KAMARAJ COLLEGE

SELF FINANCING COURSES

(Reaccredited with “A+” Grade by NAAC)

(Affiliated to Manonmaniam Sundaranar University, Tirunelveli.)

THOOTHUKUDI – 628003.

STUDY MATERIAL FOR B.B.A

BUSINESS STATISTICS

SEMESTER –I

ACADEMIC YEAR 2022-2023

PREPARED BY

BBA DEPARTMENT

BUSINESS STATISTICS – CABA11 & CASL11

SYLLABUS

UNIT I MEASURE OF CENTRAL TENDENCY

Measures of Central value- characteristics of an ideal measure- Measures of Central tendency – mean, median, mode – Application in Business decisions – Measures of Dispersion – absolute and relative measures of dispersion – Range, Quartile Deviation, Mean Deviation, Standard Deviation, Co-efficient of Variation – Moments, Skewness, Kurtosis - (Conceptual frame work only)

UNIT II CORRELATINON ANALYSIS

Correlation analysis: Meaning and Significance – Correlation and Causation, Types of Correlation, Methods of studying Simple Correlation – Scatter diagram, Karl Pearson's Coefficient of Correlation, Spearman's Rank Correlation co-efficient.

UNIT III REGRESSION ANALYSIS

Regression Analysis – Regression Vs Correlation, Linear Regression, Regression lines, Standard error of estimates.

UNIT IV TIME SERIES ANALYSIS

Time Series-Meaning and significance – utility, components of Time series- Measurement of Trend: Method of least squares, Parabolic Trend and Logarithmic trend.

UNIT V INDEX NUMBERS

Meaning and significance, problems in construction of index numbers, methods of constructing index numbers – weighted and unweighted, test of adequacy of index numbers, chain index numbers, base shifting, splicing and deflating index numbers

Table of Content

UNIT-I

MEASURES OF CENTRAL TENDANCY

Introduction

In the study of a population with respect to one in which we are interested we may get a large number of observations. It is not possible to grasp any idea about the characteristic when we look at all the observations. So it is better to get one number for one group. That number must be a good representative one for all the observations to give a clear picture of that characteristic. Such representative number can be a central value for all these observations. This central value is called a measure of central tendency or an average or a measure of locations.

There are five averages. Among them mean, median and mode are called simple averages and the other two averages geometric mean and harmonic mean are called special averages.

Definitions

The meaning of average is nicely given in the following definitions.

"A measure of central tendency is a typical value around which other figures congregate."

"An average stand for the whole group of which it forms a part yet represents the whole."

"One of the most widely used set of summary figures is known as measures of location."

Characteristics for a good or an ideal average

The following properties should possess for an ideal average.

1. It should be rigidly defined.
2. It should be easy to understand and compute.
3. It should be based on all items in the data.
4. Its definition shall be in the form of a mathematical formula.
5. It should be capable of further algebraic treatment.
6. It should have sampling stability.

7. It should be capable of being used in further statistical computations or processing.

Besides the above requisites, a good average should represent maximum characteristics of the data; its value should be nearest to the most items of the given series.

Arithmetic mean or mean

Arithmetic means or simply the mean of a variable is defined as the sum of the observations divided by the number of observations. If the variable x assumes n values $x_1, x_2 \ldots x_n$ then the mean, X, is given by

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots X_n}{N}$$

Example 1:

Calculate the mean for 2, 4, 6, 8, 10

Solution:

$X = \frac{2+4+6+8+10}{5}$

$X = \frac{30}{5}$

$X = 6$

Under this method an assumed or an arbitrary average (indicated by A) is used as the basis of calculation of deviations from individual values. The formula is

$$\bar{X} = A \pm \frac{\sum d}{n}$$

Where, A = the assumed mean or any value in x

d = the deviation of each value from the assumed mean

Example 2:

A student" s marks in 5 subjects are 75, 68, 80, 92, 56. Find his average mark.

X	d= X-A
75	(75-68) =7
68	0
80	12
92	24
56	-21
N=5	$\sum d = 31$

$\bar{X} = A \pm \frac{\sum d}{n}$

$= 68+ \frac{31}{5}$

= 68+6.2

= 74.2

Grouped Data: ***Discrete Series***

Direct Method

In case of *discrete series,* frequency against each of the observations is multiplied by the value of the observation. The values, so obtained, are summed up and divided by the total number of frequencies. Symbolically,

$$\bar{X} = \frac{\sum f x}{N}$$

Where *x* = the mid-point of individual class

f = the frequency of individual class

N = the sum of the frequencies or total frequencies

Example 3:

Given the following frequency distribution, calculate the arithmetic mean

Marks	64	63	62	61	60	59
Number of Students	8	18	12	9	7	6

Solution:

X	**F**	**Fx**	**d = X-A**	**fd**
64	8	512	2	16
63	18	1134	1	18
62 A	12	744	0	0
61	9	549	–1	– 9
60	7	420	–2	–14
59	6	354	–3	– 18
	N =60	**Σfx=3713**		**Σfd=– 7**

$$\bar{X} = \frac{\sum fx}{N}$$

$$\bar{x} = \frac{3713}{60}$$

$\bar{x} = 61.88$

Continuous Series

Here, class intervals are given. The process of calculating arithmetic mean in case of continuous series is same as that of a discrete series. The only difference is that the midpoints of various class intervals are taken. You should note that class intervals may be exclusive or inclusive or of unequal size. Example of exclusive class interval is, say, 0–10, 10–20 and so on. Example of inclusive class interval is, say, 0–9, 10–19 and so on. Example of unequal class interval is, say, 0–20, 20–50 and so on. In all these cases, calculation of arithmetic mean is done in a similar way.

$$\bar{X} = \frac{\sum fm}{N}$$

Step deviation method

$$\bar{X} = A \pm \frac{\sum fd}{N} \times C$$

Where

$d = \frac{X-A}{C}$

A = any value in x

N = total frequency

c = width of the class interval

Example 4:

Following is the distribution of persons according to different income groups. Calculate arithmetic mean.

