings

- Allen, R.G.D.; Index Numbers of Theory and Practice, MacMillan, 1975.
- Croxton, F.E. and Cowden, D.J.; Applied General Statistical, Prentice Hall, 1967.
- Greenwald, W.I., Statistics for Economics, C.E. Merit Books; 1963.
- Mills, F.C.; Statistical Methods, Henry Holt, 1955.

U.P. Rajarshi Tandon Open University, Prayagraj

UGSTAT – 104
Applied Statistics

Block: 2 Time Series Analysis

Unit – 4 : Time Series

Unit – 5 : Determination of Trend

Unit – 6 : Determination of Seasonal Indices

Course Design Committee

Dr. Ashutosh Gupta **Chairman**
Director, School of Sciences, U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anup Chaturvedi **Member**
Department of Statistics, University of Allahabad, Prayagraj

Prof. S. Lalitha **Member**
Department of Statistics, University of Allahabad, Prayagraj

Prof. Himanshu Pandey **Member**
Department of Statistics, D. D. U. Gorakhpur University, Gorakhpur.

Dr. Shruti **Member-Secretary**
School of Sciences, U.P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Prof. Ramji Tiwari **Writer**
Department of Statistics, University of Jammu, Jammu

Prof. B. P. Singh **Writer**
Department of Statistics, Banaras Hindu University, Varanasi

Dr. Alok Kumar **Writer**
Department of Statistics, Banaras Hindu University, Varanasi

Dr. D. C. Singh **Reviewer**
Department of Statistics, Harish Chandra P. G. College, Varanasi

Prof. Umesh Singh **Editor**
Department of Statistics, Banaras Hindu University, Varanasi

Dr. Shruti **Course/ SLM Coordinator**
School of Sciences, U. P. Rajarshi Tandon Open University, Prayagraj

UGSTAT – 104 **APPLIED STATISTICS**

First Edition: *March 2008* (Published with the support of the Distance Education Council, New Delhi)
Second Edition: ***July 2021***
ISBN :

Printed By : K.C. Printing & Allied Works, Panchwati, Mathura - 281003

Block & Units Introduction

The ***Block - 2 – Time Series Analysis*** is the second block with three units. the time series is a sequence of data values of some variables corresponding to successive point of time. In this block the analysis of time series data is discussed.

Unit – 4 – Time Series, deals with difference component of time series like trend, seasonal components, cyclical variation and random components.

Unit – 5 – Determination of Trend, deals with the various methods of determination of trend.

Unit – 6 – Determination of Seasonal Indices, dealt with various methods of determination of seasonal component of the time series data.

At the end of every block/unit the summary, self assessment questions and further readings are given.

Unit-1: Analysis of Time Series

Structure

4.1 Introduction

A time series refers to statistical data which are collected or observed over successive period of time. According to Croxtan and Cowdeu "A *time series consists of data arranged chronologically."* A time series data may be define as a sequence of values of some variable corresponding to successive points of time. In time Series data, the observation are not timeless as in the case of random variables occurring through some experiment, say, the yield of a particular crop in different plots of different blocks, but are real variables where observations are ordered with respect to successive point of time. Examples of time series are: yearly production of a particular crop in a particular region, quarterly inventories is an industry, monthly advances by banks, weekly sale by some store etc. Time series data may be categorized into period data and point data. Period data refers to the accumulated value of a flow variable during a particular period. For example rice production of 125 million tones during 1985-86. This data refers to total rice production during the period 01 April 1985 to 31 March 1986. In contrast we have point data, refers to the value of variable at particular point of time we shall now deal with such series in which the main problem is the analysis of the effects of different factors spread over a long period of time. We shall discuss the technique of the measurement of chronological variations. In the study of economic problems such series have a unique place of importance. In various type of time series data it is necessary to find out the changes in time, such studies which relates to analysis of series spread over a period of time come under analysis of time series. Although the term "time series" usually refers to economic data, and we too shall be concern here economic data, it equally applies to data arising in the natural and the other social sciences. In time series analysis, we analyses the past in order to understand the future better. One of the most important task before economists and businessman these days to make estimates for the future. For example a businessman is interested in finding out his likely sales in year 2009 or year 2015 so that he could adjust his production accordingly and avoid the possibility of either unsold sticks of inadequate production to meet the demand. In such situation time series analysis is very useful.

Symbolically, Y_t denotes the value of variable at time t (t = 1, 2… n) in case the figures relates to the successive periods and not points of time, t is to be taken as the mid-point of the t^{th} period.

4.2 Objectives

After going through this unit you should be able to understand:

- Concept of time series data, and their utility.
- Various components of time series
- Additive and multiplication models.

4.3 Utility of Time Series Analysis

The analysis of time series is useful not only to the economist and businessman but also to the scientists, geologists, sociologists, biologists, administrators, planners, research workers etc. for the following reasons:

1. It helps in understand past behavior. By collecting the data over a period of time one can easily understand what changes have taken place in the past. This analysis is extremely helpful in predicting the future behavior.
2. It helps in planning future operations. Plans for future actions cannot be made without forecasting event and relationship they will have. If the regularity of occurrence of any features over a sufficient long period could be clearly established than, within limits, prediction of probable future variations would become possible.
3. The review and evaluation of progress in any field in economics and business activity is largely done on the basis of the time series data collected in regard to a given phenomena. For example, the progress of plans is judged by the yearly rates of growth in gross national products.
4. It facilitates comparison. Once the data is systematically recorded, comparisons over time i.e. between one period and the order become easy. Thus exercise may go on regularly. Actually the time series analysis is a mean of more scientific comparison after considering the various components of the series to known how they have behaved over a period of time. Different time series are often compared and important conclusion drawn there form.

4.4 Component of a Time Series

A graphical representation of a time series will reveal the change over time. A series which exhibits no change during the period under consideration will give a horizontal line. However, usually we shall come across time series showing continual changes over time, giving us an overall impression of haphazard movement. A critical study of the series will, however reveal that the changes is not totally haphazard and a part of it, at least can be accounted for. The part which can be accounted for is the systematic part and the remaining part is the unsystematic or irregular.

Thus the various forces at work affecting the values of a phenomenon in a time series can be broadly classified into the following four categories commonly know as the components as:

i. Secular Trend
ii. Seasonal Variation
iii. Cyclical Variation
iv. Irregular Variation

In a given time series some or all of the above components may be resent.

4.4.1 Secular Trend

By secular trend (or simply, trend) of a time series we can mean the smooth regular, long-term movement of the series if observed long enough. Some series may exhibit and upwards or a downwards trend or may remain more or less at constant level. A gain some series after a period of growth (decline) may reserve their course and enter a period of decline (growth). But sudden or frequent changes are incompatible with the idea of trend.

4.4.2 Seasonal Variation

By seasonal fluctuations we mean a periodic movement in a time series where the period is not longer than one year. A periodic movement in a time series is one which recurs or repeats at regular intervals of times (or periods). Example of seasonal fluctuations may be found in the passenger traffic during the 24 hours of a day, sales of departmental store during the 12 month of a year, issue of library books during the seven days of a week, and so on. The factors which mainly cause this year, issue of library books during the seven days of a week and so on. The factors which mainly cause this type of variation in economic time series are the climate changes of the different variation in economic time series are the climate changes of the different seasons and the custom and habits which the people follow different times. For example, the occurrence of the festival in a particular month will increase the sale of certain consumer goods in the month.

The study and measurement of this component is of prime importance in certain cases. The efficient running of any department store, for example would necessitate a careful study of seasonal variation of demands of the goods.

4.4.3 Cyclical Fluctuation

By cyclical fluctuations we mean the oscillatory movement in a time series the period of oscillation being more than a year. One complete period is called cycle. The cyclical fluctuations are not necessarily periodic, since the length of the cycle as also the intensity of fluctuations may be change from one cycle to another in an irregular manner.

4.4.4 Irregular Fluctuations

Irregular fluctuations are those which are either wholly unaccountable or are caused by such unforeseen events as wars, floods, strikes, etc. This category of movements includes all times of variation that are accounted by secular trend, or season or cyclical fluctuations.

4.5 Mathematical Models for Time Series Analysis

Separation of the different components of a time series is of importance of time series is of importance, because it may be that we want to study the series after eliminating the effect of a particular component. It may be noted that it is the system part of the time series which may be used in forecasting.

In the classical or traditional approach, it is assumed that there is a multiplicative relation among the four components; that is, any particular value (Y_t) is considered to be the product of factors attributable to secular trend (T_t) seasonal components (S_t), cyclical component (C_t), and irregular (It), component. Thus the multiplicative model of time series can be expressed as

$$Y_t = T_t \times C_t \times I_t \qquad \dots\dots\dots\dots\dots (2.1)$$

Another approach is to assume Y_t to be the sum of the four components, i.e. time series is decomposed by additive hypothesis thus additive model of time series is given by

$$Y_t = T_t \times C_t \times I_t \qquad \dots\dots\dots\dots\dots (2.2)$$

This model is not generally used since it considered inappropriate for most economic data. However, if Y_t represent the logarithm of the original variable, then one may use this simpler, additive model instead of the multiplicative model i.e.

$$log\, Y_t = logT_t + logS_t + logC_t + logI_t \qquad \dots\dots\dots\dots\dots (2.3)$$

In practice, most of the series relating to economic data confirm to the multiplicative model (2.1)

4.6 Exercises

1. What is meant by a time series? How it is used for forecasting? Indicate its importance in business and economics.
2. Define a time series. Explain its uses in different fields. How it is useful to understand future better?
3. What is a time series? What are its components? Explain secular trend with one example.
4. What are the various component of a time series analysis? Explain these with examples.
5. Explain additive and multiplication models used in time series analysis. Why is the multiplicative model most commonly used as compared to additive model?

4.7 Summary

A time series a data collected and arranged chronologically. In this unit, it has been explained that the variation in time series data can be attributed to smooth long term changes, periodic (long or short term) changes or irregular changes. Thus we can describe four components of a time series namely trend, seasonal, cyclic and irregular (random). A time series may or may not have all the components. Mathematical model for time series has also been discussed.

4.8 Further Readings

- Corxton, F.E. and Cowden, D.J.: Applied General Statistics, Prentice Hall 1967.
- Goon A.N., Gupta M.K. & Das Gupta B (1987) Fundamentals of Statistics Vol. I, The World Press Pvt. Ltd., Kolkata.

Unit-5: Determination of Trend

Structure

5.1 Introduction

The analysis of time is important not only to businessmen and economists but also scientists, social scientist. The time series analysis as statistical tool is very useful. In this lessen we introduction the method which are generally used for the study of measurement of the trend component in a time series.

5.2 Objectives

After going through this unit you should be able to understand:

- The graphic method to measure trend.
- The method of semi averages to measure trend
- The method of moving average to measure trend

5.3 Determination of Trends

The following are the four methods which are generally used for the study of measurement of the trend component in a time series

- Graphic (or Free- hand curve Fitting) Method.
- Method of Semi average
- Method of curve fitting by the Principle of Least-es
- Method of Moving Averages.

5.4 Graphic (or Free Hand Curve Fitting) Method

This is the simplest and most flexible method of estimating the secular trend and consists in first obtaining the histogram and by plotting the time series values on a graph paper and then drawing a free hand smooth curve through these points so that it accurately reflects the long term tendency of the data. The smoothing of the curve eliminates the other components, viz seasonal, cyclical and random variations. In order to obtain the proper trend line or curve the following points may be born in mind:

(i) It should be smooth.

(ii) The number of points above the trend Curve/line should be more or less equals to the number of points below it.

(iii) The sum of vertical deviations of the given points above the trend should be approximately equals to the sum of vertical deviations of the points below the trend line so that the total negative deviations.

(iv) The sum of squares of vertical deviations of the given points from the trend line/ curve is minimum possible.

(v) If the cycles are present in the data then the trend line should be so drawn that:

(a) It has equals number of cycles above and below it.

(b) It bisects the cycles so that the areas of cycles above and below the trend line are approximately the same.

(vi) The minor short-term fluctuations or abrupt and sudden variations may be ignored

5.4.1 Merits

(i) It is very simple and time saving method and does not require and mathematical calculations.

(ii) It is a very flexible method in the sense that it can be used to describe all types of trend- linear as well as non- linear.

5.4.2 Demerits

(i) The strongest objections to this method is that it is highly subjective in nature. The trend curve so obtained will very much depend on personal bias and judgment of the investigator handing the data and consequently different person will obtain different trend curves for the same set of data. Thus, a proper and judicious use of this method requires great skill and expertise on the part of the investigator and this very much restricts the popularity and utility of this method. This method, though simple and flexible, is seldom used in practice because of the inherent bias of the investigator.

(ii) It does not help to measure trend.

(iii) Because of the subjective mature of the free hand trend curve, it will be dangerous to use it for forecasting or making predictions.

5.5 Method of Semi – Averages

As compared with the graphic method, this method has more objective approach. In this method, the whole time series data is classified into two equal parts with respect to time. For example, if we are given the time series values for 10 years from 1965 to 1974 then the two equal parts will be the data corresponding to periods 1965 to 1969 and 1970 to 1974. However, in case of the odd number of the years, the two equal parts are obtained on omitting the values for the middle period. Thus, for example, for the data for 9 years from 1970 to 1978, the value of the middle year, viz., 1974 being omitted. Having divided the given series into two equal parts, we next compute the arithmetic mean of the time series values for each half separately. These means are called semi-averages. Then these are called semi-averages. Then these semi averages are plotted as points against the middle point of the respective time periods covered by each part. The line joining these points gives the straight line trend fitting the given data.

As an illustration of the time series data for 1965- 1974 we have:

	Part- I	Part-II

Period:	1965 to 1969	1970 to 1974
Semi-Averages:	$\overline{x_1} = \frac{y_1 + y_2 + y_3 + y_4 + y_5}{5}$	$\overline{x_1} = \frac{y_1 + y_2 + y_3 + y_4 + y_5}{5}$
Middle of time period	1967	1972

$\overline{x_1}$ is plotted against 1967 and is plotted against 1972. The trend line is obtained on joining the points so obtained viz., the points (1967, $\overline{x_2}$) and (1972, $\overline{x_2}$) by a straight line. In the above case the two parts consists of an odd number of year viz., 5 and hence the middle time period is computed easily. However, if the two halves consist of even numbers of years as in the next case given above; viz., the year 1970 to 1973 and 1975 to 1978, the centering of average time period is slightly difficult. In this case XI (the mean of the values of the years 1970 to 1973) will be plotted against the mean of two middle periods of the period 1970 to 1973, viz., the mean of the years 1971 and 1972. Similarly $\overline{x_2}$ will be plotted against the men of the years 1976 and 1977.

5.5.1 Merits

(i) An obvious advantages of this method is its objectivity in the sense that it does' not depend on the personal judgment and everyone who uses this method gets the same trend line and hence the same trend values.

(ii) It is easy to understand and apply as compared with the moving averages or least square methods of measuring trend.

(iii) The line can be extended both ways to obtain future or past estimates.

5.5.2 Limitations

(i) This method assumes the presence of linear trends (in the time series values) which may not exist.

(ii) The use of arithmetic mean (for obtaining semi-average) may also questioned because of its limitations.

Accordingly, the trend values obtained by this methods and the predicted values for future are not precise and reliable.

Examples 2.1: Apply the method if semi-averages for determining trend of the following data and estimate the value for 1990:

Years	Sales (thousand units)	Years	Sales (thousand units)

1973	20	1996	30
1974	24	1997	28
1995	22	1998	32

`If the actual figure of sales for 1980 is 35,000 units how do you accounts for the difference between the figures you obtain and the actual figures given to you?

Solution: Here= 6, and hence the two parts will be 1993 to 1975 and 1976 to 1978.

Years	Sales (thousand units)	3- years Semi-Totals	Semi- Averages (A.M.)
1973	20		
1974	24	66	66/3=32
1995	22		
1996	30		
1997	28	90	90/3=30
1998	32		

Here the semi-average 22 is to be plotted against the mid year of the first part, i.e., 1994 and the semi-average 30 is to be plotted against the mid-year of the second part, viz., 1997

Now the trend values for 1994 is the average of first part, viz., 22 ('000 units) and for 1997 is 30 ('000 units). Hence using the fact that the yearly increment in sales is 2.667 ('000 units), the trend values for sale of various years can be obtained as shown below.

COMPUTING OF TREND VALUES

Year	Trend Values ('000 units)	Years	Trend Values ('000 units)
1993	22-2.667=19.333	1997	30
1994	22	1998	30+2.667=32.667
1995	22+2.667=24.667	1999	32.667+2.667=35.334
1976	24.667+2.667=27.334	1-98 & 1999-00	35.334+2.667=38.001

Thus the estimated (trend) value for sales in 1990 is 38,001 units. This trend value differs from the given values of 35,000 units because it has been obtained under the assumptions that there is linear relationship between the given time series values which in the case is not true. Moreover, in computing the trend value the effects of seasonal, cyclical and irregular variation have been completely ignored while the observed values are affected by these factors.

5.6 Method of Curve Fitting by the Principle of Least Squares

The principle of least squares provides us an analytical or mathematical device to obtain an objective fit to the trend of given time series. Most of the data relating to economic and business time series confirm to definite laws of growth or decay and accordingly in such a situation analytical trend fitting will be more reliable for forecasting and predictions. This technique can be used to fit linear as well as non- linear trends.

5.6.1 Fitting of Linear Trend

Let the straight line trend between the given time series values (y) and the time (t) are given by the equation:

$$Y= a+bt \qquad (1)$$

Then the given time 't' and the estimated value Ye of y as given by

$$Ye = a+bt \qquad (2)$$

The principle of least squares consist in estimating the values of a and b in (1) so that the sum of squares of errors of estimates

$$E = \sum (y - y_e)^2 = \sum (y - a - bt)^2 \qquad (3)$$

Is minimum the summation being taken over given values of the time series. This will be so if

$$\frac{\partial E}{\partial a} = 0 = -2\sum (y - a - bt)$$

$$\frac{\partial E}{\partial a} = 0 = -2\sum t(y - a - bt)$$

Which on simplification gives the normal equations or least square equations for estimating a and b as

$$\sum y = na + b\sum t \qquad (4)$$

$$\sum ty = a\sum t + b\sum t^2 \qquad (5)$$

Where n is the number of the time series pair (t,y). it may be seen that equation (4) is obtained on taking of both sides in equation (1). Equation (5) is obtained on multiplying equation (1) by t and then summing both sides over the given values of the series.

Solving (4) and (5) for a and b and substituting these values in (1), we finally get the equation of the straight line trend.

5.6.2 Fitting a Second Degree (Parabolic) Trend

Let the second degree parabolic trend be given by the equations:

$$Y = a + bt + ct\ 2 \qquad (6)$$

Then for any given value of t, the trend value is given by:

$$Y_e = a + bt + ct^2 \qquad (7)$$

Thus if ye is trend value corresponding to an observed value y, then according to the principle of least squares we have to obtain the values of a, b and c in (6) so that

$$E = \sum (y - y_e)^2 = \sum (y - a - bt - ct)^2$$

is minimum for variations in a, band c. Thus the normal or least square equations for estimating a, b and c are given by:

$$\frac{\partial E}{\partial a} = 0 = -2\sum (y - a - bt - ct^2)$$

$$\frac{\partial E}{\partial a} = 0 = -2\sum t(y - a - bt - ct^2)$$

$$\frac{\partial E}{\partial a} = 0 = -2\sum t^2\ (y - a - bt - ct^2)$$

The summation being taken over the values of the time series. On simplifying and transposing we get the normal equations as:

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

$$\sum ty = n\sum t + b\sum t^2 + c\sum t^2$$

$$\sum t^2 y = n\sum t^2 + b\sum t^3 + c\sum t^4 \qquad (7)$$

The first equation in (7) is obtained on summing both sides of (6). The second equation is obtained on multiplying (6) with t, [the coefficient of the second constant bin (6)] and then summing both sides. The third equation is obtained on multiplying both sides of (6) with t2 [the coefficient of c, the third constant in (6)] and then summing over values of the series.

For given times series $\sum y$, $\sum ty$, c $\sum t^2 y$, c $\sum t$, $\sum t^2$, $\sum t^3$, and $\sum t^4$,c can be calculated and equations (7) can be solved for a, b and c. with these values of a, b and c the parabolic curve (6) is the trend curve best fit.

Remark: **Change of Origin:** Usually the values of are for different years, say 1970, 1971,….1979 and thus computation of $\sum t$, $\sum t^2$, $\sum t^3$, etc. and hence the solution of equation (4) and (5) for linear trend or equations (7) for parabolic trend is quite tedious and time consuming. However, it may be remarked that the time variable t in the time series has no magnitudinal value but it ha only positional or locational importance. Hence we can shift the origin in the time variable according to our convention and assign it the time variable according to our convention and assign it the consecutive values 0, 1, 2, 3,….etc, the time period allotted the value 0 is known as the period of origin. This might slightly facilitate the solution of normal equations. However the algebraic computation can be simplified to a great extent by shifting the origin in time variable t to a new variable x in such a way that we always get.

$\sum x = \sum x^3$ = 0.The technique is explained below and can be applied only if the values of t are given to be equidistant, say at an interval h.

If n, the number of time series values is odd, then the transformation is:

$$x = \frac{t - middle\ value}{Interval\ (h)} \qquad (8)$$

Thus, if we are given yearly figures for say, 1970, 1971, 1972,….1976, i.e., n = 7 then

$$x = \frac{t - middle\ value}{Interval\ (h)} = t - 1973$$

Putting t= 1970, 1972,….1976 in (*) we get x=-3, -2, -1, 0, 1, 2, and 3 respectively so that $\sum x = \sum x^3$ = 0.

If n is even then the transformation is :

$$x = \frac{t - (Arithmetic\ mean\ of\ two\ middle\ values)}{\frac{1}{2} Interval\ (h)} \qquad (9)$$

Thus if we are given the yearly for, say, 1965, 1966, 1967,……………1972 then

$$x = \frac{t - \frac{1}{2}(1968 + 1969)}{\frac{1}{2}} = 2t - 3937$$

Putting t= 1965, 1966,…., 1972 in (**) we get respectively:

x=-7, -5, -3, -1, 1,3,5,7 so that $\sum x = \sum x^3$ = 0

The transformations (*) or (**) will always give $\sum x = \sum x^3$ = 0 and this reduces the algebraic calculations for the solution of normal equations to be great extent. For example for the linear trend

$$Y = a+bx \qquad (10)$$

Where x is determined either by (8) or (9) according as n is odd or even the normal equations for estimating a and b becomes:

$$\sum y = na + b\sum x\,y + a\sum x + b\sum x^2$$

But $\sum x = 0$. Hence these equations give:

$$\sum y = na \; and \; \sum x\,y + b\sum x^3$$

$$\Rightarrow a = \frac{\sum y}{n} \quad and \quad b = \frac{xy}{\sum x^2} \qquad (11)$$

With these values of b and b, (10) gives the equation of the trend line.

Similarly, for the parabolic trend

$$Y = a+bx+cx^2, \qquad (12)$$

The normal equations for estimating a, b and care

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

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

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

But $\sum x = 0 \; \sum x^3$ Hence these reduces to:

$$(i) \sum y = na + c \sum x^2$$

$$(ii) \sum xy = b \sum x^2$$

$$(iii) \sum x^2 y = a \sum x^2 + c \sum x^4$$

Equations (ii) gives the value of $b = \frac{xy}{\sum x^2}$ and equation (i) and (ii) can be solved simultaneously for a and c. With these values of a, b and c the curve (12) becomes the parabolic trend curve of best fit.

5.6.3 Fitting of Exponential Trend

The exponential curve is given by the equation:

$Y=ab^t$ (13)

Taking logarithm of both sides we get

log y = log a + t log b (14)

Y = A + Bt

Where Y = log y; A = log a and B = log b (15)

(14) is a straight line trend between Y and t. Hence the normal equations for estimating A and B are

$$\sum y = nA + B \sum t$$

$$\sum ty = A \sum t + B \sum t^2$$

These equations can be solved for A and B and we finally get on using (15):

a = Antilog (A) and b = Antilog (B) (16)

With these values of a and b, the curve (13) becomes best exponential trend values.

Remark: As already explained in the fitting of linear and parabolic trend, we can change the origin in t to new variable x such that $\sum x = 0$ and then considering the trend curve $y= ab^x$, the calculation can be reduced to a great extent.

5.6.4 Second Degree Curve Fitted to Logarithms

Suppose the trend curve is

$$y = ab^t c^{t^2} \qquad (17)$$

Taking logarithm of both sides we get:

$$\log y = \log a + t \log b + t^2 \log c$$

$$\Rightarrow \quad Y = A + Bt + Ct^2 \qquad (18)$$

Where Y = log y; A = log a, B= log b , C= log c (19)

Now (18) is a second degree parabolic curve in Y and t and can be fitted by the technique already explained, we can finally obtain:

a=antilog (A); b = Antilog (B) and c= Antilog (C)

with the values of a band c the curve (17) becomes the best degree curve fitted to logarithms.

5.6.5 Merits

The method of least squares is the most popular and widely used method of fitting mathematical function to a given set of observations. It has the following advantages.

(i) Because of its analytical and mathematical character, this method completely eliminates the element of subjective judgment or personal biased on the part of the investigators.

(ii) Unlike the method of moving averages, this method enables us to compute the trend values for all the given time period in the series.

(iii) 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 period of the series and the forecasted values are also quite reliable.

(iv) 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. If we fit the linear trend y = a + bx, where x is obtained from t by change of origin such that $\sum x = 0$ then for the yearly data, the annual rate of growth is b or 2b according the number of year is odd or even respectively.

5.6.6 Demerits

(i) The most serious limitation of the method is the determination of the type of the trend of the curve to be fitted, that is whether we should fit a linear or parabolic trend or some other more complicated trend curve. Assumptions about the type of trend to be fitted might introduce some bias.

(ii) The addition of even a single new observation necessitates all the calculations' to be a fresh which is not so in the case of moving average method.

(iii) This method requires more calculations and is quite tedious and time consuming as compared with other methods. It is rather difficult for a nonmathematical person to understand and use.

(iv) 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.

(v) It can not be fit growth curves (Modified exponential curve, Gompertz curve and Logistic curve) to which most of the economic and business time series conform.

5.6.7 Examples

Example 2.2: Fit a trend line to the following data by the least squares method.

Year :	1975	1977	1979	1981	1983
Production (in '000 tons) :	18	21	23	27	16

Estimate the production in 1980 and 1985

Solution: Let the trend line given by the equation:

$y = a + bx$ (*)

Where x= t-1979, i.e., origin is at 1979 and x units = 1 year and y is production (in '000 tons)

COMPUTATION FOR STRATIGHT LINE TREND

Year (t)	Production (in '000 tons) (y)	X= t- 1979	xy	Xt	Trend Values ('000 tons) Ye=21+O.lx
1975	18	-4	-72	16	21-0.4=20.6
1977	21	-2	-42	4	21.02=20.8
1979	23	0	0	0	21.0

1981	27	2	54	4	21+0.2=21.2
1983	16	4	64	16	21+0.4=21.4
Total	$\sum y = 105$	$\sum x = 0$	$\sum xy = 4$	$\sum x^2 = 40$	

The normal equation for estimation a and b in (*) are given by:

$$\sum y = na + b\sum x\,, \sum xy = a\sum x + b\sum x^2$$

105 = 5a + 0 , 4= a* 0+40b

$$a = \frac{105}{5} = 21, \qquad b = \frac{4}{40} = 0.1$$

Substituting in (*), the trend line is given by the equation:

Ye = 21+0.1x (**)

Substituting x= -4,-2, 0, 2, 4 in (**), we obtain the trend values for the years 1975 to 1983 respectively. the trend values are given in the last column of the above table.

The estimated production is 1980 is obtained on taking

X= t-1979 = 1989- 1979=1. in (**). Thus

(y_e) 1980= 2+0.1*1=21+0.6=21.1 ('000tons)

The estimated production in 1985 is obtained on taking in (**). Thus

(Ye) 1985= 2+0.1*6=21+0.6= 21.6 ('000 tons).

Example 2.3: Given below are, given the figures of the production (in '000tons) of a sugar factory:-

Year	1969	1970	1971	1972	1973	1974	1975
Production	77	88	94	85	91	98	90

(i) Fit a straight line by the method of "least squares" and show the trend values

(ii) What is the monthly increase in production

(iii) Eliminate the trend.

Solution:

COMPUTION OF THE STRATIGHT LINE TREND

Year (t)	Production (in '000 tons) (y)	X= t-1979	xy	x^2	Trend Values ('000 tons) Ye=21+O.lx
1969	77	-3	-231	9	83
1970	88	-3	-176	4	85
1971	94	-1	-94	1	87
1972	85	0	0	0	89
1973	91	1	91	1	91
1974	98	2	196	4	93
1975	90	3	270	9	95
Total	$\sum y = 623$	$\sum x = 0$	$\sum xy = 56$	$\sum x^2 = 28$	$\sum y_e = 623$

Let the straight line trend is given by the:

y= a + bx (*)

Where the origin in July 1972 and x unit = 1 year. The normal equations for estimating a and b in (*) are:

$$\sum y = na + b\sum x \, , \sum xy = a \sum x + b \sum x^2$$

$$a = \frac{\sum y}{n} \, , \qquad b = \frac{\sum xy}{\sum x^2} \quad [because \sum x = 0]$$

$$a = \frac{623}{7} = 89, \qquad b = \frac{56}{28} = 2$$

Hence the straight line trend is given by the equation:

Y= 89+2x (**)

Putting x= -3 -2, -1, 0,1,2,3 in (**) we get the trend values for the years 1969 to 1975 respectively and are shown in the last column of the above table. It may be checked that

$\sum y = \sum y_e$ as required by the principle of least squares.

(ii) From (*) it is obvious that the values increase by a constant amount 'b' units every. Thus the yearly increase in production is 'b' units, i.e. 2* 1000= 2000 tons. Hence the monthly increase in production = 2000/12= 166.67 tons. Assuming multiplicative model, the trend values are eliminated on dividing the given values (y) by the trend values (y_e). However if we assume the additive model, the trend eliminated values are given by (y-ye). The resulting values contains short-term (cyclic) variations and irregular variations. Since the data are annual the seasonal variations are absent.

