QUANTITATIVE METHODS
FOR ECONOMIC ANALYSIS-II

BA ECONOMICS
2011 Admission onwards

IV Semester

CORE COURSE

UNIVERSITY OF CALICUT
SCHOOL OF DISTANCE EDUCATION
CALICUT UNIVERSITY.P.O., MALAPPURAM, KERALA, INDIA – 673 635

265
UNIVERSITY OF CALICUT
SCHOOL OF DISTANCE EDUCATION
STUDY MATERIAL

IV SEMESTER

B.A. ECONOMICS
(2011 ADMISSION)

CORE COURSE:

QUANTITATIVE METHODS FOR ECONOMIC ANALYSIS-II

Prepared by:

Module I:  Smt.Prajisha.P,
Assistant Professor,
PG Department of Economics,
Kozhikode – 673 580.

Module II:  Sri.Rahul.K,
Assistant Professor,
Department of Economics,
Govt. College Kalpetta,
Wayanad.

Module III:  Smt.Rasheeja.T.K,
ICSSR Research Associate,
P.G. Department of Economics,
Govt. College, Kodanchery,
Kozhikode – 673 580.

Module IV:  Dr.K.Rajan,
Associate Professor,
PG Department of Economics,
MD College, Pazhanji,
Thrissur.

Edited by:  Dr. C. Krishnan,
Associate Professor,
Govt. College, Kodenchery,
Kozhikode – 673 580.

Layout & Settings
Computer Section, SDE

© Reserved
## CONTENTS

<table>
<thead>
<tr>
<th>MODULE</th>
<th>TITLE</th>
<th>Page No</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODULE I</td>
<td>MEANING OF STATISTICS AND DESCRIPTION OF DATE</td>
<td>05-54</td>
</tr>
<tr>
<td>MODULE II</td>
<td>CORRELATION AND REGRESSION</td>
<td>55-75</td>
</tr>
<tr>
<td>MODULE III</td>
<td>INDEX NUMBER AND TIME SERIES ANALYSIS</td>
<td>76-95</td>
</tr>
<tr>
<td>MODULE IV</td>
<td>VITAL STATISTICS</td>
<td>96-117</td>
</tr>
</tbody>
</table>
MODULE I
MEANING OF STATISTICS AND DESCRIPTION OF DATE


Meaning of Statistics

The word statistics is understood by different people in different manner. For some people, it is tables, charts etc. For some it is a science. To other it is a method of studying quantitative information regarding some phenomena. Thus we see that generally the word is used in two different senses. In the first sense it refers to numerical facts called statistical data. In the second sense it refers to some theories, method, principles etc. In this sense, statistics is a body of methods known as statistical methods. The methods and techniques provided by statistics range from the most elementary devices which may be understood by even a laymen to the complicated mathematical procedures which can be understood only by experts.

Origin and growth of statistics

Statistics is a very old branch of knowledge. The world statistics seems to have been derived from the Latin word ‘Status’ or Italian word ‘Statista’ or German word ‘Statistik’, all mean political state. The origin of statistics was due to administrative requirements of the state. Administration of the state required the collection and analysis of data relating to population and material wealth of the country. One of the earliest Censuses of population and wealth was held in Egypt as early as 3050 BC, for the creation of pyramids. Although statistics originated as a science of kings, now it has emerged into a very important and useful science of all human beings. There is hardly any branch of human activity where the statistical methods are not made use of. Statistics has grown in to a very powerful science.

Definition of “Statistics”

Statistics has been defined either as a singular noun or a plural noun in various ways by different authors.

a) Definitions of statistics as a plural noun (or as numerical facts)

Statistics as a plural noun stands for numerical facts collected. Though different authors have defined the term statistics differently, the definition given by Horace Secrist is most exhaustive. According to Horace Secrist “Statistics are aggregates or fact affected to a marked extent by multiplicity of causes numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other”.
Characteristics of Statistics:-

As per the definition given by Horace Secrist, Statistics – should possess the following characteristics.

- Statistics should be aggregates of facts.
- They must be numerically expressed.
- They should be enumerated or estimated according to reasonable standard of accuracy.
- They should be collected in a systematic manner.
- They should be collected for a predetermined purpose.
- Statistics collected should facilitate comparison. So they must be homogeneous.

b) Definition of statistics as a singular noun (or as a method)

The word statistics as a singular noun stands for a body of methods known as statistical methods. Statistics is a method for obtaining and analyzing numerical facts and figures, in order to arrive at some decisions. In this sense, the word statistics has been defined by many statisticians. Of these definitions, the definition given by Seligman is short and simple and yet quite comprehensive.

According to Seligman “Statistics is the science which deals with the method of collecting, classifying, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry”. The definition points out the various statistical methods or features of science of statistics. They are collection of data, classification of data, presentation of data and analysis and interpretation of these presented data.

Porf.A.L.Bowley has given three definitions. At one place, he says “Statistics may be called the science of counting”. At another place Bowley says “Statistics may rightly be called the science of averages”. Still another definition given by the same author is “Statistics is the science of measurements of social organism, regarded as a whole in all its manifestations.

Boddington defines statistics as “the science of estimates and probabilities.

Functions of Statistics

The following are the important functions of statistics.

1. **It simplifies the complexity:** In our studies we collect huge facts and figures. They can’t be easily understood. Statistical methods make these large numbers of facts easily and readily understandable. In statistics there are methods like – graphs and diagrams, classifications, average etc which render complex data very simple.

2. **It presents fact in a definite Form:** One of the most important functions of statistics is to present general statements in a precise and definite form.

3. **It facilitates comparison:** Unless figures are comparing with other of the same kind, they are often devoid of any meaning. When we say that the price of a commodity has increased very much, the statement does not make the positions very clear. But when we say that last year price was ₹10 but now it is ₹11 the comparison becomes easy. Statistics provides a number of suitable methods like ratios, percentages, averages etc for comparison.
4. **It helps in formulating and testing of hypothesis:** Statistical methods are extremely helpful in formulating and testing hypothesis and to develop new theories.

5. **It helps in prediction:** Plans and policies of organizations are invariably formulated well in advance of the time of their implementation. Knowledge of future tendencies is very helpful in framing suitable policies and plans. Statistical methods provide helpful means of forecasting future events.

6. **Formulation of policies:** Statistics provide the basic material for framing suitable policies. The statistical tools like collection of data help much in this regard.

### Scope of statistics

In the early period of development of statistics, it had only limited scope. But in modern times the scope of statistics has become so wide that it includes all quantitative studies and analysis relating to any department of enquiry. Thus it is hardly possible to find a field where statistical methods are not used. The universality of statistics is enough to indicate its importance, utility and indispensability to the modern world. We shall discuss below the importance of statistics in various fields.

- **Statistics and business:**

  Statistics is an aid to business and commerce. When a person starts business, he enters into the profession of forecasting. Modern statistical devices have made business forecasting more precise and accurate. A business man needs statistics right from the time he proposes to start business. He should have relevant fact and figures to prepare the financial plan of the proposed business. Statistical methods are necessary for these purposes. In industrial concern statistical devices are being used not only to determine and control the quality of products manufactured by also to reduce wastage to a minimum. The technique of statistical control is used to maintain quality of products.

- **Statistics and Economics:**

  In the year 1890 Prof. Alfred Marshall, the famous economist observed that “statistics are the straw out of which I, like every other economist, have to make bricks”. This proves the significance of statistics in economics. Economics is concerned with production and distribution of wealth as well as with the complex institutional set-up connected with the consumption, saving and investment of income. Statistical data and statistical methods are of immense help in the proper understanding of the economic problems and in the formulation of economic policies. In fact these are the tools and appliances of an economist’s laboratory. In the field of economics it is almost impossible to find a problem which does not require an extensive uses of statistical data. As economic theory advances use of statistical methods also increase. The laws of economics like law of demand, law of supply etc can be considered true and established with the help of statistical methods. Statistics of consumption tells us about the relative strength of the desire of a section of people. Statistics of production describe the wealth of a nation. Exchange statistics through light on commercial development of a nation. Distribution statistics disclose the economic conditions of various classes of people. There for statistical methods are necessary for economics.
• **Statistics and Physical Science:**

The physical sciences, especially astronomy, geology and physics are among the fields in which statistical methods were first developed and applied, but until recently these sciences have not shared the 20th century development of statistics to the same extent as the biological and social science. Currently however the physical science seem to be making increasing use of statistics, especially in astronomy, chemistry, engineering, geology and certain branches of physics.

• **Statistics and Research:**

Statistics is an indispensable tool of research. Most of the advancement in knowledge has taken place because of experiments conducted with the help of statistical methods. For example, experiments about crop yield and different types of fertilizers and different types of soils of the growth of animals under different diets and environments are frequently designed and analysed according to statistical methods. Statistical methods are also useful for the research in medicine and public health. In fact there is hardly any research work today that one can find complete without statistical data and statistical methods.

• **Statistics and other uses**

Statistics are also useful to various institutions such as bankers, brokers, insurance companies, auditors, social workers, trade associations and chamber of commerce.

One must understand that statistics is not a dry, abstract and unrealistic pursuit followed by a small group of highly trained mathematicians, but rather a vitally important part of the economic and business life of the community.

**Limitations of Statistics**

- Statistic does not deal with individual items.
- Statistics is only one method of studying a problem.
- Statistics deals only with quantitative characteristics.
- Statistics is liable to be misuse.
- Statistics may mislead to wrong conclusion in the absence of details.

**Graphs of frequency distribution**

A frequency distribution can be represented graphically in any of the following ways. The most commonly used graphs and curves for representation a frequency distribution are

- Histogram
- Frequency Polygon
- Smoothened frequency curve
- Ogives or cumulative frequency curves.

**Histogram:**

A histogram is a set of vertical bars whose one as are proportional to the frequencies represented. While constructing histogram, the variable is always taken on the X axis and the frequencies on the Y axis. The width of the bars in the histogram will be proportional to the class interval. The bars are drawn without leaving space between them. A histogram generally represents a continuous curve. If the class intervals are uniform for a frequency distribution, then the width of all the bars will by equal.
Example:

<table>
<thead>
<tr>
<th>Marks</th>
<th>No. of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-15</td>
<td>5</td>
</tr>
<tr>
<td>15-20</td>
<td>20</td>
</tr>
<tr>
<td>20-25</td>
<td>47</td>
</tr>
<tr>
<td>25-30</td>
<td>38</td>
</tr>
<tr>
<td>30-35</td>
<td>10</td>
</tr>
</tbody>
</table>

Frequency Polygon (or line graphs)

Frequency Polygon is a graph of frequency distribution. There are two ways of constructing a frequency polygon.

a) Draw histogram of the data and then join by straight lines the mid points of upper horizontal sides of the bars. Join both ends of frequency polygon with x axis. Then we get frequency polygon.

b) Another method of constructing frequency polygon is to take the mid points of the various class intervals and then plot frequency corresponding to each point and to join all these points by a straight line. Here we have not to construct a histogram:

Example:

Draw a frequency polygon to the following frequency distribution

<table>
<thead>
<tr>
<th>Marks</th>
<th>No. of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-20</td>
<td>5</td>
</tr>
<tr>
<td>20-30</td>
<td>8</td>
</tr>
<tr>
<td>30-40</td>
<td>15</td>
</tr>
<tr>
<td>40-50</td>
<td>20</td>
</tr>
<tr>
<td>50-60</td>
<td>12</td>
</tr>
<tr>
<td>60-70</td>
<td>7</td>
</tr>
</tbody>
</table>
A continuous frequency distribution can be represented by a smoothed curve known as frequency curve. The mid values of classes are taken along the x axis and the frequencies along y axis. The points thus plotted are joined by smoothened curve. When the points of a frequency polygon are joined by free hand method curve and not by a straight line, we get frequency curve. The curve is drawn freehand in such a manner that the area included under the curve is approximately same as that of the frequency polygon. If the class intervals are not uniform, adjust they y co-ordinate so that the frequencies are proportional to the area of the rectangle contained by plotted points.

Example:

<table>
<thead>
<tr>
<th>Marks</th>
<th>No. of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-20</td>
<td>5</td>
</tr>
<tr>
<td>20-30</td>
<td>8</td>
</tr>
<tr>
<td>30-40</td>
<td>15</td>
</tr>
<tr>
<td>40-50</td>
<td>20</td>
</tr>
<tr>
<td>50-60</td>
<td>12</td>
</tr>
<tr>
<td>60-70</td>
<td>7</td>
</tr>
</tbody>
</table>
Difference between frequency polygon and frequency curve

Frequency polygon is drawn to frequency distribution of discrete or continuous nature. Frequency curves are drawn to continuous frequency distribution. Frequency polygon is obtained by joining the plotted points by straight lines. Frequency curves are smooth. They are obtained by joining plotted points by smooth curve.

Ogives (Cumulative frequency curve)

A frequency distribution when cumulated, we get cumulative frequency distribution. A series can be cumulated in two ways. One method is frequencies of all the preceding classes one added to the frequency of the classes. This series is called less than cumulative series. Another method is frequencies of succeeding classes are added to the frequency of a class. This is called more than cumulative series. Smoothed frequency curves drawn for these two cumulative series are called cumulative frequency curve or Ogives. Thus corresponding to the two cumulative series we get two ogive curves, known as less than ogive and more than ogive.

Less than ogive curve is obtained by plotting frequencies (cumulated) against the upper limits of class intervals. More than ogive curve is obtained by plotting cumulated frequencies against the lower limits of class intervals. Less than ogive is an increasing curve, sloping upwards from left to right. More than ogive is a decreasing curve and slopes from left to right.

Example:

Draw less than and more than cumulative frequency distribution for the following frequency distribution.

<table>
<thead>
<tr>
<th>Marks</th>
<th>No. of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-20</td>
<td>4</td>
</tr>
<tr>
<td>20-30</td>
<td>6</td>
</tr>
<tr>
<td>30-40</td>
<td>10</td>
</tr>
<tr>
<td>40-50</td>
<td>20</td>
</tr>
<tr>
<td>50-60</td>
<td>18</td>
</tr>
<tr>
<td>60-70</td>
<td>2</td>
</tr>
</tbody>
</table>

Cumulative frequency distribution:

<table>
<thead>
<tr>
<th>Marks less than</th>
<th>No. of Students</th>
<th>Marks More than</th>
<th>No. of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0</td>
<td>10</td>
<td>60</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>20</td>
<td>56</td>
</tr>
<tr>
<td>30</td>
<td>10</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>40</td>
<td>20</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>50</td>
<td>40</td>
<td>50</td>
<td>20</td>
</tr>
<tr>
<td>60</td>
<td>58</td>
<td>60</td>
<td>2</td>
</tr>
<tr>
<td>70</td>
<td>60</td>
<td>70</td>
<td>0</td>
</tr>
</tbody>
</table>
Pie diagrams are used when the aggregate and their division are to be shown together. The aggregate is shown by means of a circle and the division by the sectors of the circle. For example: to show the total expenditure of a government distributed over different departments like agriculture, irrigation, industry, transport etc. can be shown in a pie diagram. In constructing a pie diagram the various components are first expressed as a percentage and then the percentage is multiplied by 3.6, so we get angle for each component. Then the circle is divided into sectors such that angles of the components and angles of the sectors are equal. Therefore one sector represents one component. Usually components are with the angles in descending order are shown.

Example: Draw pie diagram to represents the distribution of the certain blood group ‘O’ among Gypsies, Indians and Hungarians.

Answer

<table>
<thead>
<tr>
<th>Race</th>
<th>%</th>
<th>Angle (% x 3.6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gypsies</td>
<td>34.3</td>
<td>123.48</td>
</tr>
<tr>
<td>Indians</td>
<td>31.3</td>
<td>112.68</td>
</tr>
<tr>
<td>Hungarians</td>
<td>34.4</td>
<td>123.84</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>360</td>
</tr>
</tbody>
</table>
Measures of Central Tendency

One of the most important objectives of statistical analysis is to get one single value that describes the characteristics of the entire mass of unwieldy data. Such a value is called central value or an average or the expected value of variable. Clark defines “Average is an attempt to find one single figure to describe whole of figures”. It is clear from this definition that an average is a single value that represents the group of values. Such a value is of great significance because it depicts the characteristics of the whole group. Since an average represents the entire data, its value lies somewhere in between the two extremes. For this reason an average is frequently referred to as a measure of central tendency.

Objective of the study of averages

There are two main objectives of the study of averages.

a) **To get a single value that describes the characteristics of the entire group:** Average enables us to get a birds-eye view of the entire date. 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 take the average, we will get one single value that represents the entire population. Such a figure throws lights on the standard of living of an average Indian.

b) **To facilitate comparison:** Measure of central value, by reducing the mass of data to one single figure, enables comparison to be made. Comparison can be made either at a point of time or over a period of time. However while making comparisons one should also take into consideration the multiplicity of forces that might be affecting the data. For Example, if per capita income is rising in absolute terms from one period to another, it should not lead one to think that the standard of living is necessarily improving because the prices might be rising faster than the rise in per capita income and so in real terms people might be worse off. Moreover, the same measure should be used for making comparison between two or more groups. For example, we should not compare the mean wage of one factory with the median wage of another factory for drawing any inference about the wage levels.
Requisites of a good average

Since an average is a single value representing a group of values, it is desired that such a value satisfies the following properties.

1. **Easy to understand**: Since statistical methods are designed to simplify the complexities.
2. **Simple to compute**: A good average should be easy to compute so that it can be used widely. However, though case of computation is desirable, it should not be sought at the expense of other averages. Ie, if in the interest of greater accuracy, use of more difficult average is desirable.
3. **Based on all items**: The average should depend upon each and every item of the series, so that if any of the items is dropped, the average itself is altered.
4. **Not unduly affected by Extreme observations**: Although each and every item should influence the value of the average, none of the items should influence it unduly. If one or two very small or very large items unduly affect the average, ie, either increase its value or reduce its value, the average can’t be really typical of entire series. In other words, extremes may distort the average and reduce its usefulness.
5. **Rigidly defined**: An average should be properly defined so that it has only one interpretation. It should preferably be defined by algebraic formula so that if different people compute the average from the same figures they all get the same answer. The average should not depend upon the personal prejudice and bias of the investigator, otherwise results can be misleading.
6. **Capable of further algebraic treatment**: We should prefer to have an average that could be used for further statistical computation so that its utility is enhanced. For example, if we are given the data about the average income and number of employees of two or more factories, we should able to compute the combined average.
7. **Sampling stability**: Last, but not least we should prefer to get a value which has what the statisticians called “sampling stability”. This means that if we pick 10 different groups of college students, and compute the average of each group, we should expect to get approximately the same value. It does not mean, however that there can be no difference in the value of different samples. There may be some differences but those samples in which this difference is less that are considered better than those in which the difference is more.

Types of Averages

The important types of averages are:

- Arithmetic mean
- 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.
I. Arithmetic Mean

It is the most common type and widely used measure of central tendency. Arithmetic mean of a series is the figure obtained by dividing the total value of the various items by their number. There are two types of arithmetic mean.

1. Simple arithmetic mean
2. Weighted arithmetic mean

1. **Simple arithmetic average** – Individual observation

   a. **Direct Method**: steps
      - Add up all the values of the variables \( X \) and find out \( \sum X \)
      - Divide \( \sum X \) by their number of observations \( N \)

\[
\overline{X} = \frac{\sum X}{N} = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N}
\]

   b. **Short cut method**: The arithmetic mean can also be calculated by short cut method. This method reduces the amount of calculation. It involves the following steps

   i. Assume any one value as an assumed mean, which is also known as working mean or arbitrary average (A).
   ii. Find out the difference of each value from the assumed mean \( (d = X - A) \).
   iii. Add all the deviations \( (\sum d) \)
   iv. Apply the formula

\[
\overline{X} = A + \frac{\sum d}{N}
\]

Where \( \overline{X} \rightarrow \text{Mean}, \frac{\sum d}{N} \rightarrow \text{Sum of deviation from assumed mean, } A \rightarrow \text{Assumed mean}

Example: Calculate arithmetic mean

<table>
<thead>
<tr>
<th>Roll No</th>
<th>Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>55</td>
</tr>
<tr>
<td>4</td>
<td>78</td>
</tr>
<tr>
<td>5</td>
<td>58</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Roll Nos.</th>
<th>Marks</th>
<th>( d = X - 55 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40</td>
<td>-15</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>-5</td>
</tr>
<tr>
<td>3</td>
<td>55</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>78</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>58</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>5</td>
</tr>
</tbody>
</table>

\[ \sum d = 11 \]
Calculation of arithmetic mean - Discrete series

To find out the total items in discrete series, frequency of each value is multiplies with the respective size. The value so obtained is totaled up. This total is then divided by the total number of frequencies to obtain arithmetic mean.

Steps

1. Multiply each size of the item by its frequency \( fX \)
2. Add all \( fX \) i.e. (\( \sum fX \))
3. Divide \( \sum fX \) by total frequency \( N \).

The formula is

\[
\bar{X} = A + \frac{\sum d}{N}
\]

Example

<table>
<thead>
<tr>
<th>( X )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f )</td>
<td>10</td>
<td>12</td>
<td>8</td>
<td>7</td>
<td>11</td>
</tr>
</tbody>
</table>

Solution

<table>
<thead>
<tr>
<th>( X )</th>
<th>( f )</th>
<th>( fX )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>24</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>55</td>
</tr>
</tbody>
</table>

\[ N = \sum fX = 141 \]

\[ \bar{X} = \frac{\sum fX}{N} = \frac{141}{4.8} \]

\[ \bar{X} = 29.3 \]

Short cut Method

Steps:

- Take the value of assumed mean (\( A \))
- Find out deviations of each variable from \( A \) i.e \( d \)
- Multiply \( d \) with respective frequencies \( (fd) \)
- Add up the product \( (\sum fd) \)
- Apply formula

\[
\bar{X} = A \pm \frac{\sum fd}{N}
\]
Continuous series

In continuous frequency distribution, the value of each individual frequency distribution is unknown. Therefore an assumption is made to make them precise or on the assumption that the frequency of the class intervals is concentrated at the center that the mid-point of each class intervals has to be found out. In continuous frequency distribution, the mean can be calculated by any of the following methods.

a. Direct method
b. Short cut method
c. Step deviation method

a. Direct Method

Steps:
1. Find out the mid value of each group or class. The mid value is obtained by adding the lower and upper limit of the class and dividing the total by two. (symbol = m)
2. Multiply the mid value of each class by the frequency of the class. In other words m will be multiplied by f.
3. Add up all the products - $\sum fm$
4. $\sum fm$ is divided by N

Example:

From the following find out the mean profit

<table>
<thead>
<tr>
<th>Profit/Shop:</th>
<th>100-200</th>
<th>200-300</th>
<th>300-400</th>
<th>400-500</th>
<th>500-600</th>
<th>600-700</th>
<th>700-800</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of shops:</td>
<td>10</td>
<td>18</td>
<td>20</td>
<td>26</td>
<td>30</td>
<td>28</td>
<td>18</td>
</tr>
</tbody>
</table>

Solution

<table>
<thead>
<tr>
<th>Profit (x)</th>
<th>Mid point - m</th>
<th>No of Shops (f)</th>
<th>fm</th>
</tr>
</thead>
<tbody>
<tr>
<td>100-200</td>
<td>150</td>
<td>10</td>
<td>1500</td>
</tr>
<tr>
<td>200-300</td>
<td>250</td>
<td>18</td>
<td>4500</td>
</tr>
<tr>
<td>300-400</td>
<td>350</td>
<td>20</td>
<td>7000</td>
</tr>
<tr>
<td>400-500</td>
<td>450</td>
<td>26</td>
<td>11700</td>
</tr>
<tr>
<td>500-600</td>
<td>550</td>
<td>30</td>
<td>16500</td>
</tr>
<tr>
<td>600-700</td>
<td>650</td>
<td>28</td>
<td>18200</td>
</tr>
<tr>
<td>700-800</td>
<td>750</td>
<td>18</td>
<td>13500</td>
</tr>
</tbody>
</table>

\[\sum f = 150 \quad \sum fm = 72900\]

\[\bar{X} = \frac{\sum fd}{N} = \frac{72900}{150} = 486\]
b) **Short cut method**

Steps:
1. Find the mid value of each class or group (m)
2. Assume any one of the mid value as an average (A)
3. Find out the deviations of the mid value of each from the assumed mean (d)
4. Multiply the deviations of each class by its frequency (fd).
5. Add up the product of step 4: \( \sum fd \)
6. Apply formula

\[
\bar{X} = A + \frac{\sum fd}{N}
\]

Example: (solving the last example)

Solving: Calculation of Mean

<table>
<thead>
<tr>
<th>Profit ($)</th>
<th>M</th>
<th>d = m - 450</th>
<th>f</th>
<th>fd</th>
</tr>
</thead>
<tbody>
<tr>
<td>100-200</td>
<td>150</td>
<td>-300</td>
<td>10</td>
<td>-3000</td>
</tr>
<tr>
<td>200-300</td>
<td>250</td>
<td>-200</td>
<td>18</td>
<td>-3600</td>
</tr>
<tr>
<td>300-400</td>
<td>350</td>
<td>-100</td>
<td>20</td>
<td>-2000</td>
</tr>
<tr>
<td>400-500</td>
<td>450</td>
<td>0</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>500-600</td>
<td>550</td>
<td>100</td>
<td>30</td>
<td>3000</td>
</tr>
<tr>
<td>600-700</td>
<td>650</td>
<td>200</td>
<td>28</td>
<td>5600</td>
</tr>
<tr>
<td>700-800</td>
<td>750</td>
<td>300</td>
<td>18</td>
<td>5400</td>
</tr>
<tr>
<td>( \sum f = 150 )</td>
<td>( \sum fd = 5400 )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\bar{X} = A + \frac{\sum fd}{N} = 450 + \frac{5400}{150} = 486
\]

c) **Step deviation method**

The short cut method discussed above is further simplified or calculations are reduced to a great extent by adopting step deviation methods.

Steps:
1. Find out the mid value of each class or group (m)
2. Assume any one of the mid value as an average (A)
3. Find out the deviations of the mid value of each from the assumed mean (d)
4. Deviations are divided by a common factor (d')
5. Multiply the d' of each class by its frequency (fd')
6. Add up the products \( \sum fd' \)
7. Then apply the formula

\[
\bar{X} = A + \frac{\sum fd'}{N} \times c
\]

Where \( c \) = Common factor
Example:

Calculate mean for the last problem

Solution

<table>
<thead>
<tr>
<th>Profit</th>
<th>m</th>
<th>f</th>
<th>d</th>
<th>d'</th>
<th>f d'</th>
</tr>
</thead>
<tbody>
<tr>
<td>100-200</td>
<td>150</td>
<td>10</td>
<td>-300</td>
<td>-3</td>
<td>-30</td>
</tr>
<tr>
<td>200-300</td>
<td>250</td>
<td>18</td>
<td>-200</td>
<td>-2</td>
<td>-36</td>
</tr>
<tr>
<td>300-400</td>
<td>350</td>
<td>20</td>
<td>-100</td>
<td>-1</td>
<td>-20</td>
</tr>
<tr>
<td>400-500</td>
<td>450</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>500-600</td>
<td>550</td>
<td>30</td>
<td>100</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>600-700</td>
<td>650</td>
<td>28</td>
<td>200</td>
<td>2</td>
<td>56</td>
</tr>
<tr>
<td>700-800</td>
<td>750</td>
<td>18</td>
<td>300</td>
<td>3</td>
<td>54</td>
</tr>
</tbody>
</table>

\[
\sum f = 150 \\
\sum f d' = 540
\]

\[
\bar{X} = A + \frac{\sum fd'}{N} \times c
\]

\[
450 + \frac{540}{150} \times 100
\]

\[
450 + (0.36 \times 100) = 486
\]

**Merits of Arithmetic mean**

1. It is easy to understand and easy to calculate
2. It is used in further calculation
3. It is rigidly defined
4. It is based on value of every item in the series
5. It is provide a good basis for comparison
6. Its formula is rigidly defined.
7. The mean is more stable measure

**Demerits**

1. The mean is unduly affected by extreme items
2. It is unrealistic
3. It may lead to a false conclusion
4. It can’t be accurately determined even if one of the value is not known
5. It is not use full for the study of qualities like intelligence, honesty etc.
6. It can’t be located by observation or the graphic method.

