SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – I – INTRODUCTION TO STATISTICS – SMT1304

I. INTRODUCTION TO STATISTICS

Origin and development of statistics – Definition of statistics-Importance and Scope of Statistics – Limitations of statistics – Misuse of statistics. Presentation of Data-Diagrammatic representation of data – Bar diagrams – Pie diagrams – histogram- Frequency Polygon and frequency curve – Pictogram and Cartogram.

History

Statistics is a very well-known term in the history, be it ancient or medieval. However, there are still a few unanswered questions. One such question is – "origin of the word 'statistics'." There are several views related to the same. One such view is that it has a Latin origin and the word that it comes from is 'status.' On the contrary, another view speaks of its Italian origin and that it comes from 'statista.' According to scholars, the origin is German and the word it comes from is 'statistik.' Similarly, according to more suggestion, the origin is traced back to a French word called 'statistique.' In the past, statistics was all about "collection" of data. Also, the goal was to maintain the data for the welfare of everyone in the area. According to various calculations, there were several predictions that led to one or the other answer. Statistics play a very vital role in any domain. It helps in collecting data, be it in any field. Along with that, it also helps in analyzing data using statistical techniques.

What is Statistics?

Statistics can come forward in two ways: singular and plural. In plural form, statistics is quantitative as well as qualitative. In the plural sense, data is generally taken into account keeping in mind the statistical analysis. Singularly, it is more like a scientific method that helps in presenting, collecting, as well as analyzing data. All of this brings some major characteristics into the limelight.

Statistics is the study of the collection, organisation, analysis, interpretation and presentation of data. It is built up from the field of mathematics known as probability. Probability gives us a way to determine how likely an event is to occur. It also gives us a way to talk about randomness. It can be used in every field of scientific research, such as psychology, economics, medicine, advertising, demography and many more. Statistical course will teach students on the basic concepts of logic, mathematics, statistical reasoning, analyse data, evaluate data and research methods.

Basically, there are two branches of statistics. They are – Descriptive and Inferential. Here is a brief knowledge of both of the branches.

- **Descriptive:** This branch deals with the basic and major aspects related to numeric. The numeric's and data contain graphs, tables, and many more quantities. These quantities help with serving information.
- **Inferential:** This branch deals with making inferences about the large data group. The knowledge for making inferences generally comes from samples. Sample evidence brings out inferences.

What is the importance of statistics?

- **Statistical** knowledge helps to use the proper methods to collect the data, employ the correct analyses, and effectively present the results. **Statistics** is a crucial process behind how we make discoveries in science, make decisions based on data, and make predictions.

The scopes of the statistics are as follows:

- Statistics being indispensable in the modern world has been of utmost use to the government as they are using statistics constantly researching to improve the economic development of countries.
- Statistics in the industry are widely used for equality control.
- In education also statistics are widely used because now research has become a common feature in all branches of activities and studies.
- In the field of Medical sciences too statistical tools play a very vital role, for example, it is used to test the efficiency of a new drug or medicine.

Statistics is indispensable in this modern age aptly termed as "the age of planning". The governments of most countries around the world are constantly researching to improve its economic development. Statistical data and techniques of statistical analysis are immensely useful in solving economical problems such as wages, price, time series analysis, demand analysis. It is an irreplaceable tool of production control. Business executives are relying more and more on statistical techniques for studying the preference of the customers. Industry statistics are widely used in equality control. In production engineering, statistical tools such as inspection plan, control chart etc. are extensively used to find out whether the product is confirming to the specifications or not. Statistics are useful to banker, insurance companies, social workers, labour unions, trade associations, chambers and to the politicians.

Limitations:

Limitations come a lot before directly applying the statistical methods. It is necessary to be aware of it in order to move ahead. Some of the primary limitations of statistics are:

- Statistics is all about “aggregates.” Be it an individual or a statistician, they are all a part of the aggregate.
- It also deals with quantitative data. However, it is not a very difficult task to do a conversion from qualitative to quantitative. All that is needed is the numerics and description related to the qualitative data.
- In order to propose specific projections, i.e. sales, price, quantity and so on, there is a requirement of a set of conditions. So, if, by any chance, these conditions turn out to be wrong or are violated, there is a chance that the projections and its outcome will be inaccurate.
- Statistical inferences make use of random sampling options. Hence, not following the rules for sampling would be a very bad idea as it can lead to wrong results. The conclusions coming off would have errors. So, the idea here is to consult the experts before hopping into the sampling scheme, directly.

Misuse of Statistics:

Statistics, when used in a misleading fashion, can trick the casual observer into believing something other than what the data shows. That is, a misuse of statistics occurs when a statistical argument asserts a falsehood. In some cases, the misuse may be accidental. In others, it is purposeful and for the gain of the perpetrator. When the statistical reason involved is false or misapplied, this constitutes a statistical fallacy.

The false statistics trap can be quite damaging for the quest for knowledge. For example, in medical science, correcting a falsehood may take decades and cost lives.

Misuses can be easy to fall into. Professional scientists, even mathematicians and professional statisticians, can be fooled by even some simple methods, even if they are careful to check everything.

Many misuses of statistics occur because:

- The source is a subject matter expert, not a statistics expert. The source may incorrectly use a method or interpret a result.
- The source is a statistician, not a subject matter expert. An expert should know when the numbers being compared describe different things. Numbers change, as reality does not, when legal definitions or political boundaries change.
- The subject being studied is not well defined.
- Data quality is poor.

Presentation of Data- Diagrammatic representation of data

Diagrammatic Presentation of Data gives an immediate understanding of the real situation to be defined by data in comparison to the tabular presentation of data or textual representations. Diagrammatic presentation of data translates pretty effectively the highly complex ideas included in numbers into more concrete and quickly understandable form. Diagrams may be less certain but are much more efficient than tables in displaying the data.

Concept of Diagrammatic Presentation

- Diagrammatic presentation is a technique of presenting numeric data through Pictograms, Cartograms, Bar Diagrams & Pie Diagrams etc. It is the most attractive and appealing way to represent statistical data. Diagrams help in visual comparison and have a bird's eye view.
- Under Pictograms, we use pictures to present data. For example, if we have to show the production of cars, we can draw cars. Suppose, production of cars is 40,000. We can show it by a picture having four cars, where 1 Car represents 10,000 units.
- Under Cartograms, we make use of maps to show the geographical allocation of certain things.
- Bar Diagrams are rectangular in shape placed on the same base. Their height represents the magnitude/value of the variable. Width of all the bars and gap between the two bars is kept the same.
- Pie Diagram is a Circle which is sub-divided or partitioned to show the proportion of various components of the data.

Advantages of Diagrammatic Presentation

(1) Diagrams Are Attractive and Impressive:

Data presented in the form of diagrams are able to attract the attention of even a common man.

(2) Easy to Remember

Diagrams have a great memorizing effect. The picture created in the mind by diagrams last much longer than those created by figures presented through the tabular form.

(3) Diagrams save Time

It presents complex mass data in a simplified manner. Data presented in the form of diagrams can be understood by the user very quickly.

(4) Diagrams Simplify Data

Diagrams are used to represent a huge mass of complex data in a simplified and intelligible form, which is easy to understand.

(5) Diagrams Are Useful in Making Comparisons

It becomes easier to compare two sets of data visually by presenting them through diagrams.

(6) More Informative

Diagrams not only depict the characteristics of data but also bring out other hidden facts and relations which are not possible from the classified and tabulated data.

Types of One-dimensional Diagram:

One dimensional diagram is that diagram in which the only length of the diagram is considered. It can be drawn in the form of a line or in various types of bars.

Following Are the Types of One-dimensional Diagram:

(1) Simple Bar Diagram

Simple Bar diagram comprises of a group of rectangular bars of equal width for each class or category of data.

(2) Multiple Bar Diagram

This diagram is used when we have to make a comparison between two or more variables like income and expenditure, import and export for different years, marks obtained in different subjects in different classes, etc.

(3) Sub-divided Bar Diagram

This diagram is constructed by sub-dividing the bars in the ratio of various components.

(4) Percentage Bar Diagram

Sub-divided bar diagram presented on a percentage basis is known as Percentage Bar Diagram.

(5) Broken-scale Bar Diagram

This diagram is used when the value of one observation is very high as compared to the others. In order to gain space for the smaller bars of the series, the largest bars may be broken. The value of each bar is written at the top of the bar.

(6) Deviation Bar Diagram

Deviation bars are used for representing net changes in data like Net Profit, Net Loss, Net Exports, Net Imports, etc.

Meaning of Pie Diagram:

A Pie Diagram is a circle divided into sections. The size of the section indicates the magnitude of each component as a part of the whole.

Steps Involved in Constructing Pie Diagram

1. Convert the given values in percentage form and multiply it with 3.6’ to get the amount of angle for each item.
2. Draw a circle and start the diagram at 12’O clock position.
3. Take the highest angle first with protector (D) and mark lower angles successively.
4. Shade different angles differently to show distinction in each item.

Bar diagram

There are two types of bar diagrams namely, Horizontal Bar diagram and Vertical bar diagram. While horizontal bar diagram is used for qualitative data or data varying over space, the vertical bar diagram is associated with quantitative data or time series data.

Bars i.e. rectangles of equal width and usually of varying lengths are drawn either horizontally or vertically.

We consider Multiple or Grouped Bar diagrams to compare related series. Component or sub-divided Bar diagrams are applied for representing data divided into a number of components. Finally, we use Divided Bar charts or Percentage.

Bar diagrams for comparing different components of a variable and also the relating of the components to the whole. For this situation, we may also use Pie chart or Pie diagram or circle diagram.

Example: 1

The total number of runs scored by a few players in one-day match is given.

PLAYERS	1	2	3	4	5	6
RUNS SCORED	30	60	10	50	70	40

Draw bar graph for the above data.

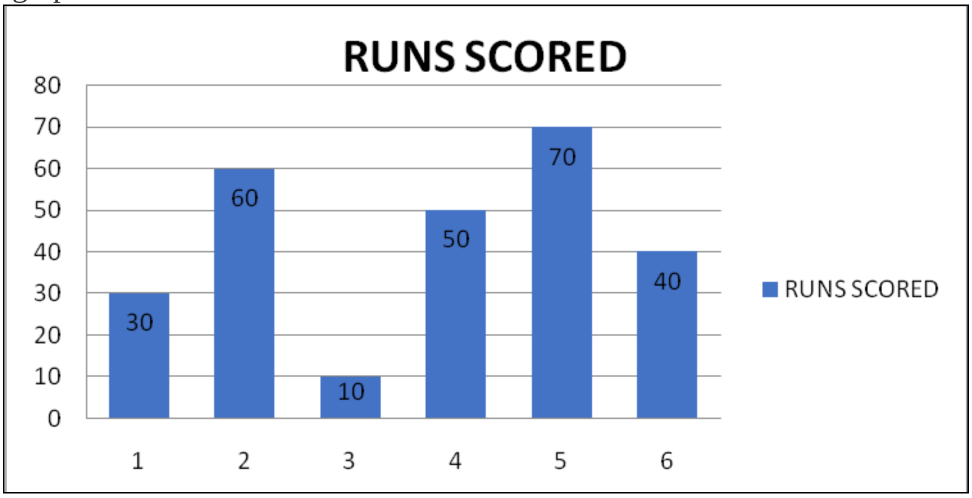

FIGURE: 1

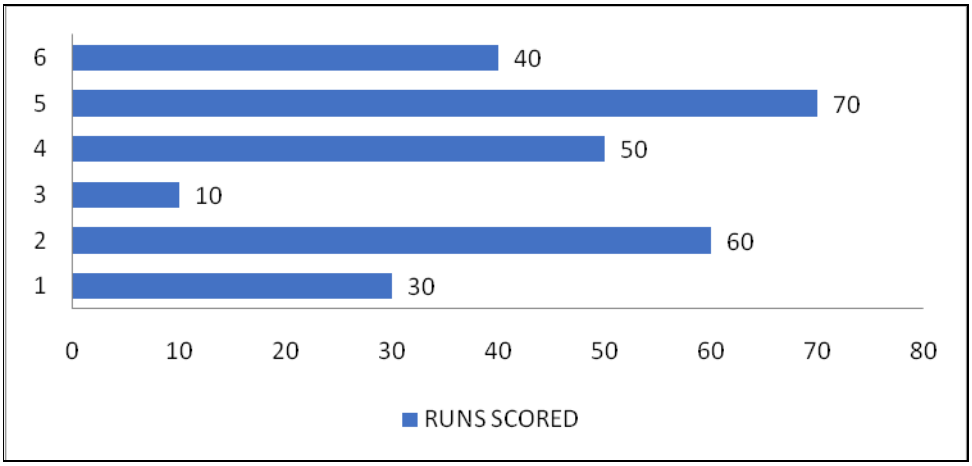

FIGURE: 2

Example: 2

The total number of runs scored by a few players in one-day match is given.

PLAYERS	1	2	3	4	5	6
RUNS SCORED INNINGS 1	30	60	10	50	70	40
RUNS SCORED INNINGS 2	42	50	50	35	40	15

Draw multiple bar graph for the above data.

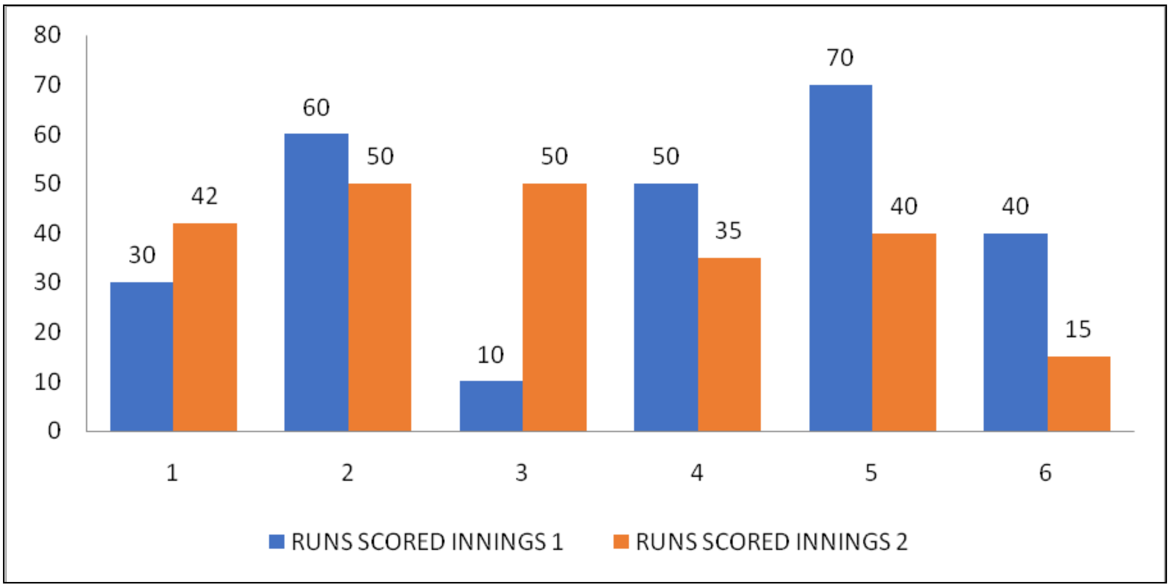

FIGURE: 3

Example: 3

The total number of runs scored by a few players in one-day match is given.

PLAYERS	1	2	3	4	5	6
RUNS SCORED INNINGS 1	30	60	10	50	70	40
RUNS SCORED INNINGS 2	42	50	50	35	40	15

Draw Component bar graph or Sub divided Bar graph for the above data.

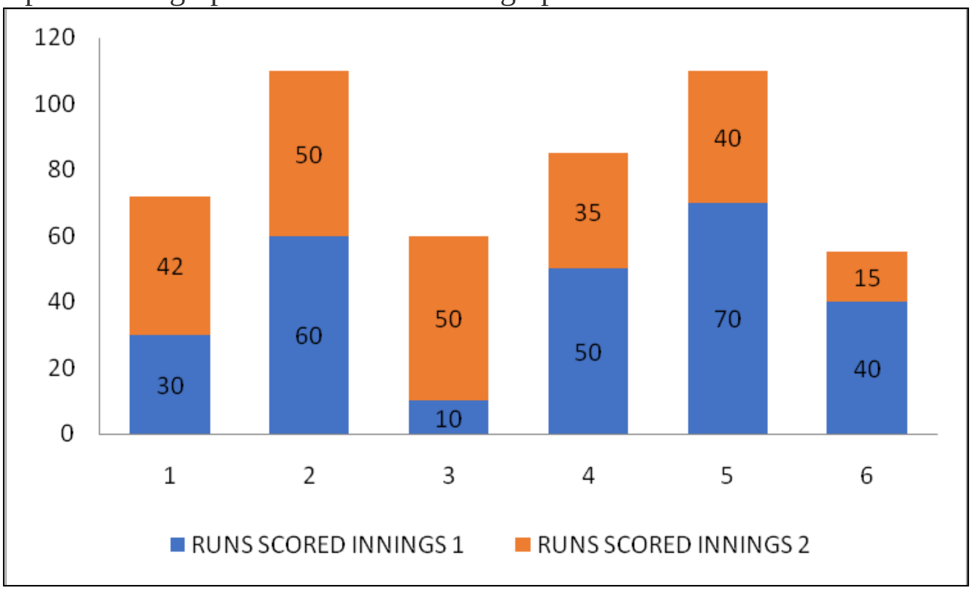

FIGURE: 4

Example: 4

The total number of runs scored by a few players in one-day match is given.

PLAYERS	1	2	3	4	5	6
RUNS SCORED INNINGS 1	30	60	10	50	70	40

Draw Pie Chart for the above data.

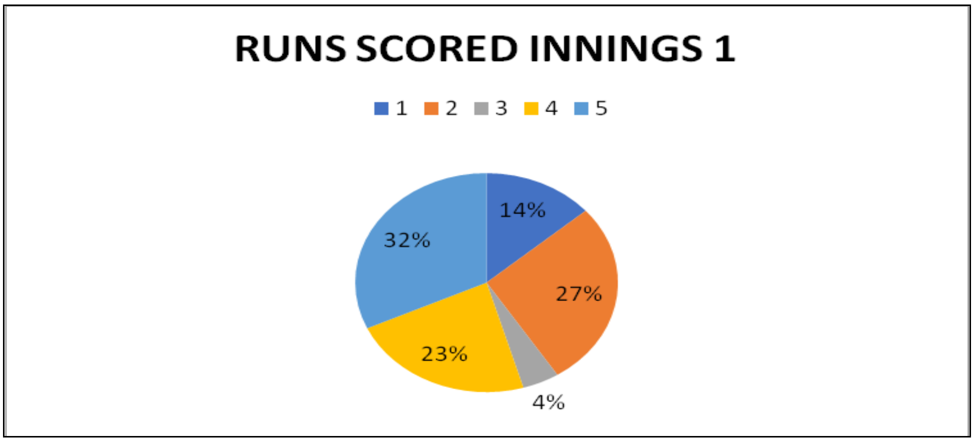

FIGURE: 5

Example: 5

Represent the following data by a percentage bar diagram.