Income Rs.(100)	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Number of persons	6	8	10	12	7	4	3

Income C.I	Number of Persons (f)	Mid M	$d = \frac{M-A}{C}$	Fd
0-10	6	5	-3	-18
10-20	8	15	-2	-16
20-30	10	25	-1	-10
30-40	12	35	0	0
40-50	7	45	1	7
50-60	4	55	2	8
60-70	3	65	3	9
	N=50			**Σfd= -20**

$$\bar{X} = A \pm \frac{\Sigma fd}{N} \times C$$

$\bar{x} = 35 - \frac{20}{50} \times 10$

$\bar{x} = 35 - \frac{200}{50}$

$\bar{x} = 35 - 4$

$\bar{x} = 31$

Merits and demerits of Arithmetic mean:

Merits:

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. If the number of items is sufficiently large, it is more accurate and more reliable.
4. It is a calculated value and is not based on its position in the series.
5. It is possible to calculate even if some of the details of the data are lacking.
6. Of all averages, it is affected least by fluctuations of sampling.
7. It provides a good basis for comparison.

Demerits:

1. It cannot be obtained by inspection nor located through a frequency graph.
2. It cannot be in the study of qualitative phenomena not capable of numerical measurement i.e., Intelligence, beauty, honesty etc.,
3. It can ignore any single item only at the risk of losing its accuracy.
4. It is affected very much by extreme values.
5. It cannot be calculated for open-end classes.
6. It may lead to fallacious conclusions, if the details of the data from which it is computed are not given.

Weighted Arithmetic mean

For calculating simple mean, we suppose that all the values or the sizes of items in the distribution have equal importance. But, in practical life this may not be so. In case some items are more important than others, a simple average computed is not representative of the distribution. Proper weightage has to be given to the various items. For example, to have an idea of the change in cost of living of a certain group of persons, the simple average of the prices of the commodities consumed by them will not do because all the commodities are not equally important, e.g rice, wheat and pulses are more important than tea, confectionery etc.,

It is the weighted arithmetic average which helps in finding out the average value of the series after giving proper weight to each group.

Definition:

The average whose component items are being multiplied by certain values known as "weights" and the aggregate of the multiplied results are being divided by the total sum of their "weight".

If $x_1, x_2 \ldots x_n$ be the values of a variable x with respective weights of $w_1, w_2 \ldots w_n$ assigned to them, then

$$\text{Weighted A.M} = \bar{X}_w = \frac{W_1 X_1 + W_2 X_2 + \cdots + W_i X_i}{W_1 + W_2 + \cdots W_n} = \frac{\sum W_i X_i}{\sum W_i}$$

Uses of the weighted mean:

Weighted arithmetic mean is used in:

a. Construction of index numbers.

b. Comparison of results of two or more universities where number of students differ.

c. Computation of standardized death and birth rates.

Example 5:

Calculate weighted average from the following data

Designation	Monthly salary (in Rs)	Strength of the cadre
Class 1 officers	1500	10
Class 2 officers	800	20
Subordinate staff	500	70
Clerical staff	250	100
Lower staff	100	150

Solution

Designation	Monthly salary x	Strength of the cadre, w	wx
Class 1 officer	1500	10	15,000
Class 2 officer	800	20	16,000
Subordinate staff	500	70	35,000
Clerical staff	250	100	25,000
Lower staff	100	150	15,000
		$\Sigma W=350$	$\Sigma wx= 1,06,000$

Weighted A.M= $\bar{X}_w = \frac{\sum wx}{\sum w}$

$$\bar{X}_w = \frac{106000}{350}$$

$$\bar{X}_w = 302.86$$

Harmonic mean (H.M)

Harmonic mean of a set of observations is defined as the reciprocal of the arithmetic average of the reciprocal of the given values. If $x_1, x_2 \ldots.. x_n$ are n observations,

$$H.M = \frac{n}{\sum_{i=1}^{n} f\left(\frac{1}{x_i}\right)}$$

For frequency distribution

$$H.M = \frac{N}{\sum_{i=1}^{n} f\left(\frac{1}{x_i}\right)}$$

Example 6:

From the given data calculate H.M 5, 10,17,24,30

X	$\frac{1}{x}$
5	0.2000
10	0.1000
17	0.0588
24	0.0417
30	0.0333
Total	**0.4338**

$$\text{H.M} = \frac{n}{\sum \frac{1}{x}}$$

$$\text{H.M} = \frac{5}{0.4338} = 11.526$$

Example 7:

The marks secured by some students of a class are given below. Calculate the harmonic mean.

Marks	20	21	22	23	24	25
Number of Students	4	2	7	1	3	1

Solution:

Marks (X)	No. of students (f)	$\frac{1}{x}$	$f(\frac{1}{x})$
20	4	0.0500	0.2000
21	2	0.0476	0.0952
22	7	0.0454	0.3178
23	1	0.0435	0.0435
24	3	0.0417	0.1251
25	1	0.0400	0.0400
	N=18		$\sum f\left(\frac{1}{x}\right) = 0.8216$

$$\text{H.M} = \frac{n}{\sum f\frac{1}{x}}$$

$$\text{H.M} = \frac{18}{0.8216} = 21.91$$

Merits of H.M :

1. It is rigidly defined.
2. It is defined on all observations.
3. It is amenable to further algebraic treatment.
4. It is the most suitable average when it is desired to give greater weight to smaller observations and less weight to the larger ones.

Demerits of H.M :

1. It is not easily understood.
2. It is difficult to compute.
3. It is only a summary figure and may not be the actual item in the series
4. It gives greater importance to small items and is therefore, useful only when small items have to be given greater weightage.

Geometric mean

The geometric mean of a series containing n observations is the n[th] root of the product of the values. If x_1, x_2…, x_n are observations then

$$G.M = \sqrt[n]{x_1 . x_2 \ldots x_n}$$

$$= (x_1 . x_2 \ldots x_3)^{\frac{1}{n}}$$

$$\text{Log } G.M = \frac{1}{n} \log (x_1 . x_2 \ldots x_n)$$

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

$$= \frac{\sum \log x_i}{n}$$

$$GM = \text{Antilog } \frac{\sum \log x_i}{n}$$

For grouped data

$$GM = \text{Antilog } [\frac{\sum f \log x_i}{N}]$$

Example 8:

Calculate the geometric mean of the following series of monthly income of a batch of families 180, 250, 490, 1400, 1050.

X	log x
180	2.2553
250	2.3979
490	2.6902
1400	3.1461
1050	3.0212
N=5	∑log x=13.5107

$$GM= \text{Antilog } \frac{\sum log\ x_i}{n}$$

$$= \text{Antilog } \frac{13.5107}{5}$$

$$= \text{Antilog } \frac{13.5107}{5}$$

$$= \text{Antilog } 2.7021 = 503.6$$

Example 9:

Calculate the average income per head from the data given below. Use geometric mean.