ELEMINANTION OF TREND

Year	Trend eliminated values based on	
	Additive Model (y-ye)	Multiplication Model
1969	77-83=-6	77/83=0.93
1970	88-85=3	88/85=1.04
1971	94-87=7	94/87=1.08
1972	85-89=-4	85/89=0.96
1973	91-91=0	91/91=1.00
1974	98-93=5	98/93=1.05
1975	90-95=-5	90/95=0.95

Example 2.4: Fit an equation of the form Y= A+bt+Ct 2 to the data given below:

X:	1	2	3	4	5
Y:	25	28	33	39	46

Solution: Here n, the number of pairs is odd. Hence we take

t= X-(middle value)= X-3, (*)

so that values of t corresponding to X= 1,2,3,4 and 5 are -2, -,0,1, and 2 respectively. Let the second degree trend equation between Y and t be:

Y= A+Bt+Ct^2
Where t= X-3 (**)

CALCULATION FOR SECOND DEGREE TREND

X	Y	T=X-3	t^2	t^3	t^4	tY	t^2Y
1	25	-2	4	-8	16	-50	100
2	28	-1	1	-1	1	-28	28
3	33	0	0	0	0	0	0
4	39	1	1	1	1	39	39
5	46	2	4	8	16	92	184
Total	$\sum y = 171$	$\sum t = 0$	$\sum t^2 = 10$	$\sum t^3 = 0$	$\sum t^4 = 34$	$\sum ty = 53$	$\sum t^2 y = 35$

The normal equation for estimation A, B and C in (**) are:

$$\sum y = nA + B\sum t + C\sum t^2$$

$$\sum ty = A\sum t + B\sum t^2 + C\sum t^3$$

$$\sum t^2y = A\sum t + B\sum t^3 + C\sum t^4$$

⇒ 171 = 5A+10C (i)

53-10B (ii)

351= log A+34C (iii)

(ii) ⇒ B= 53/10= 5.3

Multiplying (i) by 2 and then subtracting from (iii) we get:

351-2*171= (10A+34C)-(10A+20C)

⇒ 14C = 351 – 342 = 9

⇒ C = 9/14 = 064

Substituting in (i) we get

A = (171-10C)/5 = (171-6.4)/5 = 164.6/5= 32.92

Substituting the values of A, B and C in (**) we get the trend equation as:

$Y= 32.92+5.3t + 0.64t^2$ (iv)

Where t= X-3.

Hence the second degree trend equation of Y and X becomes:

Y= 32.92+5.3 (X-3) + 0.64 (X-3)

$= 32.92+5.3X-15.9+0.64(X^2-6X+9)$

$= (32.92+15.90+5.76)+(5.30-3.48)X+0.64X^2$

$= 22.78+1.46X+0.64X^2$

Remark: The trend values of Y for X = 1,2,3,4 and 5 can be computed as given below

Computation of Trend Values

X	Y	t=X-3	t^2	5.3t	0.64 t^2	Trend values Ye= 32.92+53t+0.64 t^2
1	25	-2	4	-10.6	2.56	32.92-10.6+2.56=24.88
2	28	-1	1	-5.3	0.64	32.92-5.3+0.64=8.26
3	33	0	0	0	0	32.92-0+0=32.92
4	39	1	1	5.3	0.64	32.92+5.3+0.64=38.86
5	46	2	4	10.6	2.56	32.92+10.6+2.56=46.08

If we compare the original value (Y) and the corresponding trend values (Ye), we observe that they are very close. Hence, we may conclude that the parabolic trend (iv) is a very good fit to the given data.

5.7 Methods of Moving Averages

Method of moving averages is a very simple and flexible method of measuring trend. It consists in obtaining a series of moving averages (arithmetic means) of the successive overlapping groups or section of the time series. The averaging process smoothens out fluctuations and the ups and downs in the given data. The moving average is characterized by a constant known as the period or extent of moving averages. Thus, the moving average of period "m" is a series of successive averages (A.M's) of m overlapping values at a time, starting with 1^{st}, 2^{nd} , 3^{rd} value and so on. Thus for the time series values y_1, y_2, y_3,………..for different time periods the moving average (M.A.) values of period 'm' are given by:

$$1^{st}\,M.A.=\frac{1}{m}(y_1+y_2+\cdots\ldots..+y_m)$$

$$2^{nd}\,M.A.=\frac{1}{m}(y_2+y_3+\cdots\ldots..+y_{m+1})$$

$$3^{rd}\,M.A.=\frac{1}{m}(y_3+y_4+\cdots\ldots..+y_{m+2})$$

and so on

We shall discusses tow cases

Case (i) ***When Period is odd:*** If the period 'm' of the moving average is odd, then the successive values of the moving averages are placed against the middle values of the corresponding the corresponding time intervals. For example, if m= 5, the first moving average value is place against the middle period, i.e., 3rd, the second M.A. value is placed against the time period 4 and so on.

Case (ii) ***When Period is even:*** If the period 'm' of the moving averages even then there are two middle periods and the M.A. value are placed in between the two middle periods of the time intervals it covers. Obviously, in this case, the M.A. value will not coincide with a period of the given time series and an attempt is made to synchronies them with the original data by taking two-period average of moving averages and placing them in between the corresponding time periods. This technique is called centering and the corresponding moving average values are called centered moving averages. In particular if the period m=4, the first moving average value is placed against the middle of 2nd and 3rd time intervals; the second average value is placed in between 3rd and 4th time period and so on. These are given by

$$\overline{y_1} = \frac{1}{4}(y_1 + y_2 + y_3 + y_4),$$

$$\overline{y_2} = \frac{1}{4}(y_2 + y_3 + y_4 + y_5),$$

$$\overline{y_3} = \frac{1}{4}(y_3 + y_4 + y_5 + y_6),$$

And so on. The centered moving averages are obtained on taking 2-period M.A. of and so on. Thus,

First Centered M.A.$= \frac{1}{2}(\overline{y_1} + \overline{y_2})$

$$= \frac{1}{2}\left[\frac{1}{4}(y_1 + y_2 + y_3 + y_4) + \frac{1}{4}(y_2 + y_3 + y_4 + y_5)\right]$$

$$= \frac{1}{8}[(y_1 + y_2 + y_3 + y_4) + (y_2 + y_3 + y_4 + y_5)]$$

$$= \frac{1}{8}[(y_1 + y_2 + y_3 + y_4 + y_5)]$$

Similarly, are placed against the time periods 3,4,5,…..and so on

Second M.A. $= \frac{1}{8}[(y_2 + y_3 + y_4 + y_5 + y_6)]$

And so on. These centered moving average are placed against the time periods 3,4,5, …. and so on.

Equation (1) may be regarded as a weighted of y_1, y_2, y_3, y_4, y_5 the corresponding weights being 1,2,2,2,1, i.e.,

$$\bar{Y} = \frac{w_1 y_1 + w_2 y_2 + w_3 y_3 + w_4 y_4 + w_5 y_5}{w_1 + w_2 + w_3 + w_4 + w_5} = \frac{\sum wy}{\sum w},$$

$$where \quad w_1 = w_2 = 1 \; and \; w_2 = w_3 = w_4 = 2.$$

Similar interpretation can be given to (2)

From (1) and (2) we that a centered moving average of period 4 is equivalent to weighted moving average of period 5, the corresponding weighted being 1,2,2,2,1.

The moving average values plotted time period give the trend curve. The basic problem in M.A. method is the determination of period ‘m’ is discussed in the remark below:

Remark:

Moving Average and Linear Trend: If the time series data does not contain any movement except the trend which when plotted on a graph gives a straight line curve, then the moving average will produce the series. The following example will clarify the point.

Example 2.5

Year (t)	Values	3-Yearly M.A.	5-Yearly M.A.	7-Yearly M.A.
1	10		-	-
2	14	14	-	-
3	18	18	18	-
4	22	22	22	22
5	26	26	26	26
6	30	30	30	30
7	34	34	34	34
8	38	38	38	38
9	42	42	42	-
10	46	46	-	-
11	50	-	-	-

Thus the trend values by the moving average of extent 3,5,7 and so on coincide with the original series.

(ii) Moving Averaging and Curvilinear Trend: If the data does nor contain any oscillatory or irregular movements and has only general trend and the histogram (graph) of the time series gives a curve which is convex (concave) to the base, then the trend values computed by moving average method will give another curve parallel to the given curve but above (below) it. In other words, if there are no variations in the data except the trend which is curvilinear then the moving average values, when plotted, will exhibit the same curvilinear pattern but slightly away from the given historigram. Further greater the period of the moving, the farther will be trend curve from the original histogram. In other words, the difference between the trend values and the original values becomes larger as the period of moving average increases.

(iii) ***Period of Moving Average:*** The moving average will completely eliminate the oscillatory movements if:

(i) The period of the moving average is equal to or a multiple of the period of oscillatory movement provided they are regular in period or amplitude and

(ii) The trend is linear or approximately so.

Hence to compute correct trend values by the method of moving averages, the period or. Extent of the moving average should be same as the period of the cyclic movement in the series. However, if the period of moving average is less or more than the period of the cycle movement then it (M.A.) will only reduce their effect.

Quite often we come across time series data which does not exhibit regular cyclic movement and might reflect cycles with varying periods which may be determined on drawing the historigram of the given time series and observing the time distance between various peaks.

In such a situation, the period of the moving average is taken as the average period of the various cycles present in the data.

(iv) ***Moving Average and Polynomial Trend:*** In most of the economic and business time series the trend is rarely linear and accordingly, if the trend is curvilinear, the moving average values will give the distorted picture of the trend. In such a case the correct trend values are obtained by taking a weighted moving average of the given values. Thus in case of polynomial trend (2^{nd} degree parabolic, cubic or higher degree) the moving average is characterized by two constants [m, p], m being the extent or period of the moving average and p, the degree of the polynomial trend to be fitted. The choice of the period of moving average has already been discussed in the remark 3, The degree of polynomial trend to be fitted is decided by variate difference method. The weights to be used will depend upon the period odd M.A. and the degree of polynomial trend to be fitted. The determination of the weights of the moving average for the techniques and is thus not discussed here. For example, the weights of moving average [5,2], i.e. a moving average of extent 5 for a parabolic trend is given by:

$$-\left(\frac{3}{35},\frac{12}{35},\frac{17}{35},-\frac{3}{35}\right)$$

Thus the first moving average value for the series $y_1, y_2, y_3, \ldots..$ is given by

$$\frac{1}{35}(-3y_1 + 12y_2 + 17y_3 - 3y_5).$$

The weights for the moving average [7,2,], i.e. a M.A. of the period 7 for parabolic trend are:

$$\left(-\frac{2}{21},\frac{3}{21},\frac{6}{21},\frac{7}{21},-\frac{6}{21},\frac{3}{21},-\frac{2}{21}\right),$$

And the first trend value is given by:

$$\frac{1}{21}(-2y_1 + 3y_2 + 6y_3 + 7y_4 + 6y_5 + 3y_6 - 2y_7).$$

It may be observed that:

(i) The weights of the M.A. are symmetric about the middle values and

(ii) The sum of weights in unity.

(v) ***Effect of Moving Average on Irregular Fluctuations:*** The moving average smoothens the ups and downs present in the original data and therefore reduces the intensity or irregular fluctuations to some extent. It can't eliminate. Them completely. However greater the period of moving average, the greater is the amount of reduction in their intensity. Thus from point of view of reducing irregular variations, long period moving average is recommended. However we have pointed out in Remark 2, that greater the period of moving average, farther are the trend values from original values. In order words longer period of moving average is likely to give distorted picture of trend values. Accordingly, as a compromise the period of moving average should neither be too large nor too small. The optimum period of the moving average is the one that coincides with or is multiple of the period of cycles in the time series as it would completely eliminate cyclical variation, reduce the irregular variations and therefore gives the best possible values of the trend.

5.7.1 Merits and Demerits of Moving Average Method

Merits

i) This method does not require any mathematical complexities and is quite simple to understand and use as compared with squares method.

ii) Unlike the ‘Free hand curve’ method this method does not involve any methods of subjectivity since the choice of period of the moving average is determined by the oscillatory movements in the data and not by personal judgment of investigator.

iii) Unlike the method of trend fitting by the principle of least squares the moving average method is quite flexible in the sense that few more observations may be added to the given data without affecting the trend vales already obtained. The addition of some new observations will simply result in some more trend values at the end.

iv) The oscillatory movement can be completely eliminated by choosing the period of the M.A. equals to or multiple of the period of cycle movement in the given series. A proper choice of the period also reduces the irregular fluctuations to some extent.

v) In additions to the measurement of the trend, the method of moving average is also used for measurement of cyclical, seasonal and irregular fluctuations.

Limitations

i) An obvious limitation of the moving average in that we can not obtain the trend values for all the given observations. We have to forego the tend values for some observations at both the extremes depending on the period of moving average. For example for moving average of period 5,7,and 9 we loose the trend value for the first and last 2,3, and 4 values respectively.

ii) Since the trend values obtained moving average method can not be expressed by any fluctuation relationship, this method cannot be used for forecasting or prediction future values which is the main objective of trend analysis.

iii) The selection of the period of moving average is vey important and is not easy to determine particularly when the time series does not exhibit cycles which are regular in period and amplitude. In such a case the moving average will not completely eliminate the oscillatory movements and consequently the moving average values will not represent the true picture of the general trend.

iv) In case of non-linear trend, which is generally the case in most of the business and economic time series the trend values given by the moving average methods are biased and they lie either above or below the true sweep of the data.

Keeping in view the limitations, the moving average method is recommended under the following situations:

(i) If the is linear or approximately so,

(ii) The trend oscillatory movement describing the given time series are regular both in period and amplitude.

(iii) If forecasting is not required.

5.7.2 Examples

Example 2.6: Calculate three yearly moving average for the following data.

Year	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960
Y:	242	250	252	249	253	255	251	257	260	265	262

Solution:

Computation of 3- Yearly Moving Average

Year (1)	Year (2)	3-yearly moving totals (3)	3-yearly moving averages (Trend values) (4)=(3)/3
1950	242	-	-
1951	250	744	248.0
1952	252	751	250.3
1953	249	754	251.3
1954	253	757	252.3
1955	255	759	253.0
1956	251	763	254.3
1957	257	768	256.0
1958	260	782	260.7
1959	265	787	262.3
1960	262	-	-

Example 2.7: Choose an appropriate period of the moving average for the following data and calculate moving average for that period:

Year	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Values	130	127	124	135	140	132	129	127	145	158	153	146	145	164	170

Solution: Since the peaks of the given data occur at the years 1,5,10 and 15; the data clearly exhibits a regular cyclic movement with period 5. Hence the period of the moving average for determining the trend values is also 4; viz., the period of cyclic variations.

Computation of Five Yearly Averages

Year (1)	Year (2)	5-yearly moving totals (3)	5-yearly moving averages (Trend values) (4)=(3)/5
1	130	-	-
2	127	-	-
3	124	656	131.2
4	135	658	131.6
5	140	660	132.0
6	132	663	132.6
7	129	673	134.6
8	127	691	138.2
9	145	712	142.8
10	158	729	145.8
11	153	747	149.4
12	146	766	153.2
13	145	778	155.6
14	164	-	-
15	170	-	-

5.8 Summary

In this lesson we introduce to the students the basic required for understanding the concept of Determination of trends. The basic has been explained with the help of examples.

5.9 Self Assessment Exercises

P-1) Determine the trend of the following data by method of semi averages:

Year:	1995	1996	1997	1998	1999	2000
Sales (Lakhs Rs.):	38.5	45.0	44.0	43.0	47.0	51.0

P-2) Calculate trend values by 3- yearly moving average and 4 yearly moving average methods:

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Sales (000Rs.)	5	7	9	12	11	10	8	12	13	17

P-3) Fit a straight line trend to the following data by method of least spaces:

Year	1990	1991	1992	1993	1994	1995	1996
Production (‘000tons)	80	90	92	83	94	99	92

P-4) Fit a parabola of the second degree to the following data.

Year	1981	1982	1983	1984	1985	1986
Y	144	148	154	172	186	252

Also estimate the value of y for 1998.

5.10 Answers

P-1) Annual increment = 1.5

Trend values: 41.0 42.5 44.0 45.5 47.0 48.5

P-2) Trend values by 3- yearly moving average method:

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Trend Values	-	7.00	9.33	10.67	11.0	9.67	10.00	11.0	14.00	-

Trend values by 4- yearly moving average method:

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Trend Values	-	-	9.00	10.13	10.38	10.25	10.50	11.63	-	-

P-3)

Year	1984	1986	1988	1990	1992	1994	1996
Trend Value	84	86	88	90	92	94	96

P-4)

Year	1981	1982	1983	1984	1985	1986
Y	148.4	143.1	150.1	169.3	200.7	244.4

Estimate the value of y for 1988 is 363.3

5.11 Further Readings

- Corxton, F.E. and Cowden, D.J.: Applied General Statistics, Prentice Hall 1967.
- Goon A.N., Gupta M.K. & Das Gupta B (1987) Fundamentals of Statistics Vol. I The World Press Pvt. Ltd., Kolkata.

Unit-6: Determination of Seasonal Indices

Structure

6.1 Introduction

To introduce the concept of the seasonal variation in a time series data are those short term variations contained with in a year that recur regularly. They are periodic in nature and the periods may be days, weeks, months quarters or the seasonal like summer, rainy, winter etc. In this unit we introduce the commonly used method of isolating seasonal variations.

6.2 Objectives

After studying this unit you should be able to:

- Understand the concept of measurement of seasonal indices
- Use the method of simple Averages to measures seasonal indices
- Use the ratio to trend method to measure seasonal indices
- Use the ratio to moving average method to measure seasonal indices

- Use the method of link relatives to measure seasonal indices

6.3 Measurement of Seasonal Indices

A seasonal variation in time series means the variations due to such forces which operate in regular periodic manner with period not longer than one year. The study of such variations which are predominantly exhibited by most of the economics and business man or sales, manager for planning paramount importance to a businessman or sales manager for planning future production and in scheduling purchase, inventory control, personal requirements and selling and advertising programmes. The objectives for studying seasonal patterns in a time series may be classified as follows:

i) To isolate the seasonal variations, i.e. to determine the effect of seasonal swings on the value of given phenomenon, and

ii) To eliminate them i.e. to determine the value of the phenomenon if there were no seasonal ups and downs I the series. This is called de-seasonalising the given data and is necessary for the study of cyclic variations.

Obviously for the study of seasonal variations, the time series data must be given for 'part' of a year, viz,. monthly, weekly, quarterly, daily or hourly. The study of seasonal pattern is superimposed on the values of a given series independently in the sense that a particular month (for a monthly data), or quarter (for quarterly data) will always exert a particular effect on the values of the series. Seasonal variations are measured as relative effective effects in terms of ratios or percentages, assuming multiplicative model and occasionally absolute changes assuming additive model or time series. The following are the different methods of measuring seasonal variations:

6.4 Method of Simple Averages

This is a simplest method of measuring seasonal variation in a time series. It is based on the assumption that the series contains neither a trend nor cyclical variations but only seasonal and irregular variations and it involves the following steps: (we shall explain the steps for monthly data. They can be modified accordingly for quarterly, weekly or daily data.)

(1) Arrange the data by years and months.

(2) Compute the average (Arithmetic Mean) $\bar{x}_i$ for i^{th} month, i= 1,2,3,….12. Thus $\bar{x}_1, \bar{x}_2, \dots \dots . \bar{x}_{12}$ are the average values for January, February, December respectively, the average being taken over different years, say k in number.

The irregular variation may be eliminated by averaging.

(3) Obtain the overall average $\bar{x}$ of these averages obtained in step (2). This is given by

$$\bar{x} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots \ldots . + \bar{x}_{12}}{12}$$

(4) Seasonal indices for different months are obtained on expressing each monthly average as a percentage of the overall average, X i.e.

$$Seasonal\ index\ for\ any\ month = \frac{monthly\ average}{\bar{x}} \times 100$$

$$Thus\ Seasonal\ Index\ for\ January = \frac{\bar{x}_1}{\bar{x}} \times 100$$

$$Seasonal\ Index\ for\ Fabruary = \frac{\bar{x}_2}{\bar{x}} \times 100.$$

.

.

$$Seasonal\ Index\ for\ December = \frac{\bar{x}_{12}}{\bar{x}} \times 100$$

Remarks: (i) if we are given quarterly for different years then we compute average value $\bar{x}_i (i = 1,2,3,4)$ for each quarters over different years and then

$$\bar{x} = \frac{1}{4}(\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \bar{x}_4)$$

Finally seasonal index numbers for different quarters are given by the formula:

$$Seasonal\ Index\ for\ quarter = \frac{\bar{x}_{12}}{\bar{x}} \times 100, \qquad i = 1,2, \ldots . ,4.$$

(ii) The sum of seasonal index must be 1200 for monthly data and 400 for quarterly data.

(iii) From computational point of view, a somewhat convenient formula for computing the seasonal index is obtained on substituting the value of $\bar{x}$ in

Thus we get

$$\boldsymbol{Seasonal\ Index\ for\ any\ month} = \frac{monthly\ average}{\frac{\bar{x}_1 + \bar{x}_2 + \cdots \ldots . + \bar{x}_{12}}{12}} \times 100$$

$$= Monthly\ Average \times \frac{1200}{sum\ of\ monthly\ averges}$$

Similarly we shall have

$$Seasonal\ Index\ for\ quarter = Quarterly\ Average \times \frac{400}{sum\ of quarterly\ averges}$$

A more simplified formula is as follows:

It T_i is the total of i^{th} seasonal (i = 1, 2, 3,….12 for monthly data), over the given k different years then

$$\bar{x}_i = \frac{T_i}{k}, (i = 1,2, \dots .12)$$

$$and\ \bar{x} = \frac{1}{12}\sum \bar{x}_i = \frac{1}{12}\sum_i \left(\frac{T_i}{k}\right)$$

$$= \frac{1}{12k}\sum_i T_i = \frac{1}{12}\bar{T}$$

$$Seasonal\ index\ for\ ith\ Season = \frac{\bar{x}_i}{\bar{x}} \times 100$$

$$= \left(\frac{\frac{T_i}{k}}{\frac{T}{12k}}\right) \times 100 = \left(\frac{T_i}{T}\right) \times 1200, where\ T = \sum_i T_i$$

So instead of seasonal means we may use seasonal totals.

6.4.1 Merits and Demerits

The method of simple average is very simple method of isolating of seasonal variations in this series. It assumes that the data do not contain any trend and cyclical fluctuations at all or their effect on time series values are not quite significant. This is very serious limitation since most of the economics or business time series exhibit definite trend and is affected to a great extent by cycles. Accordingly, the indices obtain by this method do not truly represent the seasonal swing in the data because they include the influence of trend and cyclical variations also. This method tries to eliminate the random or irregular component by averaging the monthly (or quarterly) values over different years. In order to arrive at any meaningful seasonal indices, first of all trend effects should be eliminated from the given values. This is done in the next two methods, viz,. ‘ratio to trend’ method and ‘ratio to moving average’ method.

6.5 Examples

Example 3.1: Use the method of monthly averages to determine the monthly indicates for the following data of production of a commodity for the year 1979, 1980, 1981

Month	1979	1980	1981
	(production in lakhs of tones)		
January	12	15	16
February	11	14	15
March	10	13	14
April	14	16	16
May	15	16	15
June	15	15	17
July	16	17	16
August	13	12	13
September	11	13	10
October	10	12	10
November	12	13	11
December	15	14	15

Solution:

Computation of Seasonal Indices

year	Months											
	Jan.	Feb.	Mar.	Apr.	May.	June	Jul.	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
Total	43	40	37	46	46	47	49	38	34	32	36	44
Monthly Average ($\bar{x}_i$)	14.33	13.33	12.33	15.33	15.33	15.66	16.33	12.66	11.33	10.36	12.00	14.66
Seasonal Index	104.89	97.57	90.25	112.21	112.21	114.62	119.52	92.66	82.93	78.02	87.03	107.30

$$Seasonal\ index\ for\ ith\ Season = \frac{\bar{x}_i}{\bar{x}} \times 100$$

$$where\ \bar{x} = \frac{1}{12}\sum_{i=1}^{12} \bar{x}' = 13.67$$

Example 3.2: Compute the seasonal indices for the following data assuming that there is no need to adjust the data for the trend.

Quarter:

	1970	1971	1972	1973	1974	1975
I	3.5	3.5	3.5	4.0	4.1	4.2
II	3.9	4.1	3.9	4.6	4.4	4.6
III	3.4	3.7	3.7	3.8	4.2	4.3
IV	3.6	4.8	4.0	4.5	4.5	4.7

Solution: Since we are given that there is no need to adjust the data for trend. The appropriate method for computing the seasonal indices is 'Simple Average' Method.

Computation of Seasonal Indices

Year	I Qrt	II Qrt	III Qrt	IV Qrt.
1970	3.5	3.9	3.4	3.6
1971	3.5	4.1	3.7	4.8
1972	3.5	3.9	3.7	4.0
1973	4.0	4.6	3.8	4.5
1974	4.1	4.4	4.2	4.5
1975	4.2	4.6	4.3	4.7
Total	22.8	25.5	23.1	26.1
Average (A.M.)	3.8	4.25	3.85	4.35
Seasonal Indices	$\frac{3.8}{4.06} \times 100 = 93.60$	$\frac{4.25}{4.06} \times 100 = 104.68$	$\frac{3.85}{4.06} \times 100 = 94.83$	$\frac{4.35}{4.06} \times 100 = 107.14$

The average of the average is: $\bar{x} = \frac{3.80+4.25+3.85+4.35}{4} = \frac{16.25}{4} = 4.06$

6.5 Ratio to Trend Method

This method is an important over the 'simple average' method of measuring seasonality and is based on the assumption that the seasonal fluctuations for any season (month, for monthly data and quarterly for quarterly data) are a constant factor of a trend. The following are the steps for measuring seasonal indices by this method.

(i) It compute the trend values (monthly or quarterly as the case may be), by the principle of least squares by fitting an appropriate mathematical curve (straight line, second degree parabolic curve of exponential curve, etc.)

(ii) Assuming multiplicative model of time series the trend eliminated by dividing the gives time series values for each season (month or quarter) by the corresponding trend value and multiplying by 100.

Thus

Trend eliminated values = $\frac{y}{T} \times 100$,

$$y \ denotes \ the \ value \ of \ time \ series = \frac{T \times S \times C \times I}{T} = S \times C \times I \times 100$$

The percentage will therefore include seasonal cyclical and irregular fluctuation. Further steps are more or less same as in 'Simple Average' Method.

(iii) Arrange these trend eliminated values according to years and months or quarters. An attempt is made to eliminate by cyclical and irregular variation by averaging the percentages of different months or quarters over different the given years. Arithmetic mean or median may be used for averaging. These averages give the preliminary indices of seasonal variations for different seasons (months or quarters).

(iv) These seasonal indices are adjusted to the total of 1200 for monthly data/or 400 for quarterly data by multiplying each of them with a constant factor k given by

$$k = \frac{1200}{sum \ of \ monthly \ indices}$$

$$k = \frac{400}{sum \ of \ quarterly \ indices}$$

for monthly or quarterly data respectively. This steps amount to expressing the preliminary seasonal indices as an percentage of their arithmetic mean.

6.5.1 Merits and Demerits

Since this method determines the indices of seasonal variations after eliminating the trend component, it definitely gives more representative values of seasonal swings as compared with the 'simple average' method. However the averaging process over different years will not completely eliminate the cyclical effects particularly, if the cyclical swings are obvious and pronounced in the given series. Accordingly, the indices of seasonal variations obtained by this method are mingle with cyclical effects also and are, therefore biased and not truly representative. Hence this method is recommended if the cyclical movements either absent or if present their effect is not so significant. If the data exhibits pronounced cyclical swings, then seasonal indices based on 'Ratio to moving average' method will reflect the seasonal variations better than this method. However, as compared with moving average method a distinct advantage of this method is that trend value can be obtained for each month (quarter) for which data are available where as there is loss of information of certain trend values (in the beginning and at the end) in the ratio of moving average method.

Remark: If we are given the monthly (or quarterly) figure for different years, then the fitting of trend equation to monthly (quarterly) data which involves a fairly large number of observations, by the principle of least squares is quite tedious and time consuming. In such a situation the calculations are simplified to a great extent by first fitting the trend equations to annual totals or average monthly or quarterly values and then adjusting or modifying it to monthly or quarterly values. This technique is explained in the following example.

6.5.2 Examples

Example 3.3: Using 'Ratio to Trend' method determines the quarterly seasonal indices for the following data.

Production of Coal (in Million of Tons)

Year	I Qrt	II Qrt	III Qrt	IV Qrt.
1	68	60	61	63
2	70	58	56	60
3	68	63	68	67
4	65	56	56	62
5	60	55	55	58

Solution:

Computation of Linear Trend

Year (t)	Yearly Totals	Quarterly Averages (y)	x==t-3	x^2	xy	Trend Values Ye= 61.4-1.46x
1	252	63	-2	4	-126	64.30
2	244	61.0	-1	1	-61	62.85
3	266	66.5	0	0	0	61.40
4	242	60.5	I	1	60.5	59.95
5	224	56.0	0	4	112	58.50
		$\sum y = 307$	$\sum X = 0$	$\sum X^2 = 10$	$\sum XY = -14.5$	

Let the straight line trend equations are:

Y= a+bx

Origin: 3rd Year; x units; 1 year

And y units: Average quarterly production (in Million tons)

The normal (least squares) equations for estimating a and b are:

$$\sum y = na + b\sum x \; and \sum xy = a\sum xb\sum x^2$$

Since $\sum x = 0, these\ give$:

$$a = \frac{\sum y}{n} = \frac{307}{5} = 61.4$$

$$b = \frac{\sum xy}{\sum x^2} = -\frac{14.5}{10} = -1.45$$

Hence the straight line trend is given by the equations:

Ye = 61.4-1.45x

Origin: 3rd year; x unit = 1 year;

Y unit: Average quarterly production.