Subjects	Number of Students	
	2016-17	**2017-18**
Statistics	25	30
Economics	40	42
History	35	28

Solution

Subject	**2016-17**		**2017-18**	
	Number of students (%)	**Cumulative Percentage**	**Number of students (%)**	**Cumulative Percentage**
Statistics	25	25	30	30
Economics	40	60	42	72
History	35	100	28	100

TABLE: 1

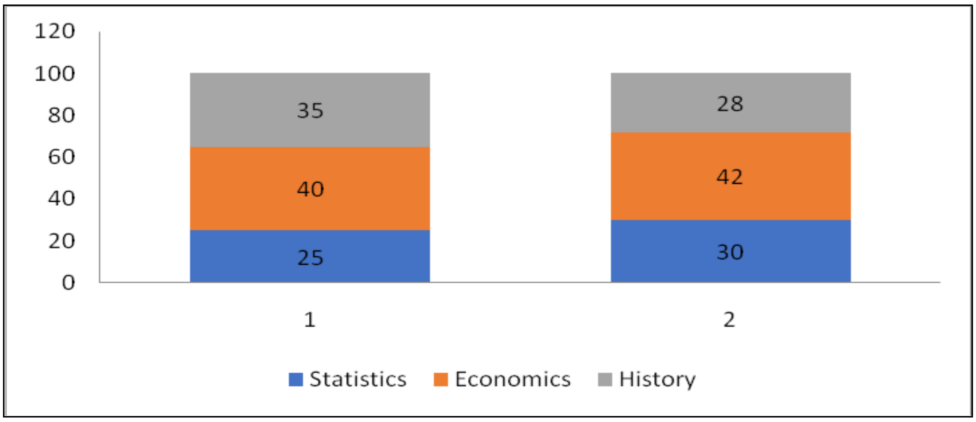

FIGURE: 6

Example: 6

Following are the data about the market share of four brands of TV sets sold in Panipat and Ambala. Present the data in the pie chart.

Brand of Sets	**Units sold in Panipat**	**Units sold in Ambala**
Samsung	480	625
Akai	360	500
Onida	240	438
Sony	120	312

Solution

Total sets sold in Place A and Place B are 1,200 and 1,875 respectively. Data are to be represented by two circles whose radii are in the ratio of square roots of total TV sets sold in each city in the ratio of : or 1:1. The calculations regarding the construction of the pie diagram are as follows.

Brands of Sets	Place A			Place B		
	Sets Sold	Sales(₹)	Sales in terms of components of 360°	Sets Sold	Sales %	Sales in terms of components of 360°
Samsung	480	40	40100×360°=144°	625	33.3	33.3100×360°=119.88°
Akai	360	30	30100×360°=108°	500	26.7	26.7100×360°=96.12°
Onida	240	20	20100×360°=72°	438	23.4	23.4100×360∘=84.24°
Sony	120	10	10100×360°=36°	312	16.6	16.6100×360°=59.76°
Total	1,200		360°	1,875		360°

TABLE: 2

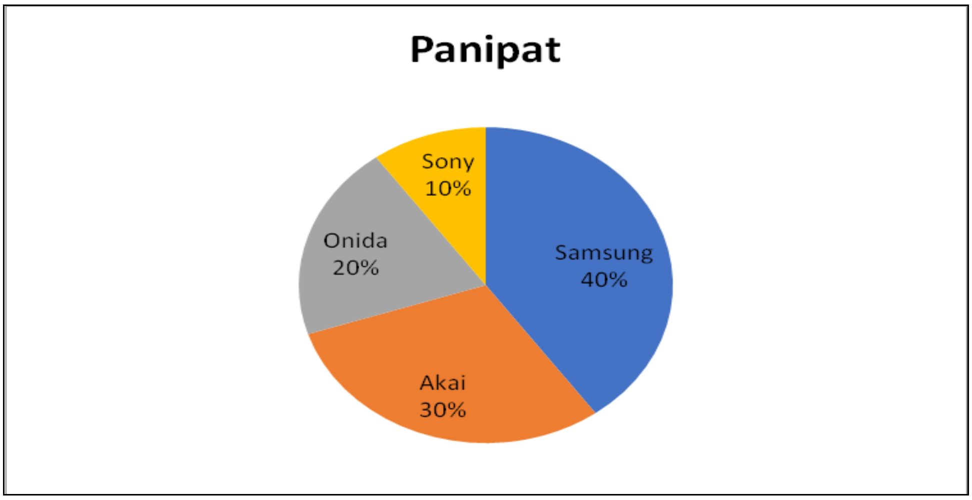

FIGURE: 7

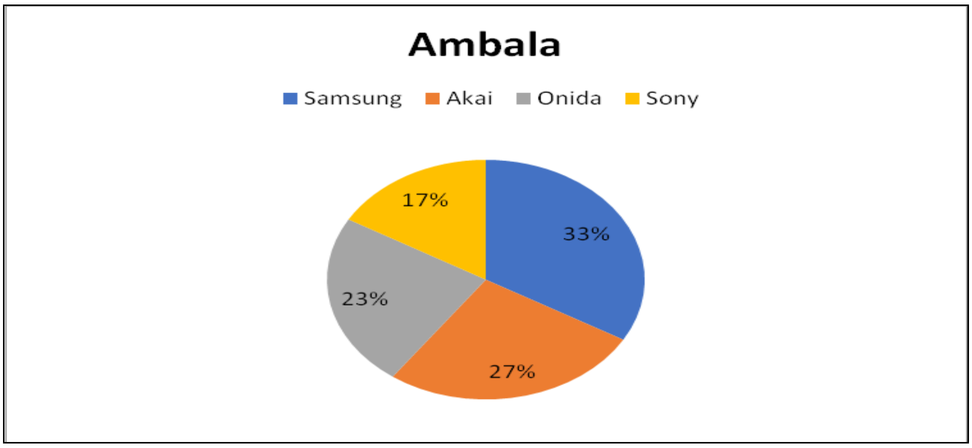

FIGURE: 8

HISTOGRAM

What is a histogram?

A histogram is a plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of continuous data. This allows the inspection of the data for its underlying distribution (e.g., normal distribution), outliers, skewness, etc. An example of a histogram, and the raw data it was constructed from, is shown below:

36	25	38	46	55	68	72	55	36	38
67	45	22	48	91	46	52	61	58	55

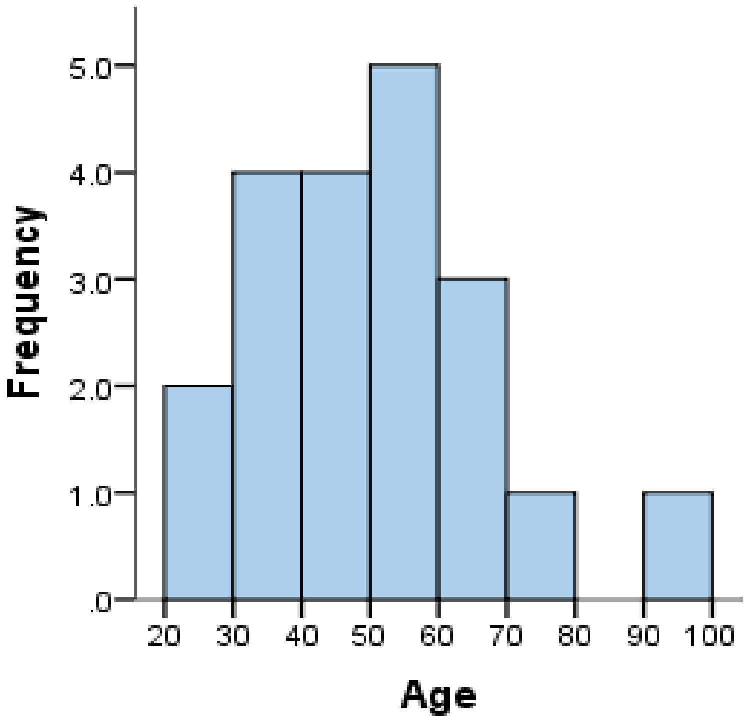

FIGURE: 9

Bin	Frequency	Scores Included in Bin
20-30	2	25,22

30-40	4	36,38,36,38
40-50	4	46,45,48,46
50-60	5	55,55,52,58,55
60-70	3	68,67,61
70-80	1	72
80-90	0	-
90-100	1	91

TABLE: 3

Example 7: The profit (in Rs crore) of a company from 1990-91 to 1999-2000 are given below:

YEAR	PROFIT	YEAR	PROFIT
1990-91	35.6	1995-96	87.2
1991-92	46.7	1996-97	113.1
1992-93	39.8	1997-98	123.6
1993-94	68.2	1998-99	119.7
1994-95	93.5	1999-2000	130.8

Represent this data by a simple bar diagram.

Solution: The simple bar diagram of the above data is given below:

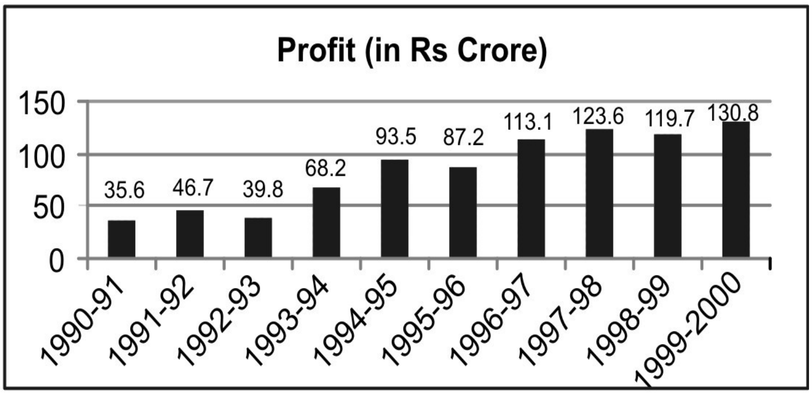

FIGURE: 10

Example 8: Represent the following data by subdivided bar diagram:

Category	Cost per chair (in Rs) year wise		
	1990	1995	1960
Cost of Raw Material	15	20	30
Labour Cost	15	18	25
Polish	5	6	15
Delivery	5	6	10

Solution: First of all we calculate the cumulative cost on the basis of the given amounts:

Category	Cost per chair (in Rs) year wise					
	1990	Cumulative Cost (in Rs)	1995	Cumulative Cost (in Rs)	1960	Cumulative Cost (in Rs)
Cost of Raw Material	15	15	20	20	30	30
Labour Cost	15	30	18	38	25	55
Polish	5	35	6	44	15	70

Delivery	5	40	6	50	10	80

TABLE: 4

On the basis of above table required subdivided bar diagram is given below:

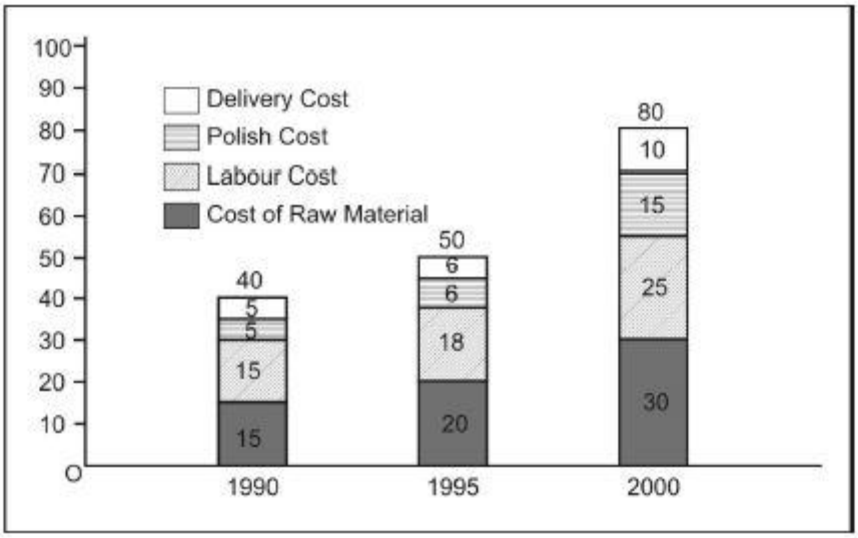

FIGURE: 11

Example 9: Draw the multiple bar diagram for the following data.

Year	Sale (in ,000 Rs)	Gross profit (in ,000 Rs)	Net profit (in, '000 Rs)
1990	100	30	10
1995	120	40	15
2000	130	45	25
2005	150	50	30
2010	200	70	30

Solution: Multiple bar diagram for the above data is given below.

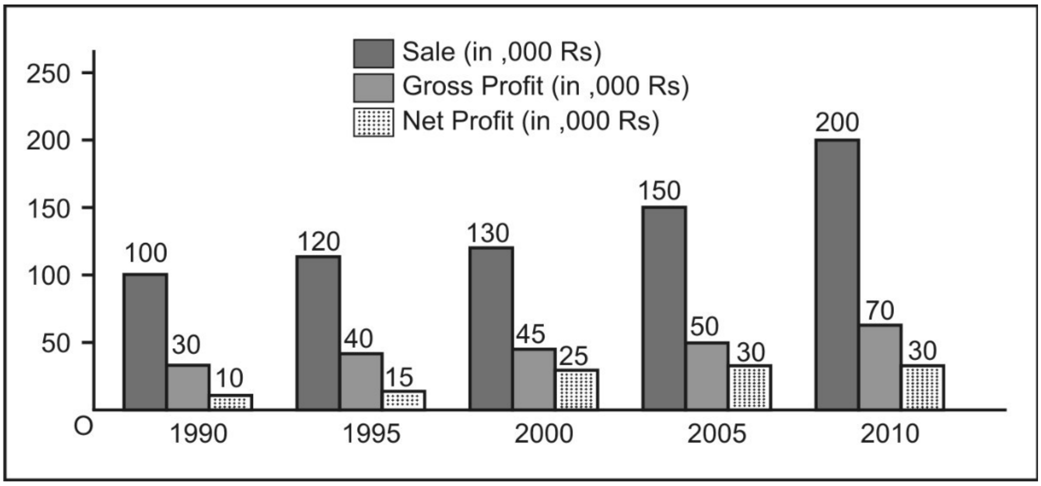

FIGURE: 12

Example 10: Draw a percentage bar diagram for the following data:

Category	Cost Per Unit (1990)	Cost Per Unit (2000)

Material	20	32
Labour	25	36
Delivery	5	12

Solution: First of all percentage and cumulative percentage are obtained for both the years in various category.

Category	Cost Per Unit (1990)	% Cost	Cumulative % Cost	Cost Per Unit (2000)	% Cost	Cumulative % Cost
Material	20	40	40	32	40	40
Labour	25	50	90	36	45	85
Delivery	5	10	100	12	15	100

TABLE: 5

On the basis of above table required percentage bar diagram is given below

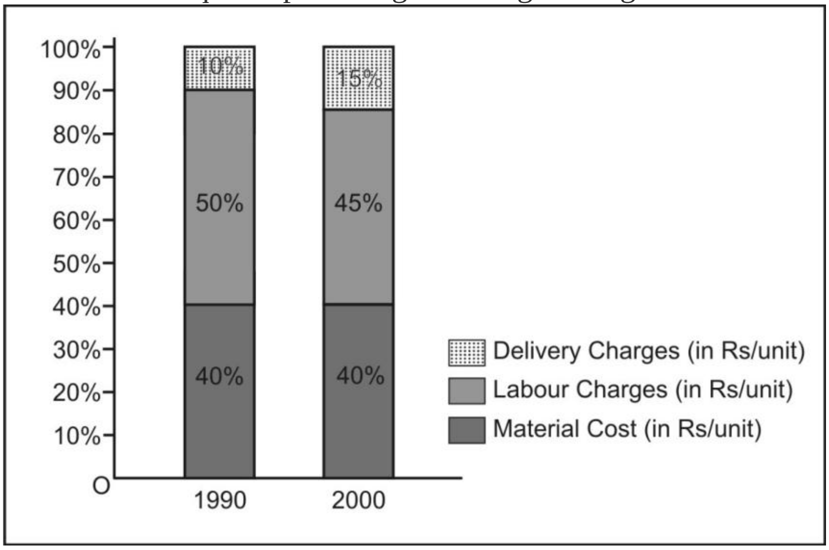

FIGURE: 13

Example 11: A company is started by the four persons A, B, C and D and they distribute the profit or loss between them in proportion of 4: 3: 2:1. In year 2010 company earned a profit of Rs 14400. Represent the shares of their profits in a pie chart.

Solution: Given ratio is 4: 3: 2:1; Sum of ratios = 4 + 3 + 2 + 1 = 10

Calculation of Degrees

Partners	Profits (in Rs)	Sector Angles (in degree)
A	14400 X 4 / 10 = 5760	5760 X 360 / 14400 = 144
B	14400 X 3 / 10 = 4320	4320 X 360 / 14400 = 108
C	14400 X 2 / 10 = 2880	2880 X 360 / 14400 = 72
D	14400 X 1 / 10 = 1440	1440 X 360 / 14400 = 36

TABLE: 6

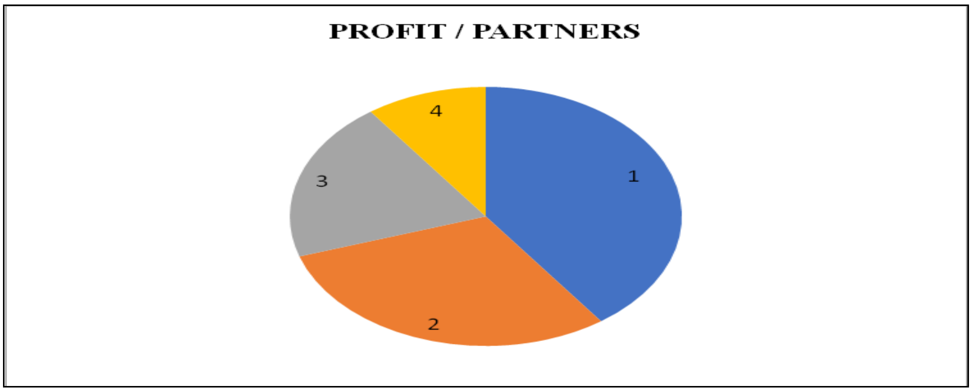

FIGURE: 14

PICTOGRAM

Pictograms, also known as picture grams, are very frequently used in representing statistical data. Pictograms are drawn with the help of pictures. These diagrams indicate towards the nature of the represented facts. Pictograms are attractive and easy to comprehend and as such this method is particularly useful in presenting statistics to the layman. The picture which is used as symbols to represent the units or values of any variable or commodity selected carefully. The picture symbol must be self explanatory in nature. For example, if the increase in number of Airlines Company is to be shown over a period of time then the appropriate symbol would be an aeroplane. The pictograms have the following merits:

(i) The magnitudes of the variables may be known by counting the pictures.