Class of people	Number of families	Monthly income per head (Rs.)
Landlords	2	5000
Cultivators	100	400
Landless-labours	50	200
Money-lenders	4	3750
Office Assistants	6	3000
Shop keepers	8	750
Carpenters	6	600
Weavers	10	300

Solution:

Class of people	**Annual income (Rs) X**	**Number of families (f)**	**Log x**	**f log x**
Landlords	5000	2	3.6990	7.398
Cultivators	400	100	2.6021	260.210
Landless-labours	200	50	2.3010	115.050
Money-lenders	3750	4	3.5740	14.296
Office Assistants 1	3000	6	3.477	20.863
Shop keepers	750	8	2.8751	23.2008
Carpenters	600	6	2.7782	16.669
Weavers	300	10	2.4771	24.771
		N=186		**482.257**

$$GM = \text{Antilog}\left[\frac{\sum f \log x_i}{N}\right]$$

$$GM = \text{Antilog}\left[\frac{482.257}{186}\right]$$

$$GM = \text{Antilog}\,[2.5928]$$

$$= \text{Rs. } 391.50$$

Merits of Geometric mean:

1. It is rigidly defined
2. It is based on all items
3. It is very suitable for averaging ratios, rates and percentages
4. It is capable of further mathematical treatment.
5. Unlike AM, it is not affected much by the presence of extreme values

Demerits of Geometric mean:

1. It cannot be used when the values are negative or if any of the observations is zero
2. It is difficult to calculate particularly when the items are very large or when there is a frequency distribution.
3. It brings out the property of the ratio of the change and not the absolute difference of change as the case in arithmetic mean.
4. The GM may not be the actual value of the series.

Median

Median is defined as the middle most observation when the observations are arranged in ascending or descending order of magnitude. That means the number of observations preceding median will be equal to the number of observations succeeding it. Median is denoted by M. The median is that value of the variate which divides the group into two equal parts, one part comprising all values greater, and the other, all values less than median

Ungrouped or Raw data:

Arrange the given values in the increasing or decreasing order. If the number of values are odd, median is the middle value. If the number of values are even, median is the mean of middle two values.

By formula

$$\text{Median (m)} = \frac{N+1}{2}^{\text{th}} \text{ item}$$

Example 10:

When odd number of values are given. Find median for the following data

25, 18, 27, 10, 8, 30, 42, 20, 53

Solution:

Arranging the data in the ascending order 8, 10, 18, 20, 25, 27, 30, 42, 53. The middle value is the 5th item i.e., 25 is the median.

Using this formula

$$\text{Median (m)} = \frac{N+1}{2}^{\text{th}} \text{ item}$$

$$\text{Median (m)} = \frac{9+1}{2}^{\text{th}} \text{ item}$$

$$\text{Median (m)} = \frac{10}{2}^{\text{th}} \text{ item}$$

$$\text{Median (m)} = 5^{\text{th}} \text{ item}$$

$$= 25$$

Example 11:

When even number of values are given. Find median for the following data 5, 8, 12, 30, 18, 10, 2, 22

Solution:

Arranging the data in the ascending order 2, 5, 8, 10, 12, 18, 22, 30. Using the formula

$$\text{Median (m)} = \frac{N+1}{2}^{\text{th}} \text{ item}$$

$$\text{Median (m)} = \frac{8+1}{2}^{\text{th}} \text{ item}$$

$$\text{Median (m)} = \frac{9}{2}^{\text{th}} \text{ item}$$

$$\text{Median (m)} = 4.5^{\text{th}} \text{ item}$$

Here median is the mean of the middle two items (ie) mean of (10,12) ie

$$= \frac{value\ \ of\ \ 4^{th}\ \ item + value\ \ of\ \ 5^{th}\ item}{2}$$

$$= \left(\frac{10+12}{2}\right) = \frac{22}{2} = 11$$

Example 12:

The following table represents the marks obtained by a batch of 10 students in certain class tests in statistics and Accountancy.

Serial No	1	2	3	4	5	6	7	8	9	10
Marks (Statistics)	53	55	52	32	30	60	47	46	35	28
Marks (Accountancy)	57	45	24	31	25	84	43	80	32	72

Indicate in which subject is the level of knowledge higher?

Solution:

For such question, median is the most suitable measure of central tendency. The mark in the two subjects is first arranged in ascending order as follows:

Serial No	1	2	3	4	5	6	7	8	9	10
Marks in Statistics	28	30	32	35	46	47	52	53	55	60
Marks in Accountancy	24	25	31	32	43	45	57	72	80	84

Median (m) = $\frac{N+1}{2}^{th}$ item

Median (m) = $\frac{10+1}{2}^{th}$ item

Median (m) = $\frac{11}{2}^{th}$ item

Median (m) = 5.5 th item

$$\frac{value\ \ of\ \ 5^{th}\ \ item + value\ \ of\ \ 6^{th}\ item}{2}$$

Median for Statistics $= \frac{46+47}{2} = 46.5$

Median for Accountancy $= \frac{43+45}{2} = 44$

Therefore, the level of knowledge in Statistics is higher than that in Accountancy.

Grouped Data:

In a grouped distribution, values are associated with frequencies. Grouping can be in the form of a discrete frequency distribution or a continuous frequency distribution. Whatever may be the type of distribution, cumulative frequencies have to be calculated to know the total number of items.

Cumulative frequency: (cf)

Cumulative frequency of each class is the sum of the frequency of the class and the frequencies of the pervious classes, i.e., adding the frequencies successively, so that the last cumulative frequency gives the total number of items.

Discrete Series:

Step1: Find cumulative frequencies.

Step2: Find $\frac{N+1}{2}$

Step3: See in the cumulative frequencies the value just greater than $\frac{N+1}{2}$

Step4: Then the corresponding value of x is median.

Example 13:

The following data pertaining to the number of members in a family. Find median size of the family.

Number of members x	1	2	3	4	5	6	7	8	9	10	11	12
Frequency F	1	3	5	6	10	13	9	5	3	2	2	1

Solution :

X	f	Cf
1	1	1
2	3	4
3	5	9
4	6	15
5	10	25
6	13	38
7	9	47
8	5	52
9	3	55
10	2	57
11	2	59
12	1	60
	N=60	

Median (m) $= \frac{N+1}{2}$ th item

Median (m) $= \frac{60+1}{2}$ th item

Median (m) $= \frac{61}{2}$ th item

Median (m) = 30.5 th item

The cumulative frequencies just greater than 30.5 is 38.and the value of x corresponding to 38 is 6.Hence the median size is 6 members per family.