Putting x = -2,-1,0,1,2 we obtain the average quarterly trend values for the years 1 to 5 respectively, which are the given in the last column of the above table.

From the trend equation, we observer that:

Yearly increment in the values = b= 1.45

Quarterly increment = -1.45/4= -.036

The negative value of b implies that we have a declining trend. Now we have to determine the quarterly trend values for each year.

The average quarterly trend value for the first year is 64.30. This is in fact the trend value of the middle quarter, i.e. half of the 2nd quarter and half of the 3rd quarter, for the 1st year. Since the quarterly increment is -0.36, we obtain the trend values for the 2nd and 3rd quarters of the first year as:

64.30-1/2(-0.36) and 64.30+112(-0.36)

i.e., 64.30+0.18 and 64.30-0.18

i.e., 64.48 and 64.12

respectively. The trend values for the first quarter, now becomes 64.48+0.36 = 64.84. Since the quarterly increment is -0.36, the trend values for the 4th quarter of first year and remaining quarters of other years are obtained on subtracting 0.36 from the values of 3rd quarter, viz., 64.12 successively. Trend values are given in the following table.

COMPUTATION OF SEASON INDICES

Trend Values

Trend Values					Trend Eliminated Values (Given values as of trend values)			
Year	I Qrt	II Qrt	III Qrt	IV Qrt.	I Qrt	II Qrt	III Qrt	IV Qrt.
1	64.84	64.48	64.12	63.76	104.87	93.05	95.13	98.81
2	63.39	63.03	62.67	62.61	110.43	92.02	89.36	96.29
3	61.94	61.58	61.22	60.84	109.78	102.31	111.07	110.09
4	60.50	60.14	59.78	59.42	107.44	98.10	93.68	104.34
5	59.06	58.70	58.34	57.98	101.59	93.70	87.42	100.03
Total					534.11	479.18	476.66	509.56
Average (A.M.) Seasonal Indices					106.82	95.84	95.33	101.91
Adjusted Seasonal Indices					106.85	95.86	95.35	101.94

Sum of indices = 106.85+95.86+95.35+101.94= 399.90

Since this is not exactly 400, the seasonal indices obtained as A.M. are adjusted to a total 400 by multiplying each of them with a constant factor, called correction factor:

k= 400/399.9 = 1.00025

Remark:

1. Since the sum of indices is 399.9 which is approximately 400, we may not apply any adjustment in this case.
2. Rounding to whole numbers the quarterly seasonal indices are 10, 96, 95, 103 respectively.
3. In obtaining the trend values, we fitted a liner trend equation to average quarterly production. However we could have fitted a straight line trend to annual (total) values and then, finally adjusted the trend equation to quarterly values.

6.6 Ratio to Moving Average Method

This is an important over the Ratio to trend method as it tries to eliminate cyclical variations which are mixed up with seasonal indices in the Ratio to Trend method. Ratio to Moving average is most widely used method of measuring seasonal fluctuations and involves the following steps:

i) Obtain centered 12-months (4 quarters) moving average values for the given series. Since the variation recurs after a span of 12 months for monthly data (4 quarters for quarter y data), a 12 months (4 quarters) moving average will completely pattern and seasonal variations provided they are constant pattern and intensity. Accordingly, the 12 months (4 quarters) moving average values may be regarded to contain trend and cyclical components, viz., T*C, as averaging process tries to eliminate the irregular component.

ii) Express the original values as percentage of centered moving average values for all months (quarters) except for 6 months (2 quarters) in the beginning and 6 months (2 quarters) at the end. Using multiplicative model of time series, these percentage give:

$$\frac{original\ values}{Moving\ Average\ values} \times 100 = \frac{T \times S \times C \times I}{T} = S \times I \times 100$$

Hence the ratio to moving average represents the seasonal and irregular components.

iii) Arrange these percentages according to years and months (quarters). Prelimiary seasonal indices are obtained on eliminating irregular component by average these percentages for each month (quarter), the average being taken over different years. If the variation in the set of values (percentages obtained in step (ii) of a month is small which is only due to irregular variations, the arithmetic mean may be used. If however there are some extreme values which are due to incomplete elimination of cyclical effect, one should use the median or modified mean the modified mean being the arithmetic mean computed after ignoring the extreme values.

iv) The sum of these indices should be 1200 (or 400) monthly (quarterly) data. If it is not so, then these seasonal indices obtained in step (iii) adjusted to a total of 1200 (or 400) by multiplying each of them with a constant factor

$$k = \frac{1200}{sum\ of\ monthly\ indices}$$

$$k = \frac{400}{sum\ of\ quarterly\ indices}$$

for monthly or quarterly data respectively. This steps amounts to expressing the preliminary seasonal indices as a percentage of their arithmetic mean.

Additive Model

If we use additive model of time series, then the method of moving averages for computing seasonal indices involves the following steps. [we shall state the steps for the monthly data and these can be modified accordingly for quarterly and other data.]

(i) Obtain 12 month average values. These will contain trend and cyclic component that is they will represent (T+C).

(ii) Trend eliminated values are obtain on subtracting these moving average values from the given time series values to give: y-M.A values = (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 these indices should be zero. In case it is not so, the preliminary indices obtained in step (iii) are adjusted to a total of zero by subtracting from each of them a constant factor, k= 1/12 (sum of monthly seasonal indices)

6.6.1 Merits and Demerits

Ratio to moving average is most satisfactory or widely used method for estimating seasonal fluctuation in a time series because it irons out both trend and cyclical component from the indices of seasonal variations. However, it should kept in mind that it will give true seasonal indicates provided the cyclical fluctuations are regular in periodicity as well as amplitude. An obvious drawback of this method is that there is loss of some trend values in the beginning and at the end and accordingly seasonal indices for first 6 months (or 2 quarters) of the first year and last 6 months (or 2 quarters) of the first year and last 6 month 9or 2 quarters) of the last year can not be determined.

Ratio to moving average method using the multiplicative model is illustrated in following example:

Example 3.4: Calculate seasonal indices by the 'ratio to moving average method' from the following data:

Year	I Qrt	II Qrt	III Qrt	IV Qrt.
1971	68	61	61	63
1972	65	58	66	61
1973	68	63	63	67

Solution:

Year/ Quarter (1)	Values (2)	4- Quarter moving totals (3)	4- quarter M.A (4)	Two point moving total of col. (4) (5)	4- quarter Centered (6)	Quarterly Values as % Centered M.A. (7)
1971						
I	68					
II	62					
	254	63.50				
III	61			126.25	63.125	96.63
		251	62.75			
IV	63			124.50	62.250	101.20
1972		247	61.75			
I	65			124.75	62.375	104.21
		252	63.00			
II	58			125.50	62.875	93.43
		250	62.50			
III	66			125.70	63.875	104.97
		253	63.25			
IV	61			127.75	64.125	95.50
1973		258	64.50			
I	68			128.25	64.500	106.04
		255	63.75			
II	63			12.900		97.67
		261	65.25			
III	63					
IV	67					

Trend Eliminated Values				
Year	I Qrt	II Qrt	III Qrt	IV Qrt.
1971	-	-	96.93	95.50
1972	104.21	97.67	-	-
1973	106.04	97.67	-	-
Total	210.25	190.10	201.60	196.70
Average (A.M.)	105.13	95.05	100.80	98.70
Adjusted seasonal indices	105.31	95.21	100.97	98.52

Sum of seasonal indices (A.M.) is:

105.31+95.05+100.80+98.35= 399.33

Which is less than 400. These indices are, therefore, adjusted to a total of 400 by multiplying each of them by a constant factor;

$$k = \frac{400}{399.33} = 1.0017$$

The adjusted seasonal indices are given in the last row of the above table.

Example 3.5: Calculate the seasonal indices by the 'ratio to moving average' method from the following data.

Year	Quarter	Y	4- Quarterly moving average
	I	75	
	II	60	
1972	III	54	63.375
	IV	59	65.375
	I	98	67.125
	II	65	70.875
1973	III	63	74.000
	IV	80	75.375
	I	90	76.625
	II	72	77.625
1974	III	66	79.500
	IV	85	81.500
	I	100	83.000
	II	78	84.750
1975	III	72	

	IV	93	

Solution:

Calculation of Seasonal Indices

Trend Eliminated Values (Given values as percentage of M.A. values, i.e. $\frac{y}{M.A.} \times 100$					
Year	I Qrt	II Qrt	III Qrt	IV Qrt.	
1972	-	-	85.2071	90.2485	
1973	128.1192	91.7108	85.2071	106.1360	
1974	117.4551	92.7536	89.0189	104.2945	
1975	120.4819	92.0354	-	-	
Total	366.0562	276.4998	253.3611	300.6790	
Average (A.M.)	122.0187	92.1666	84.4537	100.2263	Total 398.8653
Adjusted seasonal indices	122.3604	92.4247	84.6902	100.5069	399.9822 = 400

The seasonal indices obtained as average (A.M) above are adjusted to a total of 400, by multiplying each of them by a constant factor,

$$k = \frac{400}{Sum\ of\ seasonal\ indices} = \frac{400}{398.8653} = 1.0028$$

The adjusted seasonal indices are given in the last row of the above table.

6.7 Method of Link Relatives

Link relatives (L.R.) are the value of the given phenomenon in any season (month, for monthly data; quarter for quarterly data; day for weekly data and so on), express as percentage of its value in the preceding season. We shall explain the method for monthly data and it can be modified accordingly for quarterly or weekly data. Thus,

Link relative for any month = current month’s value/Previous month’s value *100

L.R. for March = Value (figure) for March Value (figure) for February*100.

The construction of indices of seasonal variation by the link relatives method also known as Pearson’s method involve the following steps:

(i) Convert the original data into link relative, i.e., express each value as a percentage of the preceding value.

(ii) As in case 'Ratio to Trend' or 'Ratio to M.A.' method, average these link relatives for each months, the average being taken over the number of years. Arithmetic mean or median may be use for averaging.

(iii) Covert these link relatives in two chain relatives C.R. on the basis of first season by the formula:

C.R. for any month = LR for that month* CR or preceding month/ 100.

The chain relative for January being taken as 100. For example,

CR for February = LR of February*CR of January/ 100

= LR of Feb (because CR of January = 100)

CR for March = LR of March * CR of February/100

.

.

.

CR for December = LR of December* CR of November/100

(iv) Obtain CR for first month, viz., January on the basis of the December chain called new (second) C.R. for January is given by:

New CR for January = LR of January*CR of December/100

Usually this will not be 100 due to the effect of long term secular trend and accordingly the chain indices are to be adjusted or corrected for the effect of trend.

(v) This adjustment is done by subtracting a 'correction factor' from each of the chain relatives. Let us write.

$$d = \frac{1}{12}[New\ CR\ for\ January - 100]$$

(vi) The indices of seasonal variation are obtained on adjusting these corrected chain relatives to a total of 1200, by expressing each of them as a percentage of arithmetic mean. This amounts to multiply each of them by a constant factors,

$$k = \frac{1200}{sum\ of\ adjusted\ C.R.'s}$$

Remark: For quarterly data we write

$$d = \frac{1}{4}[New\ CR\ for\ 1st\ Quarter - 100]$$

and the corrected CR's for 2nd, 3rd and 4 quarter are obtained on subtracting d, 2d and 3d from the CR's Obtained in step (iii).

Finally adjust these corrected CR's to a total of 400, by multiplying each of them by a constant factor,

K= 400/sum of the adjusted Quarterly C.R.s to get the indices of seasonal variations.

6.7.1 Merits and Demerits

(i) The average link relatives both cyclic and trend components. Though trend is subsequently eliminated by applying corrections, the indices will be truly representative only if the data really exhibits a straight line trend. However, this is not so in most of the economic and business time series.

(ii) Though not so easy to understand as moving average method, the actual calculation involved in this method are much less extensive than the ratio to M.A. or Ratio to trend method.

(iii) There is loss of only one link relative i.e., for the first season while in case of ratio to moving average method we loose some of the values in the beginning and at the end. Thus, 'Link Relatives' method utilizes the data more completely.

6.7.2 Examples

Example 3.6: Compute the seasonal indices by the 'Link Relative' Method for the following data:

Wheat prices (in Rs per quintal)

Quarter Year	**1970**	**1971**	**1972**	**1973**
1st (Jan.-March)	75	86	90	100
2nd (April-June)	60	65	72	78
3rd (July-Sep.)	54	63	66	72
4th (Oct.-Dec.)	59	80	85	93

Solution

COMPUTATION OF SEASONAL INDICES BY LINK RELATIVES METHOD

Link relatives

Year	I Qrt	II Qrt	III Qrt	IV Qrt.	
1970	-	80.00	90.00	109.26	Total
1971	145.76	75.85	96.92	126.26	
1972	112.50	80.00	91.67	128.79	
1973	117.65	78.00	92.31	129.17	
Total of L.R's average	375.91	313.58	370.90	494.20	
L.R. (A.M.)	125.303	78.395	92.725	123.550	
Chain Relatives	100.000	$\frac{78.395 \times 100}{100} = 78.395$	$\frac{78.395 \times 92.725}{100} = 72.69$	$\frac{123.55 \times 72.69}{100} = 89.810$	
Total Adjusted CR	100	78.395-3.135=75.26	72.690-6.270=66.42	89.810-9.405=80.41	322.09
Seasonal Indices	124.2	93.47	82.49	99.87	400.03

The New (second) CR for the first Quarter is:

(LR of 1st Qrt. * CR of the last (4th) Qrt.) / 100= (125.303*89.81)/100=112.54

We have

$$d = \frac{1}{4}[New\ CR\ of\ 1st\ Qrt. - 100]$$

$$= \frac{1}{4}(112.54 - 100) = 3.135$$

Adjusted CR's for the 2nd, 3rd and 4th quarters are obtained on subtracting d, 2d and 3d from the corresponding CR's.

Sum of adjusted CR's = 100+75.26+66.42+80.41= 322.09.

Indices of seasonal variations are obtained on adjusting these adjusted CR's to as total of 400 by multiplying each one of them with a constant factor,

k= 400/sum of adjusted CR's = 400/322.9 = 1.242.

6.8 Summary

In this lesion we introduce the commonly used method of isolating Seasonal Variations. The method has been explained with the help of examples.

6.9 Self Assessment Exercises

P-1 Compute seasonal indices by method of simple average from the data:

Year	Jan.	Feb.	Mar.	Apr.	May.	June	Jul.	Aug	Sep	Oct.	Nov	Dec.
1960	18	17	15	13	10	9	10	11	14	18	20	22
1961	20	17	14	14	12	11	12	13	16	19	21	22
1962	22	20	19	17	15	13	10	14	17	20	22	24
1963	24	22	20	18	16	14	9	15	18	18	24	25
1964	25	23	20	19	18	15	10	16	20	20	26	27

P-2) Using the data given below, calculate the seasonal indices by method simple averages.

Year/Quarter	I	II	III	IV
1990	72	68	80	70
1991	76	70	82	74
1992	74	66	84	80
1993	76	74	84	78
1994	78	74	86	82

P-3) Obtained seasonal indices by the ratio to trend method from the following data:

Year/Quarter	I	II	III	IV
1995	36	34	38	32
1996	38	48	52	42
1997	42	56	50	52
1998	56	74	68	62
1999	82	90	88	80

P-4) Using the link relative method to calculate seasonal indices from the following data:

Year/Quarter	I	II	III	IV
1981	283	258	244	260
1982	210	208	204	241
1983	194	168	159	183
1984	159	162	168	189
1985	184	179	176	197
1986	179	182	182	219
1987	200	204	207	243

6.10 Answers

P-1 Seasonal Index: 127.0, 113.8., 101.1, 93.1, 81.6, 71.3, 58.6, 79.3, 97.7, 109.2, 129.9, 138.5

P-2 Seasonal Index: 98.4, 92.1, 108.9, 100.5

P-3) Trend Value

Year/Quarter	I	II	III	IV
1995	27.5	30.5	33.5	36.5
1996	39.5	42.5	45.5	48.5
1997	51.5	54.5	57.5	60.5
1998	63.5	66.5	69.5	72.5
1999	75.5	78.5	81.5	84.5

Seasonal Index: 100.12, 109.55, 103.10, 87.23

P-4 Chain Relative: 100	97.25	95.97	109.94	
Correction Factor:	= -0.83			
Adjusted Chain Relative:	100	98.08	97.63	112.43

6.11 Further Readings

- Corxton, F.E. and Cowden, D.J.: Applied General Statistics, Prentice Hall 1967.
- Goon A.N., Gupta M.K. & Das Gupta B (1987) Fundamentals of Statistics Vol. I The World Press Pvt. Ltd., Kolkata.

U.P. Rajarshi Tandon Open University, Prayagraj

UGSTAT – 104
Applied Statistics

Block: 3 Demography

Course Design Committee

Dr. Ashutosh Gupta **Chairman**
Director, School of Sciences, U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anup Chaturvedi **Member**
Department of Statistics, University of Allahabad, Prayagraj

Prof. S. Lalitha **Member**
Department of Statistics, University of Allahabad, Prayagraj

Prof. Himanshu Pandey **Member**
Department of Statistics, D. D. U. Gorakhpur University, Gorakhpur.

Dr. Shruti **Member-Secretary**
School of Sciences, U.P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Dr. Alok Kumar **Writer**
Department of Statistics, Banaras Hindu University, Varanasi

Prof. R. C. Yadava **Reviewer**
Department of Statistics, Banaras Hindu University, Varanasi

Prof. V. P. Ojha **Editor**
Department of Statistics and Mathematics
D. D. U.Gorakhpur University, Gorakhpur

Dr. Shruti **Course/ SLM Coordinator**
School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

UGSTAT – 104 **APPLIED STATISTICS**
©UPRTOU
First Edition: *March 2008* (Published with the support of the Distance Education Council, New Delhi)
Second Edition: ***July 2021***
ISBN :

Printed By : K.C. Printing & Allied Works, Panchwati, Mathura - 281003

Block & Units Introduction

The ***Block - 3 – Demography,*** has five units. Broadly speaking demography is concerned with quantitative study of human population. It focuses its attention mainly on size of population, composition of population and territorial distribution of the population and changes there in and factors responsible for such changes such as fertility, mortality, migration and special mobility. Demography is essentially a science which heavily utilizes data for its study.

Unit – 7 – Source of Demographic Data, describes various sources of data from which relevant data are available for various demographic studies.

Unit – 8– Measures of Mortality and ***Unit – 9 – Measures of Fertility*** are concerned with describing various measures of mortality and fertility which are mainly useful for comparing levels of fertility and mortality of different populations.

In ***Unit – 10 – Life Tables***, gives a detailed description of life table. Life tables are powerful tools in studying longevity of persons of different ages and are useful in many practical situations.

Unit – 11 – Measures of Reproductivity, describes two important measures of reproduction viz. gross reproduction rate and net reproduction rate which are helpful in knowing the quantitative estimate of reproduction of females in their whole reproductive period.

At the end of unit the summary, self assessment questions and further readings are given.

Unit-7 Sources of Demographic Data

Structure

7.1 Definitions of Demography and Population Studies

Definitions:

Demography is the study of size, geographic distribution and the structure of the population, together with change and the factors that bring about that change in these aspects of the population. The subject is defined broadly or narrowly depending upon whether one includes in the definition the study of the factors causing change or not. Hauser and Duncan regard the field of demography as consists of demographic analysis (narrow definition) and the population studies (broad definition)

Population size is affected by the trends in the components of population change: births, deaths and migration.

The geographic distribution of the population is determined by the history in internal movement of people as well as variation in the components of population change.

The structure or the composition of the population is affected by demographic, economic and social processes that influence the individuals in the society (e.g. marriage, remarriage, divorce and widowhood, ageing economic activity, income education, religious affiliation ethnic background).

Many of the factors that influence the population are interrelated. Thus fertility may be regulated by economic activity, social attitudes, education, income of people. Inverse relationship may also be found: economic activity may be affected by the number of children a women or a couple has.

The field of population, therefore covers the entire spectrum of human life cycle from birth to death and the various activities that take place within it.

7.2 Demography and Vital Statistics

Population research is concerned with determining the health status of people, the health related actions people take health risk factors associated with their life style and differentials in these aspects according to specific characteristics of the population. The statistics are collected through numerous sources and summarized by statistical measures- some of which have their origin in demographic analyses. Vital statistics signifies either the data or the methods applied in the analysis of the data which provide a description of the vital events occurring in given communities. By vital events, again, we mean such events of human life as birth, death, sickness, migration, marriage, divorce, adoption, etc. This course introduces the basic demographic measures which are utilized for the analysis of health statistics.

7.3 Sources of Demographic Data

There are five vehicles which are used for demographic data collections: (i) population censuses, (ii) vital registrations, (iii) population registers, (iv) sample surveys and (v) miscellaneous sources, i.e. data collected by other government agencies as part of their routine administration. More recently ethnographic methods such as life histories and genealogies participants observation, interviews with key informants with methods used by anthropologies have been used for the collection of qualitative data in demography.

7.3.1 Census

A census of population is the total process of collecting compiling, evaluating, analyzing and publishing demographic economic and social data pertaining at a specified time to all persons in a country or in a well delimited part of a country. Some essential features of a population census are sponsorship, defined territory, universality, simultaneity defined periodicity, individual units, compilation and publication, and international simultaneity.

Most countries in the world conduct a population census once in a decade usually in a year ending in digit 0 or 1.

However additional topics are always included depending upon the requirement of data at the time of taking the census.

The census of population and housing is the largest statistical collection undertaken by a country. It collects information on the key characteristics of people and their housing in the country. The census includes all persons on census night with the exception of foreign diplomats and their families and foreign crew members on ships.

Methods of Collecting Information:

For census operations information can be collected by two methods. First method is one in which a trained enumerator approaches those from whom the information is to be collected. The enumerator keeps the questionnaire with himself, and only puts the questions. He himself records the reply of the party. An other method is that the questionnaire is distributed to the persons concerned from whom information is to be collected. The party is expected to answer all the questions mentioned in the questionnaire. In some cases the questionnaires are even filled up by the head of the family and filled in questionnaires are collected and then analyzed by the enumerators at his convenience.

Obviously in those countries where the respondents are education, second method can be adopted but not in the cases of those countries where the vast majority in illiterate. In an illiterate society only first method can work well. In India enumerators themselves are expected to fill in all the questionnaires because the people by and large are illiterate and even hesitant to provide information.

Each system, of course has its own advantages. In the first method since the enumerator himself fill up the questionnaire, therefore the chances of mistake are very few and minimum, but the disadvantage is that in many cases enumerator himself may not be in a position to get confidential information. There can be some type of confidential information which the party may not be willing to speak out but provide in writing when questionnaire is being filled. Similarly in a society where there is Pardah system the information can be collected only through writing by the party himself rather than by asking the questions.

Census Procedure

In spite of the difficulties which came on the way of census operations in the past, today need and necessity of scientific means and methods of census data collection is being increasingly realized. Almost every state has passed a legislation by which it is obligatory on each and enumerators. Those who refuse to give information can be both fined as well as put behind the bars. Almost in every state there is a separate department which deals with census work. Today census covers not only population but also figures are collected about houses, animals, schools, religious and charitable institutions, etc. In fact, with every census every possible effort is made to collect more and more data. It is very costly and there is always a desire that with this cost maximum possible return should be got. Then another reason is that census comes but once in 10 years and thus is not of a very frequent occurrence. Realizing the need and necessity of census UNO has given certain guidelines to all nations of the world to conduct census on uniform basis. The data is collected and brought to the notice of the society as a finished product through different stages.

Stages in Collection and Compilation of Census Data

Census data is collected through different stages and considerable labour is involved in that before it can be made available to the society in finished form. Some such important stages are:

1. Determination of Contents: It is the first important stage at which it is to be decided as to what type of information ought to be collected and what should be omitted. During census operation, almost all over the world information on following lines is collected:

Geographical

1. Location at the time of census.
2. Usual place of residence.

Personal

Name

Sex

Age

Marital Status

Place of birth

Date/ year of birth

Citizenship

Economic

(a) Type of Activity
    Occupation
    Industry
    Status
(b) Father's occupation
    Industry
    Status

Culture

Language

Nationality

Education

Level of Education

School/ College attended

Fertility

No. of children born

No. of living children

No. of dead children

Other information

Place of birth of:

(a) Father
(b) Mother

Place of residence during the last five years

Income

Place of work

Mode of conveyance used for going to the place of work

Age at marriage

No. of marriages

Whether arranged or love marriage(s)

Religion of:

(a) Father
(b) Mother
(c) Wife

Caste

Age of Mother/Father at the Time of Birth of First Child:

All efforts are made to accurately count the entire population at the census, but in such a gigantic exercise, some people are missed and not counted at all and some are counted several times. Overall, the population is under enumerated. The post enumeration survey conduct soon after the census provides an estimate of the extent of under enumeration. Census method can be both de jure as well as de facto. In de factor census the person is counted at a place where he is found whereas in de jure census he is counted at a place of his usual residence. Census data can also be collected by direct as well as indirect means. Where enumerator personally collects data that is called direct method but where information is collected by means of a schedule that is called indirect method of data collections.

Census data can be correct if every double counting is avoided and no person is left outside counting. Information should be collected for every unit and that too within a limited time. In it there is no place for sample data.

7.3.2 Vital Registration System

A vital statistics system irrespectively of how it is organized, is defined as the total process of (a) collecting by registration, enumeration or indirect estimation, of information on the frequency of certain vital events, as well as relevant characteristics of the events themselves and of the person(s) concerned and (b) compiling analyzing, evaluating, presenting and disseminating these data in statistical form.

The vital events on which data should be collected are live births, deaths, fetal deaths, marriage (statutory as well as non statutory formation of martial unions) divorce, annulments, judicial separation, adoption, legitimation and recognition.

There are recommended definitions (united Nations) of each of the vital events but countries may vary in their adaptability of the definitions.

In India, births, deaths and marriage statistics are generated through the compulsory registration of these events with the Registration General of India. The law prescribes responsibility of specific persons who must register the event within a fixed time period of the occurrence of the event.

Some specific data items which are generally collected on each of these vital events are as follows:

Births:

Month and year of registration, State of registration, Birth date, Sex of child, Mother's age, Usual residence of mother, Birth place of mother, Father's age, Birth place of father, Previous children of the marriage, Duration of marriage, etc.

Deaths:

Date of registration, State of registration, Date of death, Age at death, Sex, Marital status, Usual residence at death, Birth place, Occupation, Cause of death, place of death, etc.

7.3.2.1 Population Registers

A true population register system is a mechanism which provides for the continuous recording of information about the population in such a manner that data on particular events that occur to each individual, as well as selected characteristics describing him, are maintained on a current basis.

Population registers, if maintained on a current basis can furnish both the census type as well as the vital statistics type information. However maintenance of such registers is expensive. Except for a few Scandanavian countries, these are not kept elsewhere. Partial population registers which record limited information and only on select population are in operation in some other countries.

Under the system of registration generally every person is required to fill up certain forms. These are:

Birth Certificate:

Name

Father's name

Age of mother

Age of father

Legitimacy

Order of birth

Occupation of husband

Place of birth

Place of residence

No. of children already alive

Whether male of female

Name, if any

Date of Birth

Name of reporting person

Death Certificate:

Name of the deceased

Sex

Race/Caste

Age of the deceased

Place of death

Occupation

Marital Status

Place of death

Permanent residence

Place of birth

Cause of death

If foetal or still birth

Place of residence

Marriage Certificate:

Name of the bride

Father's name

Name of bridegroom

Name of husband/wife

Race of bride

Race of bridegroom

Residence of bride/bridegroom

Age of the bride

Place of birth of bride/ bridegroom

Occupation of bridegroom.

7.3.3 Population Sample Surveys (Adhoc Surveys)

Sample surveys are conducted by collecting information from part of the population to represent the whole. These are a good means of collecting data quickly and at a much reduced cost. Very many topics are included in a survey and the details of each topic are enlarged in order to provide the needed data which may answer the research questions.

Sample surveys can be designed using the theory of probability or without it. Examples of the former are: the simple random sampling, stratified sampling, systematic sampling, multi-state and cluster sampling. In these designs, each individuals for the survey has a known probability of inclusion in the sample. Statisticians can work out sampling errors attached to the population estimates (known as parameters) derived from such surveys.

The non probability sample designs are also of various types accidental, quota and purposive designs. The results from such surveys are not generalisable for the population from which the sample was selected.

Both types of surveys have non-sapling errors (response error, processing error, coverage and content error etc.), which are common to any data collection instrument. However sample surveys provide opportunity for a better control of the sampling errors than, for example, a population census.

National Sample Survey

In India National Sample Survey (NSS) is a permanent organization, which came into existence in 1950. The aim of this organization is to collect comprehensive information about socio-economic and agricultural statistics for the whole of India. Since its inception it has conducted several rounds of surveys and brought very useful publications as a result of each survey. Topics covered under survey include capital formation, indebtedness, employment and unemployment position, consumer expenditure, etc. It also collects information about labour force, mortality, fertility, family planning, urbanization, migration, etc. Since data is collected from primary sources, therefore, it is immense use for researchers, policy makers and administrators.

7.3.4 Hospital Records

Every hospital (as well as health centre or nursing home), maintains a record, for each patient, of each particulars as the age, sex, etc. of the patient, the nature of illness, the type of treatment administered, and the outcome. Some rounds of he NSS in India, for instance, have been used to collect such data.

7.3.5 Miscellaneous Sources

Data from agencies which are not directly responsible for demographic data collection are grouped in this category. As part of their routine administration, they collect information on certain groups of individuals; the data when tabulated can furnish a unique source of population information. Examples of such data are- overseas migration, records, social security data, health insurance commission data (including Medicare data), taxation data, hospital records, motor vehicle registrations, building approvals, electricity connections, telephone connections, persons on electoral roll school censuses, etc.

