# *DIRECTORATE OF DISTANCE EDUCATION*

**UNIVERSITY OF JAMMU**
**JAMMU**

## SELF LEARNING MATERIAL

## B.COM SEMESETER - V

**Subject : Business Statistics** **Unit I to IV**

**Course No. : BCG-502** **Lesson No. 1 to 12**

**Rohini Gupta Suri**
**CO-ORDINATOR**
**(M) : 94191-86716**

**Printed and Published on behalf of the Directorate of Distance Education,
University of Jammu, Jammu by the Director, DDE, University of Jammu,
Jammu.**
**http:/www.distanceeducationju.in**

# BUSINESS STATISTICS

**COURSE CONTRIBUTORS**

- Ms. Anu Rani
  NET, JRF

**REVIEW & EDITED BY :**

- Rohini G. Suri

- The Script writer shall be responsible for the lesson/script submitted to the DDE and any plagiarism shall be his/her entire responsibility.

Printed by : Jandiyal Printing Press, Jammu / 2021 / 170 Nos.

## UNIVERSITY OF JAMMU

## B.COM. FIFTH SEMESTER

## BUSINESS STATISTICS

C.No. BCG- 502 Max Marks = 100

Internal assessment = 20

External exam. = 80

**OBJECTIVE :** The basic objective of this course is to make students aware of the importance of business statistics in solving business problems.

### UNIT-I : INTRODUCTION & ANALYSIS OF QUANTITATIVE DATA

Meaning, scope, importance and limitation of statistics measures of Central Tendency Arithmetic Mean (Simple & weighted) Median, Quartiles, Deciles, Percentiles, Mode; Merits, demerits and uses of mean, median & mode;. Requirements of a good average;

### UNIT –II : MEASURES OF DISPERSION

Range, semi-inter quartile range, Quartile deviation, Mean deviation, Standard deviation and their coefficients, Coefficient of variation & skewness – Karl Pearson's, Bowley's and Kelly's coefficient of skewness.

### UNIT- III : INDEX NUMBERS

Meaning and uses of Index numbers; Methods of construction of index numbersUnweighted Indices - Simple aggregative method, simple average of price relative method, Weighted Indices-Laspeyre's method, Paasche's method, Fisher's ideal index number including time and factor reversal tests; Cost of living index: Aggregative expenditure method and family budget method.

### UNIT- IV : CORRELATION AND REGRESSION

Correlation-Meaning, types & importance; methods of studying correlation Karl Pearson's coefficient of correlation, Rank correlation and concurrent deviations, (Ungrouped data only), probable error and interpretation of data. Regression analysis: Meaning & objectives; Regression lines, Regression equation of X on Y and Y on X (Ungrouped data only)

## SKILL DEVELOPMENT (GUIDELINES FOR CLASS ROOM TEACHING AND INTERNAL ASSESSMENT)

To develop the practical knowledge about simple average and positional averagev amongst the students. State the practical importance of measure of dispersionv Construct index numbers taking examples from real life situationsv Find the practical utility of correlation and regression.v Create deep understanding of all concepts specified in the syllabus.v

## BOOKS RECOMMENDED

1. Gupta, S.C :Fundamentals of Statistics, Himalaya Pub.,New Delhi

2. Hans,Gupta & Aggarwal :Business Statistics,Kalyani Publishers,New Delhi

3. Gupta, S.C &Gupta, M.P. : Business Statistics,Sultan Chand Pub.,New Delhi

4. Tulsain & Jhunjnawalla : Business Statistics,S.Chand Pub.,New Delhi

5. Chandan ,J.S. : Statistics for Business and Economics, Vikas Pub., New Delhi

6. Sancheti and Kapoor : Statistics, S.Chand, New Delhi

7. Elhance, D.N. : Principles of Statistics, Kitab Mahal, New Delhi

8. Gupta, S.P : Statistical methods, S. Chand Pub. New Delhi

## NOTE FOR PAPER SETTER

Equal weightage shall be given to all the units of the syllabus. The external Paper shall be of the two sections viz, A & B of three hours duration. Section-A: This section shall contain four short answer questions selecting one from each unit. Each question shall carry 5 marks .A candidate shall be required to attempt all the four questions. Total weightage to this section shall be of 20 marks. Section-B: This section shall contain eight long answer questions of 15 marks each. Two questions with internal choice shall be set from each unit. A candidate shall have to attempt any four questions selecting one from each unit. Total weightage to this section shall be of 60 marks.

## MODEL QUESTION PAPER

## BUSINESS STATISTICS

Max Marks = 80

Time allowed : 3 hrs.

### Section - A (Marks 20)

**Attempt all the four questions. Each questions carries five marks.**

1. Define statistics?
2. What is coeffiecient of variation?
3. State the utility of wholesale price index number?
4. Differentiate betwen correlation and regression?

### Section - B (Marks 60)

**Attempt any four questions selections one from each unit. Each question carries 15 marks.**

1. There were 500 workets working in a factory. Their mean wage was calculated as 200. Later on it was discovered that the wags of two workers were misread as 100 and 20 in place of 80 and 220. Find the correct average?

   or

   Find out medium and mode from the following ?  
   Less than 10, 20, 30, 40, 50, 60  
   Frequency 5, 13, 25, 37, 45, 50

2. Find quartile deviation and mean deviation from the following date:  
   Size : 5 8 10 12 19 20 32  
   Frequency : 3 10 15 20 8 7 6

   Or

   Find the coefficient of variation from the date given below :

| Class : | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 |
|---|---|---|---|---|---|---|---|
| No. of Student : | 10 | 15 | 25 | 25 | 10 | 10 | 5 |

3. Calculate index numbers from the following data by :

1) Laspeyer's 2) Paasche's 3) Fisher's method.

Commodity :

| | p0 | q0 | p1 | q1 |
|---|---|---|---|---|
| A | 8.0 | 5 | 10 | 11 |
| B | 8.5 | 6 | 9 | 9 |
| C | 9.0 | 4 | 12 | 6 |

Or

Construct the cost of living index from the information given below :

| Expenses | Perentage | Price in Rs. (2013) | Prices in Rs. (2014) |
|---|---|---|---|
| Food | 35 | 125 | 150 |
| Fuel | 15 | 50 | 60 |
| Clothing | 20 | 100 | 120 |
| Miscellaneous | 30 | 60 | 90 |

4. Caculate the coefficient of correlation between x and y series from the following date :

| x : | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| y: | 3 | 4 | 6 | 7 | 10 |

Or

Solve two regression equations x and y and y on x.

| Age of husband : | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|
| Age of wife : | 20 | 22 | 25 | 26 | 27 | 28 |

Find out the age of husband when wife's age in 24.

**C. No. : BCG-502** **UNIT : I**
**SEMESTER: V** **LESSON: 1-3**

# BUSINESS STATISTICS

## STRUCTURE

1.1 Introduction

1.2 Objectives

1.3 Meaning of Statistics

 1.3.1 Definition of Statistics

 1.3.2 Characteristics of Statistics

 1.3.3 Scope of Statistics

 1.3.4 Importance of Statistics

 1.3.5 Limitations of Statistics

1.4 Summary

1.5 Self Assessment Exercises

1.6 Suggested Readings

## 1.1 INTRODUCTION

Growing complexities of human natural/social phenomenon has left no alternative to the world decision-makers but to depend upon the cardinal as well as ordinal methods of measurements and interpretation of the varied problems. Nearly, every activity-mental or physical or natural, is measured and interpreted quantitatively. Today, whole world lives in the world of numbers. Statistics reaches to conclusive decisions on the basis of Statistical knowledge. Thus, knowledge of statistics applications is very important in commerce, and to understand it, it is essential to understand meaning, characteristics, importance and limitations of Statistics.

For a layman, 'Statistics' means numerical information expressed in quantitative terms. This information may relate to objects, subjects, activities, phenomena, or regions of space. As a matter of fact, data have no limits as to their reference, coverage, and scope. At the macro level, these are data on gross national product and shares of agriculture, manufacturing, and services in GDP (Gross Domestic Product). At the micro level, individual firms, howsoever small or large, produce extensive statistics on their operations. The annual reports of companies contain variety of data on sales, production, expenditure, inventories, capital employed, and other activities. These data are often field data, collected by employing scientific survey techniques. Unless regularly updated, such data are the product of a one-time effort and have limited use beyond the situation that may have called for their collection. A student knows statistics more intimately as a subject of study like economics, mathematics, chemistry, physics, and others. It is a discipline, which scientifically deals with data, and is often described as the science of data. In dealing with statistics as data, statistics has developed appropriate methods of collecting, presenting, summarizing, and analysing data, and thus consists of a body of these methods. The material presented in this unit is a step in this direction.

## 1.2 OBJECTIVES

After studying this lesson, you should be able to understand -

- the meaning of Statistics.
- characteristics of Statistics.
- have idea about importance and limitations of Statistics.

## 1.3 MEANING OF STATISTICS

The word ‘statistics’ is used in two senses plural and singular. In the plural sense, it refers to a set of figures or data. In the singular sense, statistics refers to the whole body of tools that are used to collect data, organise and interpret them and, finally, to draw conclusions from them. It should be noted that both the aspects of statistics are important if the quantitative data are to serve their purpose. If statistics, as a subject, is inadequate and consists of poor methodology, we could not know the right procedure to extract from the data the information they contain.  Similarly, if our data are defective or that they are inadequate or inaccurate, we could not reach the right conclusions even though our subject is well developed.

Thus, statistics is the study of the collection, organization, analysis, interpretation and presentation of data. It deals with all aspects of data, including the planning of data collection in terms of the design of surveys and experiments. Statistics, like many words have different meanings in different contexts. Some people regard Statistics as data, facts or measurements while others believe it to be the study of figures. There is another group of people who consider it as analysis of figures for forecasting or drawing inferences. Besides this, the representation of facts in the form of diagrams, graphs or maps is also supposed to be Statistics. In this way Statistics imbibes three forms.

**i) As Numerical Data :** In the product form, it represents the numerical data, such as statistics of national income, unemployment, imports, exports, production, etc. Here, it is used in plural form.

**ii) As a subject :** In the process form, it is used as a subject and is in singular form like Economics, Physics, etc. In this sense, the term ‘statistics’ refers the whole field of study of which ‘Statistics’ in the plural sense are the subject-matter. In other words, it refers to the statistical principles and methods finally help in taking decisions in view of uncertainty.

**iii) Modern Connotation :** In its modern connotation, it may refer to the study of, and research into the theory and principles underlying statistical methods. It is the field of study that pioneer in, expands the frontiers of statistical methodology and uses.

### 1.3.1 Definitions of Statistics

Some of the definitions of Statistics are :-

- ***A.L. Bowley*** has defined statistics as: (i) statistics is the science of counting, (ii) Statistics may rightly be called the science of averages, and (iii) statistics is the science of measurement of social organism regarded as a whole in all its manifestations.

- ***Boddington*** defined as: Statistics is the science of estimates and probabilities. Further, *W.I. King* has defined Statistics in a wider context, the science of Statistics is the method of judging collective, natural or social phenomena from the results obtained by the analysis or enumeration or collection of estimates.

- ***Seligman*** explored that statistics is a science that deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry.

- ***Spiegal*** defines statistics highlighting its role in decision-making particularly under uncertainty, as follows: statistics is concerned with scientific method for collecting, organising, summa rising, presenting and analyzing data as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis.

- According to *Prof. Horace Secrist*, Statistics is the aggregate of facts, affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose, and placed in relation to each other.

From the above definition, we can highlight the major features of statistics as follows:

**(i)** ***Statistics are the aggregates of facts***. It means a single figure is not statistics. For example, national income of a country for a single year is not statistics but the same for two or more years is statistics.

**(ii)** ***Statistics are affected by a number of factors***. For example, sale of a product depends on a number of factors such as its price, quality, competition, the income of the consumers, and so on.

**iii)** ***Statistics must be reasonably accurate***. Wrong figures, if analysed, will lead to erroneous conclusions. Hence, it is necessary that conclusions must be based on accurate figures.

**(iv)** ***Statistics must be collected in a systematic manner***. If data are collected in a haphazard manner, they will not be reliable and will lead to misleading conclusions.

***(v) Collected in a systematic manner for a pre-determined purpose***

**(vi)** Lastly, **Statistics should be placed in relation to each other**. If one collects data unrelated to each other, then such data will be confusing and will not lead to any logical conclusions. Data should be comparable over time and over space.

### 1.3.2 Characteristics of Statistics

Some statisticians put Statistics in the category of Science, while others believe it to be Arts.

**(a) Statistics as a Science :** Science is a body of systematized knowledge. Any subject can be put in the category of Science if it possess following characteristics :-

(i) Its law and methods must be universally acceptable.

(ii) It must analyse the cause-effect relationship.

(iii) It must possess the quality of estimation and forecasting.

(iv) It should be dynamic.

Nearly, all the above characteristics are seen in Statistics. In this light, Statistics is a Science.

**(b) Statistics as an Art :** In Art, we are able to know as what to do, how to do, and what ought to be done. In this way, it is a way of doing the things.It possess the following characteristics :-

(i) Art is a group of activities to reach to an optimum solution to a problem.

(ii) It not only describes, but also prescribes.

(iii) For reaching to a definite level of perfection in arts, a special knowledge, experience and self-restriant is required.

Statistics possess the characteristics of Art, as it clearly suggests suitable methods and Laws of Statistics to reach a dersirable solution of the problems.

Thus, it is both a Science and an Art. It is Science in that its methods are basically systematic and have general application, and an Art that their applications depend to a considerable degree on the skill i.e. the special experience of the statistician and on his knowledge of the field of application.

**Divisions of Statistics:** The subject-matter of Statistics can be studied by dividing it in the following broad categories :

**i. Theoretical Statistics :** It includes the mathematical theories specially applied in Statistics.

**ii. Statistical Methods :** It includes methods or procedures used in the collection, organisation, summary, analysis, interpretation and presentation of data. These methods may again be put into two broad categories :

**a) Descriptive Methods :** It involves techniques of summarising data an presenting them in an useable form for the purpose of describing their character.

**b) Inductive or Inferential Statistics :** It encompasses the involvement of Statistics in making forecasts, estimation or judgement about some large group of data than that of actually observed or about some future happening based on a study of sample or historical data.

iii. **Applied Statistics:** It consists of application of Statistical methods and techniques to the problems and facts as they exist. Statistical quality control, Index numbers, Time Series etc. are few important tools of applied Statistics.

### 1.3.3 Scope of Statistics

Apart from the methods comprising the scope of descriptive and inferential branches of statistics, statistics also consists of methods of dealing with a few other issues of specific nature. Since these methods are essentially descriptive in nature, they have been discussed here as part of the descriptive statistics. These are mainly concerned with the following:

(i) It often becomes necessary to examine how two paired data sets are related. For example, we may have data on the sales of a product and the expenditure incurred on its advertisement for a specified number of years. Given that sales and advertisement expenditure are related to each other, it is useful to examine the nature of relationship between the two and quantify the degree of that relationship. As this requires use of appropriate statistical methods, these falls under the purview of what we call regression and correlation analysis.

(ii) Situations occur quite often when we require averaging (or totalling) of data on prices and/or quantities expressed in different units of measurement. For example, price of cloth may be quoted per meter of length and that of wheat per kilogram of weight. Since ordinary methods of totalling and averaging do not apply to such price/quantity data, special techniques needed for the purpose are developed under index numbers.

(iii) Many a time, it becomes necessary to examine the past performance of an activity with a view to determining its future behaviour. For example, when engaged in the production of a commodity, monthly product sales are an important measure of evaluating performance. This requires compilation and analysis of relevant sales data over time. The more complex the activity, the more varied the data requirements. For profit maximising and future sales planning, forecast of likely sales growth rate is crucial. This needs careful collection and analysis of past sales data. All such concerns are taken care of under time series analysis.

(iv) Obtaining the most likely future estimates on any aspect(s) relating to a business or economic activity has indeed been engaging the minds of all concerned. This is particularly important when it relates to product sales and demand, which serve the necessary basis of production scheduling and planning. The regression, correlation, and time series analyses together help develop the basic methodology to do the needful. Thus, the study of methods and techniques of obtaining the likely estimates on business/economic variables comprises the scope of what we do under business forecasting.

Keeping in view the importance of inferential statistics, the scope of statistics may finally be restated as consisting of statistical methods which facilitate decision— making under conditions of uncertainty. While the term statistical methods is often used to cover the subject of statistics as a whole, in particular it refers to methods by which statistical data are analysed, interpreted, and the inferences drawn for decision making.

Though generic in nature and versatile in their applications, statistical methods have come to be widely used, especially in all matters concerning business and economics. These are also being increasingly used in biology, medicine, agriculture, psychology, and education. The scope of application of these methods has started opening and expanding in a number of social science disciplines as well. Even a political scientist finds them of increasing relevance for examining the political behaviour and it is, of course, no surprise to find even historians statistical data, for history is essentially past data presented in certain actual format.

### 1.3.4 Importance of Statistics

The evergrowing importance of Statistics as quantitative methods is because of the functions, which it performs :-

**(i) Statistics Provides Definiteness to The Facts :** Quantitative facts can easily be believed and trusted in comparison to abstract and quantitative facts. Statistics summarises the generalised facts and presents them in a definite form. Various characteristics pertaining to some phenomena, become easily understandable, if they are expressed in numbers.

**(ii) Statistics Helps in Comparative Studies:** Statistical device like averages, ratios, coefficients, standard errors, etc. are best ways of comparison of two phenomena. The object of Statistics is to enable comparison to be made between past and present results with a view to ascertain the reasons for changes which have taken place and the effect of such changes in future.

**(iii) Statistics Helps in Studying Relationship Between Different Facts :** The numerically expressed phenomena are mere representation of hidden relationship of different factors. Association between two attributes, relationship between two/many variables like price and demand, supply and price, income and expenditure, share prices and political stability etc. can best be measured with the help of statistical tools.

**(iv) Statistics Enlarge Knowledge and Experience of Individuals :** The study of Statistics helps in creating and establishing definiteness and clarify an individuals thinking. It sharpens the facility of rational thinking and reasoning and is helpful in propounding new theories and concepts.

**(v) Statistics Helps and guides to formulate policies in diverse fields :** Formulation of policies in different fields of knowledge is essential function of any discipline whether it is a social science, business or international trade. But, no policy can be formulated in air without any data related to the field. Statistics helps in developing laws and policies by providing suitable data and methodology.

Apart from the above, There are three major functions in any business enterprise in which the statistical methods are useful. These are as follows:

**(i) The planning of operations:** This may relate to either special projects or to the recurring activities of a firm over a specified period.

**(ii) The setting up of standards:** This may relate to the size of employment, volume of sales, fixation of quality norms for the manufactured product, norms for the daily output, and so forth.

**(iii) The function of control:** This involves comparison of actual production achieved against the norm or target set earlier. In case the production has fallen short of the target, it gives remedial measures so that such a deficiency does not occur again.

A worth noting point is that although these three functions-planning of operations, setting standards, and control-are separate, but in practice they are very much interrelated.

### 1.3.5 Limitations of Statistics

Even though Statistics have served the mankind in many ways and in many fronts from peace to war and is being utilised by almost every field of knowledge for its advancement and further researches, it is not free from shortcomings which restrict its scope and usefulness. It is always advisable to use it keeping its limitations in mind. These limitations are :

1. **Statistics Does Not Deal With Individual Measurements :** Since Statistics deals with aggregate of facts, the study of individual

measurements lies outside the scope of Statistics. Data are statistical when they relate to measurement of masses, not statistical when they relate to an individual item or event as a separate entity.

2. **Statistics Deals Only With Quantitative Characteristics :** Statistics are numerical statements of facts. Such characteristics as cannot be expressed in numbers are incapable of Statistical analysis. Thus qualitative characteristics like honesty, efficiency, intelligence, blindness etc; cannot be studied directly. However, it may be possible to analyse such problems statistically by expressing them numerically.

3. **Statistical Results are True Only on Average :** The conclusions obtained statistically are not universally true. They are true only under certain conditions. This is because Statistics as a science is less exact as compared to natural sciences.

4. **There are certain phenomena or concepts where statistics cannot be used.** This is because these phenomena or concepts are not amenable to measurement. For example, beauty, intelligence, courage cannot be quantified. Statistics has no place in all such cases where quantification is not possible.

5. Statistics reveal the average behaviour, the normal or the general trend. An application of the 'average' concept if applied to an individual or a particular situation may lead to a wrong conclusion and sometimes may be disastrous. For example, one may be misguided when told that the average depth of a river from one bank to the other is four feet, when there may be some points in between where its depth is far more than four feet. On this understanding, one may enter those points having greater depth, which may be hazardous.

6. Since statistics are collected for a particular purpose, such data may

not be relevant or useful in other situations or cases. For example, secondary data (i.e., data originally collected by someone else) may not be useful for the other person.

7. Statistics are not 100 per cent precise as is Mathematics or Accountancy. Those who use statistics should be aware of this limitation.

8. In statistical surveys, sampling is generally used as it is not physically possible to cover all the units or elements comprising the universe. The results may not be appropriate as far as the universe is concerned. Moreover, different surveys based on the same size of sample but different sample units may yield different results.

9. At times, association or relationship between two or more variables is studied in statistics, but such a relationship does not indicate cause and effect' relationship. It simply shows the similarity or dissimilarity in the movement of the two variables. In such cases, it is the user who has to interpret the results carefully, pointing out the type of relationship obtained.

10. A major limitation of statistics is that it does not reveal all pertaining to a certain phenomenon. There is some background information that statistics does not cover. Similarly, there are some other aspects related to the problem on hand, which are also not covered. The user of Statistics has to be well informed and should interpret Statistics keeping in mind all other aspects having relevance on the given problem.

Apart from the limitations of statistics mentioned above, there are misuses of it. Many people, knowingly or unknowingly, use statistical data in wrong manner. Let us see what the main misuses of statistics are so that the same could be avoided when one has to use statistical data. The misuse of Statistics may take several forms some of which are explained below:

i) **Sources of data not given:** At times, the source of data is not given. In the absence of the source, the reader does not know how far the data are reliable. Further, if he wants to refer to the original source, he is unable to do so.

ii) **Defective data:** Another misuse is that sometimes one gives defective data. This may be done knowingly in order to defend one's position or to prove a particular point. This apart, the definition used to denote a certain phenomenon may be defective. For example, in case of data relating to unemployed persons, the definition may include even those who are employed, though partially. The question here is how far it is justified to include partially employed persons amongst unemployed ones.

iii) **Unrepresentative sample:** In statistics, several times one has to conduct a survey, which necessitates to choose a sample from the given population or universe. The sample may turn out to be unrepresentative of the universe. One may choose a sample just on the basis of convenience. He may collect the desired information from either his friends or nearby respondents in his neighbourhood even though such respondents do not constitute a representative sample.

iv) **Inadequate sample:** Earlier, we have seen that a sample that is unrepresentative of the universe is a major misuse of statistics. This apart, at times one may conduct a survey based on an extremely inadequate sample. For example, in a city we may find that there are 1, 00,000 households. When we have to conduct a household survey, we may take a sample of merely 100 households comprising only 0.1 per cent of the universe. A survey based on such a small sample may not yield right information.

v) **Unfair Comparisons:** An important misuse of statistics is making unfair comparisons from the data collected. For instance, one may construct an index of production choosing the base year where the production was much less. Then he may compare the subsequent year's production from this low base. Such a comparison will undoubtedly give a rosy picture of the production though in reality it is not so. Another source of unfair comparisons could be when one makes absolute comparisons instead of relative ones. An absolute comparison of two figures, say, of production or export, may show a good increase, but in relative terms it may turnout to be very negligible. Another example of unfair comparison is when the population in two cities is different, but a comparison of overall death rates and deaths by a particular disease is attempted. Such a comparison is wrong. Likewise, when data are not properly classified or when changes in the composition of population in the two years are not taken into consideration, comparisons of such data would be unfair as they would lead to misleading conclusions.

vi) **Unwanted conclusions:** Another misuse of statistics may be on account of unwarranted conclusions. This may be as a result of making false assumptions. For example, while making projections of population in the next five years, one may assume a lower rate of growth though the past two years indicate otherwise. Sometimes one may not be sure about the changes in business environment in the near future. In such a case, one may use an assumption that may turn out to be wrong. Another source of unwarranted conclusion may be the use of wrong average. Suppose in a series there are extreme values, one is too high while the other is too low, such as 800 and 50. The use of an arithmetic average in such a case may give a wrong idea. Instead, harmonic mean would be proper in such a case.

**Confusion of correlation and causation:** In statistics, several times one has to examine the relationship between two variables. A close relationship between the two variables may not establish a cause-and-effect-relationship in the sense that one variable is the cause and the other is the effect. It should be taken as something that measures degree of association rather than try to find out causal relationship.

## 1.4 SUMMARY

- **Meaning of Statistics**
  - Statistics is a body of methods for making wise decisions on the face of uncertainty.
  - Statistical methods may be defined as the collection, presentation, analysis and interpretation of numerical data.
- **Characteristics of Statistics**
  - Statistics is both a Science and an Art.
  - Statistics can be characterized as :
    1. Theoretical Statistics.
    2. Statistical Methods.
       (a) Descriptive Methods.
       (b) Inductive or Inferential Methods.
    3. Applied Statistics.
- **Importance of Statistics**
  - Statistics provides definiteness to the facts.
  - Statistics helps in comparative studies.
  - Statistics helps in studying relationship between different facts.
  - Statistics enlarges knowledge and experience of individuals.
  - Statistics helps to formulate policies.

- **Limitations of Statistics**
  - Statistics does not deal with individual measurements.
  - Statisticas deals only with quantitative characteristics.
  - Statistics results are true only on average.
  - Statistics can be misused.

## 1.5 SELF ASSESSMENT EXCERCISES

1. Comment on the following statements :

   (a) "Figures won't lie but liars figure."

   (b) "Statistics are like clay of which you can make a God or a Devil, as you please."

   ______________________________________________________________

   ______________________________________________________________

   ______________________________________________________________

   ______________________________________________________________

   ______________________________________________________________

   ______________________________________________________________

2. What is Statistics? Discuss its scope and limitations.

   ______________________________________________________________

   ______________________________________________________________

   ______________________________________________________________

   ______________________________________________________________

   ______________________________________________________________

   ______________________________________________________________

3. Discuss with help of suitable illustrations the importance of statistics.

   ______________________________________________________________

4. What are the shortcomings of Statistics? Can these shortcomings be overcome?

## 1.6 SUGGESTED READINGS

1. Yule and Kendall : Introduction to the Theory of Statistics.
2. S.P. Gupta : Statistical Methods. S. Chand and Sons, New Delhi.

C. No. : BCG-502 UNIT : I

SEMESTER: V LESSON: 2

## MEASURE OF CENTRAL TENDENCY

**Structure**

2.1 Objectives

2.2 Introduction

2.3 Meaning of Average/Mean

2.4 Objectives of Averaging

2.5 Types of Averages

2.6 Simple Mean and Weighted Mean

2.7 Geometric Mean

2.8 Summary

2.9 Self Assessment Exercises

2.10 Suggested Readings

## 2.1 INTRODUCTION

One of the most widely used set of summary figures is known as measures of location which are often referred to as average, central tendency or central location. The purpose of computing an average value for set of observations is to obtain a single value which is representative of all the items and which the mind can grasp simply. Average is very important statistical technique because it is useful for research work and for other calculation which is base for other projects.

## 2.2 OBJECTIVES

After studying this lesson, you should be able to

- Understand the concept of average
- Understand the meaning of weighted mean
- Understand the meaning of geometric mean
- Know how to calculate arithmetic mean

## 2.3 MEANING OF AVERAGE / MEAN

The word 'average' has been defined differently by various authors. Some important definitions are given below:

- "Average is an attempt to find one single figure to describe whole of figures." **-——Clark**
- "An average is a single value selected from a group of values to represent them in some way—a value which is supposed to stand for whole group, of which it is a part, as typical of all the values in the group." **—A.E. Waugh**
- "An average is a typical value in the sense that it is sometimes employed to represent all the individual values in a series or of a variable." **—Ya-Lun-Chou**
- "The average is sometimes described as a number which is typical of the whole group." **—Leabo**
- "An average value is a single value within the range of the data that is used to represent all of the values in the series. Since an average is somewhere within the range of the data, it is also called a measure of central value." **—Croxton & Cowden**

It is clear from the above definitions that an average is a single value that represents a group of values. Such a value is of great significance because it depicts the characteristic of the whole group. Since an average represents the entire data, its value lies somewhere in between the two extremes, *i.e.*, the largest and the smallest items. For this reason an averages are sometimes called 'Measure of central tendency' or 'Measures of location'.

## 2.4 OBJECTIVES OF AVERAGING

There are two main objectives of the study of Averages:

To get single value that describes the characteristics of the entire group- Measures of central tendency, by condensing the mass of data in one single value, enable us to get a bird's eye view of the entire data. Thus one value can represent thousands, lakhs and even millions of values. For example, it is impossible to remember the individual incomes of millions of earning people of India and even if one could do it there is hardly any use. But if the average income is obtained by dividing the total national income by total population we get one single value that represents the entire population. Such a figure would throw light on the standard of living of an average Indian.

1. To facilitate comparison- Measure of central value, by reducing the mass of data to one single figure, enable comparison to be made. Comparison can be made either at a point of time or over a period of time. For example, we can compare the percentage results of the students of various colleges in a certain examination, say B.Com for 2005, and thereby conclude which college is the best or we can compare the pass percentage of the same college for different time periods and thereby conclude as to whether the results are improving or deterioration. Such comparisons are of immense help in framing suitable and timely policies. For example, if the pass percentage of students in college A in B.Com was 80 in 2004 and 75 in 2005, the authorities have sufficient reason for investigating the possible cause of the deterioration in results.

## 2.5 TYPES OF AVERAGES

The following are the important types of averages-

- Arithmetic mean a) simple, and b) weighted
- Median
- Mode
- Geometric mean
- Harmonic mean

Besides these, there are less important averages like moving average, progressive average etc. these averages have a very limited field of application and are, therefore, not so popular.