Continuous Series:

The steps given below are followed for the calculation of median in continuous series.

Step1: Find cumulative frequencies.

Step2: Find $\frac{N}{2}$

Step3: See in the cumulative frequency the value first greater than $(\frac{N}{2})$, Then the corresponding class interval is called the Median class. Then apply the formula

$$median = l + \frac{\frac{N}{2} - cf}{f} \times C$$

Where

l = Lower limit of the median class

cf = cumulative frequency preceding the median

c = width of the median class

f = frequency in the median class.

N =Total frequency.

Example 14:

The following table gives the frequency distribution of 325 workers of a factory, according to their average monthly income in a certain year.

Income group (in Rs)	Number of workers
Below 100	1
100-150	20
150-200	42
200-250	55
250-300	62
300-350	45
350-400	30
400-450	25
450-500	15
500-550	18
550-600	10
600 and above	2
	325

Calculate median income

Solution :

Income group (in Rs) (Class-interval)	Number of workers (Frequency)	Cumulative frequency c.f
Below 100	1	1
100-150	20	21
150-200	42	63
200-250	55	118
250-300	62	180
300-350	45	225
350-400	30	255
400-450	25	280
450-500	15	295
500-550	18	313
550-600	10	323
600 and above	2	325
	N= 325	

$\frac{N}{2}=\frac{325}{2} = 162.5$

Here
l = 250,
N = 325,
f = 62,
c = 50,
cf = 118

$$\text{median} = l + \frac{\frac{N}{2} - cf}{f} \times c$$

$$\text{median} = 250 + \frac{162.5-118}{62} \times 50$$

$$\text{median} = 250 + \frac{44.5}{62} \times 50$$

$$\text{median} = 250 + \frac{162.5-118}{62} \times 50$$

$$\text{median} = 250 + \frac{2225}{62}$$

$$\text{median} = 250 + 35.887$$

$$\text{median} = 285.887$$

Example 15:

Calculate median from the following data

Value	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39
Frequency	5	8	10	12	7	6	3	2

Solution:

C	f	True class interval	c.f
0-4	5	0.5-4.5	5
5-9	8	4.5-9.5	13
10-14	10	9.5-14.5	23
15-19	12	14.5-19.5	35
20-24	7	19.5-24.5	42
25-29	6	24.5-29.5	48
30-34	3	29.5-34.5	51
35-39	2	34.5-39.5	53
	N=53		

$\frac{N}{2}=\frac{53}{2} = 26.5$

l = 14.5,

N = 53,

f = 12,c = 5, cf = 23

$$\text{median} = l + \frac{\frac{N}{2}-cf}{f} \times c$$

$$\text{median} = 14.5 + \frac{26.5-23}{12} \times 5$$

$$\text{median} = 14.5 + \frac{3.5}{12} \times 5$$

$$\text{median} = 14.5 + \frac{17.5}{12}$$

$$\text{median} = 14.5 + 1.46$$

$$\text{median} = 15.96$$

Note:

Since the variables are in the inclusive form, classes have to be adjusted. The difference between the upper limit of first class and lower limit of second class is one, in this problem. It is divided by 2, we get 0.5 which have been reduced lower limit of every class limit and have been added with the upper limit of every class.

Example 16:

Compute median for the following data.

Mid-value	5	15	25	35	45	55	65	75
Frequency	7	10	15	17	8	4	6	7

Solution :

The given problem is continuous series frequency distribution. Mid –values of the class limits are given. The difference between two mid – values is 10. Therefore 10/2 or 5 is reduced from each mid value to get the lower limit and 5 is added to get the upper limit of a class.

Mid x	C.I	f	c.f
5	0-10	7	7
15	10-20	10	17
25	20-30	15	32
35	30-40	17	49
45	40-50	8	57
55	50-60	4	61
65	60-70	6	67
75	70-80	7	74
		N=74	

Median = size of $\frac{N}{2}$ th item

= $\frac{74}{2}$ = 37 which lies in the class 30-40

l = 30,N = 74, f = 17, c = 10, cf = 32

$$median = l + \frac{\frac{N}{2} - cf}{f} \times C$$

$$median = 30 + \frac{37-32}{17} \times 10$$

$$median = 30 + \frac{5}{17} \times 10$$

$$median = 30 + \frac{50}{17}$$

$$median = 30 + 2.94$$

$$median = 32.94$$

Merits of Median:

1. Median is not influenced by extreme values because it is a positional average.
2. Median can be calculated in case of distribution with open-end intervals.
3. Median can be located even if the data are incomplete.
4. Median can be located even for qualitative factors such as ability, honesty etc.

Demerits of Median:

1. A slight change in the series may bring drastic change in median value.
2. In case of even number of items or continuous series, median is an estimated value other than any value in the series.
3. It is not suitable for further mathematical treatment except its use in mean deviation.
4. It is not taken into account all the observations

Mode

The mode refers to that value in a distribution, which occur most frequently. It is an actual value, which has the highest concentration of items in and around it. According to Croxton and Cowden "The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded at the most typical of a series of values".

It shows the centre of concentration of the frequency in around a given value. Therefore, where the purpose is to know the point of the highest concentration it is preferred. It is, thus, a positional measure.

Its importance is very great in marketing studies where a manager is interested in knowing about the size, which has the highest concentration of items. For example, in placing an order for shoes or ready-made garments the modal size helps because this sizes and other sizes around in common demand.

Computation of the mode:

Ungrouped or Raw Data:

For ungrouped data or a series of individual observations, mode is often found by mere inspection.

Example 17:

2, 7, 10, 15, 10, 17, 8, 10, 2

∴ Mode = M_0 = 10

In some cases, the mode may be absent while in some cases there may be more than one mode.

Example 18:

1) 12, 10, 15, 24, 30 (no mode)

2) 7, 10, 15, 12, 7, 14, 24, 10, 7, 20, 10

the modes are 7 and 10

Grouped Data:

For Discrete distribution, see the highest frequency and corresponding value of X is mode.