7.4 Error in Data Collection, Evaluation and its Adjustments

Demographic data, irrespective of how these are obtained are liable to errors. This situation arises because the entire process of collection, compilation and presentation of data depends on several variable components which may all be influenced by human performance. Thus any slip in planning organization and execution of the data collection instrument and post-enumeration processing may adversely affect the quality of data collected.

Errors in population data are of two types: those of coverage and those of content. If the data are collected through sampling the sampling errors are also present. The coverage error arises from (i) failure to cover the entire enumeration area (ii) failure to cover all the households, (iii) failure to cover every member of the household, and (iv) over counting persons.

The under enumeration rate of the population at the census is measure of the coverage error.

The content or the classification error arises from reporting or recording incorrectly the characteristics (such as age, sex, marital status, occupation) of persons who are enumerated. Imputation of missing data is a content error and so is the delay in registration of the vital events.

Unfortunately, nothing much can be done about the error in the data. Utmost care and thoroughness in enumeration with unfailing quality check on every phase of data collection can reduce these errors. Critical evaluation of data, including a review of collection and processing procedures may shed some light on 'unexpected' patterns in the data. The general procedures for detection of errors in the data are: (i) comparison of the observed data with an expected configuration, (ii) comparison with some other country's data (or within the sub divisions of the same country), (iii) comparing similar data from non demographic sources, (iv) balancing equation of directly inter-related data, and (v) direct checks. The post enumeration survey is a sort of direct check, which provides a measure of the extent of under-enumeration of the population as well as some other measures of the inconsistency in the reported characteristics of the population.

7.5 Rates and Ratios

In demographic analysis, interest also extends in certain types of rates and ratios, which utilize data from one or several different sources. Thus in understanding the population dynamics we are interested in calculated using two sources of population data- the vital statistics in the numerator and population census results in the denominator. There are many other rates and ratios which are calculated in demography.

Ratios are formed by dividing one number by another. The quotient is then multiplied by a constant – usually 100, or 1,000 or 10,000 etc. Three types of simple ratios and two types of complex ratios (known as the rates) are distinguished in demography.

The simple ratios are:

(1) The two number X and Y are distinct subgroups of data but both come from the same universe of population.

$$GRR = 5 \times 0.7079 \times \frac{100}{205} = 1.73 \; and$$

$$NRR = \frac{2{,}874.7}{1{,}000} \times \frac{100}{205} = 1.40$$

Sex ratio in the Total Population $= \frac{X}{Y} \times 100 = \frac{Number\ \ of\ males}{Number\ \ of\ females} \times 100$

$$Child\ women\ ratio = \frac{Number\ of\ Children\ Aged\ (0-4)}{Number\ of\ Females\ Aged\ (15-44)} \times 100$$

(2) The two numbers X and Y are distinct subgroups of data but they come from different universes.

$$Density\ of\ population = \frac{population\ of\ a\ country}{Area\ of\ the\ country}$$

(3) Of the two numbers X and Y, X is sub-set of Y (which means that Y includes X).

$$percentage\ of\ never\ married\ persons = \frac{Total\ never\ married}{Total\ population} \times 100$$

Note that the proportion is a ratio of type 3.

The complex ratios are of two types. These are referred to as the rates in demography. The numerator of the rates comes from the vital statistics, the number of events occurred or registered over one year (or other intervals) and the denominator is the 'population exposed to the risk' of these events. The events may be births, deaths, migration, marriages, divorces etc. The two rates distinguished as the rate or the probability, differ in terms of how the population exposed to events is calculated.

In the simplest case the population exposed to events is the average population during the year or the mid-year population.

The general definition of a rate is as follows:

$$\text{Rate of vital event} = \frac{Number\ \ of\ cases\ of\ vital\ event}{Total\ number\ \ of\ persons\ \ exposed\ \ to\ the\ risk\ of\ occur\ rence\ \ of\ the\ event}$$

It obvious that a rate refers to (a) a particular type of vital event (e.g. birth or death), (b) a particular geographical region (e.g. India or West Bengal) and (C) a particular period (e.g. the year 1995). The second and third points may not always be mentioned explicitly but may have to be understood from the context.

The number of persons exposed to the risk of a vital event is usually the population of the given area during the given period or some segment of that population. The population during any period, however, does not remain the same throughout. One will therefore, use the population either at the beginning of the period or at the end. A more correct procedure would use the mean population during the period:

$$\frac{1}{t_2 - t_1}\int_{t_1}^{t_2} P_t dt$$

Where $(t_1 + t_2)$ denotes the given period, the population P_t being assumed for simplicity to be an integrable function of time t. The mid-period population

$$P_{(t_1+t_2)/2}$$

will give an approximation to this figure (and would be equal to this figure if P_t were a linear function of t).

A rate, according to the above definition, will be a proper fraction. For ease of understanding, the fraction is generally multiplied by a constant, which for most rates is 1,000, Vital statistics rates are thus generally expressed 'per thousands of population'.

A vital statistics rate is some times looked upon as an estimate of the probability that a person exposed to the risk of the vital event during the given period will actually experience the event. This interpretation cannot however be given to all such rates.

Thus

The crude death rate for a year

$$= \frac{Number\ of\ deaths\ registered\ in\ a\ year}{Mid - year\ population\ for\ the\ same\ year} \times 1000$$

The probability based rates require that the population exposed to events is the population at the beginning of the year. Thus,

The probability of dying during a calendar year

$$= \frac{No.of\ deaths\ registered}{Population\ at\ the\ beginning\ of\ the\ year} \times 1000$$

7.6 Exercises and Questions

1.1 The following table gives the population figure according to 1901, 1911, 1921,……1991 censuses.

Census Year	Population (millions)
1901	238.3
1911	25.0
1921	251.2
1931	278.9
1941	318.5
1951	361.0
1961	439.1
1971	547.0
1981	683.3
1991	846.3

Write a detailed note on patterns of population growth in India during 20th century.

1.2. Describe various sources of demography data.

1.3 Mention difference between rates and ratios.

1.4 Name some important sample surveys conducted in India.

1.5 Differentiate between a sample survey and census.

Unit-8 Measures of Mortality

Structure

8.1 Introduction

Mortality analysis begins with good quality data on deaths and population. These data are conventionally obtained from vital registration systems and population censuses respectively. The crude death rate and the specific death rates (age, sex, age-sex, age-sex-cause of death specific) are simple measures of mortality. The other measures are based on the life tables. The method of its construction is dealt with in this chapter. The life table methodology has been used in many other applications in demography, such as in the analysis of marriage patterns (nuptiality tables), labour force (working life tables) school enrollment etc. The life tables are used in population estimation, population projections, and to show the impact of disability, cause of death elimination etc. on the survival of the population.

8.2 Measures of Mortality

There are some following measurements of mortality:

8.2.1 Crude Death Rate

The crude death rate is calculated by dividing the number of registered deaths in a year by the mid-year population for the same year. The rate is expressed as per 1,000 population.

$$Crude\ death\ rate = \frac{D}{P} \times 1000$$

Where D and P are deaths registered in a year and P is the mid year population.

This rate has a simple interpretation, for it gives the number of deaths that occur, on the average, per 1000 people in the community. Further it is relatively easy to compute requiring only the total population size and the total number of deaths. Besides it is a probability rate in the true sense of the term. It represents an estimate of the chance of dying for a person belonging to the given population, because the whole population may be suppose to be exposed to the risk of dying of something or the other.

However it has also some serious drawbacks. In using the CDR, we ignore the fact that the chance of dying is not the same for the young and the old or for males and females, and the fact that it may also vary with respect to race, occupation or locality of dwelling.

8.2.2 Specific Death Rate

The crude death for specific causes of death are calculated in a similar way by selecting deaths due to specific cause as the numerator and mid-year population as the denominator. Thus,

$$Cause - specific\ death\ rate = \frac{D^c}{P} \times 1000$$

Where symbol ‘c’ represents a specific cause.

The rate could be made specific to sex selecting the numerator and the denominator for each sex of the population.

Age specific Death Rates (ASDR)

The age-specific death rates are calculated from deaths and population both specific death rate (${}_mM_d$) = $\frac{nD_x}{nP_x} \times 1000$

Where ‘x’ indicates the age and ‘n’ the class interval of age.

The age-cause-specific death rate $(nM_x^c) = \frac{nD_x^c}{nP_x} \times 1000$

The age-cause-specific death rates can be calculated similarly.

It should be noted that the sum of the cause-specific rates over all causes equals the crude death rate. Similarly, the sum of the age-cause-specific death rates equals the age-specific death rate at given age.

Standardization is a technique which provides a summary measure of the rats (similar to the crude rates) while controlling for the compositional variation between the populations being compared. Thus a comparison of the standardized rates gives a 'true' comparison of the phenomenon studied. We shall illustrate the calculations of the standardized rates with the help of the deaths rates.

The ASDR is type of central death rate that is a rate relating to the events in a given category during a year to the average population of the category. In a high mortality situation, the death rates by age, that is the age of specific deaths, form a U-shaped curve indicating a high mortality in early and old ages. At low levels of mortality, the pattern of ASDR changes to J-shaped indicating a relatively higher mortality in the very early period of life, which drops to a low level after the hazards of early life and extends over a long period of life, and finally it rises sharply in old ages.

Crude Death Rate as Weighted Average of Age- Specific Death Rates

Consider the following table in which the symbols represent the following:

D (i) = Number of deaths registered at age I in one year.

P (i) = Population exposed to the deaths at age i. It is the mid year population at age i.

M (i) = Age-specific death rate at age i per person (i.e. we do not multiply the rate by the constant).

D, P and M are respectively the total number of deaths, the total population and the crude death rate per person (i.e., we do not multiply the rate by the constant).

Calculation of the Crude and Age Specific Death rates.

Age (i)	Number of Deaths D (i)	Population P (i)	Age-Specific Death Rate $M(i) = \frac{D(i)}{P(i)}$
0	-	-	-
1	-	-	-
.			
.			
85+	-	-	-
Total	D	P	M

Crude Death Rate = D/P (ignore the constant multiplier)

Age-Specific Death Rate at age (i) M(i)=$\frac{D(i)}{P(i)}$, $\therefore \quad D(i) = M(i).P(i)$

Sum over all ages, $\sum D(i) = D = \sum M(i).P(i)$

$$\therefore Crude\ Death\ Rate = \frac{D}{P} = \sum \frac{P(i)}{P}.M(i)$$

i.e. Crude death rate =

$$\sum [\Pr oportion\ f\ population\ at\ age\ (i) \times Age\ specific\ death\ rate\ at\ age\ (i)]$$

Similarly it can be shown that:

$$\text{Crude death rate} = \sum \left[\left\{ \begin{matrix} Proportion\ Males\ in\ Total \\ Population\ Age\ (i) \end{matrix} \times \begin{matrix} Age\ Specific\ Death\ Rate \\ of\ Male\ at\ Age\ (i) \end{matrix} \right\} \right]$$

$$+ \left\{ \begin{matrix} Proportion\ Females\ in\ Total \\ Population\ Age\ (i) \end{matrix} \times \begin{matrix} Age\ Specific\ Death\ Rate \\ of\ female\ at\ age\ (i) \end{matrix} \right\}$$

Thus the crude death rate is weighted average of the age specific (or the age-sex specific death rates), the weight being the age.

Taking just one composition variable, say age we can see that the crude death rate of the two population A and B could be different because (1) there are differences in the proportionate age distribution of the two populations and / or (2) there are differences in the age specific death rates of the two populations. Following the symbols used above and introducing A and B to represent the two population we can write:

Crude death rate of Population A = $\sum_{'} \frac{P(i,A)}{P(A)}.M(i,A)$

Crude death rate of Population B = $\sum_{'} \frac{P(i,A)}{P(A)}.M(i,B)$

8.2.3 Standardized Death Rates

Method of Direct Standardization

In this method the distributions of the compositional variables (age, sex, marital status etc.) of the populations being compared are made identical and the standardized rates (similar to the crude rates) are calculated such that the difference between them is only due to the variation in the age specific rates of their population.

A standard population is selected which is employed for deriving all the standardized rates in a set to be compared.

Data Needed

(1) For one compositional variable (say age) standardization age distribution of the standard population and

(2) Age specific death rates in all population to be compared.

Calculations

If M (i, x) represents the age specific death rate at age (i) for population (x) and P (i,s) is the standard population at age (i)

The standardized death rate for population x = $\sum_{'} \frac{P(i,s)}{P(s)} . M(i,x) \ or = \ \sum_{'} \frac{P(i,A)M(i,x)}{P(s)}$

The numerator is the number of expected deaths in the standard population had the age specific death rates of population (x) applied to the standard population, and the denominator is the total standard population. The rate is multiplied by 1,000 to express the rate as per 1,000 population.

(All the calculations are done with the rates per person. Finally, the standardized death rate is multiplied by the constant 1,000).

B. If the standardized death rate is required after controlling for the two characteristics of the population say age and sex the data needed will be the same as on the previous page but split sex as well.

Thus the standardized death rate for population x will be:

$$\sum \frac{[P(i,s,males).M(i,males) + P(i,s,females.M(i,females)]}{[P(s,males) + P(s,females)]}$$

C. If the death rates of males and females are to be compared, these are two different populations, and the method given under A is to be used. Thus,

$$The\ standardized\ death\ rate\ of\ males = \frac{\sum_{,} P(i,s).M(i,males)}{P(s)}$$

$$The\ standardized\ death\ rate\ of\ females = \frac{\sum_{,} P(i,s).M(i,females)}{P(s)}$$

The Standard Population

The selection of the standard population is in theory, arbitrary. However this population should be similar to the ones for whom the rates are being compared. The population of India at a most recent census data is appropriate for measuring state- differentials in mortality, or for comparing mortality trends over time for India. If two country's rate are to be compared either one country's age distribution or the average of the two country's distributions is appropriate.

Method of Indirect Standardization

This method is used when the age specific death rates for the populations to be compared cannot be calculated because of the distribution of the number of deaths by age is unavailable or not reliably available, but the total number of deaths and the age distribution of populations whose rates are to be compared are available.

Data Needed

(1) Observed number of deaths in all populations whose death rates are to be compared.

(2) Age distribution of all population whose death rates are to be compared.

(3) Age specific death rates for a population to be used as standard.

(4) Crude death rate in the standard population.

Calculations

A. If P (i,x) represents the population x at age (i), M(i,s) is age-sepcific death rate at age I in standard population, M(s) is the crude death rate in the standard population, O(x) is the observed number of deaths in population x and E(x) is the expected deaths in population x, then

The standardized death rate for population x= $\frac{O(x)}{E(x)}.M(s)$

The expected deaths in population x = E(x) =$\sum_{,} P(i,x).M(i,s)$

(All the calculations are done with the rates per person. Finally the standardized death rate is multiplied by the constant 1,000).

Rationale behind the Formula for Indirectly Standardize Death Rate:

Following the symbols we have used so far, and adding a few more, P(i, A) and P (i, B) representing two populations are A and B at age(i), P(A) and P(B) as the total populations A and B, P(S) as the total standard population, EM(A) and EM(B) as the expected death rates in population A and B, and O(S) as the crude death rate in the standard population, we have:

$$Expected\ death\ rate\ in\ population\ A = EM(A) = \frac{\sum_{,} P(i, A).M(i, s)}{P(A)}$$

$$Expected\ death\ rate\ in\ population\ B = EM(B) = \frac{\sum_{,} P(i, B).M(i, s)}{P(B)}$$

$$Observed\ death\ rate\ in\ the\ standard\ population\ O(S) = \frac{\sum_{,} P(i, s).M(i, s)}{P(s)}$$

Note that the different between EM(A) and EM(B) is only due to the fact that different age structures of the two populations are used. Both EM(A) and EM(B) are different from O(S) again due to the different age structure.

The ratio of O(S)/E(A) and O(S)/EM(B) is the effect on the death rates of varying age structures of the two populations A and B.

The crude death rates of the two populations are adjusted for age effect by multiplying the crude rates by the respective adjustment factor O(S)/EM(A) and O(S)/EM(B), to allow for the variation in their population age structures. The results are the standardized rates as can be seen from the following:

$$Standardised\ Death\ Rate\ for\ Population\ A$$
$$= \begin{pmatrix} Crude\ Death \\ Rate\ of\ Population\ A \end{pmatrix} . \frac{Death\ Rate\ in\ Standard\ Popultion}{Expected\ Death\ Rate\ in\ Population\ A}$$

$$= \left(\frac{\begin{matrix} Observed\ Deaths\ in \\ Population\ A \end{matrix}}{\begin{matrix} Expected\ Deaths\ in \\ Population\ A \end{matrix}} \right) \begin{pmatrix} Crude\ Death\ Rate\ in \\ Standard\ Population \end{pmatrix}$$

Similarly,

Standardized death rate for Population B =

$$= \left(\frac{\begin{matrix}Observed\ Deaths\ in\\ Population\ B\end{matrix}}{\begin{matrix}Expected\ Deaths\ in\\ Population\ B\end{matrix}}\right)\left(\begin{matrix}Crude\ Death\ Rate\ in\\ Standard\ Population\end{matrix}\right)$$

8.2.4 Maternal Mortality Rate

This rate is defined by the formula $1000 \times \frac{D^P}{B}$

Where D^P= total number of deaths from puerperal causes among the female population in the given community.

And B = Total number of live birth occurring in the given period in the community.

This rate may be looked upon as an alternative to, or a refined version of the corresponding cause-of-death rate.

First, here note is taken of the fact that only the part of the female population that goes through conception some time during the period, and not whole population is expressed to the risk of dying from puerperal causes (i.e, causes relating to child-birth). This population may be taken to be approximately the number of months giving birth to live born children plus the number of those delivered of dead fetuses.

8.2.5 Infant Mortality rate

The infant mortality rate (IMR), too, is an alternative to, and in a sense an improvement upon, the age-specific death rate for age 01.b.d. (last birth day) – in other words, upon the death rate for infants (i.e. children under 1 year of age). It is defined as

$$IMR = 1{,}000 \times \frac{D^P}{B}$$

Where Do = Number of Deaths among Children of age 01.b.d. (Last Birth Day) and B = Number of Live Births.

8.3 Examples, Exercises and Questions

Questions:

2.1 In the second and third columns of the following table are given the age specific death rates for Kerala and West Bengal for the year 1993. The figures in the fourth column give the estimated age distribution of the Indian population for the same year.

Age	Death Rate (per thousands)		Percentage in Standard Population (India)
(Years 1.b.d.)	Kerala	West Bengal	
0-4	3.4	17.0	12.8
5-9	0.1	1.5	12.1
10-14	0.3	0.9	11.2
15-19	0.8	1.7	10.5
20-24	0.9	2.3	9.7
25-29	1.1	1.7	8.2
30-34	1.7	2.4	6.9
35-39	1.7	2.3	6.2
40-44	2.4	4.0	5.0
45-49	4.1	4.8	4.4
50-54	7.4	10.1	3.6
55-59	12.2	16.9	3.0
60-64	21.6	24.6	2.4
65-69	27.3	40.5	1.8
70+	85.5	79.4	2.2
All ages	6.0	7.4	100.0

Sources: SRS system: Fertility and Mortality Indicators, 1993 (Tables 1,7). Office of the Registrar General, India.

Compute standardized death rates for Kerala and West Bengal.

2.2 Describe different measures of mortality giving their data requirements.

2.3 What is the major limitation of CDR? How standardized death rate is superior to CDR?

2.4 With the help of the following data relating to New Zealand, 1958, determine the crude death rate and the age-specific death rates, separately for males and females.

Age	Population (000)		Number of Deaths	
	Male	Female	Male	Female
0-	29.8	28.5	807	609
1-4	109.3	104.9	192	138
5-9	126.1	120.7	88	65
10-19	198.2	189.7	182	82
20-29	150.8	142.7	247	117
30-39	156.9	151.0	284	203
40-49	139.5	138.3	565	425
50-59	110.0	106.7	1230	746
60-69	70.1	80.9	2083	1464
70-79	45.4	54.5	3308	2650
80-	13.7	18.1	2195	2621
Total	1,149.8	1,136.8	11,181	9,120

Unit-9 Measures of Fertility

Structure

9.1 Introduction

Just like the mortality analysis, fertility analysis also begins with good quality data on births and population. These data are conventionally obtained from vital registration systems and population censuses respectively. The crude birth rate and the specific birth rates (generally woman, parity of woman, birth order of child etc.) are simple measures of fertility rates, the measures of fertility. Based on the specific fertility rates, the measures of reproductively are calculated. Some measures of fertility are based on the census or survey data alone. Fertility studies involving effect of contraception, post-partum abstinence, and breastfeeding on fertility, and birth interval dynamics make use of many other population measures.

9.2 Measures of Fertility

There are some following measures of fertility:

9.2.1 Crude Birth Rate

The crude birth rate is calculated by dividing the number of registered births in a year by the mid-year population for the same year. The rate is expressed as per 1,000 population.

$$Crude\ Birth\ Rate = \frac{Live\ births\ registered\ in\ a\ year}{Mid - year\ total\ population\ for\ the\ same\ year} \times 1000$$

This simple rate is, however not an adequate measure of fertility, as it is calculated without paying any regard to the age and sex composition of the community.

For one thing, it cannot be called a probability rate, since the whole population cannot be supposed to be at the risk of experiencing the particular type of vital event we are considering here. Only female and only those between certain ages really liable to this risk. Among such females, again the risk varies from one group to another: a woman of 25 is certainly under a greater risk than a woman of 40.

9.2.2 General Fertility Rate

The general fertility rate is calculated by dividing the number of registered births in a year by the mid-year population of females aged 15-44 years for the same year. The rate is expressed as per 1,000 female population in reproductive ages.

$$Generally\ fertility\ rate = \frac{Live\ births\ registered\ in\ a\ year}{Mid-year\ total\ female\ population\ aged\ 15-44} \times 1000$$

The formula for the GFR is thus,

$$i = 1{,}000 \times \frac{B}{\sum_{\omega i}^{\omega i} fP_x}$$

Where i= general fertility rate per 1,000 females in child bearing ages;

B= number of live births in the given region during the given period;

${}^{f}P_x$= number of females of age x 1.b.d. in the given region during the given period; and

$\omega_1\ \omega_2$ = lower and upper limits of the female reproduction period.

The computation of the GFR requires that a decision be taken before hand as to which year of life of a woman should be included in the children bearing (or reproductive) period. Although the practice varies in this respect the generally adopted method is to take ω_2= 49. Births to mothers under 15 and above 49 are so rare that they are not recorded separately but are included in the age-groups 15 and 49, respectively.

9.2.3 Age Specific Fertility Rate

T he age specific fertility rates are calculated from births and females population both specific to each age (or age group) of woman. Thus,

$$\text{Age-specific fertility rate} = \frac{Live\ birth\ to\ women\ aged\ x,x+n}{Mid-year\ female\ populaiton\ aged\ x,x+n} \times 1000$$

Where ‘x’ indicates the age and ‘n’ the class interval of age.

Thus the specific fertility rate for the age-group x to x+n-1 is

$$_{n}i_{x}=1{,}000 \times \frac{nB_x}{f_nP_x}$$

Where nB_x= number of live births to women of age x to x+n-1 in the given region during the period and

f_nP_x= number of women of age x to x+n-1 in the region during the given period.

In the case of an annual age specific fertility rate, n= 1 and here one writes simply

$$i_{x}=1{,}000 \times \frac{B_x}{fP_x}$$

All the three rates can be made 'more' specific to various groups of women. Usually, the numerator is made 'more' specific while keeping the denominator unchanged (i.e., same as defined in the above three rates). Thus births could be classified by sex, birth order of child, legitimacy status of the child-the idea being that the total of 'more' specific fertility rates is the specific fertility rate. If data permit, both the numerator and the denominator of the rates are made 'more' specific. Fertility rates by marital status of woman, parity specific age parity specific fertility rates qualify this category. Thus,

Age specific fertility rate for parity 1 women at age x =

$$\frac{Number\ of\ births\ occurred\ in\ a\ year\ to\ mothers\ of\ parity\ 1\ at\ age\ x}{Mid\ year\ population\ of\ women\ of\ parity\ 1\ and\ at\ x}$$

The quotient is multiplied by 1,000. Women are classified by parity in indicate their past number of children ever born before the birth of the current child. Children are classified by the birth order.

From the birth statistics for a year one can calculate the mean or median age of mother giving a birth in that year.

9.2.4 Total Fertility Rate

Mean or median age of the fertility schedule (a set of the age specific rates from the minimum age to the upper age of the woman's reproductive life span) is calculated from the age-specific fertility rates at single or five year age groups of women. Note that these are better measures of the average age of mother giving a birth as these take into account the age distribution of the female population in that year.

Total fertility rate is a summary measure of the age-specific fertility rates. It is calculated as the sum of the age-specific fertility if ages rates are in single years of age. If the rates are for each five year age group of woman, the sum of the rates is multiplied by 5, on the assumption that the five year age group rate will hold at each single age within the age interval of five years. The total fertility rate can be calculated for each year (or period) or for birth cohorts. The formulae are as follows:

Total fertility rate = $\sum_x f_x,\ where\ f_x\ is\ the\ age\ specific\ fertility\ rate\ at\ age\ x$

Or

$$Total\ fertility\ rate = 5.\sum_x 5f_x$$

9.3 Examples, Exercises and Questions

Example 3.1: The population (mid year population) in a community development block for the year 2007 was 167350. The number of births in the block for the year 2007 was 4130. There are 185 females in the reproductive age group per thousand population. Compute Crude birth Rate, and General Fertility Rate.

$$CBR = \frac{4130}{167350} \times 1000 = 24.7\ per\ thousands\ population.$$

As per given information, number of females in the reproductive age group will be

$$167350 \times \frac{185}{1000} = 30960$$

$$so\ GFR = \frac{4130}{30960} \times 1000 = 133.4\ per\ thousands\ females\ in\ the\ reproductive\ age\ group$$

Example 3.2: In a population there were 1277 and 1192 females in the age groups 20-24 years and 25-29 years respectively in the year 2002. There were 210 and 165 births to females of above two age groups respectively. Compute age specific fertility rates for the two age groups for the year 2002.

$$ASFR\ for\ age\ group\ (20-24) = \frac{210}{1277} \times 1000 = \frac{164.4}{1000} females$$

$$ASFR\ for\ age\ group\ (25-29) = \frac{165}{1192} \times 1000 = \frac{138.4}{1000} females$$

Exercise 3.1: The number of births occurring in Assam in 1988 is shown here classified according to age of mother, together with the female population in each age-group of the child bearing period:

Age	Female Population	Number of births to mothers in the age group
15-19	200	4,000
20-24	173	26,000
25-29	161	32,000
30-34	160	23,000
35-39	155	11,000
40-44	125	2,000
45-49	87	125
Total		

The total population of Assam in 1988 was 4,000.5 thousands.

Determine (a) the crude birth rate, (b) the general fertility rate, (C) the age specific rates and (d) the total fertility rate for 1988.

Exercise 3.2: Match the following:

A	B
(i) Crude birth Rate (CBR)	$(a) \dfrac{Legimate\ Births}{Married\ female\ populations\ (15-49)} \times k$
(ii) General Fertility Rate (GFR)	$(b)\ \dfrac{B}{P} \times K$
(iii) General Marital Fertility Rate (GMFR)	$(C) \dfrac{B}{F_{(15-49)}} \times K$
(iv) Age- Specific Fertility Rate (ASFR)	$(d)\ \dfrac{nB_x}{nF_x} \times K$

Questions 3.1 Write whether the following statements are true or false:

(a) CBR is not very sensitive to small fertility change.

(b) Although CBR is affected by age composition of population and level of fertility but not by the age pattern of fertility.

(c) GFR is a more acceptable measure of fertility level because it controls the variations in age composition within the reproductive age range.

(d) ASFRs are widely affected by variations in population composition.

(Ans.: (a) True (b) False (C) False (D) False**)**

Unit-10 Life Tables

Structure

10.1 Introduction

In its simplest form, a life table is a mathematical model for depicting mortality situation experienced by the population. Consider a group of children born in a year. They will all be of exact age 0 in the year they were born; exact age 1 in the next year and so on. Of course not all will survive to age 1; some would die between exact ages 0 and 1. If we continue to follow these children throughout their lives until no body remains, we have complete data on their survival status at each exact age from 0 (when all were alive) to the end age of life (when none of this group was left alive). These data can be placed on a Lexis diagram-along a diagonal- and the probabilities of dying from one exact age to another can be calculated. These probabilities of dying from one exact age to another can be calculated. These probabilities are the sole basis of the construction of the life table. Separate life tables are made for males and females.

The life table just described is called the cohort (or generation) life table as we traced the mortality experience of a real cohort (children born in a year). The other life table is a cross-sectional (or period) is traced. The cross-sectional life tables are said to be tracing the mortality experience of a hypothetical cohort. It should be noted that the data for the cohort life table are not easy to assemble as one will have to follow the original cohort from its origin (birth) until each member of the cohort has died. Such data covers the whole life span of the cohort. The cross- sectional life tables, on the other hand, utilize death data for a year (or over a period) from vital registrations and the mid-year population estimation from census enumerations or the post census estimates; the data which are easy to assemble.

A life table is mathematical tool that portrays the mortality conditions at a particular point of time among the population. Keyfitz defined a life table "a scheme for expressing the forms of mortality in terms of probabilities." According to Barclay, the life table is a life history of a hypothetical group, or cohort of people, as it is diminished gradually by deaths. The record begins at the birth of each member, and continues until all have died. The cohort loses a predetermined proportion at each age, and thus represents a situation that is artificially contrived.

Types of Life Tables:

Current and Cohort Life Table: Life tables can be categorized into two types according to the reference year of the table. One is the current life table and the other is the cohort or generation L.T. The current L.T.is based upon the mortality experience of a community for a short period of time such as one year, three years of an inter-censual period during which the mortality of a community has not changed substantiality. The current L.T. does not depict the mortality experiencing the ASDRs observed during a particular time. A current life table, therefore may be viewed as a "snapshot" of current mortality. It is an excellent summary description of mortality in a year or short period. The cohort or generation life table is based on the mortality experience of birth cohort, i.e., person born during one particular year.