**Weighted Mean**

Simple arithmetic mean gives equal importance to all items. Sometimes the items in a series may not have equal importance. So the simple arithmetic mean is not suitable for those series and weighted average will be appropriate.
Weighted means are obtained by taking into account these weights (or importance). Each value is multiplied by its weight and sum of these products is divided by the total weight to get weighted mean.

Weighted average often gives a fair measure of central tendency. In many cases it is better to have weighted average than a simple average. It is invariably used in the following circumstances.

1. When the importance of all items in a series is not equal, we would associate weights to the items.
2. For comparing the average of one group with the average of another group, when the frequencies in the two groups are different, weighted averages are used.
3. When relations, percentages and rates are to be averaged, weighted averages is used.
4. It is also used in the calculations of birth and death rate index number etc.
5. When average of a number of series is to be found out together weighted average is used.

Formula: Let \( x_1 + x_2 + x_3 + \ldots + x_n \) be in values with corresponding weights \( w_1 + w_2 + w_3 + \ldots + w_n \). Then the weighted average is

\[
\text{Weighted Average} = \frac{\sum w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}{\sum w_1 + w_2 + \ldots + w_n}
\]

### II Median

Median is the value of item that goes to divide the series into equal parts. Median may be defined as the value of that item which divides the series into two equal parts. Arranging the data is necessary to compute median. As distinct from the arithmetic mean, which is calculated from the value of every item in the series, the median is what is called a positional average. The term position refers to the place of value in a series.

**Calculation of median:** - Individual observations

**Steps:**

- Arrange data in ascending or descending order of magnitude.
- In a group composed of an odd number of values such as 7, add to the total number of values and divide by 2 \( \left[ \frac{7+1}{2} \right] \). It gives the answer 4, the number of value starting at either end of the numerically arranged groups will be the median value. In the form of formula.

\[
\text{Median} = \text{Size of } \left[ \frac{N+1}{2} \right]^{th} \text{item.}
\]

**Example**

Find out the median from the following items

<table>
<thead>
<tr>
<th>X</th>
<th>10</th>
<th>15</th>
<th>9</th>
<th>25</th>
<th>19</th>
</tr>
</thead>
</table>


<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Size of the item ascending order</th>
<th>Descending order (X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>15 ← Median → 15</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>9</td>
</tr>
</tbody>
</table>

Median = Size of the $\frac{N+1}{2}^{th}$ item

$= $ Size of the $\frac{5+1}{2}^{th}$ item

$= 3^{rd}$ item $= 15$

**Calculation of Median: Discrete series**

Steps:
- Arrange the data in ascending or descending order
- Find cumulative frequencies
- Apply the formula

\[
\text{Median} = \text{Size of } \left[ \frac{N+1}{2} \right]^{th} \text{ item}
\]

Example: Calculate median from the following

<table>
<thead>
<tr>
<th>Size of shoes</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>5.5</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>28</td>
</tr>
<tr>
<td>6.5</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
</tr>
<tr>
<td>7.5</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>34</td>
</tr>
</tbody>
</table>

Solution

<table>
<thead>
<tr>
<th>Size</th>
<th>f</th>
<th>Cumulative f (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>5.5</td>
<td>16</td>
<td>26</td>
</tr>
<tr>
<td>6</td>
<td>28</td>
<td>54</td>
</tr>
<tr>
<td>6.5</td>
<td>15</td>
<td>69</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>99</td>
</tr>
<tr>
<td>7.5</td>
<td>40</td>
<td>139</td>
</tr>
<tr>
<td>8</td>
<td>34</td>
<td>173</td>
</tr>
</tbody>
</table>

\[
\text{Median} = \text{Size of } \left[ \frac{N+1}{2} \right]^{th} \text{ item}
\]
\[ N = 173 \]

\[ \text{Median} = \frac{173 + 1}{2} = 87^{\text{th}} \text{ item} = 7 \]

\[ \text{Median} = 7 \]

**Calculation of median** – Continuous series

Steps:
- Find out the median by using \( N/2 \)
- Find out the class which median lies
- Apply the formula

\[ L + \frac{N - c f}{f} \times i \]

Where
- \( L \) = lower limit of the median class
- \( f \) = frequency of the median class
- \( i \) = class interval
- \( c f \) = cumulative frequency of the proceeding median class

Example: Calculate median from the following data

<table>
<thead>
<tr>
<th>Marks</th>
<th>Frequency</th>
<th>Cumulative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-25</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>25-40</td>
<td>20</td>
<td>26</td>
</tr>
<tr>
<td>40-55</td>
<td>44</td>
<td>70</td>
</tr>
<tr>
<td>55-70</td>
<td>26</td>
<td>96</td>
</tr>
<tr>
<td>70-85</td>
<td>3</td>
<td>99</td>
</tr>
<tr>
<td>85-100</td>
<td>1</td>
<td>100</td>
</tr>
</tbody>
</table>

\[ \text{Median item} = \frac{N}{2} = \frac{100}{2} = 50 \]

It lies in 40-55 marks group. So median is
Median = \[ L + \frac{\frac{N}{2} - cf}{f} \times i \]
\[ = 40 + \frac{50 - 26}{44} \times 15 \]
\[ = 40 + 8.18 \]
\[ = 48.18 \text{ marks} \]

**Merit of median:**
1. It is easy to understand and easy to compute
2. It is quite rigidly defined
3. It eliminates the effect of extreme values.
4. It is amenable to further algebraic treatment
5. The value is generally lies in a distribution
6. Since it is a positional average, median can be computed even if the item at the extremes are unknown.
7. Helpful for open end classes.

**Demerits:**
1. It ignores the extreme value
2. It is more affected by sampling fluctuations than mean.
3. For calculating median it is necessary to arrange the data; other averages do not need any arrangement.
4. It is not capable for further algebraic treatment.

### III. Mode

Mode is the common item of a series. It is the value which occurs the greatest number of frequency in a series. It is derived from the French word “La mode” meaning fashion. Mode is the most fashionable or typical value of a distribution, because it is repeated the highest number of times in the series.

**Calculation of Mode:**

Mode can be often be found out by mere inspection in case of individual observations. The data have to be arranged in the form of array so that the value which has the highest frequency can be known for example to persons has the following income.

\[ \text{₹} \ 850, 750, 600, 825, 850, 725, 600, 850, 640, 530 \]

Here 850 repeated three times; therefore the mode salary is \text{₹}850.

In certain cases that there may not be a mode or there may be more than one mode.

For example:

a) 40, 44, 45, 48, 52 (No mode)

b) 45, 55, 25, 28, 32, 55, 45 (bi modal mode is i) 45 ii) 55
When we calculate the mode from data, if there is only one mode in the series, it is called unimodal. If there are two modes, it is called bimodal; if there are three modes, it is called trimodal and if there are more than three modes it is called multimodal.

**Calculation of mode: discrete series**

In discrete series the value having highest frequency is taken as Mode. A glance at a series can reveal which is the highest frequency. So we get mode by mere inspection. So this method is also called inspection method.

Example:

Find the mode from the following data

<table>
<thead>
<tr>
<th>Size</th>
<th>5</th>
<th>15</th>
<th>16</th>
<th>25</th>
<th>37</th>
<th>45</th>
<th>56</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>10</td>
<td>16</td>
<td>28</td>
<td>15</td>
<td>30</td>
<td>40</td>
<td>38</td>
</tr>
</tbody>
</table>

Ans: the value 45 has highest frequency

**Grouping and Analysis method**

In the case of certain series, there may be more than one highest frequency. In the case of some other series, frequency may not be increasing and decreasing in a systematic manner. In these cases inspection method may not be a suitable method. Therefore, we apply a method is called grouping and analysis method.

Example:

Find the mode from the following data

<table>
<thead>
<tr>
<th>Size</th>
<th>3</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>2</td>
<td>7</td>
<td>15</td>
<td>27</td>
<td>12</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Ans:

<table>
<thead>
<tr>
<th>size</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td>27</td>
<td>42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td>12</td>
<td></td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>43</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

Ans:
The highest number in the last column is 6. It refers to the size 12.

**Preparation of grouping and analysis table**

In column 1 → given frequencies are shown

In column 2 → frequencies added in twos starting from the top. (ie 2 + 7 = 9, 15 + 27 = 42, 12 + 4 = 16, 3 + 2 = 5)

In column 3 → frequencies added in twos leaving the first (7 + 15 = 22, 27 + 12 = 39, 4 + 3 = 7)

In column 4 → frequencies added in threes starting from the top. (2 + 7 + 15 = 24, 27 + 12 + 4 = 19)

In column 5 → frequencies added in threes, leaving the first frequency. (7 + 15 + 27 = 49, 12 + 4 + 3 = 19)

In column 6 → frequencies added in threes, leaving the first and the second (15 + 27 + 12 = 54, 4 + 3 + 2 = 9)

**Calculation of Mode: Continuous Series**

Step 1: By preparing grouping table and analysis table or by inspection ascertain in the modal class.

Step 2: Determine the value of mode by applying the following formula

\[ \hat{M}_o = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i \]

Where \( L \) = Lower limit of the modal class, \( \Delta_1 \) = difference between the frequency of the modal class and the frequency of the pre-modal class; ie, preceding class (ignoring signs); \( \Delta_2 \) = difference between the frequency of the modal class and the frequency of the post modal class, ie; succeeding class (ignoring signs); \( i \) = the size of the class interval of the modal class.

Another form of this formula is

<table>
<thead>
<tr>
<th>size</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
\[ M_0 = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i \]

Where \( L \) = Lower limit of the modal class;
\( f_1 \) = frequency of the modal class
\( f_0 \) = frequency of the class preceding the modal class.
\( f_2 \) = frequency of the class succeeding the modal class.

While applying the above formula for calculating mode it is necessary to see that the class intervals are uniform throughout. If they are unequal they should first be made equal on the assumption that the frequencies are equally distributed throughout the class, otherwise we will get misleading results.

**Merits of Mode**
1. It is easy to understand as well as easy to calculate. In certain cases, it can be found out by inspection.
2. It is usually an actual value as it occurs most frequently in the series.
3. It is not affected by extreme values in the average
4. It is simple and precise
5. It is most representative average.
6. The value of mode can be determined by graphic method.

**Demerits of mode**
1. It is suitable for further mathematical treatment
2. It may not give weight to extreme items.
3. It is difficult to compute; when there are both positive and negative items in a series and when there are one or more items is zero.
4. It is stable only when sample is large
5. It will not give the aggregate value as in average.

**IV Geometric Mean:**
Geometric mean is defined as the \( n^{th} \) root of the product of \( N \) items of series. If there are two items, take the square root; if there are three items, we take the cube root; and so on. Symbolically;

\[ GM = \sqrt[n]{(X_1)(X_2) \ldots \ldots (X)_n} \]

Where \( X_1 \), \( X_2 \) \ldots \ldots \( X_n \) are referring to the various items of the series.

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. \( GM \) then is calculated as follows.
log G.M = \frac{\log X_1 + \log X_2 + \ldots + \log X_N}{N}

G.M. = \frac{\sum \log X}{N}

G.M. = \text{Antilog} \left[ \frac{\sum \log X}{N} \right]

In discrete series GM = \text{Antilog} \left[ \frac{\sum f \log X}{N} \right]

In continuous series GM = \text{Antilog} \left[ \frac{\sum f \log m}{N} \right]

\text{Where } f = \text{frequency}

\text{M = mid-point}

\textbf{Merits of G.M}

1. It is based on each and every item of the series.
2. It is rigidly defined.
3. It is useful in averaging ratios and percentages and in determining rates of increase and decrease.
4. It is capable of algebraic manipulation.

\textbf{Limitations}

1. It is difficult to understand
2. It is difficult to compute and to interpret
3. It can’t be computed when there are negative and positive values in a series or one or more of values are zero.
4. G.M has very limited applications.

\textbf{V - Harmonic Mean}

The Harmonic Mean is based on the reciprocals of the numbers averaged. It is defined as the reciprocal of the arithmetic mean of the reciprocal of the individual observation. i.e;

H.M = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + \ldots + \frac{1}{X_n}}

When the number of items is large the computation of H.M in the above manner becomes tedious. To simplify calculations we obtain reciprocals of the various items from the tables and apply the following formula:

In individual observations, H.M = \frac{N}{\sum \frac{1}{X}}

In discrete series, H.M = \frac{N}{\sum f \frac{1}{X}}

In continuous series, H.M = \frac{N}{\sum f \frac{1}{m}} = \frac{N}{\sum \frac{f}{m}}
Merits of Harmonic mean:

1. Its value is based on every item of the series.
2. It lends itself to algebraic manipulation.

Limitations

1. It is not easily understood
2. It is difficult to compute
3. It gives large weight to smallest item.

Relationship among the averages:

In any distribution when the original items differ in size, the value of AM, GM and HM would also differ and will be in the following order.

\[ A.M \geq G.M \geq H.M \]

Example

Calculate Geometric mean from the following data

<table>
<thead>
<tr>
<th>Size</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
</tr>
</tbody>
</table>

Solution

<table>
<thead>
<tr>
<th>x</th>
<th>f</th>
<th>log x</th>
<th>fx log x</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2</td>
<td>0.6990</td>
<td>1.3980</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>0.9031</td>
<td>2.7093</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>1.0006</td>
<td>4.0000</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1.0792</td>
<td>1.0792</td>
</tr>
</tbody>
</table>

\[ G.M = \text{Antilog} \left[ \frac{\sum fx \log x}{N} \right] \]
\[ = \text{Antilog} \left[ \frac{9.1865}{10} \right] = 8.292 \]

Example:

Calculate Harmonic Mean of the following values.

<table>
<thead>
<tr>
<th>Size</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
</tr>
<tr>
<td>14</td>
<td>30</td>
</tr>
<tr>
<td>18</td>
<td>10</td>
</tr>
</tbody>
</table>
Solution

<table>
<thead>
<tr>
<th>x</th>
<th>f</th>
<th>log x</th>
<th>fx log x</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>20</td>
<td>0.1667</td>
<td>3.334</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
<td>0.1000</td>
<td>4.000</td>
</tr>
<tr>
<td>11</td>
<td>30</td>
<td>0.0714</td>
<td>2.142</td>
</tr>
<tr>
<td>18</td>
<td>10</td>
<td>1.0556</td>
<td>0.556</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>10.032</strong></td>
</tr>
</tbody>
</table>

\[ \text{H.M} = \frac{\sum x^2}{\sum f} \]
\[ = \frac{30}{1.006} = 29.82 \]

**Quadratic Mean: (Root mean Square)**

Quadratic mean of a set of values is defined as the square root of the mean of the squares of those values. It is useful when some values are negative and others are positive because in such cases the Arithmetic mean is not a good representative.

If \( X_1, X_2 \ldots X_n \) are observations, then,

\[ \text{QM} = \sqrt{\frac{x_1^2 + x_2^2 + x_3^2 + \ldots + x_n^2}{n}} \]

Example:

Find the Quadratic mean of the following items 15, 20, 27, 35 and 40.

\[ \text{QM} = \sqrt{\frac{15^2 + 20^2 + 27^2 + 35^2 + 40^2}{5}} = 28.91 \]

**Progressive Averages.**

Progressive average is calculated with the help of simple arithmetic average. It is a cumulative average. In the calculation of this average, the figures of all previous years are added and no figure is left out. Thus the progressive average of the second year would be equal to the arithmetic average of the figures of the first two years. The progressive average of the third year would be equal to the arithmetic average of the figure of the first three years and so on.

Eg: Calculate the progressive average of the following data.

<table>
<thead>
<tr>
<th>Year:</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales (in crores):</td>
<td>15</td>
<td>20</td>
<td>23</td>
<td>22</td>
<td>20</td>
</tr>
</tbody>
</table>
Progressive average is used by business houses with a view to compare the current profit with those of past.

**Positional Values – Quartiles, deciles and percentiles**

Median divides the distribution into two equal parts. There are other values also which divide the series into a number of equal parts and they are called partition values or positional value. There are 4 types of positional values. They are median, quartiles, deciles and percentile.

1. **Quartiles**

A measure while divides an array into four equal parts is known as quartiles. Each portion contains equal number of items. The first, second and third points are termed as first quartile (Q₁), second quartile (Q₂) or median and third quartile (Q₃). The first quartile (Q₁) or lower quartile, has 25% of the items of the distribution below it and 75% of the items are greater than it. Q₂ (median has 50% of the observations above it and 50% of the observations below it. The upper quartile (Q₃), has 75% of the items of the distribution below it and 25% of the items are above it.

**Calculation of quartiles – (Individual and Discrete series)**

The method for locating the quartiles is the same as that for median. The following steps may be noted

1. Find out the cumulative frequency
2. Then apply the formula.

First Quartile Q₁ = Size of \( \left(\frac{N+1}{4}\right)^{th} \) item.

Third quartile Q₃ = Size of \( 3 \left(\frac{N+1}{4}\right)^{th} \) item.

**Continuous series**

\[
Q₁ = L₁ + \frac{N - cf}{f} \times i \\
Q₃ = L + \frac{3N - cf}{f} \times i
\]

where:
- \( L₁ \) → Lower limit of the class
- \( N \) → Total number of observations
- \( cf \) → Cumulative frequency
- \( f \) → Class intervals
- \( i \) → Class intervals
2. Deciles

Deciles are the values which divide the series in to ten parts. We get nine dividing positions to be called deciles. There are nine quartiles. \(D_5\) is the median. To compute deciles in the individual and discrete series method is the same i.e

\[D_2 = \text{Size of } \left(\frac{2N}{10}\right)^{\text{th}} \text{ item}\]

To compute the deciles in the continuous series, the following formula is applied, say \(D_2\)

\[D_2 = L_1 + \frac{2N}{10} \times i\]

3. Percentile

Percentile value divides the distribution into 100 parts. We get 99 dividing positions. Eg. \(P_1, P_2, P_3\) etc. To compute the percentile in individual and discrete series the method is same. For instance,

\[P_{25} = \text{Size of } \left(\frac{25N}{100}\right)^{\text{th}} \text{ item}\]

\[P_{25} = Q; \ P_{50} = \text{median}; \ P_{75} = Q_3\] etc.

In continuous series, to estimate percentile the formula is

\[P_{25} = L_1 + \frac{25N}{100} \times i\]

Example:

Find lower quartile \((Q_1)\), median, deciles, seven and sixtieth percentile for the following data.

<table>
<thead>
<tr>
<th>Wages</th>
<th>f</th>
<th>cf</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-20</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>20-30</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>30-40</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>40-50</td>
<td>21</td>
<td>36</td>
</tr>
<tr>
<td>50-60</td>
<td>43</td>
<td>79</td>
</tr>
<tr>
<td>60-70</td>
<td>32</td>
<td>111</td>
</tr>
<tr>
<td>70-80</td>
<td>9</td>
<td>120</td>
</tr>
</tbody>
</table>
\[ Q_1 = \text{Size of } \left( \frac{N}{4} \right)^{th} \text{ item} \]
\[ = \frac{120}{4^{th}} \text{ item} = 30^{th} \text{ item} \]

\[ Q_1 \text{ lies in the class 40-50} \]
\[ Q_1 = L_1 + \frac{\frac{N}{4} - cf}{f} \times i \]
\[ = 40 + \frac{30-15}{21} \times 10 \]
\[ = 47.14 \]

\[ \text{Median } = \text{Size of } \left( \frac{N}{2} \right)^{th} \text{ item } = \frac{120}{2} = 60^{th} \text{ item} \]

\[ \text{Median lies in the class 50-60} \]
\[ \text{Median } = L_1 + \frac{\frac{N}{2} - cf}{f} \times i \]
\[ = 50 + \frac{60-36}{43} \times 10 \]
\[ = 55.58 \]

\[ D_7 = \text{Size of } \frac{7N}{10}^{th} \text{ item } = 7 \times \frac{120}{10}^{th} \text{ item } = 84^{th} \text{ item} \]

\[ D_7 \text{ lies in the class 60-70} \]
\[ D_7 = L_1 + \frac{\frac{7N}{10} - cf}{f} \times i \]
\[ = 60 + \frac{84-79}{32} \times 10 = 61.56 \]

\[ P_{60} = \text{Size of } \frac{60N}{100}^{th} \text{ item } = 60 \times \frac{120}{100}^{th} \text{ item} \]

\[ = \text{Size of } 72^{nd} \text{ item } : P_{60} \text{ lies in the class 50-60} \]
\[ P_{60} = L_1 + \frac{\frac{60N}{100} - cf}{f} \times i \]
\[ = 50 + \frac{72-36}{43} \times 10 \]
\[ = 58.37 \]
MEASURES OF DISPERSION

Various measures of central tendency give us one single figure that represents the entire data. But the average alone can’t adequately describe a set of observations, unless all the observations are the same. It is necessary to describe the variability or dispersion of the observations. A.L. Bowley defines “dispersion is a measure of variation of the items”. It is clear from this definition that dispersion (also known as scatter, spread or variation) measures the extent to which the items vary with the central value. Since measures of dispersion give an average of the differences of various item from an average, they are also called average of the ‘second order’.

Significance of measuring variation

Measures of dispersion are needed for four basic purposes

1. To determine the reliability of an average.
2. To serve as a basis for the control of the variability.
3. To compare two or more series with regard to their variability.
4. To facilitate the use of other statistical measures.

Properties of a good measure of variation

A good measure of dispersion should possess as far as possible the following properties.

a. It should be simple to understand and easy to compute.
b. It should be rigidly defined.
c. It should be based on each and every item of distribution
d. It should amenable to further algebraic treatment.
e. It should have sampling stability.
f. It should not be unduly affected by extreme items.

Methods of studying variation/dispersion

The following are the important methods of studying variation.

1. The Range
2. The interquartile Range and quartile deviation.
3. Mean deviation or average deviation
4. Standard deviation
5. Lorenz curve.

Of these the first two, namely the range and quartile deviations are positional measures, because they depend on the values at a particular position in the distribution. The average deviation and standard deviation are called calculation measures of deviation because all of the values are employed in their calculation and the Lorenz curve is a graphic method.

Absolute and relative measures of variation: Measures of dispersion may be either absolute or relative. Absolute measures of dispersion are expressed in same statistical unit in which the original data are given such as rupees, kilograms, tones etc. These values may be used to compare the variations in two distributions provided the variables are expressed in the same unit and same average size. In the case of two sets of data which are expressed in different units, however such as quintals of sugar versus tons of sugarcane, the absolute measures of dispersion are not comparable. Here the relative and measure of dispersion should be used.
A measure of relative dispersion is the ratio of a measure of absolute dispersion to an appropriate average. It is sometimes called co-efficient of dispersion, because, “coefficient” is a pure number that is independent of unit of measurement. It should be remembered that while computing the relative dispersion the average used as base should be the same one from which the absolute deviations are measured. This means that the arithmetic mean should be used with the standard deviation and either median or arithmetic mean with the mean deviation.

1. Range

Range is the simplest method of studying dispersion. It is defined as the difference between the value of smallest item (S) and the value of largest item (L) included in the distribution.

\[
\text{Range} = L - S
\]

The relative measure corresponding to range, called coefficient of range is obtain by applying the following formula.

\[
\text{Coefficient of range} = \frac{L - S}{L + S}
\]

Eg: The following are the marks obtained by ram in five different subjects

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Maths</th>
<th>History</th>
<th>Science</th>
<th>English</th>
<th>Economics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marks</td>
<td>83</td>
<td>76</td>
<td>92</td>
<td>70</td>
<td>80</td>
</tr>
</tbody>
</table>

Ans:

Range = L – S = 92 – 70 = 22

Coefficient of range = \[
\frac{22}{92 + 70} = 0.135
\]

Merits and Limitations

**Merits**

- Amongst all the methods of studying dispersion, range is the simplest to understand easiest to compute.
- It takes minimum time to calculate the value of range hence if one is interested in getting a quick rather than very accurate picture of variability one may compute range.

**Limitation**

- Range is not based on each and every item of the distribution.
- It is subject to fluctuation of considerable magnitude from sample to sample.
- Range can’t tell us anything about the character of the distribution with the two.
- According to kind “Range is too indefinite to be used as a practical measure of dispersion

**Uses of Range**

- Range is useful in studying the variations in the prices of stocks, shares and other commodities that are sensitive to price changes from one period to another period.
- The meteorological department uses the range for weather forecasts since public is interested to know the limits within which the temperature is likely to vary on a particular day.
II. INTER–QUARTILE RANGE OR QUARTILE DEVIATION

By eliminating the lowest 25% and the highest 25% of items in a series, we are left with the central 50% which are ordinarily free of extreme values. To obtain a measure of variation, we use the distance between the first and the third quartiles – i.e., semi inter quartile range. Inter quartile range is computed by deducting the value of the first quartile from the value of third quartile.

Symbolically;
Interquartile range  = \( Q_3 - Q_1 \)

Semi interquartile range or quartile deviation is defined as half of the distance between the third and first quarter.

Symbolically,
Quartile Deviation = \( \frac{Q_3 - Q_1}{2} \)

Quartile Deviation is an absolute measure of dispersion. The relative measure of dispersion, known as coefficient of quartile deviation, is calculated as follows.

Coefficient of Q.D. = \( \frac{Q_3 - Q_1}{Q_3 + Q_1} \)

Quartile deviation is an improved measure over the range, as it is not calculated from extreme items, but on quartiles. For symmetrical distribution we have

\[ \text{Median} + \text{Q. D} = Q_3 \]

Eg: 2:

Calculate the quartile deviation and its coefficient of as monthly earnings for a year.

<table>
<thead>
<tr>
<th>Months</th>
<th>Months earnings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>239</td>
</tr>
<tr>
<td>2</td>
<td>250</td>
</tr>
<tr>
<td>3</td>
<td>251</td>
</tr>
<tr>
<td>4</td>
<td>251</td>
</tr>
<tr>
<td>5</td>
<td>257</td>
</tr>
<tr>
<td>6</td>
<td>258</td>
</tr>
<tr>
<td>7</td>
<td>260</td>
</tr>
<tr>
<td>8</td>
<td>261</td>
</tr>
<tr>
<td>9</td>
<td>262</td>
</tr>
<tr>
<td>10</td>
<td>262</td>
</tr>
<tr>
<td>11</td>
<td>273</td>
</tr>
<tr>
<td>12</td>
<td>275</td>
</tr>
</tbody>
</table>
Solution:

\[ Q_1 = \text{earnings of} \left( \frac{N+1}{4} \right)^{th} \text{item} \]
\[ = \frac{12+1}{4} \times \frac{13}{4} = 3.25 \text{ month} \]
\[ \text{i.e.}, = \text{Rs. 251} \]

\[ Q_3 = \text{earnings of} \left( \frac{3(N+1)}{4} \right)^{th} \text{item} \]
\[ \text{i.e., 9.75 months = Rs: 262} \]

Quartile Deviation (Q.D) = \[ \frac{Q_3 - Q_1}{2} \]
\[ = \frac{262 - 251}{2} = \]
\[ = \frac{11}{2} = 5.5 \text{ Rupees.} \]

Quartile Coefficient of dispersion = \[ \frac{Q_3 - Q_1}{Q_3 + Q_1} \]
\[ = \frac{262 - 251}{262 + 251} = \]
\[ = \frac{11}{513} = 0.0214 \]

Example 3:

Calculate the range and Quartile deviation of wages.