Continuous distribution:

See the highest frequency then the corresponding value of class interval is called the modal class. Then apply the formula.

$$\text{Mode} = M_0 = l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C$$

l = Lower limit of the model class

$\Delta_1 = f_1 - f_0$

$\Delta_2 = f_1 - f_2$

f_1 = frequency of the modal class

f_0= frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class

Example 19:

Calculate mode for the following:

C.I	F
0-50	5
50-100	14
100-150	40
150-200	91
200-250	150
250-300	87
300-350	60
350-400	38
400 and above	15

*

Solution:

The highest frequency is 150 and corresponding class interval is 200 – 250, which is the modal class.

l = Lower limit of the model class 200

f_1 = frequency of the modal class 150

f_0= frequency of the class preceding the modal class 91

f_2 = frequency of the class succeeding the modal class 87

$\Delta_1 = f_1 - f_0$ 150-91 = 59

$\Delta_2 = f_1 - f_2$ 150-87 =63

C=50

$$\text{Mode} = M_0 = l + \frac{\Delta_1}{\Delta_1+\Delta_2} \times c$$

$$= 200 + \frac{59}{59+63} \times 50$$

$$= 200 + \frac{2950}{122}$$

$$= 200 + 24.18$$

$$= 224.18$$

*

MEASURES OF DISPERSION

Introduction

The measure of central tendency serves to locate the centre of the distribution, but they do not reveal how the items are spread out on either side of the centre. This characteristic of a frequency distribution is commonly referred to as dispersion. In a series all the items are not equal. There is difference or variation among the values. The degree of variation is evaluated by

various measures of dispersion. Small dispersion indicates high uniformity of the items, while large dispersion indicates less uniformity. For example
consider the following marks of two students.

Student I	Student II
68	85
75	90
65	80
67	25
70	65

Both have got a total of 345 and an average of 69 each. The fact is that the second student has failed in one paper. When the averages alone are considered, the two students are equal. But first student has less variation than second student. Less variation is a desirable characteristic.

Characteristics of a good measure of dispersion

An ideal measure of dispersion is expected to possess the following properties

1. It should be rigidly defined
2. It should be based on all the items.
3. It should not be unduly affected by extreme items.
4. It should lend itself for algebraic manipulation.
5. It should be simple to understand and easy to calculate

Absolute and Relative Measures:

There are two kinds of measures of dispersion, namely

1. Absolute measure of dispersion
2.Relative measure of dispersion.

Absolute measure of dispersion indicates the amount of variation in a set of values in terms of units of observations. For example, when rainfalls on different days are available in mm, any absolute measure of dispersion gives the variation in rainfall in mm. On the other hand relative measures of dispersion are free from the units of measurements of the observations They are pure numbers. They are used to compare the variation in two or more sets, which are having different units of measurements of observations.

The various absolute and relative measures of dispersion are listed below.

Absolute measure	**Relative measure**
1. Range	1. Co-efficient of Range
2. Quartile deviation	2. Co-efficient of Quartile deviation
3. Mean deviation	3. Co-efficient of Mean deviation
4. Standard deviation	4. Co-efficient of variation

Range and coefficient of Range

Range:

This is the simplest possible measure of dispersion and is defined as the difference between the largest and smallest values of the variable.

In symbols, Range = L – S.

Where

L = Largest value.

S = Smallest value.

In individual observations and discrete series, L and S are easily identified. In continuous series, the following two methods are followed.

Method 1:

L = Upper boundary of the highest class

S = Lower boundary of the lowest class.

Method 2:

L = Mid value of the highest class.

S = Mid value of the lowest class.

Co-efficient of Range:

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

Example 1:

Find the value of range and its co-efficient for the following data. 7, 9, 6, 8, 11, 10

Solution:

L = 11, S = 4.

Range = L - S = 11- 4 = 7

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

Co-efficient of Range $=\frac{11-4}{11+4}$

Co-efficient of Range $=\frac{7}{15}$

Co-efficient of Range = 0.4667

Example 2:

Calculate range and its co efficient from the following distribution.

Size: 60-63 63-66 66-69 69-72 72-75

Number: 5 18 42 27 8

Solution:

L = Upper boundary of the highest class. = 75

S = Lower boundary of the lowest class. = 60

Range = L – S = 75 – 60 = 15

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

Co-efficient of Range $=\frac{75-60}{75+60}$

Co-efficient of Range $=\frac{15}{135}$

Co-efficient of Range = 0.1111

Merits and Demerits of Range:

Merits:

1. It is simple to understand.
2. It is easy to calculate.
3. In certain types of problems like quality control, weather forecasts, share price analysis, etc., range is most widely used.

Demerits:

1. It is very much affected by the extreme items.
2. It is based on only two extreme observations.
3. It cannot be calculated from open-end class intervals.
4. It is not suitable for mathematical treatment.
5. It is a very rarely used measure.

Quartile Deviation and Co efficient of Quartile Deviation

Quartile Deviation (Q.D):

Definition:

Quartile Deviation is half of the difference between the first and third quartiles. Hence, it is called Semi Inter Quartile Range.

In Symbols, Q .D = $\frac{Q3-Q1}{2}$ Among the quartiles Q1, Q2 and Q3, the range Q3 – Q1 is called inter quartile range and $\frac{Q3-Q1}{2}$, semi- inter quartile range.

Co-efficient of Quartile Deviation:

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

Example 3:

Find the Quartile Deviation for the following data: 391, 384, 591, 407, 672, 522, 777, 733, 1490, and 2488.

Solution:

Arrange the given values in ascending order.

384, 391, 407, 522, 591, 672, 733, 777, 1490, 2488.

Position of Q_1 is $\frac{N+1}{4} = \frac{10+1}{4} = \frac{11}{4} = 2.75$th item

Q1 = 2nd value + 0.75 (3rd value – 2nd value)

= 391 + 0.75 (407 – 391)

= 391 + 0.75 × 16

= 391 + 12

= 403

Position of Q3 is $3\frac{N+1}{4} = 3 \times 2.75 = 8.25$

Q3 = 8th value + 0.25 (9th value – 8th value)

= 777 + 0.25 (1490 – 777)

= 777 + 0.25 (713)

= 777 + 178.25 = 955.25

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

Q. D $= \frac{955.25-403}{2}$

*Q. D $= \frac{552.25}{2}$

Q.D = 276.125

Example 4:

Weekly wages of labours are given below. Calculate Q.D and Coefficient of Q.D.