(ii) An illiterate person can also get the information.

(iii) The facts represented in a pictorial form can be remembered longer.

Example 11: Draw a pictogram for the data of production of tea (in hundred kg) in a particular area of Assam from year 2006 to 2010.

Year	2006	2007	2008	2009	2010
Production of Tea (in 100 kg.)	2.5	3.0	4.0	5.5	7.0

|**Solution:** Pictogram for the production of tea in a particular area of Assam from year 2006 to 2010 is shown below:

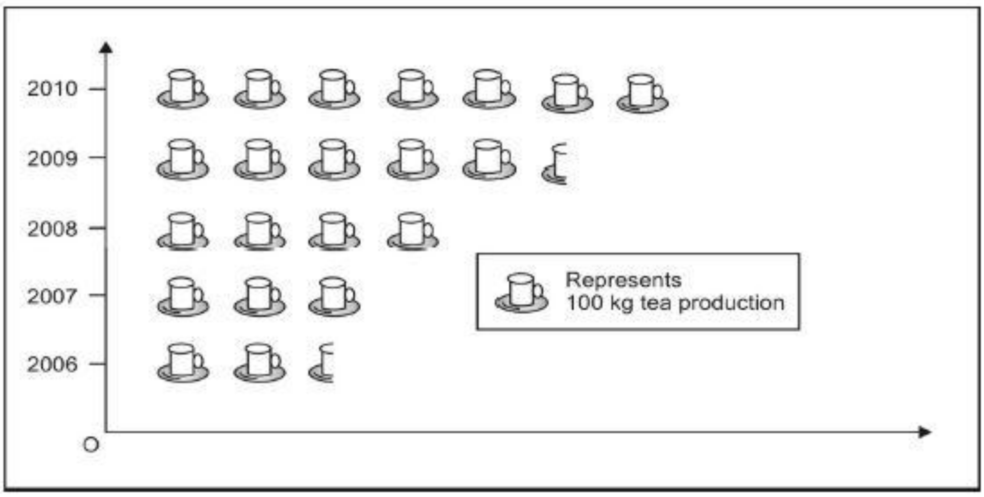

FIGURE: 15

CARTOGRAM

Representation of the numerical facts with the help of a map is known as cartogram. By representing the facts by maps, the impact of the results on different geographical area may be shown and to be compared also. Maps are helpful in comparative study of various districts of a state or different states of a country. For example, the production of wheat in different geographical areas can also be represented by cartogram. The quantities on the map can be shown in many ways, such as through shads or colours or by dots or by placing pictograms in each geographical area or by the appropriate numerical figure in each geographical area.

Example: 1 Cartogram of Germany, with the states and districts resized according to population
Example: 2 States and union territories of India on (Left) an equal-area map, and (Right) a Flow-Based Cartogram where areas are proportional to GDP.

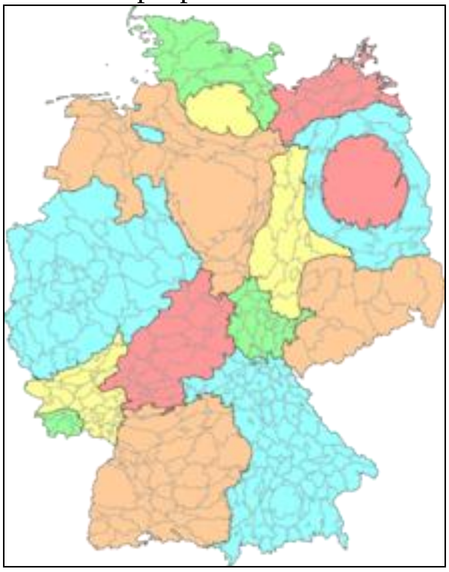

FIGURE: 16

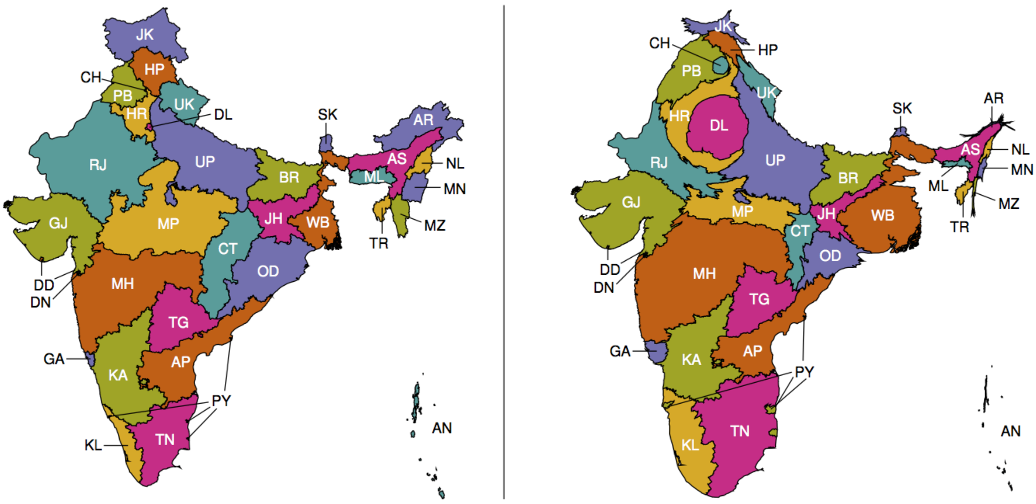

FIGURE: 17

REFERENCES:

1. Vittal.P.R., Business Statistics, Margham publications, Chennai, 2008.
2. Gupta.S.C. Statistics, Himalaya Publishers, Mumbai, 2005.
3. S.P.Gupta, Business Statistics, Sultan Chand & Sons, New Delhi, 2008.
4. Beri.G., Business Statistics, Tata McGraw Hill Publishing Company Ltd, New Delhi, 2009.
5. Yule and Kendall (1993), Introduction to theory of Statistics. Universal Book Stall, New Delhi.

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – II – MEASURES OF CENTRAL TENDENCY– SMT1304

MEASURES OF CENTRAL TENDENCY

Simple averages – mean, median, mode – Geometric mean and Harmonic mean – Weighted Arithmetic mean – Measures of Dispersion- Range – Quartile deviation – Mean deviation – Standard deviation –Coefficient of variation – Combined mean and standard deviation. Skewness- Karl Pearson and Bowley's Coefficient of Skewness- Moments- Kurtosis.

Introduction

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

Measures of central tendency which are also known as averages, gives a single value which represents the entire set of data. The set of data may have equal or unequal values.

Measures of central tendency are also known as "Measures of Location".

It is generally observed that the observations (data) on a variable tend to cluster around some central value. For example, in the data on heights (in cms) of students, majority of the values may be around 160 cm. This tendency of clustering around some central value is called as central tendency. A measure of central tendency tries to estimate this central value.

Various measures of Averages are

(i) Arithmetic Mean
(ii) Median
(iii) Mode
(iv) Geometric Mean
(v) Harmonic Mean

Averages are important in statistics Dr.A.L.Bowley highlighted the importance of averages in statistics as saying "Statistics may rightly be called the Science of Averages".

Mean (Arithmetic)

The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data. The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have n values in a data set and they have values x1,x2, …,xn, the sample mean, usually denoted by $\bar{x}$ (pronounced "x bar"), is: $\bar{x} = \sum x / n$.

Median

The median is the middle score for a set of data that has been arranged in order of magnitude.

Mode

The mode is the most frequent score in our data set.

For individual observations $x_1, x_2, \ldots x_n$

(i) Mean $= \bar{X} = \frac{\Sigma x}{n}$

(ii) Median = Middle value if 'n' is odd
= Average of the two middle values if 'n' even

(iii) Mode = Most frequent value

Example 1

Find Mean, Median and Mode for the following data
3, 6, 7, 6, 2, 3, 5, 7, 6, 1, 6, 4, 10, 6

Solution:

Mean $= \bar{X} = \frac{\Sigma x}{n}$

$= \frac{3+6+7+\ldots +4+10+6}{14} = 5.14$

Median :

Arrange the above values in ascending (descending) order

1, 2, 3, 3, 4, 5, 6, 6, 6, 6, 6, 7, 7, 10

Here n = 14, which is even

$\therefore$ Median= Average two Middle values
= 6

Mode = 6 ($\because$ the values 6 occur five times in the above set of observation)

Grouped data (discrete)

For the set of values (observation) $x_1, x_2, \ldots x_n$ with corresponding frequences $f_1, f_2, \ldots f_n$

(i) Mean $= \overline{X} = \frac{\Sigma fx}{N}$, where $N = \Sigma f$

(ii) Median = the value of x, corresponding to the cumulative frequency just greater than $\frac{N}{2}$

(iii) Mode = the value of x, corresponding to a maximum frequency.

Example 2

Obtain Mean, Median, Mode for the following data

Value (x)	**0**	**1**	**2**	**3**	**4**	**5**
Frequency (f)	**8**	**10**	**11**	**15**	**21**	**25**

Solution:

x	0	1	2	3	4	5
f	8	10	11	15	21	25
fx	0	10	22	48	80	125
cf	8	18	29	44	65	90

$N = \Sigma f = 90$

$\Sigma fx = 285$

$\therefore$ Mean $= \frac{\Sigma fx}{N}$

$= 3.17$

Median :

$N = \Sigma f = 90$

$\frac{N}{2} = \frac{90}{2} = 45$

the cumulative frequency just greater than $\frac{N}{2} = 45$ is 65.

$\therefore$ The value of x corresponding to c.f. 65 is 4.

$\therefore$ Median = 4

Mode :

Here the maximum frequency is 25. The value of x, which corresponding to the maximum frequency (25) is 5.

$\therefore$ Mode = 5

Arithmetic mean for continuous distribution

The formula to calculate arithmetic mean under this type is

$$\overline{X} = A + \left(\frac{\Sigma fd}{N} \times c\right)$$

where A = arbitrary value (may or may not chosen from the mid points of class-intervals.

d = $\frac{x - A}{c}$ is deviations of each mid values.

c = magnitude or length of the class interval.

N = Σf = total frequency

Example 3

Calculate Arithmetic mean for the following

Marks	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**	**70-80**
No.of Students	**5**	**8**	**12**	**15**	**6**	**4**

Solution:

Marks	No. of Students	Mid value x	$d = \frac{x-A}{c}$ A=55, c=10	fd
20-30	5	25	-3	-15
30-40	8	35	-2	-16
40-50	12	45	-1	-12
50-60	15	55	0	0
60-70	6	65	1	6
70-80	4	75	2	8
	N = Σf = 50			**Σfd = -29**

∴ Arithmetic mean,

$$\overline{X} = A + \left(\frac{\Sigma fd}{N} \times c\right)$$

$$= 55 + \left(\frac{-29}{50} \times 10\right) = 49.2$$

Example 4

Calculate the Arithmetic mean for the following

Wages in Rs. :	100-119	120-139	140-159	160-179	180-199
No.of Workers :	18	21	13	5	3

Solution:

Wages	No. of workers f	Mid value x	$d = \frac{x-A}{c}$ A=149.5, c=20	fd
100-119	18	109.5	-2	-36
120-139	21	129.5	-1	-21
140-159	13	149.5	0	0
160-179	5	169.5	1	5
180-199	3	189.5	2	6
	N = Σf = 60			**Σfd = -46**

$$\overline{X} = A + \left(\frac{\Sigma fd}{N} \text{ x } c\right)$$

$$= 149.5 + \left(\frac{-46}{60} \text{ x } 20\right) = 134.17$$

Median for continuous frequency distribution

$$\text{Median} = l + \left(\frac{\frac{N}{2} - m}{f} \text{ x} c\right)$$

where l = lower limit of the Median class.

m = c.f. of the preceding (previous) Median class

f = frequency of the Median class

c = magnitude or length of the class interval corresponding to Median class.

N = Σf = total frequency.

Example 5

Find the Median wage of the following distribution

Wages (in Rs.) :	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**
No.of labourers:	**3**	**5**	**20**	**10**	**5**

Solution :

Wages	No. of labourers f	Cumulative frequency c.f.
20-30	3	3
30-40	5	8
40-50	20	28
50-60	10	38
60-70	5	43
	$N = \Sigma f = 43$	

Here $\frac{N}{2} = \frac{43}{2} = 21.5$

cumulative frequency just greater than 21.5 is 28 and the corresponding median class is 40-50

$$\Rightarrow l = 40, \text{m} = 8, \text{f} = 20, \text{c} = 10$$

$$\therefore \text{Median} = l + \left(\frac{\frac{N}{2} - m}{f} \times c \right)$$

$$= 40 + \left(\frac{21.5 - 8}{20} \times 10 \right) = \text{Rs. } 46.75$$

Example 6

Calculate the Median weight of persons in an office from the following data.

Weight (in kgs.)	:	**60-62**	**63-65**	**66-68**	**69-71**	**72-74**
No.of Persons	:	**20**	**113**	**138**	**130**	**19**

Solution:

Weight	No. of persons	c.f.
60-62	20	20
63-65	113	133
66-68	138	271
69-71	130	401
72-74	19	420
	N=Σf = 420	

Here $\frac{N}{2} = \frac{420}{2} = 210$

The cumulative frequency (c.f.) just greater than $\frac{N}{2} = 210$ is 271 and the corresponding Median class 66 - 68. However this should be changed to 65.5 - 68.5

$\Rightarrow l = 65.5, \quad m = 133, \quad f = 138, \quad c = 3$

Mode for continuous frequency distribution:

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - (f_0 + f_2)} \times c\right)$$

where l = lower limit of the modal class.

f_1 = frequency of the modal class.

f_0 = frequency of the class just preceding the modal class.

f_2 = frequency of the class just succeeding the modal class.

c = class magnitude or the length of the class interval corresponding to the modal class.

Observation :

Some times mode is estimated from the mean and the median. For a symmetrical distribution, mean, median and mode coincide. If the distribution is moderately asymmetrical the mean, median and mode obey the following empirical relationship due to Karl Pearson.

Mean - mode = 3(mean - median)

=> mode = 3 median - 2mean.

Example 7

Calculate the mode for the following data

Daily wages (in Rs.) :	50-60	60-70	70-80	80-90	90-100
No. of Workers :	35	60	78	110	80

Solution :

The greatest frequency = 110, which occurs in the class interval 80-90, so modal class interval is 80-90.

$\therefore\ l = 80,\ f_1 = 110,\ f_0 = 78;\ f_2 = 80;\ c = 10.$

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - (f_0 + f_2)} \times c\right)$$

$$= 80 + \left(\frac{110 - 78}{2(110) - (78 + 80)} \times 10\right)$$

= Rs. 85.16

Geometric mean:

(i) Geometric mean of n values is the nth root of the product of the n values. That is for the set of n individual observations $x_1, x_2 \ldots\ldots x_n$ their Geometric mean, denoted by G is

$$\sqrt[n]{x_1 \cdot x_2 \cdot x_3 \ldots x_n} \quad \text{or} \quad (x_1\ x_2 \ldots\ldots x_n)^{1/n}$$

Observation:

$$\log G = \log (x_1, x_2 \ldots\ldots\ldots x_n)^{1/n}$$

$$= \frac{1}{n} \log (x_1, x_2 \ldots\ldots\ldots x_n)$$

$$\log G = \frac{1}{n} \sum_{i=1}^{n} \log x_i$$

$$\Rightarrow \log G = \frac{\Sigma \log x}{n}$$

$$\therefore \text{Geometric Mean} = G = \text{Antilog}\left(\frac{\Sigma \log x}{n}\right)$$

Example : 8

Find the Geometric Mean of 3, 6, 24, 48.

Solution :

Let x denotes the given observation.

x	**log x**
3	0.4771
6	0.7782
24	1.3802
48	1.6812
	Σ log x= 4.3167

G.M. = 11.99

(ii) In case of discrete frequency distrisbution i.e. if $x_1, x_2 \ldots x_n$ occur f_1, f_2, f_n times respectively, the Geometric Mean, G is given by

$$G = \left(X_1^{f_1}\ X_2^{f_2}\ \ldots\ X_n^{f_n}\right)^{\frac{1}{N}}$$

where $N = \Sigma f = f_1 + f_2 + \ldots + f_n$

Observation:

$$\log G = \frac{1}{N} \log \left(X_1^{f_1} \; X_2^{f_2} \; \ldots \; X_n^{f_n} \right)$$

$$= \frac{1}{N} [f_1 \log x_1 + f_2 \log x_2 + \ldots\ldots + f_n \log x_n]$$

$$= \frac{1}{N} \Sigma f_i \log x_i$$

$$\Rightarrow \log G = \frac{\Sigma f_i \log x_i}{N}$$

$$\therefore G = \text{Antilog} \left(\frac{\Sigma f_i \log x_i}{N} \right)$$

Example 9

Calculate Geometric mean for the data given below

x	:	**10**	**15**	**25**	**40**	**50**
f	:	**4**	**6**	**10**	**7**	**3**

Solution :

x	f	log x	f logx
10	4	1.0000	4.0000
15	6	1.1761	7.0566
25	10	1.3979	13.9790
40	7	1.6021	11.2147
50	3	1.6990	5.0970
	N = Σf = 30		**Σf logx = 41.3473**

$$\therefore G = \text{Antilog} \left(\frac{\Sigma f \log x}{N} \right)$$

$$= \text{Antilog} \left(\frac{41.3473}{30} \right)$$

$$= \text{Antilog}\ (1.3782)$$

$$= 23.89$$

(iii) In the case of continuous frequency distribution,

$$\therefore G = \text{Antilog} \left(\frac{\Sigma f \log x}{N} \right)$$

where $N = \Sigma f$ and x being the midvalues of the class intervals

Example 10

Compute the Geometric mean of the following data

Marks :	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**
No. of students :	**5**	**7**	**15**	**25**	**8**

Solution :

Marks	No. of Students f	Mid value x	log x	f log x
0 – 10	5	5	0.6990	3.4950
10 – 20	7	15	1.1761	8.2327
20 – 30	15	25	1.3979	20.9685
30 – 40	25	35	1.5441	38.6025
40 – 50	8	45	1. 6532	13.2256
	$N = \Sigma f = 60$			$\Sigma f \log x = 84.5243$

$$\therefore G = \text{Antilog}\left(\frac{\text{Óflogx}}{N}\right)$$

$$= \text{Antilog}\left(\frac{84.5243}{60}\right)$$

$$= \text{Antilog}\,(1.4087) = 25.63$$

Observation:

Geometric Mean is always smaller than arithmetic mean i.e. G.M. ≤ A.M. for a given data

Harmonic Mean

(i) Harmonic mean of a number of observations is the reciprocal of the arithmetic mean of their reciprocals. It is denoted by H.