<table>
<thead>
<tr>
<th>Wages ( ₹)</th>
<th>Labourers</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 – 32</td>
<td>12</td>
</tr>
<tr>
<td>32 – 34</td>
<td>18</td>
</tr>
<tr>
<td>34 – 36</td>
<td>16</td>
</tr>
<tr>
<td>36 – 38</td>
<td>14</td>
</tr>
<tr>
<td>38 – 40</td>
<td>12</td>
</tr>
<tr>
<td>40 – 42</td>
<td>8</td>
</tr>
<tr>
<td>42 - 44</td>
<td>6</td>
</tr>
</tbody>
</table>

Solution

Range : = L – S

Calculation of Quartiles:
\[ Q_1 = \text{Size of } \left( \frac{N}{4} \right)^{th} \text{ item} \]
\[ = \frac{96}{4} = 21.5 \]

i.e. Q. lies in the group 32 – 34
\[ Q_1 = L + \frac{N - c.f}{f} \times i \]
\[ = 32 + \frac{96 - 12}{18} \times 2 \]
\[ = 32 + \frac{84}{18} \]
\[ = 32 + 4.67 \]
\[ = 36.67 \]

\[ Q_3 = \text{Size of } \left( \frac{3N}{4} \right)^{th} \text{ item} \]
\[ = 3 \times \frac{96}{4} = 64.5^{th} \text{ item} \]

Q.3 lies in the group 38 – 40
\[ Q_3 = L + \frac{3N - c.f}{f} \times i \]
\[ = 38 + \frac{192 - 60}{12} \times 2 \]
\[ = 38 + 13 \]
\[ = 51 \]

Q.D = \[ \frac{Q_3 - Q_1}{2} \]
\[ = \frac{51 - 36.67}{2} \]
\[ = \frac{14.33}{2} \]
\[ = 7.165 \]
Coefficient of Q.D. = \( \frac{Q_3 - Q_1}{Q_3 + Q_1} \)

= \( \frac{38.75 - 33.6}{38.75 + 33.06} \)

= \( \frac{5.69}{71.81} \)

= 0.08

Merits of Quartile Deviation

1. It is simple to understand and easy to calculate.
2. It is not influenced by extreme values.
3. It can be found out with open end distribution.
4. It is not affected by the presence of extreme values.

Demerits

1. It ignores the first 25% of the items and the last 25% of the items.
2. It is a positional average: hence not amenable to further mathematical treatment.
3. The value is affected by sampling fluctuations.

III. MEAN DEVIATION OR AVERAGE DEVIATION

The two methods of dispersion discussed above, namely range and quartile deviation, are not based on all observations. They are positional measures of dispersion. They do not show any scatter of the observations from an average. The mean deviation and standard deviations are based on all observations.

Mean deviation (M.D) is the average difference between the items in a distribution and the median or mean of that series. Theoretically there is an advantage in taking the deviations from median because, the sum of deviations from the items from median is minimum when signs are ignored. However, in practice, the arithmetic mean is more frequently used in calculating the value of an average deviation and this is the reason why it is more commonly called mean deviation.

Computation of M.D – Individual Observations

If \( X_1, X_2, X_3, \ldots \ldots \ldots, X_n \) are \( N \) given observations then the deviations about an average \( A \) is given by:

\[
M.D. = \frac{1}{N} \sum |X - A|
\]

\[
M.D. = \frac{\sum |D|}{N}
\]

Where \( |D| = |X - A| \)

Steps:

- Compute median of the series
- The deviations of items from median ignoring \( + \) signs and denote these by \( |D| \).
- Obtain the total of these observations ie., \( \sum |D| \).
- Divide the total obtained by the number of observations ie., \( \sum |D| \).
The relative measure corresponding to the M.D called the coefficient of mean deviation; is obtained by dividing mean deviation by particular average used in computing mean deviation. Thus, if mean deviation has been computed from median, the coefficient of M.D shall be obtained by dividing MD by median

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

If MD is obtained from mean, the coefficient of mean deviation will be;

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

Example:
Calculate M.D. from mean and median for the following data; and also calculate coefficient of M.D. 100, 150, 200, 250, 360, 490, 500, 600, 671

Solution:

\[
\begin{array}{ccc}
X & |D| = |X - \bar{X}| & |D| = X - \text{Median} \\
100 & 269 & 260 \\
150 & 219 & 210 \\
200 & 169 & 160 \\
250 & 119 & 110 \\
360 & 9 & 0 \\
490 & 121 & 130 \\
500 & 131 & 140 \\
600 & 231 & 240 \\
671 & 302 & 311 \\
\end{array}
\]

\[
\Sigma X = 3321 \quad \Sigma |D| = 1570 \quad \Sigma |D| = 1561
\]

Mean

\[
\bar{X} = \frac{\Sigma X}{N} = \frac{3321}{9} = 369
\]

\[
\bar{X} = 369
\]

M.D. from mean

\[
= \frac{\Sigma |D|}{N} = \frac{1570}{9} = 174.44
\]

Coefficient of MD = \( \frac{\text{M.D}}{\bar{X}} \)

\[
\frac{174.44}{369} = 0.47
\]

Median

\[
\text{Median} = \text{value of } \left[ \frac{\text{The } \frac{9+1}{2} \text{ item}}{\frac{5}{2}} \right] \text{th item} = 360
\]

M.D from median

\[
= \frac{\Sigma |D|}{N} = \frac{1561}{9} = 173.44
\]

Coefficient of MD = \( \frac{\text{M.D}}{\text{Median}} \)

\[
\frac{173.44}{9} = 0.48
\]
Calculation of M.D – Discrete series

In discrete series the formula of calculating mean deviation is:

\[ M.D = \frac{\sum f|D|}{N} \]

Steps:

- Calculate mean or median
- Take the deviations of the item from mean or median ignoring signs and denoted they by \( |D| \).
- Multiply these deviations by respective frequencies and obtain the total \( \sum f|D| \).
- Divide this total by number of observations.

This gives us the value of mean deviation.

Example:

Calculate mean deviation from the following series

\[
\begin{array}{cccccc}
  x & 10 & 11 & 12 & 13 & 14 \\
  f & 3 & 12 & 18 & 12 & 3 \\
\end{array}
\]

Solution:

\[ M.D = \frac{\sum f|D|}{N} \]

Median = Size of \( \left[ \frac{N+1}{2} \right]^{th} \) item = \( \frac{48+1}{2} = 24.5 \)

So median is = 12

\[
\begin{array}{cccc}
  X & f & |D| & f|D| & c.f \\
  |X - 12| & & & & \\
\end{array}
\]

\[
\begin{array}{cccccc}
  10 & 3 & 2 & 6 & 3 \\
  11 & 12 & 1 & 12 & 15 \\
  12 & 18 & 0 & 0 & 33 \\
  13 & 12 & 1 & 12 & 45 \\
  14 & 3 & 2 & 6 & 48 \\
\end{array}
\]

\[ N = 48 \quad \sum |D| = 36 \]

\[ M.D = \frac{\sum f|D|}{N} = \frac{36}{48} = 0.75 \]
Calculation of M.D – Continuous Series

For calculation mean deviation, in continuous series, the procedure remains same as discussed above. The only difference is that we have to obtain the mid-point of various classes and take deviations of these points from mean or median. The formula is same:

\[
M.D = \frac{\sum f |D|}{N}
\]

**Example:**
Calculate M.D from mean for the following data

<table>
<thead>
<tr>
<th>Class :</th>
<th>2-4</th>
<th>4-6</th>
<th>6-8</th>
<th>8-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>f:</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

| Class | Mid value m | Frequency | fm | \(|D|= X- \bar{X}\) | \(f|D|\) |
|-------|-------------|-----------|----|-------------------|------|
| 2-4   | 3           | 3         | 9  | 2.2               | 6.6  |
| 4-6   | 5           | 4         | 20 | 0.2               | 0.8  |
| 6-8   | 7           | 2         | 14 | 1.8               | 3.6  |
| 8-10  | 9           | 1         | 9  | 3.8               | 3.8  |

\[\bar{X} = \frac{\sum fm}{N} = \frac{52}{10} = 5.2\]

\[M.D = \frac{\sum f |D|}{N} = \frac{148}{10} = 14.8\]

**Merits of M.D:**

i. It is simple to understand and easy to compute.

ii. It is not much affected by the fluctuations of sampling.

iii. It is based on all items of the series and gives weight according to their size.

iv. It is less affected by extreme items.

v. It is rigidly defined.

vi. It is a better measure for comparison.

**Demerits**

i. It is a non-algebraic treatment

ii. Algebraic positive and negative signs are ignored. It is mathematically unsound and illogical.

iii. It is not as popular as standard deviation.
Uses:

It will help to understand the standard deviation. It is useful in marketing problems. It is used in statistical analysis of economic, business and social phenomena. It is useful in calculating the distribution of wealth in a community or nation.

IV. STANDARD DEVIATION (σ)

The concept, standard deviation was introduced by Karl Pearson in 1893. It is the most important measure of dispersion and is widely used in many statistical formulas. It is also known as root mean square deviation, for the reason that it is the square root of the mean of the squared deviations from arithmetic mean.

Standard deviation measures the absolute dispersion (or variability of the distribution). The greater the standard deviation, the greater will be magnitude of deviation.

Calculation of standard deviation

Individual Observations – In the case of individual observations, standard deviation may be computed by applying any of the following two methods.

1. By taking deviation of the items from actual mean.
2. By taking deviations from an assumed mean.

Deviations taken from actual mean:

When deviations are taken from actual mean the following formula is applied.

\[ \sigma = \sqrt{\frac{\sum X^2}{N}} \]

Where; \( x = (X - \bar{X}) \)

Steps:

- Calculate the actual mean of the series ie. \( \bar{X} \)
- Take the deviations of the items from mean ie. \( (X - \bar{X}) \) denoted by \( x \).
- Square these \( x \) and obtain \( \sum x^2 \)
- Divide the \( \sum x^2 \)by \( N \) and extract the square root. This gives the value of standard deviation.

Deviations taken from an assumed mean: This method is adopted when the arithmetic mean is a fraction value. Take deviations from fractional value would be a very difficult task. To save time we apply short cut method; deviation taken from assumed mean.

\[ \sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2} \]

Where \( d \) stands for deviation from assumed mean (X-A)
Steps:

- Assume any one of the item in the series as an average (A).
- Find out the deviations from assumed mean $X - A = d$
- Find out total of the deviations i.e. $\Sigma d$
- Square the deviations i.e., $d^2$ and add up the squares of deviations, i.e. $\Sigma d^2$
- Then substitute the values in the following formula

$$\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}$$

Example:

Calculate the standard deviation from the data given below

$240, 260, 290, 245, 255, 288, 272, 263, 277, 251$

Solution:

<table>
<thead>
<tr>
<th>$X$</th>
<th>$(X-264) = d$</th>
<th>$d^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>240</td>
<td>$-4$</td>
<td>576</td>
</tr>
<tr>
<td>260</td>
<td>$-4$</td>
<td>16</td>
</tr>
<tr>
<td>290</td>
<td>$+26$</td>
<td>676</td>
</tr>
<tr>
<td>245</td>
<td>$-19$</td>
<td>361</td>
</tr>
<tr>
<td>255</td>
<td>$-9$</td>
<td>81</td>
</tr>
<tr>
<td>288</td>
<td>$+24$</td>
<td>576</td>
</tr>
<tr>
<td>272</td>
<td>$+8$</td>
<td>64</td>
</tr>
<tr>
<td>263</td>
<td>$-1$</td>
<td>1</td>
</tr>
<tr>
<td>277</td>
<td>$+13$</td>
<td>169</td>
</tr>
<tr>
<td>251</td>
<td>$-13$</td>
<td>169</td>
</tr>
</tbody>
</table>

$\Sigma X = 2641 \quad \Sigma d = 0 \quad \Sigma d^2 = 2689$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{2641}{10} = 264.1$$

Since it is a fraction we take 264 as assumed mean.

$$\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}$$

$$= \sqrt{\frac{2689}{10} - \left(\frac{0}{10}\right)^2}$$

$$= \sqrt{268.9 - 0.01}$$

$$= 16.398$$
Calculation of Standard Deviation: - Discrete Series

There are three methods for calculating standard deviation in a discrete series.

a. Actual mean method
b. Assumed mean method
c. Step deviation method

a. Actual Mean Method

Steps:
- Calculate the mean of the series.
- Find deviations for various items from mean, i.e., \( X - \bar{X} \)
- Square the deviations \( (d^2) \) and multiply with respective frequencies \( (f) \), we get \( f d^2 \).
- Total the product \( \sum f d^2 \). Then apply the formula

\[
\sigma = \sqrt{\frac{\sum f d^2}{N}}
\]

b. Assumed Mean Method: Here deviations are taken

Steps:
1. Assume anyone of the items in the series as an average and this is called assumed mean.
2. Find out the deviations from assumed mean \( X - A \) and denoted by \( d \).
3. Multiply these deviations by respective frequencies and get \( \sum f d \).
4. Square the deviations \( (d^2) \).
5. Multiply the squared deviations \( (d^2) \) by respective frequencies \( (f) \) and get \( \sum f d^2 \).
6. Substitute the values in the following formula

\[
\sigma = \sqrt{\frac{\sum f d^2}{N} - \left(\frac{\sum f d}{N}\right)^2}
\]

where \( d = X - A \)

Example: Calculate standard deviation

Marks: 10 20 30 40 50 60
Students: 8 12 20 10 7 3

Solution

<table>
<thead>
<tr>
<th>Marks</th>
<th>( f )</th>
<th>( d = X - 30 )</th>
<th>( fd )</th>
<th>( f d^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8</td>
<td>-2( \hat{e} )</td>
<td>-16( \hat{e} )</td>
<td>3200</td>
</tr>
<tr>
<td>20</td>
<td>12</td>
<td>-1( \hat{e} )</td>
<td>-12( \hat{e} )</td>
<td>1200</td>
</tr>
<tr>
<td>30</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>40</td>
<td>10</td>
<td>10</td>
<td>100</td>
<td>1000</td>
</tr>
<tr>
<td>50</td>
<td>7</td>
<td>20</td>
<td>140</td>
<td>2800</td>
</tr>
<tr>
<td>60</td>
<td>3</td>
<td>30</td>
<td>90</td>
<td>2700</td>
</tr>
</tbody>
</table>

\( N = 60 \) \n\( \sum f d = 50 \) \n\( \sum f d^2 = 10900 \)
Standard Deviation = \( \sqrt{\frac{\sum fd^2}{N} - \left( \frac{\sum fd}{N} \right)^2} \)

\[ = \sqrt{\frac{10900}{60} - \left( \frac{50}{60} \right)^2} \]

\[ = \sqrt{181.67 - 0.69} \]

\[ = 13.45 \]

**C. Step deviation Method:**

Here we take a common factor for all item of the series. In this method the calculation becomes easy and simple. The formula for this is

\[ \sigma = \sqrt{\frac{\sum fd'^2}{N} - \left( \frac{\sum fd'}{N} \right)^2} \times C \]

where \( d' = \frac{x-A}{C} \) and \( C = \) Common factor or Class interval

**Example:**

Calculate Standard deviation

<table>
<thead>
<tr>
<th>Marks</th>
<th>f</th>
<th>( x-30 )</th>
<th>fd'</th>
<th>fd'^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8</td>
<td>-2</td>
<td>-16</td>
<td>32</td>
</tr>
<tr>
<td>20</td>
<td>12</td>
<td>-1</td>
<td>-12</td>
<td>12</td>
</tr>
<tr>
<td>30</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>40</td>
<td>10</td>
<td>1</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>50</td>
<td>7</td>
<td>2</td>
<td>14</td>
<td>28</td>
</tr>
<tr>
<td>60</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>27</td>
</tr>
</tbody>
</table>

\( N = 60; \quad \sum fd' = 5; \quad \sum fd'^2 = 109 \)

\( \sum fd'^2 = 109; \quad \sum fd' = 5; \quad N = 60; \quad C = 10 \)

\[ \sigma = \sqrt{\frac{\sum fd'^2}{N} - \left( \frac{\sum fd'}{N} \right)^2} \times 10 \]

\[ = \sqrt{\frac{109}{60} - \left( \frac{5}{60} \right)^2} \times 10 \]

\[ = \sqrt{1.817 - 0.0069} \times 10 \]

\[ = \sqrt{1.345} \times 10 \]

\[ = 13.45 \]
Calculation of standard deviation – Continuous series

In the continuous series the method of calculating standard deviation is almost same as in a discrete series. But here, the mid values of class intervals are to be found out. The step deviation method is widely used.

\[ \sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2 \times C} \]

where, \( C = \text{Class interval} \)

\[ \begin{align*}
  d' &= \frac{M - A}{C} \\
  m &= \text{mid value}
\end{align*} \]

Steps

1. Find out the mid points of various classes.
2. Assume one of the mid values as an average and denote it by A.
3. Find out deviations of each mid-value from the assumed average and denote by d.
4. If the class intervals are equal, then take the common factor, and find out \( d' = \frac{X - A}{C} \).
5. Multiply these \( d' \) by respective frequencies and get \( \sum fd' \).
6. Square the deviations and get \( \sum f d' \)².
7. Multiply \( d' \)² by respective frequencies and obtain \( \sum fd' \)².
8. Substitute values in the formula.

Example

Calculate standard deviation of the following.

<table>
<thead>
<tr>
<th>Marks</th>
<th>0-10</th>
<th>10-20</th>
<th>20-30</th>
<th>30-40</th>
<th>40-50</th>
<th>50-60</th>
<th>60-70</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Students</td>
<td>5</td>
<td>12</td>
<td>30</td>
<td>45</td>
<td>50</td>
<td>37</td>
<td>21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Marks</th>
<th>Mid, ( 'm' )</th>
<th>( M - A )</th>
<th>( \frac{X - A}{C} )</th>
<th>( fd' )</th>
<th>( fd' )²</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10</td>
<td>5</td>
<td>0</td>
<td>( \frac{x-35}{10} )</td>
<td>-15</td>
<td>45</td>
</tr>
<tr>
<td>10-20</td>
<td>15</td>
<td>-2</td>
<td>( \frac{x-35}{10} )</td>
<td>-24</td>
<td>48</td>
</tr>
<tr>
<td>20-30</td>
<td>25</td>
<td>-1</td>
<td>( \frac{x-35}{10} )</td>
<td>-30</td>
<td>30</td>
</tr>
<tr>
<td>30-40</td>
<td>35</td>
<td>0</td>
<td>( \frac{x-35}{10} )</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>40-50</td>
<td>45</td>
<td>1</td>
<td>( \frac{x-35}{10} )</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>50-60</td>
<td>55</td>
<td>2</td>
<td>( \frac{x-35}{10} )</td>
<td>74</td>
<td>148</td>
</tr>
<tr>
<td>60-70</td>
<td>65</td>
<td>3</td>
<td>( \frac{x-35}{10} )</td>
<td>60</td>
<td>189</td>
</tr>
</tbody>
</table>

N=200 \hspace{1cm} \sum fd'=118 \hspace{1cm} \sum fd'^2=510
\[ \sigma = \sqrt{\frac{\sum f d'^2}{N} - \left(\frac{\sum f d'}{N}\right)^2} \times C \]

\[ = \sqrt{\frac{510}{200} - \left(\frac{118}{200}\right)^2} \times 10 \]
\[ = \sqrt{2.55 - 3.481} \times 10 \]
\[ = 1.4839 \times 10 \]
\[ = 14.839 \]

**Coefficient of Variation**

Standard deviation is the absolute measure of dispersion. It is expressed in terms of the units in which the original figures are collected and stated. The relative measure of standard deviation is known as coefficient of variation.

Variance: Square of Standard deviation

Symbolically:

\[ \text{Variance} = \sigma^2 \]

\[ \sigma = \sqrt{\text{Variance}} \]

Coefficient of standard deviation = \( \frac{\sigma}{\bar{x}} \times 100 \)

In the above formula, coefficient of standard deviation will be in fraction and as such not very good for comparison. Therefore, the coefficient of standard deviation is multiplied by 100 gives coefficient of variation.

Coefficient of variation (c.v) = \( \frac{\sigma}{\bar{x}} \times 100 \)

**Example:**

Particulars regarding income of two villages are given below.

<table>
<thead>
<tr>
<th></th>
<th>Village A</th>
<th>Village B</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of people</td>
<td>600</td>
<td>500</td>
</tr>
<tr>
<td>Average income</td>
<td>175</td>
<td>186</td>
</tr>
<tr>
<td>Variance of income</td>
<td>100</td>
<td>81</td>
</tr>
</tbody>
</table>

In which village, the variation in income is greater?

**Solution:**

Village A’s variation in income distribution

\[ \text{C.V} = \frac{\sigma}{\bar{x}} \times 100 \]
\[ = \frac{100}{175} \times 100 \]
\[ = 5.7 \]
Village B’s variation in income distribution

\[ \text{C.V} = \frac{\sigma}{\bar{X}} \times 100 \]

\[ = \frac{\sqrt{81}}{186} \times 100 \]

\[ = 4.84 \]

Village A has greater variation than B

**Merits of Standard Deviation**

1. It is rigidly defined and its value is always definite and based on all observation.
2. As it is based on arithmetic mean, it has all the merits of arithmetic mean.
3. It is possible for further algebraic treatment.
4. It is less affected by sampling fluctuations.

**Demerits**

1. It is not easy to calculate.
2. It gives more weight to extreme values, because the values are squared up.

**V. Graphic Method of Dispersion**

- **LORENZ CURVE**

A graphic method of showing dispersion is adopted by Dr. Max O. Lorenz, a famous economic statistician, who studied the distribution of wealth and the curve used by him came to be known as Lorenz curve. Lorenz curve is a device used to show the measurement of economic inequalities as the distribution of income, and wealth. Lorenz curve can also be used for the study of distribution of profit, wage etc.

The following are the method for constructing Lorenz Curve.

1. The size of the item and their frequencies are to be cumulated.
2. Percentage must be calculated for each cumulating value of the size and frequency of items.
3. Plot the percentage of the cumulated values of the variable against the percentage of the corresponding cumulated frequencies. Join these points with as smooth free hand curve. This curve is called Lorenz curve.
4. Zero percentage on the X axis must be joined with 100% on Y axis. This line is called the line of equal distribution.
The greater the distance between the curve and the line of equal distribution, the greater the dispersion. If the Lorenz curve is nearer to the line of equal distribution, the dispersion or variation is smaller.

**Uses of Lorenz Curve**
1. To study the variability in a distribution.
2. To compare the variability relating to a phenomenon for two regions.
3. To study the changes in variability over a period.

**Disadvantages of Lorenz Curve**
1. A disadvantage of Lorenz Curve is that it gives only a relative idea of the dispersion as compared with the line of equal distribution. It doesn’t give a numerical value of the variability for the given distribution.
2. When there are two Lorenz curves drawn for two data, sometimes the two curves cross and recross each other and thus making it difficult for comparative purpose to say which Lorenz curve represents greater inequality.

**Gini’s Coefficient**
Gini has devised a concentration ratio based on his mean difference measure. Co-efficient of concentration is obtained by dividing the mean difference by twice the arithmetic mean.
\[ G = \frac{\text{Mean difference}}{2\overline{X}} \]

This value lies between zero and one. In equal distribution its value is zero with the increase in the inequality the value of coefficient goes up.

**SKEWNESS AND KURTOSIS**

**Skewness:**
A distribution which is not symmetrical is called a skewed distribution and in such distributions, the mean the median and the mode will not coincide, but the values are pulled apart. If the curve has a longer tail towards right, it is said to be positively skewed. If the curve has a longer tail towards the left, it is said to be negatively skewed. The following illustration will clarify;

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

- The value of mean median and mode do not coincide.
- When the data are plotted on a graph, they do not give the normal bell-shaped form.
- The sum of positive deviation from the median is not equal to the sum of the negative deviations.
- Quartiles are not equi distant from the median.
- Frequencies are not distributed at points of equal deviation from the mode.

**Measures of Skewness**
The measures of asymmetry are usually called measure of skewness. They may be either absolute or relative.

**Absolute measure of Skewness:**
Skewness can be measured in absolute terms by taking the difference between mean and mode.

Absolute skewness = \( \overline{X} - \text{mode} \)

If the value of the mean is greater than mode, the skewness is positive

If the value of mode is greater than mean, the skewness is negative

The greater the amount of skewness (negative or positive), more the tendency towards asymmetry. The absolute measure of skewness will be proper measure for comparison, and hence, in each series a relative measure or coefficient of skewness has to be computed.

**Relative measure of skewness**
There are three important measures of relative skewness.

1. Karl Pearson’s coefficient of skewness.
2. Bowley’s coefficient of skewness.
3. Kelly’s coefficient of skewness.
I. Karl Pearson’s Coefficient of skewness:

According to him, absolute skewness = Mean – Mode. This measure is not suitable for making valid comparison of skewness in two or more distributions, because

a) Unit of measurement may be different in different series.

b) The same size of skewness has different significance with small or large variation in two series.

Therefore, to avoid the difficulties, an absolute measure is adopted. This is done by dividing the difference between the mean and mode by standard deviation. The resultant coefficient is called Pearsonian Coefficient of skewness ($Skp$)

$$Skp = \frac{\bar{X} - Mode}{\sigma}$$

In case mode is ill defined, the coefficient can be determined by changed the formula:

$$Skp = 3 \frac{\bar{X} - Mode}{\sigma}$$

$$= 3 \frac{\Sigma d - \Sigma d}{\sigma}$$

Example:

Calculate Karl Pearson’s Co-efficient of skewness for the following data.
25, 15, 23, 40, 27, 25, 23, 25, 20

<table>
<thead>
<tr>
<th>Size</th>
<th>Deviation from A = 25(d)</th>
<th>$d^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>-10</td>
<td>100</td>
</tr>
<tr>
<td>23</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>40</td>
<td>15</td>
<td>225</td>
</tr>
<tr>
<td>27</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>23</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>-5</td>
<td>25</td>
</tr>
</tbody>
</table>

$\Sigma d = -2$ \hspace{1cm} $\Sigma d^2 = 362$

$$\text{Mean} = A + \frac{\Sigma d}{N}$$

$$= 25 - \frac{2}{9}$$

$$\text{Standard Deviation}$$

$$\text{SD} = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}$$
\[ = 25 - 0.22 \]
\[ = 24.78 \]
Mode = 25

Karl Pearson’s Coefficient of skewness
\[ \text{Skp} = \frac{\text{Mean} - \text{Mode}}{\sigma} \]
\[ = \frac{24.78 - 25}{6.3} \]
\[ = -0.03 \]

II. Bowley’s Coefficient of skewness:

In the above method of measuring skewness, the whole of the series is needed. Prof. Bowley has suggested a formula based on relative position of quartiles. In a symmetrical distribution, the quartiles are equidistant from the value of the mean, i.e., median - \( Q_1 = Q_3 - \) median. This means that the value of median is the mean of \( Q_1 \) & \( Q_2 \). But in a skewed distribution, the quartiles will not be equidistant from the median. Hence Bowley has suggested the following formula.

Absolute Sk = \( (Q_3 - \text{median}) - (\text{median} - Q_1) \)
\[ = Q_3 + Q_1 - 2 \text{ median} \]

Coefficient of Sk = \( \frac{Q_3 + Q_1 - 2 \text{ median}}{Q_3 - Q_1} \)

Example:
Calculate coefficient of skewness of the following frequency distribution:

<table>
<thead>
<tr>
<th>No. of children/family</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of families</td>
<td>7</td>
<td>10</td>
<td>16</td>
<td>25</td>
<td>18</td>
<td>11</td>
<td>8</td>
</tr>
</tbody>
</table>

Solution

<table>
<thead>
<tr>
<th>No. of children (x)</th>
<th>No. of families (f)</th>
<th>c.f</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>33</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>58</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>76</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>87</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>95</td>
</tr>
</tbody>
</table>
II. Kelly’s Coefficient of skewness

Bowley’s measure neglects the two extreme quartiles. To measure the skewness, it would be better to consider the entire data or the more extreme items. His measure of skewness is based on two deciles, i.e., 10th and 90th percentiles or first or ninth deciles. Symbolically,

\[ Sk_B = \frac{Q_3 + Q_1 - 2 \times \text{median}}{Q_3 - Q_1} \]

So, \( Q_1 = 2 \)

\[ Q_3 = 3 \left[ \frac{N + 1}{4} \right]^{th \text{item}} = \left[ \frac{96}{4} \right]^{th \text{item}} = 72^{th \text{item}} \]

So, \( Q_3 = 4 \)

Median = Size of \( \left[ \frac{N + 1}{2} \right]^{th \text{item}} \)

\[ = \frac{96}{2} = 48^{th \text{item}} \]

Size of 48th is 3. Hence median = 3

\[ Sk_B = \frac{4 + 2 - 2 \times (3)}{4 - 2} = 0 \]

III. Kurtosis

The expression Kurtosis is used to describe the peakedness of a curve. The three measures – central tendency, dispersion and skewness describe the characteristics of frequency distributions. But these studies will not give a clean picture of the characteristics of a distribution.