## 2.6 SIMPLE MEAN AND WEIGHTED MEAN

The most popular and widely used measure of representing the entire data by one value is what most laymen call an 'average' and what the statisticians call the arithmetic mean. Its value is obtained by adding together all the items and by dividing this total by the number of items. Arithmetic mean may either be:

(i) Simple arithmetic mean, or

(ii) Weighted arithmetic mean.

**Calculation of Simple Arithmetic Mean—Individual Observations**

The process of computing mean in case of individual observations (i.e., where frequencies are not given) is very simple. Add together the various values of the variable and divide the total by the number of items. Symbolically:

Here = Arithmetic Mean, $\Sigma X$ = Sum of all the values of the variable X, i.e. $X_1$, $X_2$, $X_3$ ...... $X_n$; N = Number of observations.

*Steps.* The formula involves two steps in calculating mean:

(i) Add together all the values of the variable X and obtain the total, *i.e.*, ΣX.

(ii) Divide this total by the number of the observations, *i.e.*, N.

**Illustration 1.** The following table gives the monthly income of 10 employees in an office :

Income 1,780 1,760 1,690 1,750 1,840 1,920 1,100 1,810 1,050 1,950 (Rs.)

Calculate the arithmetic mean of incomes.

Let income be denoted by the symbol X.

**Solution.**

**Calculation of Arithmetic Mean**

| *Employee* | *Monthly Income (Rs.)* | *Employee* | *Monthly Income (Rs.)* |
|---|---|---|---|
| 1 | 1,780 | 6 | 1,920 |
| 2 | 1,760 | 7 | 1,100 |
| 3 | 1,690 | 8 | 1,810 |
| 4 | 1,750 | 9 | 1,050 |
| 5 | 1,840 | 10 | 1,950 |
| | | N=10 | ΣX=16,650 |

Here, ΣX=16,650, N=10

=1,665

Hence the average income is Rs. 1,665

**Short-cut Method.** The arithmetic mean can be calculated by using what is known as an arbitrary origin. When deviations are taken from an arbitrary origin, the formula for calculating arithmetic mean is

where A is the assumed mean and *d* is the deviation of items from assumed mean, *i.e.*, *d*=(X–A).

**Steps.**

1. Take an assumed mean.
2. Take the deviations of items from the assumed mean and denote these deviations by *d*.
3. Obtain the sum of these deviations, i.e., Σ*d*.
4. Apply the formula:

From Illustration 1 calculate arithmetic mean by taking 1,800 as the assumed mean.

**Solution.**

**Calculation of Arithmetic Mean**

| Employee | Income | (X–1800) |
|---|---|---|
| 1 | 1,780 | –20 |
| 2 | 1,760 | –40 |
| 3 | 1,690 | –110 |
| 4 | 1,750 | –50 |
| 5 | 1,840 | +40 |
| 6 | 1,920 | +120 |
| 7 | 1,100 | –700 |
| 8 | 1,810 | +10 |

| | | |
|---|---|---|
| 9 | 1,050 | –750 |
| 10 | 1,950 | +150 |
| N=10 | | Σd=–1350 |

$A = 1800, \Sigma d = -1350, N = 10$

$= 1800 - = 1,800 - 135 = 1,665$

Hence the average income is Rs. 1,665.

**Note.** The learner will find that the calculations here are more than what we had when we used the formula

$X = A + \Sigma fX/N$

This is true for ungrouped data. But for grouped data considerable saving in time is possible by adopting the short-cut method.

### Calculation of Arithmetic Mean—Discrete Series

In discrete series arithmetic mean may be computed by applying

(i) Direct method, or

(ii) Short-cut method.

### Direct Method

The formula for computing mean is

$X = A + \Sigma fX/N$

Where $f$ = Frequency ; X = The variable in question; N = Total number of observations, *i.e.*, $\Sigma f$.

*Steps:*

(i) Multiply the frequency of each row with the variable and obtain the total $\Sigma fX$.

(ii) Divide the total obtained by step (i) by the number of observations, i.e., total frequency.

**Illustration 2.** From the following data of the marks obtained by 60 students of a class, calculate the arithmetic mean :

| Marks | No. of Students | Marks | No. of Students |
|---|---|---|---|
| 20 | 8 | 50 | 10 |
| 30 | 12 | 60 | 6 |
| 40 | 20 | 70 | 4 |

Let the marks be denoted by X and the number of students by *f*.

**Solution.**

**Calculation of Arithmetic Mean**

| **Marks**<br>**X** | **No. of Students**<br>***f*** | ***f*X** |
|---|---|---|
| 20 | 8 | 160 |
| 30 | 12 | 360 |
| 40 | 20 | 800 |
| 50 | 10 | 500 |
| 60 | 6 | 360 |
| 70 | 4 | 280 |
| | N = 60 | *f* X = 2,460 |

Mean= $\Sigma f$X/N =2460/60= 41

Hence the average marks = 41.

**Short-cut Method.** According to this method,

X= A+$\Sigma$ *f* d/N

Where A = Assumed mean; d=(X – A); N = Total Number of observations, i.e., Σ*f*.

*Steps:*

(i) Take an assumed mean.

(ii) Take the deviations of the variable X from the assumed mean and denote the deviations by *d*.

(iii) Multiply these deviations with the respective frequency and take the total Σ*fd*.

(iv) Divide the total obtained in third step by the total frequency.

**Illustration 3.** Calculate arithmetic mean by the short-cut method using frequency distribution of illustration 2.

**Solution.**

**Calculation of Arithmetic Mean**

| **Marks**<br>**X** | **No. of Students**<br>***f*** | **(X–40)**<br>***d*** | ***fd*** |
|---|---|---|---|
| 20 | 8 | –20 | –160 |
| 30 | 12 | –10 | –120 |
| 40 | 20 | 0 | 0 |
| 50 | 10 | +10 | +100 |
| 60 | 6 | +20 | +120 |
| 70 | 4 | +30 | +120 |
| | **N = 60** | | **Σ*fd* = 60** |

X= A+Σ *f* d/N

= 40+60/60= 40+1 = 41

## Calculation of Arithmetic Mean—Continuous Series

In continuous series, arithmetic mean may be computed by applying any of the following methods:

(i) Direct Method.

(ii) Short-cut method.

### Direct Method

When direct method is used

$X= \Sigma fm/N$

where $m$ = mid-point of various classes; $f$ = the frequency of each class. N = the total frequency.

***Steps :***

(i) Obtain the mid-point of each class and denote it by $m$.

(ii) Multiply these mid-points by the respective frequency of each class and obtain the total $\Sigma fm$.

(iii) Divide the total obtained in step (i) by the sum of the frequency, *i.e.* N.

**Illustration 4.** From the following data compute arithmetic mean by direct method :

| Marks | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 |
|---|---|---|---|---|---|---|
| No. of Students | 5 | 10 | 25 | 30 | 20 | 10 |

**Solution.**

**Calculation of Arithmetic Mean by Direct Method**

| Marks | Mid-point | No. of Students | |
|---|---|---|---|
| | *m* | *f* | *fm* |
| 0-10 | 5 | 5 | 25 |
| 10-20 | 15 | 10 | 150 |
| 20-30 | 25 | 25 | 625 |
| 30-40 | 35 | 30 | 1,050 |
| 40-50 | 45 | 20 | 900 |
| 50-60 | 55 | 10 | 55 |
| | | N = 100 | $\Sigma fm$ = 3,300 |

X= $\Sigma fm/N$

= 3300/100=33

### Short-cut Method

When short-cut method is used, arithmetic mean is computed by applying the following formula :

where A = assumed mean; *d* = deviations of mid-points from assumed mean, *i.e.* (*m* – A); N = total number of observations.

*Steps :*

(i) Take an assumed mean.

(ii) From the mid-point of each class deduct the assumed mean.

(iii) Multiply the respective frequencies of each class by these deviations and obtain the total $\Sigma fd$.

(iv) Apply the formula :

Calculate arithmetic mean by the short-cut method from the data of Illustration 4.

**Solution.**

**Calculation of Arithmetic Mean**

| **Marks** | **Mid-point**<br>***m*** | **No. of Students**<br>***f*** | **(m–35)**<br>***d*** | ***fd*** |
|---|---|---|---|---|
| 0-10 | 5 | 5 | –30 | –150 |
| 10-20 | 15 | 10 | –20 | –200 |
| 20-30 | 25 | 25 | –10 | –250 |
| 30-40 | 35 | 30 | 0 | 0 |
| 40-50 | 45 | 20 | +10 | +200 |
| 50-60 | 55 | 10 | +20 | +200 |
| | | **N = 100** | | **$\Sigma fd$ = –200** |

$$X = A + \Sigma fd/N = 35 + (-200)/100 = 35 - 2 = 33$$

In order to simplify the calculations, we can divide the deviations by class intervals i.e. calculate (m-A)/I and then multiply by i in the formula for getting mean. The formula becomes:

$$X = A + \Sigma fd / N * i$$

It may be pointed out that when class intervals are unequal we can simplify calculations by taking a common factor. In such a case we should use (m-A)/c instead of (m-A)/i while making calculations.

**Step Deviation Method**

When step deviation method is used, arithmetic mean is computed by applying the following formula:

$$X = A + \Sigma fd/N * i$$

where A= assumed mean, d = deviations of mid-points from assumed mean, I = with of the class interval and N = total number of observations.

Steps:

1. Take an assumed mean.
2. From the mid-point of each class deduct the assumed mean.
3. Multiply the respective frequencies of each class of these deviations and obtain the total Σfd.
4. Apply the formula: A + Σfd/N * i

**From the data of illustration 4, Compute Arithmetic Mean by Step Deviation Method**

**Solution**

**Calculation of Arithmetic Mean**

| **Marks** | **mid-point**<br>**m** | **No. of Students**<br>***f*** | **(m-35)**<br>**d** | **(m-35)/10** | **fd** |
|---|---|---|---|---|---|
| 0-10 | 5 | 5 | -30 | -3 | -15 |
| 10-20 | 15 | 10 | -20 | -2 | -20 |
| 20-30 | 25 | 25 | -10 | -1 | -25 |
| 30-40 | 35 | 30 | 0 | 0 | 0 |
| 40-50 | 45 | 20 | 10 | 1 | 20 |
| 50-60 | 55 | 10 | 20 | 2 | 20 |
| | | N=100 | | Σfd= -20 | |

X= A+ Σfd/N * i = 35-20/100*10= 35-2 = 3

**Merits and Demerits of Arithmetic Mean**

**Merits**

1. It is the simplest average to understand and easiest to compute.
2. It is affected by the value of every item in the series.
3. It is defined by a rigid mathematical formula with the result that

everyone who computes the average gets the same answer.

4. It is calculated value and not based on position in the series.

### Demerits

1 In a distribution with open ended classes, the value of mean cannot be computed without making assumptions regarding the size of the class interval in the open ended classes. If such classes contain a large proportion of the values, them mean may be subject to substantial error. However, the values of the median and mode can be computed when there are open ended classes without making any assumptions about size of class interval.

2 The arithmetic mean is not always a good measure of central tendency. The mean provides a 'characteristic' value in the sense of indicating where most of the values lie, only when the distribution of the variable is reasonably normal (bell shaped). In case of a U-shaped distribution the mean is not likely to serve a useful purpose.

### Weighted Arithmetic Mean

One of the limitations of the arithmetic mean discussed above is that it gives equal importance to all the items. But there are cases where the relative importance of the different items is not the same. When this is so, we compute weighted arithmetic mean. The term 'weight' stands for the relative importance of the different items. The formula for computing weighted arithmetic mean is:

Where represents the weighted arithmetic mean; X represents the variable values, *i.e.*, $X_1$, $X_2$, ......., $X_n$.

W represents the weights attached to variable values, *i.e.*, $w_1$, $w_2$, ..., $w_n$, respectively.

(i) Multiply the weights by the variable X and obtain the total ΣWX.

(ii) Divide this total by the sum of the weights, i.e., ΣW.

In case of frequency distribution, if $f_1$, $f_2$, ........., $f_n$ are the frequencies of the variable values $X_1$, $X_2$, ..... $X_n$ respectively then the weighted arithmetic mean is given by : X= ΣW(fx)/ ΣW

Form the expanded form

$X= W_1(fx_1)+W_2\ (f_2x_2)+\ldots\ldots W_n\ (f_nx_n)/\ \Sigma W_1+W_2+\ldots\ldots+W_n$

An important problem that arises while using weighted mean is regarding selection of weights. Weights may be either actual or arbitrary, *i.e.*, estimated. Needless to say, if actual weights are available, nothing likes this. However, in the absence of actual weights, arbitrary or imaginary weights may be used. The use of arbitrary weights may lead to some error, but it is better than no weights at all. In practice, it is found that if weights are logically assigned keeping the phenomena in view, the error involved will be so small that it can be easily over looked.

It should be noted that:

(i) Simple arithmetic mean shall be equal to the weighted arithmetic mean if the weights are equal. Symbolically,

If W1 = $W_2$

(ii) Simple arithmetic mean shall be less than the weighted arithmetic mean if and only if greater weights are assigned to greater values and smaller weights are assigned to smaller values. Symbolically,

If $(w_2-w_1)\ (X_1-X_2) < 0$

(iii) Simple arithmetic mean is greater than the weighted arithmetic mean if and only if smaller weight is attached to the higher values and greater weight is attached to the smaller values. Symbolically,

If $(w_2-w_1)\ (X_1-X_2) < 0$

It may be noted that weighted arithmetic mean is specially useful in problems relating to:

(i) Construction of index numbers, and

(ii) Standardized birth and death rates.

**Illustration 5.** (*a*) A contractor employs three types of workers—male, female and children. To a male he pays Rs. 40 per day, to a female worker Rs. 32 per day and to a child worker Rs. 15 per day. What is the average wage per day paid by the contractor?

**Solution.** The average wage is not the simple arithmetic mean, i.e., =Rs. 29 per day. If we assume that the number of male, female and child workers is the same, this answer would be correct. For example, if we take 10 workers in each case then the mean wage would be

However, the number of male, female and child workers employed is generally different. If we know how many workers of each type are employed by the contractor in question, nothing like this. However, in the absence of this we take assumed weights. Let as assume that the number of male, female and child workers employed is 20, 15 and 5 respectively. The average wage would be the weighted mean calculated as follows :

| **Wages per day (Rs.)**<br>**X** | **No. of Workers**<br>**W** | **WX** |
|---|---|---|
| 40 | 20 | 800 |
| 32 | 15 | 480 |
| 15 | 5 | 75 |
| | ΣW=40 | ΣWX=1,355 |

=33.875 or 33.88

## 2.7 GEOMETRIC MEAN

Geometric mean is defined as the N*th* root of the product of N items or values. If there are two items, we take the square root; if there are three items, the cube root; and so on. Symbolically‘

Where $X_1$, $X_2$, $X_3$, etc., refer to the various items of the series.

Thus the geometric mean of 3 values 2, 3, 4, would be :

= 2.885

When the number of items is three or more the task of multiplying the numbers and of extracting the root becomes excessively difficult. To simplify calculations logarithms are used. Geometric mean then is calculated as follows:

$$\log G.M. = \log X_1 + \log X_2 + \ldots\ldots\ldots + \log X_n / N$$

or $\log G.M. = \Sigma \log X/N$

$\Sigma$ G.M. $=$Antilog $\Sigma \log X/N$

In discrete series G.M. = Antilog $\Sigma f \log X/N$

In continuous series G.M. = Antilog $\Sigma f \log X/N$

**Properties of Geometric Mean**

The following are two important mathematical properties of geometric mean :

- The product of the value of series will remain unchanged when the value of geometric mean is substituted for each individual value. For example, the geometric mean for series 2, 4, 8 is 4 ; therefore, we have

  $$2 \times 4 \times 8 = 64 = 4 \times 4 \times 4$$

- The sum of the deviations of the logarithms of the original observations above or below the logarithm of the geometric mean is equal. This also means that the value of the geometric mean is such as to balance the ratio deviations of the observations from it. Thus, using the same previous numbers, we find that

  Because of this property, this measure of central value is especially adapted to average ratios, rates of change, and logarithmically distributed series.

**Calculation of Geometric Mean-Individual Observation**

G.M. = Antilog Σ*f*logX/N

*Steps :*

(i) Take the logarithms of the variable X and obtain the total Σ log X.

(ii) Divide Σ log X by N and take the antilog of the value so obtained. This gives the value of geometric mean.

**Illustration 6.** Daily income of ten families of a particular place is given below. Find out G.M.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 85 | 70 | 15 | 75 | 500 | 8 | 45 | 250 | 40 | 36 |

**Solution.**

**Calculation of Geometric Mean**

| X | Log X | X | Log X |
|---|---|---|---|
| 85 | 1.9294 | 8 | 0.9031 |
| 70 | 1.8451 | 45 | 1.6532 |
| 15 | 1.1761 | 250 | 2.3979 |
| 75 | 1.8751 | 40 | 1.6021 |
| 500 | 2.6990 | 36 | 1.5563 |
| | | | Σ log X = 17.6373 |

## 1.7 Geometric Mean- Discrete Series

G.M = Antilog (Σ f log x / N)

Steps:

1. Find the logarithms of the variable X.

2. Multiply these logarithms with the respective frequencies and obtain the total f log x.

3. Divide f log x by the total frequency and take the antilog of the value so obtained

$$G.M = \text{Antilog } (\Sigma f \log x / N)$$

**Illustration 7**: Find the geometric mean for the data given below:

| Marks | 4-8 | 8-12 | 12-16 | 16-20 | 20-24 | 24-28 | 28-32 | 32-36 | 36-40 |
|---|---|---|---|---|---|---|---|---|---|
| Frequency | 6 | 10 | 18 | 30 | 15 | 12 | 10 | 6 | 2 |

**Calculation of Geometric Mean**

| **Marks** | **mid-point** | **log m** | **f x log m** | |
|---|---|---|---|---|
| | **m** | **f** | | |
| 4-8 | 6 | 6 | 0.7782 | 4.6692 |
| 8-12 | 10 | 10 | 1.0000 | 10.000 |
| 12-16 | 14 | 18 | 1.1461 | 20.6298 |
| 16-20 | 18 | 30 | 1.2553 | 37.6590 |
| 20-24 | 22 | 15 | 1.3424 | 20.1360 |
| 24-28 | 26 | 12 | 1.4150 | 16.9800 |
| 28-32 | 30 | 10 | 1.4771 | 14.7710 |
| 32-36 | 34 | 6 | 1.5315 | 9.1890 |
| 36-40 | 38 | 2 | 1.5798 | 3.1596 |
| N=109 | | | Σ f * log m= 137.1936 | |

$$G.M= A.L (\Sigma f \log x / N) = A.L ( 137.1936 / 109) = A.L\ 1.2587 = 18.1$$

## 2.8 SUMMARY

Central value condenses the mass data in one single value. Central value enable us to get a bird's eye view of entire data. Thus one value can represent thousands, lakhs and even millions of values. So, this one value is very important because it depicts the characteristics of the whole group. Moreover, this value lies somewhere in between the two extreme values.

## 2.9 SELF ASSESSMENT EXERCISES

1. Define average with the help of example.
2. Explain the meaning of weighted mean.
3. Explain with the help of data, meaning of geometric mean.
4. Calculate average from the following data:

| Marks | Students |
|---|---|
| 0-10 | 5 |
| 10-20 | 10 |
| 20-30 | 15 |
| 30-40 | 20 |
| 40-50 | 25 |
| 50-60 | 30 |
| 60-70 | 35 |

Ans 45

5.

| Marks | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 |
|---|---|---|---|---|---|---|---|
| No. of Students | 5 | 12 | 30 | 45 | 50 | 37 | 21 |

Ans 40.9

## 2.10 SUGGESTED READINGS

1. Yule and Kendall : Introduction to the Theory of Statistics.
2. S.P. Gupta : Statistical Methods. S. Chand and Sons, New Delhi.

**C. No. : BCG-502** **UNIT : I**

**SEMESTER: V** **LESSON: 3**

# MEDIAN, MODE AND PARTITION VALUES

**Structure**

3.1 Introduction

3.2 Objectives

3.3 Median

3.4 Mode

3.5 Partition Values

3.6 Harmonic Mean

3.7 Summary

3.8 Self Assessment Exercises

3.9 Suggested Readings

## 3.1 INTRODUCTION

In the last lesson Mean and Geometric Mean were covered. In this lesson Mode, Median and Harmonic Mean, the remaining three measures of central tendency are covered. First their concept is explained along with the requisite formula and then some examples are given to illustrate their applications. Also discussed in Section 5.6 are various partition values like Deciles, Percentiles and Quartiles. In Section 5.9 are given some exercises for practice of the students.

## 3.2 OBJECTIVES

After studying this lesson, you should be able to :

- understand the concept of Median, Mode and Harmonic Mean.
- calculate Median, Mode and Harmonic Mean.
- understand the concept and compute partition values.

## 3.3 MEDIAN

The median by definition refers to the middle value in a distribution. It is a positional average. It is that value which divides the data into two equal parts when it is arranged in increasing or decreasing order.

For example, if the age of five employees is Rs. 8000, Rs. 7500, Rs. 16000, Rs. 7800, and Rs. 6000, then the median would be Rs. 7800 as Rs. 6000, Rs. 7500 are below it and Rs. 8000, Rs. 16000 are above it. In case number of observations is even, then median is taken to be arithmetic mean of two middle most values. For example, if the data is 16000, 18000, 12000, 10000, 22000, 15000, 20000 and 19000, then the median would be average of 16000 and 18000 i.e. 17000.

**Calculation of Median- Discrete Series**

median = Size of th item. For example, if given

| *x* | *f* | *c.f.* |
|---|---|---|
| 800 | 16 | 16 |
| 1000 | 24 | 40 |
| 1500 | 26 | 66 |
| 1800 | 30 | 96 |
| 2000 | 20 | 116 |
| 2500 | 6 | 122 |
|  | N = 122 |  |

Median = 61.5th item *i.e.* 1500

In case of frequency distribution (continuous) the median is obtained by the formula.

Median = L + N/2 – c.f / f * i

Where L is lower limit of median class i.e. the class in which the middle item of the distribution lies, *c.f.* is cumulative frequency of the class preceding the median class, *f* is simple frequency of the median class and *i* is class interval of the median class.

For example, if we have to calculate median for the following frequency distribution:

| **Marks** | **No. of students** |
|---|---|
| 5–10 | 7 |
| 10–15 | 15 |
| 15–20 | 24 |
| 20–25 | 31 |
| 25–30 | 42 |
| 30–35 | 30 |
| 35–40 | 26 |
| 40–45 | 15 |
| 45–50 | 10, |

Then first obtain table as follows:

| ***C.I*** | ***f*** | ***c.f.*** |
|---|---|---|
| 5–10 | 7 | 7 |
| 10–15 | 15 | 22 |
| 15-20 | 24 | 46 |

|  |  |  |
|---|---|---|
| 20–25 | 31 | 77 |
| 25–30 | 42 | 119 |
| 30–35 | 30 | 149 |
| 35–40 | 26 | 175 |
| 40–45 | 15 | 190 |
| 45–50 | 10 | 200 |
| Total | N = 200 |  |

Here N = 200, and  = 100th item i.e. median class is 25–30. Hence, L = 25, C.F. = 77, $f$ = 42, $i$ = 5 and Median = 25+×5

= 27.74 Ans.

**Merits and Limitations**

**Merits :**

1. It is especially useful in case of open-end classes since only the position and not the values of the items must be known.
2. Extreme values do not affect the median as strongly as they do the mean.
3. It is most appropriate average in dealing with qualitative data.
4. Median can be determined graphically whereas the mean cannot be graphically ascertained.

**Limitations:**

1. For calculating median, it is necessary to arrange the data; other averages do not need any arrangement.
2. Since it is a positional average, its value is not determined by each and every observation.
3. It is not capable of algebraic treatment.

4. It is erratic if the number of items is small.

## 3.4 MODE

The mode or the modal value is that value in a series of observations which occurs with the greatest frequency. For example, the mode of the series 30, 50, 80, 50, 40, 50, 90, 30, 50, 60, 70, 50, 80, 50, 40, 30 would be 50, since this value occurs more frequently than any of the others.

A set of data may have a single mode in which case it is said to be unimodal. It may have two modes which make it bimodal or it may have several modes and be called multimodal.

There are many situations in which authentic mean and median fail to reveal the true characteristic of data. For example, when we talk of most common wage, most common income, most common size of shoe or readymade garments we have in mind mode and not mean or median.

For discrete series or raw data it is very easy to find mode while for continuous series we apply the formula

Mode= L+ $f_m$-$f_1$/ 2 $f_m$- $f_1$ $f_2$*i where,

L = Lower limit of modal class (class with highest frequency)

$f_m$ = frequency of modal class,

$f_1$ = frequency of class preceding modal class,

$f_2$ = frequency of class succeeding modal class

and

$i$ = class interval of the modal class

When mode is ill defined, its value may be ascertained by the following formula based upon the relationship between mean, median and mode:

Mode = 3 Median–2 Mean

This measure is called the empirical mode.

**Illustration 1**

**Calculate mode from the following data**

| Age (in years) | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 | 80-90 |
|---|---|---|---|---|---|---|---|---|
| No. of persons | 5 | 9 | 13 | 21 | 20 | 15 | 8 | 3 |

**Solution:**

**GROUPING TABLE**

| Age (in years) | I | II | III | IV | V | VI |
|---|---|---|---|---|---|---|
| 10-20 | 5 | | | | | |
| | | 14 | | | | |
| 20-30 | 9 | | | 27 | | |
| | | | 22 | | | |
| 30-40 | 13 | | | | **43** | |
| | | 34 | | | | |
| 40-50 | **21** | | | | | **54** |
| | | | **41** | | | |
| 50-60 | 20 | | | **56** | | |
| | | **35** | | | **43** | |
| 60-70 | 15 | | | | | |
| | | | 23 | | | 26 |
| 70-80 | 8 | | | | | |
| | | | 11 | | | |
| 80-90 | 3 | | | | | |

## ANALYSIS TABLE

| **Column No.** | Class in which mode is expected | | | |
|---|---|---|---|---|
| | **30-40** | **40-50** | **50-60** | **60-70** |
| **I** | | 1 | | |
| **II** | | | 1 | 1 |
| **III** | | 1 | 1 | 1 |
| **1V** | 1 | 1 | 1 | |
| **V** | | 1 | 1 | 1 |
| **VI** | | | | |
| | **1** | **5** | **5** | **4** |

This is bi-modal series and hence mode shall be determined by the formula:

Mode = 3 Median - 2 Mean

**Calculation of Mean and Median**

| **Age** | **No. of Persons**<br>**f** | **mid-point**<br>**m** | **(m-55 / 10)**<br>**d** | **fd** | **c.f** |
|---|---|---|---|---|---|
| 10-20 | 5 | 15 | -4 | -20 | 5 |
| 20-30 | 9 | 25 | -3 | -27 | 14 |
| 30-40 | 13 | 35 | -2 | -26 | 27 |
| 40-50 | 21 | 45 | -1 | -21 | 48 |
| 50-60 | 20 | 55 | 0 | 0 | 68 |
| 60-70 | 15 | 65 | 1 | 15 | 83 |
| 70-80 | 8 | 75 | 2 | 16 | 91 |
| 80-90 | 3 | 85 | 3 | 9 | 4 |
| | N = 94 | | | Σfd =-54 | |

Calculation of Mean = A + Σfd/N * i

A = 55, Σfd = -54, N = 94, I = 10

Mean = 55- 54/94*10 = 55 – 5.75 = 49.25

**Calculation of Median**= size of N/2th item = 94/2 = 47th item. It lies in the class 40-50

Median = L + N/2 – c.f / f * i

L = 40, N/2 = 47, c.f = 27, f=21, I = 10

Median = 40 + 47-27/21 * 10 = 40 + 9.52 =- 49.52

Mode= 3 (49.52) – 2 (49.25) = 148.56 – 98.5 = 50.06

**Merit and Limitations of Mode**

**Merits :**

1. Mode is the most typical or representative value of a distribution.
2. Its value can be determined in open-end distributions without ascertaining the class limits.
3. It can be used to describe qualitative phenomenon.
4. Value of mode can also be determined graphically.

**Limitations**

1. The value of mode cannot always be determined.
2. It is not capable of algebraic manipulations.
3. Its value is not based on each and every item of the series.

4 It is not a rigidly defined measure.

## 3.5 PARTITION VALUES

The partition values are those which divide data into equal parts. They are relational positional measures. Just like Median which divides the series into two equal parts, other partition values are Quartiles, Deciles and Percentiles.

**Quartiles:** The three values which divide a data/series into four equal parts are known as Quartiles. They are denoted by Q1, Q2 and Q3. Q1 stands for first quarter i.e. there are 25% values of the series which are below it and 75% values which are above it. Similarly, Q2 implies that 50% values are below it and 50% values are above it and Q3 implies that 75% values are below it and 25% are above it. Mathematically, for discrete series.

Q1 = Size of th item

Q2 = Size of th item

Q3 =Size of th item

And for Continuous series (Grouped distribution),

Q*h* = L+hN/4-c.f/f×*i* ; *h* = 1, 2, 3,

Where by putting *h* = 1, 2 and 3, respectively we shall obtain Q1, Q2 and Q3. Here L is lower limit of Quartile class, *f* is frequency of Quartile class, N = ?*fi* and C.F is cumulative frequency of class proceeding Quartile class and *i* is width of class interval.

**Deciles:** Deciles are those values which divides the data/series into 10 equal parts. Thus, there are nine deciles namely, D1, D2, D3, ……., D9; and the general formula for obtaining them in case of continuous series (Grouped distribution) is

Dh = L+hN/10-c.f/f×*i* ; *h* = 1, 2, 3, 4, ……, 9;

Where L is Lower limit of decile class, *f* is frequency of decile class, *c.f.* is cumulative frequency of class preceding decile class, N = ?*fi* and *i* is width of class interval.

**Percentiles:** Percentiles are those values which divide the data/series into

100 equal parts. Thus, there are 99 percentiles namely,

P1, P2,......,P99; and they can be calculated using the following general formula in case of continuous series (Grouped distribution) as :

$$Ph = L+hN/100\text{-}c.f/f\times i;\ h = 1, 2, \ldots, 99;$$

Where, L is lower limit of percentile class, f is frequency of percentile class, c.f. is cumulative frequency of class preceding percentile class, *i* is width of class interval and N = Σ*fi*.