Complete and Abridged Life Table: There are two usual ways of presenting a L.T., namely, complete and abridged life tables, according to the length of the age interval in which the basic data are presented. In a complete L.T., information is given for every single year of age from birth until the last applicable age. In abridged L.T.s., information is given only for broader age intervals such as x to x +5 years. The single abridged L.T. is usually prepared rather than the more elaborate complete life table since the abridged is less laborious to prepare and sufficient reliable for most purposes and often more convenient to use.

Conventionally the life tables are made by single years of age or by five years age group while splitting the age group 0-4 into 0 and 1-4 years. A life table calculated for single years of age is called a complete life table, whereas that based on the five years age grouping an abridged life table. If the death and population data are reliable, the complete life tables are more accurate than the abridged life tables; but there are ways of converting the abridged life tables into complete life tables and vice-versa.

10.2 Description of Complete Life Table

Let us consider the age of the person in yearly intervals. We begin with a cohort of babies who are all born in the same year. The following notations or relationships are used.

1_x The number of survivors at exact age x. Note that l_0 (known as the radix of the life table) will be the size of the original cohort (the survivors at an exact age 0). It is usually assumed to be 10,000 or 100,000.

d_x The number of deaths between two exact ages x and x + 1. It is also the number of deaths at central age x in the life table.

$$d_x = 1_x - 1_{x+1}$$

q_x Probability of dying between exact ages x and x + 1.

$q_x = d_x / 1_x$

p_x Probability of survival between exact ages x and x + 1.

$p_x = 1_{x+1}/1_x$ $\frac{l_x - d_x}{l_x} = 1 - \frac{d_x}{l_x} = 1 - q_x$

Lx Person years of exposure to the risk of dying between exact ages x and x+1

$$l_x = l_{x+1}.1 + d_x \times 0.5$$

$$= l_{x+1} + (l_x + l_{x+1}) \times 0.5$$

$$L_x = 0.5(l_x + l_{x+1})$$

$$Also\ l_x = l_x - \frac{dx}{2}$$

Persons aged x + 1 lived complete one year during the interval x to x + 1 and those who died during the interval , d_x on average they lied for half of the year (on the assumption of the uniform distribution of deaths during the year).

At age 0 the distribution of deaths within the parallelogram is not uniform only a few of the deaths at age 0 are expected to contribute to persons- years of exposure in the interval x, x+1.

L_0 is calculated in a different way.

$$L_0 = l_1.1 + d_{0.}f = l_1.1 + (l_0 - l_1).f$$

$$= f.l_0 + (1 - f).l_1$$

Where f is a separation factor.

L_x : function is also called the life table population.

M_x. : Life table death rate between exact ages x and x+1

$$M_x = d_x / L_x$$

T_x : Total number of person years lived beyond exact age x.

$$T_x = \sum_{x}^{\mu} L_x;\ \ u\ is\ the\ upper\ age\ at\ which\ no\ body\ survives$$

Note that T_0 will be the total number of person-years lived by l_0 persons.

e_x^o : Expectation of life at exact age x.

$e_x^o = \frac{T_x}{l_x}$ *the expection of life birth* $= T_0/l_{0.}$

The expectation of life at an exact age x gives a summary measures of mortality. It tells on average, how many years a person can look forward to live having lived to an exact age of x years.

P_x: Survivorship ratio from central age x to x+1.

$$P_x = L_{x+1}/L_x$$

$P_x = L_0 /l_0$; it is the survivorship ratio from birth to central age 0.

This is another summary measure of mortality based on the life table. It is calculated as the ratio of total deaths to total population in the life table. Thus,

Life Table Rate = l_0/T_0; note that l_0 is total births which is equal to total deaths, and T_0 is total life table population.

= $1/ e_0^o$

The rate is then multiplied by 1,000.

10.3 Construction of a Life Table

The pivotal column of a life table is the q_x column, as will be apparent from the following discussion. Suppose we have the value of q_x for every x from 0 upwards. We can then start with a suitable cohort – say, one or 100,000 (l_0) births. Multiplying l_0 by q_0 we get .Then $l_1 = l_0 - d_0$ Again, $d_1 = l_1\ \ q_1, q_1,\ l_2 = l_1 - d_1$. and so on. Having obtained the values in the l_x column, we can then fill in the other columns viz. L_x, T_x (for which we start from the bottom of the table and get the values successively by using the relation $T_x = L_x + T_{x+1}$) and e_x^o by means of the relations stated above.

Relationship between probability of dying and life table death rate at age x.

$$q_x = \frac{d_x}{1_x} = \ [d_x/L_x]/[1_x/L_x]$$

$$\frac{d_x}{L_x} = m_x$$

Now

$$\frac{l_x}{L_x} = \frac{L_x + \frac{1}{2}d_x}{L_x} = 1 + \frac{1}{2}m_x$$

$$\therefore \quad q_x = \frac{m_x}{1+\frac{1}{2}m_x} = \frac{2m_x}{2+m_x}$$

Proof of $\quad l_x = L_x + \frac{1}{2}d_x$

Since $\quad l_x = \frac{1}{2}(l_x + l_{x+1}) + \frac{1}{2}(l_x - l_{x+1}) = \ l_x$

10.4 Uses of Life Table

Although the primary purpose of a life table is to present a clear picture of mortality of a population, it may be put to other important uses also. These are:

(i) Life tables are used by life insurance companies in determining rates of premium for policies of persons of different ages. In fact, here the probabilities of survival of a person of given age are computed upto different ages and accordingly the premium rate is computed.

(ii) Government or firm also life tables for the determination of rates of retirement benefit for its employees.

(iii) Life tables are also used for population projection for future dates.

(iv) Life tables are used in computation of net reproduction rate and probability of widowhood , orphan-hood etc.

10.5 Examples, Exercise and questions

Example 4.1: In a population if the number of males at the age 55 was 10,53 and the probability of dying between ages 55 and 56 was 0.01 then after one year how many males will reach an age 56?

Here

$$l_{55} = 1056$$

$$q_{55} = 0.01$$

$$d_{55} = l_{55} \times q_{55} = 1056 \times 0.01 = 1056 \cong 11$$

Therefore, the number of males at age 56 will be

$$l_{56} = l_{56} - d_{55} = 1056 - 11 = 1046$$

Example 4.2: In a sample survey of a locality number of males between ages 45 and 46 were 30, 450 and 30, 320 respectively. Calculated q_{45}

Here

$$l_{46} = 30320 \; and \; l_{45} = 30450$$

$$\therefore \; d_{45} = 30450 - 30320 = 130$$

$$\therefore \; m_{45} \cong \frac{d_x}{\frac{l_x + l_x + 1}{2}} = \frac{130}{\frac{30450 + 30320}{2}} = \frac{130}{30385} = .0046$$

$$\therefore \; q_{45} = \frac{2 \times .0046}{2 + .0046} = .\frac{0092}{2.0046} \cong .00458 \; Ans.$$

$$(* \; Here \; D_x \; is \; appxomimtted \; by \; d_x)$$

Fill in the blanks:

(i) In a complete Life Table , the information are given for ____________, however, in an Abridged Life Table, these functions are computed for ____________.

(ii) A life table is based on the assumption that the cohort is closed against ______________________.

(iii) The cohort originates with some standard number of births that is called __________________________of the life table.

(iv) The column q_x represents ___________between ages x to x+1.

(v) T_x represents total number of years lived beyond age ___________ and is symbolically written as T_x = ______________+ L_x.

(vi) e_x^o Represents expectations of lie at _____________ and is written as e_x^o____________/ 1_x.

Questions 4.1: A part of a life table is given here with most of the entries missing. On the basis of the available figures, supply the missing ones.

Age x	1x	d_x	q_x	L_x	T_x	e_x^o
10	85,000	-	0.590	-	-	-
11			0.631			
12			0.680			
13	-	-	0.742			

14			0.820			
15			0.915	-	-	-
16	-	-	1.058			
17			1.115	-		-
18	-	-	1.219			
19	-	-	1.334	-	4,081,000	-

Based on above data determine the probability (a) that a child of age 10 will live at lest 5 years more, (b) that two children aged 10 and 11 will each live at least 5 year more and (c) that of two children aged 10 and 11 at least one will die within 9 years.

4.2 Fill up the several columns of the life table below assuming that deaths are linearly distributed within the age groups:

Age x	1x	d_x	q_x	L_x	T_x	e_x^o
20	0.00419	82,284	-	-	3796,020	-
21	0.00443	-	-	-	-	-
22	0.00462	-	-	-	-	-

4.3 Give a description of complete life table and discuss the relationship among various columns of a life table.

4.4 Write uses of life tables

4.5 if the values of q_x column are known, how other columns can be computed.

Unit-11 Measures of Reproductively (Population Growth)

Structure

11.1 Introduction

In fertility studies, interest extends to finding out as to how the population will be replaced in the future. More specifically, as women reproduce children how one woman will be replaced, i.e. whether one woman will be replaced by 1 woman or lesser or more number of women.

11.2 Gross Reproduction Rate

The total fertility rat calculated per woman, gives a measure of reproductively of both sexes of children. A total fertility rate of 1.8 children per women implies that if women continue to reproduce at the current level of the age-specific fertility rates (which gives a total fertility rate of 1.8), she will give birth 1.8 children. As the sex ratio at birth is usually in favour of males (it varies from population to population, but can be taken as 105 male births to 100 females births) only 0.88 female babies would be born. Thus one female will be replaced by 0.88 female in the 'long run'.

For a proper measure of population growth, it is necessary to take into account the age sex composition of the population. it is also appropriate that we should take female births alone, since it is mainly females through which a population increases. In this case age specific fertility rates will then be given by

$$f_{ix} = \frac{f_{B_x}}{fP_x} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5.1)$$

Where f_{B_x} is the number of female births to women of age x during the give period in the given community. Summing over these rates for all ages in the reproductive period, we get a measure of population growth, known as gross reproduction rate (GRR), as

$$GRR = \sum_{\omega_2}^{\omega_1} fi_x \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots (5.2)$$

GRR indicates the number of daughter who would be born, on the average, to each a group of females beginning life together, supposing non of them died before reaching the end of the child bearing period, if they experienced throughout this period the current level of fertility as represented by the series of rates fi_x.

If the fertility rates are in 5 yearly age-group, viz

$$f_5 i_x = \frac{f_5 B_x}{f_5 P_x} \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots . (5.3)$$

Then the GRR will be given by

$$GRR = 5 \times \sum_{\omega_2}^{\omega_1} fi_x \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots (5.4)$$

In some cases births may not be available according to age of mother and according to sex. Here formula cannot be applied, however, approximate value of the GRR can still be obtained if sex ratio at birth, i.e., the ratio of the number of male births to the number of female births is available. Here we shall have, approximately,

$$\frac{fB_x}{B_x} = a\ constant\ (say, k) \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots (5.5)$$

$$fB_x \rightarrow females\ births$$

$$B_x \rightarrow total\ births\ to\ females\ of\ age\ x.$$

Then

$$k = \frac{\sum_{\omega_2}^{\omega_1} f\ B_x}{\sum_{\omega_2}^{\omega_1} B_x} = \frac{f_B}{B} \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots . . (5.6)$$

So that

$$fB_x = B_x \times \frac{f\ B}{B}\ and fi_x = i_x \times \frac{f\ B}{B} \dots \dots \dots \dots \dots \dots \dots \dots \dots (5.7)$$

Therefore, an estimate of the GRR will be

$$\frac{f\,B}{B}\sum_{\omega_2}^{\omega_1} i_x \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(5.8)$$

$\sum_{\omega_2}^{\omega_1} i_x$ is just the TFR except for the usual multiplier 1,000.

For India, the sex ratio at birth may be taken to be 105 males to 100 females. Hence for a year for which the TFR is approximately 3.54, the GRR Will be

$$3.54 \times \frac{100}{205} = 1.73$$

Thus this rate is calculated as follows:

GRR = Total fertility rate ×sex-ratio at births

A gross reproduction rate of less than unity indicates a declining population in the future.

One limitation of the gross reproduction rate is that it assumes that all women who enter the reproductive age (say 15 years) will live to the end of their reproductive life (to age 49 for example). This assumption is not valid as some women die during their reproductive life span.

11.3 Net Reproduction Rate

As mentioned above GRR does not take cognizance of the fact that some of the females who are assumed to begin life together may die before reaching age 15, some may die between ages 15 and 16 and so on.

That is the GRR takes into account current fertility but ignores current mortality.

Taking this fact into account the factor of mortality in measuring population growth, a life table for females on the basis of the observed age-specific death rates for females,

f_m is constructed. The values in the L_x column of the table (denoted by fi_x in this case) give the mean size of the cohort of fi_0 females in the age interval x to x+1 for varying x. Hence,

$fi_0 \times {}^{f}L_x$…………………(5.9)

gives the number of female children that would be born to the cohort at age x l.b.d. Summing (5.9) over reproductive period we have

$$\sum_{\omega_2}^{\omega_1} fl_0 \times f_{L_x} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(5.10)$$

is the total number of female children that are excepted to be born to the fl_0 female during their life time. Thus a new measure of population growth is

$$\frac{1}{fl_0}\sum_{\omega_2}^{\omega_1} fi_x \times f_{L_x} \text{ ………………………………..} (5.11)$$

Which is called the net reproduction rate (NRR). The NRR shows how many females would be born on the average per member of a group of females beginning life together, if they were observed rates of mortality and fertility throughout their life time.

The NRR is usually computed by the formula

$$\frac{1}{fl_0}\sum_{\omega_2}^{\omega_1} fi_x \times f_{L_x} = \sum_{\omega_2}^{\omega_1} fi_x \times f_{P_0} \text{ ………………………………..} (5.12)$$

The quantities $fi_x/fl_0 = f_x P_o$ are called the survivorship values for females.

If quinquennial fertility rates $f_5 i_x$ are given then an estimate of the NRR is

$$\frac{1}{fl_o}\sum_5^f i_x \times f_5 i_x \text{ ………………………………….} (5.13)$$

Where

$$f_5 L_x = f_{L_x} + f_{L_{x+1}} + \cdots \text{ ………..} f_{L_{x+4}}$$

Obviously, the NRR cannot be greater than the GRR.

Thus NRR indicates how many future mothers would be born to present mother according to the current levels of fertility and morality. If the NRR = 1, then it may said that current fertility and mortality are such that a group of newly born females will easily replace itself in the next generation. In such case the population may be said to have a tendency to remain constant in size. It may be said to show a tendency to increase or decrease according as the NRR > or < 1, for in that case a group of females is expected to be replaced by a larger or a smaller number of females in the next generation. In this way the NRR may be looked upon as a good index of population growth.

Example 5.1: Determination of Gross and Net Reproduction Rates for India, 1993

(1) Age in Year	(2) Age-specific fertility rate	(3) Female life-table stationary population	(4) Col. (2) * col. (3)

15-19	0.0696	4180	290.9
20-24	0.2346	4123	967.3
25-29	0.1897	4063	770.8
30-34	0.1143	4001	457.3
35-39	0.0611	3934	240.4
40-44	0.0285	3860	110.0
45-49	0.0101	3763	38.0
			2,874.7
Total	0.7079	-	

The sex ratio at birth for the country may be supposed to be 105 males to 100 females. Hence the above table, we get

$$GRR = 5 \times 0.7079 \times \frac{100}{205} = 1.73$$

$$and\ NRR = \frac{2874.7}{1000} \times \frac{100}{205} = 1.40$$

Example 5.2: The procedure for computing the GRR and NRR is also shown in the following Table:

Methods for Computing GRR and NRR, U.P. 1971

Age group	Female population (1)	Live Birth (2)	ASFRs (per women) (3)=(2)/(1)	Female live birth (4)	Age specific maternity rates (5)=(4)/(1)	Age-specific survival rates* ($_5L_x/l_0$) (e_x^s=68) (6)	Expected Female births per women (7)=(5)x(6)
15-19	79865	8813	0.1103	4293	0.0538	0.90530	0.0487
20-24	63315	17620	0.2783	8582	0.1355	0.90048	0.1220
25-29	51680	14585	0.2812	7104	0.1370	0.89472	0.1226
30-34	44440	10235	0.2303	4986	0.1122	0.88799	0.0996
35-39	38795	7569	0.1951	3686	0.1950	0.88009	0.0836
40-44	32250	2760	0.058	1344	0.0417	0.87013	0.0363
45-49	26720	381	0.0143	186	0.0070	0.85705	0.0060
Total	-	61963	1.1951	30181	0.5822	-	0.5188
TFR	-	-	5,9755	-	2.9110	-	-
GRR*	-	-	-	-		-	2.5940
NRR	-	-	-	-		-	-

Here

$$NRR = \frac{1}{1_0}\sum_{x=15}^{49} L_x f_x^w = (.5188) \times 5 = 2.60$$

* Probability of surviving from birth to the mid point of the age group,

$$** \, GRR = TFR \times \frac{1}{1 + S.R. at\ birth}$$

$$= 5.9755 \times .4878 = 2.91$$

Or

$$GRR = \sum_{x=15}^{45} 5f_x^w = (.5822) \times 5 = 2.91$$

$$NRR = \sum_{x=15}^{45} 5f_x^w \times \frac{5l_x^w}{l_0} = (.5188) \times 5 = 2.91$$

11.4 Exercises and Questions

5.1 The quinquennial fertility rates (computed on the basis of female births alone) for Kerala 1961 are shown in the following table, together with the survival factor for each 5 year age-group (which is the probability for a newborn female to survive till the mid point of the age group and is approximately equal to $f_5 L_x / 5^f\, l_o$):

Age	Fertility rate (female births)	Survival factor
15-19	0.0106	0.968
20-24	0.0660	0.968
25-29	0.0673	0.964
30-34	0.0410	0.957
35-39	0.0214	0.953
40-44	0.0065	0.944
45-49	0.0006	0.929

Compute the GRR and NRR for Kerala for 1961 on the basis of the above data.

5.2 Fill in the blanks:

(a) Computation of TFR is based on births of both sexes whereas GRR is based on ……………………………………only

(b) Like TFR, GRR also assumes that women in reproductive age groups ……………………till the end of their reproductive period.

(c) GRR in a population is 1.8 means if 100 mothers follow the current schedule of fertility they will be replaced by ………………..daughters.

(d) NRR is nothing but a refinement over GRR where the is introduced.

(e) Reproductive survival ratio is defined as the ratio of …………….

(f) Fertility of replacement level corresponds to the value of N.R.R. = ………………………….

(Ans.: (a) female births, (b) survives (c) 180 (d) mortality (e) $\frac{5l_x^w}{l_0}$ (f) 1)

5.3 Explain GRR and NRR and shown that NRR ≤ GRR. When GRR will be equal to NRR?

U.P. Rajarshi Tandon Open University, Prayagraj

UGSTAT – 104
Applied Statistics

Block: 4 Statistical Quality Control

Course Preparation Committee

Prof. S. K. Pandey **Writer**
Department of Statistics, Lucknow University, Lucknow

Dr. Shruti **Writer**
School of Sciences, U.P. Rajarshi Tandon Open University, Prayagraj

Prof. V. P. Ojha **Reviewer**
Department of Statistics and Mathematics
D. D. U.,Gorakhpur University, Gorakhpur

Prof. Umesh Singh **Editor**
Department of Statistics, Banaras Hindu University, Varanasi

Dr. Shruti **Course/ SLM Coordinator**
School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

UGSTAT – 104 **APPLIED STATISTICS**

First Edition: *March 2008* (Published with the support of the Distance Education Council, New Delhi)
Second Edition: ***July 2021***
ISBN :

Printed By : K.C. Printing & Allied Works, Panchwati, Mathura - 281003

Block & Units Introduction

The ***Block - 4 – Statistical Quality Control*** deals with the theory of statistical quality control, has four units.

Unit – 12 - Introduction to Statistical Quality Control is introductory and gives the concept of control charts, control limits and sub groupings.

Unit – 13 - Control Charts for Variables describes control charts for variables.

Unit – 14 - Control Charts for Attributes discuss control charts for attributes.

Unit – 15 - Principles of Acceptance Sampling, Presents the principles of acceptance sampling along with single sampling plan, its OC, ASN, AQL, LTPD and AOQL functions.

At the end of unit the summary, self assessment questions and further readings are given.

Unit-12: Introduction to Statistical Quality Control

Structure

12.1 Introduction

In any production process, regardless of how well designed and carefully maintained, a certain amount of inherent or natural variability will always exist. This natural variability is the cumulative effect of many small essential uncontrollable causes, often called as "stable system of chance causes" or "allowable causes". A process which is operating with only allowable causes of variation present is said to be in "Statistical control".

Other kinds of variability may occasionally be present in the output of a process. These are large variations that are attributed to special causes like differences among machines or operators or raw material used or interaction among them. Such variability usually represent an unacceptable level of process performance. We refer to these sources of variability as "assignable causes."

It is very typical for production process to operate in –control state, producing acceptable products for relatively long periods of time. Occasionally, however assignable causes will occur seemingly at random, resulting in a "shift" to an out of control state where a larger proportion of the process output does not confirm to requirements. A major objective of statistical quality control is to quickly detect the occurrence of assignable causes or process shift so that the investigation of the process and the corrective action may be undertaken before very many non conforming units are manufactured.

In the above type of problem, our aim is to control the production process so that the proportion of defective (or non-conforming) items is not excessively large. This is known as "process control". In other type of problem, we like to ensure that the lots of manufactured goods do not contain an excessively large proportion of defective items. This is known as "product control of lot control". The two are different problems, because even when the process is in control so that the proportion of defective products for the entire output over a long period may not be large, it is possible that the entire output over a long period may not be large, it is possible that the individual lot of items may contain excessively technique of CONTROL CHART, whereas product control is achieved through SAMPLING INSPECTION.

12.2 Objectives

After studying this unit you should be able to understand

- The need of Statistical Quality Control
- Advantages of Statistical Quality Control
- General theory of Control Charts
- Obtaining control limits (3-σ limits)
- Out of control criteria

12.3 Advantages of Quality Control

Every manufacturer wants the process of manufacturing to be cost effective. Also a consistent quality of product provides a better movement in the market rather than a product with variable characteristics. So, the quality control measures the utmost requirement of any production process. Main advantages of quality control are:-

(1) With a regular use of quality control any defect in the production process can be detected at an early stage so that damage due to the defect is minimum.

(2) With the use of quality control techniques, we get better quality assurance at lower inspection cost.

(3) Statistical quality control provided better time management and hence waste of time and material is minimized.

(4) Wherever testing is of destructive type, by means of quality control only, one can get a sample of appropriate size so that quality can be checked with minimum waste of money.

(5) By means of quality control, any product reaches to the market with a consistent quality characteristics and so captives the market early. Also with a maintained quality, movement of the product is faster and the chances that product will come back to supplier due to poor quality are minimum.

12.4 Quality Characteristics

By quality characteristic we mean any characteristic of the product which is of interest in determing its quality. Many quality characteristics are measurable qualitatively (or numerically) and may be looked upon as "variables" which may be continuous eg:- Length, area, thickness, volume, density or chemical composition of a product or discrete e.g. the number of defects in a piece of cloth.

Often the quality characteristic cannot be measured and is expressed as an "attribute". Here each item may be classified as 'good' (or non-defective) or 'bad' (or defective). Also an item which has one or more defects is defective. Again although a characteristic may be measureable, one may treat is as an tribute for simplicity, e.g. a manufacture producing rods may classify a rod as defective of it is too long or too short and thus avoid measuring its actual length.

12.5 Control Charts: Basic Principles

Suppose samples of a given are taken from a process at more or less regular intervals and suppose for each sample statistic T is computed. For example, T may be the sample mean, or sample range or sample standard deviation or sample fraction defective. Being a sample result T will be subject to sampling fluctuations. If the process is in control, i.e., no assignable causes of variation are present, the sampling fluctuations of T should be due to chance variation alone. Supposing in such a case

$$E(T) = \mu_T \text{ and } var(T) = \sigma_T^2$$

We may take any values of T lying outside the limits $(\mu_T - 3\sigma_T)$ and $(\mu_T + 3\sigma_T)$ as an indication of the presence of systematic variation, i.e., variation due to assignable causes. The reason behind this argument is that in case T is normally distributed with mean μ_T and variance σ_T^2 (and the process is stable), then

$$P[|T - \mu_T| \leq 3\sigma_T]$$

$$= P[-3\sigma_T \leq T - \mu_T \leq 3\sigma_T]$$

$$= P\left[-\frac{3\sigma_T}{\sigma_T} \leq \frac{T-\mu_T}{\sigma_T} \leq \frac{3\sigma_T}{\sigma_T}\right]$$

$$= P[-3 \leq Z \leq 3]$$

$$where - Z = \frac{T-\mu_T}{\sigma_T} \sim N(0,1)$$

$$= 2P[0 < -z < 3]$$

$$= 2 \times 0.49865 \approx 0.9973$$

$$i.e., \qquad P[|T-\mu_T| \leq 3\sigma_T] \approx 0.9973$$

Even when T is non normal from the Chebychev's Inequality

$$P[|x-E(x)| \leq \varepsilon] \geq 1-\frac{V(x)}{\varepsilon^2},$$

$$We\ have\ for\ \varepsilon = 3\sigma_T$$

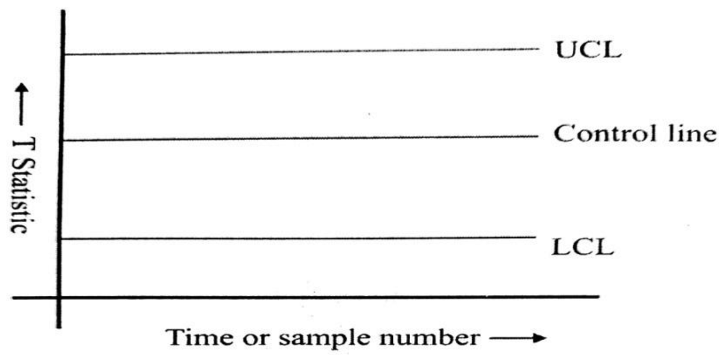
$$P[|T-\mu_T| \leq 3\sigma_T] \geq 1-\frac{\sigma_T^2}{9\sigma_T}$$

$$P[|T-\mu_T| \leq 3\sigma_T] \geq 1-\frac{1}{9}$$

$$P[|T-\mu_T| \leq 3\sigma_T] \geq \frac{8}{9}$$

Thus if the observed T_i (where i stands for the i^{th} sample) lies between the limits $(\mu_T - 3\sigma_T)$ and $(\mu_T + 3\sigma_T)$, it is taken to be a fairly good indication of non existence of assignable causes of variation at the time i^{th} sample was taken. If the observed Ti wanders or falls outside the limits, one suspects the existence of assignable causes of variation and the process is supposed to be out of control. He obvious action is then to stop the process and to hunt for and remove assignable causes.

The theory of control charts was developed by Dr. Walter Shewhart. It is a horizontal chart where time or sample number is plotted on the abscissa and the values of statistics T are plotted as ordinates. There is a 'central line' corresponding to the mean value μ_T a 'lower control limit' (LCL) corresponding to $(\mu_T - 3\sigma_T)$ and an upper control limit (UCL) corresponding to$(\mu_T + 3\sigma_T)$ as shown in the following diagram.



If μ_T and σ_T are not known , it is possible to estimate them. If several samples are taken, the mean of T is estimated from the mean of sample and standard deviation of T is estimated from the within sample variation of the samples.

According to Dr. Shewhart a control chart ma serve, first to define the goal or standard for a process that the management might strive to attain, second it may be used as an instrument for attaining that goal; and third, it may serve as a means of judging whether the goal has been attained. It is thus an instrument to be used in specification, production and inspection and when so used, bring these three phases of industry into an independent whole. Let us elaborate this.

If the sample values of T are plotted for a significant range of output and time and if these values all fall within the control limits and show no systematic pattern, we say that the process is in control at the level indicate by the chartered thus a control chart may be used to specify the goal of management.

The control chart may also be used to attain certain goals with respect to process quality. The central line and control limits may be standard values chosen by the management such that they want the process to be control at that level of quality. Sample data are plotted on the chart and if departures from the in control state are investigated and corrected, then eventually the process may be brought into control at the target or standard values.

Lastly a control chart may be used for judging whether the state of control has been attained. If the sample values of T all fail within the control limits without varying in a non-random manner within the limits, then the process may be judged to be within control at the level indicated by the chart. Likewise, if a process has been judged to be in control and new sample results continue to fall within the limits on the chart (without being in any non random pattern) the process may be judged to be continuing in a state of statistical control at the given level.

12.6 Operating Characteristic of a Control Chart

There is a close connection between control chart and hypothesis testing. Essentially the control chart is a test of the hypothesis that the process is in a state of statistical control. A sample point falling within the control limits is equivalent to accepting the hypothesis of statistical control and a point falling outside the limits is equivalent to rejecting the hypothesis of statistical control. Just as in hypothesis of statistical control. Just as in hypothesis testing, we may think of probability of a type I error of the control chart (concluding that the process is out of the control when it is really in control) and the probability of type II error of the control chart (concluding that the process is in control when it is not).

It is occasionally useful to use the operating characteristic (OC) curve of control chart to display the probability to type II error. An OC curve shows the probability (or risk) of inferring that process is in control at the designated level, when the process is actually open at a different level. This would be an indication of the ability of a control chart to detect process shifts of different magnitudes.

12.7 Choice of Control Limits

Specifying the control limits is one of the critical decisions that must be in designing control chart. By moving the control limits further from the central line we decrease the risk of type I error but, simultaneously, increase the risk type II error. If we move the control line close to the central line, the risk of type I error increase while the risk of type II error decreases.

In the above discussion we have considered control limits. We could as well as use 0.001 probability limits, k*, such that

$$P[|T-\mu_T| > k^*\sigma_T] = 0.002$$

If T is normally distributed, k* = 3.09. Some analysts suggest using two sets of limits on control chart. The outer limits at 3σ all the usual action limits that is when a point falls outside these limits, a search for an assignable cause is made and a corrective action is taken if necessary. The inner limits usually at 2σ are called ‘warning limits’. When probability limits are used the action limits are 0.001 limits and the warning limits are 0.025 limits. If one or more points fall between the warning limits and action limits or very close to the warning limits, then we should be suspicious that the process may not be operating properly. One possible action then is to increase the sampling frequency and use the additional data in conjunction with the suspicious points to investigate the state of control of the process.