As far as the measurement of shape is concerned, we have two characteristics – skewness which refers to the asymmetry of a series and Kurtosis which measures the peakedness of a normal curve.
Measures of Kurtosis tell us the extent to which a distribution is more peaked or more flat tapped than the normal curve. A normal curve which is symmetrical and bell shaped is designated as mesokurtic because it is kurtic at the centre. If a curve is more flat than normal curve is platy kurtic. If a curve is relatively more narrow and peaked at the top it is designated as leptokurtic.

**Measures of Kurtosis**

The measure of Kurtosis is based upon the forth moment about the mean of the distribution. Symbolically

\[ \beta_2 = \frac{M_4}{M_2^2} \]

- \( M_4 \) = 4\text{th moment}
- \( M_2 \) = 2\text{nd moment}

If \( \beta_2 = 3 \), the distribution is said to be normal. (ie mesokurtic)
If \( \beta_2 > 3 \), the distribution is more peaked to curve is leptokurtic.
If \( \beta_2 < 3 \), the distribution is said to be flat topped and the curve is platy kurtic.
MODULE II
CORRELATION AND REGRESSION

CORRELATION-MEANING, TYPES AND DEGREES OF CORRELATION- METHODS OF MEASURING CORRELATION- GRAPhICAL METHODS: SCATTER DIAGRAM AND CORRELATION GRAPH; ALGEBRAIC METHODS: KARL PEARSON’S COEFFICIENT OF CORRELATION AND RANK CORRELATION COEFFICIENT - PROPERTIES AND INTERPRETATION OF CORRELATION COEFFICIENT-SIMPLE LINEAR REGRESSION-MEANING, PRINCIPLE OF ORDINARY LEAST SQUARES AND REGRESSION LINES.

CORRELATION

In a bivariate distribution (distribution in which each unit of the series assumes two values), we may be interested to find if there is any relationship between the two variable under study. The Correlation is a statistical tool which studies the relationship between two variables and correlation analysis involves various methods and techniques used for studying and measuring the extent of the relationship between the two variables.

According to Croxton and Cowden, “when the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation”.

In the words of A.M. Tuttle “Correlation is an analysis of the covariation between two or more variables.”

In the words of A.M. Tuttle “Correlation is an analysis of the covariation between two or more variables.”

The study of correlation deals with the degree of natural statistical relationship between two or more variables, that is, correlation studies correspondence of movement (going togetherness) between two variables or series of paired items. For example,

1. If the price increases, demand decreases
2. If the price increases, supply increases
3. If the income of a family increases, the expenditure will increase
4. Cases of lung cancer may increase with increased smoking

In the above examples, the two variables move together either in the same direction or in opposite direction. In correlation we do not deal with one series but rather with the association or relationship between two series, and we do not measure variation in one series but rather compare variation in two or more series.

To measure the association of series through correlation we must have sufficient number of items in the series. If we have only two or three pairs of values, we cannot generalize concerning the way in which they vary together. Moreover, there must not be a blank in one series where there is a value in the other series, that is, there must be pairing throughout.

**Importance (or utility) of Correlation**

1. The Correlation Coefficient helps in measuring the extent of relationship between two variables.
2. Existence of relationship between two or more variables enables us to predict what will happen in future. For example, if the production of rice has increased, other factors remaining constant, we may expect a fall in the price of rice.

3. If two variables are closely related, we can estimate the value of one variable given the value of another variable. This is done with the help of regression equation.


**KINDS OR TYPES OF CORRELATION**

1. Based on direction of change – positive and negative correlation.
2. Based on change in proportion – linear and non-linear correlation.
3. Based on number of variables – simple, partial and multiple correlations.

1. **Positive and Negative Correlation**

   If two variables move together in the same direction, the correlation between them is said to be positive. If two variables move together in opposite directions, the correlation between them is said to be negative. If they do not move together at all, then there is no correlation between them.

2. **Linear and Non-linear Correlation**

   When the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable, then the correlation is said to be linear. In such a case, if the value of the variables are plotted on a graph paper, a straight line is obtained.

   But when the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable, then the correlation is said to be non-linear or curvilinear. In such a situation, we obtain a curve if the values of the variables are plotted on a graph paper.

3. **Simple, Partial and Multiple Correlation**

   The study of correlation for two variables (of which one is independent and the other is dependent) involves application of simple correlation. When more than two variables are involved in a study of correlation, then it can either be as of multiple correlation or of partial correlation. Multiple correlation studies the relationship between a dependent variable and two or more independent variables. In partial correlation we measure the correlation between a dependent variable and one particular independent variable, assuming that all other independent variables remain constant.

**METHODS OF MEASURING CORRELATION**

I. **Graphical Method**
   (a) Scatter Diagram
   (b) Correlation Graph

II. **Algebraic Method (Coefficient of Correlation)**
   (a) Karl Pearson’s Product Moment Correlation Coefficient
   (b) Spearman’s Rank Correlation Coefficient
   (c) Coefficient of concurrent deviations
I. (a) Scatter Diagram

This is a simple diagrammatic method to establish correlation between a pair of variables. Scatter diagram can be applied for any type of correlation – linear as well as non-linear. Scatter diagram can distinguish between different types of correlation although it fails to measure the extent of relationship between the variables.

Each data point, that is, a pair of values \((x_i, y_i)\) is represented by a point in the rectangular axes of ordinates. The totality of all the plotted points forms the scatter diagram. The pattern of the plotted points reveals the nature of correlation. In case of a positive correlation, the plotted points lie from lower left corner to upper right corner, in case of a negative correlation the plotted points concentrate from upper left to lower right, and in case of zero correlation, the plotted points would be equally distributed without depicting any particular pattern.

The scatter diagram is a visual aid to show the presence or absence of correlation between two variables. A line of best fit can be drawn using the method of least squares. This line will be as close to the points as possible. If the points are very close to this line, there is very high degree of correlation. If they lie very much away from this line it shows that the correlation is not much. The following figures shows different types of correlation.

![Figure 2.1: Positive Correlation](image1)

![Figure 2.2: Perfect Positive Correlation](image2)
Figure 2.3: Negative Correlation

Figure 2.4: Perfect negative correlation

Figure 2.5: No Correlation

Figure 2.6: Curvilinear correlation
Merits of Scatter Diagram Method:

1. It is an easy way of finding the nature of correlation between two variables.
2. By drawing a line of best fit by free hand method through the plotted dots, the method can be used for estimating the missing value of the dependent variable for a given value of independent variable.
3. Scatter diagram can be used to find out the nature of linear as well as non-linear correlation.
4. The values of extreme observations do not affect the method.

Demerits of Scatter Diagram Method:

It gives only rough idea of how the two variables are related. It gives an idea about the direction of correlation and also whether it is high or low. But this method does not give any quantitative measure of the degree or extent of correlation.

II. (b) Correlation Graph:

One way of detecting the nature of correlation is to draw correlation graphs and read the direction of curves. Under this method, separate curves are drawn for the X variable and Y variable on the same graph paper. The values of the variable are taken as ordinates of the points plotted. From the direction and closeness of the two curves we can infer whether the variables are related. If both the curves move in the same direction (upward or downward), correlation is said to be positive. If the curves are moving in the opposite direction, correlation is said to be negative.

But correlation graphs are not capable of doing anything more than suggesting the fact of a possible relationship between two variables. We can neither establish any casual relationship between two variables nor obtain the exact degree of correlation through them. They only tell us whether the two variables are positively or negatively correlated.

III. Karl Pearson’s Product Moment Correlation Coefficient

This is by far the best method for finding correlation between two variables, provided the relationship between the two variables is linear. Pearson’s correlation coefficient may be defined as the ratio of covariance between the two variables to the product of the standard deviation of the two variables. If the two variables are denoted by x and y and if the corresponding bivariate data are \((x_i, y_i)\) for \(i = 1, 2, 3 \ldots \ldots \ldots n\), then the coefficient of correlation between x and y, due to Karl Pearson, is given by

\[
r = r_{xy} = \frac{\text{cov}(x,y)}{S_x \cdot S_y}
\]

where

\[
\text{cov}(x,y) = \frac{\sum (x_i-x)(y_i-y)}{n} = \frac{\sum xy_i}{n} - \bar{x} \bar{y}
\]

\[
S_x = \sqrt{\frac{\sum (x_i-x)^2}{n}} = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}
\]

and

\[
S_y = \sqrt{\frac{\sum (y_i-y)^2}{n}} = \sqrt{\frac{\sum y^2}{n} - \bar{y}^2}
\]

A simple formula for computing correlation coefficient is given by

\[
r = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}
\]
Properties of Correlation Coefficient

1. The coefficient of correlation is a unit free measure. This means that if X denotes height of a group of students expressed in cm and Y denotes their weight expressed in kg, then the correlation coefficient between height and weight would be free from any unit.

2. The coefficient of correlation remains invariant under a change of origin (adding or subtracting a constant number from all observations) and/or change of scale (multiplying or dividing all observations by a constant number) of the variables under consideration. This property states that if the original pair of variables X and Y is changed to a new pair of variables U and V by effecting a change of origin and scale for both X and Y, that is

\[ U = \frac{X - a}{b} \]

\[ V = \frac{Y - c}{d} \]

Now finding the correlation coefficient between U and V

\[ r_{uv} = \frac{n \Sigma U V - \Sigma U \Sigma V}{\sqrt{n \Sigma U^2 - (\Sigma U)^2} \cdot \sqrt{n \Sigma V^2 - (\Sigma V)^2}} \]

Then we have,

\[ r_{xy} = r_{uv} \]

The two correlation coefficients remain equal and they would have opposite signs only when b and d, the two scales, differ in sign.

3. The coefficient of correlation always lies between -1 and +1, including both the limiting values, that is

\[-1 \leq r \leq 1\]

Interpretation of Correlation Coefficient (Degree of Correlation Coefficient)

1. Perfect Positive Correlation: If the value of r is equal to one, then correlation is said to be perfect positive.

2. Perfect Negative Correlation: There is perfect negative correlation if the value of r is equal to -1.

3. Positive Correlation: If the value of r lies between 0 and 1 (that is, 0<r<1), then correlation is said to be positive. If the value is close to one, there is high positive correlation and if the value is close to zero, there is low positive correlation.

4. Negative Correlation: If the value of r lies between 0 and -1 (that is, -1<r<0), then correlation is said to be negative. If the value is close to -1, there is high negative correlation and if the value is close to 0, there is low negative correlation.

5. No correlation: If the value of r is equal to zero, then we can conclude that there exists no linear correlation between the variables concerned.

PROBABLE ERROR (P.E) OF THE COEFFICIENT OF CORRELATION

Probable Error of r is very useful in interpreting the value of r and is worked out as under for Karl Pearson’s Coefficient of Correlation.

\[ P.E. = 0.6745 \sqrt{\frac{1-r^2}{n}} \]

If r is less than its PE, it is not at all significant. If r is more than six times its PE and greater than ± 0.5, then it is considered significant.
Merits of Pearson’s Correlation Coefficient
1. It gives a precise and quantitative value indicating the degree of relationship existing between the two variables. The value of r is easily interpretable.
2. It measures the direction as well as the degree of relationship between the two variables.

Demerits of Pearson’s Correlation Coefficient
1. The value of the coefficient is affected by extreme items.
2. It assumes linear relationship between the two variables.

Example 1:
Compute the coefficient of correlation between x and y from the following data
n = 10, xy = 220, \(\Sigma x^2 = 200\), \(\Sigma y^2 = 262\), \(\Sigma x = 40\) and \(\Sigma y = 50\)

Solution:
Applying the formula
\[
\text{r} = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{n\Sigma y^2 - (\Sigma y)^2}}.
\]
\[
= \frac{10 \times 220 - 40 \times 50}{\sqrt{10 \times 200 - (40)^2} \cdot \sqrt{10 \times 262 - (50)^2}}.
\]
\[
= \frac{2000 - 2000}{\sqrt{1000} \cdot \sqrt{2400}}.
\]
\[
= \frac{0}{200}.
\]
\[
= 0.91.
\]
Thus there is a good amount of positive correlation between the two variables x and y.

Example 2:
Find the product moment correlation coefficient from the following information.

\[\begin{array}{ccccccc}
X & 2 & 3 & 5 & 5 & 6 & 8 \\
Y & 9 & 8 & 8 & 6 & 5 & 3 \\
\end{array}\]

Solution:
In order to find the covariance and the two standard deviations, we prepare the following table:

<table>
<thead>
<tr>
<th>X_i</th>
<th>Y_i</th>
<th>X_iY_i</th>
<th>X_i^2</th>
<th>Y_i^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>9</td>
<td>18</td>
<td>4</td>
<td>81</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>24</td>
<td>9</td>
<td>64</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>40</td>
<td>25</td>
<td>64</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>30</td>
<td>25</td>
<td>36</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>30</td>
<td>36</td>
<td>25</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>24</td>
<td>64</td>
<td>9</td>
</tr>
<tr>
<td>29</td>
<td>39</td>
<td>166</td>
<td>163</td>
<td>279</td>
</tr>
</tbody>
</table>

Computation of Correlation Coefficient
Using the formula
\[ r = \frac{n \sum XiY_i - \sum X_i \sum Y_i}{\sqrt{n \sum X_i^2 - (\sum X_i)^2} \cdot \sqrt{n \sum Y_i^2 - (\sum Y_i)^2}}. \]
\[ = \frac{6 \times 163 - (29)^2}{\sqrt{6 \times 279 - (39)^2}}. \]
\[ = -0.93 \]

Thus there is high degree of negative correlation between x and y.

**Example 3:**
The following data relate to the test scores obtained by eight salesmen in an aptitude test and their daily sales in thousands of rupees:

<table>
<thead>
<tr>
<th>Salesman</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scores</td>
<td>60</td>
<td>55</td>
<td>62</td>
<td>56</td>
<td>62</td>
<td>64</td>
<td>70</td>
<td>54</td>
</tr>
<tr>
<td>Sales</td>
<td>31</td>
<td>28</td>
<td>26</td>
<td>24</td>
<td>30</td>
<td>35</td>
<td>28</td>
<td>24</td>
</tr>
</tbody>
</table>

**Solution:**
Let the scores and sales be denoted by x and y respectively.

<table>
<thead>
<tr>
<th>Scores X&lt;sub&gt;i&lt;/sub&gt;</th>
<th>Sales Y&lt;sub&gt;i&lt;/sub&gt;</th>
<th>U&lt;sub&gt;i&lt;/sub&gt; = X&lt;sub&gt;i&lt;/sub&gt; - 22</th>
<th>V&lt;sub&gt;i&lt;/sub&gt; = Y&lt;sub&gt;i&lt;/sub&gt; - 30</th>
<th>U&lt;sub&gt;i&lt;/sub&gt; . V&lt;sub&gt;i&lt;/sub&gt;</th>
<th>U&lt;sub&gt;i&lt;/sub&gt;&lt;sup&gt;2&lt;/sup&gt;</th>
<th>V&lt;sub&gt;i&lt;/sub&gt;&lt;sup&gt;2&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>31</td>
<td>-2</td>
<td>1</td>
<td>-2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>55</td>
<td>28</td>
<td>-7</td>
<td>-2</td>
<td>14</td>
<td>49</td>
<td>4</td>
</tr>
<tr>
<td>62</td>
<td>26</td>
<td>0</td>
<td>-4</td>
<td>0</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>56</td>
<td>24</td>
<td>-6</td>
<td>-6</td>
<td>36</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>62</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>64</td>
<td>35</td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>4</td>
<td>25</td>
</tr>
<tr>
<td>70</td>
<td>28</td>
<td>8</td>
<td>-2</td>
<td>-16</td>
<td>64</td>
<td>4</td>
</tr>
<tr>
<td>54</td>
<td>24</td>
<td>-8</td>
<td>-6</td>
<td>48</td>
<td>64</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>-13</td>
<td>-14</td>
<td>90</td>
<td>221</td>
<td>122</td>
<td></td>
</tr>
</tbody>
</table>

Since r is unaffected by a change in origin
\[ r = \frac{n \sum UiVi - \sum Ui \cdot \sum Vi}{\sqrt{n \sum Ui^2 - (\sum Ui)^2} \cdot \sqrt{n \sum Vi^2 - (\sum Vi)^2}}. \]
\[ = \frac{8 \times 90 - (-13)(-14)}{\sqrt{8 \times 221 - (-13)^2} \cdot \sqrt{8 \times 122 - (-14)^2}}. \]
\[ = \frac{538}{1768-169 \cdot 976-196}. \]
\[ = 0.48 \]

**Example 4:**
Coefficient of correlation between $x$ and $y$ for 20 items is 0.4. The AM’s and SD’s of $x$ and $y$ are known to be 12 and 15 and 3 and 4 respectively. Later on, it was found that the pair (20, 15) was wrongly taken as (15, 20). Find the correct value of the correlation coefficient.

Solution:

We are given that $n=20$ and the original $r=0.4$, $\bar{x} = 20$, $\bar{y} = 15$ $s_x$ = 3 and $s_y$ = 4.

$$r = \frac{\text{cov}(x, y)}{s_x s_y} = 0.4$$

$$= \frac{\text{cov}(x, y)}{3 \times 4} = 0.4$$

$$\text{Cov}(x, y) = 4.8$$

$$\frac{\sum xy}{n} - \bar{x}\bar{y} = 4.8$$

$$\frac{\sum xy}{20} - 12 \times 15 = 4.8$$

$$\sum xy = 3696$$

Hence, \text{correlated} = 3696 - 20 \times 15 + 15 \times 20 = 3696

Also $s_x^2 = 9$

$$\frac{\sum x^2}{n} - \bar{x}^2 = 9$$

$$\frac{\sum x^2}{20} - 12^2 = 9$$

$$\Sigma x^2 = 3060$$

Corrected $\Sigma x^2 = 3060 - 15^2 + 20^2 = 3235$

Similary $s_y^2 = 16$

$$\frac{\sum y^2}{n} - \bar{y}^2 = 16$$

$$\frac{\sum y^2}{20} - 15^2 = 16$$

$$\Sigma y^2 = 4820$$

Corrected $\Sigma y^2 = 4820 - 20^2 + 15^2 = 4645$

Corrected $\Sigma x = n\bar{x} - \text{Wrong value} + \text{Correct value}$

$= 20 \times 12 - 15 + 20$

$= 245$

Corrected $\Sigma y = n\bar{y} - \text{Wrong value} + \text{Correct value}$

$= 20 \times 15 - 20 + 15$

$= 295$

Therefore,

Corrected $r = \frac{20 \times 3696 - 245 \times 295}{\sqrt{20 \times 3235 - (245)^2} \times \sqrt{20 \times 4645 - (295)^2}}.$

$= 0.31$
IV. Spearman’s Rank Correlation Coefficient

Karl Pearson’s Correlation Coefficient is based on the magnitudes of the variables. But in many situations it is not possible to find the magnitude of the variables. For example, we cannot measure beauty or intelligence quantitatively. But it may be possible, in their case, to rank the individuals in some order. The correlation coefficient obtained from the ranks so obtained is called rank correlation. Rank Correlation can also be applied to find the level of agreement (or disagreement) between two judges so far as assessing a qualitative characteristic is concerned. According to Spearman, the formula for Rank Correlation Coefficient is

\[ r_R = 1 - \frac{6 \sum d_i^2}{n(n^2-1)} \]

Where \( r_R \) denotes rank correlation coefficient

\[ d_i = \text{difference in ranks} \]

\[ n = \text{Number of items} \]

Example 1:

The ranking of 10 individuals at the start and at the finish of a course of training are as follows:

<table>
<thead>
<tr>
<th>Individuals:</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank before:</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>9</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>10</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Rank after:</td>
<td>6</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>10</td>
<td>5</td>
<td>9</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

Calculate rank correlation coefficient.

Solution:

In order to find out the rank correlation coefficient, we construct the following table.

<table>
<thead>
<tr>
<th>Rank before</th>
<th>Rank after</th>
<th>Rank difference(di)</th>
<th>di²</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>7</td>
<td>49</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>8</td>
<td>64</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
<td>9</td>
</tr>
</tbody>
</table>

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>176</td>
</tr>
</tbody>
</table>

Table 2.3

Computation of rank correlation coefficient
Example 2:

Compute the coefficient of rank correlation between sales and advertisement expressed in thousands of rupees from the following data.

Sales : 90 85 68 75 82 80 95 70
Advertisement : 7 6 2 3 4 5 8 1

Solution:

We note that since the highest sale as given in the data is 95, it is to be given rank 1, the second highest sales 90 is to be given rank 2 and so on. We have given rank to the other various advertisements in a similar manner.

<table>
<thead>
<tr>
<th>Rank for sales</th>
<th>Rank for Advertisement</th>
<th>Rank difference(di)</th>
<th>di²</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ r_R = 1 - \frac{6\sum di^2}{n(n^2-1)} \]
\[ = 1 - \frac{6 \times 4}{8(64-1)} \]
\[ = 1 - 0.0476 \]
\[ = 0.95 \]

The high position value of the rank correlation coefficient indicates that there is a very good amount of agreement between sales and advertisement.

**Repeated Rank**

When the values repeat in one or both series, rank correlation coefficient is obtained using the formula

\[ r_R = 1 - \frac{6\left[\sum di^2 + \frac{\sum (t_j^3 - t_j)}{12}\right]}{n(n^2-1)} \]
Here, \( t_j \) represents the \( j^{th} \) tie length and the summation \( \sum (t_j^3 - t_j) \) extends over the length of all the ties for both the series.

Example 1:

Obtain the coefficient of rank correlation between economics marks and statistics marks as given below.

<table>
<thead>
<tr>
<th>Economics Marks</th>
<th>80</th>
<th>56</th>
<th>50</th>
<th>48</th>
<th>50</th>
<th>62</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics Marks</td>
<td>90</td>
<td>75</td>
<td>75</td>
<td>65</td>
<td>65</td>
<td>50</td>
<td>65</td>
</tr>
</tbody>
</table>

Solution:

This is a case of tied ranks as more than one student share the same mark both for Economics and Statistics. For Economics, the student receiving 80 marks gets rank 1, student getting 62 marks received rank 2, the student with 60 receives rank 3, student with 56 marks gets rank 4 and since there are two students, each getting 50 marks, each would be receiving a common rank, the average of the next two ranks 5 and 6, that is, \( \frac{5+6}{2} = 5.5 \) and lastly the last rank. In a similar manner, we award to the students with statistics marks.

<table>
<thead>
<tr>
<th>Economics Marks</th>
<th>Statistics Marks</th>
<th>Rank for Economics</th>
<th>Rank for Statistics</th>
<th>Rank difference(di)</th>
<th>( d_i^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>90</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>56</td>
<td>75</td>
<td>4</td>
<td>2.5</td>
<td>1.5</td>
<td>2.25</td>
</tr>
<tr>
<td>50</td>
<td>75</td>
<td>5.5</td>
<td>2.5</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>48</td>
<td>65</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>50</td>
<td>65</td>
<td>5.5</td>
<td>5</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>62</td>
<td>50</td>
<td>2</td>
<td>7</td>
<td>-5</td>
<td>25</td>
</tr>
<tr>
<td>60</td>
<td>65</td>
<td>3</td>
<td>5</td>
<td>-2</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 2.5: Computation of Rank Correlation between Economics Marks of Statistics with Tied marks.

For Economics mark there is one tie of the length 2 and for Statistics marks, there are two ties of lengths 2 and 3 respectively.

\[
\frac{\sum(t_j^3 - t_j)}{12} = \frac{(z^3 - 2)+(z^3 - 2)+(z^3 - 3)}{12} = \frac{36}{12} = 3
\]

Therefore

\[
r_R = 1 - \frac{\delta \sum d_i^2 + \frac{\sum(t_j^3 - t_j)}{12}}{n(n^2 - 1)}
\]

\[
= 1 - \frac{6[44.5 + 3]}{7(7^2 - 1)}
\]

\[
= 1 - \frac{6 \times 47.5}{7 \times 48} = 0.15
\]
V - Coefficient of Concurrent Deviations

A very simple and causal method of finding correlation when we are not serious about the magnitude of the two variables is the application of concurrent deviation. This method involves in attaching a positive sign for a $X$ – value (except the first) if this value is more than the previous value and assigning a negative value if this value is less than the previous value. This is done for the $Y$ – series as well. The deviation in the $X$ – value and the corresponding $Y$ – value is known to be concurrent if both the deviations have the same sign.

Denoting the number of concurrent deviations by $c$ and total number of deviations as $m$ (which must be one less than the number of pairs of $X$ and $Y$ values), the coefficient of concurrent deviation is given by

$$r_c = \pm \sqrt{\pm \frac{(2c-m)}{m}}$$

If $(2c - m) > 0$, then we take the positive sign both inside and outside the radical sign and if $(2c - m) < 0$, we take the negative sign both inside and outside the radical sign.

Like Pearson’s correlation coefficient and Spearman’s Rank correlation coefficient, the coefficient of concurrent deviations also lies between -1 and 1, both inclusive.

Example 1:

Find the coefficient of concurrent deviation from the following data.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>25</td>
<td>28</td>
<td>30</td>
<td>23</td>
<td>35</td>
<td>38</td>
<td>39</td>
<td>42</td>
</tr>
<tr>
<td>Demand</td>
<td>35</td>
<td>34</td>
<td>35</td>
<td>30</td>
<td>29</td>
<td>28</td>
<td>26</td>
<td>23</td>
</tr>
</tbody>
</table>

Solution:

Table 2.6

<table>
<thead>
<tr>
<th>Year</th>
<th>Price</th>
<th>Sign of deviation from the previous figure (a)</th>
<th>Demand</th>
<th>Sign of deviation from the previous figure (b)</th>
<th>Product of Deviation (a×b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>25</td>
<td></td>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>+</td>
<td>34</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>+</td>
<td>35</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>-</td>
<td>30</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>+</td>
<td>29</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>38</td>
<td>+</td>
<td>28</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>39</td>
<td>+</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>+</td>
<td>23</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In this case, \( m = \) number of pairs of deviation = 7
\( c = \) number of positive signs in the product of deviation column = number of concurrent deviation = 2

Thus

\[
 r_c = \pm \sqrt{\pm \frac{(2c-m)}{m}}
\]

\[
 = \sqrt{\frac{4-7}{7}}
\]

\[
 = -\frac{\sqrt{3}}{7} = -0.65
\]

(Since \( \frac{2c-m}{m} = \frac{3}{7} \) we take negative sign both inside and outside of the radical sign)

**REGRESSION ANALYSIS**

The term ‘regression’ was first used in 1877 by Sir Francis Galton who made a study that showed that the height of children born to tall parents will tend to move or ‘regress’ towards the mean height of the population. In other words, the height of the children of unusually tall or unusually short parents tends to move toward the average height of the population. Galton’s ‘law of universal ‘regression’ was confirmed by his friend Karl Pearson, who collected more than a thousand records of height of members of family groups. He found that the average height of sons of a group of tall fathers was less than their fathers’ height and the average height of sons of a group of short fathers was greater than their fathers’ height, thus “regressing” tall and short sons alike toward the average height of all men. In the words of Galton, this was “regression to mediocrity”.

The modern interpretation of regression is, however, quite different. Broadly speaking, we may say “regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables, with a view to estimating and or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter”.