**ILLUSTRATION 2**.

1. Calculate Median and Mode from the following data :

| **Value** | **Frequency** | **Value** | **Frequency** |
|---|---|---|---|
| Less than 10 | 4 | Less than 50 | 96 |
| Less than 20 | 16 | Less than 60 | 112 |
| Less than 30 | 40 | Less than 70 | 120 |
| Less than 40 | 76 | Less than 80 | 125 |

**Solution:**

Since cumulative frequencies are given, we first find simple frequencies :

**Calculation**

| **Value** | **c.f.** | **Frequency** |
|---|---|---|
| 0–10 | 4 | 4 |
| 10–20 | 16 | 12 |
| 20–30 | 40 | 24 |
| 30-40 | 76 | 36 |
| 40–50 | 96 | 20 |
| 50–60 | 112 | 16 |

| | | |
|---|---|---|
| 60–70 | 120 | 8 |
| 70–80 | 125 | 5 |

Median = Size of th item

= Size of th item

= 62·5th item.

Hence Median class is 30–40.

∴ Median = 36.25

And for mode, we observe that highest frequency is for class 30–40 and thus it becomes modal class.

Hence

Mode = 34.28

### 3.6 Harmonic Mean

The harmonic mean is based on the reciprocals of numbers averaged. It is based as the reciprocals of the arithmetic mean of the reciprocal of the individual observations. Thus, by definition

$$H.M = N / ( 1/x_1 + 1/x_2 +1/ x_3 +.......+ 1/x_n$$

Where $x_1$, $x_2$, $x_3$ ...$x_n$ refers to the various items of the variable

**Illustration3:** Find the harmonic mean from the following

2574 475 75 5 0.8 0.08 0.005 0.0009

**Solution**

**Calculation of Harmonic Mean**

| X | 1/X |
|---|---|
| 2574 | 0.0004 |
| 475 | 0.0021 |

|  |  |
|---|---|
| 75 | 0.0133 |
| 5 | 0.2000 |
| 0.8 | 1.2500 |
| 0.08 | 12.5000 |
| 0.005 | 200.0000 |
| 0.0009 | 1111.1111 |
|  | **Σ(1/x)= 1325.0769** |

**H.M = N**/ Σ ( 1/x) = 8 / 1325.0769 = 0.006

**Illustration4:** Calculate Harmonic Mean, given

| **Class Interval :** | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 |
|---|---|---|---|---|---|
| **Frequency :** | 4 | 6 | 10 | 7 | 3 |

**Solution :**

**Calculation**

| Class Interval | *f* | *x* | *f/x* |
|---|---|---|---|
| 10–20 | 4 | 15 | 0.267 |
| 20–30 | 6 | 25 | 0.240 |
| 30–40 | 10 | 35 | 0.286 |
| 40–50 | 7 | 45 | 0.156 |
| 50–60 | 3 | 55 | 0.055 |
|  | **N = 30** |  | **Σ*f*/*x* = 1.004** |

H.M = **N**/ Σ ( 1/x) =30/1.004 = 29.88

**Illustration5:** An automobile driver travels from plain to hill station 100km distance at a speed of 30km/hour. He then makes the return trip at speed of 20 km/hour. What is his average speed over the entire distance?

**Solution:**

Since the harmonic mean is a measure of central tendency for data expressed as rates such as kms. Per hour, kms. Per litre, hours per semester etc.; we use it for this problem.

Here $x1 = 30$, $x2 = 20$, $n = 2$

= 24km/hour.

1. Given

| Marks | Frequency |
|---|---|
| 0–100 | 5 |
| 100–200 | 12 |
| 200–300 | 16 |
| 300–400 | 25 |
| 400–500 | 13 |
| 500–600 | 10 |
| 600–700 | 9 |

Find Quartile 3rd, Decile 5th and Percentile 20th.

**Solution**

**Table**

| Marks | Frequency | Cum. Frequency |
|---|---|---|
| 0–100 | 5 | 5 |
| 100–200 | 12 | 17 |
| 200–300 | 16 | 33 |
| 300–400 | 25 | 58 |
| 400–500 | 13 | 71 |
| 500–600 | 10 | 81 |
| 600–700 | 9 | 90 |
| Total | N = 90 | |

(i) $Q3 = L + 3N/4 – c.f / f * i$

Since = $3N/4 = 67.5$, the Q3 class is 400–500.

Hence, $Q3 = 400+(67.5-58)/13×100$

$= 473.07$

(ii) $D5 = L + 5N/10 – c.f / f * i$

Since =$5N/10 = 45$, the D5 class is 300–400.

Hence, $D5 = 300+(45-33)/25×100$

$= 348$

(iii) $P20 = L + 20N/100 – c.f / f * i$

Since = = 18, the P20 class is 200–300.

Hence

$P20 = 200+(18-17)/16×100$

$= 206.25$

## 3.7 SUMMARY

- **Median**

Central value which divides data into two equal parts when data is arranged in decreasing or increasing order.

**Formula**

Median = $L + N/12 – c.f / f * i$ (for frequency distribution)

- **Mode**

Central value which repeats maximum number of times in a frequency distribution.

**Formula**

Mode = $L+ f_m-f_1/ 2 f_m- f_1 f_2*i$ (for frequency distribution)

In case mode is ill defined,

Mode = 3 Median–2 Mean.

- **Harmonic Mean**

Reciprocal of the arithmetic mean of the reciprocals of individual observations.

**Formula**

H.M. = **N**/ Σ ( f/x)  (for frequency distribution)

- **Partition Values**

**1. Quartiles (Q1, Q2, and Q3) :**

Divides data into four equal parts.

**Formula**

Qh = L + hN/4 – c.f / f ×*i* (*h* = 1, 2, 3).

**2. Deciles (D1, D2 ..., D9)**

Divides data into 10 equal parts.

**Formula**

Dh = L + hN/10 – c.f / ×*i* (*h* = 1, 2 …., 9)

**3. Percentiles (P1, P2 ......, P99)**

Divides data into 100 equal parts.

**Formula**

P*h* = L + hN/100 – c.f / ×*i* (*h* = 1, 2 …., 99).

## 3.8 SELF ASSESSMENT EXERCISES

1. Find Mean and Median and Mode for data :

| **Age** | **No. of Deaths** |
|---|---|
| 0–10 | 16 |
| 10–20 | 18 |
| 20–30 | 20 |
| 30–40 | 22 |
| 40–50 | 18 |
| 50–60 | 50 |
| 60–70 | 24 |
| 70–80 | 12 |

Mean= 42.44

Median= 47.77

2. Find the harmonic mean of the following data :

| Marks: | 10 | 20 | 25 | 40 | 50 |
|---|---|---|---|---|---|
| No. of Students : | 20 | 30 | 50 | 15 | 5 |

Ans= 20.08

3. An aero plane covers four sides of a square at speeds of 1000, 2000, 3000, 4000 kms per hour, respectively. What is the average speed of the plane in its flight around the square? Ans = 1920

4. The geometric mean of 10 observations was calculated as 28.6. It was later discovered that one of the observations was recorded as 23.4 instead of 32.4. Apply appropriate correction and calculate the correct geometric mean.

Ans= 29.54

5.Calculate $Q_1$, $D_5$ and $D_7$, and $P_{90}$ for the following data :

| Loss: | 1000–2000 | 2000–3000 | 3000–4000 |
|---|---|---|---|
| (Rs.) | | | |

|  |  |  |  |  |
|---|---|---|---|---|
| No. of Days: | 5 | 7 | 12 |  |
| Loss:<br>(Rs.) | 4000–5000 | 5000–6000 | 6000–7000 | 7000–8000 |
| No. of Days : | 20 | 14 | 10 | 7 |

Ans= $Q_{1= 3562.5}$, $D_5$= 4675,

$D_7$= 5607.14, P90$_{= 6950}$

6. Calculate Mode from the following data

| Marks | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 |
|---|---|---|---|---|---|---|---|---|
| No. of Students | 2 | 18 | 30 | 45 | 35 | 20 | 6 | 4 |

Ans 36

## 3.9 SUGGESTED READINGS

1. Elehance : Advanced Statistics
2. S.P. Gupta : Statistics
3. Vishwanathan : Business Statistics : An Applied Orientation.

**C. No. : BCG-502** **UNIT : I**

**SEMESTER: V** **LESSON: 4**

# MEASURES OF DISPERSION—I

## Structure

4.1 Introduction

4.2 Objectives

4.3 Concept and Significance of Dispersion

4.4 Absolute and Relative Measures of Dispersion

4.5 Various Measures of Dispersion

4.6 Range

4.7 Quartile Deviation

4.8 Mean Deviation

4.9 Summary

4.10 Self Assessment Exercises

4.11 Suggested readings

## 4.1 INTRODUCTION

In Unit-I, you have learnt about the central tendency, the first main characteristic in analysing any data, which tells us where the centre of the set of data lies but does not tells us how the set of observations (data) is scattered around this central value. Many times the values of central tendency of two or more distributions (observations) may be equal, but the observations/distributions may scatter widely around its central value. In this way, the central value alone cannot describe the distribution adequately. If in a business there is a high degree of

variability in the raw material, then it could not find mass production economical. Suppose an investor is looking for a suitable equity share for investment. While examining the movement of share prices, he should avoid those shares that are highly fluctuating-having sometimes very high prices and at other times going very low. Such extreme fluctuations mean that there is a high risk in the investment in shares. The investor should, therefore, prefer those shares where risk is not so high. In this lesson, you will study about the degree or extent to which data tend to spread around central value (dispersion). There are various measures of dispersion viz., range, quartile deviation, mean deviation and standard deviation. The first three measures of dispersion will be discussed in this lesson and standard deviation will be considered in next lesson.

## 4.2 OBJECTIVES

After studying this lesson, the students will be able to :

- understand the concept of dispersion,
- differentiate between central tendency and dispersion,
- distinguish between absolute and relative measures of dispersion,
- compute several measures of dispersion such as range, quartile deviation and mean deviation for different types of data, and
- choose the use of an appropriate measures under different situations.

## 4.3 CONCEPT AND SIGNIFICANCE OF DISPERSION

Dispersion is a measure of the extent to which the individual item vary from a central value Dispersion is used in two senses, (i) difference between the extreme items of the series and (ii) average of deviation of items from the mean.

**Absolute Measure :** The figure showing the limit or magnitude of dispersion is known as absolute measure and it is shown in the same unit as ;those of the original data, example measures of dispersion in the age of students, their height, weight etc.

**Relative Measure :** For comparative study the concerning absolute measure is divided by the corresponding mean or some other characteristic value to obtain a ratio or percentage, which is known as the relative measure.

In other words, the word dispersion is used to denote the degree of hetrogeneity in the data. The word dispersion is used in two senses in Statistics :

I. The scatteredness of the values of a variable due to variation among themselves is called dispersion.

II. The deviations from a measure of central tendency or any other fixed value are not uniform in their size. The scatteredness of these deviations is also dispersion.

The terms like, dispersion, spread, variation, scatter, deviation give the idea of homogeneity of the data. When we compare two or more series and found that their means are same, we do not consider them similar because their dispersion may differ.

Some important **definitions** of dispersion are given below:

1. "Dispersion is the measure of the variation of the items." **-A.L. Bowley**
2. "The degree to which numerical data tend to spread about an average value is called the variation of dispersion of the data." **-Spiegel**
3. Dispersion or spread is the degree of the scatter or variation of the variable about a central value." **-Brooks & Dick**

4. “The measurement of the scatterness of the mass of figures in a series about an average is called measure of variation or dispersion.”

**-Simpson & Kajka**

It is clear from above that dispersion (also known as scatter, spread or variation) measures the extent to which the items vary from some central value. Since measures of dispersion give an average of the differences of various items from an average, they are also called averages of the second order. An average is more meaningful when it is examined in the light of dispersion. For example, if the average wage of the workers of factory A is Rs. 3885 and that of factory B Rs. 3900, we cannot necessarily conclude that the workers of factory B are better off because in factory B there may be much greater dispersion in the distribution of wages. The study of dispersion is of great significance in practice as could well be appreciated from the following example:

**Example 1.** The daily sales of three different firms (‘) is given in the following table :

| Firm A | Firm B | Firm C |
|---|---|---|
| 40000 | 38500 | 90000 |
| 40000 | 41500 | 15000 |
| 40000 | 39000 | 40500 |
| 40000 | 41000 | 30500 |
| 40000 | 38000 | 25000 |
| 40000 | 42000 | 30000 |
| 40000 | 37500 | 22000 |
| 40000 | 42000 | 42000 |
| 40000 | 40500 | 65000 |
| Mean = 40000 | Mean = 40000 | Mean = 40000 |

Since the average sales of these firms are same, thus we may conclude that the distributions of sales of these firms are similar. But we observe that the variations in the sales are different from firm to firm. Daily sales of Firm A are constant for all the days whereas there is some variation in sales of Firm B and greater amount of variation for

Firm C.

Further, we also observe that in case of Firm B, the deviations of individual sales from the average sale are much smaller than the deviations of firm C. This mean that the average of the deviations from the average sale will be smaller for Firm B as compared to Firm C. In other words, Firm B has smaller dispersion than Firm C.

The measures of dispersion enable comparisons of two or more distributions with regard to their variability. Measuring variability also facilitates the use of the other statistical measures like skewness, kurtosis, correlation, regression, statistical inference etc.

## SIGNIFICANCE AND PROPERTIES OF MEASURING VARIATION

Measures of variation are needed for four basic purposes:

1. Measures of variation point out as to how far an average is representative of the mass. When dispersion is small, the average is a typical value in the sense that it closely represents the individual value and it is reliable in the sense that it is a good estimate of the average in the corresponding universe. On the other hand, when dispersion is large, the average is not so typical, and unless the sample is very large, the average may be quite unreliable.

2. Another purpose of measuring dispersion is to determine nature and cause of variation in order to control the variation itself. In matters of health variations in body temperature, pulse beat and blood pressure are the basic guides to diagnosis. Prescribed treatment is designed to control their variation. In industrial production efficient operation requires control of quality variation the causes of which are sought through inspection is basic to the control of causes of variation. In social

sciences a special problem requiring the measurement of variability is the measurement of "inequality" of the distribution of income or wealth etc.

3. Measures of dispersion enable a comparison to be made of two or more series with regard to their variability. The study of variation may also be looked upon as a means of determining uniformity of consistency. A high degree of variation would mean little uniformity or consistency whereas a low degree of variation would mean great uniformity or consistency.

4. Many powerful analytical tools in statistics such as correlation analysis. The testing of hypothesis, analysis of variance, the statistical quality control, regression analysis is based on measures of variation of one kind or another.

A good measure of dispersion should possess the following properties

1. It should be simple to understand.
2. It should be easy to compute.
3. It should be rigidly defined.
4. It should be based on each and every item of the distribution.
5. It should be amenable to further algebraic treatment.
6. It should have sampling stability.
7. Extreme items should not unduly affect it.

## 4.4. ABSOLUTE AND RELATIVE MEASURES OF DISPERSION

There are two types of measures of dispersion, namely absolute measure of dispersion and relative measures of dispersion. The measures of dispersion which are expressed in terms of the original units of data are termed as Absolute Measures. These measures are not suitable for comparison of various distributions or series expressed in different units of measurements. Relative measures of dispersion, on the other hand, are expressed as ratio or percentage. Therefore, relative measures are pure numbers i.e. independent of the unit of measurement. It is also known as coefficient of dispersion. Thus, the relative measures are useful for comparing variability in two or more distributions where units of measurements may be different .

## 4.5 VARIOUS MEASURES OF DISPERSION

There are five measures of dispersion: Range, Inter-quartile range or Quartile Deviation, Mean deviation, Standard Deviation, and Lorenz curve. Among them, the first four are mathematical methods and the last one is the graphical method. The most common measures of dispersion are :

(i) Range—spread of entire data.

(ii) Quartile Deviation/Inter Quartile Range-spread of middle 50% data.

(iii) Mean Deviation—mean of the absolute deviation from the central value.

(iv) Standard Deviation or Root Mean Square Deviation about arithmetic mean.

The first two measures of dispersion are based on selected items of the data. However, the last two measures of dispersion are based on all the items of the data.

The relative of measures of dispersion corresponding to the absolute measures of dispersion are :

| **Absolute Dispersion** | **RelativeMeasures of Measures of Dispersion** |
|---|---|
| Range | Coefficient of Range |
| Quartile Deviation | Coefficient of Quartile Deviation |
| Mean Deviation | Coefficient of Mean Deviation |

Standard Deviation		Coefficient of Standard Deviation

**NOTE :-** Coefficient of standard deviation when expressed in terms of percentage is called coefficient of variation.

## 4.6 RANGE

Range is the simplest measure of dispersion. For a given set of observations, the range is defined as the difference between the largest (maximum) and smallest (minimum) observation.

Thus,

Range = $X_{max}$ .- $X_{min}$.

where $X_{max}$ = highest value and $X_{min}$.= lowest value.

In case of grouped data, the range is defined as the difference between the upper limit of the highest class and lower limit of the lowest class. The relative measure corresponding to range, called coefficient of range, is obtained by expressing range as the ratio of sum of two extreme values.

**Merits** of Range :

- Simple and easy to be computed.

- It takes minimum time to calculate.

- Not necessary to know all the values, only smallest and largest value is required.

- Helpful in quality control of products.

**Demerits o**f Range :

- Not based on all the items.

- Subject to fluctuation/uncertain measure.

- Cannot be computed in case of open-end distributions.

- As it is not based on all the values it is not considered as a good or appropriate measure.

Thus, the coefficient of range is given by :

$$\text{Coefficient of Range} = \frac{\text{Xmax} \; . - \text{Xmin}}{\text{Xmax} \; . + \text{Xmin}}$$

For example, consider the daily sales data for the three firms as given in example 1. The range for three firms is

For Firm A, Range = 40000 - 40000 = 0

For Firm B, Range = 42000 - 37500 = 4500

For Firm C, Range = 90000 - 15000 = 75000

Thus, on the basis of these values of range, we may conclude, that the variation is zero for Firm A, the variation is small in case of Firm B as compared to Firm C.

**Example 2.** Find the range and coefficient of range from the following data:

| Sales (Rs. In lacs) | No. of Days |
|---|---|
| 20-30 | 8 |
| 30-40 | 3 |
| 40-50 | 18 |
| 50-60 | 16 |
| 60-70 | 14 |
| 70-80 | 12 |
| 80-90 | 10 |

**Solution :**

Range = Xmax .- Xmin

= 90-20

= 70

Coefficient of Range = $\frac{Xmax\ . - Xmin}{Xmax\ . + Xmin}$

$= \frac{90-20}{90+20}$

$= \frac{70}{110}$

= 0.6364

The range is a crude measure of dispersion, since its value depends only on two extreme observations. It provides only some idea about variability of the data. Range is helpful in studying variations in the prices of shares, debentures and agricultural commodities which are very sensitive to price change. The range is a good indicator for weather fore.cast. It is also useful in statistical quality control.

**Inter-Quartile Range** : Inter-quartile range represents the difference between the third-quartile and the first quartile. It is also known as the range of middle 50% values.

Inter-quartile range = Q3 – Q1

**Merits :**

- It is easy to calculate.

- Can be measured in open end distributions.

- It is least affected by the uncertainty of the extreme values.

**Demerits** :

- It does not represent all the values.

- It is an uncertain measure.

- It is very much affected by sampling fluctuations.

iii) Percentile Range : It is the difference between the values of the 90$^{th}$ and 10 th percentile. It is based on the middle 80% items of the series.

Percentile Range = P90 – P10

**Merits and Demerits** : Its use is limited. Percentile range has almost the same merits and demerits as those of inter-quartile range.

## 4.7 QUARTILE DEVIATION

Range is based on two extreme items and it does not take into account the variation within the range. For this reason, quartile deviation is defined. Quartile Deviation gives the average amount by which the two quartiles differ from the median. Quartile deviation is an absolute measure of dispersion. Quartile deviation is defined as half the difference between the upper and lower quartile. Thus, Quartile deviation is defined as

$$\text{Quartile Deviation (Q.D.)} = \frac{Q3 - Q1}{2}$$

where $Q_3$ and $Q_1$ are the third and first quartiles. It is also known Semi-Interquartile Range.

**Merits** :

- It is easy to calculate and understand.

- It has a special utility in measuring variation in open end distributions.

- QD is not affected by the presence of extreme values.

**Demerits** :

- It is very much affected by sampling fluctuations.

- It does not give an idea of the formulation of the series.

- It is not capable of further algebraic treatment.

For comparing two or more distributions in respect of variation, the coefficient of quartile deviation is required which is defined as

$$\text{Coefficient of Quartile Deviation} = \frac{Q3 - Q1}{Q3 + Q1}$$

Quartile deviation is dependent on the two quartiles and does not take into account the variability of the largest 25% and smallest 25% of observations. It is, therefore, unaffected by extreme values. Another advantage of quartile deviation is that it is the only measure of dispersion which can be used for open-end distribution. The main drawback of quartile deviation is that it does not depend upon the magnitudes of all the observations and it is based on the middle 50% of the observations.

**Example 3.** Calculate quartile deviation and coefficient of quartile deviation from the following data :

| | | | | | | |
|---|---|---|---|---|---|---|
| Age group: | 20 -25 | 25-30 | 30-35 | 35-40 | 40-45 | 45-50 |
| No. of Workers: | 50 | 70 | 100 | 180 | 150 | 120 |

**Solution :** For determining the Q.D. and computation of $Q_1$ and $Q_3$

| Classes | No. of Workers<br>(f) | Cum. Freq.<br>c.f. |
|---|---|---|
| 20-25 | 50 | 50 |
| 25-30 | 70 | 120 |
| 30-35 | 100 | 220 |
| 35-40 | 180 | 400 |
| 40-45 | 150 | 55 |
| 45-50 | 120 | 670 |
| **Total** | **670** | |

Its coefficient value, we first determine the values of $Q_1$ and $Q_3$ from the given distribution.

The $Q_1$ is value of N/4$^{th}$= 670/4$^{th}$= 167.5th item.

Thus, 30–35 is the first quartile class. Hence

$$Q_1 = L + \frac{\frac{N}{4} - c.f.}{f} X\ i$$

$$Q_1 = 30 + \frac{167.5 - 120.}{100} X\ 5$$

$$Q_1 = 30 + 2.375$$

$$Q_1 = 32.375 \text{ years}$$

Similarly, $Q_3$ is the value of 3N/4th = (502.5)th item.

Thus, 40–50 is the third quartile class. Hence,

$$Q_3 = L + \frac{\frac{3N}{4} - c.f.}{f} X\ i$$

$$Q_3 = 150 + \frac{502.5 - 400.}{150} X\ 5$$

$$Q_3 = 150 + 3.42$$

$$Q_3 = 153.42 \text{ years}$$

Therefore,

$$Q.D. = \frac{Q3 - Q1}{2}$$

$$= \frac{153.42 - 32.375}{2}$$

= 60.52 years

And

Coefficient of Quartile Deviation= $\frac{Q3 - Q1}{Q3 + Q1}$

$= \frac{121.045}{187.795}$

= 0.6515

**Example 4.** Compare dispersion in the following two series using Q.D.

Series A (Height in Inches) : 58, 56, 62, 61, 63, 64, 65, 59, 62, 65, 55

Series B (Weight in Kgs) : 117, 112, 127, 123, 125, 130, 106, 119, 121, 132, 108

**Solution :** Since the units of measurement in two series are different, the dispersion

can be compared only by using relative measure of dispersion, i.e. coefficient of Q.D. To obtain it, first we find $Q_1$ and $Q_3$

**Computation of $Q_1$ and $Q_3$**

| Sr. No. | Height (A) | Weight (B) |
|---|---|---|
| 1 | 55 | 106 |
| 2 | 56 | 108 |
| 3 | 58 | 112 |
| 4 | 59 | 117 |
| 5 | 61 | 119 |
| 6 | 62 | 121 |
| 7 | 62 | 123 |
| 8 | 63 | 125 |
| 9 | 64 | 127 |
| 10 | 65 | 130 |
| 11 | 65 | 132 |

Computation of $Q_1$ and $Q_3$ for series A :

$Q_1$ = N+1/4 th value = 11+1/4 th value

= $3^{rd}$ value = 58 inches

$Q_3$ = 3N+1/4th value =3x11+1/4 th value= 36/4th value

= 9th value = 64 inches

Thus, coefficient of Q.D. for series A = $\frac{Q3 - Q1}{Q3 + Q1}$

$\frac{64 - 58}{64 + 58}$

= 0.049

Computation of $Q_1$ and $Q_3$ for series B.

$Q_1$ = N+1/4th value = 3rd value

= 112 kgs

$Q_3$ = 3N+1/4 th value = 9th value

= 127 kgs

And coefficient of Q.D. is

$= \frac{Q3 - Q1}{Q3 + Q1}$

$\frac{127 + 112}{127 - 117}$

= 0.063

Since, the coefficient of Q.D. for series B is higher than that of A, hence its dispersion is also higher.

## 4.8 MEAN DEVIATION

As we have discussed in previous sections that both the range and quartile deviation are not ideal measures because they are not based on all the observations. More so, these are not measures of dispersion in the strict sense of the term as they do not measure scatteredness in observations around an average. But, the mean deviation is an ideal measure in this sense as it is based on all the observations of the data. Mean Deviation is also known as average deviation or first measure of dispersion. It is the average difference between the items in a distribution and the median mean or mode of that series.and also it is computed as the arithmetic mean of the absolute deviations of the individual observations from the average of the given data. The average which is frequently used in computing the mean deviation is mean or median, though sometimes mode can also be used. Absolute deviations means the deviations are treated as positive regardless of the actual sign. Symbolically, the mean deviation about mean, median, or mode can be expressed as follows :

(a) Mean deviation from mean $(\bar{X})$

M.D. $(\bar{X}) = \frac{1}{n}\sum|X - \bar{X}|$

(b) Mean deviation from median (Me)

M.D. (Me) $= \frac{1}{n}\sum|X - Me|$

(c) Mean deviation from mode (Mo)

M.D. (Mo) $= \frac{1}{n}\sum|X - Mo|$

In case of frequency distributions or grouped data, the above formulae can be expressed as

M.D. $(\bar{X}) = \frac{1}{n}\sum \mathrm{f}\,|X - \bar{X}|$

M.D. (Me) $= \frac{1}{n}\sum \mathrm{f}|X - Me|$

M.D. (Mo)$= \frac{1}{n}\sum f|X - Mo|$

where *f* is the frequency of a particular class and N be the total of frequencies of all the classes.

**Merits :**

- A major advantage of mean deviation is that it is simple to understand

and easy to calculate.

- It takes into consideration each and every item in the distribution. As a result, a change in the value of any item will have its effect on the magnitude of mean deviation.

- The values of extreme items have less effect on the value of the mean deviation.

- As deviations are taken from a central value, it is possible to have meaningful comparisons of the formation of different distributions.

**Demerits :**

- Signs are ignored therefore mathematically it is incorrect and not a significant measure.

- Cannot be compared if mean deviations of different series are based on different averages.

- Does not give accurate results.

An important property of Mean Deviation is that it has the minimum value when deviations are taken from median. The relative measures corresponding to the mean deviation called the coefficient of Mean Deviation. It is obtained by dividing mean deviation by the particular average used in computing the mean deviation. Thus, coefficient of Mean Deviation from mean is

$$\frac{\text{M.D. } (\bar{X})}{(\bar{X})}$$

Similarly,

$$\text{Coefficient of M.D. } = \frac{\text{M.D. (Me)}}{((Me)}$$

$$\text{Coefficient of M.D.} = \frac{\text{M.D. (Mo)}}{((Mo)}$$

**Example 5.** Compute the mean deviation of the following data from mean, median and mode :

18, 25, 63, 59, 29, 72, 17, 25, 105, 87

**Solution :** In order to determine the mean deviation about mean, median and mode, we first find these values. Since, there are 10 observations which is an even number, thus the median will be the average of two middle most values when data is arranged in order of magnitude :

17, 18, 25, 25, 29, 59, 63, 72, 87, 105

Thus, median is $\frac{\frac{n}{2th} item+n+1/2th\ item}{2} = \frac{5th\ item+6th\ item}{2}$

$= \frac{29 + 59}{2} = 44$

Similarly,

Mean $\bar{X}$ = 1/N “X = 500/10 = 50

and mode is 25.

To obtain the mean deviation about mean, median and mode, we prepare prepare the following table :

| **X** | **lX-**$\bar{X}^{l}$ | **lX-**$M_e$**l** | **lX-**$M_o$**l** |
|---|---|---|---|
| 18 | 32 | 26 | 7 |
| 25 | 25 | 19 | 0 |
| 63 | 13 | 19 | 38 |
| 59 | 9 | 15 | 34 |
| 29 | 21 | 15 | 4 |
| 72 | 22 | 28 | 47 |
| 17 | 33 | 27 | 8 |

|  |  |  |  |
|---|---|---|---|
| 25 | 25 | 19 | 0 |
| 105 | 55 | 61 | 80 |
| 87 | 37 | 43 | 62 |
| Total | 272 | 272 | 280 |

Thus, mean deviation about mean in

M.D. M.D. ($\bar{X}$)= $\frac{1}{n}$“f $|X - \bar{X}|$

= 272/10 = 27.2

Similarly,

M.D. (Me) = $\frac{1}{n}\sum \mathrm{f}|X - Me|$

= 272/10 = 27.2

M.D. (Mo) = $\frac{1}{n}\sum f|X - Mo|$

= 280/10 = 28.0

**Example 6.** For the following data, find the coefficient of mean deviation from median :

| **Sales (` . ‘000)** | **No. of firms** |
|---|---|
| 30-40 | 10 |
| 40-50 | 20 |
| 50-60 | 30 |
| 60-70 | 35 |
| 70-80 | 25 |
| 80-90 | 10 |

**Solution :** To determine the coefficient of mean deviation from median, we first find median and then the coefficient of mean deviation about median.

| **Sales (` '000)** | **Mid Value** | **No. of firms** | ***c.f.*** | **lX-$M_e$l** | **flX-$M_o$l** |
|---|---|---|---|---|---|
| 30-40 | 35 | 10 | 10 | 26.43 | 264.30 |
| 40-50 | 45 | 20 | 30 | 16.43 | 328.60 |
| 50-60 | 55 | 30 | 60 | 6.43 | 192.90 |
| 60-70 | 65 | 35 | 95 | 3.57 | 124.95 |
| 70-80 | 75 | 25 | 120 | 13.57 | 339.25 |
| 80-90 | 85 | 10 | 130 | 23.57 | 235.70 |
| Total | | 130 | | 137.1963 | 1485.70 |

$$\text{Median (Me)} = L+ \frac{\frac{N}{2}-c.f.}{f} X\ i$$

$$=60+ \frac{65-60}{35} X\ 10$$

$$= 60 + \frac{50}{35}$$

$$= 61.43$$

## 4.9 SUMMARY

Dispersion means the spread or the scatteredness of the data. It is also used to describe the average of deviation of items from some measure of central tendency. A good measure of dispersion should be based on all observations, should easily be calculated, least affected by sampling fluctuations and capable to further algebric treatment. The various measures of dispersion in common use are Range, Quartile Deviation, Mean Deviation and Standard Deviation. The first three measures of dispersion have been discussed in this lesson while the fourth one i.e. Standard Deviation will be considered in Lesson 7.