Weekly Wage (Rs.):	100	200	400	500	600
No. of Weeks:	5	8	21	12	6

Weekly Wage (Rs.)	No. of Weeks	Cum. No. of Weeks
100	5	5
200	8	13
400	21	34
500	12	46
600	6	52
Total	**N = 52**	

Position of Q1 is $\frac{N+1}{4} = \frac{52+1}{4} = \frac{53}{4}$ = 13.25th item

Q1 = 13th value + 0.25 (14th Value – 13th value)

= 13th value + 0.25 (400 – 200)

= 200 + 0.25 (400 – 200)

= 200 + 0.25 (200)

= 200 + 50 = 250

Position of Q3 is $3\frac{N+1}{4} = 3 \times 13.25 = 39.75$

Q3 = 39th value + 0.75 (40th value – 39th value)

= 500 + 0.75 (500 – 500)

= 500 + 0.75 × 0

= 500

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

Q. D $=\frac{500-250}{2}$

Q. D $=\frac{250}{2}$

Q. D = 125

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

Co-efficient of Quartile Deviation = $\frac{500-250}{500+250}$

Co-efficient of Q.D = $\frac{250}{750}$

Co-efficient of Quartile Deviation = 0.3333

Example 5:

For the data given below, give the quartile deviation and coefficient of quartile deviation.

X :	351 – 500	501 – 650	651 – 800	801–950	951–1100
f :	48	189	88	47	28

Solution :

X	F	True class Intervals	Cumulative frequency
351-500	48	350.5-500.5	48
501-650 (Q_1)	189	500.5-650.5	237
651-800 (Q_3)	88	650.5-800.5	325
801-950	47	800.5-950.5	372
951-1100	28	950.5-1100.5	400
Total	**N = 400**		

Q1= l ± $\frac{\frac{N}{4}-Cf}{f}$× c

$\frac{N}{4} = \frac{400}{4} = 100$ which lies in $500.5 - 650.5$

l=500.5

64

cf= 48

f=189

c=150

Q1= 500.5 ± $\frac{100-48}{189}$× 150

Q1= 500.5 ± $\frac{52}{189}$× 150

Q1= 500.5 ± $\frac{7800}{189}$

$Q1 = 500.5 + 41.27$

$Q1 = 5431.77$

Q1= l ± $\frac{3\frac{N}{4}-Cf}{f}$× c

$3\frac{N}{4} = 3 \times 100 = 300$ which lies in $650.5 - 800.5$

Q3 Class is 650.5 – 800.5

l = 650.5,

cf = 237,

f = 88,

C = 150

Q3= 650.5 ± $\frac{300-237}{88}$× 150

Q3= 650.5 ± $\frac{63}{88}$× 150

Q3= 650.5 ± $\frac{300-237}{88}$× 150

$Q3 = 650.5 + 107.39$

$Q3 = 751.89$

Therefore

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

Q.D= $\frac{751.89-541.77}{2}$

Q.D= $\frac{216.12}{2}$

$\text{Q. D} = 108.06$

Co-efficient of Quartile Deviation = $\frac{751.89-541.77}{751.89+541.77}$

Co-efficient of Quartile Deviation = $\frac{216.12}{1299.66}$

Co-efficient of Quartile Deviation = 0.1663

Merits and Demerits of Quartile Deviation

Merits :

1. It is Simple to understand and easy to calculate
2. It is not affected by extreme values.
3. It can be calculated for data with open end classes also.

Demerits:

1. It is not based on all the items. It is based on two positional values Q_1 and Q_3 and ignores the extreme 50% of the items
2. It is not amenable to further mathematical treatment.
3. It is affected by sampling fluctuations.

Mean Deviation and Coefficient of Mean Deviation

Mean Deviation:

The range and quartile deviation are not based on all observations. They are positional measures of dispersion. They do not show any scatter of the observations from an average.

The mean deviation is measure of dispersion based on all items in a distribution.

Definition:

Mean deviation is the arithmetic mean of the deviations of a series computed from any measure of central tendency; i.e., the mean, median or mode, all the deviations are taken as positive i.e., signs are ignored. According to Clark and Schekade,

"Average deviation is the average amount scatter of the items in a distribution from either the mean or the median, ignoring the signs of the deviations".

We usually compute mean deviation about any one of the three averages mean, median or mode. Sometimes mode may be ill defined and as such mean deviation is computed from mean and median. Median is preferred as a choice between mean and median. But in general practice and due to wide applications of mean, the mean deviation is generally computed from mean. M.D can be used to denote mean deviation.

Coefficient of mean deviation:

Mean deviation calculated by any measure of central tendency is an absolute measure. For the purpose of comparing variation among different series, a relative mean deviation is required. The relative mean deviation is obtained by dividing the mean deviation by the average used for calculating mean deviation.

$$co\text{ efficient of mean deviation} = \frac{\text{mean deviation}}{\text{mean or median or mode}}$$

If the result is desired in percentage, the coefficient of mean deviation

$$\text{co efficient of mean deviation} \frac{\text{mean deviation}}{\text{mean or median or mode}} \times 100$$

Computation of mean deviation – Individual Series:

1. Calculate the average mean, median or mode of the series.
2. Take the deviations of items from average ignoring signs and denote these deviations by |D|.
3. Compute the total of these deviations, i.e., Σ |D|
4. Divide this total obtained by the number of items.

Symbolically: $M.D = \frac{\Sigma|D|}{n}$

Example 6:

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

Solution:

MEAN

$$\bar{X} = \frac{\sum X}{N} \quad ; \quad \bar{X} = \frac{3321}{9} = 369$$

MEDIAN

Now arrange the data in ascending order 100, 150, 200, 250, 360, 490, 500, 600, 671

Median= $\frac{N+1}{2}^{\text{th}}$ item Median= $\frac{9+1}{2}^{\text{th}}$ item Median= $\frac{10}{2}^{\text{th}}$ item

$Median$ = 5 th item $Median$ = 360

X	$\lvert D \rvert = \lvert X - \bar{X} \rvert$	$\lvert D \rvert = \lvert X - median \rvert$
100	269	260
150	219	210
200	169	160
250	119	110
360	9	0
490	121	130
500	131	140
600	231	240
671	302	311
3321	**$\sum D$ =1570**	**$\sum D$ =1561**

$$\text{M. D from mean} = \frac{\sum|D|}{n}$$

$$\text{M. D from mean} = \frac{1570}{9} = 174.44$$

$$\text{co efficient of mean deviation} = \frac{\text{mean deviation}}{\text{mean}}$$

$$\text{co efficient of mean deviation} = \frac{174.44}{369}$$

$$\text{co efficient of mean deviation} = 0.47$$

$$\text{M. D from median} = \frac{\sum|D|}{n}$$

$$M.D = \frac{1561}{9} = 173.44$$

$$co\ efficient\ of\ mean\ deviation = \frac{mean\ deviation}{median}$$

$$co\ efficient\ of\ mean\ deviation = \frac{173.44}{360} = 0.48$$

Mean Deviation – Discrete series:

Steps: 1. Find out an average (mean, median or mode).