When there are two variable \( X \) and \( Y \), and if \( y \) is influenced by \( X \), that is, if \( Y \) depends on \( X \), then we get a simple linear regression or simple regression. Here \( Y \) is known as dependent variable or explained variable and \( X \) is known as independent variable or explanatory variable. However, if we are studying the dependence of one variable on more than one explanatory variable, it is known as multiple regression analysis. In other words, in simple regression there is only one explanatory variable, whereas in multiple regression there is more than one explanatory variable.

**Assumptions in Regression Analysis**

While making use of the regression technique for making predictions it is always assumed.

i) That there is an actual relationship between the dependent and independent variables;

ii) That the values of the dependent variable are random but the values of the independent variable are fixed quantities without error,
iii) That there is clear indication of direction of the relationship. This means that dependent variable is a function of independent variable;
iv) That is conditions are the same when the regression model is being used. In other words, it simply means that the relationship has not changed since the regression equation was computed; and 
v) That the analysis be used to predict values within the range (and not for values outside the range for which it is valid.

**Regression Versus Causation**

Although regression analysis deals with the dependence of one variable on other variables, it does not necessarily imply causation. In the words of Kendall and Stuart, “A statistical relationship, however strong and however suggestive, can never establish causal connection: our ideas of causation must come from outside statistics, ultimately from some theory or other”. Thus, a statistical relationship per se cannot logically imply causation. To ascribe causality, one must appeal to a prior or theoretical considerations.

**Regression V. Correlation**

Closely related to but conceptually very much different from regression analysis is correlation analysis, where the primary objective is to measure the strength or degree of linear association between two variables. The correlation coefficient measures this strength of linear association. For example, we may be interested in finding the correlation coefficient between smoking and lung cancer, between scores on statistics and mathematics examinations, between high school grades and college grades, and so on. In regression analysis, we are not primarily interested in such a measure. Instead, we try to estimate or predict the average value of one variable on the basis of the fixed values of other variables. Thus, we may want to know whether we can predict the average score on a statistics examination by knowing a student’s score on a mathematics examination.

Regression and correlation have some fundamental difference that are worth mentioning. In regression analysis there is an asymmetry in the way the dependent and explanatory variables are treated the dependent variable is assumed to be statistical, random, or stochastic, that is, to have a probability distribution. The explanatory variable, on the other hand, are assumed to have fixed values (in repeated sampling). In correlation analysis, on the other hand we treat any two variables symmetrically; there is no distinction between the dependent and explanatory variables. After all, the correlation between X and Y is the same as that between Y and X. Moreover, both variables are assumed to be random.

**Simple Linear Regression Model**

In case of simple linear regression model, a single variable is used to predict another variable on the assumption of linear relationship between the given variables. The variable to be predicted is called the dependent variable and the variable on which the prediction is based is called the independent variable. If variable Y depends on X, then the regression line of Y on X is given by

$$ Y = a + bX $$

Here a and b are two constants and they are also known as regression parameters. Furthermore, b is also known as the regression coefficient of Y on X, and also denoted by $b_{yx}$. We may define the regression line of Y on X as the line of best fit obtained by the method of least squares.
Similarly, it variable X depends on Y, then the regression line of X on Y is given by

\[ X = a_{xy} + b_{xy} Y \]

Here \( a_{xy} \) and \( b_{xy} \) are parameters and the subscript \( xy \) is used to distinguish them from those of Y on X regression line.

**Estimation of Parameters: The Method of Least Squares**

The method of least squares or ordinary lease squares (OLS) is attributed to Corl Friedrich Gauss, a German Mathematician. Under certain assumptions, the method of least squares has some very attractive statistical properties that have made it one of the most powerful and popular methods of regression analysis. To understand this method, we first explain the least square principle.

In order to explain the principle of least squares, we first define the error or disturbance term \( \hat{u} \) as the difference between the actual or individual value of the dependent variable Y and the estimated value \( \hat{y} \) (estimated on the basis of sample values). The methods of least squares states that the regression line should be drawn through the plotted points in such a way that the sum of the squares of the of the vertical deviations of the actual and estimated Y values (that is, \( \hat{u} \)) is minimum.

That is,

\[
\text{Minimize } \sum \hat{u}^2 = \sum (y - \hat{y})^2
\]

Thus, the line of regression becomes the line of best fit.

According to the principle of least squares, the normal equations for estimating parameters \( a \) and \( b \) of Y on X regression line are

\[
\sum Y = na + n \sum X \quad \text{and} \quad \sum XY = a \sum X + b \sum X^2
\]

Solving these normal equations, we find

\[
b = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2} \quad \text{and} \quad a = \bar{Y} - b \bar{X}
\]

Similarly, the normal equations for the regression line X on Y are

\[
\sum X = na_{xy} + b_{xy} \sum Y \quad \text{and} \quad \sum XY = a_{xy} \sum Y + b_{xy} \sum Y^2
\]

And solving these equations we find

\[
b_{xy} = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2} \quad \text{and} \quad a_{xy} = \bar{X} - b_{xy} \bar{Y}
\]

**Example 1:**

The data given below is collected from 7 persons from a department referring to years of service and their monthly income. Based on this data, calculate regression equation Y on X and regression equation X on Y.
Solution:

Table 2.7
Calculation of Regression Coefficient

<table>
<thead>
<tr>
<th>Employee</th>
<th>Year of Service</th>
<th>Income (in 1000 Rs.)</th>
<th>X²</th>
<th>Y²</th>
<th>XY</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>7</td>
<td>18</td>
<td>49</td>
<td>324</td>
<td>126</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>20</td>
<td>64</td>
<td>400</td>
<td>160</td>
</tr>
<tr>
<td>C</td>
<td>11</td>
<td>35</td>
<td>121</td>
<td>1225</td>
<td>385</td>
</tr>
<tr>
<td>D</td>
<td>5</td>
<td>20</td>
<td>25</td>
<td>400</td>
<td>100</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
<td>15</td>
<td>9</td>
<td>225</td>
<td>45</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>12</td>
<td>1</td>
<td>144</td>
<td>12</td>
</tr>
<tr>
<td>G</td>
<td>15</td>
<td>45</td>
<td>225</td>
<td>2025</td>
<td>675</td>
</tr>
<tr>
<td>ΣX = 50</td>
<td>ΣY = 165</td>
<td>ΣX² = 494</td>
<td>ΣY² = 4743</td>
<td>ΣXY = 1503</td>
<td></td>
</tr>
</tbody>
</table>

(1) Regression equation of Y on X is given by

\[ Y = a + bX \]

Using OLS methods

\[ b = \frac{n \Sigma XY - \Sigma X \Sigma Y}{n \Sigma X^2 - (\Sigma X)^2} \]

Substituting the values

\[ b = \frac{7 \times 1503 - 50 \times 165}{7 \times 494 - (50)^2} = 2.371 \]

\[ a = \frac{\Sigma Y}{n} - b \frac{\Sigma X}{n} \]

\[ = \frac{165}{7} - 2.371 \times \frac{50}{7} = 6.639 \]

Therefore, the regression equation of Y on X is

\[ Y = 6.639 + 2.371 X \]

(2) Regression equation of X on Y is given by

\[ X = a_{xy} + b_{xy}Y \]

Using OLS method
Regression Coefficients

Regression equation of Y on X is given as

\[ Y = a + bX \]

The quantity ‘b’, the slope of the line of regression Y on X is called the “slope coefficient” or “regression coefficient”. The regression coefficient represents the increment value of the dependent variable for a unit change in the value of the independent variable.

Similarly, the regression line of X on Y is given as

\[ X = a_{xy} + b_{xy}Y \]

The quantity ‘\(b_{xy}\)’, the slope of the line of regression X on Y is called the “slope coefficient” or “regression coefficient”.

The regression coefficient of Y on X is also given by

\[ b_{xy} = r \frac{\sigma_Y}{\sigma_X} \]

The regression coefficient of X on Y is also given by

\[ b_{xy} = r \frac{\sigma_X}{\sigma_Y} \]

Where ‘r’ is the coefficient of correlation between X and Y, \(\sigma_X\) is the population standard deviation of X and \(\sigma_Y\) is the population standard deviation of Y.

Properties of Regression Lines

1. The regression coefficients remain unchanged due to a shift of origin, but change due to a shift of scale. This property states that if the origin pair of variables is (x, y) and if they are changed to the pair (u, v) where

\[ u = \frac{x - d}{p} \] and \[ v = \frac{y - c}{q} \]
Then

\[ b_{yx} = \frac{a}{p} \times b_{vu} \] and

\[ b_{xy} = \frac{p}{a} \times b_{uv} \]

2. The two lines of regression interest at the point \((\bar{X}, \bar{Y})\), the mean values of X and Y. According to this property, the point of intersection of the regression line of Y on X and the regression line of X on Y is \((\bar{X}, \bar{Y})\).

3. The coefficient of correlation between two variables X and Y is the simple geometric mean of the two regression coefficients. This property states that if the two regression coefficients are denoted by \(b_{yx}\) and \(b_{xy}\), then the coefficient of correlation \(r\) is given by

\[ r = \frac{\pm b_{yx} \times b_{xy}}{\sqrt{b_{yx}^2 + b_{xy}^2}} \]

If both regression coefficients are negative, \(r\) would be negative and if both are positive, \(r\) would assume a positive value.

From this, the following can be concluded.

(a) If \(b_{yx}\) is positive, then \(b_{xy}\) should also be positive and vice versa. That is, both regression coefficient should have the same sign.

(b) If one of the regression coefficients is greater than one, the other must be less than one, since the value of the coefficient of correlation cannot be greater than one.

(c) Arithmetic mean of \(b_{yx}\) and \(b_{xy}\) is equal to or greater than the coefficient of correlation. Symbolically,

\[ \frac{b_{yx} + b_{xy}}{2} \geq r \]

Example 1:

If the relationship between two variable x and u is \(u + 3x = 10\) and between y and v is \(2y + 5v = 25\), and the regression coefficient of y on x is known as 0.8, what would be the regression coefficient of v on u ?

Solution:-

\[ u + 3x = 10 \]

Therefore, \(u = -3x + 10\)

That is, \(u = \frac{x - \frac{10}{3}}{-1/3} \) and

\[ 2y + 5v = 25 \]

\[ v = \frac{-2y + 25}{5} \]

That is, \(v = \frac{y - 25}{2} \) and \(\frac{-25}{2} \)

Using properly 1, we have

\[ b_{yx} = \frac{a}{p} \times b_{vu} \]
here q = -5/2 and p = -1/3, and given $b_{yx} = 0.8$

$$0.8 = \frac{-5/2}{-1/3} b_{vu}$$

$$0.8 = \frac{15}{2} b_{vu}$$

$$b_{vu} = \frac{2}{15} \times 0.8$$

$$= \frac{8}{75}$$

Example 2:-

For the variables x and y, the regression equations are given as $7x - 3y - 18 = 0$ and $4x - y - 11 = 0$

(i) Find the arithmetic mean of x and y
(ii) Identify the regression equation of y on x
(iii) Compute the correlation coefficient between x and y.
(iv) Given the variance of x as 9, find the standard deviation of y.

Solution:-

(i) Since the two lines of regression interest at the point $(\bar{x}, \bar{y})$, replacing x and y by $\bar{x}$ and $\bar{y}$ respectively in the given regression equations, we get.

$$7 \bar{x} - 3 \bar{y} - 18 = 0$$
$$4 \bar{x} - \bar{y} - 11 = 0$$

Solving these equations, we get

$$\bar{x} = 3 \text{ and } \bar{y} = 1$$

Thus the arithmetic mean of x and y is given by 3 and 1 respectively.

(ii) Let us assume that $7x - 3y - 18 = 0$ represents the regression line of Y on X and $4x - y - 11 = 0$ represents the regression line of X on Y.

Now

$$y = -6 + \frac{7}{3} x$$

Therefore

$$b_{yx} = \frac{7}{3}$$

Again $4x - y - 11 = 0$

$$x = \frac{11}{4} + \frac{1}{4} y$$

Therefore $b_{xy} = \frac{1}{4}$

Now taking the product

$$b_{yx} \times b_{xy} = \frac{7}{3} \times \frac{1}{4} = \frac{7}{12} < 1$$
Since the product is less than one, our assumptions are correct. Thus, $7x - 3y - 18 = 0$ truly represents the regression line of $Y$ on $X$.

(iii) \[ r = \pm \sqrt{b_{yx} \times b_{xy}} \]

\[ b_{yx} = \frac{7}{3} \quad \text{and} \quad b_{yx} = \frac{1}{4} \]

Therefore

\[ r = \sqrt{\frac{7}{3} \times \frac{1}{4}} \]

\[ = \sqrt{\frac{7}{12}} = 0.7638 \]

(We take the sign of $r$ as positive since both the regression coefficients are positive)

(iv) \[ b_{yx} = r \frac{\sigma_y}{\sigma_x} \]

\[ \frac{7}{3} = 0.7638 \times \frac{\sigma_y}{3} \]

(since $\sigma_x^2 = 9$ is given, $\sigma_x = 3$)

Therefore

\[ \sigma_y = \frac{7}{0.7638} = 9.1647 \]

INDEX NUMBERS

Index Number is a Composite device used to measure and compare the relative changes in the magnitude of certain heterogeneous items, in two or more distinct situations. In Economics we use index numbers to measure overtime changes in the Macro variables like price, output, employment, etc. Historically, the first index was constructed in 1764 to compare the Italian price Index is in 1750 with the price level in 1500. Though originally developed for measuring the effect of change in Prices, Index numbers have become today one of the most widely used statistical device and there is hardly any field where they are not used. Index numbers are described as ‘barometers of economic activity’. i.e. one wants to get an idea as to what is happening to an economy. We should look to important indices like the index numbers of Industrial production, agricultural production business activity etc.

Definition and Meaning

In the words of Croxton & Cowden, “Index numbers are devices for measuring differences in the magnitude of a group of related variables”. According to Spiegel “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.” According to Wessel Willet and Simone, “An Index number is a special type of average that provides a measurement of relative changes from time to time or from place to place.”

An Index number, therefore, measure changes in group of related variables overtime. An Index number which is computed from a Single variable is called a univariate index, whereas an index number which is constructed from a group of variables is considered a composite index.

Characteristics of Index Numbers

1) Index numbers are modified form of averages: Generally an average is based on the terms which are comparable and the units of measurement of all the terms are the same. But in the case of Index numbers, we have different type of terms and different types of units. For example, while constructing a consumer price index the various items are divided into broad heads, namely (i) Food (ii)Clothing, (iii) Fuel & Lighting (iv) House Rent and Miscellaneous. These items are expressed in different units. Thus under the head ‘food’ wheat and rice may be quoted per quintal, ghee per Kg. etc. Similarly cloth may be measure in terms of metres. An average of all these terms expressed in different units is obtained by using the technique of Index numbers.
2) Index numbers measure the net change in a group of related variables. Since Index numbers are essentially arranged, they describe in one single figure the increase or decrease in a group of related variables under study. The group of variables may be the period of a specified set of commodities, the volume of production in different sectors etc. Thus, if the consumer price index working class for Mumbai has gone up to 116 in March 19997 compared to March 1996. It means that there is a net increase of 16% in the prices of commodities included in the index.

3) Index numbers measure the effect of changes over a period of time. Index numbers are most widely used for measuring changes over a period of time. Thus we can find out the net changes in agricultural prices from the beginning of First Plan period to the end of the Eighth Plan period. Similarly we can compose the agricultural production, industrial production, imports, exports, wages etc. at two different times.

**Uses of Index Numbers:**

Index Numbers are extensively used in fields of economic and business management to compare data relating to production sales, revenues and other financial matters. Some of the important uses of index numbers are as follows:

1. They help in framing suitable policies: Many of the economic and business policies are guided by Index numbers. For example, while deciding the increase in dearness allowance of the employees, the employers have to depend primarily upon the cost of living index.

   There is a large number of other fields where Index numbers are useful. For example, Sociologists may speak of population indices, health authorities prepare indices to display changes in the adequacy of hospital facilities and educational research organizations have devised formulae to measure changes in the effectiveness of school systems.

2. They reveal trends and tendencies: Index numbers are not widely used for measuring changes over a period of time. They enable us to study the general trend of the Phenomenon under study. By examining the Index Numbers of Industrial production, business activity etc. for the last few years we can conclude about the trend of production and business activity. By examining important conclusions as to how much change is taking place due to the effect of seasonality, clinical forces, irregular forces etc.

3. They are important in forecasting future economic activity: Index numbers can be used to forecast the future events. They are often used in time series analysis, the historical study of long term trend seasonal variations and business cycle development, so that business leader may keep the pace with changing economic and business conditions and have better information available for decision making purpose.

4. Index numbers are very useful in deflating: Index numbers are used to adjust the original data for price changes, or to adjust wages for cost of living changes and thus transform nominal wages into real wages. Moreover, nominal income can be transformed into real income and nominal sales into real sales through appropriate Index Number.

**Methods of Constructing Index Numbers**

Methods of constructing Index numbers can be grouped under two heads:

(a) Unweighted Indices: and
(b) Weighted Indices
In the unweighted indices weights are not assigned whereas in the weighted indices weights are assigned to the various items. Each of these types may be further divided under two heads.

(i) Simple Aggregative: and
(ii) Simple Average of Relatives

**Unweighted Index Numbers**

**I. Simple Aggregative Method**

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

\[ P_{01} = \frac{\sum P_1}{\sum P_0} \times 100 \]

\[ \sum P_1 = \text{Total of Current year prices for various commodities} \]

\[ \sum P_0 = \text{Total of base year prices for various commodities} \]

**Eg : From the following data construct an Index for 1995 taking 1994 as base**

<table>
<thead>
<tr>
<th>Commodity</th>
<th>Price in 1994</th>
<th>Price in 1995</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>50</td>
<td>70</td>
</tr>
<tr>
<td>B</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>C</td>
<td>80</td>
<td>90</td>
</tr>
<tr>
<td>D</td>
<td>110</td>
<td>120</td>
</tr>
<tr>
<td>E</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

\[ \sum P_0 = 300 \quad \sum P_1 = 360 \]

\[ P_{01} = \frac{360}{300} \times 100 = 120 \]

This means that as compared to 1994, in 1995 here is a net increase in the prices of commodities included in the Index to the extent of 20%.

There are two main limitations of this method. They are:

1) The units used in the price or quantity quotations can exert a big influence on the value of the index, and
2) No consideration is given to the relative importance of the Commodities.

**II. Simple Average of price Relatives Method**

When this method is used to construct a price index, first of all price relatives are obtained for the various items included in the index and then arrange of these relatives is obtained using any one of the measures of central value, ie, arithmetic mean, median, mode, geometric or harmonic mean. When arithmetic mean is used for averaging the relatives, the formula for computing the index is:
Where ‘N’ refers to the number of items whose price relatives are thus averaged. When geometric mean is used for averaging the price relatives the formula for obtaining the index becomes

\[ P_{01} = \frac{\sum \left( \frac{P_1}{P_0} \times 100 \right)}{N} \]

\[ \log P_{01} = \frac{\sum \left( \frac{P_1}{P_0} \times 100 \right)}{N} \] or \[ \frac{\sum \log P}{N} \]

where \( P = \frac{P_1}{P_0} \times 100 \) OR

\[ P_{01} = \text{Antilog} \left[ \frac{\sum (P_1 \times 100)}{N} \right] = \text{Antilog} \frac{\sum \log P}{N} \]

Other measures of Central value are not common use for averaging relatives.

Eg : From the following data construct an index for 1995 taking 1994 as base by the average of relatives method by using (a) Arithmetic mean & (b) Geometric Mean.

<table>
<thead>
<tr>
<th>Commodity</th>
<th>Price in 1994</th>
<th>Price in 1995</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>50</td>
<td>70</td>
</tr>
<tr>
<td>B</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>C</td>
<td>80</td>
<td>90</td>
</tr>
<tr>
<td>D</td>
<td>110</td>
<td>120</td>
</tr>
<tr>
<td>E</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

Commodity Price in 2010 Price in 2012 Price relatives \( \frac{P_1}{P_0} \times 100 \) Log P

<table>
<thead>
<tr>
<th>Commodity</th>
<th>( P_0 )</th>
<th>( P_1 )</th>
<th>( \frac{P_1}{P_0} \times 100 )</th>
<th>( \log P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>50</td>
<td>70</td>
<td>140.0</td>
<td>2.1461</td>
</tr>
<tr>
<td>B</td>
<td>40</td>
<td>60</td>
<td>150.0</td>
<td>2.1761</td>
</tr>
<tr>
<td>C</td>
<td>80</td>
<td>90</td>
<td>112.5</td>
<td>2.0512</td>
</tr>
<tr>
<td>D</td>
<td>110</td>
<td>120</td>
<td>109.1</td>
<td>2.0378</td>
</tr>
<tr>
<td>E</td>
<td>20</td>
<td>20</td>
<td>100.0</td>
<td>2.0000</td>
</tr>
</tbody>
</table>

\[ \sum P_0 = 300 \] \[ \sum P_1 = 360 \] \[ \sum \left( \frac{P_1}{P_0} \times 100 \right) = 611.6 \] \[ \sum \log P = 10.4112 \]

\[ P_{01} = \frac{\sum \left( \frac{P_1}{P_0} \times 100 \right)}{N} \]

\[ = 122.32 \]

\[ P_{01} = \text{Antilog} \left[ \frac{\sum \left( \frac{P_1}{P_0} \times 100 \right)}{N} \right] \]

\[ = \text{Antilog} \left[ \frac{10.4112}{5} \right] \]

\[ = \text{Antilog} 2.0822 = 120.9 \]
A. Weighted Index Numbers

Construction of useful index numbers requires a conscious effort to assign to each commodity a weight in accordance with its importance in the total phenomenon that the index is supposed to describe. Weighted index numbers are of two types:

I. Weighted Aggregative Indices and
II. Weighted Average of relatives

I. Weighted Aggregative Index Numbers

In this method appropriate weights are assigned to various commodities to reflect their relative importance. The amounts of the quantity consumed, purchased or marketed are the weights used. Some of the important formulae used under this method are:

1. Laspeyre’s Method
2. Paasche’s Method
3. Dorbish & Bowley’s method
4. Fisher’s ideal Method
5. Marshall–Edgeworth Method, and
6. Kelley’s Method

1. Laspeyre’s Method: The Laspeyres price index is a weighted aggregate price index, where the weights are determined by quantities in the base period. The formula for constructing the index is:

\[ P_{01} = \frac{\sum p_1q_0}{\sum p_0q_0} \times 100 \]

The primary disadvantage of the Laspeyres Method is that it does not take into consideration the consumption pattern. The Laspeyres Index has an upward bias. When the prices increase, there is a tendency to reduce the consumption of higher priced items. Similarly when prices decline, consumers shift their purchase to those items which decline the most.

2. Paasche’s Method: Under this method weights are determined by quantities in the given year.

\[ P_{01} = \frac{\sum p_1q_1}{\sum p_0q_1} \times 100 \]

The Difficulty of this method is that revised weights or quantities must be computed each year or each period, adding to the data collection expense in the preparation of the index. When the numbers of commodities is large, this index is not frequently used.

3. Dorbish & Bowley’s Method: Dorbish & Bowley have suggested simple arithmetic mean of Laspeyres and Paasche index so as to take into account the influence of both current and base periods.

\[ P_{01} = \frac{L + P}{2} \]

Where L = Laspeyres Index P = Paasche’s Index

OR

\[ P_{01} = \frac{\sum p_1q_0 + \sum p_1q_1}{\sum p_0q_0 + \sum p_0q_1} \times 100 \]
4. **Fisher’s Ideal Index**: Fisher’s ideal index is the geometric mean of the Laspeyres and Paasche’s indices.

\[
P_{01} = \sqrt{ \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100}
\]

OR

\[
P_{01} = \sqrt{L \times P}
\]

Fisher’s Index is known as ‘ideal’ because (1) it is based on geometric mean, which is considered to be the best average for constructing index numbers. (2) It takes into account both current as well as base year prices and quantities (3) It satisfies both time reversal as well as the factor reversal list and (4) it is free from bias.

It is not, however, a practical index to compute because it is excessively laborious. The data, particularly for the Paasche segment of the index, are not readily available.

5. **Marshall-Edgeworth Method**: In this method also both the current year as well as base year prices and quantities are considered.

\[
P_{01} = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100
\]

It is a simple, readily constructed measure, giving a very close approximation to the results obtained by the ideal formula.

6. **Kelley’s Method**: According to Truman L. Kelly the formula for constructing index numbers.

\[
P_{01} = \frac{\sum p_1 q}{\sum p_0 q} \times 100
\]

Where \( q \) refer to some period, not necessarily the base year or current year.

Eg: Construct index number of price from the following data by applying 1) Laspeyres Method, 2) Paasche’s Method, 3) Bowleys Method 4) Fisher Method and 5) Marshall Edgeworth Method.

<table>
<thead>
<tr>
<th>Commodity</th>
<th>1994</th>
<th>1995</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Price</td>
<td>Quantity</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>19</td>
</tr>
</tbody>
</table>
### Laspeyres Method

\[
P_01 = \left( \frac{\sum p_1 q_0}{\sum p_0 q_0} \right) \times 100
\]

\[
= \frac{200}{160} \times 100 = 125
\]

### Paasche’s Method

\[
P_01 = \left( \frac{\sum p_1 q_1}{\sum p_0 q_1} \right) \times 100
\]

\[
= \frac{130}{103} \times 100 = 126.21
\]

### Bowley’s Method

\[
P_01 = \left( \frac{L + P}{2} \right)
\]

\[
= \frac{125 + 126.21}{2} = 125.6
\]

### Fisher’s Method

\[
P_01 = \left( \frac{\sum p_1 q_0}{\sum p_0 q_0} \right) \times \left( \frac{\sum p_1 q_1}{\sum p_0 q_1} \right) \times 100
\]

\[
P_01 = \frac{200}{160} \times \frac{130}{103} \times 100
\]

\[
= 1.578 \times 100
\]

\[
= 1.256 \times 100 = 126.6
\]

### Marshall – Edgeworth Method

\[
P_01 = \left( \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \right) \times 100
\]

\[
= \frac{200 + 130}{160 + 103} \times 100 = \frac{330}{263} \times 100 = 125.47
\]

### II. Weighted Average of Price Relatives Method

In this method, appropriate weights are assigned to the commodities according to the relative importance of those commodities in the group. Thus the index for the whole group is obtained on taking the weighted average of the price relatives. To find the average, Arithmetic Mean or Geometric Mean can be used.
When AM is used, the index is \( P_{01} = \frac{\sum PV}{\sum V} \)

Where \( P = \) Price relative
\( V = \) Value of weights i.e. \( p_0 q_0 \)

When GM is used, the index is \( P_{01} = \text{Antilog} \left( \frac{\sum V \log P}{\sum V} \right) \)

Where \( P = \frac{p_1}{p_0} \times 100 \)
\( V = \) value of weight

Eg: From the following data compute price index by supplying weighted average of price relatives method using Arithmetic Mean and Geometric Mean.