12.8 Sample Size and Sample Frequency

In designing a control chart one must specify both the sample size and sampling frequency. In general larger samples will make it easier to detect small shifts in the process. This can be seen from the OC curves of a control chart of different sample sizes. If the process shift is relatively large than those that would be employed if the shift of interest is small.

The most desirable situation from the point of view of detecting shift would be to take large sample very frequently. However, this is not economically feasible, i.e., either we take small samples at short intervals or large samples at larger intervals. Current industry practices to favour the former among the two.

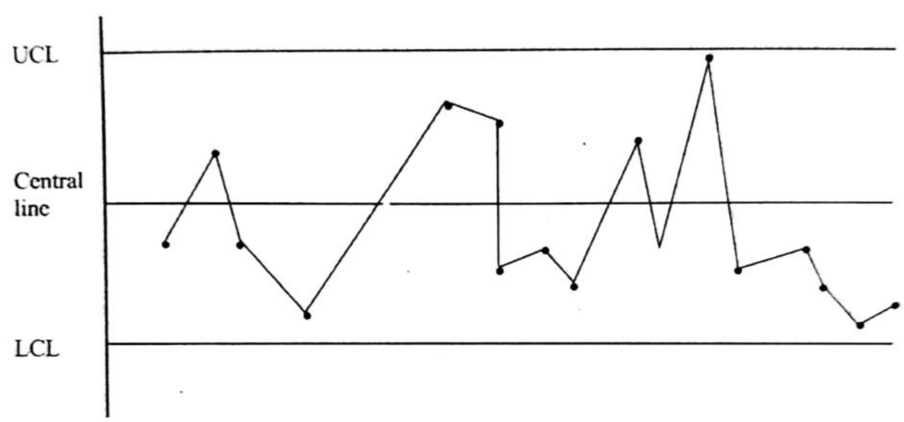
12.9 Rational Subgroups

A fundamental idea in the use of control charts is the collection of sample data according to what Shewart called 'rational subgroup' concept. Generally, this means that the subgroups or sample should be so selected that if assignable causes are present, the choice for difference between the subgroup will be minimized. In other words, the products within a subgroup should be very homogeneous, while the difference between subgroup should be very homogeneous, while the differences between subgroups will indicate the presence of systematic variations.

The most obvious bases for selection of subgroups is the order of production. Each subgroup will then consist of products of a machine or a homogenous group of machines for a short period of time, so that there cannot be any unmarkable in the cause system within the period, when there are two or more machines having different pattern of variation, it may necessary to have different subgroups for different machines. It may, therefore be necessary to have different subgroups for different machines or different operators or different shifts etc.

12.10 Analysis of Pattern on Control Charts

A control chart may indicate an out of control condition either when one or more points fall beyond the control limits, or when the points falling within the control limits exhibit some non-random or systematic pattern of behavior for example see the following control chart:

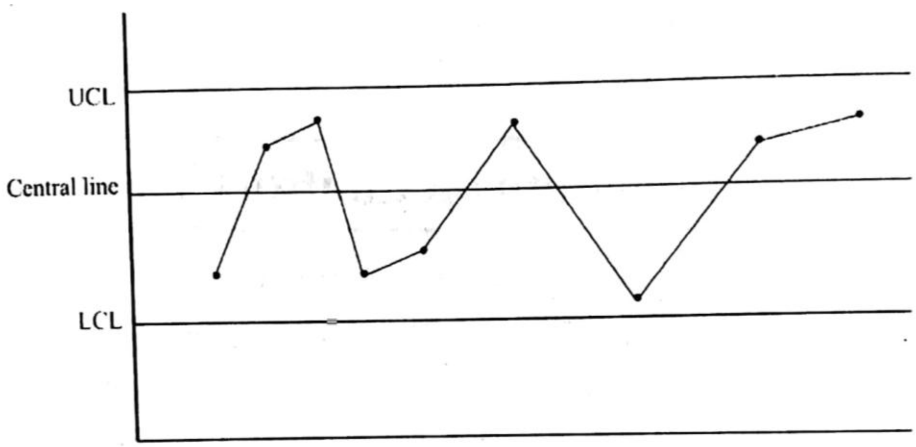


Although all the points fall within the control limits, the points do not indicate statistical control because pattern is very non-random in appearance. Most of the points fall away the central line. We observe that there are case of long run up and long run down.

We define run as a sequence of observation of the same type (increasing or decreasing). In addition to runs up the down, we could define the types of observations as those above and below the central line, respectively, so that two in a row above the central line would a run of length two.

A run of length 8 or more points has a very low probability of occurrence in a random sample of points. Consequently any type of run of length 7 or more should be taken as single of an out of control condition, For example 8 consecutive points on one side of the central line would indicate that the process is out of control.

Other types of patterns may also indicate an out of control situation. See the following chart:



Note that the points exhibit a cycle behavior although they fall within control limits. Such a pattern would indicate a problem with the process, such as operator fatigue, raw material deliveries etc. While the process is not out of control the yield may be improved by elimination or reduction of sources of variability causing cyclic behavior.

The problem is one of pattern recognition that is recognizing systematic or non-random patterns on the control chart and indentifying the reason for this behavior. The ability to interpret a particular pattern of the process.

Several different criteria may be applied simultaneously to the control chart to determine whether the process is out of control. Some such criteria used in practice are as follows.

1) One or more points outside the control limits.

2) A run of at least seven or eight points where the run may be of any kind.

3) Two of three consecutive points outside the 2σ warning limits, but still inside the control limits.

4) Four of five consecutive points beyond the 1σ limits.

5) An unusual or non random pattern on the data

6) One or more points near a warming or control limits.

Suppose an analyst uses k-test criteria, criteria i having type I error of false alarm probability for the decision α_i the probability of type I error based on all k tests is

$$\alpha = 1 - \prod_{i=1}^{k}(1 - \alpha_i)$$

assuming independent of tests.

12.11 Rate of Detection of Change in Average Level (Average Run Length or A.R.L. Function

The A.R.I. is the average number of sample points which must be plotted on a chart before a point indicates an out of control situation. This indicates how quickly the chart will detect any shift in the process average.

The OC function of the chart gives the probability P_a of deciding that the process is in control , P_a as a function of process average μ Let.

$$p=1-P_a$$

be the probability of deciding that the process is out of control.

If N is the number of points necessary before taking a decision that the process is out of control,

$$A.R.L. = E(N) = \sum_{n=1}^{\infty} n(1-p)^{n-1}$$

$$= p[1-(1-p)]^{-2}$$

$$= p[1-(1-p)]^{-2} = \frac{p}{p^2}$$

This as a function of process average gives the A.R.I. curve of the chart for true average μ but should be small for shift in .

12.12 Key Words

Allowable cause of variation: The variation which is due to random causes or chance causes. It is natural to the process and can not be prevented.

Preventable causes of variation: This is the variation which occurs because of any defect in the production process. Hence it is preventable.

Quality: In statistical termed, the quality is that specification which a manufacture desire to attain through the process. This specification may be quantitative (measurable) or qualitative (attributive or non measurable).

Control limits: Control limits provide us an interval in which we desire our product specification to lie.

3σ *Control limits:* If the process is in control then the probability that the outcome will lie within 3σ control limits in 99.73%.

12.13 Exercises

(1) Explain the theoretical basis of control charts.

(2) What are operating characteristics of a control chart.

(3) Discuss how will you analyze the pattern on a control chart.

12.14 Summary

The statistical quality control or SQC is a statistical method for separating allowable variation from preventable variation in any production process. In this way appropriate step can be taken early as possible whenever assignable causes are operating in the process. This is achieved with the use of Shewart's Control Chart which in turn are based on the 3-σ control limits. Thus upper control limit and lower control limit are obtained for any ongoing process and we want to output of the process to lie strictly within control limits. If it is not so, process is termed as out of control. Various types of control charts are described in the next unit.

12.15 Further Readings

- Burr, I.W. Engineering Statistics and Quality Control Mcgraw Hill.
- Cowden, D.J. Statistical Methods in Quality Control, Prentica Hall.
- Goon A.N., Gupta M.K. & Das Gupta B (1987) Fundamentals of Statistics Vol. I The World Press Pvt. Ltd., Kolkata.

Unit-13: Control Charts for Variables – X, R and σ Charts

Structure

13.1 Introduction

Let us start with a quality characteristic (x). Suppose it is a continuous variable (length or diameter or breaking strength). For manufactured products which are solely subject to random variation may be supposed to be normally distributed, (From the central Limit Theorem), the sum of a large number of independent components each of which contributes a relatively negligible proportion to the total variability of x will be normally distributed. Thus the different distributions of X will be normally distribution for the different sub-groups are supposed to follow the normal distribution. Let the ith sub-group has mean μ_1 and standard deviation (s.d.) σ_i^2 . To examine whether the process is in control, we require to see whether all the $\mu' s$ and the $\sigma' s$ are same. The four types of situation may arise:

(a) The process is in control,

(b) The mean is out of control not the s.d.,

(c) The s.d. is out of control not the mean,

(d) Both the mean and the s.d. are out of control.

The appropriate estimator corresponding $u \text{ and } \sigma \text{ are } \bar{x} \text{ and } s$ respectively. hence the whole judgment regarding control or lack of it is based on control charts for $\bar{x} \text{ and } s$. it is to be remembered, however, that the range R in spite of its inferiority to s from the theoretical point of view, is simpler and easier to compute. Hence in quality control the range is often preferred to the S.D. and one would frequently use charts for $\bar{x} \text{ and } R$ instead of using charts for $\bar{x} \text{ and } s$.

13.2 Objectives

After reading this unit, you should be able to understand.

- The preparation of control charts for mean, range and standard deviation.
- How to prepare control chart when standards are specified or unspecified.
- How to check the process is in control.

13.3 Control Charts for Mean ($\bar{X}$)

To check whether mean of the products is in control, we use this chart. In the following paragraph we will give you the various control limits.

13.3.1 Standards Given

For samples of size n per sub-group, we have for a stable system with mean μ and standard deviation σ, we know that if $\bar{x}$ is the sample mean then

$$E(\bar{x}) = \mu$$

$$and\ \sigma_x = \frac{\sigma}{\sqrt{n}},$$

Further, the observation in each sub-group are supposed to be mutually independent.

If the values for $\mu \text{ and } \sigma$ are specified as $\mu' \text{ and } \sigma'$ the control chart for $\bar{x}$ mean is given by

$$LCL = \mu' - 3\frac{\sigma'}{\sqrt{n}} = \mu' - A_1\sigma'$$

$$Central\ line = \mu'$$

$$UCL = \mu' + 3\frac{\sigma'}{\sqrt{n}} = \mu' + A_1\sigma' \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.1)$$

Where,

$$A_1 = 3/\sqrt{n}$$

13.3.2 Standards not Given

Suppose there are m sub-groups and let the successive sample means are $\bar{x}_1, \bar{x}_2, \dots \dots \bar{x}_m$ and the successive standard deviations be s_1, s_2, s_m and the successive ranges be R_1, R_2, ……R_m. Since μ *and* σ are not specified these are estimated from the sample themselves. Let

$$\left.\begin{aligned} \bar{x} &= \sum_i \frac{\bar{x}_i}{m} \\ \bar{s} &= \sum_i \frac{s_i}{m} \\ &\text{and} \\ \bar{R} &= \sum_i \frac{R_i}{m} \end{aligned}\right\} \dots \dots \dots \dots \dots \dots \dots \dots \dots (2.2)$$

Which are the pooled mean the mean of sample standard deviations and the mean of sample ranges, respectively.

The relations

$$E(\bar{x}) = \mu \qquad \dots \dots \dots \dots \dots \dots \dots . (2.3)$$

$$E(\bar{s}) = c_2\sigma (valid\ for\ a\ normal\ variable\ x) \dots \dots \dots \dots \dots \dots (2.4)$$

and

$$E(\bar{R}) = d_2\sigma \ \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots (2.5)$$

Where c_2 and d_2 are the function of n.

Thus an estimate of μ *is* $\hat{\mu} = \bar{x}$ ………………(2.6)

$$\hat{\sigma} = \frac{\bar{s}}{c_2} \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots . (2.7)$$

and

$$\hat{\sigma} = \frac{R}{d_2} \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots . (2.8)$$

In case one uses the estimates (2.6) and (2.7) the chart for mean will be based on

$$LCL = x - A_1\sigma'$$

$$Central\ line = x$$

$$UCL = \bar{x} + A_1\bar{s} \quad \ldots\ldots\ldots\ldots\ldots\ldots..(2.9)$$

Where, $A_1 = \frac{3}{c_2\sqrt{n}}$ and is tabulated together with c_2 for different values of n in statistical tables books and quality control tables.

On the other hand, if one uses the estimates (2.6) and (2.8), the chart for mean will be given by

$$\left.\begin{aligned} LCL &= \bar{x} - 3\frac{\bar{R}}{d_2\sqrt{n}} = \bar{x} + A_2\bar{R} \\ Central\ line &= \bar{x} \\ and\ UCL &= \bar{x} + 3\frac{\bar{R}}{d_2\sqrt{n}} = \bar{x} + A_2\bar{R} \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(2.10)$$

Where $A_2 = \frac{3}{d_2\sqrt{n}}$ and is again, given for different values of n in the statistical tables book and quality control tables.

13.4 Control Charts for Range (R)

Range is the simplest measure of dispersion, thus to check whether the variation in the variability of quality characteristic is within control or not may use control for s.d. or range. Control chart for range is discussed below:

13.4.1 Standard Given

For a normally distributed variable x,

$$\left.\begin{aligned} E(R) &= d_2\sigma \\ and\ \sigma_R &= D_\sigma \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.11)$$

Where both D and d_2 are the function of n.

When the standard value of σ is given by σ' then the chart for R will be;

$$\left.\begin{aligned} LCL &= d_2\sigma - 3D\sigma = D_1\sigma' \\ Central\ line &= d_2\sigma' \\ and\ UCL &= d_2\sigma + 3D\sigma = D_2\sigma' \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.12)$$

Where D_1 = (d_2-3D) and D_2 = (d_2+2D). The values of D_1, D_2 and d_2 are obtained for different values of n from the quality control tables.

13.4.2 Standards not Given

When the value of σ is not specific then we estimate it by $\frac{R}{d_2}$ and the chart for R will be based on

$$\left.\begin{array}{c} LCL = \bar{R} - 3\frac{D}{d_2}\bar{R} = D_3\bar{R} \\ Central\ line = \bar{R} \\ and\ UCL = \bar{R} + 3\frac{D}{d_2}\bar{R} = D_4\bar{R} \end{array}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. (2.13)$$

Where $D_2 = \left(1 - 3\frac{D}{d_2}\right)$ $and\ D_4 = \left(1 + 3\frac{D}{d_2}\right)$.The values of D_3 and D_4 are, available from the quality control table.

In either case if LCL comes out to be negative, it is taken to be zero Rang R, by its very nature can not be a negative quantity.

13.5 Control Charts for S.D. σ

As state above, S.D. is widely used measure of dispersion and thus to check whether S.D. of process is in control or not we may use this chart.

13.5.1 Standards Given

For normally distributed r.v. x., we know that

$$\left.\begin{array}{c} E(s) = C_2\sigma \\ and\ \sigma_s = \sigma\sqrt{\frac{n-1}{n} - c_2^2} \end{array}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.14)$$

If the standard value of σ is σ' is given then the chart will be based on

$$\left.\begin{array}{c} LCL = C_2\sigma' - 3\sigma'\sqrt{\frac{n-1}{n} - c_2^2} = B_1\sigma' \\ Central\ line = c_2\sigma' \\ and\ UCL = C_2\sigma' - 3\sigma'\sqrt{\frac{n-1}{n} - c_2^2} = B_2\sigma' \end{array}\right\} \ldots\ldots\ldots\ldots\ldots\ldots. (2.15)$$

Where

$$B_1 = C_2 - 3\sqrt{\frac{n-1}{n} - c_2^2} \; and \; B_2 = c_2 - 3\sqrt{\frac{n-1}{n} - c_2^2}$$

The values of B_1 and B_2 and for different values of n may be obtained from the quality control table.

13.5.2 Standard not Given

If the standard value of σ is not given the we use the estimates $\frac{\bar{s}}{c_2} for\ \sigma$ and the control chart will be based on

$$\left.\begin{aligned} LCL = \bar{s} - 3\frac{\bar{s}}{c_2}\sqrt{\frac{n-1}{n} - c_2^2} = B_3\bar{s} \\ Central\ line = \bar{s} \\ and\ UCL = \bar{s} + 3\frac{\bar{s}}{c_2}\sqrt{\frac{n-1}{n} - c_2^2} = B_4\bar{s} \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots..(2.16)$$

Where

$$B_3 = 1 - \frac{3}{c_2}\sqrt{\frac{n-1}{n} - c_2^2} \; and \; B_3 = 1 + \frac{3}{c_2}\sqrt{\frac{n-1}{n} - c_2^2}$$

The values of B_3 and B_4 for different values of n can be obtained from the quality control table.

Here also if LCL, to be comes out negative then it is taken as zero Since in no case s be a negative quality.

It should be noted that $\bar{X}, R\ and\ \sigma$ charts are all based on the assumption that the parent distribution of the quality characteristic is normal. Sometimes this assumption may not be valid. Sometimes it may be possible to use the appropriate non normal distribution to find exact limits.

13.7 Examples

Example 2.1: The following data gives readings for 10 samples of size 6 each in the production of a certain component.

Sample	1	2	3	4	5	6	7	8	9	10
Mean	383	508	505	582	557	337	514	614	707	753
Range	95	128	100	91	68	65	148	28	37	80

Draw control for $\bar{X}$. One can assume that all the samples are from homogeneous lot. (given for n= 6; $A_2 = 0.483$; $D_4 = 2.004$).

Solution:

Table: Calculation for Control Charts for $\bar{X}$

Sample No.	Sample Mean $\bar{X}$	Sample Range R
1	383	95
2	508	128
3	505	100
4	582	91
5	557	68
6	337	55
7	514	148
8	614	28
9	707	37
10	653	80
Total		

Control limits for $\bar{X}$ chart:

$$\bar{\bar{X}} = \frac{\sum \bar{X}}{n} = \frac{5460}{10} 546; \ \bar{R} = \frac{\sum R}{n} = \frac{840}{10} = 84$$

$$U.C.L. = \ \bar{\bar{X}} = A_2\bar{R} = 546 + 0.483 \ \times 84 = 546 + 40.57 = 586.57.$$

$$C.L. = \ \bar{\bar{X}} = 546$$

$$L.C.L. = \ \bar{\bar{X}} = A_2\bar{R} = 546 - 0.483 \ \times 84 = 546 - 40.57 = 505.43$$

The control chart for is given below:

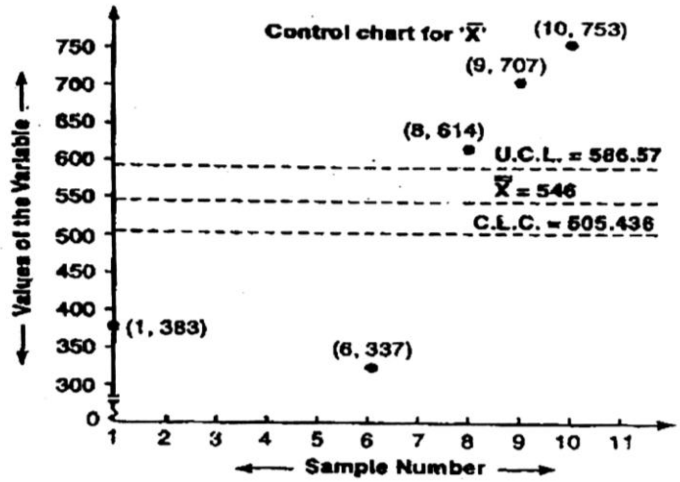


Fig. 2.1 Control Chart for $\overline{X}$

Since some of the sample points fall outside the control limits, so the process is out control.

Example 2.2: Construct a control chart for mean and the range for the following data on the basis of fuses, samples of 5 being taken every hour (each set of 5 has been arranged in ascending order of magnitude).

1	2	3	4	5	6	7	8	9	10	11	12
42	42	19	36	42	51	60	18	15	69	64	61
65	45	24	54	51	74	60	20	30	109	90	78
75	68	80	69	57	75	72	27	39	113	93	94
78	72	81	77	59	78	95	42	62	118	109	109
87	90	81	84	78	132	138	60	84	153	112	136

$[given\ for\ n = 5, A_2 = 0.577, D_3 = 0\ and\ D_4 = 2.115]$

Solution:

Table: Calculation for $\overline{X}$- chart and R- Chart

Sample No. (1)	Sample observations (2)					Total (3)	Sample Mean X (4)= (3)+5	Sample Range (R) (5)
1	42	65	75	78	87	347	69.4	45

2	42	45	68	72	90	317	63.4	48
3	19	24	80	81	81	285	57.0	62
4	36	54	69	77	84	320	64.0	48
5	42	51	57	59	78	287	57.4	36
6	51	74	75	78	132	410	82.0	81
7	60	60	72	95	138	425	85.0	78
8	18	20	27	42	60	167	33.4	42
9	15	30	39	62	84	230	46.0	69
10	69	109	113	118	153	562	112.4	84
11	64	90	93	109	112	468	93.6	48
12	61	78	94	109	136	478	95.6	75
						Total	$\sum \bar{X} = 859.2$	$\sum R = 716$

(i) **Mean of the Sample Means:**

$$\bar{X} = \frac{\sum \bar{X}}{N} = \frac{859.2}{12} = 17.6 \left[\sum \bar{X} = 859.2\,, N = 12\right]$$

(ii) **Value of R:** It is computed from the values of R shown in column (5). For example , the value of R for the first sample is computed as follows;

R = 87-42 = 45.

(iii) **The Mean of Sample Ranges R is**

$$R = \frac{\sum R}{N} = \frac{716}{12} = 59.66 \qquad [\,N = 12]$$

We are given that for n =5, A_2 = 0.577, D_3 = 0 and D_4 = 2.115

Control Limits for X Chart.

$$U.C.L. = \bar{\bar{X}} = A_2\bar{R} = 71.6 + 0.577\,(59.66) = 71.6 + 34.4238 = 106.024$$

$$C.L. = \bar{\bar{X}} = 71.6$$

$$L.C.L. = \bar{\bar{X}} = A_2\bar{R} = 71.6 - 0.577(59.66) = 71.6 - 34.4238 = 37.1766$$

Since the sample point (means) corresponding to sample number 10 lie outside the control. This suggest the presence of some assignable causes of variations which must be detected and corrected.

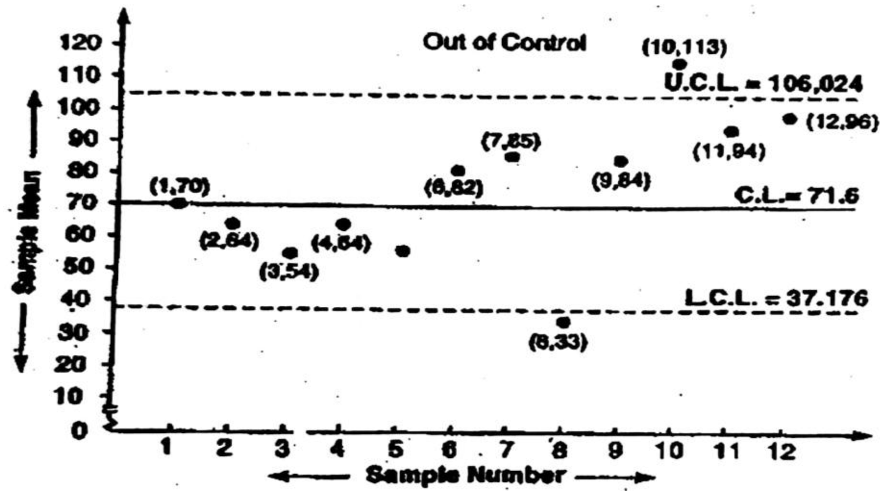


Fig. 2.2 Control Chart for $\overline{\overline{X}}$

Control limits for R-Chart:

$U.C.L. = D_4 R = 2.115 \times 59.66 = 126.18.$

$U.C.L. = D_4 R = 0.$

$C.L. = \ R = 59..66$

The control chart for R is given below:

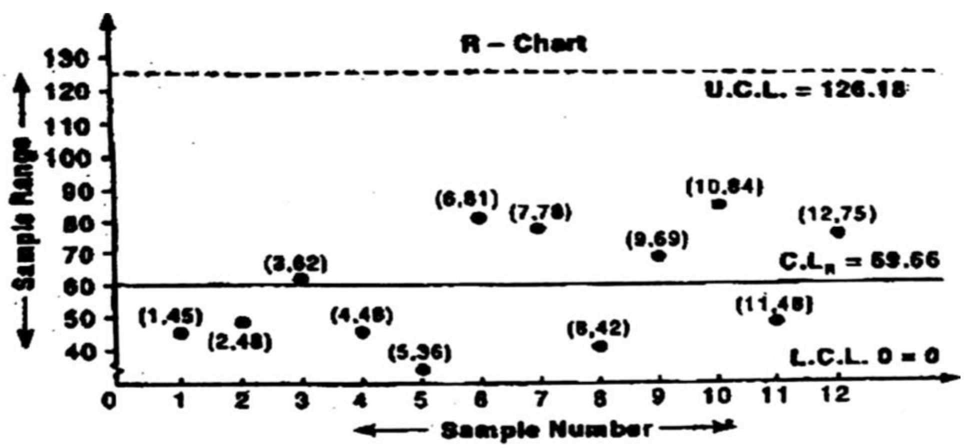


Fig. 2.2 Control Chart for R

This chart shows that the process of variability is under control all the sample points (Ranges) are lying within the control limits.

Example 2.3: From the following data calculate the control limits for the mean and range control charts.

Sample	Dimensions in inches Units of samples

	1	2	3	4	5
1	50	55	52	49	54
2	51	50	53	50	46
3	50	53	48	52	47
4	48	53	50	51	53
5	46	50	44	48	47
6	55	51	56	47	51
7	45	48	53	48	51
8	50	56	54	53	57
9	47	53	49	52	54
10	56	53	55	54	52

[For a sample of size 5: A_2D 0.577, D_4= 2.115 and D_3 = 0]

Solution;

Calculation for X and R-Chart

Sample No. (1)	Sample observations (2)					Total (3)	Sample Mean X (4)= (3)+5	Sample Range (R) (5)
1	50	55	52	49	54	260	52	6
2	51	50	53	50	46	250	50	7
3	50	53	48	52	47	250	50	6
4	48	53	50	51	53	250	51	5
5	46	50	44	48	47	255	47	4
6	55	51	56	47	51	260	52	9
7	45	48	53	48	51	245	49	8
8	50	56	54	53	57	270	54	7
9	47	53	49	52	54	255	51	7
10	56	53	55	54	52	280	56	4
						Total	$\sum X = 859.2$	$\sum R = 716$

1. **Control Chart for Mean (X):**

Central Line : $\bar{X} = \frac{\sum \bar{X}}{N} = \frac{512}{10} = 51.2$, $\bar{R} = \frac{\sum R}{N} = \frac{63}{10} = 6.3$

$$U.C.L. = \bar{\bar{X}} = A_2\bar{R} = 51.2 + 0.577\ (6.3) = 51.2 + 3.65 = 54.84$$

$$L.C.L. = \bar{\bar{X}} = A_2\bar{R} = 51.2 - 3.64 = 47.56$$

Graph of $\overline{X}$ Chart

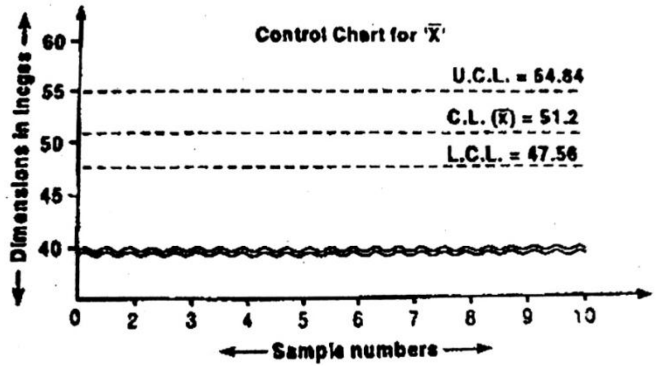


Fig. 2.4 Control Chart for Mean

2. **CONTROL CHART FOR R**
   Central Line : R = 6.3
   U.C.L. = D_4 R = 2.115 x 6.3 = 13.32
   L.C.L. = D_3 R = 0 X 6.3 = 0.

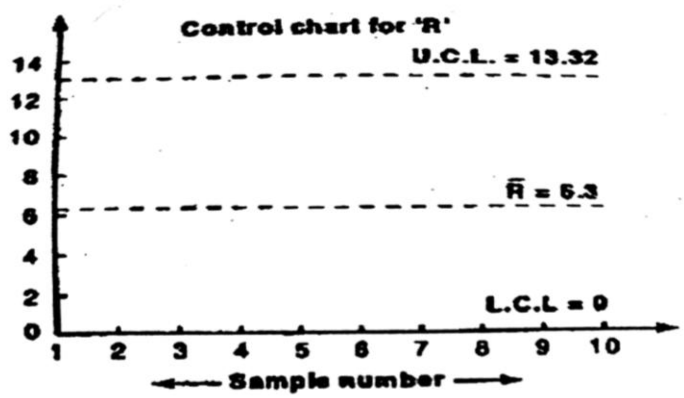


Fig. 2.5 Control Chart for Range

Since the sample mean for sample number 5 and 10 falls outside control limit and a number of sample ranges falls outside the control limit for range we may conclude that process is out of control.