Thus, if $x_1, x_2 \ldots x_n$ are the observations, their reciprocals are $\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_n}$. The total of the reciprocals is $= \Sigma\left(\frac{1}{x}\right)$ and the mean of the reciprocals is $= \frac{\Sigma \frac{1}{x}}{n}$

$\therefore$ the reciprocal of the mean of the reciprocals is $= \frac{n}{\Sigma\left(\frac{1}{x}\right)}$

$$H = \frac{n}{\Sigma\left(\frac{1}{x}\right)}$$

EXAMPLE: 11

Find the Harmonic Mean of 6, 14, 21, 30

Solution :

x	$\frac{1}{x}$
6	0.1667
14	0.0714
21	0.0476
30	0.0333
	$\Sigma \frac{1}{x} = 0.3190$

$$H = \frac{n}{\Sigma \frac{1}{x}} = \frac{4}{0.3190} = 12.54$$

$\therefore$ Harmonic mean is $H = 12.54$

(ii) In case of discrete frequency distribution, i.e. if $x_1, x_2 \ldots. x_n$ occur $f_1, f_2, \ldots. f_n$ times respectively, the Harmonic mean, H is given by

$$H = \frac{1}{\frac{\frac{f_1}{x_1}+\frac{f_2}{x_2}+\ldots+\frac{f_n}{x_n}}{N}} = \frac{1}{\frac{1}{N}\Sigma\left(\frac{f}{x}\right)} = \frac{N}{\Sigma\left(\frac{f}{x}\right)}$$

where $N = \Sigma f$

Example 12

Calculate the Harmonic mean from the following data

x :	**10**	**12**	**14**	**16**	**18**	**20**
f :	**5**	**18**	**20**	**10**	**6**	**1**

Solution :

x	f	$\frac{f}{x}$
10	5	0.5000
12	18	1.5000
14	20	1.4286
16	10	0.6250
18	6	0.3333
20	1	0.0500
	$N = \Sigma f = 60$	$\Sigma \frac{f}{x} = 4.4369$

$$H = \frac{N}{\Sigma\left(\frac{f}{x}\right)}$$

$$= \frac{60}{4.4369} = 13.52$$

Example 13

Calculate the Harmonic mean for the following data.

Size of items	50-60	60-70	70-80	80-90	90-100
No. of items	12	15	22	18	10

Solution :

size	f	x	$\frac{f}{x}$
50-60	12	35	0.2182
60-70	15	65	0.2308
70-80	22	75	0.2933
80-90	18	85	0.2118
90-100	10	95	0.1053
	$N = \Sigma f = 77$		$\Sigma \frac{f}{x} = 1.0594$

$$H = \frac{N}{\Sigma \frac{f}{x}} = \frac{77}{1.0594} = 72.683$$

Observation:

(i) For a given data H.M. $\leq$ G.M.

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

(iii) (A.M.) x (H.M.) = $(\text{G.M.})^2$

Measures of Dispersion

In a group of individual items, all the items are not equal. There is difference or variation among the items. For example, if we observe the marks obtained by a group of studens, it could be easily found the difference or variation among the marks.

The common averages or measures of central tendency which we discussed earlier indicate the general magnitude of the data but they do not reveal the degree of variability in individual items in a group or a distribution. So to evaluate the degree of variation among the data, certain other measures called, measures of dispersion is used.

Measures of Dispersion in particular helps in finding out the variability or Dispersion/Scatteredness of individual items in a given distribution. The variability (Dispersion or Scatteredness) of the data may be known with reference to the central value (Common Average) or any arbitrary value or with reference to other vaues in the distribution. The mean or even Median and Mode may be same in two or more distribution, but the composition of individual items in the series may vary widely. For example, consider the following marks of two students.

Student I	Student II
68	82
72	90
63	82
67	21
70	65
340	340
Average 68	Average 68

It would be wrong to conclude that performance of two students is the same, because of the fact that the second student has failed in one paper. Also it may be noted that the variation among the marks of first student is less than the variation among the marks of the second student. Since less variation is a desirable characteristic, the first student is almost equally good in all the subjects.

It is thus clear that **measures of central tendency are insufficient to reveal the true nature and important characteristics of the data.** Therefore we need some other measures, called measures of Dispersion. Few of them are Range, Standard Deviation and coefficient of variation.

Range:

Range is the difference between the largest and the smallest of the values.

Symbollically,

Range = L - S

where L = Largest value

S = Smallest value

Co-efficient of Range is given by = $\frac{L - S}{L + S}$

Example 14

Find the value of range and its coefficient for the following data

6 8 5 10 11 12

Solution:

$$L = 12 \quad \text{(Largest)}$$
$$S = 5 \quad \text{(Smallest)}$$
$$\therefore \text{Range} = L - S = 7$$

$$\text{Co-efficient of Range} = \frac{L - S}{L + S} = 0.4118$$

Example : 15

Calculate range and its coefficient from the following distribution.

Size	**20 - 22**	**23 - 25**	**26 - 28**	**29 - 31**	**32 - 34**
Number	**7**	**9**	**19**	**42**	**27**

Solution:

Given is a continuous distribution. Hence the following method is adopted.

Here, L = Midvalue of the highest class

$$\therefore \quad L = \frac{32+34}{2} = 33$$

S = Mid value of the lowest class

$$\therefore \quad S = \frac{20+22}{2} = 21$$

$$\therefore \text{Range} = L - S = 12$$

Standard deviation:

Standard Deviation is the root mean square deviation of the values from their arithmetic mean.

S.D. is the abbreviation of standard Deviation and it is represented by the symbol σ (read as sigma). The square of standard deviation is called variance denoted by σ^2

(i) Standard Deviation for the raw data.

$$\sigma = \sqrt{\frac{\Sigma d^2}{n}}$$

Where $d = x - \overline{X}$

n = number of observations.

Example 16

Find the standard deviation for the following data

75, 73, 70, 77, 72, 75, 76, 72, 74, 76

Solution :

x	$d = x - \overline{X}$	d^2
75	1	1
73	-1	1
70	-4	16
77	3	9
72	-2	4
75	1	1
76	2	4
72	-2	4
74	0	0
76	2	4
$\Sigma x = 740$	$\Sigma d = 0$	$\Sigma d^2 = 44$

$$\overline{X} = \frac{\Sigma x}{n} = \frac{740}{10} = 74$$

$\therefore$ Standard Deviation,

$$\sigma = \sqrt{\frac{\Sigma d^2}{n}} = \sqrt{\frac{44}{10}} = 2.09$$

(ii) Standard deviation for the raw data without using Arithmetic mean.

The formula to calculate S.D in this case

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

Example : 17

Find the standard deviation of the following set of observations. 1, 3, 5, 4, 6, 7, 9, 10, 2.

Solution :

Let x denotes the given observations

x	:	1	3	5	4	6	7	9	8	10	2
x^2	:	1	9	25	16	36	49	81	64	100	4

Here $\Sigma x = 55$

$\Sigma x^2 = 385$

$$\therefore \sigma = \sqrt{\left(\frac{\Sigma x^2}{n}\right) - \left(\frac{\Sigma x}{n}\right)^2}$$

$$= \sqrt{\left(\frac{385}{10}\right) - \left(\frac{55}{10}\right)^2} = 2.87$$

(iii) S.D. for the raw data by Deviation Method

By assuming arbitrary constant, A, the standard deviation is given by

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

where d $= x - A$

A = arbitrary constant

Σd^2= Sum of the squares of deviations

Σd = sum of the deviations

n = number of observations

Example 18

For the data given below, calculate standard deviation 25, 32, 53, 62, 41, 59, 48, 31, 33, 24.

Solution:

Taking A = 41

x	25	32	53	62	41	59	48	31	33	24
d = x -A	-16	-9	12	21	0	18	7	-10	-8	-17
d^2	256	81	144	441	0	324	49	100	64	289

Here $\Sigma d = -2$

$\Sigma d^2 = 1748$

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

$$= \sqrt{\left(\frac{1748}{10}\right) - \left(\frac{-2}{10}\right)^2} = 13.21$$

(iv) Standard deviation for the discrete grouped data

In this case

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N}} \text{ where } d = x - \overline{X}$$

Example 19

Calculate the standard deviation for the following data

x	**6**	**9**	**12**	**15**	**18**
f:	**7**	**12**	**13**	**10**	**8**

Solution:

x	f	fx	$d = x - \overline{X}$	d^2	fd^2
6	7	42	-6	36	252
9	12	108	-3	9	108
12	13	156	0	0	0
15	10	150	3	9	90
18	8	144	6	36	288
$N = \Sigma f = 50$		**$\Sigma fx = 600$**			**$\Sigma fd^2 = 738$**

$$\overline{X} = \frac{\Sigma fx}{N} = \frac{600}{50} = 12$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N}} = \sqrt{\frac{738}{50}} = 3.84$$

(v) Standard deviation for the continuous grouped data without using Assumed Mean.

In this case

$$\sigma = c \times \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \text{ where } d = \frac{x - A}{c}$$

Example 20

Compute the standard deviation for the following data

Class interval :	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency :	8	12	17	14	9	7	4

Solution :

Taking A = 35

Class Intervals	Frequency f	Mid value x	$d = \frac{x-A}{c}$	fd	fd^2
0-10	8	5	-3	-24	72
10-20	12	15	-2	-24	48
20-30	17	25	-1	-17	17
30-40	14	A35	0	0	0
40-50	9	45	1	9	9
50-60	7	55	2	14	28
60-70	4	65	3	12	36
	$N = \Sigma f = 71$			$\Sigma fd = -30$	$\Sigma fd^2 = 210$

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

$$= 10 \times \sqrt{\frac{210}{71} - \left(\frac{-30}{71}\right)^2}$$

$$= 16.67$$

CO-EFFICIENT OF VARIATION:

Co-efficient of variation denoted by C.V. and is given by

$$C.V. = \left(\frac{\sigma}{\bar{x}} \times 100\right)\%$$

Observation:

(i) Co-efficient of variation is a **percentage expression**, it is used to compare two or more groups.

(ii) The group which has less coefficient of variation is said to be **more consistent** or **more stable**, and the group which has more co-efficient of variation is said to be **more variable** or **less consistent.**

Example 21

Prices of a particular commodity in two cities are given below.

City A :	40	80	70	48	52	72	68	56	64	60
City B :	52	75	55	60	63	69	72	51	57	66

Which city has more stable price

Solution :

City A	City B	$d_x=x-\overline{X}$	$d_y=y-\overline{y}$	$d^2_x=(x-\overline{X})^2$	$d^2_y=(y-\overline{y})^2$
40	52	-21	-10	441	100
80	75	19	13	361	169
70	55	9	-7	81	49
48	60	-13	-2	169	4
52	63	-9	1	81	1
72	69	11	7	121	49
68	72	7	10	49	100
56	51	-5	-11	25	121
64	57	3	-5	9	25
60	66	-1	-4	1	16
$\Sigma x=610$	$\Sigma y=620$			$\Sigma d^2_x = 1338$	$\Sigma d^2_y = 634$

$$\overline{X}= \frac{\Sigma x}{n} = \frac{610}{10} =61$$

$$\overline{y}= \frac{\Sigma y}{n} = \frac{620}{10} = 62$$

$$\sigma_x = \sqrt{\frac{1338}{10}} = 11.57$$

$$\sigma_y = \sqrt{\frac{634}{10}} = 7.96$$

$$\text{C.V. (x)} = \frac{\sigma_x}{\overline{x}} \times 100$$

$$= \frac{11.57}{61} = 18.97\%$$

$$\text{C.V. (y)} = \frac{\sigma_y}{\overline{y}} \times 100$$

$$= \frac{7.96}{62}\ 12.84\%$$

Conclusion

Comparatively, C.V. (y) < C.V (x)

⇒ City B has more stable price.

Combined Variance

If there are two sets of data consisting of n_1 and n_2 observations with s_1^2 and s_2^2 as their respective variances, then the variance of the combined set consisting of n_1+n_2 observations is :

$$S^2 = [n_1(s_1^2 + d_1^2) + n_2(s_2^2 + d_2^2)] / (n_1 + n_2)$$

Where d_1 and d_2 are the differences of the means, $\overline{x_1}$ and $\overline{x_2}$, from the combined mean $\overline{x}$ respectively.

Example : Find the combined standard deviation of two series A and B

	Series A	Series B
Mean	50	40
Standard deviation	5	6
No. of items	100	150

Solution:

Given $\overline{x_1} = 50$ and $\overline{x_2} = 40$, $s_1^2 = 25$ and $s_2^2 = 36$, $n_1 = 100$ and $n_2 = 150$

Combined mean $\overline{x} = \dfrac{100 \times 50 + 150 \times 40}{100 + 150} = 44,$

$d_1 = \overline{x_1} - \overline{x} = 50 - 44 = 6$, and $d_2 = \overline{x_2} - \overline{x} = 40 - 44 = -4$

Combined variance $= \dfrac{100(25+36)+150(36+16)}{100+150}$

$= 55.6$

Therefore, combined SD $= \sqrt{55.6} = 7.46$

MEAN DEVIATION

You have seen that range is a measure of dispersion, which does not depend on all observations. Let us think about another measure of dispersion, which will depend on all observations.

One measure of dispersion that you may suggest now is the sum of the deviations of observations from mean. But we know that the sum of deviations of observations from the A.M is always zero. So we cannot take the sum of deviations of observations from the mean as a measure.

One method to overcome this is to take the sum of absolute values of these deviations. But if we have two sets with different numbers of observations this cannot be justified. To make it meaningful we will take the average of the absolute deviations. Thus mean deviation (MD) about the mean is the mean of the absolute deviations of observations from arithmetic mean.

If $x_1, x_2, \ldots, x_n$ are n observations, then, MD $= \dfrac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$

EXAMPLE 22

Find the MD for the following data 12, 15, 21, 24, 28

Solution:

$$\overline{X} = \frac{12+15+21+24+28}{5} = 20$$

x	$\lvert x_i - \bar{x} \rvert$
12	8
15	5
21	1
24	4
28	8
Total	26

$$\text{MD} = \frac{26}{5} = 5.2$$

Mean deviation about mean for a frequency table

Let $x_1, x_2, \ldots, x_n$ be the values and $f_1, f_2, \ldots, f_n$ are the corresponding frequencies. Let N be the sum of the frequencies. Then, $\text{MD} = \frac{1}{N}\sum_{i=1}^{n} \lvert x_i - \bar{x} \rvert f_i$

In the case of a grouped frequency table, take the mid-values as x-values and use the same method given above.

Example : Find the mean deviation of the heights of 100 students given below:

Height in cm	frequency
160 – 162	5
163 – 165	18
166 – 168	42
169 – 171	27
172 - 174	8

Solution:

Height in cm	Mid-value (x)	Frequency (f)	fx	$\lvert x_i - \bar{x} \rvert$	$f_i \lvert x_i - \bar{x} \rvert$
160 – 162	161	5	805	6.45	32.25
163 – 165	164	18	2952	3.45	62.10
166 – 168	167	42	7014	0.45	18.90
169 – 171	170	27	4590	2.55	68.85
172 - 174	173	8	1384	5.55	44.40
Total		100	16745		226.50

$$\overline{X} = \frac{16745}{100} = 167.45$$

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

$$= \frac{226.5}{100} = 2.265$$

QUARTILE DEVIATION

Quartile deviation (Semi inter-quartile range) is one-half of the difference between the third quartile and first quartile.

That is, Quartile deviation, $Q.D = \frac{Q_3 - Q_1}{2}$

Example : Estimate an appropriate measure of dispersion for the following data:

Income (Rs.)	No. of persons
Less than 50	54
50 – 70	100
70 – 90	140
90 – 110	300
110 – 130	230
130 – 150	125
Above 150	51
	1000

Solution:

Since the data has open ends, Q.D would be a suitable measure

Income (Rs.) x	No. of persons f	Cumulative frequency
Less than 50	54	54
50 – 70	100	154
70 – 90	**140**	**294**
90 – 110	300	594
110 – 130	**230**	**824**
130 – 150	125	949
Above 150	51	1000
	1000	

$$Q_1 = l_1 + (\frac{N}{4} - m_1)\frac{c_1}{f_1}$$

$$Q_3 = l_3 + (\frac{3N}{4} - m_3)\frac{c_3}{f_3}$$

Here $N = 1000$, $\frac{N}{4} = 250$, $\frac{3N}{4} = 750$

The class 70 – 90 is the first quartile class and 110 – 130 is the third quartile class

$l_1 = 70$, $m_1 = 154$, $c_1 = 20$, $f_1 = 140$

$l_3 = 110$, $m_3 = 594$, $c_3 = 20$, $f_3 = 230$

$$Q_1 = 70 + (250 - 154)\frac{20}{140}$$
$$= 83.7$$

$$Q_3 = 110 + (750 - 594)\frac{20}{230}$$
$$= 123.5$$

$$Q.D = \frac{123.5 - 83.7}{2} = 19.9 \text{ Rs.}$$

RELATIVE MEASURES:

The absolute measures of dispersion discussed above do not facilitate comparison of two or more data sets in terms of their variability. If the units of measurement of two or more sets of data are same, comparison between such sets of data is possible directly in terms of absolute measures. But conditions of direct comparison are not met, the desired comparison can be made in terms of the *relative measures*.

Coefficient of Variation is a relative measure of dispersion which express standard deviation(σ) as percent of the mean. That is Coefficient of variation, $C.V = (\sigma / \overline{x})100$. Another relative measure in terms of quartile deviations is **Coefficient of quartile deviation** and is defined as $Q_r = \frac{Q_3 - Q_1}{Q_3 + Q1} \times 100$.

Example: An analysis of the monthly wages paid to workers in two firms A and B, belonging to the same industry, gives the following results:

	Firm A	Firm B
Number of workers	586	648
Average monthly wage	52.5	47.5
Standard deviation	10	11

In which firm, A or B, is there greater variability in individual wages?