<table>
<thead>
<tr>
<th>Commodity</th>
<th>Sugar</th>
<th>Flour</th>
<th>Milk</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{p_0}{q_0} )</td>
<td>3.0</td>
<td>1.5</td>
<td>1.0</td>
</tr>
<tr>
<td>( \frac{p_1}{q_0} )</td>
<td>20 Kg</td>
<td>40 Kg</td>
<td>10 Lit.</td>
</tr>
<tr>
<td>( \frac{p_1}{p_0} \times 100 )</td>
<td>4 ( \frac{1}{3} ) x 100</td>
<td>1.6 ( \frac{1}{1.5} ) x 100</td>
<td>1.5 ( \frac{1}{1.0} ) x 100</td>
</tr>
</tbody>
</table>

By using Arithmetic Mean

<table>
<thead>
<tr>
<th>Commodity</th>
<th>( p_0 )</th>
<th>( q_0 )</th>
<th>( p_1 )</th>
<th>( p_0 q_0^c )</th>
<th>( \frac{p_1}{p_0} \times 100 )</th>
<th>( pv )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sugar</td>
<td>3.0</td>
<td>20 Kg</td>
<td>4</td>
<td>60</td>
<td>8000</td>
<td></td>
</tr>
<tr>
<td>Flour</td>
<td>1.5</td>
<td>40 Kg</td>
<td>1.6</td>
<td>60</td>
<td>6400</td>
<td></td>
</tr>
<tr>
<td>Milk</td>
<td>1.0</td>
<td>10 Lit.</td>
<td>1.5</td>
<td>10</td>
<td>1500</td>
<td></td>
</tr>
</tbody>
</table>

\( \sum V = 130 \)
\( \sum V = 15000 \)

\( P_{01} = \frac{\sum PV}{\sum V} = \frac{15900}{130} = 122.31 \)

3) By using Geometric Mean

<table>
<thead>
<tr>
<th>Commodity</th>
<th>( p_0 )</th>
<th>( q_0 )</th>
<th>( p_1 )</th>
<th>( p_0 q_0^c )</th>
<th>( p )</th>
<th>( \log p )</th>
<th>( V \log p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sugar</td>
<td>3.0</td>
<td>20 Kg</td>
<td>4</td>
<td>60</td>
<td>133.3</td>
<td>2.1249</td>
<td>127.494</td>
</tr>
<tr>
<td>Flour</td>
<td>1.5</td>
<td>40 Kg</td>
<td>1.6</td>
<td>60</td>
<td>106.7</td>
<td>2.0282</td>
<td>121.692</td>
</tr>
<tr>
<td>Milk</td>
<td>1.0</td>
<td>10 Lit.</td>
<td>1.5</td>
<td>10</td>
<td>150.0</td>
<td>2.1761</td>
<td>21.761</td>
</tr>
</tbody>
</table>

\( \sum V = 130 \)
\( \sum V. \log p = 170.947 \)
\[ P_{01} = \text{Antilog} \left( \frac{\sum v log p}{\sum v} \right) = \text{Antilog} \left( \frac{270.947}{130} \right) = \text{Antilog} 2.084 = 120.9 \]

Merits of weighted Average of Relative Indices

- When different index numbers are constructed by the average of relatives method, all of which have the same base, they can be combined to form a new index.
- When an index is computed by selecting one item from each of the many subgroups of items, the values of each subgroup may be used as weights. Then only the method of weighted average of relatives is appropriate.

Test of Index Numbers:

The following are the most important tests through which one can list the consistency of index numbers.

1. The time reversal test
2. The factor reversal test

The Time Reversal Test

If \( P_{01} \) stands for the juice relative for year ‘1’ with base year ‘0’ with base year ‘1’, then the time reversal test requires,

\[ P_{01} \times P_{10} = 1 \]

This test is not satisfied by both Laspeyres and Paasche’s index numbers. Fisher’s formula satisfies this test

\[ P_{01} \times P_{10} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \neq 1 \]

Paasche’s Method \( = P_{01} \times P_{10} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \neq 1 \)

Fisher’s Method \( = P_{01} \times P_{10} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0} = 1 \)

The Factor Reversal Test:

If \( P_{01} \) stands for the price relative for the year ‘1’ with base year ‘0’ and \( Q_{01} \) stands for quantity relative for the year ‘1’ with base year ‘0’, then the condition is

\[ P_{01} \times Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0} \]

Laspeyre’s Formula \( = P_{01} \times Q_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum q_1 p_0}{\sum q_0 p_0} \neq \frac{\sum p_1 q_1}{\sum p_0 q_0} \)

Paasche’s formula \( = P_{01} \times Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum q_1 p_1}{\sum q_0 p_1} \neq \frac{\sum p_1 q_1}{\sum p_0 q_0} \)
Hence the condition is not satisfied:

\[
\text{Fisher’s Formula} \quad = P_{01} \times Q_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_1}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_1 q_0}
\]

Fisher’s formula satisfies both time reversal and factor reversal test. This is why the Fisher’s formula is often called Fisher’s Ideal Index Number.

**Base Shifting, Splicing and Deflating the Index Numbers:**

For a variety of reasons, it frequently becomes necessary to change the reference base of an Index number series from one time period to another without returning to the original raw data and recomputing the entire series. This change of reference base period is referred as “shifting the base”. There are two important reasons for base shifting. They are, the previous base has become too old and is almost useless for the purpose of Comparison. And which have been computed on different base periods, it may be describe for them to have the same base period. This may necessitates a shift in the base period.

The formula for shifting the base is,

\[
\frac{\text{Index number of an year}}{\text{Index number of base year}} \times 100
\]

Eg: The Index Numbers given below are worked out with 1990 = 100. Shift the base to 2005.

<table>
<thead>
<tr>
<th>Year</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index</td>
<td>215</td>
<td>300</td>
<td>312</td>
<td>350</td>
<td>400</td>
<td>420</td>
<td>450</td>
<td>460</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Given Index 1990 = 100</th>
<th>Index 2005 = 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>215</td>
<td>( \frac{215}{420} \times 100 = 51.2 )</td>
</tr>
<tr>
<td>2001</td>
<td>300</td>
<td>( \frac{300}{420} \times 100 = 71.4 )</td>
</tr>
<tr>
<td>2002</td>
<td>312</td>
<td>( \frac{312}{420} \times 100 = 74.3 )</td>
</tr>
<tr>
<td>2003</td>
<td>350</td>
<td>( \frac{350}{420} \times 100 = 83.3 )</td>
</tr>
<tr>
<td>2004</td>
<td>400</td>
<td>( \frac{400}{420} \times 100 = 95.2 )</td>
</tr>
<tr>
<td>2005</td>
<td>420</td>
<td>( \frac{420}{420} \times 100 = 100.0 )</td>
</tr>
<tr>
<td>2006</td>
<td>450</td>
<td>( \frac{450}{420} \times 100 = 107.1 )</td>
</tr>
<tr>
<td>2007</td>
<td>460</td>
<td>( \frac{460}{420} \times 100 = 109.5 )</td>
</tr>
</tbody>
</table>
Splicing:

By Splicing we mean combining two or more overlapping series of Index Number with different base year into one with a common base year. It is very useful for enabling comparison between the new and old index numbers.

\[
\text{Spliced Index No.} = \frac{\text{Index Number of Current Year} \times \text{Old Index Number of New base year}}{100}
\]

Eg: An Index Number was constructed with 1980 as base and continued up to 1996. From 1996 Index was computed with 1996 as base.

<table>
<thead>
<tr>
<th>Year</th>
<th>Index with 1980 as base</th>
<th>Index with 1996 as base</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>1981</td>
<td>105</td>
<td></td>
</tr>
<tr>
<td>1985</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td>1986</td>
<td>112</td>
<td></td>
</tr>
<tr>
<td>1988</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>1990</td>
<td>125</td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td>132</td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td>150</td>
<td>100</td>
</tr>
<tr>
<td>1997</td>
<td></td>
<td>120</td>
</tr>
<tr>
<td>1998</td>
<td></td>
<td>140</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>150</td>
</tr>
</tbody>
</table>

Therefore Spliced Series:

<table>
<thead>
<tr>
<th>Year</th>
<th>Spliced series (1980 = 100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>100</td>
</tr>
<tr>
<td>1981</td>
<td>105</td>
</tr>
<tr>
<td>1985</td>
<td>110</td>
</tr>
<tr>
<td>1986</td>
<td>112</td>
</tr>
<tr>
<td>1988</td>
<td>120</td>
</tr>
<tr>
<td>1990</td>
<td>125</td>
</tr>
<tr>
<td>1995</td>
<td>132</td>
</tr>
<tr>
<td>1996</td>
<td>150</td>
</tr>
<tr>
<td>1997</td>
<td>(\frac{150}{100} \times 120 = 180)</td>
</tr>
<tr>
<td>1998</td>
<td>(\frac{150}{100} \times 140 = 210)</td>
</tr>
<tr>
<td>2000</td>
<td>(\frac{150}{100} \times 150 = 225)</td>
</tr>
</tbody>
</table>

Deflating

By deflating we mean making allowances for the effect of changing price levels. A rise in price level causes a reduction in the purchasing power of money. When we express the value of money giving allowances to the changes in price level, the process is known as deflating.

Purchasing power of money is the reciprocal of an appropriate price index.
Current purchasing power of Money = \( \frac{1}{\text{Price Index}} \times 100 \)

Since the value of money goes down with rising prices the workers are interested not in how much they earn, but how much their wage will buy. This is called the real wage.

\[
\text{Real wage} = \frac{\text{Current Cash wage}}{\text{Current price index}} \times 100
\]

Given below are the price indices for a number of years. Find the purchasing power of money.

<table>
<thead>
<tr>
<th>Year</th>
<th>Price Index</th>
<th>Purchasing power of money</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>100</td>
<td>( \frac{1}{100} \times 100 = 1.00 )</td>
</tr>
<tr>
<td>2006</td>
<td>150</td>
<td>( \frac{1}{150} \times 100 = 0.67 )</td>
</tr>
<tr>
<td>2007</td>
<td>165</td>
<td>( \frac{1}{165} \times 100 = 0.61 )</td>
</tr>
<tr>
<td>2008</td>
<td>180</td>
<td>( \frac{1}{180} \times 100 = 0.56 )</td>
</tr>
</tbody>
</table>

Eg: The following table gives the money wages and cost of living index number based on 2002. Calculate real wages.

<table>
<thead>
<tr>
<th>Year</th>
<th>Wage (in Rs.)</th>
<th>Index No. (2002 = 100)</th>
<th>Deflated money or real wages</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>65</td>
<td>100</td>
<td>( \frac{65}{100} \times 100 = 65.00 )</td>
</tr>
<tr>
<td>2003</td>
<td>70</td>
<td>110</td>
<td>( \frac{70}{110} \times 100 = 63.64 )</td>
</tr>
<tr>
<td>2004</td>
<td>75</td>
<td>120</td>
<td>( \frac{75}{120} \times 100 = 62.50 )</td>
</tr>
<tr>
<td>2005</td>
<td>80</td>
<td>130</td>
<td>( \frac{80}{130} \times 100 = 61.54 )</td>
</tr>
<tr>
<td>2006</td>
<td>90</td>
<td>150</td>
<td>( \frac{90}{150} \times 100 = 60.00 )</td>
</tr>
<tr>
<td>2007</td>
<td>100</td>
<td>200</td>
<td>( \frac{100}{200} \times 100 = 50.00 )</td>
</tr>
<tr>
<td>2008</td>
<td>120</td>
<td>250</td>
<td>( \frac{120}{250} \times 100 = 48.00 )</td>
</tr>
</tbody>
</table>
Consumer Price Index Number

The Consumer price index numbers are generally intended to represent the average change over time in the prices paid by the ultimate consumer of a specified basket of goods and services. The need for constructing such an index arises because the general index numbers fail to give an exact idea of the effect of the change in the general price level on the cost of living of different classes of people. The construction of such an index is of great significance because very often the demand for a higher wage is based on the cost of living index and the wages and salaries in most countries are adjusted in accordance with the consumer price index. It is also known as cost of living index numbers.

The cost of living indices are of great significance. Automatic adjustments of wages or dearness allowance component of wages are governed in many countries by such indices. At government level index numbers are used for wage policy, price policy, rent control, taxation and general economic policies. It is also used to measure changing purchasing power of currency, real income, etc. and it is also used for analyzing markets for particular kinds of goods and services.

Construction of a Consumer Price Index

The following are the steps in constructing a Consumer Price Index.

1. Decision about the class of people for whom the index is meant. That is whether it relates to industrial workers, teachers, officers etc. The scope of the index numbers must be clearly defined. It is also necessary to divide the geographical areas covered by the index.

2. Conducting family budget enquiry. The object of conducting a family budget enquiry to determine the amount that an average family of the group included in the index spends on different items of consumption. While conducting such an enquiry, the quantities consumed and their prices are taken into account.

3. Obtaining price quotations. The collection of retail prices is a very important and, at the same time, very tedious and difficult task because such prices may vary from place to place, shop to shop and person to person. Price quotations should be obtained from the localities in which the class of people concerned reside or from where they usually make their purchases. The retail price should relate to a fixed list of items and for each item the quantity should be fixed by means of suitable specifications. Retail prices should be those actually charged from consumers. Discount should be taken into account if it is given to all customers, In a period of price control or rationing, where illegal prices are charged openly such prices should be taken into account along with the controlled prices.

Method of Constructing the Index

The index may be constructed by applying any of the following methods:

1) Aggregate Expenditure Method or Aggregation Method
2) Family Budget Method or the Method of Weighted Relatives.

1. Aggregate Expenditure Method.

When this method is applied the quantities of commodities consumed by the particular group in the base year are estimated which constitutes the weight.

\[
\text{Consumer Price Index} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100
\]
2. Family Budget Method

When this Method is applied the family budgets of a large number of people for whom index is meant are carefully studied and the aggregate expenditure of an average family on various items is estimated.

\[ \text{Consumer Price Index} = \frac{\sum PV}{\sum v} \]

Where \( P = \frac{P_1}{P_0} \times 100 \)

\( V = \text{Value of weights i.e. } P_0q_0 \)

Eg: Construct the Consumer price index number of 1998 on the basis of 1997 from the following data using 1) the aggregate expenditure method and 2) the family budget method.

<table>
<thead>
<tr>
<th>Commodity</th>
<th>Quantity Consumed in 1997</th>
<th>Unit</th>
<th>Price in 1997</th>
<th>Price in 1998</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>Quintal</td>
<td>5.75</td>
<td>6.00</td>
</tr>
<tr>
<td>B</td>
<td>6</td>
<td>Quintal</td>
<td>5.00</td>
<td>8.00</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>Quintal</td>
<td>6.00</td>
<td>9.00</td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>Quintal</td>
<td>8.00</td>
<td>10.00</td>
</tr>
<tr>
<td>E</td>
<td>4</td>
<td>Kg.</td>
<td>2.00</td>
<td>1.50</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>Quintal</td>
<td>20.00</td>
<td>15.00</td>
</tr>
</tbody>
</table>

\[
\begin{array}{c|c|c|c|c|c}
\text{Commodity} & \text{Quantity} & \text{Unit} & \text{Price in 1997} & \text{Price in 1998} \\
\hline
A & 6 & Quintal & 5.75 & 6.00 \\
B & 6 & Quintal & 5.00 & 8.00 \\
C & 1 & Quintal & 6.00 & 9.00 \\
D & 6 & Quintal & 8.00 & 10.00 \\
E & 4 & Kg. & 2.00 & 1.50 \\
F & 1 & Quintal & 20.00 & 15.00 \\
\end{array}
\]

\[
\begin{align*}
\text{Commodity} & \quad q_0 \quad p_0 \quad p_1 \quad P_1q_0 \quad p_0q_0 \quad p \quad v \quad p_0v \\
A & \quad 6 \quad 5.75 \quad 6.00 \quad 36 \quad 34.50 \quad 104.34 \quad 34.50 \quad 3600 \\
B & \quad 6 \quad 5.00 \quad 8.00 \quad 48 \quad 30.00 \quad 160.00 \quad 30.00 \quad 4800 \\
C & \quad 1 \quad 6.00 \quad 9.00 \quad 9 \quad 6.00 \quad 150.00 \quad 6.00 \quad 900 \\
D & \quad 6 \quad 8.00 \quad 10.00 \quad 60 \quad 48.00 \quad 125.00 \quad 48.00 \quad 6000 \\
E & \quad 4 \quad 2.00 \quad 1.50 \quad 6 \quad 8.00 \quad 75.00 \quad 8.00 \quad 600 \\
F & \quad 1 \quad 20.00 \quad 15.00 \quad 15 \quad 20.00 \quad 75.00 \quad 20.00 \quad 1500 \\
\end{align*}
\]

\[
\text{Consumer Price Index} = \frac{\sum P_1q_0}{\sum p_0q_0} \times 100 = \frac{17400}{146.5} \times 100 = 118.77
\]

\[
\text{Family Budget method} = \frac{\sum PV}{\sum v} = \frac{17400}{1465} = 118.77
\]

Precautions while using Consumer price Index

Following factors are to be born in mind while constructing consumer Price Index Number
1) Consumer price Index measures changes in the retail prices only in the given period compared to the base period.
2) Consumption pattern also changes from time to time
3) Qualities of goods consumed also change.

Stock Market Index Number

A Stock Index or Stock Market Index is a method of measuring the value of a section of the stock market. It is computed from the prices of selected stocks. It is a tool used by investors and financial managers to describe the market and to compare the return on specific investments.

BSE & NSE

BSE, the Bombay Stock Exchange is the oldest stock exchange in Asia. In October 2007, the equity market capitalization of the companies if the largest stock exchange in South Asia and the tenth largest in the world.

BSE was established in 1875. There are around 4800 Indian Companies listed with the stock exchange. SENSEX is the bench mark Index of the BSE. It is an indicator of all the major Companies of the BSE. If the Sensex goes up, it means that the prices of the stocks of most of the major companies on the BSE have gone up. The sensitivity Index, Sensex has got 30 listed companies.

NSE, the National Stock Exchange of India Ltd., is the largest stock exchange in India and the third largest in the World in terms of volume of transactions. NIFTY is the major index of NSE and it comprises of 50 listed companies. In October 2007, the equity market capitalization of the companies listed on the NSE was US $ 1.46 trillion, making it the second largest stock exchange in South Asia behind the BSE and the eleventh largest in the world.

TIME SERIES ANALYSIS

A time series is a sequence of observations of a certain variables at regular time intervals. Analysis of time series is a statistical device which can be used to understand, interpret and evaluate changes in economic phenomena over time with the hope of more correctly anticipatory the course of future events. For example, a business man is interested in finding out his likely sales in the year 1998 or as a long term planning in 2000 or the year 2010 so that he could adjust his production accordingly and avoid possibility of either unsold stocks of inadequate production to meet the demand. In the time series analysis, time is the most important factor because the variable is related to time which may be either year, month, week, day, hour or even minutes or seconds.

The analysis of time series if of great significance not only to the economists and business man but also to the scientist, astronomist, geologist etc. for the reasons given below.

1) It helps in understanding past behavior. It helps to understand what changes have taken place in the past. Such analysis is helpful in predicting the future behavior.
2) It helps in planning future operations: Statistical techniques have been evolved which enable time series to be analysed in such a way that the influence which have determined the form of that series may be ascertained. If the regularity of occurrence of any feature over a sufficient long period could be clearly established then. Within limits, prediction of probable future variations would become possible.
3) It helps in evaluating current accomplishments. The actual performance can be compared with the expected performance and the cause of variation analysed. For example, if expected sale for 2000-01 was 10,000 washing machine and the actual sale was only 9000. One can investigate the cause for the shortfall in achievement.

4) It facilitates comparison. Different time series are often compared and important conclusions drawn therefrom.

Components of Time Series

The fluctuations of time series can be classified into four basic type of variations. They are often called components or elements of a time series. They are:

1) Secular Trend
2) Seasonal Variations
3) Cyclical Variations
4) Irregular Variations

Secular Trend

Trend, also called secular for long term trend is the basic tendency of production, sales, income, employment etc. to grow or decline over a period of time. It include steady movements over a long time. Secular trend movements are attributable to factors such as population change, technological progress and large scale shifts in consumer tastes.

Seasonal Variation

Seasonal variations are those periodic movements in business activity which occur regularly every year and have their origin in the nature of the year itself. Seasonal variation is evident when the data are recorded at weekly or monthly or quarterly intervals. Climate & weather conditions, customs, traditions and habits are the important factors that cause seasonal variation.

Cyclical Variation

Cyclical fluctuations are long term movements that represent consistently recurring rises and declines in activity. These variations usually last for two or more years and are regular neither in amplitude nor in length. There are four different phases in a business or the economy called business cycle. They are:

- Prosperity or Boom
- Recession
- Depression
- Recovery

Most of the series of data relating to price, investment, income, wage, production etc, in an economy exhibit this type of cycle.

Irregular Variations

Irregular variations do not repeat in any specific pattern. They are also called erratic, accidental, episodic variations. These variations are caused by accidental and random factors like earthquakes, famines, flood, wars, strikes, lockouts, epidemics and revolutions. They include variations which are not attributable to secular seasonal or cyclical variations.
Measurement of Trend: Moving Average and the Method of least squares:

1. Method of Moving Averages

When a trend is to be determined by the method of moving average value for a number of years is secured and this average is taken as the normal or trend value for the unit of time falling at the middle of the period covered in the calculation of the average. While applying this method, it is necessary to select a period for moving average such as 3 yearly, 5 yearly or 8 yearly moving average etc.

The 3 yearly moving average shall be computed as follows:

$$\frac{a + b + c}{3}, \frac{b + c + d}{3}, \frac{c + d + e}{3}, \frac{d + e + f}{3}$$ ...

5 yearly moving average

$$\frac{a + b + c + d + e}{5}, \frac{b + c + d + e + f}{5}, \frac{c + d + e + f + g}{5}$$ ...

Calculate the 3 yearly moving average of the producing figures given below and draw the trend.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Production in Tonnes</td>
<td>15</td>
<td>21</td>
<td>30</td>
<td>36</td>
<td>42</td>
<td>46</td>
<td>50</td>
<td>56</td>
<td>63</td>
<td>70</td>
<td>74</td>
<td>82</td>
<td>90</td>
<td>95</td>
<td>102</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Production</th>
<th>3 Yearly totals</th>
<th>3 Yearly Movin Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1993</td>
<td>15</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>1994</td>
<td>21</td>
<td>66</td>
<td>22</td>
</tr>
<tr>
<td>1995</td>
<td>30</td>
<td>87</td>
<td>29</td>
</tr>
<tr>
<td>1996</td>
<td>36</td>
<td>108</td>
<td>36</td>
</tr>
<tr>
<td>1997</td>
<td>42</td>
<td>124</td>
<td>41.33</td>
</tr>
<tr>
<td>1998</td>
<td>46</td>
<td>138</td>
<td>46</td>
</tr>
<tr>
<td>1999</td>
<td>50</td>
<td>152</td>
<td>50.67</td>
</tr>
<tr>
<td>2000</td>
<td>56</td>
<td>169</td>
<td>56.33</td>
</tr>
<tr>
<td>2001</td>
<td>63</td>
<td>189</td>
<td>63</td>
</tr>
<tr>
<td>2002</td>
<td>70</td>
<td>207</td>
<td>69</td>
</tr>
<tr>
<td>2003</td>
<td>74</td>
<td>226</td>
<td>75.33</td>
</tr>
<tr>
<td>2004</td>
<td>82</td>
<td>246</td>
<td>82</td>
</tr>
<tr>
<td>2005</td>
<td>93</td>
<td>267</td>
<td>89</td>
</tr>
<tr>
<td>2006</td>
<td>95</td>
<td>287</td>
<td>95.67</td>
</tr>
<tr>
<td>2007</td>
<td>102</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>
Merits of Moving Average Method

- It is simple as compared to the method of least squares.
- It is flexible. If a few more figures are added to the data, the entire calculations are not changed.
- It has the advantage that it follows the general movements of the data and that its shape is determined by the data rather than statisticians choice of a mathematical function.
- It is particularly effective if the trend of a series is very irregular.
Limitations:

- Trend values cannot be computed for all the years. The moving averages for the first few years and last few years cannot be obtained. It is often these extreme years in which we may be interested.
- Selection of proper period is a great difficulty. If a wrong period is selected, there is ever likelihood that conclusions may be misleading.
- Since the moving average is not represented by a mathematical function, this method cannot be used for forecasting.
- It can be applied only to those series which show periodicity.

Method of Least Squares:

The principle of least squares provides us with a mathematical or analytical device to obtain a mathematical curve which will be fit to the given series. The technique of obtaining this mathematical curve by the principle of least squares is known as curve fitting.

The principle of least squares states that the sum of squares of the deviation between the observed values and the trend values is least. The technique can be used to fit the linear as well as non-linear trends. Linear trend is one which gives the straight line when plotted on a graph paper, curves like parabola, exponential curve, etc.

Linear Trend

A straight line can be fitted to the data by the method of curve fitting based on the most popular principle called the principle of least squares. Such a straight line is also known as Line of Best fit. Let the line of best fit be described by an equation of the type \( y = a + bx \) where \( y \) is the value of dependent variable, \( a \) and \( b \) are two unknown constants whose values are to be determined.

To find \( a \) and \( b \), we apply the method of least squares. Let \( E \) be the sum of the squares of the deviations of all the original values from their respective values derived from the equations. So that \( E = \sum (y - (a + bx))^2 \)

By Calculus method, for minimum \( \frac{\partial E}{\partial b} = 0 \). Thus we get the two equations known as Normal equations. They are:

\[
\sum y = na + b \sum x \\
\sum xy = a \sum x + b \sum x^2
\]

Solving these two normal equations, we get \( a \) and \( b \). Substituting these values in the equation \( y = a + bx \), we get the trend equation.

Eg: Fit a straight line trend to the following series by the method of least squares.

<table>
<thead>
<tr>
<th>Year</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production of Steel (000’ tones)</td>
<td>10</td>
<td>13</td>
<td>12</td>
<td>14</td>
<td>12</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>Year</td>
<td>Production of Steel y</td>
<td>X = L-2004</td>
<td>xy</td>
<td>X²</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-----------------------</td>
<td>------------</td>
<td>----</td>
<td>----</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>10</td>
<td>-3</td>
<td>-30</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>13</td>
<td>-2</td>
<td>-26</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>12</td>
<td>-1</td>
<td>-12</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>12</td>
<td>1</td>
<td>12</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>16</td>
<td>2</td>
<td>32</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>14</td>
<td>3</td>
<td>42</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>91</td>
<td>0</td>
<td>18</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Equation of straight line is y = a + bx

Normal Equations are \( \sum y = na + b \sum x \) and \( \sum xy = a \sum x + b \sum x^2 \)

Since \( \sum x = 0 \), \( a = \frac{\sum y}{n} \) and \( b = \frac{\sum xy}{\sum x^2} \)

ie. \( a = \frac{91}{7} = 13 \) and \( b = \frac{18}{28} = .64 \)

\( \therefore \) The equation of straight line trend is \( y = a + bx \) ie, \( y = 13 + 0.64x \)

The trend values are obtained by putting various values for \( x \) in the equation.

Exercise:

1. Calculate 5 yearly moving average for the following data of number of commercial and industrial failures in a country 1992-2007

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of failures</td>
<td>23</td>
<td>26</td>
<td>28</td>
<td>32</td>
<td>20</td>
<td>12</td>
<td>12</td>
<td>10</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of failures</td>
<td>13</td>
<td>11</td>
<td>14</td>
<td>12</td>
<td>9</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

2. Fit a Straight line trend by the method of least square and find the trend values

<table>
<thead>
<tr>
<th>Year</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>68</td>
<td>72</td>
<td>73</td>
<td>80</td>
<td>90</td>
<td>92</td>
<td>90</td>
<td>110</td>
<td>114</td>
</tr>
</tbody>
</table>
MODULE IV
VITAL STATISTICS

VITAL STATISTICS: MEANING AND USES- FERTILITY RATES: CRUDE BIRTH RATE, GENERAL FERTILITY RATE, SPECIFIC FERTILITY RATE, GROSS REPRODUCTION RATE AND NET REPRODUCTION RATE - MORTALITY RATES: CRUDE DEATH RATE, SPECIFIC DEATH RATE, STANDARDISED DEATH RATE, INFANT MORTALITY RATE AND MATERNAL MORTALITY RATE-SEX RATIO AND COUPLE PROTECTION RATIO.

Meaning

Vital statistics forms the most important branch of statistics and it deals with mankind in the aggregate. It is the science of numbers applied to the life history of communities and nations. Thus, it may refer to a government database (government records) recording the births and deaths of individuals within that government's jurisdiction. The term signifies either the data or the methods applied in the analysis of the data which provide a description of the vital events occurring in given communities. By vital events we mean such events of human life as birth, death, sickness, marriage, divorce, adoption, legitimation, recognition, separation, etc. In short, all the events which have to do with an individual’s entrance into or departure from life together with the changes in civil status which may occur to him during his lifetime may be called vital statistics. Thus, the system of counting births, marriages, migration, diseases and disabilities and death is known as vital statistics. Therefore, vital statistics signifies the data or methods applied in the analysis of data pertaining to vital events occurring in a community. Vital statistics have to do with people rather than things. The population census is the most fundamental statistical inquiry provides a picture of the population and its characteristics at one moment of time: Hence, vital statistics provide the tools for measuring the changes which continuously occur in a community or country.