## 4.10 SELF ASSESSMENT EXERCISES

1. What do you understand by dispersion? What is the need of studying it?

2. Explain briefly the essentials of a good measures of dispersion.

3. Calculate mean deviation from median of the prices given below:

Prices (`.): 210, 220, 225. 225. 225, 235, 240. 250, 255, 260.

Also find coefficient of mean deviation.

4. Calculate quartile deviation and mean deviation about mean, median for the following data :

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Age (in years)** : | 20 | 30 | 40 | 50 | 60 | 70 | 80 |
| **No. of members**: | 3 | 61 | 132 | 153 | 140 | 51 | 3 |

5. Compute the range, quartile deviation, mean deviation, coefficient of mean deviation from mean, median and mode of the following data :

| No. of Shares Applied for | No. of Applicants |
|---|---|
| 0-100 | 500 |
| 100-200 | 350 |
| 200-300 | 2500 |
| 300-400 | 1500 |
| 400-500 | 1000 |
| 500-600 | 400 |
| 600-700 | 150 |
| 700-800 | 200 |

## 4.11 SUGGESTED READINGS

1. Elehance : Advanced Statistics
2. S.R Gupta : Statistics
3. Vishwanathan : Business Statistics : An Applied Orientation.

**C. No. : BCG-502** **UNIT : II**

**SEMESTER: V** **LESSON: 5**

# MEASURES OF DISPERSION – II

**Structure**

5.1 Introduction

5.2 Objectives

5.3 Standard Deviation

5.4 Computation of Standard Deviation

5.5 Properties of Standard Deviation

5.6 Summary

5.7 Self Assessment Exercises

5.8 Suggested Readings

## 5.1 INTRODUCTION

In lesson 6, we have discussed the concept of measure of dispersion and we have also leamt about three measures of dispersion viz., Range, Quartile Deviation and Mean Deviation. The range and quartile deviation are based on selected items of the data while mean deviation is based on all the observations of the data. However, in mean deviation we always consider the absolute deviation from central value irrespective of whether the deviation is negative or positive. Now, to consider the sign of these deviations, we have another measure of dispersion which take care of the problem of signs. Thus, in this lesson we shall study about standard deviation, its computation, properties. The coefficient of variation will also be discussed in this lesson. .

## 5.2 OBJECTIVES

The following are the main objectives of this lesson.

- to introduce the concept of standard deviation and coefficient of variation,
- to provide the computational procedure of standard deviation,
- to explain the various properties of standard deviation.
- to explain the various merits and limitations of standard deviation, and
- to make the comparison of the different measures of dispersion.

## 5.3 STANDARD DEVIATION

Standard Deviation was introduced by Karl Pearson in 1823. It is the most important and widely used measure of studying dispersion, as it is free from those defects from which the earlier methods suffer and satisfies most of the properties of a good measure of dispersion. Standard Deviation is the square root of the average of the square deviations from the arithmetic mean of a distribution. It is one of the most popular and important measure of dispersion. It satisfies most of the properties of a good measure of dispersion. The standard deviation is defined as the positive square root of the arithmetic mean of the squares of deviations of the observations from the arithmetic mean. It is also known as "Root Mean Square Deviation" and is generally denoted by Greek letter  (sigma). Symbolically, for ungrouped data or individual series the standard deviation is defined as

$$\sigma = \sqrt{\frac{1}{n}\sum(X - \overline{X})^2}$$

In case of a frequency distribution or grouped data, it is given by

$$\sigma = \sqrt{\frac{1}{n}\sum f\,(X - \overline{X})^2}$$

where N is the sum of the frequencies.

The square of the standard deviation is called variance. It is usually denoted

by $\sigma^2$.

The standard deviation gives us an idea about the extent to which observations are scattered around their mean. Thus, two or more distributions having the same mean can be compared directly for their variability with the help of corresponding standard deviations. Now, the following two situations may arise :

(i) When two or more distributions having unequal means are to be compared in respect of their variability.

(ii) When two or more distributions having observations expressed in different units of measurements are to be compared in respect of their variability.

For making comparisons in the above two situations, we use a relative measure of dispersion, called coefficient of variation (C.V.). It is defined as

$$\text{C.V.} = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$$\text{Or, C.V.} = \frac{\sigma}{\bar{X}} \text{ x } 100$$

Thus, C.V. is a pure number independent of the units of measurements. The distribution/series having greater C.V. is considered to be more variable than the other, and the distribution with lesser C.V, shows greater consistency, homogeneity and uniformity.

**Merits of Standard Deviation** :

- Based on all the items.
- Well-defined and definite measure of dispersion.
- Least affected by sample fluctuations.
- Suitable for algebraic treatment.

**Demerits :**

- Standard Deviation is comparatively difficult to calculate.
- Much importance is given to the extreme values.

## 5.4 COMPUTATION OF STANDARD DEVIATION

**For Ungrouped Data (Individual Series);**

The following two methods arc used for computing standard deviation.

*(i)* Direct Method, *(ii)* Short-Cut Method

*(i)* **Direct Method** : In this procedure, the following formula is used for standard deviation :

$$\sigma = \sqrt{\frac{1}{n}\sum(X-\overline{X})^2}$$

**Example 1.** Calculate the standard deviation from the following data :

X : 10, 11, 17, 25, 7, 13, 21, 10, 12, 14

**Solution :**

| X | X-$\bar{X}$ | (X-$\bar{X}$)$^2$ | X$^2$ |
|---|---|---|---|
| 10 | -4 | 16 | 100 |
| 11 | -3 | 9 | 121 |
| 17 | 3 | 9 | 289 |

| | | | |
|---|---|---|---|
| 25 | 11 | 121 | 625 |
| 7 | -7 | 49 | 49 |
| 13 | -1 | 1 | 169 |
| 21 | 7 | 49 | 441 |
| 10 | -4 | 16 | 100 |
| 12 | -2 | 4 | 144 |
| 14 | 0 | 0 | 196 |
| 140 | | 274 | 2234 |

Here , $n=10$, $\Sigma X= 140$, $= 140/10$, $\Sigma(X-)^2= 274$ and $\Sigma X^2= 2234$

$$\sigma = \sqrt{\frac{1}{n}\Sigma(X-\overline{X})^2}$$

$$= \sqrt{\frac{274}{10}}$$

$$= 5.24$$

**(i) Short-Cut Method :** In most of the cases the mean of the given series happens to be a fractional value and then the process of taking deviations and squaring them becomes quite difficult and time consuming. To overcome this diffifulty, short-cut method can be used which involves deviations from assumed mean. The short-cut formula for S.D. is

**(ii)** $$\sigma = \sqrt{\frac{1}{n}\sum d2-(\sum d/n)^2}$$

where, $d$ is the deviation from assumed mean say A, *i.e.*,

$d = X - A$

$\Sigma d$ = the sum of deviations

$\Sigma d2$ = the sum of squares of deviations

**Example 2.** Find S.D. from the data given below :

X : 20, 22, 27, 30, 31, 32, 35, 40, 45, 51

**Solution :**

| Sr. No. | X | d = X-32 | $d^2$ |
|---|---|---|---|
| 1 | 20 | -12 | 144 |
| 2 | 22 | -10 | 100 |
| 3 | 27 | -5 | 25 |
| 4 | 30 | -2 | 4 |
| 5 | 31 | -1 | 1 |
| 6 | 32 | 0 | 0 |
| 7 | 35 | 3 | 9 |
| 8 | 40 | 8 | 64 |
| 9 | 45 | 13 | 169 |
| 10 | 51 | 19 | 361 |
| Total | 333 | 13 | 877 |

Here n = 10, $\Sigma X =$ 333, $\bar{X} =$ 33.3

Since $\bar{X}$ is a fractional, thus let assumed mean (A) is 32, then

$$\sigma = \sqrt{\frac{1}{n}\sum d2 - (\sum d/n)}^2$$

$$\sigma = \sqrt{877/10 - (13/10)}^2$$

$$= \sqrt{86.01}$$

= 9.27

**Grouped Data :** Any of the following three procedures may be applied to find S.D. for grouped data :

I. Direct Method, II. Short-Cut Method,

III. Step-deviation Method.

**I. Direct method :** In direct method, the formula is used to calculate S.D.

$$\sigma = \sqrt{\frac{1}{n}\sum f(X-\bar{X}}\,)^2$$

or, $$\sigma = \sqrt{\frac{1}{n}\sum fX2-(\sum fX}/n)^2$$

**II. Short-cut Method :** When data is huge or arithmetic mean comes out fractions then it will be more appropriate to use short-cut method to determine S.D. Here

$$\sigma = \sqrt{\frac{1}{n}\sum fd2-(\sum fd}/n)^2$$

where $d$ is the deviation taking from assumed mean 'A', i.e. $d$ = X–A.

**III. Step-Deviation Method :** In short-cut method our main aim is to simplify the deviations so that calculations become easier. In grouped data, specially in continuous frequency distributions, we observe that the calculations can be further simplified if the deviations ($d$) are divided by a common factor, say $h$, which is usually the size of the class intervals. In this method the formula for S.D. is thus becomes :

$$\sigma = h\sqrt{\frac{1}{n}\sum fd2-(\sum d}/n)^2$$

where d= $\frac{X-A}{h}$

**Example 3.** Find the S.D. and C.V. of the following data.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Size (X): | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Frequency : | 6 | 12 | 15 | 28 | 20 | 14 | 5 |

**Solution.**

| X | *f* | *fx* | X-$\bar{X}$ | $f(X-\bar{X})^2$ | d = X-7 | *f*d | *f*d$^2$ |
|---|---|---|---|---|---|---|---|
| 4 | 6 | 24 | -3.06 | 56.1816 | -3 | -18 | 54 |
| 5 | 12 | 60 | -2.06 | 50.9232 | -2 | -24 | 48 |
| 6 | 15 | 90 | -1.06 | 16.8540 | -1 | -15 | 45 |
| 7 | 28 | 196 | -0.06 | 0.1008 | 0 | 0 | 0 |
| 8 | 20 | 160 | 0.94 | 17.6720 | 1 | 20 | 20 |
| 9 | 14 | 126 | 1.94 | 52.6904 | 2 | 28 | 56 |
| 10 | 5 | 50 | 2.94 | 43.2180 | 3 | 15 | 45 |
| Total | 100 | 706 | | 237.6400 | | 6 | 238 |

**Direct Method :** Here N = 100, ΣfX = 706

Since, $\bar{X} = \frac{\Sigma fx}{N} = \frac{706}{100} = 7.06$

Σf = ( X-$\bar{X}$ )$^2$ = 237.64, thus

$$\sigma = \sqrt{\frac{1}{n}\Sigma f(X-\bar{X})^2}$$

$$= \sqrt{\frac{237.64}{100}} = 1.54$$

**Short-Cut Method :** Here d = X–7, $\sum fd$ = 6 and $\sum fd^2$ = 238, thus

$$\sigma = \sqrt{\frac{1}{n}\sum fd2 - (\sum fd/n)^2}$$

$$= \sqrt{\frac{238}{100} - (6/100)^2}$$

$$= \sqrt{2.3764}$$

= 1.54

Thus C.V. = = σ / $\bar{X}$ X 100

= 1.5 / 7.06 x 100

= 21.81

**Example 4.** Use step-deviation method for calculating S.D. of the following distribution:

| | | | | | | |
|---|---|---|---|---|---|---|
| Age group/year : | 25-30 | 30-35 | 35-40 | 40-45 | 45-50 | 50-55 |
| No. of workers : | 10 | 12 | 25 | 40 | 10 | 3 |

**Solutions.**

| **Age group** | **No. of Workers*(f)*** | **Mid Value *X*** | ***d* = $\frac{\mathbf{X-42.5}}{\mathbf{5}}$** | ***fd*** | ***fd*2** |
|---|---|---|---|---|---|
| 25-30 | 10 | 27.5 | -3 | -30 | 90 |
| 30-35 | 12 | 32.5 | -2 | -24 | 48 |
| 35-40 | 25 | 37.5 | -2 | -25 | 25 |
| 40-45 | 40 | 42.5 | 0 | 0 | 0 |
| 45-50 | 10 | 47.5 | 1 | 10 | 10 |
| 50-55 | 3 | 52.5 | 2 | 6 | 12 |
| 706 | 100 | 706 | | -63 | 185 |

Here A = 42.5, $h$ = 5, then

$$\sigma = h\sqrt{\frac{1}{n}\sum fd2 - (\sum fd/n)}^{2}$$

$$\sigma = 5\sqrt{\frac{185}{100} - (-63/100)}^{2}$$

$$= 5\sqrt{1.4531}$$

$$=6.027$$

## 5.5 PROPERTIES OF STANDARD DEVIATION

The following are the main properties of the standard deviation :

I. The value of S.D. remains unchanged if each of the observations in a series is

increased or decreased by a constant value. Thus, if S.D. of X is , then S.D. of Y = X±b

S.D. (Y) = S.D. (X±b)

= S.D. (X)

where b is any constant.

II. If each observation is multiplied or divided by a constant value, then S.D. will also be multiplied or divided by the same constant. Thus, if Y= AX, where A is constant then S.D. of Y = (S.D. of X). A.

II. For a given set of observations, the S.D. is never less than the mean deviation about mean and quartile deviation.

IV. Root mean square deviation calculated about a value other than arithmetic mean will always be higher than S.D.

V. If two groups contain $n_1$ and $n_2$ observations with mean $\bar{X}_1$ and $_2$ and S.D.

$\sigma_1$ and $\sigma_2$ respectively, then the S.D. of the combined group can be determined on using the following formula :

$$\sigma = \frac{\sqrt{n1\,(\sigma 1^2 + d1)^2 + n2\,(\sigma 1^2 + d1)^2}}{n_1+n_2}$$

where

$d_1 = \bar{X}_1 - \bar{X}, d_2 = \bar{X}_2 - \bar{X}$ and

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

## 5.6. SUMMARY

In this lesson we have discussed a measure of dispersion standard deviation, which is considered to be the best measure of dispersion. It possesses all the qualities and properties of a good measure of dispersion. Thus, it is widely used in stastical analysis and treatment.

## 5.7 SELF ASSESSMENT EXERCISES

**1.** What is standard deviation? Explain its superiority over other measures of dispersion.

______________________________________________

______________________________________________

______________________________________________

______________________________________________

______________________________________________

______________________________________________

2. Distinguish between various measures of dispersion and explain their relative merits and demerits.

______________________________________________

______________________________________________

3. What is coefficient of variation? What is its role as a measure of dispersion?

4. Find the different measures of dispersion from the following data :

| **Marks** | **No. of Students** |
|---|---|
| less than 10 | 10 |
| less than 20 | 30 |
| less than 30 | 50 |
| less than 40 | 80 |
| less than 50 | 120 |
| less than 60 | 280 |
| less than 70 | 320 |
| less than 80 | 360 |
| less than 90 | 390 |

5. Calculate standard deviation by all the methods for the following data :

| Sales (Rs. In Lacs) | No. of firms |
|---|---|
| 0-10 | 12 |
| 10-20 | 16 |
| 20-30 | 20 |
| 30-40 | 50 |
| 40-50 | 40 |
| 50-60 | 30 |
| 60-70 | 20 |
| 70-80 | 16 |

## 7.8 SUGGESTED READINGS

1. Yule and Kendall : Introduction to the Theory of Statistics.
2. S.P.Gupta : Statistical Methods, S.Chand and Sons, New Delhi

**C. No. : BCG-502** **UNIT : II**
**SEMESTER: V** **LESSON: 6**

# MEASURES OF SKEWNESS

**Structure**

6.1 Introduction

6.2 Objectives

6.3 Meaning of Skewness

6.4 Measures of Skewness

6.5 Difference Between Skewness and Dispersion

6.6 Summary

6.7 Self Assessment Exercises

6.8 Suggested Readings

## 6.1 INTRODUCTION

We have already talked about the central tendency and dispersion of frequency distribution and also seen their different computational procedures. As we have discussed that central tendency tells about a value around which many items of the data congregate. However, dispersion tells that how much the items deviate from central tendency. In this lesson, we shall study an additional feature of frequency distribution which tells us how far the frequency curve of the given frequency distribution deviates from a symmetric one. This feature is known as skewness.

## 6.2 OBJECTIVES

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

- understand the concept of skewness,
- differentiate between symmetrical, positively skewed and negatively skewed data.
- calculate skewness by different methods,
- distinguish between dispersion and skewness, and
- choose an appropriate measure of skewness which should be suitable in a given situation.

## 6.3 MEANING OF SKEWNESS

Skewness is the lack of symmetry. A frequency distribution is said to be symmetrical if the frequencies are symmetrically distributed about central value. When a frequency distribution is not symmetrical it is called a skewed frequency distribution. There are two types of skewness; positive skewness and negative skewness.

**I. Positive Skewness :** A distribution is said to be positively skewed or skewed to the right if the observations pile up at the lower values of the variable. That is, in the positively skewed distribution the frequencies are spread out over a greater range of values on the higher value end of the curve than they are on the low value end. Thus, the curve of such a distribution has longer tail to the right.

**II. Negative Skewness :** A distribution is said to be negatively skewed or skewed to the left if the observations pile up at the higher values of the variable. In the negatively skewed distributions, the frequencies are spread out over a greater range of values on the lower value end of the curve then they are on the highly value end. Thus, the curve of such a distribution has longer tail to the left.

Some important **definition**s of skewness are as follows:

1. "When a series is not symmetrical it is said to be asymmetrical or skewed." **-Croxton & Cowden.**

2. "Skewness refers to the asymmetry or lack of symmetry in the shape of a frequency distribution**."** **-Morris Hamburg.**

3. "Measures of skewness tell us the direction and the extent of skewness. In symmetrical distribution the mean, median and mode are identical. The more the mean moves away from the mode, the larger the asymmetry or skewness." **-Simpson & Kalka**

4. "A distribution is said to be 'skewed' when the mean and the median fall at different points in the distribution, and the balance (or centre of gravity) is shifted to one side or the other-to left or right." **–Garrett**

The above definitions show that the term 'skewness' refers to lack of symmetry" i.e., when a distribution is not symmetrical (or is asymmetrical) it is called a skewed distribution.

## TESTS OF SKEWNESS

In order to ascertain whether a distribution is skewed or not the following tests may be applied. Skewness is present if:

1. The values of mean, median and mode do not coincide.
2. When the data are plotted on a graph they do not give the normal bell-shaped form i.e. when cut along a vertical line through the centre the two halves are not equal.
3. The sum of the positive deviations from the median is not equal to the sum of the negative deviations.
4. Quartiles are not equidistant from the median.
5. Frequencies are not equally distributed at points of equal deviation from the mode.

On the contrary, when skewness is absent, i.e. in case of a symmetrical distribution, the following conditions are satisfied:

## 6.4 MEASURES OF SKEWNESS

There are various criteria to check whether a given frequency distribution is skewed or not. On the basis of these criteria, we find absolute measures of skewness; but for comparing two or more distributions we find relative measures. The relative measures of skewness are called Coefficient of Skewness. The following are the main criteria for measuring skewness:

1. The values of mean, median and mode coincide.
2. Data when plotted on a graph give the normal bell-shaped form.
3. Sum of the positive deviations from the median is equal to the sum of the negative deviations.
4. Quartiles are equidistant from the median.
5. Frequencies are equally distributed at points of equal deviations from the mode.

**(I) Karl Pearson's Coefficient of Skewness :** In a symmetrical distribution the mean, median and mode occur at the same points i.e. mean, median and mode coincides. In a positively skewed distribution, the value of mean is maximum, that of mode least and median lies in between the two i.e. mean> median> mode.

However, in a negatively skewed distribution, the value of mode is maximum, that of mean least and median lies in between mean and mode i.e. Mode> Median>

Mean. Consequently, the distance between mean and mode is used to measure skewness ; the greater the distance, the more skew the distribution. Symbolically,

If mean-mode = 0, no skewness

If mean-mode > 0, positive skewness

If mean-mode < 0, negative skewness

The relative measure of skewness corresponding to this absolute measure, defined by Karl Pearson, is given by

$$SK_p = \frac{Mean-Mode}{Standard\ deviation}$$

It is a pure number and is zero for symmetrical distributions. This Coefficient of Skewness lies between ± 1.

If mode is ill defined, then using the approximate relationship :

Mode = 3 Median–2 Mean

The above formula reduces to

$$SK_p = \frac{3(Mean-Median)}{Standard\ deviation}$$

**Example 1.** From the marks secured by students in Section I and II of a class of 100 students, the following information is obtained.

Section I : = 56.83, S.D. = 12.8, Mode = 61.47

Section II : = 57.83, S.D. = 12.8, Mode = 57.60

Determine which distribution of marks is more skewed.

**Solution:** We determine Karl Pearson's Coefficient of Skewness for sections :

**Section I**

$$SK_p = \frac{Mean-Mode}{Standard\ deviation}$$

$$= \frac{56.83-61.47}{12.8}$$

$$= - \ 0.3625$$

**Section II**

$$SK_p \text{ (II)} = \frac{57.83-57.60}{12.8}$$

$$= \quad 0.0179$$

Thus, the distribution of marks in Section I is more skewed. The skewness for Section I is negative while for Section II is positive.

**Example 2.** Find out the S.D. if coefficient of skewness is –0.475, mean is 60 and median is 65 of a particular distribution.

**Solution :** As we know that

$$SK_p = \frac{3(\mathit{Mean}-\mathit{Median})}{\mathit{Standard\ deviation}}$$

on putting the given values, we get

$$0.475 = \frac{3(\ 60-65)}{\mathit{Standard\ deviation}}$$

or

$$\text{Standard Deviation} = \frac{3(\ 60 - 65)}{0.475}$$

$$= \frac{-3\ X\ 5}{-0.475}$$

$$= 31.58$$

**(2) Bowley's Coefficient of Skewness :** This method is based on quartiles. As we know that in a symmetrical distribution the first quartile $Q_1$ and the third quartile $Q_3$ are at equal distances from the second quartic $Q_2$ i.e. median, and they are not at equal distances from median ($Q_2$) in a skewed distribution. Bowley considered this fact and suggested a measure of skewness thus, if

$Q_3–Q_2 = Q_2–Q_1,$
no skewness

$(Q_3–Q_2) > Q_2–Q_1,$ positive skewness

$(Q_3–Q_2) < Q_2–Q_1,$ negative skewness

Corresponding to this absolute measure of skewness the relative measure of skewness (Bowley's Coefficient of Skewness) is given by

$$SKB = \frac{(Q_3-Q_2)-(Q_2-Q_1)}{(Q_3-Q_2)-+(Q_2-Q_1)}$$

$$\frac{Q_3-2Q_2+Q_1}{(Q_3-Q_1)}$$

This is a pure number and is zero for symmetrical distributions and lies between ± 1.

This method is particularly useful in case of open-end distributions and where extreme values are present or when class intervals are unequal. But the main disadvantage of this method is that it is based on only central 50% of the data and it ignores the remaining 25% of data below $Q_1$ and 25% of data above $Q_3$.

**Example 3.** Calculate Bowley's Coefficient of Skewness for the following frequency distribution

| | | | | | |
|---|---|---|---|---|---|
| Marks Obtained: | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 |
| No. of Students: | 2 | 7 | 10 | 5 | 3 |

**Solution :** To determine Bowley's Coefficient of Skewness, we first calculate $Q_1$, $Q_2$, and $Q_3$

| Class | Frequency (f) | Cum frequency (c.f.) |
|---|---|---|
| 0-10 | 2 | 2 |
| 10-20 | 7 | 9 |
| 20-30 | 10 | 19 |
| 30-40 | 5 | 24 |
| 40-50 | 3 | 27 |
| Total | 27 | |

Here,

$$Q_1 = L_1 + \frac{\frac{N}{4} - c.f.}{f} X\ h$$

$$= 10 + 6.75\text{-}2/ 7 \ X \ 10$$

$$= 10 + 6.786$$

$$= 16.786$$

$$Q_2 = L_1 + \frac{\frac{N}{2} - c.f.}{f} X\ h$$

$$= 20 + 13.5\text{-}9/ 10 \ X \ 10$$

$$= 20 + 4.5$$

$$= 24.5$$

$$Q_3 = L_1 + \frac{\frac{3N}{4} - c.f.}{f} X\ h$$

$$= 30 + 20.25\text{-}19/ 5 \ X \ 10$$

$$= 30 + 2.5$$

$$= 32.5$$

$$SK_B = \frac{Q_3 - 2Q_2 + Q_1}{(Q_3 - Q_1)}$$

$$= \frac{32.5-2(24.5)+16.786}{32.5-16.786}$$

$$= \frac{0.286}{15.714}$$

$$= 0.0182$$

**Example 4.** The following table gives the distribution of monthly income of 500 workers in a factory :

| **Monthly Income(`)** | **No. of employees** |
|---|---|
| Below Rs. 1000 | 10 |
| 1000-1500 | 25 |
| 1500-2000 | 145 |
| 2000-2500 | 220 |
| 2500-3000 | 70 |
| 3000 and above | 30 |

(i) Obtain the limits of income of central 50% of the observed employees.

(ii) Calculate Bowley's Coefficient of Skewness.

**Solution :** We first find $Q_1$, $Q_2$, and $Q_3$

**Calculation of Quartiles**

| **Monthly Income** | ***f*** | ***c.f.*** |
|---|---|---|
| < 1000 | 10 | 10 |
| 1000-1500 | 25 | 35 |
| 1500-2000 | 145 | 180 |

| | | |
|---|---|---|
| 2000-2500 | 220 | 400 |
| 2500-3000 | 70 | 470 |
| 3000 > | 30 | 500 |

$Q_1 \quad = L_1 + \frac{\frac{N}{4} - c.f.}{f} X\ h$

= 1500 + 125-35/ 145 X 500

= 1500 + 310.35

= 1810.35

$Q_2 \quad = L_1 + \frac{\frac{N}{2} - c.f.}{f} X\ h$

= 2000 + 250-180/ 220 X 500

= 2000 + 159.09

= 2159.09

and

$Q_3 \quad = L_1 + \frac{\frac{3N}{4} - c.f.}{f} X\ h$

= 2000 + 375-180/ 220 X 500

= 2000 + 443.18

= 2443.18

(i) The income of central 50 percent of workers lies between Rs. 1810.35 and 2443.18.

(ii) Bowley's Coefficient of Skewness is given by

$SK_B = \underline{Q_3 - 2Q_2 + Q_1}$

$_{-}(Q_3\text{-}Q_1)$

$$= \frac{2443.18-2X\ 2159.09+1810.35}{2443.18-1810.35}$$

$$= \frac{-64.65}{632.83}$$

= -0.102

**(3) Coefficient of Skewness Based on Moments :** An absolute measure of

skewness based on moments is measured by third central moment i.e. $\mu_3$. If,

| | |
|---|---|
| $\mu_3 = 0$ | no skewness |
| $\mu_3 > 0$ | positive skewness |
| $\mu_3 < 0$ | negative skewness |

Relative measures (Coefficient of Skewness) based on central moments are

(i) $\beta_1 = \mu_3^2$; but it does give the sign of skewness.

(ii) $\gamma_1 = \beta_1 = \mu_{3''}\mu T! = \mu^3/ \sigma^3$

Obviously this coefficient is a pure number and zero for symmetrical distribution. Symbolically,

| | |
|---|---|
| $\gamma_1 = 0$ | no skewness |
| $\gamma_1 > 0$ | positive skewness |
| $\gamma_1 < 0$ | negative skewness |

## 6.5 DIFFERENCE BETWEEN SKEWNESS AND DISPERSION

Dispersion and skewness are two different characteristics of a frequency distribution. The main difference between them are :

(i) Dispersion indicates the amount of variation rather than its direction.

But skewness indicates the direction of variation.

(ii) The measure of skewness are dependent upon the amount of dispersion.

(iii) Dispersion is related with the composition of the distribution where as skewness is related with its shape.

(iv) For a symmetrical distribution, the skewness is always zero whereas dispersion may take any value.