Solution: Coefficient of variation for firm A $= \frac{10}{52.5} \times 100$

$= 19\%$

Coefficient of variation for firm B $= \frac{11}{47.5} \times 100$

$= 23\%$

There is greater variability in wages in firm B.

Very often it becomes necessary to have a measure that reveals the direction of dispersion about the center of the distribution. Measures of dispersion indicate only the extent to which individual values are scattered about an average. These do not give information about the direction of scatter. *Skewness* refers to the direction of dispersion leading departures from symmetry, or lack of symmetry in a direction.

If the frequency curve of a distribution has longer tail to the right of the center of the distribution, then the distribution is said to be positively skewed. On the other hand, if the distribution has a longer tail to the left of the center of the distribution, then distribution is said to be negatively skewed. Measures of skewness indicate the magnitude as well as the direction of skewness in a distribution.

Empirical Relationship between Mean, Median and Mode

The relationship between these three measures depends on the shape of the frequency distribution. In a symmetrical distribution the value of the mean, median and the mode is the same. But as the distribution deviates from symmetry and tends to become skewed, the extreme values in the data start affecting the mean.

In a positively skewed distribution, the presence of exceptionally high values affects the mean more than those of the median and the mode. Consequently the mean is highest, followed, in a descending order, by the median and the mode. That is, for a *positively skewed distribution*, Mean > Median> Mode. In a negatively skewed distribution, on the other hand, the presence of exceptionally low values makes the values of the mean the least, followed, in an ascending order, by the median and the mode. That is, for a negatively skewed distribution, Mean < Median < Mode.

Empirically, if the number of observations in any set of data is large enough to make its frequency distribution smooth and moderately skewed, then, Mean – Mode = 3(Mean – Median)

MEASURES OF SKEWNESS

3. ***Karl Pearson's measure of skewness***: Prof. Karl Pearson has been developed this measure from the fact that when a distribution drifts away from symmetry, its mean, median and mode tend to deviate from each other.

   Karl Pearson's measure of skewness is defined as, $S_{kP} = \dfrac{\text{Mean - Mode}}{\text{SD}}$

4. ***Bowley's measure of skewness***: developed by Prof. Bowley, this measure of skewness is derived from quartile values.

   It is defined as $S_{kB} = \dfrac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q1}$

5. ***Moment measure of skewness***:

   If $x_1, x_2, \ldots, x_n$ are n observations, then the r^{th} moment about mean is defined as

   $$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r$$

   The moment measure of skewness is defined as $\beta_1 = m_3/(SD)^3$

   In a perfectly symmetrical distribution $\beta_1 = 0$, and a greater or smaller value of β_1 results in a greater or smaller degree of skewness.

REFERENCES:

1. Vittal.P.R., Business Statistics, Margham publications, Chennai, 2008.
2. Gupta.S.C. Statistics, Himalaya Publishers, Mumbai, 2005.
3. S.P.Gupta, Business Statistics, Sultan Chand & Sons, New Delhi, 2008.
4. Beri.G., Business Statistics, Tata McGraw Hill Publishing Company Ltd, New Delhi, 2009.
5. Yule and Kendall (1993), Introduction to theory of Statistics. Universal Book Stall, New Delhi.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – III – CURVE FITTING – SMT1304

CURVE FITTING

Fitting a straight line and second degree parabola. Correlation- Scatter diagram – Limits of correlation coefficient – Spearman's Rank correlation coefficient- Simple problems – Regression- Properties of Regression coefficients and regression lines.

Fitting curves by Method of Least Squares

Curve Fitting: Let (x_i, y_i); $i = 1, 2 \ldots n$ be a given set of n pairs of values, X being independent variable and Y being the dependent variable. The general problem in curve fitting is to find, if possible, an analytic expression of the form y = f(x), for the functional relationship suggested by the given data. Fitting of curves to a set of numerical data is of considerable importance theoretical as well as practical. Moreover, it may be used to estimate the values of one variable which would correspond to the specified values of the other variable.

Fitting a straight line

Let y = a + bx be the equation of the line to be fitted. To estimate the values of a and b we have, the following normal equations.

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2}$$

Here n is the number of observations, and the quantities $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i y_i$ and $\sum_{i=1}^{n} x_i^{\,2}$ can be obtained from the given set of points (x_i, y_i); i = 1, 2, …, n and the above equations can be solved for a and b.

Solved Examples:

Example 1: Fit a straight line to the following data:

X	1	2	3	4	6	8
Y	2.4	3	3.6	4	5	6

Solution: Let the straight line to be fitted is y = a + bx

X	**Y**	**XY**	**X^2**
1	2.4	2.0	1
2	3	6.0	4
3	3.6	10.8	9
4	4	16.0	16
6	5	30.0	36
8	6	48.0	64
24	**24**	**113.2**	**130**

Using the normal equations, $\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$

$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2}$ we get,

24 = 6a + 24b and
113.2 = 24a + 130b
Solving above two equations, we get
a = 1.976 and b = 0.506

Example 2 : Fit a straight line to the following data:

X	0	5	10	15	20	25
Y	12	15	17	22	24	30

Solution: Let the straight line to be fitted is y = a + bx

X	**Y**	**XY**	**X^2**
0	12	0	0
5	15	75	25
10	17	170	100
15	22	330	225
20	24	480	400
25	30	750	625
75	**120**	**1805**	**1375**

Using the normal equations, $\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$

$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\ 2}$ We get,

120 = 6 a + 75 b and
1805 = 75 a + 1375 b
Solving above two equations, we get
a = 11.29 and b = 0.697

Example 3: Fit a straight line of the form y = a + bx for the following data and estimate the value of y when x is 40

X	2	4	6	10	20	24
Y	6	8	13	12	35	42

Solution: Here n = 6

X	**Y**	**XY**	**X^2**
2	6	12	4
4	8	24	16
6	13	78	36
10	12	120	100
20	35	700	400
24	42	1008	576
66	**116**	**1950**	**1132**

Using the normal equations, $\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$

$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\ 2}$ We get, 116 = 6 a + 66 b and

1950 = 66 a + 1132

Solving the above two equations, we get

a = 1.073 and

b = 1.66

Now to estimate the value of y when x is 40, we substitute the value of x in the fitted equation

y = a + b x

(i.e.) y = 1.07 + 1.66 x

= 1.07 + 1.66 x 40

y = 67.47

Fitting a parabola

Let $y = a + bx + cx^2$ be the equation of the line to be fitted. To estimate the values of a and b and c, we have, the following normal equations.

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3$$

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4$$

Here n is the number of observations, and the quantities $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i y_i$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} x_i^3$, $\sum_{i=1}^{n} x_i^4$ and $\sum_{i=1}^{n} x_i^2 y_i$ can be obtained from the given set of points (x_i, y_i); i = 1, 2, …, n and the above equations can be solved for a, b and c.

Solved Examples:

Example 4: Fit a parabola to the following data:

X	0	1	2	3	4
Y	1	1.8	1.3	2.5	6.3

Solution: Let $y = a + bx + cx^2$ be the second degree parabola to be fitted, n = 5

X	**Y**	**X^2**	**X^3**	**X^4**	**XY**	**X^2Y**
0	1.0	0	0	0	0	0
1	1.8	1	1	1	1.8	1.8
2	1.3	4	8	16	2.6	5.2
3	2.5	9	27	81	7.5	22.5
4	6.3	16	64	256	25.2	100.8
10	**12.9**	**30**	**100**	**354**	**37.1**	**130.3**

Using normal equations $\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2} + c\sum_{i=1}^{n} x_i^{\,3}$$

$$\sum_{i=1}^{n} x_i^{\,2} y_i = a\sum_{i=1}^{n} x_i^{\,2} + b\sum_{i=1}^{n} x_i^{\,3} + c\sum_{i=1}^{n} x_i^{\,4}$$ We get,

12.9 = 5a + 10b + 30c
37.1 = 10a + 30b + 100c
130.3 = 30a + 100b + 354c
Solving the above equations, we get a = 1.42; b = – 1.07; c = 0.55.
Thus the required equation of parabola is $y = 1.42 – 1.07x + 0.55x^2$

Example 5: Fit a parabola to the following data and estimate y when x is 6

X	1	3	4	5	7
Y	2	3	6	15	39

Solution: Let $y = a + bx + cx^2$ be the second degree parabola to be fitted, n = 5

X	**Y**	**X^2**	**X^3**	**X^4**	**XY**	**X^2Y**
1	2	1	1	1	2	2
3	3	9	27	81	9	27
4	6	16	64	256	24	96
5	15	25	125	625	75	375
7	39	49	343	2401	273	1911
20	**65**	**100**	**560**	**3364**	**383**	**2411**

Using normal equations $\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^{\,2}$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2} + c\sum_{i=1}^{n} x_i^{\,3}$$

$$\sum_{i=1}^{n} x_i^{\,2} y_i = a\sum_{i=1}^{n} x_i^{\,2} + b\sum_{i=1}^{n} x_i^{\,3} + c\sum_{i=1}^{n} x_i^{\,4}$$ we get,

65 = 5a + 20b + 100c
383 = 20a + 100b + 560c
2411 = 100a + 560b + 3364c
Solving the above equations, we get a = 6.54; b = – 5.93; c = 1.51.
Thus the required equation of parabola is $y = 6.54 – 5.93x + 1.51x^2$
Now to estimate the value of y when x is 6, we substitute the value of x in the fitted equation
$y = a + bx + cx^2$
$= 6.54 – 5.93 \times 6 + 1.51 \times 6^2$
y = 25.32

Correlation and Regression

Correlation and Regression analyses are based on the relationship, or association, between two (or more) variables. In correlation, we consider the linear relationship between two variables, the sample observations are obtained by selecting a random sample of the units of association (which may be persons, places, animals, points in time, or any other element on which the two measurements are taken) and by taking on each a measurement of X and a measurement of Y. The objective is solely to obtain a measure of the strength of the relationship between the variables. For example, the relationship between the wing length and tail length of a particular species of birds can be studied by the correlation analysis.

Karl Pearson Coefficient of Correlation

As a measure of intensity or degree of linear relationship between two variables, Karl Pearson (1867-1936), a British Biometrician, developed a formula called Correlation coefficient (also called as product moment correlation coefficient).

Correlation coefficient between two random variables X and Y usually denoted by r(X, Y) or simply r_{XY}, is a numerical measure of linear relationship between them and is defined as

$$r(X,Y)=\frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

(i.e.) $$r(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2.\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{\frac{1}{2}}} \quad \ldots\ldots(3.1)$$

(i.e.) $$r(X,Y)=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n}x^2-n\bar{x}^2\right)\left(\sum_{i=1}^{n}y^2-n\bar{y}^2\right)}} \quad \ldots\ldots(3.2)$$

(i.e) $$r(X,Y)=\frac{n\sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sqrt{\left(n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\right)\left(n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2\right)}} \quad \ldots\ldots(3.3)$$

The correlation coefficient is a dimensionless number; it has no units of measurement. The maximum value r can achieve is 1, and its minimum value is –1. Therefore, for any given set of observations, $-1 \le r \le 1$. The values r = 1 and r = – 1 occur when there is an exact linear relationship between x and y. As the relationship between x and y deviates from perfect linearity, r moves away from 1 or – 1 and closer to 0. If y tends to increase in magnitude as x increases, r is greater than 0 and x and y are said to be positively correlated; if y decreases as x increases, r is less than 0 and the two variables are negatively correlated. If r = 0, there is no linear relationship between x and y and the variables are uncorrelated.

Note: The formula (3.1) can be used when the values of $\bar{x}$ and $\bar{y}$ are integral values and formula (3.2) can be used when the values of $\bar{x}$ and $\bar{y}$ are in decimals. For a continuous series of values the mid points of the class intervals will be used for x and y values.

Solved Examples:

Example 1: Calculate the correlation coefficient between X and Y from the following data:

$\sum_{i=1}^{15}\left(X_i-\bar{X}\right)^2=136$ $\quad$ $\sum_{i=1}^{15}\left(Y_i-\bar{Y}\right)^2=138$ $\quad$ $\sum_{i=1}^{15}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)=122$

Solution:

We have $r(X,Y)=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2.\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{\frac{1}{2}}}$

$=\dfrac{122}{\sqrt{136}\sqrt{138}}$

r(X,Y) = 0.89

Example 2. Some health researchers have reported an inverse relationship between central nervous system malformations and the hardness of the related water supplies. Suppose the data were collected on a sample of 9 geographic areas with the following results:

C.N.S. malformation	9	8	5	1	4	2	3	6	7
Water hardness(ppm)	120	130	90	150	160	100	140	80	200

Calculate the Correlation Coefficient between the C.N.S. malformation rate and Water hardness.

Solution:

Let us denote the C.N.S. malformation rate by x and water hardness by y. The mean of the x series $\bar{x}$ = 5 and the mean of the y series $\bar{y}$ =130, hence we can use the formula (2.1)

Calculation of correlation coefficient

x	**y**	**(x – $\bar{x}$) = x – 5**	**(y – $\bar{y}$) = y – 130**	**(x – $\bar{x}$)^2**	**(y – $\bar{y}$)^2**	**(x – $\bar{x}$) (y – $\bar{y}$)**
9	120	4	– 10	16	100	– 40
8	130	3	0	9	0	0
5	90	0	– 40	0	1600	0
1	150	– 4	20	16	400	– 80
4	160	– 1	30	1	900	– 30
2	100	– 3	– 30	9	900	90
3	140	– 2	10	4	100	– 20
6	80	1	– 50	1	2500	– 50
7	200	2	70	4	4900	140
				Σ(x – $\bar{x}$)^2=60	**Σ(y – $\bar{y}$)^2=11400**	**Σ(x – $\bar{x}$) (y – $\bar{y}$)=10**

$$r(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2.\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{\frac{1}{2}}} = \frac{10}{[60x11400]^{\frac{1}{2}}}$$

r(X,Y) = 0.012

Therefore, the correlation coefficient between the C.N.S. malformation rate and water hardness is 0.012.

Example 3: Find the product moment correlation for the following data

X	57	62	60	57	65	60	58	62	56

Y	71	70	66	70	69	67	69	63	70

Solution:

X	**Y**	**XY**	**X^2**	**Y^2**
57	71	4047	3249	5041
62	70	4340	3844	4900
60	66	3960	3600	4356
57	70	3990	3249	4900
65	69	4485	4225	4761
60	67	4020	3600	4489
58	69	4002	3364	4761
62	63	3906	3844	3969
56	70	3920	3136	4900
537	**615**	**36670**	**32111**	**42077**

Thus we have, n = 9, $\sum X = 537$, $\sum Y = 615$, $\sum XY = 36670$, $\sum X^2 = 32111$, $\sum Y^2 = 42077$

$$r(X,Y) = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{n\sum X^2 - \left(\sum X\right)^2}\sqrt{n\sum Y^2 - \left(\sum Y\right)^2}}$$

$$= \frac{9\times 36670 - 537\times 615}{\sqrt{9\times 32111 - 537^2}\sqrt{9\times 42077 - 615^2}}$$

$$r(X,Y) = -0.414$$

Example 4: A computer operator while calculating the coefficient of correlation between two variables X and Y for 25 pairs of observations obtained the following constants: $\sum X = 125$, $\sum Y = 100$, $\sum XY = 508$, $\sum X^2 = 650$, $\sum Y^2 = 460$. However it was later discovered at the time of checking that he had copied two pairs as (6,14) and (8,6) while the correct pairs were (8,12) and (6,8). Obtain the correct correlation coefficient.

Solution:

The formula involved with the given data is,

$$r(X,Y) = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{n\sum X^2 - \left(\sum X\right)^2}\sqrt{n\sum Y^2 - \left(\sum Y\right)^2}}$$

The Corrected $\sum X$ = Incorrect $\sum X$ – (6+8) + (8+6) = 125

Corrected $\sum Y$ = Incorrect $\sum Y$ – (14+6) + (12+8) = 100

Corrected $\sum X^2$ = Incorrect $\sum X^2$ – $(6^2+8^2) + (8^2+6^2)$ = 650

Corrected $\sum Y^2$ = Incorrect $\sum Y^2$ – $(14^2+6^2) + (12^2+8^2)$ = 436

Corrected $\sum XY$ = Incorrect $\sum XY$ – (84+48) + (96+48) = 520

Now the correct value of correlation coefficient is,

$$r(X,Y) = \frac{25\times 520 - 125\times 100}{\sqrt{25\times 650 - 125^2}\sqrt{25\times 436 - 100^2}} = 0.67$$

Spearman's Rank Correlation Coefficient

If X and Y are qualitative variables then Karl Pearson's coefficient of correlation will be meaningless. In this case, we use Spearman's rank correlation coefficient which is defined as follows:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2-1)}$$ where d is the difference in ranks.

Solved Examples:

Example 5: The ranks of same 16 students in Mathematics and Physics are as follows. The numbers within brackets denote the ranks of the students in Mathematics and Physics. (1,1), (2,10), (3,3), (4,4), (5,5), (6,7), (7,2), (8,6), (9,8), (10,11) (11. 15), (12,9), (13,14), (14,12), (15,16), (16,13). Calculate the rank correlation coefficient for the proficiencies of this group in Mathematics and Physics.

Solution:

Ranks in Maths (X)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
Ranks in Physics (Y)	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13	
d = X – Y	0	–8	0	0	0	–1	5	2	1	–1	–4	3	–1	2	–1	3	0
d^2	0	64	0	0	0	1	25	4	1	1	16	9	1	4	1	9	136

Spearman's Rank Correlation Coefficient is given by, $\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2-1)}$

$$= 1 - \frac{6 \times 136}{16(16^2-1)}$$ = 0.8

Example 6: Ten competitors in a musical test were ranked by the three judges A, B and C in the following order:

Ranks by A	1	6	5	10	3	2	4	9	7	8
Ranks by B	3	5	8	4	7	10	2	1	6	9
Ranks by C	6	4	9	8	1	2	3	10	5	7

Using rank correlation coefficient method, discuss which pair of judges has the nearest approach to common likings in music.

Solution: Here n = 10

Ranks by A (X)	Ranks by B (Y)	Ranks by C (Z)	D_1 =X–Y	D_2 =X–Z	D_3 =X–Y	D_1^2	D_2^2	D_3^2
1	3	6	– 2	– 5	– 3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	– 3	– 4	– 1	9	16	1
10	4	8	6	– 2	– 4	36	4	16
3	7	1	– 4	2	6	16	4	36
2	10	2	– 8	0	8	64	0	64
4	2	3	2	1	– 1	4	1	1
9	1	10	8	– 1	– 9	64	1	81
7	6	5	1	2	1	1	4	1

8	9	7	– 1	1	2	1	1	4
Total			0	0	0	200	60	214

$$\rho(X,Y) = 1 - \frac{6\sum_{i=1}^{n} D_1^{\,2}}{n(n^2-1)} = 1 - \frac{6 \times 200}{10 \times 99} = -\frac{7}{33}$$

$$\rho(X,Z) = 1 - \frac{6\sum_{i=1}^{n} D_2^{\,2}}{n(n^2-1)} = 1 - \frac{6 \times 60}{10 \times 99} = \frac{7}{11}$$

$$\rho(Y,Z) = 1 - \frac{6\sum_{i=1}^{n} D_2^{\,2}}{n(n^2-1)} = 1 - \frac{6 \times 214}{10 \times 99} = -\frac{49}{165}$$

Since $\rho(X,Z)$ is maximum, we conclude that the pair of judges A and C has the nearest approach to common likings in music.