Definitions of Vital Statistics

Vital statistics have been variously defined. Some of the important definitions of Vital Statistics are:

1. According to Arthur Newsholme, Vital Statics may be interpreted in two ways - in a broader sense and in a narrow sense. In a broader sense, it refers to all types of population statistics by whatever mode collected. In a narrower sense it refers only to the statistics derived from registrations of births, deaths and marriages.

2. According to B.Benjamin, “Vital Statistics are conventionally numerical records of marriage, births, sickness, and deaths by which the health and growth of a community may be studied”.

3. According to Arthur Newsholme, Vital Statistics is “That branch of biometry which deals with data and the laws of human mortality, morbidity and demography”.

A statistical study of human population has two aspects. 1. A study of the composition of the population at a point of time and 2. A study of the changes that occur during a given period, i.e. growth or decline of the population. Changes in the population are the outcome of the events like births, deaths, migration, marriages, divorces, etc., called vital events. Vital statistics is the application of statistical methods to the study of those facts and has been defined as the registration, preparation, transcription, collection, compilation and preservation of data pertaining to the dynamics of the population.
Thus, it is clear from the above definitions that in a broader sense, vital statistics refers to all types of population statistics. The purpose of such statistics is to find out the changing composition of communities (nations) with reference to sex, age, education, birth and death rates, marriage, economics and civic status, etc.

METHODS OF OBTAINING VITAL STATISTICS (Sources of Vital Statistics)

There are three methods of obtaining vital statistics and they are: 1. Registration Method, 2. Census Enumeration and 3. Analytical Method. We will explain the estimation of vital rates using census data

1. Registration Method:
   The registration method is the most important source of obtaining vital statistics. It may be defined as the continuous and permanent recording of the occurrence of vital events pertaining to births, deaths, marriages, migration etc. These data have their values as legal documents and they are useful as a source of statistics. In most countries there is a system of registering the occurrence of every important vital event under legal requirements. For example, when a child is born, the matter has to be reported to the proper authorities, together with such information as the age of mother, religion of parents, etc. similarly, when a man dies, the death is to be recorded with appropriate authorities and a certificate is to be obtained before the body is cremated.

   Continuous permanent recording of vital events can best be ensured by means of legislation which makes registration compulsory. Such legislation should also provide sanctions for the enforcement of the obligation. Thus it will be seen that the registration method is characterized not only by the continuous character of the observations but also by the compulsory nature of the method. Both provisions are fundamental, Registration of vital events for legal purposes is an almost universal requirement. Data on births and deaths can also be obtained from the hospital record.

2. Census Enumeration: In most countries of the world population census is undertaken generally at ten years interval. A census is an enumeration at a specified time of individuals inhabiting a specified area, during which particulars are collected regarding age, sex, marital status, occupation, religion, etc. The fundamental deficiency of the census method for collecting vital statistics is that it can, at best, produce returns for the census year and no other. Census years are usually ten years apart. For the intercensal years, current vital statistics are not produced by the census method, and thus, that method fails in the first and minimum requisite for vital statistics, i.e., the production of data on a current basis. Not only does the census method fail to provide intercensal data but it fails also to record completely the occurrence of births and deaths even for the census year. Periodic surveys have been employed to secure ad hoc information on births and deaths in areas where the registration method has not been established or where it is very defective. In such situations, survey has the distinct advantages of making available some vital statistics not otherwise obtainable and of securing at the same time the corresponding population.

3. Analytical Method: Estimation of Vital Rates using Census Data. It is assumed that the derivation of birth, death and marriage rates is the object of collecting vital statistics. Hence, there is another method which could be employed to yield these basic facts. This method is mathematical one based on an analysis of the returns of two consecutive censuses of population. The census returns employed must of necessity be the result of very accurate and dependable enumerations, which have produced reliable age and marital status distributions of the population. If certain assumptions are made regarding migration and the reliability of the enumeration is ensured, data from censuses of population can be used to derive information on the approximate
numbers of births. Deaths and marriages which have occurred in the population over the intercensal period. This indirect method yields aggregates only and that too solely for the year of the census. It does not, therefore, justify, its consideration as a method of developing vital statistics which by definition must be current and continuous. However, it is a method which has been developed for estimating vital statistics in Brazil, for example and as such should be mentioned for its applicability to the relatively rare areas which have non-existent or deficient registration statistics but a reliable census of population.

In the following discussion, we shall be concerned with births and deaths – the two most important vital events. It will be assumed that we have from census data for the given community the total size of the population and also its distribution with respect to different characters (i.e. age and sex) corresponding to different points of time: while from registers we have data regarding the number of births and deaths occurring during different periods.

In order to determine the population at a time (say, t) subsequent to a census or between two censuses one may use a number of procedures. A very common method is to make use of statistics of births, deaths, immigration and emigration. The population \( P_t \) at time \( t \) is then obtained as

\[
P_t = P_0 + (B - D) + (I - E)
\]

Where

- \( P_0 \) = total population recorded at last census
- \( B \) = total number of births during the given period
- \( D \) = total number of deaths during the given period
- \( I \) = total number of immigrants
- \( E \) = total number of emigrants

**Uses of Vital Statistics (Importance of Vital Statistics)**

Vital statistics are highly useful and they are useful to individuals, operating agencies, in research, demographic and medical field, in public administration and internationally.

1) **Use to individuals**: Records of births, deaths, marriages and divorce etc. are highly useful to the individuals. The basic registration document has legal significance to the person concerned.
2) **Use to operating Agencies**: Records of birth, deaths and marriages are useful to governmental agencies for a variety of administrative purpose. For example, the control programmes for infectious diseases within the family and community often depend on the death registration report for their initiation. Public health programmes of post-natal care for the mother and child usually have their starting point to the birth register and the corresponding birth indices. Public safety, accident prevention, and crime eradication programmes make use of the death registration records.
3) **Use in Research**: Vital statistics are indispensable in demographic research. The study of population movement and of the interrelationships of demographic with economic and social factors is of fundamental importance to society, and will become increasingly more so as the advances in technology and public health focus attention on demographic problems. The three directions which such an analysis takes are: 1) population estimation (2) population projection, and
(3) analytical studies. Very closely allied to the role of vital statistics in demographic research is their use by the medical profession engaged in research. Medical and pharmaceutical research, like demographic research, requires a certain number of guideposts. This guidance may be found in part at least in mortality and natality statistics.

4. **Use in Public Administration:** Vital statistics are vital to public health. As vital statistics include information on births, deaths and a lot of other health information which are highly useful to public health officials. The health officials must have data on the prevalence of disease and major health issues. Most national governments by law mandate the collection of vital statistics. Thus vital statistics are fundamental elements in public administration, which is the machinery and methods underlying all official programmes of economic and social development in either ‘developed’ or ‘under-developed’ areas. The role of vital statistics in overall planning and evaluation of economic and social development is the most important use to which this body of data may be placed. To monitor current demographic trends and action programmes, for scientific research to study interrelationship between fertility and mortality trends and socio-economic development, vital statistics are indispensable.

5. **International use of vital Statistics:** Vital statistics are also useful from the international viewpoint. Only by a sufficiently wide survey of human facts can the required norms of all sorts be established, norms which represent the character of the great unit constituted by the aggregation of all the nations.

6. **For policy making:** Vital statistics form the basis for policy guidance, planning and projections.

   However, it should be noted that vital statistics, like all statistics and multiple vital records, are not ends in themselves but tools for the study and understanding of other phenomena.

**FERTILITY**

The term fertility refers to the actual production of children. In the measurement of growth of population of a country, fertility has an important place. Population of a country may go on changing and changes in the population are due to the changes in births and deaths. Thus, the growth of population in a country is the net result of obtained as the difference between total births and total deaths.

**MEASUREMENT OF FERTILITY**

Fertility rates vary according to the level of development achieved by a country or region. Generally, developed countries have a much lower fertility rate due to greater wealth, education, and urbanization. Mortality rates are low, birth control is understood and easily accessible, and costs are often deemed very high because of education, clothing, feeding, and social amenities. With wealth, contraception becomes affordable. However, in countries like Iran where contraception was subsidised before the economy accelerated, birth rate also rapidly declined. Further, longer periods of time spent getting higher education often mean women have children later in life. The result is the demographic-economic paradox. Female labor participation rate also has substantial negative impact on fertility. However, this effect is neutralized among Nordic or liberalist countries. In undeveloped countries on the other hand, families desire children for their labour and as caregivers for their parents in old age. Fertility rates are also higher due to the lack of access to contraceptives, generally lower levels of female education, and lower rates of female employment in industry. In order to study the speed at which the population is increasing, fertility rates are used which are of various types. Important amongst these are:
Crude Birth Rate (Birth Rate)

It is the simplest method of measuring fertility. The crude birth rate, computed as the ratio of the number of births to the total population, is more affected by population differences in age and sex ratio. Therefore the crude birth rate is a better measure of tax burden and other economic statistics than the general fertility rate. The crude birth rate may be measured as the number of births in a given population during a given time period (such as a calendar year), divided by the total population and multiplied by 1,000. According to the United Nation’s World Population Prospects of the 2008 Revision Population Database, the crude birth rate is the number of births over a given period, divided by the person-years lived by the population over that period. It is expressed as the number of births per 1,000 population. It acts as population through childbirth. In this method the number of births is related to the total population. Since it is only a live birth that signifies an addition to the existing population, live births alone are considered in measuring fertility, thus excluding still births.

The annual crude birth rate is defined as:

\[
\text{Crude Birth Rate} = \frac{\text{Annual Births}}{\text{Annual mean population}} \times 1000
\]

In this measure the births are related to the mean population and not to the population at a particular date. The crude birth rate of a given year tells us at what rate births have augmented the population over the course of the year.

The crude birth rate usually lies between 10 and 55 per 1,000. The level of the crude birth rate is determined by:

i) The sex and age distribution of the population; and
ii) The fertility of the population. i.e., the average rate of child-bearing of females.

A relatively high crude birth rate can be recorded if the sex and age distribution is favourable even though fertility is low. Thus, countries with a relatively large proportion of population in the 15-50 years age group will have a relatively high crude birth rate, other things being equal. Hence, crude birth rate is not a suitable measure of fertility as the fertility differs from one age group to another. Therefore, we have the fertility rates like i) General Fertility Rate ii) Specific Fertility Rate and iii) Total Fertility Rate.

General Fertility Rate

The General Fertility Rate (GFR) is the birth rate of women of child bearing age (age 15-44). While births to women less than 15 or more than 44 years are included in the general fertility rate, the population for those ages are not. This rate refers to the proportion of the number of children born per 1,000 of females, the reproductive or child-bearing age. Thus the numerator of this rate would remain the same as the crude rate, but the denominator would be limited to the age-sex group of the population able to contribute to the birth rate. The general fertility rate is calculated by dividing the total number of births in a given year by the number of women aged 15 through 44 and multiplying by 1,000. The formula for such a rate is:

\[
\text{G.F.R} = \frac{\text{Number of live births which occurred among the Population of a given geographic area during a given year}}{\text{Mid-year female population of ages 15 to 49 in the given geographic area during the same year}} \times 1000
\]
The computation of the F.G.R. requires that a decision be taken before hand as to which years of the life a woman should be included in the child-bearing period. Although the practice varies in this respect, generally the child-bearing age is taken 15 to 50 years. Births to mothers under 15 and above 50 are so rare that they are not recorded separately but are included in the age-group 15 and 49 respectively.

The G.F.R. shows how much the women in child-bearing ages have added to the existing population through births. It takes into account the sex composition of the population and also the age composition to a certain extent. Yet it is calculated without proper regard to the age composition of the female population in child-bearing ages. The fecundity of women differs according to age-groups. In our country it is low in the age group 15-19 after which it gradually declines. In U.S.A. fecundity reaches its peak in the age-group 20-24 and thereafter declines. For this reason even if the general fertility rate of two populations may correspond to each other, we cannot assume that the fertility rate is really identical unless different age-groups are also taken into consideration in its calculation. It should be noted that the calculation of the general fertility rate is limited solely to live births. It is not a pregnancy rate and does not include induced abortions, fetal deaths (stillbirths), or spontaneous abortions miscarriages). The general fertility rate is the best overall indicator of reproductive behavior and success.

**Specific Fertility Rate**

*Age-specific fertility rate* refers to the number of births to females in a particular age category in a particular year compared to the number of females in that age category. *Age-specific fertility rate* is usually expressed as births per woman or births per 1,000 women in the age category. It is usually calculated for the age range 15 to 49 as only a very small proportion of births occur to women outside of that age range. *Age-specific fertility rates* are usually calculated for single years of age or for 5-year age categories.

The concept of specific fertility arises out of the fact that fertility is affected by a number of factors such as age, marriage, state or region, urban-rural characteristics, etc. When fertility rate is calculated on the basis of age distribution, it is called the age-specific fertility rate. While calculating age-specific fertility rate women of different ages in the child-bearing age are placed in small age groups so as to put them at part with others of child-bearing capacity. The fertility of women differs from age to age and, therefore, the grouping of women of different ages is essential. The capacity to bear children is much higher in the age-group 20 to 25 than in the age-group 40 to 45.

\[
S.F.R. = \frac{\text{Number of live births which occurred to female of a specified age-group of the population of a given geographic area during a given year}}{\text{Mid-year female population of the specified age-group in the given geographic area during the same year}} \times 1000
\]

**Total Fertility Rate**

Another frequently-used indicator is the total fertility rate, the average number of children born to a woman during her lifetime. The total fertility rate is generally a better indicator of current fertility rates because unlike the crude birth rate, it is not affected by the age distribution of the population. Fertility rates tend to be higher in less economically-developed countries and lower in more economically-developed countries. The TFR (or TPFR—total period fertility rate) is a better index of fertility than the Crude birth rate (annual number of births per thousand population) because it is independent of the age structure of the population, but it is a poorer estimate of actual
completed family size than the total cohort fertility rate, which is obtained by summing the age-specific fertility rates that actually applied to each cohort as they aged through time. In particular, the TFR does not necessarily predict how many children young women now will eventually have, as their fertility rates in years to come may change from those of older women now. However, the TFR is a reasonable summary of current fertility levels. The total fertility rate (TFR) are sometimes also called the fertility rate, period total fertility rate (PTFR) or total period fertility rate (TPFR) of a population is the average number of children that would be born to a woman over her lifetime if:

1. she were to experience the exact current age-specific fertility rates (ASFRs) through her lifetime, and
2. she were to survive from birth through the end of her reproductive life. It is obtained by summing the single-year age-specific rates at a given time.

The TFR is a synthetic rate, not based on the fertility of any real group of women since this would involve waiting until they had completed childbearing. Nor is it based on counting up the total number of children actually born over their lifetime. Instead, the TFR is based on the age-specific fertility rates of women in their "child-bearing years," which in conventional international statistical usage is ages 15–44 or 15–49. Replacement fertility is the total fertility rate at which newborn girls would have an average of exactly one daughter over their lifetimes. That is, women have just enough female babies to replace themselves (or, equivalently, adults have just enough total babies to replace themselves). If there were no mortality in the female population until the end of the childbearing years (generally taken as 44 or 49, though some exceptions exist) then the replacement level of TFR would be very close to 2.0 (actually slightly higher because of the excess of boy over girl births in human populations). However, the replacement level is also affected by mortality, asexuality, genetic disorders inhibiting procreation, and by women without the desire to have children. The replacement fertility rate is roughly 2.1 births per woman for most industrialized countries (2.075 in the UK for example), but ranges from 2.5 to 3.3 in developing countries because of higher mortality rates. Taken globally, the total fertility rate at replacement is 2.33 children per woman. At this rate, global population growth would trend towards zero.

The TFR is, therefore, a measure of the fertility of an imaginary woman who passes through her reproductive life subject to all the age-specific fertility rates for ages 15–49 that were recorded for a given population in a given year. The TFR represents the average number of children a woman would have were she to fast-forward through all her childbearing years in a single year, under all the age-specific fertility rates for that year. In other words, this rate is the number of children a woman would have if she was subject to prevailing fertility rates at all ages from a single given year, and survives throughout all her childbearing years.

In order to measure correctly the population growth, we calculate the number of children born per thousand females in the child-bearing age divided into different age-groups. This leads to the total fertility rate which is calculated by adding up the specific fertility rates belonging to different age-groups. The total fertility rate is the mean number of children which a female aged 15 can expect to bear if she lives until at least the age of 50. Provided she is subject to the given fertility conditions over the whole of her child-bearing period. The total fertility rate for a particular area during a given period is a summary measure of the fertility conditions operating in that area during that period. In order to calculate the total fertility rate we shall have to calculate specific fertility rates and then add them. Total fertility rate is thus the sum of the age-specific fertility rates from a given age to the last point of child-bearing age of a female. In practice, we can shorten this procedure by working in quinquennial age groups.
Specific Fertility Rate = \frac{\text{Annual births to females aged x and under (x+5)}}{\text{Mean number of females aged x and under (x+5)}} \times 1000

Such a specific fertility rate is the rate per 1,000 per annum at which the females in the particular age-group produce offspring.

If we add the quinquennial specific fertility rates and multiplies by 5, we shall have the total number of children which 1.000 females aged 15 will bear over their lifetimes. A calculation based on quinquennial age-groups involves only one-fifth of the arithmetic of one based on single age groups and is very nearly as accurate, Symbolically.

\text{TFR} = \sum \text{SFR} \times t

Where \( t \) = the magnitude of the age class.

Limitations of Fertility Rates

A few limitations of the Fertility rates are to be mentioned here. Fertility rates do not give an idea of the rate of population growth as it only refers to the number of children. If majority of births are of male children, then the female population will be reduced so that fertility rate does not reveal correct position of population growth.

REPRODUCTION RATES

The fertility rates are unsuitable for giving an idea of the rate of population growth because they ignore the sex of the newly born children and their mortality. If the majority of births are of male children, then the female population will be reduced so that fertility rate does not reveal correct position of population growth.

Gross Reproduction Rate

The gross reproduction rate (GRR) is the average number of daughters that would be born to a woman (or a group of women) if she survived at least to the age of 45 and conformed to the age-specific fertility rate of a given year. It is often regarded as the extent to which the generation of daughters replaces the preceding generation of females. The GRR is particularly relevant where sex ratios are significantly affected by the use of reproductive technologies. Gross reproduction rate measures the rate at which a new born female would, on an average, add to the total female population, if they remained alive and experienced the age-specific fertility rate till the end of the child-bearing period. It is the sum of fertility rate till the end of the child-bearing period. It is the sum of age-specific fertility rates calculated from female births for each single year of age. It shows the rate at which mothers would be replaced by daughters and the old generation by the new if no mother died or migrated before reaching the upper limits of the child-bearing age, i.e., 49 years. Another underlying assumption is that the same fertility rate continued to be in operation. If the gross reproduction rate of a population is exactly 1. It indicates that the sex under consideration is exactly replacing itself; if it is less than 1, the population would decline, no matter how the death rate may be and if it is more than 1. The population would increase, no matter how low the death rate may be. The gross reproduction rate is computed by the following formula:
G.R.R. = \frac{\text{Number of female births}}{\text{Total number of births}} \times \text{Total Fertility Rate}

Also, G.R.R. = \frac{\text{Number of female children born to 1,000 women}}{1,000}

The G.G.R. is used as a measure of the fertility in a population. It is useful for comparing fertility in different areas or in the same area at different time periods. The G.R.R. could in theory range from 0 to about 5. Gross reproduction rate has an advantage over the total fertility rate because in its computation we take into account only the female babies who are the future mothers whereas in the total fertility rate we include both male and female babies that are born.

An important limitation of the gross reproduction rate is that it ignores the current mortality. All the girls born do not survive till they reach the child –bearing age. Hence the gross reproduction rate is misleading in that it inflates the number of potential mothers. This defect is removed by computing the net reproduction rate. The accuracy of gross reproduction rate depends on the accuracy with which age-specific fertility rates can be computed. The principal sources of error are: (1) under-registration of births, (2) mis-statements or inadequate statements of the age of mother at registration, and (3) errors in enumeration or estimation of the female population by age-groups.

SURVIVAL FACTOR
Survival factor indicates the number of females surviving in a particular age group. For example, if the survival factor is 956, it means that out of 1000 live born females for that age, 956 only survive. Survival factor = \frac{\text{No. of female children born to thousand women} \times \text{No. of survivals out of thousand female children}}{1000}.

Net Reproduction Rate
An alternative fertility measure is the net reproduction rate (NRR). Gross reproduction rate adjusted for the effects of mortality is called the net reproduction rate. It measures the number of daughters a woman would have in her lifetime if she were subject to prevailing age-specific fertility and mortality rates in the given year. Thus, the net reproduction rate (NRR) is the average number of daughters that would be born to a female (or a group of females) if she passed through her lifetime conforming to the age-specific fertility and mortality rates of a given year. This rate is similar to the gross reproduction rate but takes into account that some females will die before completing their childbearing years. An NRR of one means that each generation of mothers is having exactly enough daughters to replace themselves in the population. The NRR is particularly relevant where sex ratios at birth are significantly affected by the use of reproductive technologies, or where life expectancy is low.

When the NRR is exactly one, then each generation of women is exactly reproducing itself. The NRR is less widely used than the TFR, and the United Nations stopped reporting NRR data for member nations after 1998. But the NRR is particularly relevant where the number of male babies born is very high – see gender imbalance and sex selection. This is a significant factor in world population, due to the high level of gender imbalance in the very populous nations of China and India. The gross reproduction rate (GRR), is the same as the NRR, except that - like the TFR - it ignores life expectancy. Though gross reproduction rate gives an idea about the growth of population, it excludes the effect of the mortality on the birth rate. The rate estimates the average number of daughters that would be produced by women throughout their lifetime if they
were exposed at each age to the fertility and mortality rates on which the calculation is based. It thus indicates the rate at which the number of female births would eventually grow per generation if the same fertility and mortality rates remained in operation. A net reproduction rate of 1 indicates that on the basis of the current fertility and female mortality, the present female generation is exactly maintaining itself. Both fertility and mortality are taken into account while calculating net reproduction rate. In its calculation it is assumed that 1,000 mothers give birth to a certain number of girls of whom a percentage dies in infancy and certain percentage does not marry. Of married girls some would become widow and it is only the balance that passes through fertility period and adds to the population growth. The N.R.R. measures the rate at which female population is replacing itself. Thus, the net reproduction represents the rate of replenishment of that population.

The net reproduction rate is obtained by multiplying the female specific fertility rate of each age by the population of female survivors to that age in a life table and adding up the products. An allowance is thus made for mortality.

\[
\text{N.R.R} = \sum \left( \text{Number of female births X Survival rate} \right)
\]

In other words NRR is obtained by dividing number of female birth to thousand newly born females on the basis of current fertility and mortality rates by 1000.

Net Reproduction Rate = \frac{\text{Number of female births to thousand newly born females on the basis of current Fertility and mortality rates}}{1000}

The Net reproduction rate is always less than Gross reproduction rate. Both the rates will be equal when all the newly born daughters reached the child bearing age and passed through it. The net reproduction rate in theory can range from 0 to 5. If the net reproduction rate is one, it indicates that on the basis of current fertility and mortality rates, a group of newly born females will exactly replace itself in the new generation. This means that the population will be constant. If the net reproduction rates are below one, it indicates a declining population and if it is more than one, the population has a tendency to increase.

- If NRR = 1, the female population remains constant
- If NRR< 1, the female population is declining
- If NRR> 1, the female population is rising

MEASUREMENT OF MORTALITY

Mortality Rates (Death Rates)

Mortality rate

Mortality rate is a measure of the number of deaths (in general, or due to a specific cause) in a population, scaled to the size of that population, per unit of time. Mortality rate is typically expressed in units of deaths per 1000 individuals per year; thus, a mortality rate of 9.5 (out of 1000) in a population of 100,000 would mean 950 deaths per year in that entire population. It is distinct from morbidity rate, which refers to the number of individuals in poor health during a given time period (the prevalence rate) or the number of newly appearing cases of the disease per unit of time (incidence rate). The term "mortality" is also sometimes inappropriately used to refer to the number of deaths among a set of diagnosed hospital cases for a disease or injury, rather than for the general population of a country or ethnic group. This disease mortality statistic is more precisely referred to as "case fatality rate" (CFR).
The mortality conditions of population are studied by measuring the crude death rate, the specific death rate, standardised death rate and infant mortality rate.

**The Crude Death Rate**

The crude death rate is the number of deaths from all causes in a given period (year) per 1000 of population, in a given community (locality). The crude death rate for a given year tells us at what rate deaths have depleted the population over the course of the year. We can calculate the crude death rate for males and females separately. The crude death rate usually lies between 8 and 30 per 1,000. The female rate is generally lower than the male rate. In most countries, crude death rates have fallen substantially over the years. The annual crude death rate is defined as:

$$\text{Crude Death Rate} = \frac{\text{Annual Deaths}}{\text{Annual Mean Population}} \times 1,000$$

The level of the crude death rate is determined by:

i) The sex and age distribution of population; and

ii) The mortality of the population, i.e., average longevity of the population.

An old population can exhibit a relatively high crude death rate even if longevity is high (i.e. mortality is low). The crude death rate, which measures the decrease in a population due to deaths, is perhaps the most widely used of any vital statistics rate. This is so for two reasons:

1) It is relatively easy to compute

2) It has value as an index in numerous demographics and public health problems.

However, death-rate so computed is likely to be misleading especially when it is required to compare the death rates in two areas or in two occupations. It is because of the fact that mortality varies with sex age whereas the crude death rate marks all age differentials. It assumes that age-sex structures of the populations being compared are the same. However, in practice it is not so. Population composed of a high proportion of persons at the older ages where mortality is higher will naturally show a higher crude death rate than younger population.

The crude death rate may be used for comparing the mortality situations of the same place at different times, provided the periods compared are not too far apart, because in a stable, large community the age and sex compositions of the population change very slowly. If the time trend is studied for a long period of years the effect of population changes must be examined. Greater caution is necessary for comparison between areas, since rather significant differences in crude death rates may arise entirely from differences in the age-sex distribution of the populations. However, where it is known that the population distributions are approximately similar, or where the crude rate differences are large, as in any international comparisons, the crude rate has great value as an index of mortality.

**Specific Death Rates**

By themselves, crude death rates are not enough for a detailed study of the mortality conditions in a community. We often need to know more about deaths occurring in different section of the population. For instance, people interested in infant or child welfare work study death taking place under 1 year of age or in such age groups as 1-4 years, 5-9 years, etc. Those interested in maternal health, study how many deaths occurred among women of child-bearing age. Insurance companies are interested in deaths occurring at different ages of the population.
The formula for computing specific death rate is:

Annual death rate Specific for age  = 

\[
\frac{\text{Number of deaths which occurred among a specific age group of the population of a given geographic area during a given year}}{\text{Mid-year population of the specified age group in the given geographic area during the same year}} \times 1,000
\]

These rates measure the risk of dying in each of the age groups selected for the computation. Usually such rates are computed for the entire span of years, and are further specified by sex, so that rates specific for age and sex are available. The specificity by age and sex eliminates the differences which would be due to variation in population composition in respect of these characteristics, and to this extent, such rates can be compared from one geographic area to another and from one time period to another. However, it does not eliminate other variables which also may be important, such as “occupation”, “Literacy”, and the like. Nevertheless, for general analytical purposes, the death rate specific for age and sex is one of the most important and widely applicable types of death rates. It also supplies one of the essential components required for computation of net reproduction rate and life tables.