Example 2.4: Ten samples each of size 5 are draw at regular intervals from a manufacturing process. The sample mean (X) and their range (R) are given below:

Sample No.	1	2	3	4	5	6	7	8	9	10	Total
Mean (X)	49	45	48	53	39	47	46	39	51	45	462
Range (R)	7	5	7	9	5	8	8	6	7	6	68

Calculate the control limits for X of chart and R chart. Also comment on the state of control.

[Given: $A_2 = 0.58$, $D_3 = 2.115$]

Solution: Here

$$\bar{X} = \frac{\sum \bar{X}}{N} = \frac{462}{10} = 46.2\,, \qquad \bar{R} = \frac{\sum R}{N} = \frac{68}{10} = 6.8$$

For X- chart:

$$U.C.L. = \bar{\bar{X}} = A_2\bar{R} = 46.2 + 0.58\ (6.8) = 50.114$$

$$C.L. = X = 46.2$$

$$L.C.L. = \bar{\bar{X}} = A_2\bar{R} = 46.2 - 0.58\ (6.8) = 14.382$$

For R- Chart

Central Line : R = 6.3
U.C.L. = D_4 R = (2.115) (6.8) = 14.382
C.L. = R = 6.8
L.C.L. = D_3 R = (0) (6.8) = 0.

The process is not in a state of control as the sample points 4 and 9 are above U.C.L. an sample points 5 and 8 are below L.C.L as seen on the basis of limits for X chart. (show that control charts of graph paper)

Thus we conclude that the process is not in control for sample mean and hence reason for rectification may be find out.

Example 2.5: the following are sample means and ranges for 10 samples, each of size 5. Calculate the control limits for X chart and R chart and state whether the process is n a control or not. (given $A_2 = 0.577$, $D_3 = 0$, $D_4 = 2.115$)

Sample No.	1	2	3	4	5	6	7	8	9	10
Mean	5.10	4.98	5.02	4.96	5.04	4.94	4.92	4.92	4.92	4.98
Range	0.3	0.4	0.2	0.4	0.1	0.1	0.8	0.5	0.3	0.5

Solution:

Table: Computation of Control Limits for $\bar{\bar{X}}$ and R chart

Sample	Mean	Range
1	5.10	0.3

2	4.98	0.4
3	5.02	0.2
4	4.96	0.4
5	5.04	0.1
6	4.94	0.1
7	4.92	0.8
8	4.92	0.5
9	4.92	0.3
10	4.98	0.5
Total	49.82	3.6

$$Here\ ,\ \ \bar{\bar{X}} = \frac{\sum \bar{X}}{10} = \frac{49.82}{10} = 4.982;\ \bar{R} = \frac{\sum R}{N} = \frac{3.6}{10} = 0.36$$

Control Limits For X- chart:

$$U.C.L. = \ \bar{\bar{X}} = A_2\bar{R} = 4.982 + 0.577\ (0.36) = 4.982 + 0.208 = 5.19$$

$$C.L. = \ X = 4.982$$

$$L.C.L. = \ \bar{\bar{X}} = A_2\bar{R} = 4.982 - 0.577\ (0.36) = 4.982 - 0.208 = 4.77$$

Control Limits for R- Chart

U.C.L. = D_4 R = (2.11) (0.36) = 0.7614
C.L. = R = 0.36
L.C.L. = D_3 R = (0) (0.36) = 0.

Since one of the points is lying outside the control limits, so the process is said to be out of control. (Show the control limits and sample points on graph paper)

13.7 Exercises

2.1 Draw the mean and range charts from the follow data and state your conclusion.

Sample No.	1	2	3	4	5	6	7	8	9	10
Sample Mean (grams)	12.8	13.1	13.5	12.9	13.2	14.1	12.1	15.5	13.9	14.2
Sample Range (grams)	2.1	3.1	3.9	2.1	1.9	3.0	2.5	2.8	2.5	2.0

[Given. n=5, A_2 = 0.577, D_3= 0, D_4= 2.115]

2.2 A machine is set to deliver packets of a given weight. 10 samples of size 5 each were recorded. Below are given relevant data.

Sample No.	1	2	3	4	5	6	7	8	9	10
Mean	15	17	15	18	17	14	18	15	17	16
Range (R)	7	7	4	9	8	7	12	4	11	5

Calculate the value for the central line and the control limits for the mean chart and the range chart and then comment on the state of control. [Conversion Factors for n=5, are A_2= 0.58, D_3 = 0, D_4 = 2.11]

2.3 Thirty samples of 5 items each were taken from the output of a machine and a critical dimension measured. The mean of 30 samples 0 was 0.6550 inch and the average R of the 30 samples 0.0036 inch. Compute the upper and lower control limits for X and R charts, (For n=5, A_2= 0.57, D_3 = 0, D_4 = 2.115)

2.4 From a factory producing metal sheets, a sample of 5 sheets is taken every hour and the data is obtained as under. Draw a control chart for the mean and examine whether the process is under control or not.

Sample No.	1	2	3	4	5	6	7	8	9	10
Mean thickness of sheet	0.025	0.032	0.042	0.022	0.028	0.010	0.025	0.040	0.026	0.029
Sample Range	0.025	0.048	0.012	0.012	0.019	0.010	0.006	0.046	0.010	0.032

(you are given for n=5, A= 0.58)

Answer

2.1: C.L. = 5.4; U.C.L. = 15.02; L.C.L. : 12.04.

2.2: For mean chart : C.L. = 16.2; U.C.L. = 20.56; L.C.L. : 11.9 For R-Chart: C.L. = 7.4; U.C.L. = 15.6; L.C.L. : 0. The production process is in control .

2.3: For X chart: U.C.L. = 6.6571; L.C.L. : 0.6529 For R-Chart: U.C.L.= 0.0076; L.C.L.: 0.

2.4: C.L. = 0.028; U.C.L. = 0.041; L.C.L. : 0.015

13.8 Summary

If the quality of product is numerically measurable, it can be considered distributed normally (justified by central limit theorem). Under this assumption control charts for verifying the control of mean and dispersion can be developed. To check the process control, control chart for mean and range or mean and standard deviation can be prepared. 3σ limits can be obtained and if all the sample values lie with in the control limits without following any pattern, it may be concluded that process is in control.

13.9 Further Reading

- Burr, I.W. Engineering Statistics and Quality Control Mcgraw Hill.
- Cowden, D.J. Statistical Methods in Quality Control, Asia Publishing House.
- Grant, E.L.: Statistical Quality Control, Part I-IV, McGraw Hill.

Unit-14: Control Charts for Attributes: p-Chart, np-Chart and c-Chart

Structure

14.1 Introduction

If the quality characteristic is an attribute, and each item is recorded as either defective or non-defective, then to know whether the process is in control or not one has to ascertain whether the population fraction defective P is the same for all sub-groups. The judgment may be based either on the number of defectives in the sample or on the fraction defective in the sample.

14.2 Objectives

After reading this units you should be able to understand:

- The preparation of control charts for fraction defectives, number of defectives and number of defect.
- How to prepare control charts when standards are given or not given

14.3 Control Charts for Fraction Defectives (p-chart)

In this unit, we will study how to prepare control chart for fraction defectives.

14.3.1 Standard (Specification) Given

For construction of a control chart for p, we have the relation.

$$E(p) = p$$

$$and\ \sigma_p = \sqrt{P(1-P)/n} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.1)$$

Where p is sample fraction defective and P is population fraction defective.

If 'P' is the specified standard value for P then the chart will be

$$\left.\begin{array}{c} LCL = P' - 3\sqrt{\dfrac{P'(1-P')}{n}} = P' - A\sqrt{P'(1-P')} \\ Central\ line = p'\ and \\ UCL = P' + 3\sqrt{\dfrac{P'(1-P')}{n}} = P' - A\sqrt{P'(1-P')} \end{array}\right\} \ldots\ldots\ldots\ldots\ldots. (3.2)$$

Where A = $3/\sqrt{n}$

14.3.2 Standard not Given

In this case a common value for P will be estimated by $\bar{p} = \frac{\sum pi}{n}$ and the control limits will be

$$\left.\begin{array}{c} LCL = \bar{p} - A\sqrt{\dfrac{P'(1-P')}{n}} = P' - A\sqrt{P'(1-P')} \\ central\ line = \bar{p} \\ and\ UCL = \bar{p} + A\sqrt{\bar{p}(1-\bar{p})} \end{array}\right\} \ldots\ldots\ldots\ldots\ldots (3.3)$$

Here it should be noted that P can never be negative. Hence if LCL comes out to be negative then it is taken to be zero.

14.4 Control Charts for Number of Defectives (np-chart)

Now we study the how to prepare control charts for number of defectives:

14.4.1 Standard Given

Let us suppose that each random sample is taken with replacements (for even if taken without replacements practically from a infinite population. Further suppose that d=np where d is the number of defectives. Then d…….has binomial distribution and

$$\left.\begin{aligned} E(d) = E(np) = np \\ and\ \sigma_d = \sigma\sqrt{np(1-P)} \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(3.4)$$

P being the same for all sub-groups if an only if the process is in control.

If P' be the specified standard value of P, then the control chart will be based on

$$\left.\begin{aligned} LCL = nP - 3\sqrt{nP'(1-P')} \\ Central\ Line = nP' \\ and\ UCL = nP' + 3\sqrt{nP'(1-P')} \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots.(3.5)$$

14.4.2 Standard not Given

If the standard value for P, is not specified we have to estimate it from the samples themselves. An appropriate estimate for P will be

$$\bar{p} = \sum_i p_i/m.$$

And the control limits for np will be

$$\left.\begin{aligned} LCL = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})} \\ Central\ Line = n\bar{p} \\ and\ UCL = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})} \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots.(3.6)$$

Here too it should be noted that np can never be negative. Hence if LCL, comes out to be negative, then it be taken as zero.

14.4.3 Control Charts for Percent Defective

Sometimes we need control chart for percent defective. In this case we construct a control for 100p instead of p. The control lines for a percent defective chart can be written as follows:

Case 1: Standard given

$$\left.\begin{array}{c} LCL = 100p' - 100A\sqrt{p'(1-p')} \\ Central\ Line = 100p' \\ and\ UCL = 100p' - 100A\sqrt{p'(1-p')} \end{array}\right\} \ldots\ldots\ldots\ldots\ldots\ldots (3.7)$$

Case 2: Standard not given

$$\left.\begin{array}{c} LCL = 100\bar{p} - 100A\sqrt{\bar{p}(1-\bar{p})} \\ Central\ Line = 100\bar{p} \\ and\ UCL = 100p' - 100A\sqrt{\bar{p}(1-\bar{p})} \end{array}\right\} \ldots\ldots\ldots\ldots\ldots\ldots (3.8)$$

A p-chart (or np-chart or 100p-chart) is advantages because it may be used even for characters that are observed as variables.

14.5 Control Charts for Number of Defects (c-chart)

Here we are concerned with cases where each items is observed for the number of defects. A clear cut distribution between a defective and a defect is that a defective is an item that fails to fulfill one or more of the given specifications, a defect is any instance of the item's lack of conformity of specifications. Thus every defective item possesses one or more defects. The defects may be the surface defects in a roll of paper or photographic film.

It is obvious that in manufactured articles, the chances for defects to occur are numerous, even though the probability for a defect to occur in any one spot is rare. Hence the number of defects (c) may, in most cases, be supposed to follow the Poisson distribution with λ as its parameter, say.

Thus, the aim of a control chart for c to detect any differences that may exist among the Poisson distribution for the different groups or in other words among the λ values for the sub-groups.

14.5.1 Standard Given

For a Poisson variable c with parameter λ we have

$$\left.\begin{array}{c} E(c) = \lambda \\ and\ \sigma_c = \sqrt{\lambda} \end{array}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.10)$$

If a standard value for c is given as c' then the control chart for c wilt be:

$$\left.\begin{aligned} LCL = c^{'} - 3\sqrt{c^{'}} \\ Central\ Line = c^{'} \\ and\ UCL = c^{'} + 3\sqrt{c'} \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(3.11)$$

14.5.2 Standard not Given

In case no standard value for c is specified then we will have to estimate it from observed c values. Suppose c_1, c_2, …………..c_m are the values of c taken from the sample of m-groups, then the appropriate estimate of c will be

$$\bar{c} = \sum_i \frac{c_1}{m} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(3.12)$$

And then the lines for the c-chart, will be

$$\left.\begin{aligned} LCL = \bar{c} - 3\sqrt{\bar{c}} \\ Central\ Line = \bar{c} \\ and\ UCL = \bar{c} + 3\sqrt{\bar{c}} \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(3.13)$$

It be noted that c cannot be negative. Hence if LCL, come out to be negative, then it be taken equal to zero.

However, in case the different sub-groups are not of constant size. In that case, we obtain the number of defects per unit, i.e., u=c/n. Here also the limit line will not be constant, but will vary with the sub-group size n. Thus for a given specified value u' the control limits will be

$$\left.\begin{aligned} LCL = u^{'} - 3\sqrt{\frac{u^{'}}{n_i}} \\ central\ line = u^{'} \\ and\ UCL = u^{'} + 3\sqrt{\frac{u^{'}}{n_i}} \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3.14)$$

If u' is not specified then u is estimated by

$$\bar{u} = \sum_{i=1}^{m} u_i / \sum_{i=1}^{m} n_1$$

Where u_i, n_i are, respectively the number of defects and the sample size for the ith sub-group. Now we shall get the lines for the u-chart, as:

$$\left.\begin{aligned} LCL &= u' - 3\sqrt{\frac{\bar{u}}{n_i}} \\ central\ line &= \bar{u} \\ and\ UCL &= \bar{u} + 3\sqrt{\frac{\bar{u}}{n_i}} \end{aligned}\right\} \dots\dots\dots\dots\dots\dots\dots\dots (3.15)$$

14.6 Examples

Example 3.1: The following figures give the number of defectives in 20 samples, containing 2000 items.

425	430	216	341	225	322	280	306	337	305
356	402	216	264	126	409	193	280	326	389

Calculate the values for central line and control limits for p-chart

Solution: Total numbers of items in 20 samples = N= 20x2000 = 40,000.

Total number of defectives in 20 samples $\sum d = \{425 + 430 + 216 + 341 + 225 + 322 + 280 + 306 + 337 + 305 + 356 + 402 + 216 + 264 + 126 + 409 + 193 + 280 + 326 + 389 = 6{,}148\}$.

$$\therefore p = \frac{Number\ of\ defectives\ in\ all\ samples\ combined}{total\ number\ of\ items\ in\ all\ the\ sample\ combined} = \frac{\sum d}{N}$$

Mean Fraction defective: $\bar{p} = \frac{\sum d}{N} = \frac{6{,}148}{40{,}000} = 0.1537$

∴Control limits for p-chart

Central line = C.L. =$\bar{p}$ = 0.1537.

Upper Control Limit

$$U.C.L. = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.1537 + 3\sqrt{\frac{0.1537(1-0.1537)}{200}}$$

$$= 0.2537 + 3 \times 0.008 = 0.1537 + 0.02418 = 0.12952.$$

Example 3.2: The following data refers to visual defects found during the inspection of the first 10 samples of size 50 each from a lot of two wheelers manufactured by an automobile company:

Sample No:	1	2	3	4	5	6	7	8	9	10
No. of Defectives:	4	3	2	3	4	4	4	1	3	2

Draw p-chart to show that the fraction defectives are under control.

Solution:

Sample No.	No. of defectives	Fraction defectives (p)
1	4	4/50=0.08
2	3	3/50=0.06
3	2	2/50=0.04
4	3	3/50=0.06
5	4	4/50=0.08
6	4	1/50=0.08
7	4	3/50=0.08
8	1	2/50=0.06
9	3	3/50= 0.06
10	2	2/50= 0.04
	$\sum d$ =30	$\sum. d$ = 0.60

Here $\bar{p} = \frac{\sum d}{N} = \frac{0.60}{10} = 0.60, n = 50.$

$$U.C.L. = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.06 + 3\sqrt{\frac{0.06(1-0.06)}{50}} = 0.06 + 0.1008 = 0.1608$$

$$C.L. = 0.06$$

$$L.C.L. = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.06 - 0.1008 = -0.0408 \; or \; 0.$$

As fraction defectives cannot be negative so L.C.L. is taken as zero.

As all the sample point are lying within the L.C.L. (=0) and U.C.L. (=0.1608), so the process is in control.

Example 3.3 A random sample of 100 TV tubes was taken from daily production of large output of pens and the number of defective tubes was noted. On the information given prepare a p-chart and state your conclusions.

Sample No.	No. of Defective	Sample No.	No. of Defective	Sample No.	No. of Defective	Sample No.	No. of Defective
1	4	11	4	6	2	16	5
2	6	12	5	7	0	17	7
3	10	13	3	8	1	18	13
4	8	14	0	9	3	19	3
5	3	15	14	10	5	20	4

Solution:

Sample No.	No. of Defective	Fraction (p) defectives	Sample No.	No. of Defective	Fraction (p) defectives
1	4	0.04	11	4	0.04
2	6	0.06	12	5	0.05
3	10	0.10	13	3	0.03
4	8	0.08	14	0	0.00
5	3	0.03	15	14	0.14
6	2	0.02	16	5	0.05
7	0	0.00	17	7	0.07
8	1	0.01	18	13	0.13
9	3	0.03	19	3	0.03
10	5	0.05	20	4	0.04
Number of defectives in all samples combined				100	$\sum P$=1.00

$$\therefore p = \frac{Number\ of\ defectives\ in\ all\ samples\ combined}{total\ number\ of\ items\ in\ all\ the\ sample\ combined} = \frac{100}{20 \times 100} = 0.05$$

∴U.C.L. = $\bar{p} + 3\frac{\sqrt{\bar{p}(1-\bar{p})}}{n} = 0.05 + 3\frac{\sqrt{0.05(1-0.05)}}{100} = 0.05 + 3 \times 0.022 = 0.05 + 0.066 = 0.116$

C.L. = 0.05

L.C.L.= $\bar{p} - 3\frac{\sqrt{\bar{p}(1-\bar{p})}}{n} = 0.05 - 3 \times 0.022 = -0.016\ or\ 0.$

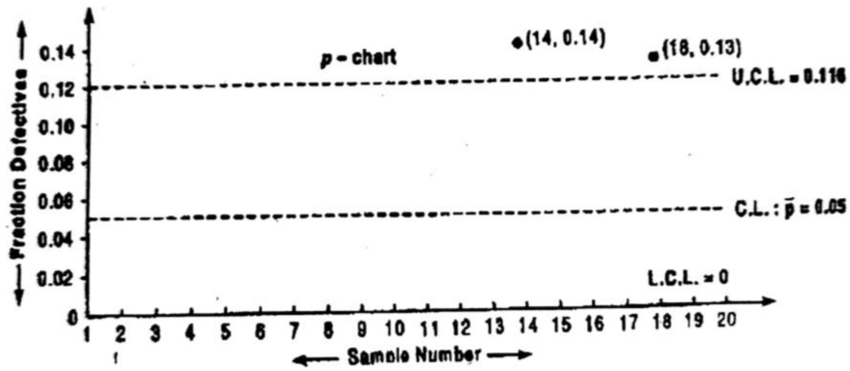


Fig. 3.1 p-Chart

From the above chart, it is clear that two points (viz. (14, 0.14) and (18, 0.13) lie outside the control limits. Hence the process seems to be out of control.

Example 3.4: The following data refer to visual defects found at inspection of the first 10 samples of size 100. Use the data 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

Solution. There are 20 defective items in 10 samples each of size 100, the np chart is the proper chart for the given data.

$$\bar{p} = Average\ fraction\ defecive = \frac{20}{10 \times 100} = 0.02$$

$$also, n = 100, \qquad \therefore n\bar{p} = 100 \times 0.02 = 2$$

$$and \sqrt{n\bar{p}(1-\bar{p})} = \sqrt{100 \times 0.02 \times 0.98} = \sqrt{1.96} = 1.4$$

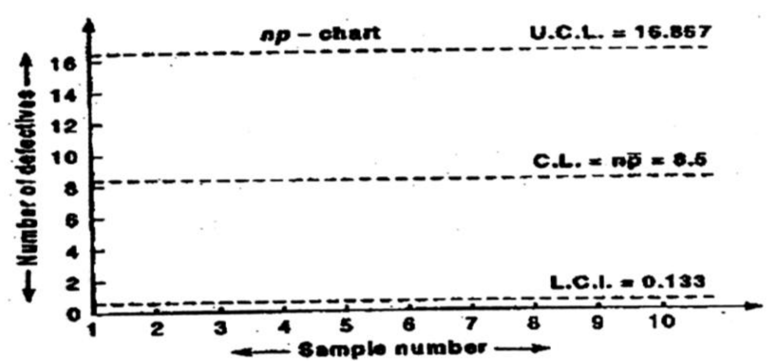


Fig. 3.2 : np - Chart

$$\therefore U.C.L. = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})} = 2 + 3(3 \times 1.4) = 6.2$$

$$C.L. = n\bar{p} = 2$$

$$L.C.L. = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})} = 2 - 3(1.4) = 0$$

Since all the points are within the control limits. The process is said to be in a state of control.

Example 3.5: The following data refer to the number of defective in 10 samples of 100 items each. Construct an appropriate control chart and interpret the control limits:

Sample No.	1	2	3	4	5	6	7	8	9	10
No. of defectives	4	8	11	3	11	7	7	16	12	6

Solution:

Np-chart shall be appropriate control chart for the given data.

Total number of defectives = [4+8+11+3+11+7+7+16+12+6]=85.

There are 85 defectives in 10 samples of 100 items each.

$$\bar{p} = Average\ fraction\ defecive = \frac{85}{10 \times 100} = 0.085$$

$$\therefore n\bar{p} = 100 \times 0.085 = 8.5$$

$$\therefore U.C.L. = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})} = 8.5 + 3(1 - 0.085) = 16.867$$

$$C.L. = n\bar{p} = 8.5$$

$$L.C.L. = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})} = 8.5 - 3 \times 2.789 = 8.5 - 8.367 = 0.133$$

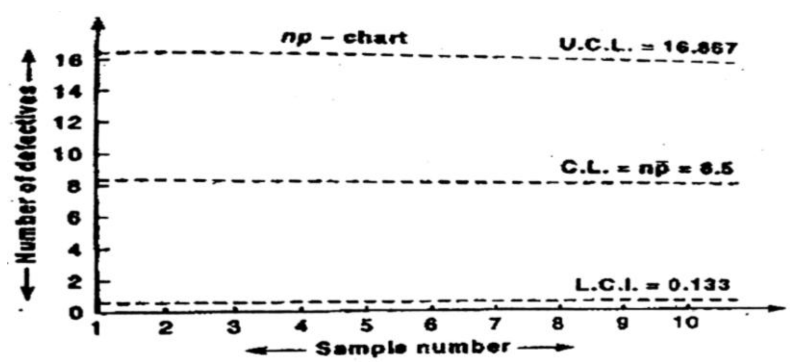


Fig. 3.3 : np - Chart

Since none of the sample points fall outside the control limits, so the process in under statistical control.

Example 3.6: The following figures give the number of defects found in 1000 items of cotton piece goods inspected every day in a certain month:

1,1,3,7,8,1,2,6,1,1,10,5,0,19,16,20,1,6,12,4,5,1,8,7,9,2,3,14,6,8 Do these data come from a controlled process?

Solution: The number of defects = $\sum c$ = 1+1+3+………14+6+8= 187; N= 30.

$$\bar{c} = \frac{\sum c}{N} = \frac{187}{30} = 6.23.$$

$$\therefore U.C.L. = \bar{c} + 3\sqrt{\bar{c}} = 6.23 + 3\sqrt{6.23} = 6.23 + 3 \times 2.496 = 13.711$$

$$C.L. = 6.23$$

$$L.C.L. = \bar{c} - 3\sqrt{\bar{c}} = 6.23 - 3\sqrt{6.23} = 6.23 - 3 \times 2.496 = -1.258 = 0.$$

[Since L.C.L. cannot be negative, so its value would be 0]. Since 4 ample points lie outside the control limits, the given data cannot be said to have come from a controlled process.

Example 3.7: Assume that 20 milk bottles are selected at random from a process. The number of air bubbles (defects) observed from the bottles is given in the following table:

[c = No. of Air bubbles (defects) in each bottle].

Bottle Number (Sample order)	Defects (c)	Bottle Number (Sample order)	Defects (c)
1	4	11	3
2	5	12	5
3	7	13	4
4	3	14	3
5	3	15	4
6	5	16	5
7	6	17	3
8	2	18	7
9	4	19	6
10	8	20	13
		Total no. of defects	$\sum c$=100

Draw a control chart for above data.

Solution:

Here will use the c-chart.

$$\bar{c} = average\ number\ of\ defects: \bar{c} = \frac{\sum c}{N} = \frac{100}{20} = 5$$

$$\therefore U.C.L. = \bar{c} + 3\sqrt{\bar{c}} = 5 + 3\sqrt{5} = 5 + 3 \times 2.236 = 5 + 6.708 = 11.708$$

$$L.C.L. = \bar{c} - 3\sqrt{\bar{c}} = 5 - 3\sqrt{5} = 5 - 3 \times 2.236 = -1.708 = 0$$

[as L.C.L. cannot be negative, so L.C.L. = 0]

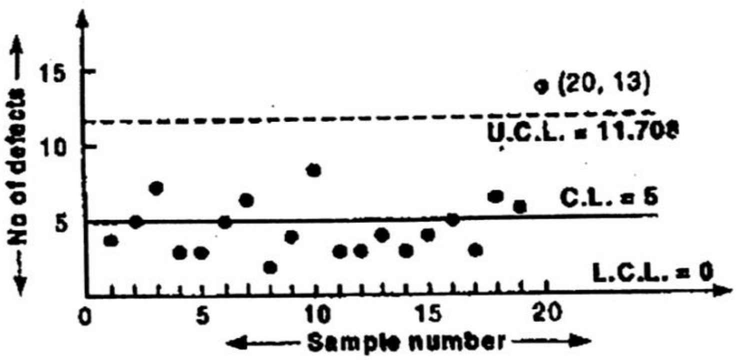


Fig. 3.4 : c - Chart

It is clear from the chart that only one point (20.13) falls outside the control limits and this is to be treated as a danger signal.

14.7 Exercises

3.1 Based on 15 sub groups each of size 200 taken at intervals of 45 minutes from a manufacturing process, the average fraction defective was found to be 0.068. Calculate the value of central line and the control limits for a p-chart.

3.2 A manufacturer of transistors found the following number of defective in 25 subgroups of 50 transistors:

3	4	4	2	3	2	7	0	2	4	2	3	4
1	2	4	8	2	4	2	6	4	3	1	4	

Construct a control chart for the fraction defective, plot the sample data on the chart and comment on the state of control.

3.3 Twenty pieces of cloth out of different rolls contained respectively 1,4,3,2,5,4,6,7,2,3,2,5,7,6,4,5,2,1,3 and 8 imperfections. Ascertain whether the process is in a state of statistical control.

3.4 The following table gives the number of defects observed in 8 woolen carpets passing as satisfactory. Construct the control chart for the number of defects:

Serial no. of Carpets	1	2	3	4	5	6	7	8
No. of defects	2	5	5	6	1	5	1	7

3.5 The following table gives the results of inspection of 10 pieces of woolen goods:

Piece no.	1	2	3	4	5	6	7	8	9	10
No. of defects	4	3	6	3	0	1	3	5	7	8

Calculate the control limits for a control chart for defects. State whether the process is in control or not.

3.6 During an examination of 10 pieces of equal length of cloth the following are the number of defects observed: 2,3,4,0,5,6,7,4,3,2. Draw a control chart for the number of defects and comment whether the process is under control or not.

Answers

3.1 C.L. = 0.068; U.C.L. = 0.1214; L.C.L. = 0.0145

3.2 C.L. = 0.066; U.C.L. = 0.171; L.C.L. = 0

3.3 c=4 U.C.L.e = 10; L.C.L. = 0; The process is under control.

3.4 C.L. = 4; U.C.L. = 10; L.C.L. = 0

3.5 U.C.L.c = 10; L.C.L. = 0; The process is under control.

3.6 C.L.c=3.6; U.C.L.c = 9.3; L.C.L.c = 0; The process is under control.

14.8 Summary

If the characteristics of quality of product is an attribute we use either p-chart or np-char or c-chart. If we wish to check whether process is in control based on the information that an item maybe defective or non-defective, we use p-chart. However if the information is available as number of defects per unit we use c-chart.

14.8 Further Readings

- Burr, I.W. Engineering Statistics and Quality Control. Mcgraw Hill.
- Cowden, D.J. Statistical Methods in Quality Control, Asia publishing House
- Grant, E.L.: Statistical Quality Control, Part I-IV, McGraw Hill.

Unit-15 Principles of Acceptance Sampling

Structure

15.1 Introduction

We have seen how the control charts enable a production process to be kept in control. But the process control does not imply lot control that is all the lots produced are good. This means that inspection of lots is required. We resort to sampling inspection, which is a procedure to determine whether a lot of manufactured items should be accepted or rejected on the base of the information supplied by random samples draw from the lot under consideration. It is also called ‘acceptance sampling’.

We usually deal with sampling inspection for attributes i.e. the items are judged good or bad by inspection and the lot quality is judged by the sample fraction defective.

A sampling plan specifies the procedure for deciding when the lot under inspection is to be rejected or accepted. Usually corrective action is taken when the lot is rejected –it is inspected fully and al its defective items are replaced by good ones. This is known as rectifying inspection. Suppose the incoming lots have fraction defective p_o, then the outgoing lots, after rectifying inspection will have fraction defective $p_1 > p_o$

Before describing sampling plans particular, we will introduce some terms relating to a sampling plans.

15.2 Objectives

After reading this section you must be able to understand

- Acceptance Sampling
- Lot Acceptance
- Producer's Risk
- Consumer's Risk
- Operating Characteristic of Single Sampling Plan
- ASN of Single Sampling Plan
- AQL., ALPD, AOQL
- Double Sampling Plans

15.3 Acceptable Quality Level (AQL)

This represents the poorest level of quality of the items produce which the consumer would consider to be acceptable as a process average. It is denoted by p.