Example 7: The coefficient of rank correlation between the marks in Statistics and Mathematics obtained by a certain group of students is 2/3 and the sum of the squares of the differences in ranks is 55. Find the number of students in the group.

Solution: Spearman's rank correlation coefficient is given by $\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2-1)}$

Here ρ = 2/3, $\sum d^2 = 55$, N = ?

Therefore $\frac{2}{3} = 1 - \frac{6 \times 55}{N(N^2-1)}$

Solving the above equation we get N = 10

Repeated Ranks:

If any two or more individuals are equal in any classification with respect to characteristic A or B, or if there is more than one item with the same value in the series then Spearman's formula for calculating the rank correlation coefficients breaks down. In this case, common ranks are given to the repeated ranks. This common rank is the average of the ranks which these items would have assumed if they are slightly different from each other and the next item will get the rank next the ranks already assumed. As a result of this, following adjustment is made in the formula: add the factor $\frac{m(m^2-1)}{12}$ to $\sum d^2$ where m is the number of items an item is repeated. This correction factor is to be added for each repeated value.

Example 8: Obtain the rank correlation coefficient for the following data:

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

Solution:

X	**Y**	**Rank X**	**Rank Y**	**D = X – Y**	**D^2**
68	62	4	5	– 1	1

64	58	6	7	– 1	1
75	68	2.5	3.5	– 1	1
50	45	9	10	– 1	1
64	81	6	1	5	25
80	60	1	6	– 5	25
75	68	2.5	3.5	– 1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
					72

In X series 75 is repeated twice which are in the positions 2nd and 3rd ranks. Therefore common ranks 2.5 (which is the average of 2 and 3) is given for each 75. The corresponding correction factor is

$$C.F = \frac{2(2^2-1)}{12} = \frac{1}{2}.$$

Also in the X series 64 is repeated thrice which are in the position 5th, 6th and 7th ranks. Therefore, a common rank 6(which is the average of 5, 6 and 7) is given for each 64. The corresponding correction factor is

$$C.F = \frac{3(3^2-1)}{12} = 2.$$

Similarly, in the Y series, 68 is repeated twice which are in the positions 3rd and 4th ranks. Therefore, common ranks(which is the average of 3 and 4) is given for each 68. The corresponding correction factor is

$$C.F = \frac{2(2^2-1)}{12} = \frac{1}{2}.$$

Now, Rank correlation coefficient is $\rho = 1 - \frac{6\left(\Sigma d^2 + TotalCorrectionFactor\right)}{n\left(n^2-1\right)}$

$$= 1 - \frac{6\left(72 + \frac{1}{2} + 2 + \frac{1}{2}\right)}{10(10-1)} \quad = 0.5454$$

Regression Analysis

Regression analysis helps us to estimate or predict the value of one variable from the given value of another. The known variable(or variables) is called independent variable(s). The variable we are trying to predict is the dependent variable. For example, in the relationship between blood pressure and age in humans, blood pressure may be considered the dependent variable and age the independent variable.

Regression equations

Prediction or estimation of most likely values of one variable for specified values of the other is done by using suitable equations involving the two variables. Such equations are known as Regression Equations

Regression equation of y on x:

$y - \bar{y} = b_{yx}(x - \bar{x})$ where y is the dependent variable and x is the independent variable and b_{yx} is given by

$$b_{yx} = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sum_{i=1}^{n}(x-\bar{x})^2} \quad \text{Or} \quad b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{n\sum_{i=1}^{n}xy - \sum_{i=1}^{n}x\sum_{i=1}^{n}y}{n\sum_{i=1}^{n}x^2 - \left(\sum_{i=1}^{n}x\right)^2}$$

Regression equation of x on y:

$x - \bar{x} = b_{xy}(y - \bar{y})$ where y is the dependent variable and x is the independent variable and b_{yx} is given by

$$b_{xy} = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sum_{i=1}^{n}(y-\bar{y})^2}$$

Or $$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{n\sum_{i=1}^{n}xy - \sum_{i=1}^{n}x\sum_{i=1}^{n}y}{n\sum_{i=1}^{n}y^2 - \left(\sum_{i=1}^{n}y\right)^2}$$

b_{yx} and b_{xy} are called as regression coefficients of y on x and x on y respectively.

Relation between correlation and regression coefficients:

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \quad \text{and} \quad b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$b_{yx} \cdot b_{xy} = r\frac{\sigma_y}{\sigma_x}\, r\frac{\sigma_x}{\sigma_y} = r^2$$

Hence $r = \pm\sqrt{b_{yx}b_{xy}}$

Note: In the above expression the components inside the square root is valid only when b_{yx} and b_{xy} have the same sign. Therefore the regression coefficients will have the same sign.

Solved Examples:

Example 9: In trying to evaluate the effectiveness of its advertising campaign a company compiled the following information. Calculate the regression line of sales on advertising.

Year	1980	1981	1982	1983	1984	1985	1986	1987
Advertisement in 1000 rupees	12	15	15	23	24	38	42	48
Sales in lakhs of rupees	5	5.6	5.8	7.0	7.2	88	9.2	9.5

Solution : Let x be advertising amount and y be the sales amount.

Here, n = 8, $\bar{x} = \frac{217}{8} = 27.1$, $\bar{y} = \frac{58.1}{8} = 7.26$

We know that, Regression equation of y on x is given by $y - \bar{y} = b_{yx}(x - \bar{x})$

Where $$b_{yx} = \frac{n\sum_{i=1}^{n} xy - \sum_{i=1}^{n} x \sum_{i=1}^{n} y}{n\sum_{i=1}^{n} x^2 - \left(\sum_{i=1}^{n} x\right)^2}$$

X	**Y**	**X²**	**XY**
12	5	144	60
15	5.6	225	84
15	5.8	225	87
23	7.0	529	161
24	7.2	576	172.8
38	8.8	1444	334.4
42	9.2	1764	386.4
48	9.5	2304	456
217	**58.1**	**7211**	**1741.6**

Therefore $b_{yx} = \frac{8 \times 1741.6 - 217 \times 58.1}{8 \times 7211 - 217^2} = 0.125$

Substituting this value in the y on x equation, we get,

$$y - 7.26 = 0.125(x - 27.1)$$

Therefore the required equation of Sales on Advertisement is y = 3.87 + 0.125 x

Example 10: In a study of the effect of a dietary component on plasma lipid composition, the following ratios were obtained on a sample of experimental anumals

Measure of dietary component (X)	1	5	3	2	1	1	7	3
Measure of plasma lipid level (Y)	6	1	0	0	1	2	1	5

(i) Obtain the two regression lines and hence predict the ratio of plasma lipid level with 4 dietary components.

(ii) Find the correlation coefficient between X and Y

Solution:

(i)

X	**Y**	**XY**	**X²**	**Y²**
1	6	6	1	36
5	1	5	25	1
3	0	0	9	0
2	0	0	4	0
1	1	1	1	1
1	2	2	1	4
7	1	7	49	1
3	5	15	9	25
23	**16**	**36**	**99**	**68**

Here n = 8 $\bar{x}$ = 2.875; $\bar{y}$ = 2

The Regression equation of y on x is given by y – $\bar{y}$ = b_{yx} (x – $\bar{x}$)

Where $$b_{yx} = \frac{n\sum_{i=1}^{n} xy - \sum_{i=1}^{n} x \sum_{i=1}^{n} y}{n\sum_{i=1}^{n} x^2 - \left(\sum_{i=1}^{n} x\right)^2}$$

$$b_{yx} = \frac{8\times 36 - 23\times 16}{8\times 99 - 23^2} = -0.304$$

Hence the regression equation of y on x is
y – 2 = – 0.304(x – 2.875)
(i.e) y = 2.874 – 0.304 x

when x = 4 (measure of dietary component) the plasmid lipid level is
y = 2.874 – 0.304 (4)
y = 1.658

The Regression equation of x on y is given by x – $\bar{x}$ = b_{xy} (y – $\bar{y}$)

Where $$b_{xy} = \frac{n\sum_{i=1}^{n} xy - \sum_{i=1}^{n} x \sum_{i=1}^{n} y}{n\sum_{i=1}^{n} y^2 - \left(\sum_{i=1}^{n} y\right)^2}$$

$$b_{xy} = \frac{8\times 36 - 23\times 16}{8\times 68 - 16^2} = -0.278$$

Hence the regression equation of x on y is
x – 2.875= – 0.278(y – 2)
(i.e) x = 3.431 – 0.278 y

(ii) The correlation coefficient between x and y is given by

$$r = \pm\sqrt{b_{yx} b_{xy}}$$

$$r = \pm\sqrt{-0.304 \times -0.278} = \pm 0.291$$

Example 11: From the data given below find (i) two regression lines (ii) coefficient of correlation between marks in Physics and marks in Chemistry (iii) most likely marks in Chemistry when marks in Physics is 78 (iv) most likely marks in Physics when marks in Chemistry is 92

Marks in Physics (X)	72	85	91	85	91	89	84	87	75	77
Marks in Chemistry (Y)	76	92	93	91	93	95	88	91	80	81

Solution:

(i)

X	**Y**	**X^2**	**Y^2**	**XY**
72	76	5184	5776	5472
85	92	7225	8464	7820
91	93	8281	8649	8463
85	91	7225	8281	7735
91	93	8281	8649	8463
89	95	7921	9025	8455
84	88	7056	7744	7395
87	91	7569	8281	7917
75	80	5625	6400	6000
77	81	5929	6561	6237
836	**880**	**70296**	**77830**	**73957**

Here n = 10 $\bar{x}$ = 83.6 $\bar{y}$ = 88

The Regression equation of y on x is given by $y - \bar{y} = b_{yx}(x - \bar{x})$

Where $$b_{yx} = \frac{n\sum_{i=1}^{n} xy - \sum_{i=1}^{n} x \sum_{i=1}^{n} y}{n\sum_{i=1}^{n} x^2 - \left(\sum_{i=1}^{n} x\right)^2}$$

$$b_{yx} = \frac{10 \times 73957 - 836 \times 880}{10 \times 70296 - 836^2} = 0.949$$

Hence the regression equation of y on x is

$y - 88 = 0.949(x - 83.6)$

(i.e) $y = 8.6 + 0.949x$

The Regression equation of x on y is given by $x - \bar{x} = b_{xy}(y - \bar{y})$

Where $$b_{xy} = \frac{n\sum_{i=1}^{n} xy - \sum_{i=1}^{n} x \sum_{i=1}^{n} y}{n\sum_{i=1}^{n} y^2 - \left(\sum_{i=1}^{n} y\right)^2}$$

$$b_{xy} = \frac{10 \times 73957 - 836 \times 880}{10 \times 77830 - 880^2} = 0.990$$

Hence the regression equation of x on y is

$x - 83.6 = 0.990(y - 88)$

(i.e) $x = -3.5 + 0.990y$

(ii) The correlation coefficient between x and y is given by

$$r = \pm\sqrt{b_{yx} b_{xy}}$$

$$r = \pm\sqrt{0.949 \times 0.990} = \pm 0.969$$

(iii) To find the most likely marks in Chemistry when marks in Physics is 78, we have to use the regression equation of y on x given by y = 8.6 + 0.949 x
Substituting the value of x as 78 in the above equation, we get,

$$y = 8.6 + 0.949 (78)$$
$$y = 73.85$$

Hence the marks in Chemistry is 82.62

(iv) To find the most likely marks in Physics when marks in Chemistry is 92, we have to use the regression equation of x on y given by x = – 3.5 + 0.990 y
Substituting the value of y as 92 in the above equation, we get,

$$x = -3.5 + 0.990 (92)$$
$$x = 87.58$$

Hence the marks in Physics is 87.58

Example 12: For a given series of values, the following data were obtained, $\bar{x}$ =36, $\bar{y}$ =85, σ_x = 11, σ_y = 8 and r = 0.66. Find (i) two regression equations (ii) estimation of x when y = 75.

Solution:

We have $b_{yx} = r\dfrac{\sigma_y}{\sigma_x} = 0.66 \times \dfrac{8}{11} = 0.4799$

and $b_{xy} = r\dfrac{\sigma_x}{\sigma_y} = 0.66 \times \dfrac{11}{8} = 0.9075$

(i) The Regression equation of y on x is given by
$y - \bar{y} = b_{yx} (x - \bar{x})$
y – 85 = 0.4799 (x – 36)
(i.e.) y = – 17.28 + 0.4799 x

The Regression equation of x on y is given by
$x - \bar{x} = b_{xy} (y - \bar{y})$
x – 36 = 0.9075 (y – 85)
(i.e.) x = – 41.35 + 0.9075 y

(ii) To estimate the value of x when y = 75, we use the regression line of x on y
x = – 41.35 + 0.9075 y Substituting y = 75, x = – 41.35 + 0.9075 (75)
Therefore x = 29.9

Example 13: For a certain X and Y series which are correlated, the regression lines are 8x – 10y = – 66 and 40x – 18y = 214. Find (i) the correlation coefficient between them and (ii) the mean of the two series.

Solution:
The given regression equations are
8x – 10y = – 66 ……………………………(1)
40x – 18y = 214 ………………………….(2)

Let us suppose that the equation (1) is the equation of line of regression of y on x and (2) as the equation of the line of regression of x on y, after rewriting (1) and (2), we get y = $\frac{66}{10} + \frac{8}{10}$x which gives the value of $b_{yx} = \frac{8}{10}$

x = $\frac{214}{40} + \frac{18}{40}$y which gives the value of $b_{xy} = \frac{18}{40}$

Now $\quad r = \pm\sqrt{b_{yx}b_{xy}} = \pm\sqrt{\frac{8}{10}\times\frac{18}{40}} = \pm 0.6$

(ii) Since both the lines of regression passes through the mean values $\bar{x}$ and $\bar{y}$, the point ($\bar{x}$, $\bar{y}$) must satisfy the given two regression lines.

Therefore, 8 $\bar{x}$ – 10 $\bar{y}$ = – 66

40 $\bar{x}$ – 18 $\bar{y}$ = 214

Solving the above two equations we get $\bar{x}$ = 13 and $\bar{y}$ = 17

Important Note: In the above problem in part (i), if we take equation (1) as the line of regression of x on y, we get, x = $-\frac{66}{8} + \frac{10}{18}$y , and hence $b_{xy} = \frac{10}{8}$ and if we take equation (2) as the line of regression of y on x, we get, y = $-\frac{214}{18} + \frac{40}{18}$x and hence $b_{yx} = \frac{40}{18}$

Therefore, $\quad r = \pm\sqrt{b_{yx}b_{xy}} = \pm\sqrt{\frac{10}{8}\times\frac{40}{18}} = \pm 1.67$

But the value of r cannot exceed unity. Hence the assumptions that line (1) is line of regression of x on y and the line (2) is line of regression of y on x are wrong.

REFERENCES:

1. Vittal.P.R., Business Statistics, Margham publications, Chennai, 2008.
2. Gupta.S.C. Statistics, Himalaya Publishers, Mumbai, 2005.
3. S.P.Gupta, Business Statistics, Sultan Chand & Sons, New Delhi, 2008.
4. Beri.G., Business Statistics, Tata McGraw Hill Publishing Company Ltd, New Delhi, 2009.
5. Yule and Kendall (1993), Introduction to theory of Statistics. Universal Book Stall, New Delhi.

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – IV – TIMES SERIES – SMT1304

TIME SERIES

Components of time series – additive and multiplicative models – Measurement of Trend- Graphical Method – Semi average method – Moving average method – least square method- Measurement of seasonal variation – Method of simple average method – Ratio trend method- Ratio to Moving average method- Method of link relatives.

A **time series** is a set of observations taken at specified times, usually at equal intervals. In other words, a *series of observations recorded over time is known as a* **time series**. Examples of time series are the data regarding population of a country recorded at the ten-yearly censuses, annual production of a crop, say, wheat over a number of years, the wholesale price index over a number of months, the daily closing price of a share on the stock exchange, the hourly temperature recorded by weather bureau of a city, the total monthly sales receipts in business establishment, and so on. *In fact, data related with business and economic activities, in general, recorded over time give* rise to a **time series.**

One of the most important tasks before the planners and administrators in the field of economic and business activities is to make future estimates based on the past behaviour of a phenomenon under consideration. For example, trade cycles are important to economists and others in business and commerce. The behaviour of the cycles and their causes are of interest to them. Such studies are to be based on the analysis of time series data collected over time. *Thus,* **the analysis of time series** *plays an important role in empirical investigations of economic, commercial, social and even biological phenomena.*

Mathematically, **a time series** is defined by the fractional relationship

$$Y_t = f(t)$$

where Y_t is the value of the variable (or phenomenon) under consideration over time t. Thus, if the values of a variable at time points $t_1, t_2, ..., t_n$ are $Y_1, Y_2, ..., Y_N$ respectively, then the series

t	:	t_1	t_2	$t_3......, t_N$
Y_t	:	Y_1	Y_2	$Y_3......, Y_N$

constitute a time series.

COMPONENTS OF TIME SERIES:

Empirical studies of a number of time series have revealed the presence of certain **characteristic movements or fluctuations** in a time series. *These characteristic movements of a time series may be classified in four different categories called* **components of time series.** In a long time series, generally, we have the following **four components :**

1. Secular Trend or long–term movements
2. Seasonal variations
3. Cyclic variations
4. Random or Irregular movements

SECULAR TREND:

Secular trend *means the general long–term tendency of a series. In fact, secular trend is that characteristic of a time series which extends consistently throughout the entire period of time under consideration. It shows a long–term tendency of an activity to grow or to decline.* For example, a time series on population shows a tendency to increase; time series ot sales of a product shows a tendency to increase, and so on. On the other hand, a downward tendency is observed in the time series on birth and death rates. The factors which remain more or less constant over a long period also produce a trend. *The term* **'long period of time'** *is a relative phenomenon and cannot be defined exactly.* For some cases, a period as small as a week may be fairly long while in other cases, a period as long as 2 years may not be assumed long. For example, an increase in agricultural production over a period of two years would not be termed as secular change, whereas if the count of bacterial population of culture every five minutes, for a week shows an increase, then we would consider it as a secular change.