## 6.6 SUMMARY

The various measures of central tendency and dispersion do not reveal all the characteristics of a data. Two distributions may have same mean and standard deviation but may differ in the shape of their distributions. If the distribution of data is not symmetrical, it is called skewed. Thus, skewness means lack of symmetry in distribution. Different methods of measuring skewness are as follows.

| Absolute Measure | Relative Measure |
|---|---|
| 1. Mean-Mode | $\frac{\text{Mean} - \text{Mode}}{\text{S.D.}}$ |
| 2. 3(Mean – Median) | $\frac{3(\text{Mean} - \text{Median})}{\text{S.D.}}$ |
| 3. $Q_3 - 2Q_2+Q_1$ | $\frac{Q_3 - 2Q_2+Q_1}{Q_3 - Q_1}$ |
| | $\beta_1 = \mu_3^2/\mu_2^3$ |
| 4. $\mu_3$ | $\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}$ |

## 6.7 SELF ASSESSMENT EXERCISES

1. What is skewness? Differentiate between skewness and dispersion.

______________________________________________
______________________________________________
______________________________________________
______________________________________________
______________________________________________
______________________________________________

2. In a positively skewed distribution mean is $\bar{X}$ and mode is M, 0< M< $\bar{X}$, what is its median?

______________________________________________
______________________________________________
______________________________________________
______________________________________________
______________________________________________
______________________________________________

3. State various methods of measuring skewness.

______________________________________________
______________________________________________
______________________________________________
______________________________________________
______________________________________________
______________________________________________

5. If the mean, mode and standard deviation are 41, 45 and 8 respectively, then find its Coefficient of Skewness.

______________________________________________
______________________________________________
______________________________________________
______________________________________________
______________________________________________

5. Calculate an appropriate Coefficient of Skewness from the following data.

| Income (` Per day) | No. of firms |
|---|---|
| Below Rs. 1000 | 20 |
| 1000-2000 | 60 |
| 2000-3000 | 30 |
| 3000-4000 | 80 |
| 4000-5000 | 40 |
| 5000-6000 | 30 |
| 6000 and above | 10 |

6. Calculate Karl Pearson's and Bowley's Coefficient of Skewness from the following data :

| Income (` Per day) | No. of firms |
|---|---|
| 1000-1500 | 15 |
| 1500-2000 | 30 |
| 2000-2500 | 50 |
| 2500-3000 | 60 |
| 3000-3500 | 55 |
| 3500-4000 | 40 |
| 4000-4500 | 30 |

6. Measures of dispersion and skewness are complimentary to one another in understanding a frequency distribution.” Elucidate the statement.

## 6.8 SUGGESTED READINGS

1. Elehance : Advanced Statistics
2. S.R Gupta : Statistics
3. Vishwanathan : Business Statistics : An Applied Orientation.

**C. No. : BCG-502** **UNIT : III**

**SEMESTER: V** **LESSON: 7**

# INDEX NUMBERS

## STRUCTURE

7.1 Introduction

7.2 Characteristics of Index numbers

7.3 Uses of Index Numbers

7.4 Classification of index numbers

7.5 Problems in constructing index numbers:

7.6 Summary

7.7 Self Assessment Exercises

7.8 Suggested Readings

## 7.1 INTRODUCTION

Historically, the first index was constructed in 1764 to compare the Italian price index in 1750 with the price level in 1500. Though originally developed for measuring the effect of change in prices, index numbers have today become one of the most widely used statistical devices and there is hardly any field where they are not used. Newspapers headline the fact that prices are going up or down, that industrial production is rising or falling, that imports are increasing or decreasing, that crimes are rising in a particular period compared to the previous period as disclosed by index numbers. They are used to feel the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies, In fact, they are described as 'barometers of economic activity', i.e., if one wants to get an idea as to what is happening to an economy, he should look to important indices like the index number of industrial production, agricultural production, business activity, etc.

Some prominent definitions of index numbers are given below:

1. 'Index numbers are devices for measuring differences in the magnitude of a group of related variables. **—Croxton & Cowdert**

2. "An index number is a statistical measure designed to show changes in a variable or a group of related variables with respect to time, geographic location or other characteristics such as income, profession, etc. **—Spiegel**

3. "In its simplest form an index number is the ratio of two index numbers expressed as a per cent. An index number is a statistical measure—a measure designed to show changes in one variable or in a group of related variables over time, or with respect to geographic location, or in terms of some other characteristics." **—Patternson**

### Definition:

***Index numbers*** are statistical devices designed to measure the relative change in the level of variable or group of variables with respect to time, geographical location etc.

In other words these are the numbers which express the value of a variable at any given period called "***current period*** "as a percentage of the value of that variable at some standard period called "***base period***".

Here the variables may be

1. The price of a particular commodity like silver, iron or group of commodities like consumer goods, food, stuffs etc.

2. The volume of trade, exports, imports, agricultural and industrial production, sales in departmental store.

3. Cost of living of persons belonging to particular income group or profession etc.

**Ex:** suppose rice sells at Rs.9/kg at BBSR in 1995 as compare to Rs. 4.50/Kg in 1985, the index number of price in 1995 compared to 1985.

Therefore the index number of price of rice in 1995 compared to 1985 is calculated as

$$\frac{Rs.9.00}{Rs.4.50} \times 100 = 200$$

***This means there is a net increase of 100% in the price of rice in 1995as compared to 1985*** **[the base year's index number is always treated as 100]**

Suppose, during the same period 1995 the rice sells at Rs. 12.00/kg in Delhi. There fore, the index number of price at Bhubaneswar compared to price at Delhi is $\frac{Rs.9.00}{Rs.12.00} \times 100 = 75$

***This means there is a net decrease of 25% in the price of rice in 1995as compared to 1985***

The above index numbers are called ***'price index numbers'***.

To take another example the production of rice in 1978 in Orissa was 44, 01,780 metric c tons compare to 36, 19,500 metric tons in 1971. So the index number of the quantity produced in 1978 compared to 1971 is

$$\frac{4401780}{3619500} \times 100 = 121.61$$

***That means there is a net increase of 21.61% in production of rice in 1978 as compared to 1971.***

The above index number is called ***'quantity index number'***

**Univariate index**: An index which is calculated from a single variable is called *univariate index.*

**Composite index**: An index which is calculated from group of variables is called *Composite index*

## 7.2 CHARACTERISTICS OF INDEX NUMBERS

1. **Index numbers are specialized averages:**

As we know an average is a single figure representing a group of figures. How ever to obtain an average the items must be comparable. For example the average weight of man, woman and children of a certain locality has no meaning at all. Further more the unit of measurement must be same for all the items. How ever this is not so with index numbers. Index numbers also one type of averages which shows in a single figure the change in two or more series of different items which can be expressed in different  units. For example while constructing a consumer price index number the various items which are use in construction are divided into broad heads namely food, clothing, fuel, lighting, house rent, and miscellaneous which are expressed in different units.

2. **Index numbers measures the net change in a group of related variables:**

Since index numbers are essentially averages, they describe in one single figure the increase or decrease in a group of related variables under study. The group of variables may be prices of set of commodities, the volume of production in different sectors etc.

3. **Index numbers measure the effect of changes over period of time:**

Index numbers are most widely used for measuring changes over a period of time. For example we can compare the agricultural production, industrial production, imports, exports, wages etc in two different periods.

## 7.3 USES OF INDEX NUMBERS

Index numbers are indispensable tools of economics and business analysis. Following are the main uses of index numbers.

1) **Index numbers  are used as economic barometers:**

Index number is a special type of averages which helps to measure the economic fluctuations on price level, money market, economic cycle like inflation, deflation etc. G.Simpson and F.Kafka say that index numbers are today one of the most widely used statistical devices. They are used to take the pulse of economy and they are used as indicators of inflation or deflation tendencies. So index numbers are called economic barometers.

**2) Index numbers helps in formulating suitable economic policies and planning etc.**

Many of the economic and business policies are guided by index numbers. For example while deciding the increase of DA of the employees; the employer's have to depend primarily on the cost of living index. If salaries or wages are not increased according to the cost of living it leads to strikes, lock outs etc. The index numbers provide some guide lines that one can use in making decisions.

**3) They are used in studying trends and tendencies.**

Since index numbers are most widely used for measuring changes over a period of time, the time series so formed enable us to study the general trend of the phenomenon under study. For example for last 8 to 10 years we can say that imports are showing upward tendency.

**4) They are useful in forecasting future economic activity.**

Index numbers are used not only in studying the past and present workings of our economy but also important in forecasting future economic activity.

**5) Index numbers measure the purchasing power of money.**

The cost of living index numbers determine whether the real wages are rising or falling or remain constant. The real wages can be obtained by dividing the money wages by the corresponding price index and multiplied by 100. Real wages helps us in determining the purchasing power of money.

**6) Index numbers are used in deflating.**

Index numbers are highly useful in deflating i.e. they are used to adjust the wages for cost of living changes and thus transform nominal wages into real wages, nominal income to real income, nominal sales to real sales etc. through appropriate index numbers.

## 7.4 CLASSIFICATION OF INDEX NUMBERS:

According to purpose for which index numbers are used are classified as below.

i) Price index number

ii) Quality index number

iii) Value index number

iv) Special purpose index number

Only price and quantity index numbers are discussed in detail. The others will be mentioned but without detail.

### Price index number:

*Price index number* measures the changes in the price level of one commodity or group of commodities between two points of time or two areas.

**Ex:** Wholesale price index numbers

Retail price index numbers

Consumer price index numbers.

### Quantity index number:

*Quantity index numbers* measures the changes in the volume of production, sales, etc in different sectors of economy with respect to time period or space.

**Note:** Price and Quantity index numbers are called *market index numbers.*

## 7.5 PROBLEMS IN CONSTRUCTING INDEX NUMBERS:

Before constructing index numbers the careful thought must be given into following problems

### i. Purpose of index numbers.

An index number which is properly designed for a purpose can be most useful and powerful tool. Thus the first and the foremost problem are to determine the purpose of index numbers. If we know the purpose of the index numbers we can settle some related problems. For example if the purpose of index number is to measure the changes in the production of steel, the problem of selection of items is automatically settled.

### ii. Selection of commodities

After defining the purpose of index numbers, select only those commodities which are related to that index. For example if the purpose of an index is to measure the cost of living of low income group we should select only those commodities or items which are consumed by persons belonging to this group and due care should be taken not to include the goods which are utilized by the middle income group or high income group i.e. the goods like cosmetics, other luxury goods like scooters, cars, refrigerators, television sets etc.

### iii. Selection of base period

The period with which the comparisons of relative changes in the level of phenomenon are made is termed as ***base period.*** The index for this period is always taken as 100. The following are the basic criteria for the choice of the base period.

i) The base period must be a normal period i.e. a period frees from all sorts of abnormalities or random fluctuations such as labor strikes, wars, floods, earthquakes etc.

ii) The base period should not be too distant from the given period. Since index numbers are essential tools in business planning and economic policies the base period should not be too far from the current period. For example for deciding increase in dearness allowance at present there is no advantage in taking 1950 or 1960 as the base, the comparison should be with the preceding year after which the DA has not been increased.

iii) Fixed base or chain base .While selecting the base a decision has to be made as to whether the base shall remain fixing or not i.e. whether we have fixed base or chain base. In the fixed base method the year to which the other years are compared is constant. On the other hand, in chain base method the prices of a year are linked with those of the preceding year. The chain base method gives a better picture than what is obtained by the fixed base method.

- **How a base is selected if a normal period is not available?**

**Ans:** Some times it is difficult to distinguish a year which can be taken as a normal year and hence the average of a few years may be regarded as the value corresponding to the base year.

**iv. Data for index numbers**

The data, usually the set of prices and of quantities consumed of the selected commodities for different periods, places etc. constitute the raw material for the construction of index numbers. The data should be collected from reliable sources such as standard trade journals, official publications etc. for example for the construction of retail price index numbers, the price quotations for the commodities should be obtained from super bazaars, departmental stores etc. and not from wholesale dealers.

**v. Selection of appropriate weights**

A decision as to the choice of weights is an important aspect of the construction of index numbers. The problem arises because all items included in the construction are not of equal importance. So proper weights should be attached to them to take into account their relative importance. Thus there are two type of indices.

i) Un weighted indices- in which no specific weights are attached

ii) Weighted indices- in which appropriate weights are assigned to various items.

**vi. Choice of average.**

Since index numbers are specialized averages, a choice of average to be used in their construction is of great importance. Usually the following averages are used.

i) A.M

ii) G.M

iii) Median

Among these averages ***G.M*** is the appropriate average to be used. But in practice G.M is not used as often as A.M because of its computational difficulties.

**vii. Choice of formula.**

A large variety of formulae are available to construct an index number. The problem very often is that of selecting the appropriate formula. The choice of the formula would depend not only on the purpose of the index but also on the data available.

## 7.6 SUMMARY

An index number is a statistical measure, designed to measure changes in a variable(s) with time/geographical location/other criteria

λ Index Numbers are of three types : (i) Price-Index Numbers (ii) Quantity Index Numbers (iii) Value-index Numbers

Method of construction of Index numbers

- Simple Aggregative method
- Simple Average of Price Relatives Method

## 7.7 SELF ASSESSMENT EXCERCISE

1. “Index Numbers are devices for measuring changes in the magnitude of a group of related variables”. Discuss this statement and point out the important uses of index numbers.

2. (a) Explain the uses of index numbers.

(b) What problems are involved in the construction of index numbers?

3. Define index number. Discuss various problems in the construction of index numbers.

4. What are the characteristics of an index number?

## 7.8 SUGGESTED READINGS

1. Elehance : Advanced Statistics
2. S.P. Gupta : Statistics
3. Vishwanathan : Business Statistics : An Applied Orientation.

**C. No. : BCG-502** **UNIT : III**

**SEMESTER: V** **LESSON: 8**

# METHODS OF CONSTRUCTING INDEX NUMBERS

## STRUCTURE

8.1 Methods of constructing index numbers:

8.2 Characteristics of Index numbers

8.3 Uses of Index Numbers

8.4 Classification of index numbers

8.5 Problems in constructing index numbers

8.6 Self-Assessment Exercises

8.7 Suggested Readings

## 8.1 METHODS OF CONSTRUCTING INDEX NUMBERS

A large number of formulae have been derived for constructing index numbers. They can be

1) Unweighted indices

   a) Simple aggregative method

   b) Simple average of relatives.

2) Weighted indices

   a) Weighted aggregative method

      **i)** Lasperey’s method

**ii)** Paasche's method

**iii)** Fisher's ideal method

**iv)** Dorbey's and Bowley's method

**v)** Marshal-Edgeworth method

**vi)** Kelly's method

b) Weighted average of relatives

## 8.2 UNWEIGHTED INDICES

### i) Simple aggregative method:

This is the simplest method of constructing index numbers. When this method is used to construct a price index number the total of current year prices for the various commodities in question is divided by the total of the base year prices and the quotient is multiplied by 100.

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

Where $P_0$ are the base year prices

$P_1$ are the current year prices

$P_{01}$ is the price index number for the current year with reference to the base year.

**Problem:**

Calculate the index number for 1995 taking 1991 as the base for the following data

| Commodity | Unit | Prices 1991 ($P_0$) | Prices 1995 ($P_1$) |
|---|---|---|---|
| A | Kilogram | 2.50 | 4.00 |
| B | Dozen | 5.40 | 7.20 |

| | | | |
|---|---|---|---|
| C | Meter | 6.00 | 7.00 |
| D | Quintal | 150.00 | 200.00 |
| E | Liter | 2.50 | 3.00 |
| Total | | 166.40 | 221.20 |

**Price index number =** $P_{01} = \frac{\sum P_1}{\sum P_0} \times 100 = \frac{221.20}{166.40} \times 100 = 132.93$

$\therefore$ There is a net increase of 32.93% in 1995 as compared to 1991.

**Limitations:**

There are two main limitations of this method

1. The units used in the prices or quantity quotations have a great influence on the value of index.
2. No considerations are given to the relative importance of the commodities.

**ii) Simple average of relatives**

When this method is used to construct a price index number, first of all price relatives are obtained for the various items included in the index and then the average of these relatives is obtained using any one of the averages i.e. mean or median etc.

When A.M is used for averaging the relatives the formula for computing the index is

$$P_{01} = \frac{1}{n} \sum \left( \frac{P_1}{P_0} \times 100 \right)$$

When G.M is used for averaging the relatives the formula for computing the index is

$$P_{01} = Anti\log\left[\frac{1}{n}\sum\log\left(\frac{P_1}{P_0}\times 100\right)\right]$$

Where n is the number of commodities

and price relative = $\frac{P_1}{P_0}\times 100$

**Problem:**

Calculate the index number for 1995 taking 1991 as the base for the following data

| **Commodity** | **Unit** | **Prices 1991 ($P_0$)** | **Prices 1995 ($P_1$)** | $\frac{P_1}{P_0}\times 100$ |
|---|---|---|---|---|
| A | Kilogram | 50 | 70 | $\frac{70}{50}\times 100 = 140$ |
| B | Dozen | 40 | 60 | 150 |
| C | Meter | 80 | 90 | 112.5 |
| D | Quintal | 110 | 120 | 109.5 |
| E | Liter | 20 | 20 | 100 |
| Total | | | | |

Price index number = $P_{01} = \frac{1}{n}\sum\left(\frac{P_1}{P_0}\times 100\right) = \frac{1}{5}\sum 612 = 122.4$

$\therefore$ There is a net increase of 22.4% in 1995 as compared to 1991.

**Merits:**

1. It is not affected by the units in which prices are quoted

2. It gives equal importance to all the items and extreme items don't affect the index number.

3. The index number calculated by this method satisfies the unit test.

**Demerits:**

1. Since it is an unweighted average the importance of all items are assumed to be the same.

2. The index constructed by this method doesn't satisfy all the criteria of an ideal index number.

3. In this method one can face difficulties to choose the average to be used.

## 8.3 WEIGHTED INDICES:

### i) Weighted aggregative method:

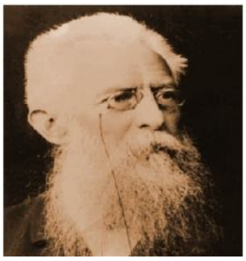
These indices are same as simple aggregative method. The only difference is in this method, weights are assigned to the various items included in the index.

There are various methods of assigning weights and consequently a large number of formulae for constructing weighted index number have been designed.

Some important methods are

**i. Lasperey's method:** This method is devised by Lasperey in year 1871.It is the most important of all the types of index numbers. In this method the base year quantities are taken weights. The formula for constructing Lasperey's price index number is

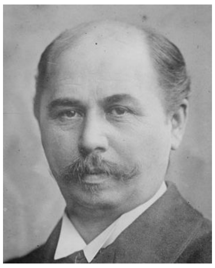
$$P_{01}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$



Ernst Louis Étienne Laspeyres
(1834-1913) , Germany

ii. **Paasche's method:** In this method the current year quantities are taken as weights and the formula is given by

$$P_{01}^{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$



Hermann Paasche
(1851-1925) Germany

iii. **Fisher's ideal method:** Fishers price index number is given by the G.M of the Lasperey's and Paasche's index numbers.
Symbolically

$$P_{01}^{F} = \sqrt{P_{01}^{La} P_{01}^{Pa}}$$

$$= \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100}$$

Sir Ronald Aylmer Fisher
(1890-1962) ,England
(England)

$$=\sqrt{\frac{\sum p_1q_0}{\sum p_0q_0}\frac{\sum p_1q_1}{\sum p_0q_1}}\times 100$$

**iv. Dorbey's and Bowley's method**

Dorbey's and Bowley's price index number is given by the A.M of the Lasperey's and Paasche's index numbers.

Symbolically

$$P_{01}^{DB}=\frac{P_{01}^{La}+P_{01}^{Pa}}{2}$$

## 8.4 QUANTITY INDEX NUMBERS:

**i. Lasperey's quantity index number:** Base year prices are taken as weights

$$Q_{01}^{La}=\frac{\sum q_1p_0}{\sum q_0p_0}\times 100$$

**ii. Paasche's quantity index number :**Current year prices are taken as weights

$$Q_{01}^{Pa}=\frac{\sum q_1p_1}{\sum q_0p_1}\times 100$$

**iii. Fisher's ideal method:** $Q_{01}^{F}=\sqrt{Q_{01}^{La}Q_{01}^{Pa}}=\sqrt{\frac{\sum q_1p_0}{\sum q_0p_0}\times\frac{\sum q_1p_1}{\sum q_0p_1}}\times 100$

**Fisher's index number is called ideal index number. Why?**

The Fisher's index number is called ideal index number due to the following characteristics.

1) It is based on the G.M which is theoretically considered as the best average of constructing index numbers.

2) It takes into account both current and base year prices as quantities.

3) It satisfies both time reversal and factor reversal test which are suggested by Fisher.

4) The upward bias of Lasperey’s index number and downward bias of Paasche’s index number are balanced to a great extent.

**Example: Compute price index numbers for the following data by**

*i) Laspeyre's method*

*ii) Paasche's method*

*iii) Fisher's ideal method.*

*iv) Dorbish-Bowley's method*

*v) Marshall -Edgeworth's method*

| Year | Commodity A | | Commodity B | | Commodity | |
|---|---|---|---|---|---|---|
| | Price | Quantity | Price | Quantity | Price | Quantity |
| *1980* | *4* | *50* | *3* | *10* | *2* | *5* |
| *1985* | *10* | *45* | *6* | *8* | *3* | *4* |

Base year : 1980

Price and quantity given in arbitrary units.

**CALCULATION OF INDICES**

| Year | 1980 | | 1885 | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Price | Quantity | Price | Quantity | $P_1q_0$ | $P_0q_0$ | $P_1q_1$ | $P_0q_1$ |
| *A* | *4* | *50* | *10* | *45* | *500* | *200* | *450* | *182* |
| *B* | *3* | *10* | *6* | *8* | *60* | *30* | *48* | *24* |
| *C* | *2* | *5* | *3* | *4* | *15* | *10* | *12* | *8* |
| *Total* | - | - | - | - | *575* | *240* | *510* | *212* |

(*i*) Laspeyre's method :

$$L_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \frac{575}{240} \times 100 = 239.58$$

(*ii*) Paasche's method :

$$P_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100 = \frac{510}{212} \times 100 = 240.57$$

(*iii*) Fisher's ideal method :

$$F_{01} = \sqrt{L_{01} \times P_{01}} = \sqrt{239.58 \times 240.57} = 240.07.$$

(*iii*) Fisher's ideal method :

$$F_{01} = \sqrt{L_{01} \times P_{01}} = \sqrt{239.58 \times 240.57} = 240.07.$$

(*iv*) Dorbish-Bowley's method :

$$DB_{01} = \frac{L_{01} + P_{01}}{2}$$

$$= \frac{239.58 + 240.57}{2} \doteq 239.82.$$

## 8.5 COMPARISON OF LASPEREY'S AND PAASCHE'S INDEX NUMBERS:-

In Lasperey's index number base year quantities are taken as the weights and in Paasche's index the current year quantities are taken as weights.

From the practical point of view Lasperey's index is often proffered to Paasche's for the simple reason that Lasperey's index weights are the base year quantities and do not change from the year to the next. On the other hand Paasche's index weights are the current year quantities, and in most cases these weights are difficult to obtain and expensive.

Lasperey's index number is said to be have upward bias because it tends to over estimate the price rise, where as the Paasche's index number is said to have downward bias, because it tends to under estimate the price rise.

When the prices increase, there is usually a reduction in the consumption of those items whose prices have increased. Hence using base year weights in the Lasperey's index, we will be giving too much weight to the prices that have increased the most and the numerator will be too large. Due to similar considerations, Paasche's index number using given year weights under estimates the rise in price and hence has down ward bias.

If changes in prices and quantities between the reference period and the base period are moderate, both Lasperey's and Paasche's indices give nearly the same values.

**Demerit of Paasche's index number:**

Paasche's index number, because of its dependence on given year's weight, has distinct disadvantage that the weights are required to be revised and computed for each period, adding extra cost towards the collection of data.

**What are the desiderata of good index numbers?**

Irving Fisher has considered two important properties which an index number should satisfy. These are tests of reversibility.

1. Time reversal test

2. Factor reversal test

If an index number satisfies these two tests it is said to be an ideal index number.

**Weighted average of relatives:**

Weighted average of relatives can be calculated by taking values of the base year ($p_0q_0$) as the weights. The formula is given by

When A.M is used $P_{01} = \frac{\sum PV}{\sum V}$

When G.M is used $P_{01} = Anti\log\frac{\sum V\log P}{\sum V}$

Where $P = \frac{p_1}{p_0}\times 100$ and $V = p_0q_0$ i.e. base year value

**Illustration 8.** From the following data compute price index by supplying weighted average of price method using :

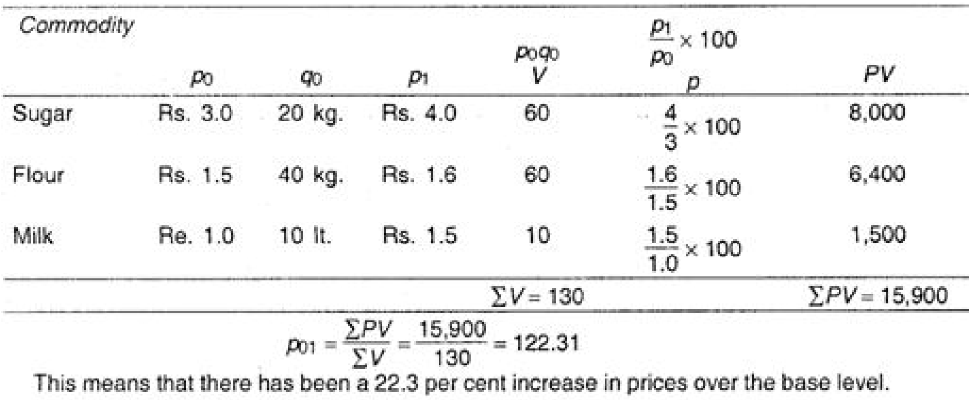
a) arithmetic means, and

b) geometric mean.

| **Commodity** | **$p_0$(`)** | **$q_0$** | **$p_1$(`)** |
|---|---|---|---|
| Sugar | 3.0 | 20 kg. | 4.0 |
| Flour | 1.5 | 40 kg. | 1.6 |
| Milk | 1.0 | 10 lt. | 1.5 |

**Solution.**

### (A) INDEX NUMBER USING WEIGHTED ARITHMETIC MEAN OF PRICE RELATIVES



| *Commodity* | $p_0$ | $q_0$ | $p_1$ | $p_0q_0$<br>$V$ | $\frac{p_1}{p_0}\times 100$<br>$p$ | $PV$ |
|---|---|---|---|---|---|---|
| Sugar | Rs. 3.0 | 20 kg. | Rs. 4.0 | 60 | $\frac{4}{3}\times 100$ | 8,000 |
| Flour | Rs. 1.5 | 40 kg. | Rs. 1.6 | 60 | $\frac{1.6}{1.5}\times 100$ | 6,400 |
| Milk | Re. 1.0 | 10 lt. | Rs. 1.5 | 10 | $\frac{1.5}{1.0}\times 100$ | 1,500 |
| | | | | $\Sigma V = 130$ | | $\Sigma PV = 15,900$ |

$$p_{01} = \frac{\Sigma PV}{\Sigma V} = \frac{15,900}{130} = 122.31$$

This means that there has been a 22.3 per cent increase in prices over the base level.

### (B) INDEX NUMBER USING GEOMETRIC MEAN OF PRICE RELATIVES

| *Commodity* | $p_0$ | $q_0$ | $p_1$ | *V* | *P* | *Log p* | *V Log p* |
|---|---|---|---|---|---|---|---|
| Sugar | ` 3.0 | 20 kg. | ` 4.0 | 60 | 133.3 | 2.1249 | 127.494 |
| Flour | ` 1.5 | 40 kg. | ` 1.6 | 60 | 106.7 | 2.0282 | 121.692 |
| Milk | ` 1.0 | 10 lt. | ` 1.5 | 10 | 150.0 | 2.1761 | 21.761 |
| | | | | *ΣV=130* | | | *ΣV log p* |
| | | | | | | | ***= 270.947*** |

$$p_{01} = \text{Antilog}\left[\frac{\Sigma V \log p}{\Sigma V}\right] = \text{Antilog}\left[\frac{270.947}{130}\right] = \text{Antilog } 2.084 = 120.9$$

## 8.6 TEST OF CONSISTENCY OR ADEQUACY

Several formulae have been suggested for constructing index numbers and the problem is that of selecting most appropriate one in a given situation. The following teats are suggested for choosing an appropriate index.

The following tests are suggested for choosing an appropriate index.

1) Unit test
2) Time reversal test
3) Factor reversal test
4) Circular test

**1) Unit test:**

This test requires that the formula for construction of index numbers should be such, which is not affected by the unit in which the prices or quantities have been quoted.

**Note:** This test is satisfied by all the index numbers except simple aggregative method.

## 2) Time reversal test

This is suggested by R.A.Fisher. Time reversal test is a test to determine whether a given method will work both ways in time i.e. forward and backward. In other words, when the data for any two years are treated by the same method, but with the bases reversed, the two index numbers secured should be reciprocals to each other, so that their product is unity. Symbolically the following relation should be satisfied.

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

Where $P_{01}$ is the index for time period 1 with reference period 0.

$P_{10}$ is the index for time period 0 with reference period 1.

**Note:** This test is not satisfied by Lasperey's method and Paasche's method. It is satisfied by Fisher's method.

**When Lasperey's method is used**

$$P_{01}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

$$P_{10}^{La} = \frac{\sum p_0 q_1}{\sum p_1 q_1} \times 100$$

Now,

$$P_{01}^{La} \times P_{10}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \text{ ‘“}1$$

Therefore this test is not satisfied by Lasperey's method

**When Paasche's method is used**

$$P_{01}^{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

$$P_{10}^{Pa} = \frac{\sum p_0 q_0}{\sum p_1 q_0} \times 100$$

Now,

$$P_{01}^{Pa} \times P_{10}^{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0} \neq 1$$

Therefore this test is not satisfied by Paasche's method

**When Fisher's method is used**

$$P_{01}^{F} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

$$P_{10}^{F} = \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \frac{\sum p_0 q_0}{\sum p_1 q_0}} \times 100$$

Now,

$$P_{01}^{F} \times P_{10}^{F} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \frac{\sum p_1 q_1}{\sum p_0 q_1}} \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \frac{\sum p_0 q_0}{\sum p_1 q_0}} = 1$$

**Value index:**

The *value* of a single commodity is the product of its price and quantity. Thus a value index 'V' is the sum of the values of the commodities of given year divided by the sum of the value of the base year multiplied by 100.

i.e. $V = \frac{\sum p_1 q_1}{\sum p_0 q_0} \times 100$

### 3) Factor reversal test:

This is also suggested by R.A.Fisher. It holds that the product of a price index number and the quantity index number should be equal to the corresponding value index. In other words the test is that the change in price multiplied by the change in quantity should be equal to change in value.

If $p_1$ & $p_0$ represents prices and $q_1$ & $q_0$ the quantities in the current year and base year respectively and if $P_{01}$ represents the change in price in the current year 1 with reference to the year 0 and $Q_{01}$ represents the change in quantity in the current year 1 with reference to the year 0.