15.4 Lot Tolerance Proportion Defective (LTPD)

This represents the poorest level of quality that the consumer is willing to accept in an individual lot. The consumer will not accept lots having proportion defective more than LTPD. It is denoted by p_t and $\bar{p} < p_t$

15.5 Producer's Risk

Suppose that the producer claims that he has standardized the quality of product at a level of fraction defective $\bar{p}$ (the producers process average). The probability of rejecting a lot under the sampling inspecting plan when the fraction defective is actually $\bar{p}$ is called producer's risk denoted by .

i.e.,

Consumer's Risk = α

P= (rejecting the lot of acceptable quality, $\bar{p}$)

15.6 Consumer's Risk

The consumer has to face the risk of accepting a lot of unsatisfactory quality on the basis of sampling inspection the probability of accepting a lot with fraction defective equal to LTPD, under a sampling plan is called the consumer's risk denoted by β .

i.e.,

Consumer's Risk = β

P= (accepting the lot of acceptable quality P_t)

15.7 O C Function

This gives the probability of accepting the lot, as a function of the lot fraction defective.

i.e.,

$$P_n(p) = L(p) = P\ (accepting\ the\ lot\ of\ quality\ p)$$

= P {lot is accepted/ (lot contains fraction defectives p)}

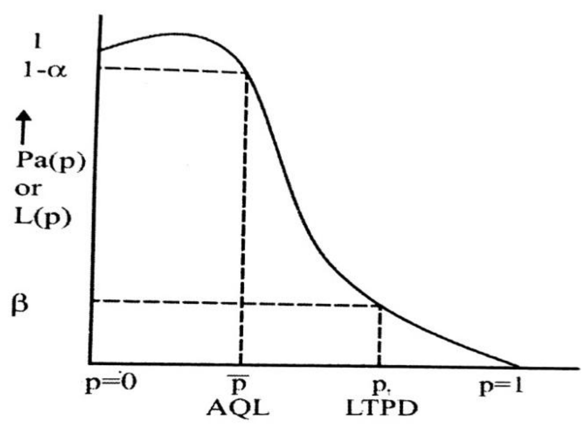
Also

$$\alpha = 1 - L(\bar{p})$$

$$or\ [L(\bar{p}) = 1 - \alpha$$

$$and$$

$$[\beta = L(p_t)$$



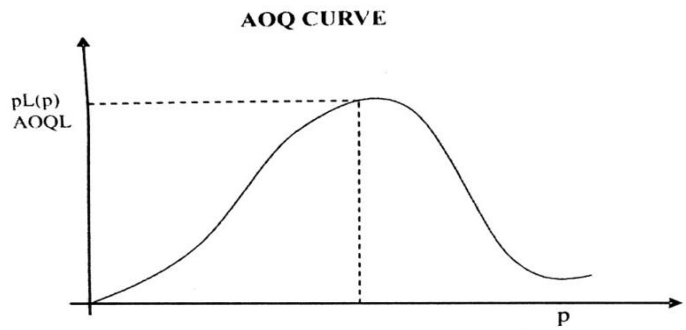
It is evident that the points on the OC curve corresponding to LTDP and AQL are β and $1 - \alpha$ respectively.

15.8 Average Outgoing Quality (AOQ)

The expected fraction defective remaining in the lot after the application of the sampling plan is called the average outgoing quality (AOQ). It is implied that rectifying inspection has been adopted. It is evident that AOQ will be the function of fraction defective in the lot.

$$\text{i.e. } AOQ \sim p\, L(p) = p\, P_a(p)$$

The maximum value of the AOQ function is known as the average outgoing quality limit AOQL. No matter how high the fraction defectives are in the incoming lots, the out going lots will never have worse quality level, on the average than AOQL.



15.9 Average Total Inspection (ATI)

It represents the average total amount of inspection per lot, including the sampling inspection or sorting. This is made up of two parts

i) The ASN of the plan and

ii) The cent percent inspection of the remainder of rejection lot

15.10 Average Sample Number (ASN) Function

This gives average (or expected) number of items inspected before a decision, regarding acceptance or rejecting, on the lot could be reached.

Now we are in a position to discuss some specific sampling plans, i.e., single and double sampling plans.

Suppose a lot of size N, having ∞ defective items is submitted for inspection. Let lot fraction defection be p = D/N.

15.11 Single Sample Plan

A single sampling plan is defined by two parameters, n- the sample size and c – the acceptance number. It is denoted by

$$\begin{pmatrix} N \\ n \\ c \end{pmatrix}$$

Let d be the number of defective items in the sample. If

(i) d ≤ c, lot is accepted (the d defective items are replaced by good ones)

(ii) d > c lot is rejected (the rejected lot is inspected fully and all its defective are replaced by good ones)

In general, d has hyper geometric distribution given by he probability function.

$$f_p(x) = P(d = x) = \frac{Np_{C_x} c_{n-x}^{N(1-p)}}{N_{C_n}}$$

If N is very large it is approximated by the Binomial Distribution.

$$f_p(x) = N_{C_x} p^x (1-p)^{n-x}$$

Moreover if n in very large and p is very small such that np = λ is finite then it is approximated by the Poisson distribution.

$$\left[f_\lambda(x) = \frac{e^{-\lambda} \lambda^x}{x!} \right]$$

The OC function of this plan is given by

L(p) = Pa (p)

= probability of accepting the lot of quality p.

= P (d ≤ c)

= P(0) + P(1) + ………………+P (C)

$= \sum_{x=0}^{c} (probability\ of\ getting\ x\ defective\ out\ of\ n)$

$= \sum_{x=0} f_p(x)$ ………………………….(4.1)

Suppose the process average is p and LTPD is pt so that the producer’s and consumer’s risk are given by

PR = α = 1- L ($\bar{p}$)

= 1- P_a ($\bar{p}$) ………………………..(4.2)

CR = β = L (p_t)

= P_a (p_t) ……………………….…..(4.3)

Usually, p is kept at AQL.

If p be the actual fraction defective in the lot of size N, the AOQ under the sampling plan is given by

$$\left[AOQ = \sum_{x=0}^{c}\left(\frac{N_p - x}{N}\right).f_p(x)\right] \quad \dots\dots\dots\dots\dots..(4.4)$$

Since the fraction defective in the lot after inspection is $\left(\frac{N_p - x}{N}\right)$ where x is the number of defective found in the accepted sample (i.e., x ≤ c) and it is zero for rejected samples (i.e. x >c)

$$AOQ = \sum_{x=0}^{c}\left(\frac{N_p - x}{N}\right).f_p(x) \quad \{since\ x = np\}$$

$$= \ p\ \left(\frac{N-n}{n}\right)\sum_{x=0}^{c} f_p(x)$$

$$= p\sum_{x=0}^{c} f_p(x) \qquad \left[\frac{N-n}{N}\sim 1\right]$$

$$= pL\ (p)$$

$$= p\ P_a(p) \qquad \dots\dots\dots\dots\dots\dots\dots..(4.5)$$

The maximum AOQ with respect to p gives AOQL.

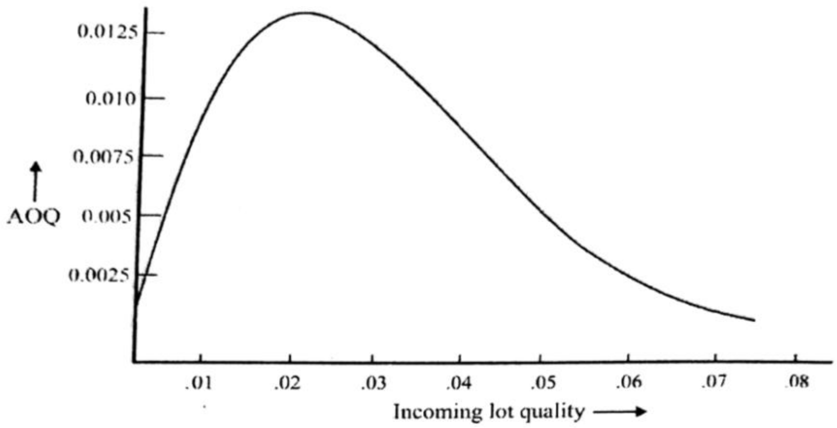
In this plan the number of items inspected is always n and therefore

ASN= n

If the lot of quality p, the average total inspection ATI is given by

ATI = np_a +N (1-P_a)

= n + (1-P_a) (N-n) ………………(4.6)



There are two approaches for this plan:

a) Lot Quality Protection: The lots size N will be specified in any while the consumer's requirement will fix the value of pi and (c.r.) β. Hence expression (4.3) gives an equation in the two unknown n and α. This equation is satisfied for various combination of value of n and c to safeguard the producer's interest too one would select that value of n and c for which A.T.I. given by (4.6) is minimum for the specified value of p. The solution however is theoretically difficult to obtain Extension tables have been prepared by Dodge and Romig who obtained the solution by numerical methods.

b) Average Quality Protection: The consumer's interests rests are taken care of by specifying the AOQL. Gives the value of N and AOQL, expression (4.4) gives an equation in n and c. In order to safeguard the producer interest, that pair of n and c satisfying (4) is selected for which A.T.I given by (4.6) is minimum for specified value of p. Extensive tables for the sampling plan under this approach are also provided by Dodge and Romig.

15.12 Double Sampling Plan

A double sampling plan is a procedure in which, under certain circumstances a second sample is required before a final decision on the lot may be made. This is defined by four parameters.

n_1= sample size of size of first sample

c_1= acceptance number of first sample

n_2 = sample size of second sample

c_2 = acceptance number for both samples

It is denoted by

$$\begin{pmatrix} N \\ n_1 \\ c_1 \\ n_2 \\ c_2 \end{pmatrix}$$

Let d_1 be the number of defectives on the first sample. If $d_1 \leq c_1$ accept the lot (replace the d_1 defective items found by good ones) $d_1 > c_2$, reject the lot (inspect 100% and replace all the defectives by non- defectives)

If $c_1 < d_1 \leq c_2$ take the second sample of size n_2 from $(N-n_1)$ remaining items. Suppose the second sample has d2 defectives. If $d_1 + d_2 \leq c_2$ accept the lot (replace the $d_1 + d_2 > c_2$, reject the lot (inspect 100% and replace all the defective by non-defectives)

Advantages & Disadvantages:

The principle advantage of a double sampling plan with respect to single sampling is that it may reduce the total amount of required inspection. Suppose, the first sample taken under a double sampling plan is smaller than the sample that would be required using a single sampling plan which offers the consumer the same protection. In all cases then in which a lot is accepted or rejected on the first sample the cost of inspection will be lower for double sampling plan. It is also possible to reject a lot without complete inspection of the second sample (curtailed sampling) i.e. to stop whenever the total number of defectives in the samples exceed c_2. Consequently, the use of double sampling can often result in cutting total inspection cost.

Double sampling has two potential disadvantages. 1. Unless curtailment is used on the second samples, double sampling may require more total inspection than would be required in a single sampling plan that offers the same protection. Thus unless double sampling is used carefully, its potential economic advantage may be lost. 2. The second disadvantage of double sampling is that it is more administratively complex which may increase the opportunity for the occurrence of inspection errors. These may be problems of storing and handling of items of the first sample which are awaiting a second sample to final decision

Let us denote

$$f(x, n; Np, N) = {}^{Np}C_x \; ({}^{N(1-p)}C_{n-x} / {}^{N}C_n)$$

$$\cong {}^{Np}C_x \, p^x \, q^{n-x}$$

$$\cong \frac{e^{-\lambda} \lambda_x}{x!}; \qquad \lambda = np$$

According to as hyper geometric or binomial model is adopted. Then the OC function of this plan is given by

L(p) = P_a (p) = Probability of accepting the lot of quality p

$$= P(d \leq c_1) + P\left(d_1 + d_2 \leq \frac{c_2}{c_1} < d_1 \leq c_2\right)$$

$$= \sum_{x=0}^{c_1}\left\{\frac{\binom{N\,p}{x}N\binom{(1-p)}{n_1-x}}{\binom{N}{n_1}}\right\} + \sum_{y=0}^{c_2}\left\{\sum_{x=c_1+1}^{c_2}\frac{\binom{Np}{x}N\binom{(1-p)}{n_1-x}}{\binom{N}{x}}\right\} \times \left\{\frac{\binom{N-n_1}{cy}\binom{(N-n_1)(1-p)}{N_2-y}}{\binom{(N-n_1)}{n_2}}\right\}$$

Or L(p) = P_a (p) = $\sum_{x=0}^{c_1} f\{x, n_1, Np, N\}$+

$$+\sum_{y=0}^{c_2-x}\left\{\sum_{x=c_1+1}^{c_2} f(x, n_1; Np, N)\right\} \times f\{y, n_2; (N-n_1)p, N-n_1\}$$

Or

$$P_a(p) - P_a^{(1)}(p) + P_a^{(2)}$$

Where

$$P_a^{(1)}(p) = \sum_{x=0}^{c_1} f\{x, n_1, Np, N\}$$

and

$$P_a^{(2)}(p) = \sum_{y=0}^{c_2-x}\left\{\sum_{x=c_1+1}^{c_2} f(x, n_1; Np, N)\right\} \times f\{y, n_2; (N-n_1)p, N-n_1\}$$

If the process average $\bar{p}$ (=AQL) and LTPD is pt, then the producer's risk and consumer's risk are given by

P.R. =α = 1-P_a ($\bar{p}$)

C.R. =β = 1-P_a (p_t)

If p is the actual fraction defective in the lot, AOQ is given by

AOQ =

$$\sum_{x=0}^{c_1}\left(\frac{N_p - x}{N}\right) f\left(x, n_1;\ n_p, N\right)$$

$$+\sum_{y=0}^{c_2-x}\sum_{x=c_1+1}^{c_2}\left[\frac{N_p-(x+y)}{N}\right] f\left(x, n_1;\ n_p, N\right) f\left(y, n_2;\ n_p, N\right) p, (N-n_1)$$

Or AOQ =

$$\frac{\left[P_a^{(1)}(p)(N-n_1)+P_a^{(2)}(p)+P_a^{(2)}(p)(N-n_1-n_2)\right]p}{N}$$

Whose maximum gives AOQL

The ASN function of this plan is given by

ASN = $n_1p_1 + (n_1+n_2)\ (1-P_1)$

Where,

P_1 = P {only one sample is necessary}

=P {lot is accepted on the first sample}+ p {lot is rejected on the first sample}

$1-P_1$ = P {the second sample is necessary)

ATI = $n_1\ P_a^{(1)}(p)(n_1+n_2)\ P_a^{(2)}(p)+N(1-P_a(p))$

As in the case of single sampling plan, there are two approaches for double sampling plan viz, (i) lot quality protection and (ii) average quality protection.

Expensive tables for both the approaches have been provided by Dodge and Romig.

15.13 Sampling Inspection by Variables Sample

In sampling inspection by variable, for each item of the sample, measurements are taken on each quality characteristic along a continuous scale.

The primary advantage of variable sampling plan is that the same operating characteristic curve can be obtained with a smaller sample size than would be required by an attribute sampling plan. Thus, a variable acceptance – sampling plan that gives the same protection as an attribute acceptance – sampling plan would required less sampling. Through the measurement data for a variable sampling plan would probably cost more per observation that the collection of attributes data but the reduction in sample size may more than offset this increased cost. When destructive testing is employed variable sampling is particularly useful in reducing the cost of inspection.

A second advantage is that measurements data usually provide more information about the manufacturing process or the lot than does attributes data. Generally numerical measurement of quality characteristic are more useful than simple classification of the item as defective or non-defective.

A final point to be emphasized is that when acceptable quality levels are very small, the sample sizes required by attributes sampling plans are very large. Under these circumstances there may be significant advantages in switching the variable measurement.

Variable sampling plans have some disadvantages. The primary disadvantage is that the distribution of the quality characteristic must be known- usually taken to be normal. If the distribution is not normal, serious departures from the advertised risks of accepting and rejecting lots of given quality may be experienced the second disadvantage of variable sampling is that a separate sampling plan must be employed for each quality characteristic.

Let x be the quality characteristic under question. It is assumed that x is normally distributed with mean μ and standard deviation σ, in the lot. There are two general types of variable sampling plans- plan that control the lot or process fraction defective, and plan that control a lot or process parameter. We shall discuss the former briefly.

Associated with x there will be specification limits, only u, only z or both u and z. If only u is given the proportion defective p, u, is given by

$$p' U = p\{x \geq U\} = 1 - \emptyset\left(\frac{U-\mu}{\sigma}\right)$$

Where, $\emptyset$ (x) is the distribution function of standard normal variable. If only L is given the proportion defection p'L, is given by

$$p' L = P\{x \leq L\} = \emptyset\left(\frac{L-\mu}{\sigma}\right) = \emptyset\left(-\frac{\mu-L}{\sigma}\right)$$

If U and L, both are specified the proportion defective is p'L + p' U

However p'L and p'U are unknown quantities, because μ *and* σ are unknown sampling inspection provides us with estimate of p'U and p'L in other words μ *and* σ of accepted or rejection.

Case 1: Known Standard Deviation

when σ is known there exist MVUE of p'L and p'U viz.

$$pL = \emptyset\left[-\sqrt{\frac{n}{n-1}}\left(\frac{\bar{x}-L}{\sigma}\right)\right]$$

$$and$$

$$pU = 1-\emptyset\left[-\sqrt{\frac{n}{n-1}}\left(\frac{U-\bar{x}}{\sigma}\right)\right]$$

A sampling plan should naturally lead to the acceptance of lot if, and is small. Thus for a given USL. U, the lot is to be accepted if

$$pU \leq M \text{ (say)}$$

or

equivalently, if

$$\frac{U-\bar{x}}{\sigma} \geq k \;\; or \; \bar{x}+k\sigma \leq U$$

Where M is a quantity determined in accordance with the specified probability of errors and k is reacted to M by

$$k = \sqrt{\frac{n-1}{n}} Z_M \dots\dots\dots\dots\dots\dots\dots\dots\dots (4.7)$$

Such that Zm is the upper 100 M% point of the standard normal variable pU≤M

Or

$$1-\emptyset\left[-\sqrt{\frac{n}{n-1}}\left(\frac{U-\bar{x}}{\sigma}\right)\right] \leq M$$

$$= \emptyset\left[-\sqrt{\frac{n}{n-1}}\left(\frac{U-\bar{x}}{\sigma}\right)\right] \geq 1-M$$

$$= \frac{U-\bar{x}}{\sigma} \geq \sqrt{\frac{n}{n-1}} Z_M$$

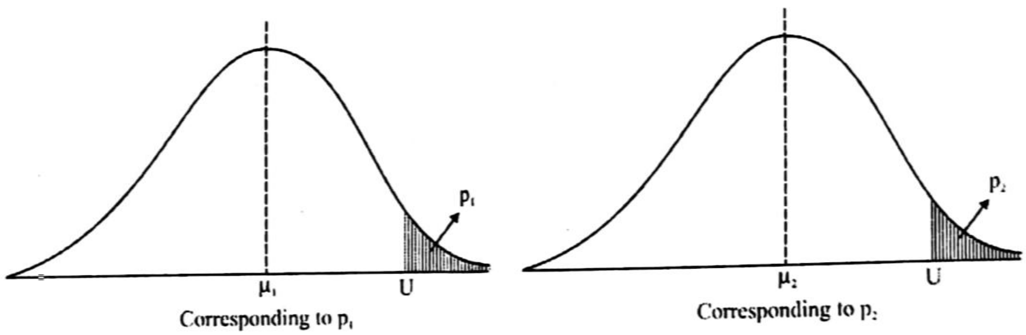
$$= \bar{x} \leq U - \sigma\sqrt{\frac{n}{n-1}} Z_M$$

Derivation of n, k and M

Suppose that we are acceptance quality level p_1 (like AQL), rejection quality level p_2 (like LTPD) producer's risk α and consumer's risk p, such that

P {acceptance of lot / p_1} = 1-α ……………….(4.8)

P {acceptance of lot / p_2} = 1-β …………..(4.9)



Evidently,

$$\mu_1 = U - Z_{p_1}\sigma \quad or \quad U = \mu_1 + Z_{p_1}\sigma$$

$$\mu_2 = U - Z_{p_2}\sigma \quad or \quad U = \mu_2 + Z_{p_2}\sigma$$

From (4.8) and (4.9) we get

$$p\left\{\bar{x}, k\sigma \leq \frac{U}{\mu_1}\right\} = 1 - \alpha \qquad \ldots\ldots\ldots\ldots..(4.10)$$

and

$$p\left\{\bar{x}, k\sigma \leq \frac{U}{\mu_1}\right\} = 1 - \beta \qquad \ldots\ldots\ldots\ldots..(4.11)$$

Equation (4.10) yields

$$P\left[\frac{(\bar{x} + k\sigma) - (\mu_1 + k\sigma)}{\sigma/\sqrt{N}} \leq \frac{U - (\mu_1 + k\sigma)}{\sigma/\sqrt{N}}\right] = 1 - \alpha$$

Or

$$P\left[\frac{(\bar{x} + k\sigma) - (\mu_1 + k\sigma)}{\sigma/\sqrt{N}} \leq \frac{Z_{p_1} - (\mu_1 + k\sigma)}{\sigma/\sqrt{N}}\right] = 1 - \alpha$$

Or

$$\frac{(Z_{p_1} + k)\sigma}{\frac{\sigma}{\sqrt{N}}} = Z_\alpha$$

$$k = Z_{p_1} - \frac{Z_\alpha}{\sqrt{n}} \qquad \dots\dots\dots\dots..(4.12)$$

Similarly equation (4.11) yields.

$$k = Z_{p_2} - \frac{Z_\beta}{\sqrt{n}} \qquad \dots\dots\dots\dots\dots..(4.13)$$

Knowing p_1, p_2, α , β we may get the value of n and k from (4.12) and (4.13)

$$n = \left[\frac{Z_\alpha + Z_\beta}{Z_{p_1} - Z_{p_2}}\right]^2 \qquad \dots\dots\dots\dots\dots\dots..(4.14)$$

and

$$k = \frac{Z_\alpha Z_{p_2} + Z_\beta Z_{p_1}}{Z_\alpha + Z_\beta} \qquad \dots\dots\dots\dots\dots\dots(4.15)$$

From these values of n and k we may easily obtain the value of M by use of (4.7).

If LsL, L, given lot is accepted if

pL ≤ M

i.e.

$$\frac{\bar{x} - L}{\sigma} \geq k \quad or\ if\ \bar{x} = k\sigma \ \geq L$$

And if u and k according to the lot size the sample size and specified AQ< are given by Bowker and Boode in “sampling inspection by variables”.

Case 2: Unknown Standard Deviation

when σ is unknown sampling inspection is based on sample mean and sample standard deviations given by s= The MVUE of p’U and p’L are given by pU and pL, functions of $\left(\frac{U-\bar{x}}{s}\right)$ and $\left(\frac{\bar{x}-L}{s}\right)$.

It can be shown that.

(i) If USL, U, is given the lot is accepted , if $\bar{x} + k^{'} s \leq U$

(ii) If LSL, L, is given the lot is accepted , if $\bar{x} + k^{'} s \geq L$

The value of k', for different cases have been tabulated.

15.14 Example

Example: Plot the operating characteristic curves for single sampling plan where N= 5000, n= 100, C= 1,2,3. Assuming consumer's risk β= .01, determine the lot tolerance fraction defective. Also plot the average outgoing quality (AOQ) curve and determine AOQL.

Solution: Here we have n/N = 1/50 which is very-very small. So the no. of defective items 'd' can be assumed to follow Poisson distribution. If Now, we use the notations already discussed, then

$$L(p) = P_a = \sum_{\lambda=0}^{c} \frac{e^{-np} (np)^{\lambda}}{\lambda!}$$

Where 'P' is the lot quality. 'pa' will be calculated using Biometrika table

We further have-

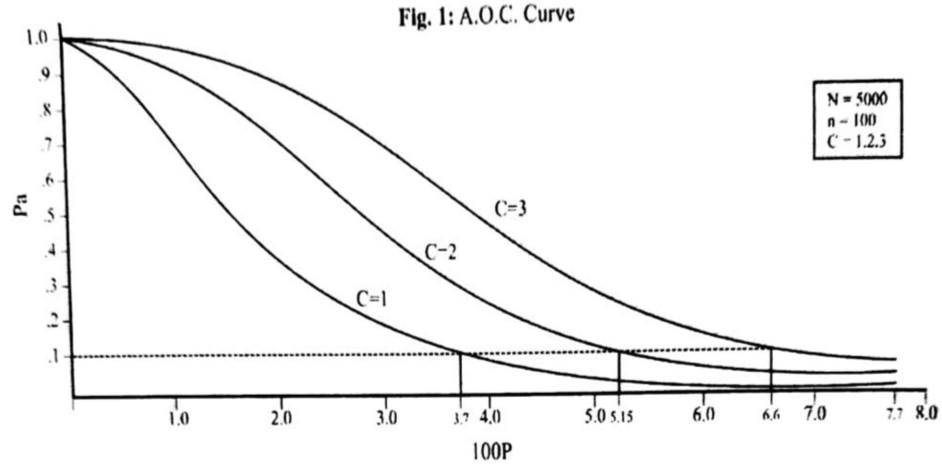
$$A.O.Q. = \frac{P(N-n)}{N} . P_a$$

To determine lot tolerance fraction defective we will draw lines parallel to 'p' axis at a distance P_a = 0.10. Then draw perpendiculars to the p-axis from the points where this line meets the O.C. curves for C= 1,2,3.

Table for calculating P_a and A.O.Q.

np	Pa			$100P\left(\frac{N-n}{N}\right)$	A.O.Q		
	C=1	C=2	C=3		C=1	C=2	C=3
.01	1.00	1.00	1.00	.01	.01	.01	.01
.2	.98	.99	1.00	.19	.19	.18	.19
.25	.97	.99	1.00	.24	.23	.24	.29
.3	.96	.99	1..00	.29	.28	.29	.29
.8	.81	.95	.99	.78	.63	.47	.77
1.0	.74	.92	.98	.98	.72	.90	.96
1.4	.59	.83	.95	1.37	.81	1.14	1.30
1.9	.43	.70	.87	1.86	.80	1.30	1.62
2.4	.31	.57	.78	2.35	.73	1.34	1.83
2.8	.23	.47	.69	2.74	.63	1.29	1.89
3.1	.18	.40	.62	3.03	.54	1.21	1.88
4.3	.07	.20	.38	4.21	29	.84	1.60
4.9	.04	.13	.28	4.80	.19	.62	1.34
5.6	.02	.08	.19	5.49	.11	.44	1.04
5.8	.02	.07	.17	5.68	.11	.40	.96
6.2	.01	.05	.13	6.07	.06	.30	.79
6.6	.01	.04	.10	6.47	.06	.26	.64
6.8	.01	.03	.09	6.66	.06	.20	.60
7.0	.01	.03	.08	6.86	.00	.20	.54
7.7	.01	.02	.05	7.55	.00	.15	.38

Fig. 1: A.O.C. Curve



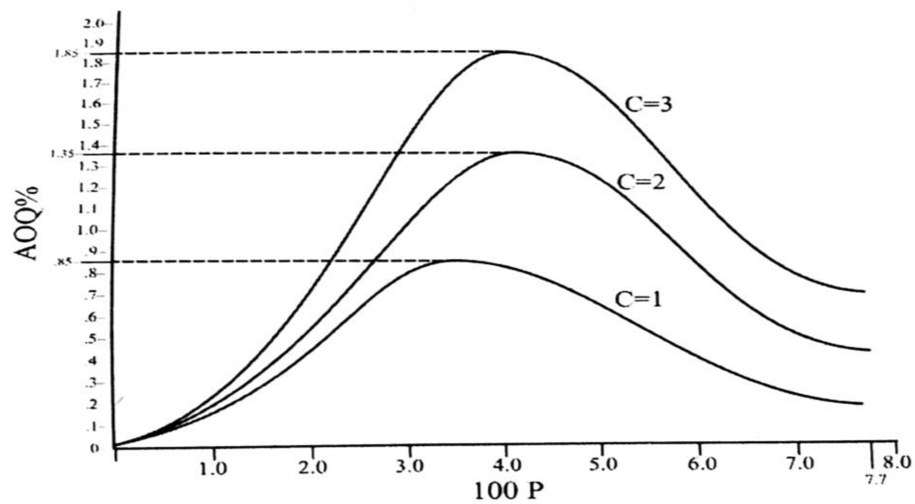
For P_a = .101 we get for Fig. (1)

For C=1 L.T.F.D. = 0.037

C=2 L.T.F.D. = 0.0515

C=3 L.T.F.

Fig. 2: The A.O.Q. Curve for single sampling



From The AOQ curve we get

For C=1 A.O.Q.L. = 0.85%

For C=2 A.O.Q.L. = 1.35%

For C=3 A.O.Q.L.= 1.85%

15.15 Exercise

1. Explain the terms producer's risk and consumer's risk.
2. Explain what is single sampling plan and what is double sampling plan.
3. For a single sampling plan, N = 2000, n= 100, C=2
   (i) Find Pa, when p = .005, .01, .10
   (ii) Find AOQL for the same

15.16 Summary

The main advantage of sampling inspection is that cost of inspection and time involved can be reduced dramatically. From economic considerations it is not practicable to inspect a full lot. So one has to opt from some sampling inspection plan whether single sampling or double sampling as the need may be.

The two main considerations on the basis of which the two plans my be compared are the operating characteristics and the average sample number. The average amount of inspection required per lot is more for single sampling plan it is for double sampling plan. Speaking generally, a double sampling plan often requires 25% to 33% less inspection on the average. For these two plans we can say that

1. It is easier for the sampling inspection to understand the technique of single sampling plan.

The psychological satisfaction gained from giving the inspected lot more than one chance is absent in single sampling.

4.17 Further Readings

- Burr, I.W. Engineering Statistics and Quality Control. Mcgraw Hill.
- Cowden, D.J. Statistical Methods in Quality Control, Prentica Hall.
- Goon, Gupta, Dasgupta; Fundamentals of Statistics, Vol. Two, The World Press Pvt. Lt.