**Standardized Death Rates**

The criticism of the crude death rate is that while making inter-area comparisons, it fails to take account of differences in the age (or age-sex) structure of the population in question and, thus, fails to reveal the “real” mortality. It has been suggested, therefore, that the crude rates, be “adjusted” to allow for the known differences in the age composition of the population involved. Several methods have been proposed and different names have been applied to the results, some workers have called these hypothetical indices “adjusted rates”. Others “standardized rates” till other “corrected rates”. Perhaps, the most appropriate term is “adjusted rates”, used with a prefix to identify the basis of the adjustment as, for example, “Age-adjusted death rate, and so forth.

The standardized death rate, abbreviated as SDR, is the death rate of a population adjusted to a standard age distribution. It is calculated as a weighted average of the age-specific death rates of a given population; the weights are the age distribution of that population. As most causes of death vary significantly with people’s age and sex, the use of standardized death rates improves comparability over time and between countries. The reason is that death rates can be measured independently of the age structure of populations in different times and countries (sex ratios usually are more stable).

There are two principal methods of age adjustment: (1) the direct method; and (2) the indirect method.

**Direct Method**

Direct method gives the death rate that would occur in some standard population if it had the mortality of the given community, or death rate that would occur in the community if its population were distributed as that of the standard. The direct method of adjusting for age consists of weighting the specific rates not by the population of the area to which they refer as its implied in the computation of the crude rate, but by the population distribution of another area, chosen as a
standard. In the direct method, the rates specific for one geographic area are multiplied by the corresponding populations of another area which, for this purpose, is considered as a “standard”. The resulting expected number for each age group is summed, and the total is divided by the total standard population to obtain an age-adjusted rate”.

The obvious defect of the direct method as a means of adjusting for the age differences of several populations is that it entails the choice of a “standard” population. The choice of this “standard” will naturally affect the magnitude of resulting adjusted rates and may change their relative positions with respect to each other. However, in eliminating bias on a national basis, it is customary to use the total population of the country as “standard” for adjusting the rates of the regions within the country. There is no generally accepted standard population for international comparisons.

Indirect Method

The indirect method adjusts the crude rate of the community by applying to it a factor measuring the relative “mortality proneness” of the population of the community. In this method the “standard” is a set of specific rates, rates than a population distributed by age. To compute an age-adjusted rate by indirect method, one requires the population of the area distributed by age. These given populations are multiplied age by age by the “standard” age specific rates to obtain the expected number of events in the standard area if it were subject to the given age distribution. The sum of these “expected events” divided by the population in the area under consideration gives an “expected rate” or “index rate” in the “standard area” – one which is dependent solely on the sex-age constitution of the population and as a rule may be treated with sufficient accuracy as remaining constant over a period of years adjacent to experience period. The “index rate” is customarily divided into the crude rate of the “standard” area and the resulting ratio is known as an “adjustment factor” which can be used to adjust the crude rate of the area under consideration. Both the direct and indirect methods of age adjustments have been criticized on the ground that the rates obtained are dependent on the age and sex structure of the standard population used and that greater gains in mortality reduction obtained at younger age are not adequately accounted for.

As method for standardization, the direct method is applied. Standardized death rates are calculated for the age group 0-64 ('premature death'), 65 years and more and for the total of ages. As most causes of death vary significantly with people's age and sex, the use of standardized death rates improves comparability over time and between countries.

Illustration 1: Compute the crude and standardized death rates of the two populations A and B from the following data:

<table>
<thead>
<tr>
<th>Age-group (years)</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Population</td>
<td>Deaths</td>
</tr>
<tr>
<td>Below 5</td>
<td>15,000</td>
<td>360</td>
</tr>
<tr>
<td>5 - 30</td>
<td>20,000</td>
<td>400</td>
</tr>
<tr>
<td>Above 30</td>
<td>10,000</td>
<td>280</td>
</tr>
<tr>
<td>Total</td>
<td>45,000</td>
<td>1,040</td>
</tr>
</tbody>
</table>
Solution.

Crude Death Rate = \( \frac{N}{P} \times 1000 \), where \( N = \text{No. of deaths}, \ P = \text{Population} \)

C.D.R. for town A = \( \frac{1,040}{45,000} \times 1,000 = 23.11 \)

C.D.R. for town B = \( \frac{2,280}{210,000} \times 1,000 = 22.80 \)

Standardised Death rate, taking population of town A as standard population:

<table>
<thead>
<tr>
<th>Age-Group (Years)</th>
<th>A</th>
<th></th>
<th></th>
<th>B</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Population</td>
<td>Deaths</td>
<td>Death Rate per Thousand</td>
<td>Population</td>
<td>Deaths</td>
<td>Death Rate per Thousand</td>
</tr>
<tr>
<td>Below 5</td>
<td>15,000</td>
<td>360</td>
<td>24</td>
<td>40,000</td>
<td>1,000</td>
<td>25</td>
</tr>
<tr>
<td>5-30</td>
<td>20,000</td>
<td>400</td>
<td>20</td>
<td>52,000</td>
<td>1,040</td>
<td>20</td>
</tr>
<tr>
<td>Above 30</td>
<td>10,000</td>
<td>280</td>
<td>28</td>
<td>8,000</td>
<td>240</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>45,000</td>
<td>1,080</td>
<td></td>
<td>100,000</td>
<td>2,280</td>
<td></td>
</tr>
</tbody>
</table>

Standardized Death Rate (town A)

\[
\frac{(15000 \times 24) + (20000 \times 20) + (10000 \times 28)}{15000 + 20000 + 10000} = \frac{360000 + 400000 + 280000}{45000} = \frac{1040000}{45000} = 23.11
\]

Standardized Death Rate (town B)

\[
\frac{(15000 \times 25) + (20000 \times 20) + (10000 \times 30)}{15000 + 20000 + 10000} = \frac{375000 + 400000 + 300000}{45000} = \frac{1075000}{45000} = 23.89
\]

We can now say that the death rate in town B is higher than in town A.

Another method of computing standardized death rate is to take some assumed population (that is the population of neither town A nor B) as standard. But this method is not popular.

**Illustration 2:** From the following table compare the death rates in two towns A and B. Which town is healthier?

<table>
<thead>
<tr>
<th>Age-group</th>
<th>Town A</th>
<th></th>
<th></th>
<th>Town B</th>
<th></th>
<th></th>
<th>Standard Population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Population</td>
<td>No of Deaths</td>
<td>Population</td>
<td>No of Deaths</td>
<td>Population</td>
<td>No. of Deaths</td>
<td></td>
</tr>
<tr>
<td>0-10</td>
<td>4000</td>
<td>36</td>
<td></td>
<td>3000</td>
<td>30</td>
<td></td>
<td>2000</td>
</tr>
<tr>
<td>20-25</td>
<td>12,000</td>
<td>48</td>
<td></td>
<td>20000</td>
<td>100</td>
<td></td>
<td>8000</td>
</tr>
<tr>
<td>25-60</td>
<td>6000</td>
<td>60</td>
<td></td>
<td>4000</td>
<td>48</td>
<td></td>
<td>6000</td>
</tr>
<tr>
<td>60 and over</td>
<td>8000</td>
<td>152</td>
<td></td>
<td>3000</td>
<td>60</td>
<td></td>
<td>4000</td>
</tr>
</tbody>
</table>
Solution.

### COMPARING DEATH RATES IN TOWNS A and B

<table>
<thead>
<tr>
<th>Age-Group</th>
<th>Town A</th>
<th>Town B</th>
<th>Standard Population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Population</td>
<td>Deaths</td>
<td>Death Rate</td>
</tr>
<tr>
<td>Below 10</td>
<td>4000</td>
<td>36</td>
<td>9</td>
</tr>
<tr>
<td>10-25</td>
<td>12000</td>
<td>48</td>
<td>4</td>
</tr>
<tr>
<td>25-60</td>
<td>6000</td>
<td>60</td>
<td>10</td>
</tr>
<tr>
<td>60 and Above</td>
<td>8000</td>
<td>152</td>
<td>19</td>
</tr>
</tbody>
</table>

Standardized death rate (Town A)  
\[
\text{S.D.R. (Town A)} = \frac{(2000 \times 9) + (4000 \times 4) + (6000 \times 10) + (4000 \times 9)}{20000} = \frac{186000}{20000} = 9.3
\]

Standardized Death Rate (town B)  
\[
\text{S.D.R. (Town B)} = \frac{(2000 \times 10) + (8000 \times 5) + (6000 \times 12) + (4000 \times 20)}{20000} = \frac{212000}{20000} = 10.69
\]

Death rate is lower in town A as compared to B, hence town A is healthier. Since the S.D.R. of town A is less than that of town B, hence town A is more healthy.

### Distinction between Crude Death Rate and Standardised Death Rate

1. Crude Death Rate is used as a measure of mortality of a population whereas Standardised Death Rate is used for comparison of mortality of two populations.

2. Crude Death Rate requires limited information like number of deaths and total population. Standardised Death Rate requires more information. It requires a standard population.

3. Crude Death Rate does not take into account the age composition of the population while Standardised Death Rate requires it.

### Age Specific Mortality Rates

The death rate calculated for a specified segment of population whose age is ‘n’ years in the last birth day is called age specific mortality rate. Different kinds of age specific mortality rates are:

#### Infant Mortality Rate

The most widely used definition of Infant mortality rate (IMR) is the number of deaths of babies under one year of age per 1,000 live births. The rate in a given region, therefore, is the total number of newborns dying under one year of age divided by the total number of live births during the year, then all multiplied by 1,000. The infant mortality rate is also called the infant death rate (per 1,000 live births). Infant mortality rates serve as one of the best indices to the general “healthiness” of a society. It is similar to age specific death rate for infants under 1 year of age. It is defined as:
Infant Mortality Rate = \frac{Number \ of \ deaths \ under \ 1 \ year \ of \ age \ which \ occurred \ among \ the \ population \ of \ a \ given \ geographic \ area \ during \ a \ given \ year}{Number \ of \ live \ births \ which \ occurred \ among \ the \ population \ of \ the \ given \ geographic \ area \ during \ the \ same \ year} \times 1000

The rate approximately measures for a given year the chances of a birth failing to survive one year of life. Still births are not included in the infant deaths. The rate can be calculated for males and females separately. The infant mortality rate varies considerably according to time and place. In countries with high standards of maternal and infant welfare it is as low as 15 to 20 per 1,000 but in some underdeveloped countries it is still well over 100 per 1,000. In many countries it has fallen spectacularly over the past sixty years or so. The male rate is appreciably higher than the female rate.

The infant mortality rate is of great value in the field of public health and its correct computation and interpretation is important. In most countries, the great risk of death at age under 1 is not equaled again in the life span until very old age is reached. But in contrast to deaths at old age, infant deaths are more responsive to improvement in environmental and medical conditions.

**Neo-Natal Mortality Rate**

The neo-natal mortality rate, like the infant mortality rate, is similar to an age specific rate. It is a rate used to measure the risk of death during the first month of life. This rate is defined as:

\[
\text{Annual Neo-natal Mortality rate} = \frac{\text{Annual deaths of infants under the age of 1 month among the population of given geographic area}}{\text{Number of live births which occurred among population of a given geographic area during the same year}}
\]

The rate measures for a given year the chance of a birth failure to survive one month of life. Most infant deaths occur within the first month of birth. The neo-natal mortality rate represents to a very large extent hard core of infant mortality. Of the neo-natal deaths more occur within the first month of birth of birth. The neo-natal mortality rate represents to a very large extent hard core of infant mortality. Of the neo-natal, deaths more occur within the first week of life.

Interpreting neo-natal rates, care must be taken to evaluate the probable effect of under-registration of live births in relation to infant deaths. It is likely that infant deaths, under 1 month, are registered less completely than any other infant deaths and the two sources of incompetencies in the rate probably compensate for each other to some extent.

**Maternal Mortality Rate**

The risk of dying from causes associated with child-birth is measured by the maternal mortality rate. For this purpose the deaths used in the numerator are those arising from puerperal causes. I.e. deliveries and complications of pregnancy, child-birth and the puerperium. The numbers exposed to the risk of dying from puerperal causes are women have been pregnant during the period. Their number being unknown the number of live births is used as the conventional base for computing comparable maternal mortality rates. The formula is:

\[
\text{Annual maternal mortality} = \frac{\text{Number of deaths from puerperal causes occurred among the female population of a given geographic area during a given year}}{\text{Number of live births which occurred among the population of the given geographic area during the same year}}
\]
The classification and coding of deaths as puerperal deaths vary from one country to another or even within the same country and hence we must be cautious in comparing maternal mortality rates for different places.

**GROWTH OF POPULATION**

**Crude rate of natural increase of population**

Crude Rate of natural increase of population in a country is the net result of difference between birth and deaths. Crude rate of natural increase = birth rate – death rate. Birth rates and death rates are expressed in 1000.

**Intercensus Estimation of Population (Estimation of Population between Two Censuses)**

The population is not counted every year but only once in 10 years. But it becomes necessary to know population for a particular year, known as mid-year estimated population. These estimates of the population are made from the preceding census figures. The three important methods of estimating population for the inter-censal and post censal years are i. Natural increase method, Arithmetic Progression Method and Geometric Method or Compound Interest method.

**Natural increase method**

Population in a year ‘1’ = Population in the year ‘0’ + (Births – Deaths) + (Immigration – Emigration) in the year 0.

\[ P_1 = P_0 + (B - D) + (I - E) \]

**Arithmetic Progression Method**

It is assumed that the population increase in A.P year by year by comparing the figures of any two census years. That is the increase in the population is assumed to be constant. The formula for estimating population for any day is given by

\[ P_e = P_1 = \frac{N}{N} (P_2 - P_1) \]

Where \( P_e \) required population estimate, \( P_1 \) and \( P_2 \) are population determined in two census, preceding and succeeding, \( N \) = Number of years between two census, \( n \) = number of years between the dates of \( P_1 \) and \( P_e \).

**Geometric Method (Compound Interest method)**

If we assume the rate of growth of population year by year, as constant, instead of the increase between two census year populations to be constant, the procedure is known as Geometric Progression Method of estimating population. So in G.P method, we assume a constant rate of increase of population over the previous year. Population at the \( n^{th} \) year is

\[ P_n = P_0 \left(1 + \frac{r}{100}\right)^n \]

where \( P_0 = \) Population at year ‘n’.

\( P_0 = \) initial population

\( r = \) rate of growth of population

\( n = \) number of years

**Sex Ratio**

**Sex ratio** is the ratio of males to females in a population. This is defined as the total number of females living out of 1000 males living in a specified locality or region.
Sex Ratio = \( \frac{\text{Total Number of females}}{\text{Total Number of Males}} \times 1000 \)

According to 2001 census, the sex ratio in Kerala is 1,058:1 and at national level it is 0.933:1. That means, in Kerala, out of 1000 males the number of females is 1058. But at national level, there are only 933 females out of 1000 males.

**Couple Protection Ratio**

The couple protection rate (CPR) is usually expressed as the percentage of women in the age group of 15-49 years, protected from pregnancy/child birth in the year under consideration for a specific area. So When the Couple Protection Rate is going up Birth rate must necessarily fall. Among the important factors influencing CPR are available methods of Birth Control, distribution of supply, service and follow-up centres and their staffing patterns. Couple Protection Ratio (CPR) or Couple Protection Rate is usually expressed as the percentage of women in the age group 15-49 years protected from pregnancy or child birth in the year under consideration for specific area. CPR is 72 % in Kerala. At the national level the CPR is 52 %as per the Health Development Indicators for the year 2007.

**Solved Problems**

**Example 1.** The Population of a town in Karnataka as per 1981 census was 37.043 millions with birth rate and death rate of 34.3 and 12.7 respectively. What were the total number of births and deaths for 1981?

Ans: Given B.R = \( \frac{B}{P} \times 1000 \) ie 34.3 = \( \frac{B}{37043000} \times 1000 \)

Therefore B = 34.3 x 37043 = 12,70,575

Therefore number of births = 1270575.

Given, DR = \( \frac{D}{P} \times 1000 \) ie 12.7 = \( \frac{D}{37043000} \times 1000 \)

Therefore, D = (12.7) (37043) = 4,70,446

Therefore number of deaths = 470446.

**Example 2.** Given the following data, estimate the population of the locality at the end of 1982.

| i. Population at the end of 1981 | : 3,42,14,320 |
| ii. Number of Births in 1982 | : 82,000 |
| iii. Number of Deaths in 1982 | : 15,000 |
| iv. Number of immigrants in 1982 | : 23,000 |
| v. Number of Emigrants in 1982 | : 18,000 |

**Solution:**

Here, \( P_0 = 3,42,14,320 \), \( B = 82,000 \), \( D = 15,000 \), \( I = 23,000 \), \( E = 18,000 \)

Therefore, the estimate of population at the end of 1982 is

\[
P_1 = P_0 + (B - D) + (I - E)
= 3, 42, 14,320 + (82,000 - 15,000) + (23,000 - 18,000)
= 3, 42, 86,320.
\]
Example 3: The mid-year population of a city in year was 4,80,500. If there were 10420 births and 6120 deaths in the year in the city, compute the Crude Birth Rate and Crude Death Rate.

Solution

The Crude Birth Rate is:

\[ CBR = \frac{\text{Total Number of births occurring in the year}}{\text{Total population at the mid point of the year}} \times 1000 \]

\[ = \frac{10420}{480500} \times 1000 = 21.69 \]

The Crude Death Rate is:

\[ CDR = \frac{\text{Number of live births occurring in the year}}{\text{Total population of the given locality at the mid point of the year}} \times 1000 \]

\[ = \frac{6120}{480500} \times 1000 = 12.74 \]

Example 4: The following table gives the age and sex distribution and the number of live births in a population.

<table>
<thead>
<tr>
<th>Age (Years)</th>
<th>0-89</th>
<th>19-20</th>
<th>20-24</th>
<th>25-29</th>
<th>30-39</th>
<th>40-59</th>
<th>60 and above</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>13,470</td>
<td>10,342</td>
<td>9,210</td>
<td>7,912</td>
<td>5,915</td>
<td>4,343</td>
<td>6,433</td>
</tr>
<tr>
<td>Female</td>
<td>12,130</td>
<td>9,942</td>
<td>9,013</td>
<td>7,913</td>
<td>5,910</td>
<td>4,413</td>
<td>6,942</td>
</tr>
<tr>
<td>Number of live births to females</td>
<td>0</td>
<td>294</td>
<td>527</td>
<td>632</td>
<td>312</td>
<td>36</td>
<td>0</td>
</tr>
</tbody>
</table>

a. Find the Crude Birth Rate

b. Find the age Specific Fertility Rates for the age group 20-24 years and 20 - 39 years.

Solution

<table>
<thead>
<tr>
<th>Age Group (Years)</th>
<th>Population</th>
<th>Number of births to Females</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td>0-9</td>
<td>13,470</td>
<td>12,130</td>
</tr>
<tr>
<td>19-20</td>
<td>10,342</td>
<td>9,942</td>
</tr>
<tr>
<td>20-24</td>
<td>9,210</td>
<td>9,013</td>
</tr>
<tr>
<td>25-29</td>
<td>7,912</td>
<td>7,913</td>
</tr>
<tr>
<td>30-39</td>
<td>5,915</td>
<td>5,910</td>
</tr>
<tr>
<td>40-59</td>
<td>4,343</td>
<td>4,413</td>
</tr>
<tr>
<td>60 and above</td>
<td>6,433</td>
<td>6,942</td>
</tr>
<tr>
<td>Total</td>
<td>57,625</td>
<td>56,263</td>
</tr>
</tbody>
</table>
a) Here, total male and female populations are 57,625 and 56,263 respectively. Therefore, total population is $57,625 + 56,263 = 113,888$. Total number of births is 1,801. Therefore, Crude Birth Rate is

$$\text{CBR} = \frac{\text{Number of live births occurring in the year}}{\text{Total population at the mid point of the year}} \times 1000$$

$$= \frac{1801}{113888} \times 1000 = 15.81$$

b) Age specific Fertility Rate for the age group 20-24 years is

$$\text{ASRF} = \frac{\text{Number of live births in the year to females aged between 20 and 24}}{\text{Total number of females aged between 20 and 24}} \times 1000$$

$$= \frac{527 \times 1000}{9013} = 58.47$$

Age Specific Fertility Rate for the age group 20-9 years is

$$\text{ASRF} = \frac{\text{Number of live births in the year to females aged between 20 and 30}}{\text{Total number of females aged between 20 and 29}} \times 1000$$

$$= \frac{(527+632+312)}{9013+791+5910} \times 1000 \approx \frac{1471}{22836} \times 1000 = 64.42$$

Example 5 The population of a village in Calicut on 1-7-2001 was 30,000. The vital events for the same year were:

The number of births: 1200, number of deaths: 495, number of infant death: 110. Compute the birth, death and infant mortality rates. Also compute the growth rate.

**Ans:**

- Birth Rate = $\frac{1200}{30000} \times 1000 = 40$ per thousand
- Death Rate = $\frac{495}{30000} \times 1000 = 16.5$ per thousand
- Infant Mortality Rate = $\frac{110}{1200} \times 1000 = 91.66$ per thousand
- Growth Rate = $\frac{1200-495-110}{1000} \times 100 = \frac{40-16.5}{1000} \times 1000 = 2.35\%$

Example 6 Calculate Age Specific Death Rates.

<table>
<thead>
<tr>
<th>Age (Years)</th>
<th>Population</th>
<th>Number of Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below 10</td>
<td>30,000</td>
<td>420</td>
</tr>
<tr>
<td>10-19</td>
<td>20,000</td>
<td>150</td>
</tr>
<tr>
<td>29-29</td>
<td>25,000</td>
<td>125</td>
</tr>
<tr>
<td>30-39</td>
<td>8,000</td>
<td>70</td>
</tr>
<tr>
<td>40-49</td>
<td>2,500</td>
<td>25</td>
</tr>
<tr>
<td>50 &amp; Above</td>
<td>2,000</td>
<td>30</td>
</tr>
</tbody>
</table>
Solution

Here, the Age Specific Death Rates are found by using the formula –

\[ ASDR = \frac{\text{Number of deaths occurring in the age group in the year}}{\text{Total population in the age group in the year}} \times 1000 \]

<table>
<thead>
<tr>
<th>Age</th>
<th>Population</th>
<th>Number of Deaths</th>
<th>ASDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below 10</td>
<td>30,000</td>
<td>420</td>
<td>14</td>
</tr>
<tr>
<td>10-19</td>
<td>20,000</td>
<td>150</td>
<td>7.5</td>
</tr>
<tr>
<td>20-29</td>
<td>25,000</td>
<td>125</td>
<td>5</td>
</tr>
<tr>
<td>30-39</td>
<td>8,000</td>
<td>70</td>
<td>8.75</td>
</tr>
<tr>
<td>40-49</td>
<td>2,500</td>
<td>25</td>
<td>10</td>
</tr>
<tr>
<td>50 &amp; Above</td>
<td>2,000</td>
<td>30</td>
<td>15</td>
</tr>
</tbody>
</table>

Example 7 Calculate Crude Death Rate and Standardised Death Rate from the following data.

<table>
<thead>
<tr>
<th>Age (Years)</th>
<th>Population (Thousand)</th>
<th>No. of Deaths</th>
<th>Standard Population (Percentage Distribution)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-9</td>
<td>21</td>
<td>350</td>
<td>22</td>
</tr>
<tr>
<td>10-24</td>
<td>30</td>
<td>102</td>
<td>30</td>
</tr>
<tr>
<td>25-44</td>
<td>37</td>
<td>229</td>
<td>28</td>
</tr>
<tr>
<td>45-64</td>
<td>17</td>
<td>54</td>
<td>15</td>
</tr>
<tr>
<td>65 &amp; Above</td>
<td>5</td>
<td>415</td>
<td>5</td>
</tr>
</tbody>
</table>

Solution

<table>
<thead>
<tr>
<th>Age</th>
<th>Population</th>
<th>Deaths</th>
<th>A = ASDR</th>
<th>Standard Population (P)</th>
<th>PA</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-9</td>
<td>21000</td>
<td>350</td>
<td>16.67</td>
<td>22</td>
<td>366.74</td>
</tr>
<tr>
<td>10-24</td>
<td>30000</td>
<td>102</td>
<td>3.40</td>
<td>30</td>
<td>102.00</td>
</tr>
<tr>
<td>25-44</td>
<td>37000</td>
<td>229</td>
<td>6.19</td>
<td>28</td>
<td>173.32</td>
</tr>
<tr>
<td>45-64</td>
<td>17000</td>
<td>54</td>
<td>20.82</td>
<td>15</td>
<td>312.30</td>
</tr>
<tr>
<td>65 &amp; Above</td>
<td>5000</td>
<td>415</td>
<td>83.00</td>
<td>5</td>
<td>415.00</td>
</tr>
<tr>
<td>Total</td>
<td>110000</td>
<td>1450</td>
<td></td>
<td>100</td>
<td>1369.36</td>
</tr>
</tbody>
</table>
The Crude Death Rate is:

\[ \text{CDR} = \frac{\text{Number of deaths in the year}}{\text{Total population at the mid point of the year}} \times 1000 \]

\[ = \frac{1450}{10000} \times 1000 = 13.18 \]

The Standardised Death Rate is:

\[ \text{STDR} = \frac{\sum \text{PA}}{\sum \text{P}} = \frac{1369.6}{100} = 13.6936 = 13.6936 \]

**Example 8:** From the data given below, calculate 1. Gross Reproduction Rate (GRR) and 2. Net Reproduction Rate (NRR).

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Female Population (000)</th>
<th>Female Births</th>
<th>Survival Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-19</td>
<td>1400</td>
<td>15130</td>
<td>.969</td>
</tr>
<tr>
<td>20-24</td>
<td>1420</td>
<td>94150</td>
<td>.967</td>
</tr>
<tr>
<td>25-29</td>
<td>1520</td>
<td>102670</td>
<td>.963</td>
</tr>
<tr>
<td>30-34</td>
<td>1750</td>
<td>72490</td>
<td>.958</td>
</tr>
<tr>
<td>.35-39</td>
<td>1450</td>
<td>31400</td>
<td>.952</td>
</tr>
<tr>
<td>40-44</td>
<td>1690</td>
<td>10640</td>
<td>.942</td>
</tr>
<tr>
<td>45-49</td>
<td>1670</td>
<td>700</td>
<td>.928</td>
</tr>
</tbody>
</table>

Ans:

<table>
<thead>
<tr>
<th>Age Group</th>
<th>FP</th>
<th>FB</th>
<th>(SFR)' (FB ÷ FP)</th>
<th>S</th>
<th>(SFR)' x S</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-19</td>
<td>1400</td>
<td>15130</td>
<td>.0108</td>
<td>.969</td>
<td>.0105</td>
</tr>
<tr>
<td>20-24</td>
<td>1420</td>
<td>94150</td>
<td>.0663</td>
<td>.967</td>
<td>.0641</td>
</tr>
<tr>
<td>25-29</td>
<td>1520</td>
<td>102670</td>
<td>.0675</td>
<td>.963</td>
<td>.0650</td>
</tr>
<tr>
<td>30-34</td>
<td>1750</td>
<td>72490</td>
<td>.0414</td>
<td>.958</td>
<td>.0397</td>
</tr>
<tr>
<td>.35-39</td>
<td>1450</td>
<td>31400</td>
<td>.0217</td>
<td>.952</td>
<td>.0207</td>
</tr>
<tr>
<td>40-44</td>
<td>1690</td>
<td>10640</td>
<td>.0063</td>
<td>.942</td>
<td>.0059</td>
</tr>
<tr>
<td>45-49</td>
<td>1670</td>
<td>700</td>
<td>.0004</td>
<td>.928</td>
<td>.0004</td>
</tr>
</tbody>
</table>

Here, c = class interval = 5

\[ \text{GRR} = \frac{\sum (SFR)}{c} \times c = .2144 \times 5 = 1.0720 = 1.072 \]

\[ \text{NRR} = \frac{\sum (SFR) \times S}{c} \times c = .2063 \times 5 = 1.0315 = 1.032 \]

*****