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

**Note:** This test is not satisfied by Lasperey's method and Paasche's method. It is satisfied by Fisher's method.

**When Lasperey's method is used**

$$P_{01}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

$$Q_{01}^{La} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$$

Now,

$$P_{01}^{La} \times Q_{01}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum q_1 p_0}{\sum q_0 p_0} \neq \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

Therefore this test is not satisfied by Lasperey's method

**When Paasche’s method is used**

$$P_{01}^{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

$$Q_{01}^{Pa} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$$

Now,

$$P_{01}^{Pa} \times Q_{10}^{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum q_1 p_1}{\sum q_0 p_1} \neq \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

Therefore this test is not satisfied by Paasche’s method

**When Fisher’s method is used**

$$P_{01}^{F} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

$$Q_{01}^{F} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100$$

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

$$= \sqrt{\frac{\left(\sum p_1 q_1\right)^2}{\left(\sum p_0 q_0\right)^2}} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

Therefore this test is satisfied by Fisher’s method

### 4) Circular test:

This is another test of consistency of an index number. It is an extension of time reversal test. According to this test, the index should work in a circular fashion.

Symbolically

$$P_{01} \; x \; P_{12} \; x \; P_{20} = 1$$

**Note:**

This test is not satisfied by Lasperey's method, Paasche's method and Fisher's method.

This test is satisfied by simple average of relatives based on G.M and Kelly's fixed base method.

**Illustration 13.** Construct a Fisher's Ideal Index from the following data and show that It satisfies time reversal and factor reversal tests:

| | **2006** | | **2007** | |
|---|---|---|---|---|
| *Items* | $p_0$ | $q_0$ | $p_1$ | $q_1$ |
| A | 10 | 40 | 12 | 45 |
| B | 11 | 50 | 11 | 52 |
| C | 14 | 30 | 17 | 30 |
| D | 8 | 28 | 10 | 29 |
| E | 12 | 15 | 13 | 20 |

**Solution .**

**CONSTRUCTION OF FISHER'S IDEAL INDEX**

| *Items* | $p_0$ | $q_0$ | $p_1$ | $q_1$ | $p_1q_0$ | $p_0q_0$ | $p_1q_1$ | $p_0q_1$ |
|---|---|---|---|---|---|---|---|---|
| A | 10 | 40 | 12 | 45 | 480 | 400 | 540 | 450 |
| B | 11 | 50 | 11 | 52 | 550 | 550 | 572 | 572 |
| C | 14 | 30 | 17 | 30 | 510 | 420 | 510 | 420 |
| D | 8 | 28 | 10 | 29 | 280 | 224 | 290 | 232 |
| E | 12 | 15 | 13 | 20 | 195 | 180 | 260 | 240 |
| | | | | | $\Sigma p_1q_0=2015$ | $\Sigma p_0q_0=1774$ | $\Sigma p_1q_1=2172$ | $\Sigma p_0q_1=1914$ |

*Fisher's Ideal Index :* $P_{01} =$

$$= \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}}$$

$$= \sqrt{\frac{2015}{1774} \times \frac{2172}{1914}} \times 100 = 1{\cdot}135 \times 100 = 113{\cdot}5$$

Time Reversal Test : Time reversal test is satisfied when :

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

$$P_{10} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}} = \sqrt{\frac{1914}{2172} \times \frac{1774}{2015}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{2015}{1774} \times \frac{2172}{1914} \times \frac{1914}{2172} \times \frac{1774}{2015}} = 1$$

Hence time reversal test is satisfied by the given data.

*Factor Reversal Test* : Factor reversal test is satisfied when:

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

$$Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} = \sqrt{\frac{1914}{1774} \times \frac{2172}{2015}}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{2015}{1774} \times \frac{2172}{1914} \times \frac{1914}{1774} \times \frac{2172}{2015}} = \frac{2172}{1774}$$

$\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$ is also equai to $\frac{2172}{1774}$. Hence factor reversal test is satisfied by the given data.

- **Prove that AM of Lasperey’s index numbers and Paasche’s index number is greater than or equal to Fisher’s index number.**

Let

Lasperey’s index number = $P_{01}^{La}$

Paasche’s index number= $P_{01}^{Pa}$

Fisher’s index number= $P_{01}^{F}$

And we have $P_{01}^{F}= \sqrt{P_{01}^{La} P_{01}^{Pa}}$

Now we have to show that

$$\frac{P_{01}{}^{La}+P_{01}{}^{Pa}}{2} \geq P_{01}{}^{F}$$

$$\Rightarrow \frac{P_{01}{}^{La}+P_{01}{}^{Pa}}{2} \geq \sqrt{P_{01}{}^{La}P_{01}{}^{Pa}}$$

$$\Rightarrow P_{01}{}^{La}+P_{01}{}^{Pa} \geq 2\sqrt{P_{01}{}^{La}P_{01}{}^{Pa}}$$

$$\Rightarrow \left(P_{01}{}^{La}+P_{01}{}^{Pa}\right)^2 \geq 4P_{01}{}^{La}P_{01}{}^{Pa}$$

$$\Rightarrow \left(P_{01}{}^{La}-P_{01}{}^{Pa}\right)^2 \geq 0$$

## 8.7 THE CHAIN INDEX NUMBERS

In fixed base method the base remain constant through out i.e. the relatives for all the years are based on the price of that single year. On the other hand in chain base method, the relatives for each year is found from the prices of the immediately preceding year. Thus the base changes from year to year. Such index numbers are useful in comparing current year figures with the preceding year figures. The relatives which we found by this method are called link relatives.

Thus $\text{link relative for current year} = \frac{Current\ \ years\ \ figure}{\text{Pr}\,evious\ \ years\ \ figure} \times 100$

And by using these link relatives we can find the chain indices for each year by using the below formula

$$\text{Chain index for current year} = \frac{Linkrelative\ of\ \ current\ \ year \times Chain\ \ index\ of\ \ previous\ \ year}{100}$$

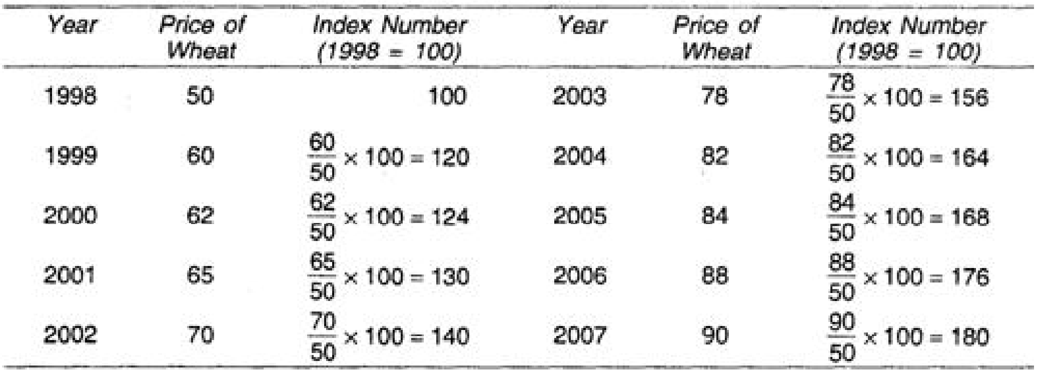
**Note:** The fixed base index number computed from the original data and chain index number computed from link relatives give the same value of the index provided that there is only one commodity, whose indices are being constructed.

**Example:** from the following data of wholesale prices of wheat for ten years construct index number taking a) 1998 as base and b) by chain base method.

| *Year* | ***Price of Wheat (Rs. per 40 kg.)*** | ***Year*** | ***Price of Wheat (Rs. per 40 kg.)*** |
|---|---|---|---|
| *1998* | *50* | *2003* | *78* |
| *1999* | *60* | *2004* | *82* |
| *2000* | *62* | *2005* | *84* |
| *2001* | 65 | 2006 | 88 |
| 2002 | 70 | 2007 | 90 |

**Solution .**

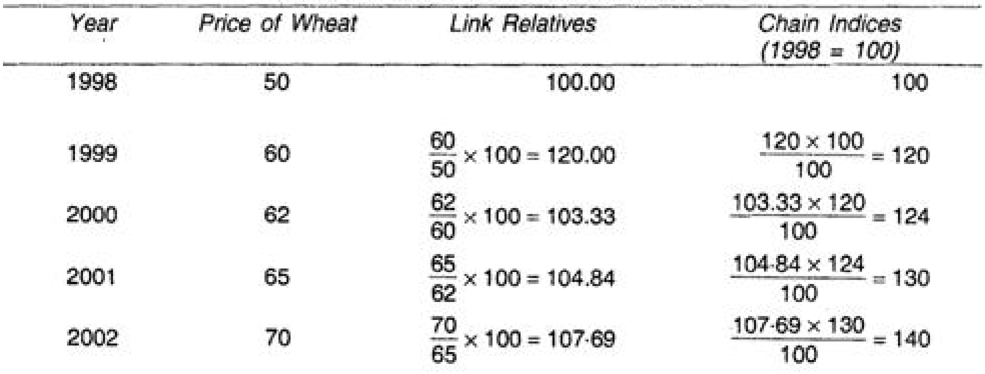
### (A) CONSTRUCTION OF INDEX NUMBERS TAKING 1998 AS BASE



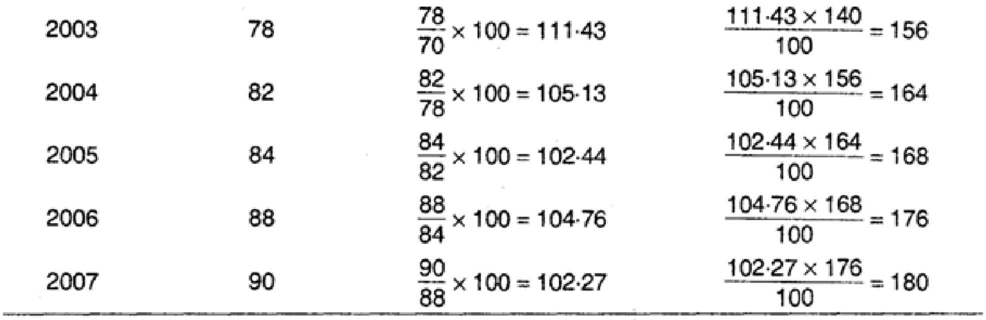
| *Year* | *Price of Wheat* | *Index Number (1998 = 100)* | *Year* | *Price of Wheat* | *Index Number (1998 = 100)* |
|---|---|---|---|---|---|
| 1998 | 50 | 100 | 2003 | 78 | $\frac{78}{50} \times 100 = 156$ |
| 1999 | 60 | $\frac{60}{50} \times 100 = 120$ | 2004 | 82 | $\frac{82}{50} \times 100 = 164$ |
| 2000 | 62 | $\frac{62}{50} \times 100 = 124$ | 2005 | 84 | $\frac{84}{50} \times 100 = 168$ |
| 2001 | 65 | $\frac{65}{50} \times 100 = 130$ | 2006 | 88 | $\frac{88}{50} \times 100 = 176$ |
| 2002 | 70 | $\frac{70}{50} \times 100 = 140$ | 2007 | 90 | $\frac{90}{50} \times 100 = 180$ |

*This means* that from 1998 to 1998 there is a 20 per cent increase; from 1999 to 2000 there is a 24 per cent increase; from 2000 to 2001 there is a 30 per cent increase. If we are interested in finding out increase from 1998 to 1999, from 1999 to 2000 from 2000 to 2001, we shall have to compute the chain indices.

(*b*) CONSTRUCTION OF CHAIN INDICES



| *Year* | *Price of Wheat* | *Link Relatives* | *Chain Indices (1998 = 100)* |
|---|---|---|---|
| 1998 | 50 | 100.00 | 100 |
| 1999 | 60 | $\frac{60}{50} \times 100 = 120.00$ | $\frac{120 \times 100}{100} = 120$ |
| 2000 | 62 | $\frac{62}{60} \times 100 = 103.33$ | $\frac{103.33 \times 120}{100} = 124$ |
| 2001 | 65 | $\frac{65}{62} \times 100 = 104.84$ | $\frac{104{\cdot}84 \times 124}{100} = 130$ |
| 2002 | 70 | $\frac{70}{65} \times 100 = 107{\cdot}69$ | $\frac{107{\cdot}69 \times 130}{100} = 140$ |



| | | | |
|---|---|---|---|
| 2003 | 78 | $\frac{78}{70} \times 100 = 111{\cdot}43$ | $\frac{111{\cdot}43 \times 140}{100} = 156$ |
| 2004 | 82 | $\frac{82}{78} \times 100 = 105{\cdot}13$ | $\frac{105{\cdot}13 \times 156}{100} = 164$ |
| 2005 | 84 | $\frac{84}{82} \times 100 = 102{\cdot}44$ | $\frac{102{\cdot}44 \times 164}{100} = 168$ |
| 2006 | 88 | $\frac{88}{84} \times 100 = 104{\cdot}76$ | $\frac{104{\cdot}76 \times 168}{100} = 176$ |
| 2007 | 90 | $\frac{90}{88} \times 100 = 102{\cdot}27$ | $\frac{102{\cdot}27 \times 176}{100} = 180$ |

**Note:** the chain indices obtained in (b) are the same as the fixed base indices obtained in (a). in fact chain index figures will always be equal to fixed index figure if there is only one series.

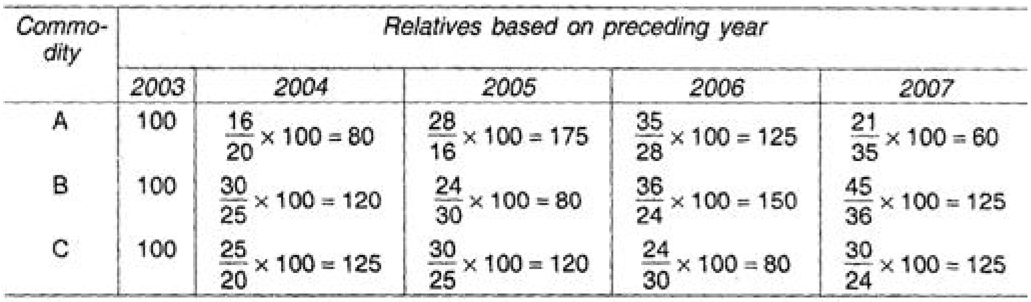
**Example-2: Compute the chain index number with 2003 prices as base from the following table giving the average wholesale prices of the commodities A, B and C for the year 2003 to 2007**

***Average wholesale prices (in Rs.)***

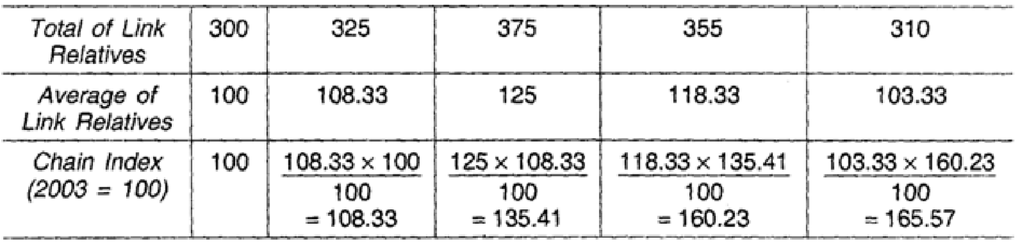
| ***Commodity*** | ***2003*** | ***2004*** | ***2005*** | ***2006*** | ***20007*** |
|---|---|---|---|---|---|
| A | 20 | 16 | 28 | 35 | 21 |
| B | 25 | 30 | 24 | 36 | 45 |
| C | 20 | 25 | 30 | 24 | 30 |

**Solution.**

**COMPUTATION OF CHAIN INDICES**



| *Commodity* | *Relatives based on preceding year* | | | | |
|---|---|---|---|---|---|
| | *2003* | *2004* | *2005* | *2006* | *2007* |
| A | 100 | $\frac{16}{20} \times 100 = 80$ | $\frac{28}{16} \times 100 = 175$ | $\frac{35}{28} \times 100 = 125$ | $\frac{21}{35} \times 100 = 60$ |
| B | 100 | $\frac{30}{25} \times 100 = 120$ | $\frac{24}{30} \times 100 = 80$ | $\frac{36}{24} \times 100 = 150$ | $\frac{45}{36} \times 100 = 125$ |
| C | 100 | $\frac{25}{20} \times 100 = 125$ | $\frac{30}{25} \times 100 = 120$ | $\frac{24}{30} \times 100 = 80$ | $\frac{30}{24} \times 100 = 125$ |



| | | | | | |
|---|---|---|---|---|---|
| *Total of Link Relatives* | 300 | 325 | 375 | 355 | 310 |
| *Average of Link Relatives* | 100 | 108.33 | 125 | 118.33 | 103.33 |
| *Chain Index (2003 = 100)* | 100 | $\frac{108.33 \times 100}{100}$<br>= 108.33 | $\frac{125 \times 108.33}{100}$<br>= 135.41 | $\frac{118.33 \times 135.41}{100}$<br>= 160.23 | $\frac{103.33 \times 160.23}{100}$<br>= 165.57 |

**Conversion of fixed based index to chain based index**

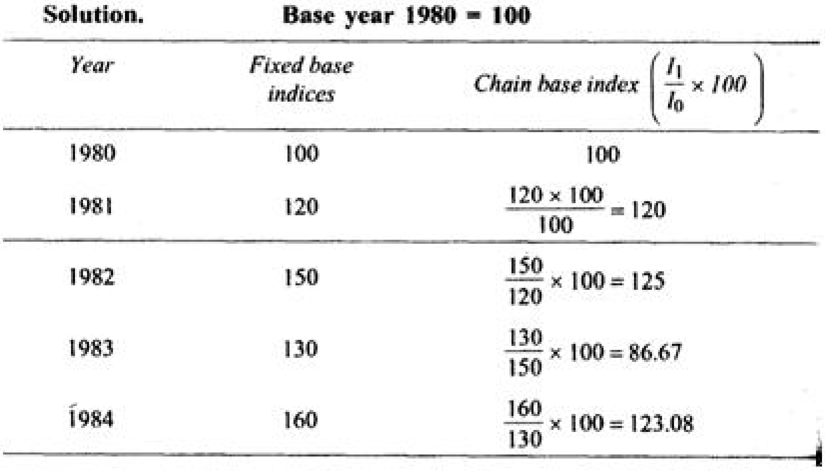
$$\text{Current year C.B.I} = \frac{\textit{Current years F.B.I}}{\text{Pr}\,\textit{evious years C.B.I}} \times 100$$

**Conversion of chain based index to fixed base index.**

$$\text{Current year F.B.I} = \frac{\textit{Current years C.B.I} \times \text{Pr}\,\textit{evious years F.B.I}}{100}$$

**Example: Compute the chain base index numbers**

| ***Year*** | ***1980*** | ***1981*** | ***1982*** | ***1983*** | ***1984*** |
|---|---|---|---|---|---|
| *Fixed base Index* | *100* | *120* | *150* | *130* | *160* |



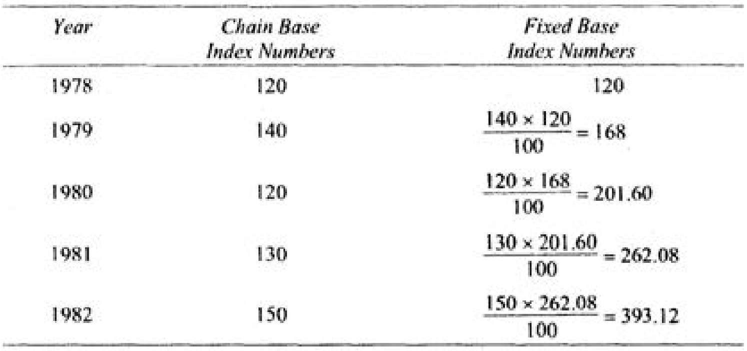
**Solution.** **Base year 1980 = 100**

| *Year* | *Fixed base indices* | *Chain base index* $\left(\frac{I_1}{I_0} \times 100\right)$ |
|---|---|---|
| 1980 | 100 | 100 |
| 1981 | 120 | $\frac{120 \times 100}{100} = 120$ |
| 1982 | 150 | $\frac{150}{120} \times 100 = 125$ |
| 1983 | 130 | $\frac{130}{150} \times 100 = 86.67$ |
| 1984 | 160 | $\frac{160}{130} \times 100 = 123.08$ |

**Example : Calculate fixed base index numbers from the following chain base index numbers**

| *Year* | *1978* | *1979* | *1980* | *1981* | *1982* |
|---|---|---|---|---|---|
| *Chain base* | | | | | |
| *Index numbers* | *120* | *140* | *120* | *130* | *150* |

**Solution. Computation of fixed base index numbers**



| *Year* | *Chain Base Index Numbers* | *Fixed Base Index Numbers* |
|---|---|---|
| 1978 | 120 | 120 |
| 1979 | 140 | $\frac{140 \times 120}{100} = 168$ |
| 1980 | 120 | $\frac{120 \times 168}{100} = 201.60$ |
| 1981 | 130 | $\frac{130 \times 201.60}{100} = 262.08$ |
| 1982 | 150 | $\frac{150 \times 262.08}{100} = 393.12$ |

**Note:** It may be remembered that the fixed base index for the first year is same as the chain base index for that year.

**Merits of chain index numbers:**

1. The chain base method has a great significance in practice, because in economic and business data we are often concerned with making comparison with the previous period.
2. Chain base method doesn’t require the recalculation if some more items are introduced or deleted from the old data.
3. Index numbers calculated from the chain base method are free from seasonal and cyclical variations.

**Demerits of chain index numbers:**

1. This method is not useful for long term comparison.
2. If there is any abnormal year in the series it will effect the subsequent years also.

**Differences between fixed base and chain base methods:**

| Chain base | Fixed base |
|---|---|
| 1. Here the base year changes | 1. Base year does not changes |
| 2. Here link relative method is used | 2. No such link relative method is used |
| 3. Calculations are tedious | 3. Calculations are simple |
| 4. It can not be computed if any one year is missing | 4. It can be computed if any year is missing |
| 5. It is suitable for short period | 5. It is suitable for long period |
| 6. Index numbers will be wrong if an error is committed in the calculation of link relatives | 6. The error is confined to the index of that year only. |

## 8.8 BASE SHIFTING

One of the most frequent operations necessary in the use of index numbers is changing the base of an index from one period to another with out recompiling the entire series. Such a change is referred to as *'base shifting'*. The reasons for shifting the base are

1. If the previous base has become too old and is almost useless for purposes of comparison.

2. If the comparison is to be made with another series of index numbers having different base.

The following formula must be used in this method of base shifting is

$$\text{Index number based on new base year} = \frac{\text{current years old index number}}{\text{new base years old index number}} \times 100$$

**Example:**

The following are the index numbers of prices with 1998 as base year

| | |
|---|---|
| 2003 | 410 |
| 2004 | 400 |
| 2005 | 380 |
| 2006 | 370 |
| 2007 | 340 |

Shift the base from 1998 to 2004 and recast the index numbers.

**Solution :**

Index number based on new base year = $\frac{\text{current years old index number}}{\text{new base years old index number}} \times 100$

Index number for 1998 = $\frac{100}{400} \times 100$ =25

…………………………………………..

Index number for 2007= $\frac{340}{400} \times 100$=85

| **Year** | **Index** |
|---|---|
| 1998 | 100 |
| 1999 | 110 |
| 2000 | 120 |
| 2001 | 200 |
| 2002 | 400 |

| **Year** | **Index number (1998as base)** | **Index number (2004 as base)** | **Year** | **Index number (1998as base)** | **Index number (2004 as base)** |
|---|---|---|---|---|---|
| 1998 | 100 | $\frac{100}{400}\times100$ =25 | 2003 | 410 | $\frac{410}{400}\times100$<br>=102.5 |
| 1999 | 110 | $\frac{110}{400}\times100$<br>=27.5 | 2004 | 400 | $\frac{400}{400}\times100$=100 |
| 2000 | 120 | $\frac{120}{400}\times100$=30 | 2005 | 380 | $\frac{380}{400}\times100$=95 |
| 2001 | 200 | $\frac{200}{400}\times100$=50 | 2006 | 370 | $\frac{370}{400}\times100$=92.5 |
| 2002 | 400 | $\frac{400}{400}\times100$<br>=100 | 2007 | 340 | $\frac{340}{400}\times100$=85 |

## 8.9 SPLICING OF TWO SERIES OF INDEX NUMBERS

The problem of combining two or more overlapping series of index numbers into one continuous series is called *splicing*. In other words, if we have a series of index numbers with some base year which is discontinued at some year and we have another series of index numbers with the year of discontinuation as the base, and connecting these two series to make a continuous series is called splicing.

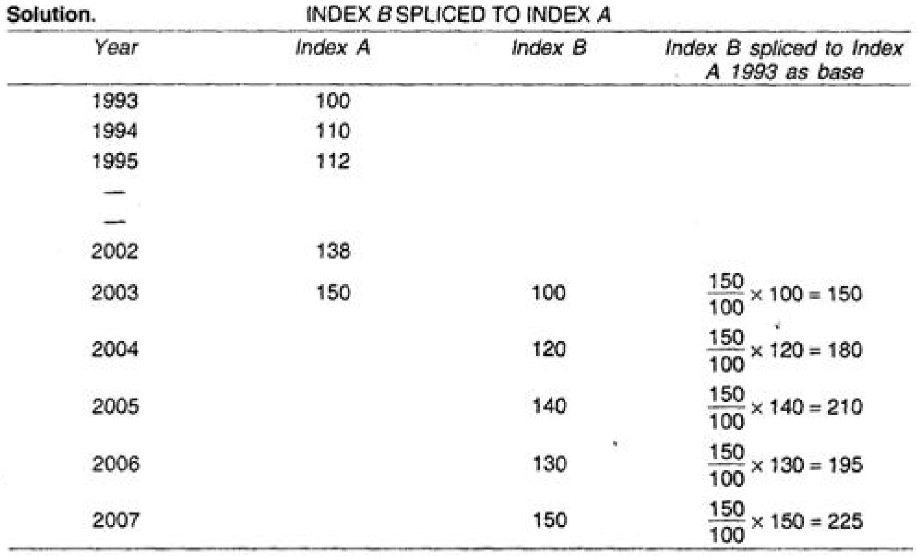
The following formula must be used in this method of splicing Index number after

$$\text{splicing } = \frac{\text{index number to be spliced} \times \text{old index number of existing base}}{100}$$

**Example:** The index A given was started in 1993 and continued up to 2003 in which year another index B was started. Splice the index B to index A so that a continuous series of index is made

| *Year* | *Index A* | *Index B* | *Year* | *Index A* | *Index B* |
|---|---|---|---|---|---|
| *1993* | *100* | | *2002* | *138* | |
| *1994* | *110* | | *2003* | *150* | |

| | | | |
|---|---|---|---|
| *1995* | *112* | *2004* | *100* |
| - | | *2005* | *120* |
| - | | *2006* | *130* |
| - | | *2007* | *150* |



**Solution.** INDEX *B* SPLICED TO INDEX *A*

| *Year* | *Index A* | *Index B* | *Index B spliced to Index A 1993 as base* |
|---|---|---|---|
| 1993 | 100 | | |
| 1994 | 110 | | |
| 1995 | 112 | | |
| — | | | |
| — | | | |
| 2002 | 138 | | |
| 2003 | 150 | 100 | $\frac{150}{100} \times 100 = 150$ |
| 2004 | | 120 | $\frac{150}{100} \times 120 = 180$ |
| 2005 | | 140 | $\frac{150}{100} \times 140 = 210$ |
| 2006 | | 130 | $\frac{150}{100} \times 130 = 195$ |
| 2007 | | 150 | $\frac{150}{100} \times 150 = 225$ |

The spliced index now refers to 1993 as base and we can make as continuous comparison of index numbers from 1993 onwards.

In the above case it is also possible to splice the new index in such a manner that a comparison could be made with 2003 as base. This would be done by multiplying the old index by the ratio 100/150. Thus the spliced index for 1993 would be $\frac{100}{150}$ x 100 = 66.7 for 1994, $\frac{100}{150}$ x 110 = 73.3, for 1995, $\frac{100}{150}$ x 112 = 74.6, etc. This process appears to be more useful because a recent year can be kept as a base. However, much would depend upon the object.

## 8.10 DEFLATING:

Deflating means correcting or adjusting a value which has inflated. It makes allowances for the effect of price changes. When prices rise, the purchasing power of money declines. If the money incomes of people remain constant between two periods and prices of commodities are doubled the purchasing power of money is reduced to half. For example if there is an increase in the price of rice from Rs10/kg in the year 1980 to Rs20/kg in the year 1982. then a person can buy only half kilo of rice with Rs10. so the purchasing power of a rupee is only 50paise in 1982 as compared to 1980.

Thus the purchasing power of money = $\frac{1}{price\ index}$

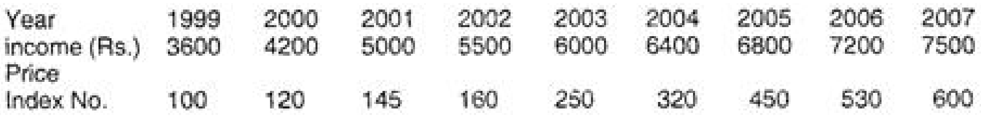
In times of rising prices the money wages should be deflated by the price index to get the figure of real wages. The real wages alone tells whether a wage earner is in better position or in worst position.

For calculating real wage, the money wages or income is divided by the corresponding price index and multiplied by 100.