SEASONAL VARIATION:

The component responsible for the regular rise and fall in the magnitude of the time series is called **seasonal variation.** *In other words* **seasonal movements** or **seasonal variations** *refer to identical, or almost identical, patterns which a time series appears to follow during corresponding months of successive years.* Such variations are due to recurring events which takes place annually, quarterly, monthly, weekly or even daily, depending on the type of data available. *But in no case this period is to exceed one year. In view of their regular nature, seasonal variations are precise and can be foreseen,* as for instance the prices of agricultural commodities fall every year during the harvesting period, the sale of umbrellas pick up very fast in a rainy season, the demand for electric fans goes up during summer. Seasonal variations in general refer to annual periodicity in business and economic activities. These are the effects of seasonal factors like climatic conditions, human habits, fashions, customs and conventions of the people in a particular society.

CYCLICAL VARIATION:

Cyclical movements or **variations** *refer to the long-term oscillations or swings about a trend line. These cycles may or may not be periodic, i.e., they may or may not follow exactly similar patterns after equal intervals of time. Such variations are of longer duration than a year and they do not show the type of regularity as observed in the case of seasonal variations.* An important example of cyclical variations are the so-called **business cycles** representing intervals of **prosperity, recession, depression** and **recovery**. Each phase changes gradually into the phase which follows it in the given order. In a business activity, these phases follow each other with steady regularity and the period from the peak of one boom to the peak of the next boom is called a **complete cycle.** The usual periods of a business cycle may be ranging between 5–11 years. Most of the economic and business series relating to income, investment, wages, production shows this tendency.The study of cyclical fluctuations is therefore very important for predicting the turning phases in a business activity which may greatly help in proper policy formation in the area.

IRREGULAR VARIATION:

Random or Irregular *movements refer to such variations in a time series which do not repeat in a definite pattern.* Irregular movements in a time series may be of two types :

(i) Random or chance variations

(ii) Episodic variations

Random or chance variations *in a real phenomenon are inevitable by nature. It does effect a series in a random way, and as such, the effect of chance or random variations on a series is small.*

On the other hand, **episodic variations** *in a time series arise due to specific events or episodes like epidemic,* fire, strike or natural calamities like flood, earthquake or late monsoon etc. In some cases, irregular variations may not have a significant importance while in others these may be so intense as to result in new cyclical variations.

MEASUREMENT OF TREND:

The main objective behind the study of the trend of a time series are :

1. to describe the long–term growing or declining trend in a phenomenon under study.
2. to eliminate the trend component in order to bring into focus the remaining components in the time series.

In order to meet these objectives, some statistical methods of **estimation** *or* **determination of trend** *are as follows* :

1. Free hand, graphic method
2. Semi-average method
3. Moving average method
4. Method of least squares

GRAPHIC METHOD:

This is the simplest method of trend determination. According to this method, we plot the graph of the series and then draw a free hand curve through the points on the graph. Smoothing of time series data with a free hand curve eliminates the other components, viz., seasonal and irregular. The method does not involve complex mathematical calculations and can be used to describe all types of trend, linear or non-linear. However, the method is very subjective and can be adopted only to have a general idea of the nature of trend.

Example 1 **:** *Using the free hand hand or graphic method, fit a straight line trend to the following time series*

Year	1983	1984	1985	1986	1987	1988	1989	1990
Sales ('000)	80	90	85	92	87	99	93	120

Solution **:** Choosing a suitable scale, years are marked along the x-axis and corresponding sales values are marked along the y-axis. The points so obtained are then joined by straight lines which show the behaviour of sale values (actual data) over the given period. Then we draw a free hand straight line through the points of actual data for smoothing the time series data to obtain the trend. The behaviour of actual data and the trend line (dotted) are shown in fig. 1.

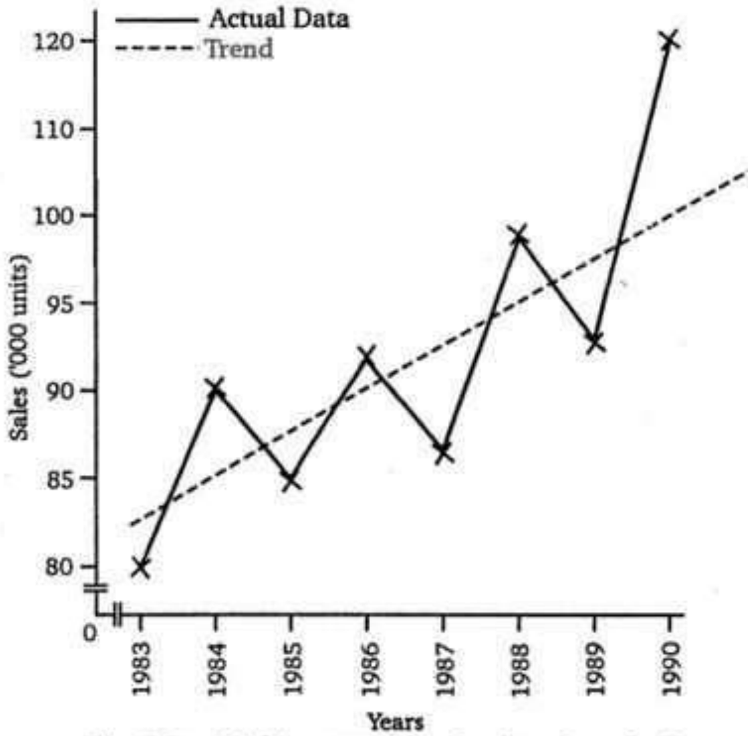

Fig. 1 Straight line trend by free hand method

SEMI- AVERAGE METHOD:

The method of semi-average is also simple. The method consists of dividing the data into two parts, preferably equal, and averaging the data in each part. In this way we obtain two points on the graph of the time series. The line obtained by joining these two points is the required trend line and may be extended in both the directions for estimating the trend values.

As compared with graphic method, the present method is better in view of its objectivity in the sense that every one who applies it would get the same results. However, the method has its limitation as it is appblicable only in a situation when the trend is linear or nearly linear. The following example will clarify the procedure.

Example 2 **:** *Determine straight line trend by semi-average method for the following time series data*

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Production ('000 units)	18	25	21	15	26	31	30	20	35	32	23

Solution **:** According to semi-average method, the given time series is divided into two parts. Here,the data about 11 years are given, thus the value corresponding to the middle year, *i. e.*, 1985 is ignored. The averages of first and the last five years are then computed as under :

	Year	**Production ('000 units)**	**Total Production**	**Semi average**	**Average year**
First five years	1980	18			
	1981	25			
	1982	21	→ 105	105 ÷ 5 = 21	1982
	1983	15			
	1984	26			
Last five years	1986	30			
	1987	20			
	1988	35	→ 140	140 ÷ 5 = 28	1988
	1989	32			
	1990	23			

MOVING AVERAGE METHOD:

The method of moving averages attempts to smooth out the irregularities in a series by a process of averaging. *By using averages of appropriate orders (or extent), cyclical, seasonal and irregular variations may be eliminated, thus leaving only the trend component.* **Moving averages of extent** *m (or period) is a series of successive averages of m terms at a time, starting from 1st, 2nd, 3rd terms and so on until we exhaust the whole time series.* if m is odd, say equal to $(2k + 1)$, then the moving average is put against the mid-value of the period it covers, *i. e.*, against $t = k + 1$. On the other hand, if m is even, say equal to $2k$, it is placed between two middle values of the period it covers. Thus when an even number of years is taken in moving average, the average does not coincide with an original time period. For overcoming this situation, moving average of extent two of these moving averages are taken and the first of such values is put against $t = k + 1$. This procedure of centering puts the moving averages against the time points of the series rather than between these points. Symbolically, the 3-yearly moving averages of a time series can be computed as shown in the following table.

Col. 1.	Col. 2.	Col. 3.	Col. 4 = Col. 3 ÷ 3
Years (t)	y_t	3-yearly moving totals	3-yearly moving averages
1	y_1	–	–
2	y_2	$\rightarrow (y_1 + y_2 + y_3)$	$(y_1 + y_2 + y_3)/3$
3	y_3	$\rightarrow (y_2 + y_3 + y_4)$	$(y_2 + y_3 + y_4)/3$
4	y_4	$\rightarrow (y_3 + y_4 + y_5)$	$(y_3 + y_4 + y_5)/3$
5	y_5	$\rightarrow (y_4 + y_5 + y_6)$	$(y_4 + y_5 + y_6)/3$
6	y_6	$\rightarrow (y_5 + y_6 + y_7)$	$(y_5 + y_6 + y_7)/3$
7	y_7	$\rightarrow (y_6 + y_7 + y_8)$	$(y_6 + y_7 + y_8)/3$
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
$N-1$	y_{N-1}	$\rightarrow (y_{N-2} + y_{N-1} + y_N)$	$(y_{N-2} + y_{N-1} + y_N)/3$
N	y_N	–	–

EXAMPLE 1:

Using three year moving averages determine the trend and short term fluctuations.

Year	: 1973	1974	1975	1976	1977	1978	1979	1980	1981	1982
Production ('000 tons)	: 21	22	23	25	24	22	25	26	27	26

Solution:

year	production	3 year moving total	3 year moving average	Short term fluctuation
1973	21	...	…	...
1974	22	66	22.00	0.00
1975	23	70	23.33	-0.33
1976	25	72	24.00	1.00
1977	24	71	23.67	0.33
1978	22	71	23.67	-1.67
1979	25	73	24.33	0.67
1980	26	78	26.00	0.00
1981	27	79	26.33	0.67
1982	26	…	…	…

Example: 2

Obtain trend for four yearly moving averages for the following data.

Year	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
Production	614	615	652	678	681	655	717	719	708	779	757

Solution : **Computation of trend by 4-yearly moving averages**

Year	Production	4-yearly moving totals	4-yearly centred moving totals	4-yearly moving averages (trend)
(1)	(2)	(3)	(4)	Col. (4) ÷ 8
1988	614			–
1989	615			–
		2559		
1990	652		5185	648.125
		2626		
1991	678		5292	661.500
		2666		
1992	681		5397	674.625
		2731		
1993	655		5503	687.875
		2772		
1994	717		5571	696.375
		2799		
1995	719		5722	715.250
		2923		
1996	708		5886	735.750
		2963		
1997	779			–
1998	757			–

In this case the following steps are followed

1. Calculate 4-yearly moving totals as usual. These are given in column (3).
2. For centring, we obtain two-yearly moving total of the 4-yearly moving totals as shown in column (4). Let us call such centred total as 4-yearly centred moving totals.
3. Finally divide the 4-yearly centred moving totals by 8 (4 × 2, *i.e.*, the period or extent of moving average × 2) to get the 4-yearly centred moving averages or Trend values.

Fitting curves by Method of Least Squares

Curve Fitting: Let (x_i, y_i); $i = 1, 2 \ldots n$ be a given set of n pairs of values, X being independent variable and Y being the dependent variable. The general problem in curve fitting is to find, if possible, an analytic expression of the form y = f(x), for the functional relationship suggested by the given data. Fitting of curves to a set of numerical data is of considerable importance theoretical as well as practical. Moreover, it may be used to estimate the values of one variable which would correspond to the specified values of the other variable.

Fitting a straight line

Let y = a + bx be the equation of the line to be fitted. To estimate the values of a and b we have, the following normal equations.

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2}$$

Here n is the number of observations, and the quantities $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i y_i$ and $\sum_{i=1}^{n} x_i^{\,2}$ can be obtained from the given set of points (x_i, y_i); i = 1, 2, …, n and the above equations can be solved for a and b.

EXAMPLE:

Below are given the figures of production (in 1000 tons) of a fertilizer factory.

Year	1997	1998	1999	2000	2001	2002	2003
Production	70	75	90	98	84	91	99

Fit a straight line trend by the method os least squares and estimate trend values for 2005.

[U.P.T.U. 2008]

Solution : We use the method of least squares to fit a straight line trend. Here, the trend line is

$$Y = a + bX$$

where Y is the production

we make the transformation

$$x = X - 2000 \qquad \text{...(i)}$$

Thus, the trend becomes

$$Y = a + bx \qquad \text{...(ii)}$$

Computation of trend by least squares method

Year (X)	Number (Y)	$x = X - 2000$	x^2	xY
1997	70	−3	9	−210
1998	75	−2	4	−150
1999	90	−1	1	90
2000	98	0	0	0
2001	84	1	1	84
2002	91	2	4	182
2003	99	3	9	297
$N = 7$	$\Sigma Y = 607$	$\Sigma x = 0$	$\Sigma x^2 = 28$	$\Sigma xY = 113$

The normal equations are

$$\Sigma Y = N a + \Sigma X$$

$$\Sigma xY = a \Sigma X + b \Sigma x^2$$

From the table, these equations becomes

$$607 = 7a + 0 \quad \Rightarrow \quad a = 86.7$$

$$113 = 0 + 28b \quad \Rightarrow \quad b = 4.03$$

Thus, the fitted trend line becomes

$$Y = 86.7 + 4.03x \quad \text{where} \quad x = X - 2000 \qquad \text{...(iii)}$$

Putting $x = -3, -2, -1, 0, 1, 2, 3$ in (iii) we can get trend values as follows :

Year	1997	1998	1999	2000	2001	2002	2003
Trend Values $Y = 86.7 + 4.03x$	74.61	78.64	82.67	86.7	90.73	94.76	98.79

Estimate of production for 2005 is

$$\hat{Y} = 86.7 + 4.03(2005 - 2000)$$

$$= 86.7 + 20.15$$

$$= 106.85$$

SEASONAL VARIATION:

As discussed earlier, *there are certain variations, called seasonal variations, which occur with certain degree of regularity within a definite period.* The period of variations may be a year, a month or even a day. A variety of causes may be listed for such variations. Some times climatic conditions affect production in agriculture and industries. For example, the sale of woollens picks up in every winter; prices of food grains come down in harvesting season; sale of cold drinks goes up during summer, etc. and so on. On the other hand, there are man-made factors which also cause such variations. For instance, the demand for consumer products goes up during the early part of month. The traffic in a city is high during the rush hours. When time series data are given in annual figures, it will not possess the seasonal variations. Thus, such variations are present only when data are given for specific periods of the year *i.e.*, the data are given quarterly, monthly, weekly, daily or hourly.

MEASURES OF SEASONAL VARIATION:

1. Method of averages
2. Moving Average Method
3. Ratio to moving average
4. Ratio to trend.

1. Method of Simple Averages

According to this method the data for each month (if monthly is given) are expressed as percentage of the average for the year. The method involves the following **steps** :

(i) Arrange the data by years and month (or quarters if quarterly data are given).

(ii) The figures for each month are added and averages are obtained by dividing the monthly totals by the number of years. Suppose the averages for the 12 months are denoted by $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_{12}$.

(iii) Then obtain the overall average of monthly averages as :

$$\bar{X} = \frac{\bar{X}_1 + \bar{X}_2 + \ldots + \bar{X}_{12}}{12}$$

(v) Obtain **seasonal indices** for different months by expressing the monthly averages as percentages of the overall average $\bar{X}$ in the following way :

$$\text{Seasonal Index for the first month} = \frac{\bar{X}_1}{\bar{X}} \times 100$$

$$\text{Seasonal Index for the second month} = \frac{\bar{X}_2}{\bar{X}} \times 100$$

... ...

$$\text{Seasonal Index for the twelfth month} = \frac{\bar{X}_{12}}{\bar{X}} \times 100$$

It should be noted that the average of the indices will always be 100, *i. e.*, the sum of the indices will be 1200 for 12 monthly data and the sum will be 400 for 4 quarterly data.

Example:

Assuming that the trend is absent, determine if there is any seasonality in the data given below

Year	Ist Quarter	2nd Quarter	3rd Quarter	4th Quarter
2004	3.7	4.1	3.3	3.5
2005	3.7	3.9	3.6	3.6
2006	4.0	4.1	3.3	3.1
2007	3.3	4.4	4.0	4.0

What are the seasonal indices for various quarters ? *(M. Com.. M.K. Univ.)*

Solution. COMPUTATION OF SEASONAL INDICES

Year	Ist Quarter	2nd Quarter	3rd Quarter	4th Quarter
2004	3.7	4.1	3.3	3.5
2005	3.7	3.9	3.6	3.6
2006	4.0	4.1	3.3	3.1
2007	3.3	4.4	4.0	4.0
Total	14.7	16.5	14.2	14.2
Average	3.675	4.125	3.55	3.55
Seasonal Index	98.66	110.74	95.30	95.30

Notes for calculating seasonal index

$$\text{The average of averages} = \frac{3.675 + 4.125 + 3.55 + 3.55}{4} = \frac{14.9}{4} = 3.725$$

$$\text{Seasonal Index} = \frac{\text{Quarterly average}}{\text{General average}} \times 100$$

$$\text{Seasonal Index for the first quarter} = \frac{3.675}{3.725} \times 100 = 98.66$$

$$\text{Seasonal Index for the second quarter} = \frac{4.125}{3.725} \times 100 = 110.74$$

$$\text{Seasonal Index for the third and fourth quarters} = \frac{3.55}{3.725} \times 100 = 95.30$$

2. Moving Average Method:

It is a method for computing trend values in a time series which eliminates the short and random fluctuations from the time series by means of moving average. Moving average of a period m is a series of successive arithmetic means of m terms at a time starting with 1 st, 2nd, 3rd so on. The first average is the mean of first m terms; the second average is the mean of 2 nd term to (m+1)th term and 3 rd average is the mean of 3rd term to (m+2)th term and so on. If m is odd then the moving average is placed against the mid value of the time interval it covers. But if m is even then the moving average lies between the two middle periods which does not correspond to any time period. So further steps has to be taken to place the moving average to a particular period of time. For that we take 2-yearly moving average of the moving averages which correspond to a particular time period. The resultant moving averages are the trend values.

Ex:1) Calculate 3-yearly moving average for the following data.

Years	Production	3-yearly moving avg (trend values)
1971-72	40	
1972-73	45	(40+45+40)/3 = 41.67
1973-74	40	(45+40+42)/3 = 42.33
1974-75	42	(40+42+46)/3 = 42.67
1975-76	46	(42+46+52)/3 = 46.67
1976-77	52	(46+52+56)/3 = 51.33
1977-78	56	(52+56+61)/3 = 56.33
1978-79	61	

Ex:1) Calculate 4-yearly moving average for the following data.