2. Find out the deviation of the variable values from the average, ignoring signs and denote them by |D|

3. Multiply the deviation of each value by its respective frequency and find out the total Σf | D|

4. Divide Σf | D| by the total frequencies N

Symbolically, $M.D = \frac{\Sigma f\ |\ D|}{N}$

Example 7:

Compute Mean deviation from mean and median from the following data:

Height in cms	158	159	160	161	162	163	164	165	166
No. of persons	15	20	32	35	33	22	20	10	8

Also compute coefficient of mean deviation

Solution :

Height X	No. of persons F	d = x-A A = 162	Fd	|D| = |X-mean|	f|D|
158	15	-4	-60	3.51	52.65
159	20	-3	-60	2.51	50.20
160	32	-2	-64	1.51	48.32
161	35	-1	-35	0.51	17.85
162	33	0	0	0.49	16.17
163	22	1	22	1.49	32.78
164	20	2	40	2.49	49.80
165	10	3	30	3.49	34.90
166	8	4	32	4.49	35.92
	N=195		**∑fd=-95**		**∑f|D|=338.59**

$\bar{X} = A \pm \frac{\sum fd}{N}$

$\bar{X} = 162 - \frac{95}{195}$

$\bar{X} = 162 - 0.49$

$\bar{X} = 161.51$

$\text{M. D from mean} = \frac{\sum|D|}{n}$

$\text{M. D from mean} = \frac{338.59}{195}$

$\text{M. D from mean} = 1.74$

$\text{co efficient of mean deviation} = \frac{\text{mean deviation}}{\text{mean}}$

$\text{co efficient of mean deviation} = \frac{1.74}{161.51}$

$\text{co efficient of mean deviation} = 0.0108$

Height x	No. of persons f	c.f.	|D| = |X – Median|	f |D|
158	15	15	3	45
159	20	35	2	40
160	32	67	1	32
161	35	102	0	0
162	33	135	1	33
163	22	157	2	44
164	20	177	3	60
165	10	187	4	40
166	8	195	5	40
	N=195			**∑ f |D|=334**

$$\text{Median (m)} = \frac{N+1}{2}^{\text{th}} \text{ item}$$

$$\text{Median (m)} = \frac{195+1}{2}^{\text{th}} \text{ item}$$

$$\text{Median (m)} = 98^{\text{th}} \text{ item}$$

$$\text{Median (m)} = 161$$

$$\text{M. D from median} = \frac{\sum|D|}{n}$$

$$\text{M. D from median} = \frac{334}{195}$$

$$\text{M. D from median} = 1.71$$

$$\text{co efficient of mean deviation} = \frac{\text{mean deviation}}{\text{median}}$$

$$\text{co efficient of mean deviation} = \frac{1.71}{161}$$

$$\text{co efficient of mean deviation} = 0.0106$$

Mean deviation-Continuous series:

The method of calculating mean deviation in a continuous series same as the discrete series. In continuous series we have to find out the mid points of the various classes and take deviation of these points from the average selected. Thus

$$\text{M.D} = \frac{\Sigma f \mid D \mid}{N}$$

Where

D = M – average

M = Mid point

Example 8:

Find out the mean deviation from mean and median from the following series.

Age in years	No. of persons
0-10	20
10-20	25
20-30	32
30-40	40
40-50	42
50-60	35
60-70	10
70-80	8

Also compute co-efficient of mean deviation.

Solution:

X	M	F	$d = \frac{M-A}{C}$ (A=35, C=10)	Fd	$\lvert D \rvert = \lvert m - \bar{x} \rvert$	f |D|
0-10	5	20	-3	-60	31.5	630.0
10-20	15	25	-2	-50	21.5	537.5
20-30	25	32	-1	-32	11.5	368.0
30-40	**35**	40	0	0	1.5	60.0
40-50	45	42	1	42	8.5	357.0
50-60	55	35	2	70	18.5	647.5
60-70	65	10	3	30	28.5	285.0
70-80	75	8	4	32	**38.5**	308.0
		N=212		**∑fd=32**		**∑f|D|=3192.5**

$\bar{X} = A \pm \frac{\Sigma fd}{N} \times c$

$\bar{X} = 35 + \frac{32}{212} \times 10$

$\bar{X} = 35 + \frac{320}{212}$

$\bar{X} = 35 + 1.51$

$\bar{X} = 36.51$

$\text{M. D from mean} = \frac{\sum|D|}{n}$

$\text{M. D from mean} = \frac{3192.5}{212}$

$\text{M. D from mean} = 15.06$

$\text{co efficient of mean deviation} = \frac{\text{mean deviation}}{\text{mean}}$

$\text{co efficient of mean deviation} = \frac{15.06}{36.51}$

$\text{co efficient of mean deviation} = 0.41$

Calculation of median and M.D. from median

X	M	F	c.f	|D| = |m-Md|	f | D |
0-10	5	20	20	32.25	645.00
10-20	15	25	45	22.25	556.25
20-30	25	32	77	12.25	392.00
30-40	35	40	117	2.25	90.00
40-50	45	42	159	7.75	325.50
50-60	55	35	194	17.75	621.25
60-70	65	10	204	27.75	277.50
70-80	75	8	212	37.75	302.00
		N=212		Total	3209.50

Median= $l \pm \frac{\frac{N}{2} - Cf}{f} \times c$

$\frac{N}{2} = \frac{212}{2} = 106 \text{ which lies in } 30 - 40$

l=30

cf=77

f=40

c=10

Median= $30 \pm \frac{106-77}{40} \times 10$

Median= $30 \pm \frac{29}{40} \times 10$

Median= $30 \pm \frac{290}{40}$

$median = 30 + 7.25$

$median = 37.25$

$\text{M. D from median} = \frac{\sum|D|}{n}$

$\text{M. D from median} = \frac{3209.5}{212} = 15.14$

$\text{co efficient of mean deviation} = \frac{\text{mean deviation}}{\text{median}}$

$\text{co efficient of mean deviation} = \frac{15.14}{37.25}$

$\text{co efficient of mean deviation} = 0.41$

Merits and Demerits of M.D :

Merits:

1. It is simple to understand and easy to compute.
2. It is rigidly defined.
3. It is based on all items of the series.
4. It is not much affected by the fluctuations of sampling.
5. It is less affected by the extreme items.
6. It is flexible, because it can be calculated from any average.
7. It is better measure of comparison.