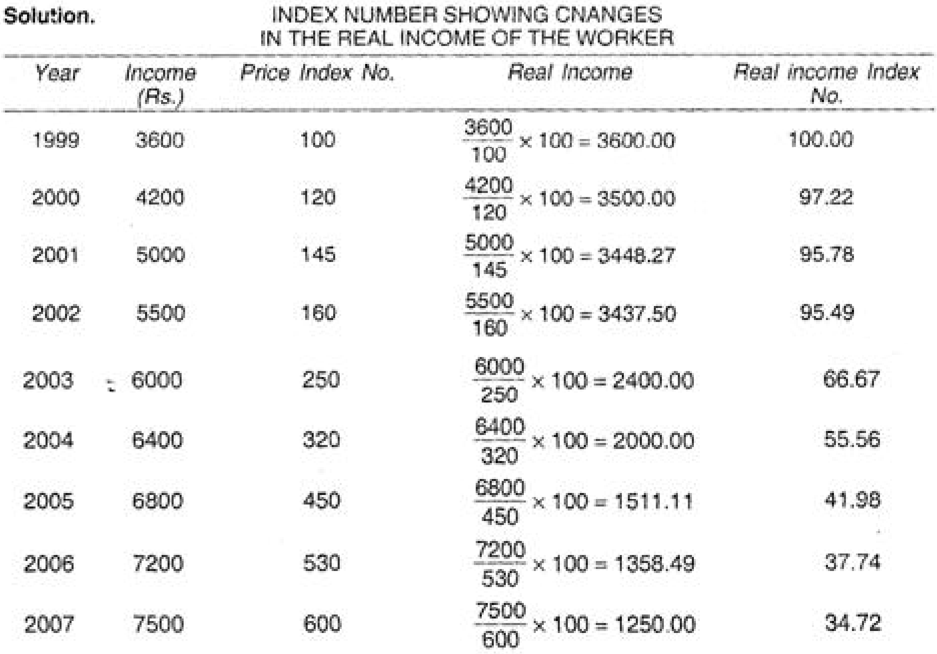
$$\text{i.e. Real wages} = \frac{Money\ \ wages}{\Pr ice\ index}\times 100$$

$$\text{Thus Real Wage Index=} \frac{\operatorname{Re} al\ wage\ of\ current\ year}{\operatorname{Re} al\ wage\ of\ base\ year}\times 100$$

**Example:** The following table gives the annual income of a worker and the general Index Numbers of price during 1999-2007. Prepare Index Number to show the changes in the real income of the teacher and comment on price increase



| Year | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|---|---|---|---|---|---|---|---|---|
| income (Rs.) | 3600 | 4200 | 5000 | 5500 | 6000 | 6400 | 6800 | 7200 | 7500 |
| Price<br>Index No. | 100 | 120 | 145 | 160 | 250 | 320 | 450 | 530 | 600 |



**Solution.** INDEX NUMBER SHOWING CNANGES IN THE REAL INCOME OF THE WORKER

| Year | Income (Rs.) | Price Index No. | Real Income | Real income Index No. |
|---|---|---|---|---|
| 1999 | 3600 | 100 | $\frac{3600}{100} \times 100 = 3600.00$ | 100.00 |
| 2000 | 4200 | 120 | $\frac{4200}{120} \times 100 = 3500.00$ | 97.22 |
| 2001 | 5000 | 145 | $\frac{5000}{145} \times 100 = 3448.27$ | 95.78 |
| 2002 | 5500 | 160 | $\frac{5500}{160} \times 100 = 3437.50$ | 95.49 |
| 2003 | 6000 | 250 | $\frac{6000}{250} \times 100 = 2400.00$ | 66.67 |
| 2004 | 6400 | 320 | $\frac{6400}{320} \times 100 = 2000.00$ | 55.56 |
| 2005 | 6800 | 450 | $\frac{6800}{450} \times 100 = 1511.11$ | 41.98 |
| 2006 | 7200 | 530 | $\frac{7200}{530} \times 100 = 1358.49$ | 37.74 |
| 2007 | 7500 | 600 | $\frac{7500}{600} \times 100 = 1250.00$ | 34.72 |

The method discussed above is frequently used to deflate individual values, value series or value indices. Its special use is in problems dealing with such diversified things as rupee sales, rupee inventories of manufacturer’s, wholesaler’s and retailer’s income, wages and the like.

**C. No. : BCG-502** **UNIT : III**
**SEMESTER: V** **LESSON: 9**

## COST OF LIVING INDEX NUMBERS

### Structure

9.1 Cost of living index numbers

9.2 Main steps or problems in construction of cost of living index numbers

9.3 Uses of cost of living index numbers

9.4 Methods for construction of cost of living index numbers

9.5 Possible errors in construction of cost of living index numbers

9.6 Problems or steps in construction of wholesale price index numbers (WPI)

9.7 Wholesale price index numbers (Vs) consumer price index numbers

9.8 Importance and methods of assigning weights

9.9 Limitations or demerits of index numbers

9.10 Summary

9.11 Self Assessment Exercises

9.12 Suggested Readings

## 9.1 COST OF LIVING INDEX NUMBERS (OR) CONSUMER PRICE INDEX NUMBERS

The ***cost of living index numbers*** measures the changes in the level of prices of commodities which directly affects the cost of living of a specified group of persons at a specified place. The general index numbers fails to give an idea on cost of living of different classes of people at different places.

Different classes of people consume different types of commodities, people's consumption habit is also vary from man to man, place to place and class to class i.e. richer class, middle class and poor class. For example the cost of living of rickshaw pullers at BBSR is different from the rickshaw pullers at Kolkata. The consumer price index helps us in determining the effect of rise and fall in prices on different classes of consumers living in different areas.

## 9.2 MAIN STEPS OR PROBLEMS IN CONSTRUCTION OF COST OF LIVING INDEX NUMBERS

The following are the main steps in constructing a cost of living index number.

### 1. 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, labors, etc. Along with the class of people it is also necessary to decide the geographical area covered by the index, such as a city, or an industrial area or a particular locality in a city.

### 2. Conducting family budget enquiry

Once the scope of the index is clearly defined the next step is to conduct a sample family budget enquiry i.e. we select a sample of families from the class of people for whom the index is intended and scrutinize their budgets in detail. The enquiry should be conducted during a normal period i.e. a period free from economic booms or depressions. The purpose of the enquiry is to determine the amount; an average family spends on different items. The family budget enquiry gives information about the nature and quality of the commodities consumed by the people. The commodities are being classified under following heads

i) Food

ii) Clothing

iii) Fuel and Lighting

iv) House rent

v) miscellaneous

**3. Collecting retail prices of different commodities**

The collection of retail prices is a very important and at the same time very difficult task, because such prices may vary from lace to place, shop to shop and person to person. Price quotations should be obtained from the local markets, where the class of people reside or from super bazaars or departmental stores from which they usually make their purchases.

## 9.3 USES OF COST OF LIVING INDEX NUMBERS

1. Cost of living index numbers indicate whether the real wages are rising or falling. In other words they are used for calculating the real wages and to determine the change in the purchasing power of money.

$$\text{Purchasing power of money} = \frac{1}{\text{Cost of living index number}}$$

$$\mathrm{R}\textit{eal Wages} = \frac{\text{Money wages}}{\text{Cost of living index umbers}} \times 100$$

2. Cost of living indices are used for the regulation of D.A or the grant of bonus to the workers so as to enable them to meet the increased cost of living.

3. Cost of living index numbers are used widely in wage negotiations.

4. These index numbers also used for analyzing markets for particular kinds of goods.

## 9.4 METHODS FOR CONSTRUCTION OF COST OF LIVING INDEX NUMBERS

Cost of living index number can be constructed by the following formulae.

1) Aggregate expenditure method or weighted aggregative method

2) Family budget method or the method of weighted relatives

### 1) Aggregate expenditure method or weighted aggregative method

In this method the quantities of commodities consumed by the particular group in the base year are taken as weights. The formula is given by

consumer price index = $\frac{\sum p_1q_0}{\sum p_0q_0} \times 100$

**Steps:**

i) The prices of commodities for various groups for the current year is multiplied by the quantities of the base year and their aggregate expenditure of current year is obtained .i.e. $\sum p_1q_0$

ii) Similarly obtain $\sum p_0q_0$

iii) The aggregate expenditure of the current year is divided by the aggregate expenditure of the base year and the quotient is multiplied by 100.

Symbolically $\frac{\sum p_1q_0}{\sum p_0q_0} \times 100$

### 2) Family budget method or the method of weighted relatives

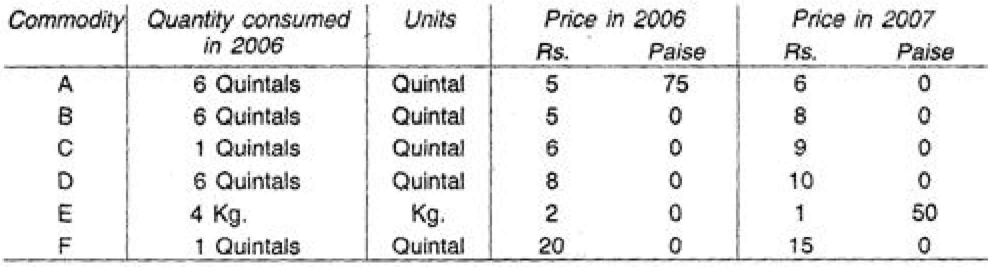
In this method cost of living index is obtained on taking the weighted average of price relatives, the weights are the values of quantities consumed in the base year i.e. . Thus the consumer price index number is given by consumer price index = $\frac{\sum pv}{\sum v}$

Where $p = \frac{p_1}{p_o} \times 100$ for each item

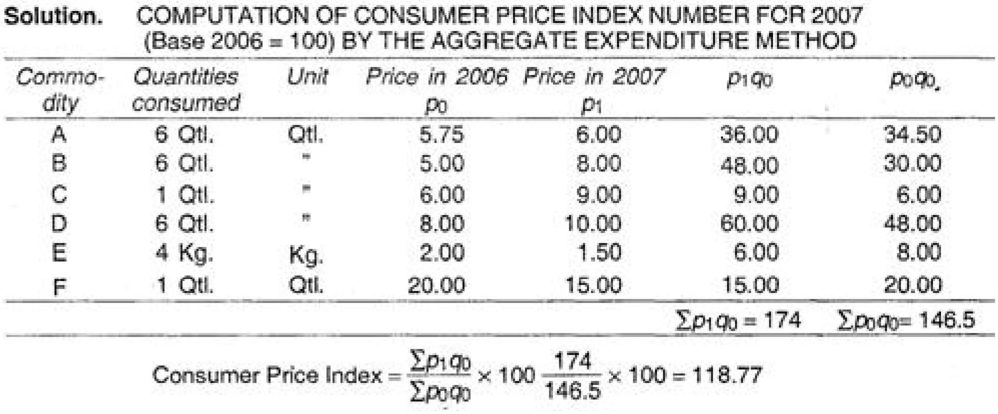
$v = p_0 q_0$, value on the base year

**Note:** It should be noted that the answer obtained by applying the aggregate expenditure method and family budget method shall be same.

**Example:** Construct the consumer price index number for 2007 on the basis of 2006 from the following data using (i) the aggregate expenditure method, and (ii) the family budget method.



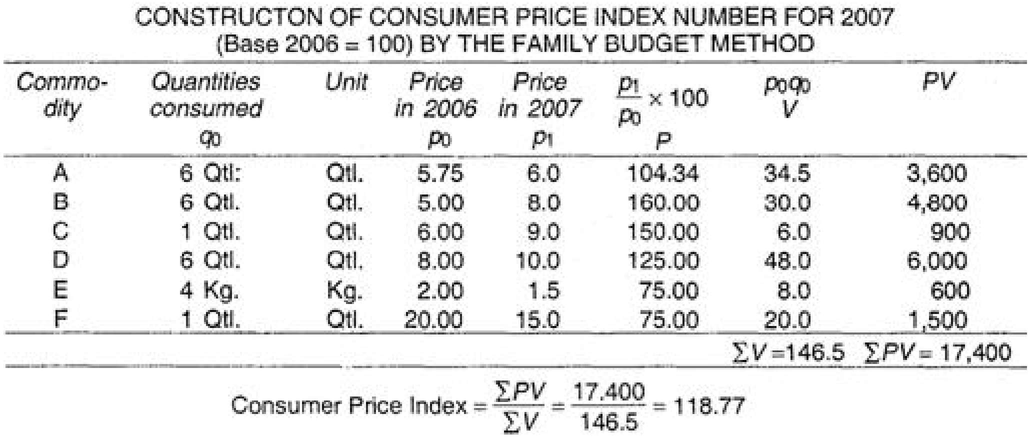
| *Commodity* | *Quantity consumed in 2006* | *Units* | *Price in 2006* | | *Price in 2007* | |
|---|---|---|---|---|---|---|
| | | | *Rs.* | *Paise* | *Rs.* | *Paise* |
| A | 6 Quintals | Quintal | 5 | 75 | 6 | 0 |
| B | 6 Quintals | Quintal | 5 | 0 | 8 | 0 |
| C | 1 Quintals | Quintal | 6 | 0 | 9 | 0 |
| D | 6 Quintals | Quintal | 8 | 0 | 10 | 0 |
| E | 4 Kg. | Kg. | 2 | 0 | 1 | 50 |
| F | 1 Quintals | Quintal | 20 | 0 | 15 | 0 |



**Solution.** COMPUTATION OF CONSUMER PRICE INDEX NUMBER FOR 2007 (Base 2006 = 100) BY THE AGGREGATE EXPENDITURE METHOD

| *Commodity* | *Quantities consumed* | *Unit* | *Price in 2006* $p_0$ | *Price in 2007* $p_1$ | $p_1q_0$ | $p_0q_0$ |
|---|---|---|---|---|---|---|
| A | 6 Qtl. | Qtl. | 5.75 | 6.00 | 36.00 | 34.50 |
| B | 6 Qtl. | " | 5.00 | 8.00 | 48.00 | 30.00 |
| C | 1 Qtl. | " | 6.00 | 9.00 | 9.00 | 6.00 |
| D | 6 Qtl. | " | 8.00 | 10.00 | 60.00 | 48.00 |
| E | 4 Kg. | Kg. | 2.00 | 1.50 | 6.00 | 8.00 |
| F | 1 Qtl. | Qtl. | 20.00 | 15.00 | 15.00 | 20.00 |
| | | | | | $\Sigma p_1q_0 = 174$ | $\Sigma p_0q_0 = 146.5$ |

Consumer Price Index = $\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100 \frac{174}{146.5} \times 100 = 118.77$



CONSTRUCTON OF CONSUMER PRICE INDEX NUMBER FOR 2007
(Base 2006 = 100) BY THE FAMILY BUDGET METHOD

| *Commo-dity* | *Quantities consumed* $q_0$ | *Unit* | *Price in 2006* $p_0$ | *Price in 2007* $p_1$ | $\frac{p_1}{p_0} \times 100$ $P$ | $p_0q_0$ $V$ | *PV* |
|---|---|---|---|---|---|---|---|
| A | 6 Qtl: | Qtl. | 5.75 | 6.0 | 104.34 | 34.5 | 3,600 |
| B | 6 Qtl. | Qtl. | 5.00 | 8.0 | 160.00 | 30.0 | 4,800 |
| C | 1 Qtl. | Qtl. | 6.00 | 9.0 | 150.00 | 6.0 | 900 |
| D | 6 Qtl. | Qtl. | 8.00 | 10.0 | 125.00 | 48.0 | 6,000 |
| E | 4 Kg. | Kg. | 2.00 | 1.5 | 75.00 | 8.0 | 600 |
| F | 1 Qtl. | Qtl. | 20.00 | 15.0 | 75.00 | 20.0 | 1,500 |
| | | | | | | $\Sigma V$ =146.5 | $\Sigma PV$ = 17,400 |

$$\text{Consumer Price Index} = \frac{\Sigma PV}{\Sigma V} = \frac{17.400}{146.5} = 118.77$$

Thus, the answer is the same by both the methods. However, the reader should prefer the aggregate expenditure method because it is far more easier to apply compared to the family budget method.

## 9.5 POSSIBLE ERRORS IN CONSTRUCTION OF COST OF LIVING INDEX NUMBERS

Cost of living index numbers or its recently popular name consumer price index numbers are not accurate due to various reasons.

1. Errors may occur in the construction because of inaccurate specification of groups for whom the index is meant.
2. Faulty selection of representative commodities resulting out of unscientific family budget enquiries.
3. Inadequate and unrepresentative nature of price quotations and use of inaccurate weights
4. Frequent changes in demand and prices of the commodity
5. The average family might not be always a representative one.

## 9.6 PROBLEMS OR STEPS IN CONSTRUCTION OF WHOLESALE PRICE INDEX NUMBERS (WPI)

Index numbers are the best indicators of the economic progress of a community, a nation and the world as a whole. Wholesale price index numbers can also be constructed for different economic activities such as Indices of Agricultural production, Indices of Industrial production, Indices of Foreign Trade etc. Besides some International organizations like the United Nations Organization, the F.A.O. of the U.N., the World Bank and International Labour Organization, there are a number of organizations in the country who publish index numbers on different aspects. These are (a) Ministry of Food and Agriculture, (b) Reserve Bank of India, (c) Central Statistical Organization, (d) Department of Commercial Intelligence and Statistics, (e) Labour Bureau, (f) Eastern Economist. The Central Statistical Organization of the Government of India publishes a Monthly Abstract of' Statistics which contains All India index numbers of Wholesale Prices (Revised series : Base year 1981-82) both commodity-wise and also for the aggregate.

### i. Purpose or object of index numbers.

A wholesale price index number which is properly designed for a purpose can be most useful and powerful tool. Thus the first and the foremost problem are to determine the purpose of index numbers. If we know the purpose of the index numbers we can settle some related problems.

### ii. Selection of commodities

Representative items should be taken into consideration. The items may be grouped into relatively homogeneous heads to make the calculation. The construction of WPI of a region or country we may group the commodities as (1) Primary Articles - (a) Food Articles (b) Non-food Articles (c) Minerals (ii) Fuel. Power, Light and Lubricants (iii) Manufactured Products (iv) Chemicals and Chemical Products (v) Machinery and Machine Equipments

(vi) Other Miscellaneous Manufacturing Industries.

### iii. Selection of base period

1. The base period must be a normal period i.e. a period frees from all sorts of abnormalities or random fluctuations such as labor strikes, wars, floods, earthquakes etc.
2. The base period should not be too distant from the given period. Since index numbers are essential tools in business planning and economic policies the base period should not be too far from the current period. For example for deciding increase in dearness allowance at present there is no advantage in taking 1950 or 1960 as the base, the comparison should be with the preceding year after which the DA has not been increased.
3. Fixed base or chain base .While selecting the base a decision has to be made as to whether the base shall remain fixing or not i.e. whether we have fixed base or chain base. In the fixed base method the year to which the other years are compared is constant. On the other hand, in chain base method the prices of a year are linked with those of the preceding year. The chain base method gives a better picture than what is obtained by the fixed base method

### iv. Data for index numbers

The data, usually the set of prices and of quantities consumed of the selected commodities for different periods, places etc. constitute the raw material for the construction of wholesale rice index numbers. The data should be collected from reliable sources such as standard trade journals, official publications etc.

### v. Selection of appropriate weights

A decision as to the choice of weights is an important aspect of the construction of index numbers. The problem arises because all items included in the construction are not of equal importance. So proper weights should be attached to them to take into account their relative importance. Thus there are two type of indices.

1. Un weighted indices- in which no specific weights are attached
2. Weighted indices- in which appropriate weights are assigned to various items.

**vi. Choice of average.**

Since index numbers are specialized averages, a choice of average to be used in their construction is of great importance. Usually the following averages are used.

iv) A.M

v) G.M

vi) Median

Among these averages G.M is the appropriate average to be used. But in practice G.M is not used as often as A.M because of its computational difficulties.

**vii. Choice of formula.**

The selection of a formula along with a method of averaging depends on data at hand and purpose for which it is used. Different formulae developed for the purpose have already been discussed in earlier sections.

## 9.7 WHOLESALE PRICE INDEX NUMBERS (VS) CONSUMER PRICE INDEX NUMBERS

1. The wholesale price index number measures the change in price level in a country as a whole. For example economic advisors index numbers of wholesale prices.Where as cost of living index numbers measures the change in the cost of living of a particular class of

people stationed at a particular place. In this index number we take retail price of the commodities.

2. The wholesale price index number and the consumer price index numbers are generally different because there is lag between the movement of wholesale prices and the retail prices.

3. The retail prices required for the construction of consumer price index number increased much faster than the wholesale prices i.e. there might be erratic changes in the consumer price index number unlike the wholesale price index numbers.

4. The method of constructing index numbers in general the same for wholesale prices and cost of living. The wholesale price index number is based on different weighting systems and the selection of commodities is also different as compared to cost of living index number

## 9.8 IMPORTANCE AND METHODS OF ASSIGNING WEIGHTS

The problem of selecting suitable weights is quite important and at the same time quite difficult to decide. The term weight refers to the relative importance of the different items in the construction of the index. Generally various items say wheat, rice, kerosene, clothing etc. included in the index are not of equal importance, proper weights should be attached to them to take into their relative importance. Thus there are two types of indices.

1) Unweighted indices - in which no specific weights are attached to various commodities.

2) Weighted indices - in which appropriate weights are assigned to various commodities.

The Unweighted indices can be interpreted as weighted indices by assuming the corresponding weight for each commodity being unity. But actually the commodities included in the index are all not of equal importance. Therefore it is necessary to adopt some suitable method of weighting, so that arbitrary and haphazard weights may not affect the results.

There are two methods of assigning weights.

i) Implicit weighting

ii) Explicit weighting

In implicit weighting, a commodity or its variety is included in the index a number of times. For example if wheat is to be given in an index twice as much times as rice then the weight of wheat is two. Where as in explicit weighting two types of weights can be assigned. i.e. quantity weights or value weights.

A quantity weight symbolized by q means the amount of commodity produced, distributed or consumed in some time period. A value weight in the other hand combines price with quantity produced, distributed or consumed and is denoted by v=pq.

For example quantity weights are used in the method of weighted aggregative like Lasperey's, Paasche's index numbers and value weights are used in the method of weighted average of price relatives.

## 9.9 LIMITATIONS OR DEMERITS OF INDEX NUMBERS

Although index numbers are indispensable tools in economics, business, management etc, they have their limitations and proper care should be taken while interpreting them. Some of the limitations of index numbers are

1. Since index numbers are generally based on a sample, it is not possible to take into account each and every item in the construction of index.

2. At each stage of the construction of index numbers, starting from selection of commodities to the choice of formulae there is a chance of the error being introduced.

3. Index numbers are also special type of averages, since the various averages like mean, median, G.M have their relative limitations, their use may also introduce some error

4. None of the formulae for the construction of index numbers is exact and contains the so called formula error. For example Lasperey's index number has an upward bias while Paasche's index has a downward bias.

5. An index number is used to measure the change for a particular purpose only. Its misuse for other purpose would lead to unreliable conclusions.

6. In the construction of price or quantity index numbers it may not be possible to retain the uniform quality of commodities during the period of investigation.

## 9.10 SUMMARY

Index numbers are the best indicators of the economic progress of a community, a nation and the world as a whole. Wholesale price index numbers can also be constructed for different economic activities such as Indices of Agricultural production, Indices of Industrial production, Indices of Foreign Trade etc. Besides some International organizations like the United Nations Organization, the F.A.O. of the U.N., the World Bank and International Labour Organization, there are a number of organizations in the country who publish index numbers on different aspects

## 10.11 SELF ASSESSMENT EXCERCISE

1. Describe a method of constructing price index number.

____________________________________________________________

____________________________________________________________

____________________________________________________________

____________________________________________________________

2. Why the cost of living index number for slum dwellers is different from that of I.A.S officers?

3. Differentiate wholesale price index number with the cost of living index numbers

## 9.12 FURTHER READINGS

1. Elehance : Advanced Statistics
2. S.P. Gupta : Statistics
3. Vishwanathan : Business Statistics : An Applied Orientation.

C. No. : BCG-502 UNIT : IV
SEMESTER: V LESSON: 10

# CORRELATION

**Structure**

10.1 Introduction

10.2 Objectives

10.3 Meaning of Correlation

10.4 Methods of Studying Correlation

10.5 Karl Pearson's Coefficient of Correlation

10.6 Limitations of Correlation Coefficient

10.7 Summary

10.8 Self Assessment Exercises

10.9 Suggested Readings

## 10.1 INTRODUCTION

In the previous lessons we have confined our discussions with the distributions of the data involving only one variable. Such a distribution is called univariate distribution. However, in practice, we come across situations where more than one variable are involved. For example, demand and supply of a commodity, volume and temperature of a gas, heights and weight of student in a class etc. In such situations, our aim is to determine whether there exists a relationship between two variables. If such a relationship can be expressed by a mathematical formula, then we shall be able to use it for an analysis of data. Correlation is the method that deals with the analysis of such relationships between two variables.

The present lesson deals with the meaning of correlation, types of correlation, methods of determining correlation, its limitations etc.

## 10.2 OBJECTIVES

After successful reading of this lesson, the students will be able to :

- understand the concept of correlation,
- draw a scatter diagram and have an idea about types of correlation,
- compute and interpret correlation, and
- know the limitations of correlation coefficient

## 10.3 MEANING OF CORRELATION SIGNIFICANCE

If for every value of a variable, X, we have a corresponding value of another variable Y, the resulting series of pairs of values of two variables is known as bivariate population and its distribution is known as bivariate distribution.

In a bivariate distribution if the change in one variable appears to be accompanied by a change in other variable and vice-versa, then the two variables are said to be correlated and this relationship is called correlation or co-variation. In other words, the tendency of simultaneous variation of the two variables is called correlation. Then correlation studies the degree of inter-dependence between two variables.

**Definitions**

"Correlation analysis deals with the association between two or more variables". **-Simpson & Kafka**

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

"Correlation is an analysis of the correlation between two or more variables". **-A.M. Tuttle**

## Types of Correlation

The correlation is of the following types:

1. **Positive and Negative Correlation:** As a first step, the correlation may be classified according to the direction of change in the two variables. When the increase (or decrease) in one variable results in a corresponding increase (or decrease) in the other, the correlation is said to be positive. Thus positive correlation means change in both variables in the same direction. For example, increase in amount spent on advertisement and sales. However, if the two variables deviate in opposite direction, they are said to be negatively correlated. In other words, if the increase (or decrease) in one variable creates a decrease (or increase) in the other variable, then the correlation between two variables is said to be negative.

2. **Simple, Partial and Multiple Correlation-** The distinction between simple, partial and multiple correlations are based upon the number of variables studied. When only two variables are studied it is a problem of simple correlation. When three or more variables are studied, it is a problem of either multiple or partial correlation. In multiple correlations three or more variables are studied simultaneously. For example, when we study the relationship between the yield of rice per acre and both the amount of rainfall and the amount of fertilizers used, it is a problem of correlation.

3. **Linear and Non-Linear (Curvilinear) Correlation-** The distribution between linear and non linear correlation is based upon the constancy of the ratio of change between the variables. If the amount of change of one variable tends to bear constant ratio to the amount of change in the other variable then the correlation is said

to be linear. Correlation would be called non-linear or curvilinear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable.

Further the correlation is perfectly positive if the change in two variables is in the same direction and same ratio. However, it is perfectly negative if the change in two variables is in opposite direction but in same ratio.

## Significance of the Study of Correlation

The study of correlation is of immense use in practical life because of the following reasons:

- Most of the variables show some kind of relationship. For example, there is relationship between price and supply, income and expenditure, etc. With the help of correlation analysis we can measure in one figure the degree of relationship existing between the variables.
- Once we know that two variables are closely related, we can estimate the value of one variable given the value of another. This is known with the help of regression analysis discussed in the next chapter.
- Correlation analysis contributes to the understanding of economic behavior, aids in locating the critically important variables on which others depend, may reveal to the economist the connection by which disturbances spread and suggest to him the paths through which stabilizing forces may become effective.

In business, correlation analysis enables the executive to estimate costs, sales, prices and other variables on the basis of some other series with which these costs, sales, or prices may be functionally related. Some of the guesswork can be removed from decisions when the relationship between a variable to be estimated and the one or more other variables on which it depends are close and reasonably invariant.

However, it should be noted that coefficient of correlation is one of the most widely used and also one of the most widely *abused* statistical measure. It is abused in the sense that one sometimes overlooks the fact that correlation measures are nothing but the strength of *linear* relationship and that it does not necessarily imply a cause-effect relationship.

- Progressive development in the methods of science and philosophy has been characterized by increase in the knowledge of relationship or correlations. In nature also one finds multiplicity of interrelated forces.
- The effect of correlation is to reduce the range of uncertainty. The prediction based on correlation analysis is likely to be more valuable and near to reality.

## 10.4 METHODS OF STUDYING CORRELATION

The following methods may be used for studying the correlation between two variables ( For ungrouped data) :

(i) Scatter Diagram Method

(ii) Karl Pearson's Coefficient of Correlation

(iii) Spearman's Coefficient of Rank Correlation.

The first two methods are discussed in the present lesson while Spearman's Coefficient of rank correlation will be discussed in Lesson 13.

**I Scatter Diagram Method :** A graphical representation of a set of pairs of values of two variables X and Y in a coordinate system is called a scatter diagram or simply dot diagram. By means of scatter diagram one can quickly judge the type of correlation between the variables. Scatter diagrams, as an example, showing various degrees of correlation are shown in the given figures

In figure (a) all the dots are lying on a straight line of positive slope, thus we have a perfect positive correlation between two variables and the value of correlation coefficient will be +1. Similarly, in fig. (b) all the dots in the diagram are lying on a straight line of negative slope and this situation shows perfect negative correlation between two variables. Here its value will be –1. In fig. (c), the dots lie close to a straight line of positive slope and this shows a high degree positive correlation. Similarly, in fig (d), the dots lie close to a straight line of negative slope which indicates that the negative correlation of high degree exists between two variables. Finally, if the dots do not follow a pattern alongwith a straight line as in fig. (e), we have no correlation or zero correlation and we may conclude that no linear relationship exists between the variables X and Y. In view of the above discussion, it is clear that the greater the scatter of dots from the straight line on the graph, the lesser the correlation.

This method has the following drawbacks :

(i) It gives only a rough idea that how the two variables are related.

(ii) It gives an idea about the direction and also whether is high or low.

(iii) It does not indicate the degree or extent of relationship existing between the two variables.

In the following section we will discuss Karl Pearson's Coefficient of Correlation which measures the correlation numerically.

## 10.5 KARL PEARSON'S COEFFICIENT OF CORRELATION

To determine the degree or extent of the linear correlation between two variables, Karl Pearson, defined a numerical measure called Correlation Coefficient denoted by r, and given by

$r$ = Cov (X, Y) / $\sigma_X \cdot \sigma_Y$

where Cov (X, Y) is the covariance between X and Y, and $?_X$ and $?_Y$ are the standard deviations of X and Y respectively. It is also known as Product Moment Correlation Coefficient or simply Coefficient of

Correlation.