Years	Production	4-yearly moving avg	2-yealry moving avg (trend values)
1971-72	40		
1972-73	45		
		(40+45+40+42)/3 = 41.75	
1973-74	40		42.5
		(45+40+42+46)/3 = 43.15	
1974-75	42		44.12
		(40+42+46+52)/3 = 45	
1975-76	46		47
		(42+46+52+56)/3 = 49	
1976-77	52		51.38
		(46+52+56+61)/3 = 53.75	
1977-78	56		
1978-79	61		

3. Ratio to Trend Method:

Ratio-to-trend method is also known as **percentage trend method.** The method overcomes the difficulty of the simple average method when trend is present in the time series data. The method involves the following **steps** in measuring the seasonal indices :

(i) Compute the trend values by fitting trend equation to observed data by the method of least squares.

(ii) Express the original time series values as percentages of corresponding trend values.

(iii) Arrange these percentages according to years and months for monthly data (or according to years and quarters for quarterly data).

EXAMPLE:

The main defect of the ratio to trend method is that if there are cyclical swings in the series, the trend whether a straight line or a curve can never follow the actual data as closely as a 12-monthly moving average does. So a seasonal index computed by the ratio to moving average method may be less biased than the one calculated by the ratio to trend method.

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2003	30	40	36	34
2004	34	52	50	44
2005	40	58	54	48
2006	54	76	68	62
2007	80	92	86	82

Solution. For determining seasonal variation by ratio-to-trend method, first we will determine the trend for yearly data and then convert it to quarterly data.

CALCULATING TREND BY METHOD OF LEAST SQUARES

Year	*Yearly totals*	*Yearly average* Y	*Deviations from mid-year* X	XY	X^2	*Trend values*
2003	140	35	-2	-70	4	32
2004	180	45	-1	-45	1	44
2005	200	50	0	0	0	56
2006	260	65	$+1$	$+65$	1	68
2007	340	85	$+2$	$+170$	4	80
$N = 5$		$\Sigma Y = 280$		$\Sigma XY = 120$	$\Sigma X^2 = 10$	

The equation of the straight line trend is $Y = a + bX$.

$$a = \frac{\Sigma Y}{N} = \frac{280}{5} = 56 \qquad b = \frac{\Sigma XY}{\Sigma X^2} = \frac{120}{10} = 12$$

Quarterly increment = $\frac{12}{4} = 3$.

Calculation of Quarterly Trend Values. Consider 2003, trend value for the middle quarter, *i.e.*, half of 2nd and half of 3rd is 32. Quarterly increment is 3. So the trend value of 2nd quarter is $32 - \frac{3}{2}$, *i.e.*, 30.5 and for 3rd quarter is $32 + \frac{3}{2}$, *i.e.*, 33.5. Trend value for the 1st quarter is $30.5 - 3$, *i.e.*, 27.5 and of 4th quarter is $33.5 + 3$, *i.e.*, 36.5. We thus get quarterly trend values as shown below :

TREND VALUES

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2003	27.5	30.5	33.5	36.5
2004	39.5	42.5	45.5	48.5
2005	51.5	54.5	57.5	60.5
2006	63.5	66.5	69.5	72.5
2007	75.5	78.5	81.5	84.5

The given values are expressed as percentage of the corresponding trend values.

Thus for 1st Qtr. of 2003, the percentage shall be $(30/27.5) \times 100 = 109.09$, for 2nd Qtr. $(40/30.5) \times 100 = 131.15$, etc.

GIVEN QUARTERLY VALUES AS % OF TREND VALUES

Year	*1st Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
2003	109.09	131.15	107.46	93.15
2004	86.08	122.35	109.89	90.72
2005	77.67	106.42	93.91	79.34
2006	85.04	114.29	97.84	85.52
2007	105.96	117.20	105.52	97.04
Total	463.84	591.41	514.62	445.77
Average	92.77	118.28	102.92	89.15
S.I. Adjusted	92.05	117.36	102.12	88.46

Total of averages = $92.77 + 118.28 + 102.92 + 89.15 = 403.12$.

Since the total is more than 400 an adjustment is made by multiplying each average by $\frac{400}{403.12}$ and final indices are obtained.

1. Ratio to moving average:

Ratio-to-moving average or percentage moving average method consists of expressing the original time series data as percentages of moving averages instead of percentages of trend values as in '**ratio-to-trend method**', while rest of the steps are essentially the same. The procedure in this method consists of the following steps :

(i) Find the centred 12-monthly-moving averages (if monthly data are given) from the given time series data.

(ii) Express the original time series values as the percentage of the corresponding centred moving average values.

(iii) Average these percentages according to years and months and find averages over the years for all the 12 months.

(iv) Find the overall average of these 12-monthly averages. If the overall average is 100, the 12 monthly averages will be taken as seasonal indices, otherwise the monthly averages expressed as percentages of the overall average will be the required seasonal indices for the 12 months.

Symbolically, the logic behind the process may be explained as under :

The 12-monthly moving averages will eliminate the seasonal and irregular components and give us an estimate of the remaining two components namely trend (T) and cyclic (C). In multiplicative model we thus get an estimate of $T \times C$. Then the second step results in :

$$\frac{Y}{T \times C} \times 100 = \frac{T \times C \times S \times I}{T \times C} \times 100 = (S \times I) \times 100$$

Now on averaging over $S \times I$ in the third step, we are able to eliminate the irregular components with a possible bias. The final step givens us the adjusted seasonal indices.

Example 1:
Obtain seasonal indices by ratio to moving average method:

Year	Quarters			
	I	II	III	IV
2007	68	62	61	63
2008	65	58	66	61
2009	68	63	63	67

Solution **:** In the 'ratio-to-moving average' method, we first calculate 4 quarterly moving averages and ratios to moving averages as under :

Computation of Ratios to Moving Averages

Year and Quarter	Original data *Y*	4-quarterly moving totals	4-quarterly centred moving totals 4	4-quarterly centred moving averages (T)	Ratio to moving averages (percentage) = Y/T × 100
2007 I	68				
II	62				
	→	254			
III	61	→	505	63.125	96.63
	→	251			
IV	63	→	498	62.250	101.20
	→	247			
2008 I	65	→	499	62.375	104.21
	→	252			
II	58	→	502	62.750	92.43
	→	250			
III	66	→	503	62.875	104.97
	→	253			
IV	61	→	511	63.875	95.50
	→	258			
2009 I	68	→	513	64.125	106.04
	→	255			
II	63	→	516	64.500	97.67
	→	261			
III	63				
IV	67				

Again, the percentage of original data to moving averages are arranged according to years and quarters to obtain the seasonal indices as shown in the following table :

Computation of Seasonal Indices

Year	Percentages to moving averages			
	I	II	III	IV
2007	–	–	96.63	101.20
2008	104.21	92.43	104.97	65.50
2009	106.04	97.67	–	–
Totals	210.25	190.10	201.60	196.70
Averages	105.125	95.05	100.80	98.35
Adjusted Quarterly Indices	$\frac{105.125}{99.83}\times 100$ **= 105.30**	$\frac{95.05}{99.83}\times 100$ **= 95.21**	$\frac{100.80}{99.83}\times 100$ **= 100.97**	$\frac{98.35}{99.83}\times 100$ **= 98.52**

$$\textbf{Overall mean} = \bar{X} = \frac{105.125 + 95.05 + 100.80 + 98.35}{4} = \mathbf{99.83}$$

REFERENCES:

1. Vittal.P.R., Business Statistics, Margham publications, Chennai, 2008.
2. Gupta.S.C. Statistics, Himalaya Publishers, Mumbai, 2005.
3. S.P.Gupta, Business Statistics, Sultan Chand & Sons, New Delhi, 2008.
4. Beri.G., Business Statistics, Tata McGraw Hill Publishing Company Ltd, New Delhi, 2009.
5. Yule and Kendall (1993), Introduction to theory of Statistics. Universal Book Stall, New Delhi.

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – V –INDEX NUMBERS – SMT1304

INDEX NUMBERS

Construction of index numbers – Unweighted index numbers- Weighted index numbers- Laspeyr's method – Paasche's method – Dorbish and Bowley method – Marshall – Edge worth method- Fishers method – Kelly's method –quality index numbers– Chain index numbers –Base shifting – Splicing and deflating the index numbers – consumer Price index numbers.

An index number is a method of evaluating variations in a variable or group of variables in regards to the geographical location, time, and other features. The base value of the index number is usually 100 and indicates either to price, date, a level of production, etc.

There are various kinds of index numbers, however, in present; the most relatable is price index numbers, which particularly indicates the changes in overall price level (or in the value of money) for a particular time. Here, the value of money is not constant; even if it falls or rises it will affect and change the price level. An increase in the price level determines a decline in the value of money and a decline in the price level means an increase in the value of money. Therefore, the differences in the value of money are indicated by the differences in the overall price level for a particular time. Therefore, changes in the overall prices can be evaluated by a statistical device known as 'index number.'

INDEX NMBER:

Index numbers is a statistical tool for measuring relative change in a group of related variables over two or more different times.

- An index number is a statistical value that measures the change in a variable with respect to time.
- Two variables that are often considered in this analysis are price and quantity.
- With the aid of index numbers, the average price of several articles in one year may be compared with the average price of the same quantity of the same articles in a number of different years
- There are several sources of 'official' statistics that contain index numbers for quantities such as food prices, clothing prices, housing, wages and so on.

Features of an Index Number

a. They are expressed in percentages.
b. They are special types of averages.
c. They measure the effect of change over a period of time

Price indexes are of two types:

a. Simple Index Number
b. Weighted price Index numbers

Uses of Index Numbers.

a. Helps us to measure changes in price level
b. Help us to know changes in cost of living
c. Help government in adjustment of salaries and allowances
d. Useful to Business Community
e. Information to Politicians
f. Information regarding foreign trade

Construction of simple Index Numbers:
There are two methods

a. Simple aggregate Method $P01 = \frac{\sum P1}{\sum P0} \times 100$

b. Simple Average of price relative method $P01 = \frac{\sum P1/P0}{N} \times 100$

Where P1 = the price of an item in the current period
Po = the price of an item in the base period

Weighted Index Numbers There are two methods:
a. Weighted Aggregate method:
In this method commodities are assigned weights on the basis of quantities purchased.

$P01 = \frac{\sum P1Q0}{\sum P0Q0}$ Where Q0 = Quantity bought or sold in the base year.

b. Weighted Average of Price Relative Method:
Under this method commodities are assigned weight or the basis of base's year value (W= P0Q0) or fixed weights (W) are used.

$P01 = \frac{\sum RW}{\sum W}$ Where R = (P1/P0) X 100

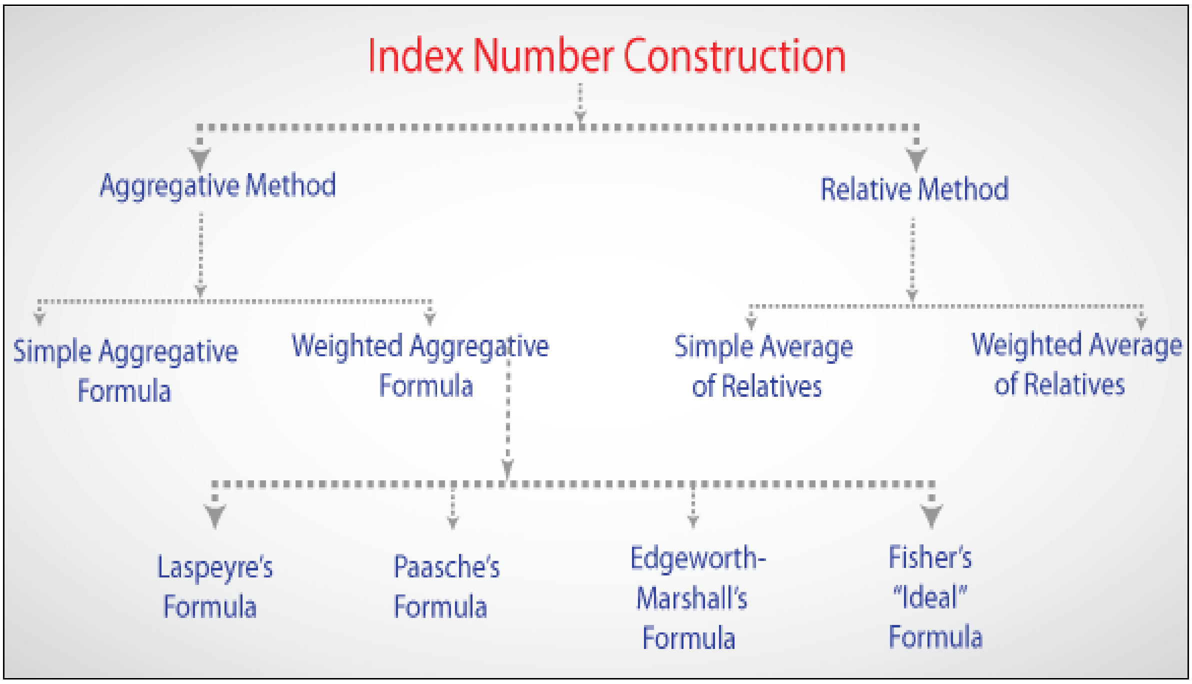

FIGURE 1

Example: 1
Simple Aggregative Method:

$$P_{01} = \frac{\Sigma P_1}{\Sigma P_0} \times 100$$

Where P_{01} Stands for the index number

ΣP_1 Stands for the sum of the prices for the year for which index number is to be found :

ΣP_0 Stands for the sum of prices for the base year.

Commodity	Prices in Base Year 1980 (in Rs.) P_o	Prices in current Year 1988 (in Rs.) P_1
A	10	20
B	15	25
C	40	60
D	25	40
Total	$\Sigma P_0 = 90$	$\Sigma P_1 = 145$

Index Number $(P_{01}) = \frac{\Sigma P_1}{\Sigma P_0} \times 100$; $P_{01} = \frac{145}{90} \times 100$; $P_{01} = 161.11$

Example: 2

Simple Average of price relative method:

$$P_{01} = \frac{\Sigma R}{N}$$

Where ΣR stands for the sum of price relatives i. e. $R = \frac{P_1}{P_0} \times 100$ and

N stands for the number of items.

Example

Commodity P_0	Base Year Prices (in Rs.) P_1	Current year Prices (in Rs.)	Price Relatives $R = \frac{P_1}{P_0} \times 100$
A	10	20	$\frac{20}{10} \times 100 = 200.0$
B	15	25	$\frac{25}{15} \times 100 = 166.7$
C	40	60	$\frac{60}{40} \times 100 = 150.00$
D	25	40	$\frac{40}{25} \times 100 = 160.0$
N = 4			$\Sigma R = 676.7$

Index Number $(p_{01}) = \frac{\Sigma R}{N}$

$$P_{01} = \frac{676.7}{4} ; P_{01} = 169.2$$

Weighted Aggregative Method:

(*i*) **Laspeyre's Formula.** In this formula, the quantities of base year are accepted as weights.

$$P_{01} = \frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times 100$$

Where P_1 is the price in the current year ; P_0 is the price in the base year ; and q_0 is the quantity in the base year.

(*ii*) **Paasche's Formula.** In this formula, the quantities of the current year are accepted as weights.

$$P_{01} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \times 100$$

Where q_1 is the quantity in the current year.

(*iii*) **Dorbish and Bowley's Formula.** Dorbish and Bowley's formula for estimating weighted index number is as follows :

$$P_{01} = \frac{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} + \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}}{2} \times 100 \quad \text{or} \quad p_{01} = \frac{L + P}{2}$$

Where L is Laspeyre's index and P is paasche's Index.

(*iv*) **Fisher's Ideal Formula.** In this formula, the geometric mean of two indices (i.e., Laspeyre's Index and paasche's Index) is taken :

$$p_{01} = \sqrt{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}} \times 100 \quad \text{or} \quad P_{01} = \sqrt{L \times P} \times 100$$

where L is Lespeyre's Index and P is paasche's Index.

Consumer Price Index: - (CPI)

The methods of constructing CPI are

Aggregate Expenditure Method: $P01 = \frac{(\sum P1Q0)X100}{\sum P0Q0}$

Family Budget Method: $P01 = \frac{\sum RW}{\sum W}$

Where R = (P1/P0) X 100 and W= P0Q0

Uses of Consumer Price Index: - (CPI)

a. It is used in calculating purchasing power of money
b. It is used for grant of Dearness Allowance.
c. It is used by government for framing wage policy, price policy etc.
d. CPI is used as price deflator of income
e. CPI is used as indicator of price movements in retail market.

Example

Comm-odity	Base Year		Current Year		P_0q_0	P_1q_0	P_0q_1	P_1q_1
	P_0	q_0	P_1	q_1				
A	10	5	20	2	50	100	20	40
B	15	4	25	8	60	100	120	200
C	40	2	60	6	80	120	240	360
D	25	3	40	4	75	120	100	160
Total					265 ΣP_0q_0	440 ΣP_1q_0	480 ΣP_0q_1	760 ΣP_1q_1

(*i*) Laspeyre's Formula :

$$p_{01} = \frac{\Sigma P_1q_0}{\Sigma P_0q_0} \times 100$$

$$p_{01} = \frac{440}{265} \times 100 = 166.04$$

(*ii*) Paasche' Formula :

$$p_{01} = \frac{\Sigma P_1q_1}{\Sigma P_0q_1} \times 100$$

$$p_{01} = \frac{700}{480} \times 100 = 158.3$$

(*iii*) Dorbish and Bowley's Formula :

$$p_{01} = \frac{\dfrac{\Sigma P_1q_0}{\Sigma P_0q_0} + \dfrac{\Sigma P_1q_1}{\Sigma P_0q_1}}{2} \times 100 = 162.2$$

$$p_{01} = \frac{\dfrac{440}{265} + \dfrac{760}{480}}{2} \times 100 = 162$$

(*iv*) Fisher's Ideal Formula :

$$p_{01} = \sqrt{\frac{\Sigma P_1q_0}{\Sigma P_0q_0} \times \frac{\Sigma P_1q_1}{\Sigma P_0q_1}} \times 100$$

$$p_{01} = \sqrt{\frac{440}{265} \times \frac{760}{480}} \times 100 = 162.1$$

REFERENCES:

1. Vittal.P.R., Business Statistics, Margham publications, Chennai, 2008.
2. Gupta.S.C. Statistics, Himalaya Publishers, Mumbai, 2005.
3. S.P.Gupta, Business Statistics, Sultan Chand & Sons, New Delhi, 2008.
4. Beri.G., Business Statistics, Tata McGraw Hill Publishing Company Ltd, New Delhi, 2009.
5. Yule and Kendall (1993), Introduction to theory of Statistics. Universal Book Stall, New Delhi.