If ($X_i$, $Y_i$), $i$ = 1, 2, 3, ...., $n$ be the set of values of size $n$ from a bivariate population of (X, Y). Let  and  be the means of X and Y respectively, then correlation coefficient is given by

$r$ = 1/N " (X-X(Y-Y)/ *h.* $\surd(1/N(X - X)2 - ((1/N(Y - Y)2$ ]

It can also be expressed as

The study of the correlation is of immense use in practical life where most of the variables show some kind of relationship. With the help of correlation coefficient we can measure the degree of relationship existing between variables.

The value of the coefficient of correlation ($r$) always lies between—1 and +1. Where $r$ = –1 or +1, the correlation is said to be perfectly negative or positive. An intermediate value of $r$ between –1 and +1 indicates the degree of linear relationship between two variables X and Y whereas its sign tells about the direction of relationship. $r$ = 0 means no linear relationship between two variables. However, if covariance is zero then the variables are said to be independent and $r$ = 0.

The correlation coefficient is a pure number, independent of the unit of measurement.

## 10.6 LIMITATIONS OF CORRELATION COEFFICIENT

The correlation coefficient is a measure of the relationship between two variables, say X and Y. While it generally, serves as a useful statistical tool, we should also be aware of its limitations. The correlation coefficient is a measure of statistical relationship, and not of casual relationship, between the two variables. This means that the value of $r$ tells us whether, and with what regularity, $y$ increases or decreases as X increase. But it cannot tell us whether that increase or decrease is due to any casual or cause-effect relationship between two variables. Further, the correlation coefficient is a measure of linear statistical relationship only, and may fail to be a proper index of statistical relationship in case it is non-linear.

## 10.7 SUMMARY

In a bivariate distribution we have two variables of observation on which values are recorded for each unit of observations. The two variables in a bivariate distribution may be interdependent. The interdependent between the variables is called covariation or correlation. Correlation analysis deals with the measurement of the extent or degree of relationship between two variables. However, it does not help in the study of cause and effect relationship between two variables. Correlation may be positive or negative depending upon the direction of change in the two variables. That is, if the two variables deviate in opposite direction, they are said to be negatively correlated and if the deviation is in same direction then there is a positive correlation.

We have three important methods of studying correlation. Scatter diagram method is a graphic device for drawing certain conclusions about the correlation. However, Karl Pearson's correlation coefficient is an important and precise method of computing the degree of correlation between the two variables. Its value lies between –1 and +1. It is a pure number independent of the unit of measurement. Third method of measuring correlation is the rank correlation method which will be discussed in Lesson 13.

## 10.8 EXERCISES

1. What do you understand by the terms :

   (i) Bivariate distribution

   (ii) Correlation

   (iii) Dot diagram

2. Discuss the meaning of correlation and distinguish between positive and negative correlations.

3. What do you mean by scatter diagram? How is scatter diagram used to determine correlation?

4. Define Karl Pearson's coefficient of correlation. What is it intend to measure?

5. Give some examples of positive and negative correlations.

6. What will be your interpretation if

   (i) $r = 0$ (ii) $r = +1$ (iii) $r = -1$

## 10.9 FURTHER READINGS

1. Elehance : Advanced Statistics
2. S.P. Gupta : Statistics
3. Vishwanathan : Business Statistics : An Applied Orientation.

**C. No. : BCG-502** **UNIT : IV**

**SEMESTER: V** **LESSON: 11**

# SPEARMAN'S COEFFICIENT OF RANK CORRELATION

**Structure**

11.1 Introduction

11.2 Objectives

11.3 Spearman's Rank Correlation

11.4 Rank Correlation For Tied Ranks

11.5 Merits and Demerits of Rank Correlation Method

11.6 Summary

11.7 Self Assessment Exercises

11.8 Suggested Readings

## 11.1 INTRODUCTION

We have discussed Karl Pearson's Coefficient of Correlation, in previous Lesson, which studied the degree of co variability of linear relationship between two variables for which the observations are definitely measured. But often we come across situations when definite measurements on the variables are not possible. For example, if a group of *n* students is arranged in order of merit or proficiency in Business Statistics and Economics without any attempt to asses numerically assigning to each student a number which indicates his position in that group; the students are then said to be ranked and the number of a particular student is his rank.

In such type of situations Spearman’s Coefficient of rank correlation is determined.

## 11.2 OBJECTIVES

The following are the main objectives of this lesson:

- to understand the meaning of ranks and rank correlation.
- to learn the computational procedure of rank correlation.
- to know the merits and demerits of rank correlation.
- to know the concurrent deviation method.

## 11.3 SPEARMAN’S COEFFICIENT OF RANK CORRELATION

Spearman suggested that the relationship between two ranks may be studied by calculating the Pearson’s Coefficient of Correlation for numerical values that happens to rank. The Spearman’s Coefficient of Rank Correlation, denoted by a Greek letter ρ (rho), is given by

$\rho = 1-6"D^2 / n(n^2-1)$

where

D = difference between paired ranks, i.e.

$Di = Rxi-Ryi$

$Rxi$ = rank of *i*th individual of variable X

$Ryi$ = rank of *i*th individual of variable Y

$n$ = the number of items ranked

**Remarks :**

1. In fact, the coefficient of rank correlation, is nothing but Karl Pearson’s coefficient of correlation between two sets of ranks.
2. In view of remark 1, its value lies between –1 and +1.

3. The value ρ = +1 stands for a perfect positive agreement between two sets of ranks, while ρ = –1 implies a perfect negative relationship.

4. The basic assumption in this correlation is that no two individuals be equal in either classification so that no ties in ranks exists.

**Example 1.** In a beauty contest two judges rank the 10 entries as follows :

| Contestant : | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Judge I : | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Judge II : | 2 | 3 | 1 | 6 | 4 | 5 | 8 | 7 | 10 | 9 |

Find the degree of agreement between ranks given by two judges.

**Solution :** In order to find the degree of agreement between ranks, we find the coefficient of rank correlation.

**Computation of Rank Correlation**

| Contestant | Rank by | Rank by | D = R1–R2 | D2 |
|---|---|---|---|---|
| Judge I | Judge II | | | |
| A | 1 | 2 | –1 | 1 |
| B | 2 | 3 | –1 | 1 |
| C | 3 | 1 | 2 | 4 |
| D | 4 | 6 | –2 | 4 |
| E | 5 | 4 | 1 | 1 |
| F | 6 | 5 | 1 | 1 |
| G | 7 | 8 | –1 | 1 |
| H | 8 | 7 | 1 | 1 |
| I | 9 | 10 | –1 | 1 |
| J | 10 | 9 | 1 | 1 |
| Total | | | | 16 |

Thus, rank correlation is given by

$\rho$ = 1– = 1–6"$D^2$/ n($n^2$-1)

= 1–6*16/10(100-1)1- 96/990 = 894/990 = 0.903

Thus, $\rho$ = 0.903 shows a high degree of agreement between ranks given by two judges.

**Example 2.** Calculate rank correlation coefficient from the following marks given out of 200 by two judges X and Y in a music competition to 8 participants:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Participant No | : | 1, | 2, | 3, | 4, | 5, | 6, | 7, | 8 |
| Marks awarded by X | : | 74, | 98, | 110, | 70, | 65, | 85, | 88, | 59 |
| Marks awarded by Y | : | 121, | 133, | 170, | 102, | 90, | 152, | 160, | 85 |

**Solution :** To determine rank correlation we first assign ranks to marks awarded by judge X and Y by allotting the first rank to the highest marks, second rank to next highest marks, and so on. The ranks so obtained for the two judges are given in the following table :

| Participant No. | Marks by X | Marks by Y | Rank for Marks X $R_1$ | Rank for Marks Y $R_2$ | D = $R_1$–$R_2$ | $D^2$ |
|---|---|---|---|---|---|---|
| 1 | 74 | 121 | 5 | 5 | 0 | 0 |
| 2 | 98 | 133 | 2 | 4 | –2 | 4 |
| 3 | 110 | 170 | 1 | 1 | 0 | 0 |
| 4 | 70 | 102 | 6 | 6 | 0 | 0 |
| 5 | 65 | 90 | 7 | 7 | 0 | 0 |
| 6 | 85 | 152 | 4 | 3 | 1 | 1 |
| 7 | 88 | 160 | 3 | 2 | 1 | 1 |
| 8 | 59 | 85 | 8 | 8 | 0 | 0 |
| Total | | | | | | 6 |

Thus, Rank correlation is

$\rho$ = $1–6\Sigma D^2 / n(n^2-1)$= 1– 6*6/8(64-1)

= 1– 36/504= 468/504 = 0.929

which shows a high degree of positive correlation between the marks awarded by the two judges.

## 11.4 RANK CORRELATION FOR TIED RANKS

In the previous section we have assumed that no two values in either series were equal (means no tie) while discussing rank correlation. However, in some cases, we may have two or more equal observations in either of the two series or in both the series. In such cases, we assign average (mean) ranks to the set of tied observations. For example, in assigning ranks to 10 observations we may note that the third largest observation is repeating three times. These three observations (3rd, 4th and 5th) are therefore tied and each is assigned average rank (3+4+5) = 4. The next individual assigned the rank 6. If we find again a tie of two observations, we assign the rank = 7.5 each and the next individual is assigned rank 9.

Obviously the formula

$$\rho = 1–6\Sigma D^2 / n(n^2-1) \qquad ..........(1)$$

cannot be used if there are ties in either one or both series. The Spearman's Coefficient of Rank Correlation is then corrected for these tied ranks and now given as follows :

$$\rho = 1–6[\ \Sigma D^{2\ +”m(m2-1/2}/\ n(n^2-1) \qquad .........(2)$$

where *m* be the number of tied observations with common ranks.

**NOTE :-** The adjustment consists of adding *m*(*m*2–1)/12 to the value ΣD2. If there are more than one set of tied observations, the correction factor *m* (*m*2–1)/12 is to be added each time to the value of ΣD2 for every value of *m*. This tendency of correction has been represented by Σ*m*(*m*2–1)/12 in formula (2).

**Example 3.** Calculate the coefficient of rank correlation from the following data :

| Marks by | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Judge I : | 48 | 33 | 40 | 09 | 16 | 16 | 65 | 24 | 16 | 57 |
| Judge II : | 13 | 13 | 24 | 06 | 15 | 04 | 20 | 09 | 06 | 19 |

**Solution :**

| Judge I<br>(X) | Judge II<br>(Y) | Rank<br>$R_1$ | Rank<br>$R_2$ | $D = R_1–R_2$ | $D^2$ |
|---|---|---|---|---|---|
| 48 | 13 | 3 | 5.5 | –2.5 | 6.25 |
| 33 | 13 | 5 | 5.5 | –0.5 | 0.25 |
| 40 | 24 | 4 | 1 | –3.0 | 9.00 |
| 09 | 06 | 10 | 8.5 | 1.5 | 2.25 |
| 16 | 15 | 8 | 4 | 4 | 16.00 |
| 16 | 04 | 8 | 10 | –2 | 4.00 |
| 65 | 20 | 1 | 2 | –1 | 1.00 |
| 24 | 09 | 6 | 7 | –1 | 1.00 |
| 16 | 06 | 8 | 8.5 | –0.5 | 0.25 |
| 57 | 19 | 2 | 3 | –1 | 1.00 |
| Total | | | | | 41 |

While ranking observations of judge I i.e. X, three observations of 16 are tied at rank 7 and, as such they are each marked the average rank of (7+8+9)/3 = 8. Similarly the observation of Judge II i.e. Y, two observations of values 13 and 6 are tied at rank 5th and 8th respectively. Therefore, these observations are assigned rank (5+6)/2 = 5.5 and (8+9)/2 = 8.5 respectively. Thus, in all there are 3 tied ranks for observation 16 of Judge I. Similarly, there are 2 tied ranks for observation 13 and 2 for observation 6 of Judge II i.e. Y. Thus, we ha three sets of tied ranks with $m$ = 3, 2 and 2 respectively.

Therefore, on using formula (2), Spearman's Coefficient of rank correlation is given by

$\rho$ = 1–6[Σ“$D^2$ +”m($m^2$-1/2/ n($n^2$-1)

= 1–6[41+{3(9-1)/12+2(4-1)/12+2(4-1)}/12] / 10(100-1)

= 1–6[41+2+1/2+1/2]/10*99

= 1–258/990= 732/990 = 0.739

**Example 4.** Calculate the rank correlation coefficient from the following data :

Marks in Economics : 45, 56, 39, 54, 45, 40, 56, 60, 30, 36

Marks in Statistics : 40, 36, 30, 44, 36, 32, 45, 42, 20, 36

**Solution :**To determine rank correlation coefficient we first assign ranks to different observations (marks) of both the subjects on using tied rank produces following table :

| Marks in Economics (X) | Marks in Statistics (Y) | RX | RY | D = $R_X$–$R_Y$ | $D^2$ |
|---|---|---|---|---|---|
| 45 | 40 | 5.5 | 4 | 1.5 | 2.25 |
| 56 | 36 | 2.5 | 6 | –3.5 | 12.25 |
| 39 | 30 | 8 | 9 | –1 | 1.00 |
| 54 | 44 | 4 | 2 | 2 | 4.00 |
| 45 | 36 | 5.5 | 6 | –0.5 | 0.25 |
| 40 | 32 | 7 | 8 | –1 | 1.00 |
| 56 | 45 | 2.5 | 1 | 1.5 | 2.25 |
| 60 | 42 | 1 | 3 | –2 | 4.00 |
| 30 | 20 | 10 | 10 | 0 | 0.00 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 36 | 36 | 9 | 6 | 3 | 9.00 |
|  |  |  |  |  | 36.00 |

_ $\rho$ = 1–6[ “D$^{2}$ +”m(m$^{2}$-1/12]/ n(n$^{2}$-1)

= 1–6[36.00+2(4-1)/12+3(9-1)/12+3(9-1)/12 / 10(100-1)

= 1–6(36.00+4.5)/990

= 747/990 = 0.76

## 11.5 MERITS & DEMERITS OF RANK CORRELATION METHOD

**Merits :**

1. This method is simple and easily understandable as compared to the Karl Pearson’s method. However, where ranks are treated as scores, ? will be equal to *r*.
2. The method is specially useful when precise measurements on the variables under study are not given or cannot be obtained, i.e., when the factors under study are qualitative in nature.
3. The method can also be applied to irregular data as it does not assume that the data should be normal.

**Demerits :**

1. This method is applied to only un-grouped data.
2. The ranking procedure involved in this procedure ignores the actual magnitude of data, and, as such, the results obtained are only approximate.
3. The computation procedure becomes difficult as the paired observations increases.

## 11.6 SUMMARY

Spearman’s Coefficient of rank correlation is given by

(a) For Untied Ranks :

□ = 1–6[ “D$^2$ / n(n$^2$-1)

(b) For Tied Ranks :

ρ = 1–6[ “D$^2$ +”m(m$^2$-1)] / n(n$^2$-1)

## 11.7 SELF ASSESSMENT EXERCISES

1. What is rank correlation? Discuss its merits and demerits.

2. Write short note on Spearman’s rank correlation coefficient.

3. The rank correlation coefficient between marks obtained by some students in two subjects is 0.80. If the sum of squares of difference of ranks is 33, then find the number of student.

4. Compute the rank correlation coefficient from the following data :

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| X : | 90, | 109, | 112, | 87, | 98, | 87, | 109, | 108 |
| Y : | 72, | 74, | 80, | 76, | 74, | 70, | 68, | 70 |

Ans= 0.867

5. Ten competitions in a beauty contest are ranked by three judges in the following order.

Judge

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| A : | 1, | 6, | 5, | 10, | 3, | 2, | 4, | 9, | 7, | 8 |
| B : | 3, | 5, | 8, | 4, | 7, | 10, | 2, | 1, | 6, | 9 |
| C : | 6, | 4, | 9, | 8, | 1, | 2, | 3, | 10, | 5, | 7 |

Use Spearman's Rank Correlation to determine which pair of judges has the nearest approach to common tastes in beauty.

R(Iand II)=-0.212

R(IIand III)=-0.297

R(1and III)=0.636

## 11.8 SUGGESTED READINGS

1. Elehance : Advanced Statistics
2. S.P. Gupta : Statistics
3. Vishwanathan : Business Statistics : An Applied Orientation.

C. No. : BCG-502 UNIT : IV
SEMESTER: V LESSON: 12

# REGRESSION ANALYSIS—I

**Structure**

12.1 Introduction

12.2 Objectives

12.3 Regression Analysis and its importance

12.4 Difference between Correlation and Regression Analysis

12.5 Lines of Regression

12.6 Fitting of Regression Line of Y on X

12.7 Summary

12.8 Self Assessment Exercises

12.9 Suggested Readings

## 12.1INTRODUCTION

In the pervious lessons, we have seen that the data giving the corresponding value of two variables can be graphically represented by a scatter diagram and a method of finding the relationship between these two variables in terms of correlation coefficient were also introduced. Very often, in the study of relationship of two variables, we come across situations where one of the two variables depends on the other. In other words, what is the possible value of the dependent variable when the value of the independent variable is known? In such situations, where one of the variables is dependent and other is independent, we can find a method of estimating the numerical relationship between two variables so that given a value of the independent variable; we can forecast the average value of the dependent variable. Regression analysis serves this purpose.

## 12.2 OBJECTIVES

After successful completion of this lesson, students will be able to:

- understand the concept of regression analysis,
- distinguish between correlation and regression,
- describe simple linear regression line, and
- explain how to fit a linear regression line.

## 12.3 REGRESSION ANALYSIS AND ITS IMPORTANCE

The word regression was first introduced by Sir Francis Galton in the study of heredity in connection with the study of height of parents and their offsprings. He found that the offspring of tall or short parents tend to regress to the average height. In other words though tall fathers do tend to have tall sons, yet the average height of sons of a group of tall feathers is less than their father's height and the average height of short fathers is less than the average height of their sons. Galton termed the line describing the average relationship between the two variables as the line of regression. Thus, by regression we mean the average relationship between two variables which can be used for estimating the value of one variable from the given values of other variable. However, the dictionary meaning of regression is "Stepping Back", but nowadays it is stand for some sort of functional relationship between two or more variables. Here the variable whose value is to be predicted is called dependent or explained variable and the variable used for prediction is called independent or explanatory variable.

### Uses of Regression Analysis

Regression analysis is a branch of statistical theory that is widely used in almost all the scientific disciplines. In economics it is the basic technique for measuring or estimating the relationship among economic variables that constitute the essence of economic theory and economic life. For example, if we know that two variables, price (X) and demand (Y), are closely related we can find out the most probable value of X for a given value of Y or the most probable value of Y for a given value of X. Similarly, if we know that the amount of tax and the rise in the price of a commodity are closely related, we can find out the expected price for a certain amount of tax levy. Thus, we find that the study of regression is of considerable help to the economists and businessmen. The uses of regression are not confined to economics and business field only. Its applications are extended to almost all the natural, physical and social sciences. The regression analysis attempts to accomplish the following:

1. Regression analysis provides estimates of values of the dependent variable from values of the independent variable. The device used to accomplish this estimation procedure is the *regression line*. The regression line describes the average relationship existing between X and Y variables, i.e. it displays mean values of X for given values of Y. The equation of this line, known as the regression equation, provides estimates of the dependent variable when values of the independent variable are inserted into the equation.

2. A second goal of regression analysis is to obtain a measure of the error involved in using the regression line as a basis for estimation. For this purpose the standard error of estimate is calculated. This is a measure of the scatter or spread of the observed values of Y around the corresponding values estimated from the regression line. If the line fits the data closely, that is, if there is little scatter of the observations around the regression line, good estimates can be made of the Y variable. On the other hand, if there is a great deal of scatter of the observations around the fitted regression line, the line will not produce accurate estimates of the dependent variable.

3. With the help of regression coefficients we can calculate the correlation coefficient. The square of correlation coefficient (*r*), called coefficient of determination, measures the degree of association of correlation that exists between the two variables. It assesses the proportion of variance in the dependent variable that has been accounted for by the regression equation. In general, the greater the value of $r^2$ the better is the fit and the more useful the regression equations as a predictive device.

## 12.4 DIFFERENCE BETWEEN CORRELATION AND REGRESSION ANALYSIS

Although the two analyses are complementary to one another, yet the choice of one or the other depends upon the purpose of statistical enquiry. The following are the main differences between correlation and regression analysis:

(i) The coefficient of correlation is used to measure the degree of co variation between the two variables, while the regression analysis provides the average relationship between these variables.

(ii) Correlation does not necessarily establish causes and effect relationship. However, in regression analysis, there is a clear indication of cause and effect relationship. Here the independent variable is the cause and dependent variable is the effect.

(iii) Whereas correlation analysis is confined only to the study of linear relationship between two variables, the regression analysis deals with linear and non-linear relationships.

(iv) In correlation analysis, $r_{xy}$ measures the linear relationship between the variables X and Y. Here $r_{xy} = r_{yx}$, i.e., it is immaterial which of the two variables is taken as dependent or independent. However, in regression analysis, the identity of variables, i.e., which is dependent and which one is independent, is important.

## 12.5 LINES OF REGRESSION

If the variables in a bivariate frequency distribution are correlated, we observe that the points in a scatter diagram cluster around a straight line, called the line of regression. In a bivariate study, we have two lines of regression, namely, regression of Y on X and regression of X on Y.

The line of regression of Y on X is used to predict or estimate or forecast the value of Y for the given value of the variable X. Thus, Y is the dependent variable and X is the independent variable. The regression line of Y on X is of the form:

$Y = a+bX$

where *a* and *b* are unknown constants to be determined by observed data on the two variables X and Y.

Similarly the regression line of X on Y is used to predict the value of X for the given value of the variable Y. Here X is dependent variable and Y is independent. The regression line of X on Y is of the from :

$X = a+bY$

where *a* and *b* are unknown constants to be determined by observed data on the two variables X and Y.

## 12.6 FITTING OF REGRESSION LINE OF Y ON X

Suppose the regression line of Y on X is

$$Y = a+bX \ . \qquad ...(1)$$

where *a* and *b* are unknown constants to be determined by observed data on the two variables X and Y.

The regression line of Y on X, given by (1) can be fitted by the Method of Least Squares. That is, we choose the constant *a* and *b* in the regression line Y = *a*+*b*X in such a way that

$$\Sigma(y_i - y_i)^2 \qquad ....(2)$$

is a minimum, where  be the estimated value of Y for X = $X_i$ i.e.  = $a+bX_i$. Here the quantity given by (2) is called sum of squares of the residuals $E_i$, $E_i = y_i - y_i$

For obtaining *a* and *b*, we minimize

$$\Sigma E_i^2 = “(y_i - y_i)^2$$

$$\Sigma(y_i\text{-a-b } X_i)^2 \qquad .....(3)$$

with respect to *a* and *b*. By using the Principle of Maxima and Minima, i.e., equating to zero the partial derivatives of $\Sigma E_i^2$ w.r.t. *a* and *b*, we get

$$\Sigma Y_i = na + b\Sigma X_i \qquad .....(4)$$

$$\Sigma X_i y_i = a\Sigma X_i + b\Sigma X_i^2 \qquad .....(5)$$

Here these two equations i.e. (4) and (5) are called normal equations. Solving these equations simultaneously for *a* and *b*, we obtain.

$$b = \Sigma X_i \ y_{i\,-\,1/n}\Sigma X_i^{2\text{-}X2} = \qquad .....(6)$$

and $$a = Y - bX \qquad .....(7)$$

The values of *b* and *a* given by (6) and (7) can also be expressed in terms of correlation coefficient. As we know that

$r =$

$=$ .....(8)

Thus from (6) and 8, we have

$b = r$ .....(9)

and

$a = -r$ .....(10)

Hence, on putting the values of *b* and *a* from equations (9) and (10) in regression line of Y on X given by equation (2), we obtain the following equation of regression line of Y on X :

Y = –*r*·+*r*··*x*

= +*r*

or

*y*– = *r*· .....(11)

The quantity *r*· is called the regression coefficient of Y on X and, in general, is denoted by $b_{yx}$. Thus

$b_{yx}$ = *r*·.

**Remark**

(i) We may ignore the lower suffix *i* from $X_i$ and $Y_i$ in the formula of *a* and *b*. Thus, we may write

*b* = , and

*a* =

(ii) If X and Y are measured from their respective means i.e.

let *x* = X– and *y* = Y–, then *b* is given by

*b* =

(iii)To find the , and $b_{YX}$ we can also use step deviation method, i.e., if we assume that $d_x$ = X–A and $d_y$ = Y–B, where A and B are assumed mean of X and Y respectively, then we have

= A+Σ*dx*

= B+Σ*dy*

and

$b_{YX}$ =

(iv)The regression line of Y on X given by equation (1) is used to estimate or predict the best value of Y for a given value of the variable X.

**Example 1.** Find the regression equation of Y on X from the following data :

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

Also estimate the value of Y when the value of X = 12.

**Solution :** Let the regression line of Y on X is

$Y = a+bX$ .....(1)

| X | Y | XY | $X^2$ |
|---|---|---|---|
| 7 | 6 | 42 | 49 |
| 4 | 5 | 20 | 16 |
| 9 | 8 | 72 | 64 |
| 6 | 8 | 48 | 36 |
| 5 | 2 | 10 | 25 |
| 30 | 30 | 192 | 190 |

Here  =  =  = 6,  =  = 6,

XY = 192, and $X^2$ = 190, thus

$b( = b_{yx}) =$

= = = 1.20

and

$a$ = –$b$ = 6–1.20×6

= –1.20

Thus regression equation of Y on X becomes

Y = –1.20+1.20X

and the estimated value of Y for X = 12 is

$Y = –1.20+1.20×12$

$= –1.20+14.40$

$= 13.20$

**Example 2.** Consider the following data on heights and weights of 10 adults :

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Height (cm) : | 178 | 176 | 170 | 174 | 165 | 162 | 178 | 165 | 174 | 172 |
| Weight (kg) : | 80 | 75 | 72 | 74 | 68 | 64 | 76 | 66 | 72 | 70 |

Predict the weight of an adult whose height is 185cm.

**Solution :** Let Y = Weight and X = Height. First we find the regression line of Y on X

| Sr. No. | X | Y | $dx$ = X–174 | $dy$ = Y–70 | $dxdy$ | $d_x^2$ |
|---|---|---|---|---|---|---|
| 1 | 178 | 80 | 4 | 10 | 40 | 16 |
| 2 | 176 | 75 | 2 | 5 | 10 | 4 |
| 3 | 170 | 72 | –4 | 2 | –8 | 16 |
| 4 | 174 | 74 | 0 | 4 | 0 | 0 |
| 5 | 165 | 68 | –9 | –2 | 18 | 81 |
| 6 | 162 | 64 | –12 | –6 | 72 | 144 |
| 7 | 178 | 76 | 4 | 6 | 24 | 16 |
| 8 | 165 | 66 | –9 | –4 | 36 | 81 |
| 9 | 174 | 72 | 0 | 2 | 0 | 0 |
| 10 | 172 | 70 | –2 | 0 | 0 | 4 |
| Total | | | –26 | 17 | 192 | 362 |

Here = A+ = 174– = 171.40

= A+ = 70+ = 71.70

$b$ ( = $b_{yx}$) =

=

= = 0.802

Thus, regression line of Y on X is

Y– = $b_{yx}$

or Y–71.70 = 0.802 (X–171.40)

Y = 0.802X–137.463+71.70

Y = –66.06+0.802X

Hence the weight (Y) of an adult, whose height (X) is 185, is given by

Y = –66.06+0.802×185 = 82.31

**Example 3.** Fit the equation of regression line of Y on X for the following data :

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| X | : | 57 | 58 | 59 | 60 | 61 | 62 | 64 |
| Y | : | 77 | 78 | 75 | 82 | 82 | 79 | 81 |

**Solution :**

| Sr. No. | X | Y | $x$ = X– | $y$ = Y– | | |
|---|---|---|---|---|---|---|
| | | | | | = $x^2$ | = $xy$ |
| 1 | 57 | 77 | –4 | –2 | 16 | 8 |
| 2 | 58 | 78 | –3 | 1 | 9 | –3 |
| 3 | 59 | 75 | –2 | –4 | 4 | 8 |
| 4 | 60 | 81 | –1 | 2 | 1 | –2 |
| 5 | 62 | 82 | 1 | 3 | 1 | 3 |
| 6 | 65 | 79 | 4 | 0 | 16 | 0 |
| 7 | 66 | 81 | 5 | 2 | 25 | 10 |
| Total | 427 | 553 | 0 | 0 | 72 | 24 |

Here = = = 61, = = = 79

Thus

$b_{yx}$=

= = 0.333

Hence, the regression line of Y on X is

Y– = $b_{YX}$

or Y–79 = 0.333(X–61)

_ Y = 58.687+0.333X

## 12.7 SUMMARY

In this lesson, we have

(i) described the concept and meaning of regression analysis,

(ii) explained the difference between correlation and regression,

(iii) described the method of fitting of regression equation of Y on X, and,

(iv) also presented some numerical examples to illustrate the computational procedure of regression line of X on Y.

## 12.8 SELF ASSESSMENT EXERCISES

Q1 What is Regression Analysis? Write its uses?

______________________________________________

______________________________________________

______________________________________________

______________________________________________

______________________________________________

______________________________________________

Q2 Difference between Correlation and Regression analysis?

______________________________________________

Q3 Calculate Regression equation of X on Y by taking deviations of X series from 5 and of Y series from 7

| | | | | | |
|---|---|---|---|---|---|
| X | 6 | 2 | 10 | 4 | 8 |
| Y | 9 | 11 | 5 | 8 | 7 |

Ans 16.4-1.3y

Q4 calculate Regression equation of Y on X and estimate when X=55 from the following

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| X | 40 | 50 | 38 | 60 | 65 | 50 | 35 |
| Y | 38 | 60 | 55 | 70 | 60 | 48 | 30 |

Ans 942X+ 6.08

Or y= 57.89

## 12.9 SUGGESTED READINGS

1. Elehance : Advanced Statistics
2. S.P. Gupta : Statistics
3. Vishwanathan : Business Statistics : An Applied Orientation