Demerits:

1. It is not a very accurate measure of dispersion.
2. It is not suitable for further mathematical calculation.
3. It is rarely used. It is not as popular as standard deviation.

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

Standard Deviation and Coefficient of variation

Standard Deviation:

Karl Pearson introduced the concept of standard deviation in 1893. It is the most important measure of dispersion and is widely used in many statistical formulae. Standard deviation is also called Root-Mean Square Deviation. The reason is that it is the square–root of the mean of the squared deviation from the arithmetic mean. It provides accurate result. Square of standard deviation is called Variance.

Definition:

It is defined as the positive square-root of the arithmetic mean of the Square of the deviations of the given observation from their arithmetic mean. The standard deviation is
denoted by the Greek letter **σ** (sigma)

Calculation of Standard deviation-Individual Series:

There are two methods of calculating Standard deviation in an individual series.

a) Deviations taken from Actual mean

b) Deviation taken from Assumed mean

a) Deviation taken from Actual mean:

This method is adopted when the mean is a whole number.

Steps:

1. Find out the actual mean of the series (x)
2. Find out the deviation of each value from the mean ($x = X - X$)
3. Square the deviations and take the total of squared deviations Σx2
4. Divide the total $\Sigma X2$ by the number of observations $\frac{\sum x^2}{2}$

The square root of $\frac{\sum x^2}{2}$ is standard deviation.

Thus $\sigma = \sqrt{\frac{\Sigma X^2}{N}}$ or $\sqrt{\frac{\Sigma(X-\bar{X})}{N}}$

b) Deviations taken from assumed mean:

This method is adopted when the arithmetic mean is fractional value. Taking deviations from fractional value would be a very difficult and tedious task. To save time and labour, we apply short–cut method; deviations are taken from an assumed mean. The
formula is:

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

Where d-stands for the deviation from assumed mean = (X-A)

Steps:

1. Assume any one of the item in the series as an average (A)
2. Find out the deviations from the assumed mean; i.e., X-A denoted by d and also the total of the
deviations Σd
3. Square the deviations; i.e., d2 and add up the squares of deviations, i.e, Σd2
4. Then substitute the values in the following formula:

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

Example 9:

Calculate the standard deviation from the following data. 14, 22, 9, 15, 20, 17, 12, 11

Solution:

Deviations from actual mean.

Values (X)	X - X	$(X-X)^2$
14	-1	1
22	7	49
9	-6	36
15	0	0
20	5	25
17	2	4
12	-3	9
11	-4	16
N=120		**$\sum(X-X)^2=140$**

$$\sigma=\sqrt{\frac{\Sigma(X-\bar{X})}{N}}$$

$$\sigma=\sqrt{\frac{\Sigma(X-\bar{X})}{N}}$$

$$\sigma=\sqrt{\frac{140}{8}}$$

$$\sigma=\sqrt{17.5}$$

$$\sigma=4.18$$

Example 10:

The table below gives the marks obtained by 10 students in statistics. Calculate standard deviation.

Student Nos :	1	2	3	4	5	6	7	8	9	10
Marks :	43	48	65	57	31	60	37	48	78	59

Solution: (Deviations from assumed mean)

Nos.	Marks (x)	d=X-A (A=57)	d^2
1	43	-14	196
2	48	-9	81
3	65	8	64
4	57	0	0
5	31	-26	676
6	60	3	9
7	37	-20	400
8	48	-9	81
9	78	21	441
10	59	2	4
N = 10		$\sum d = 44$	$\sum d^2 = 1952$

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

$$\sigma = \sqrt{\frac{1952}{10} - \left(\frac{44}{10}\right)^2}$$

$$\sigma = \sqrt{195.2 - 19.36}$$

$$\sigma = \sqrt{175.84}$$

$$\sigma = 13.26$$

Coefficient of Variation:

The Standard deviation is an absolute measure of dispersion. It is expressed in terms of units in which the original figures are collected and stated. The standard deviation of heights of students cannot be compared with the standard deviation of weights of students, as both are expressed in different units, i.e. heights in centimetre and weights in kilograms.

Therefore, the standard deviation must be converted into a relative measure of dispersion for the purpose of comparison. The relative measure is known as the coefficient of variation. The coefficient of variation is obtained by dividing the standard deviation by the mean and multiplies it by 100.

Symbolically,

$$\text{Coefficient of variation (C.V)} = \frac{\sigma}{X} \times 100$$

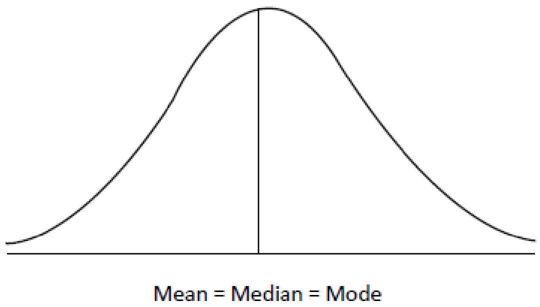
If we want to compare the variability of two or more series, we can use C.V. The series or groups of data for which the C.V. is greater indicate that the group is more variable, less stable, less uniform, less consistent or less homogeneous. If the C.V. is less, it indicates that the group is less variable, more stable, more uniform, more consistent or more homogeneous.

SKEWNESS

Introduction

Skewness means „lack of symmetry‟. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data. If in a distribution mean = median = mode, then that distribution is known as symmetrical distribution. If in a distribution mean ≠ median ≠ mode , then it is not a symmetrical distribution and it is called a skewed distribution and such a distribution could either be positively skewed or negatively skewed.

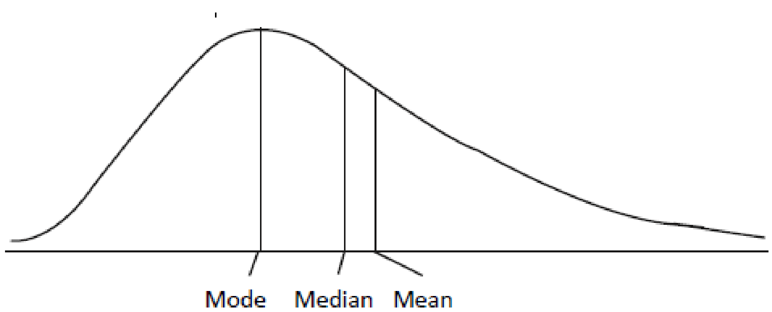
Symmetrical distribution:



It is clear from the above diagram that in a symmetrical distribution the values of mean, median and mode coincide. The spread of the frequencies is the same on both sides of the centre point of the curve.

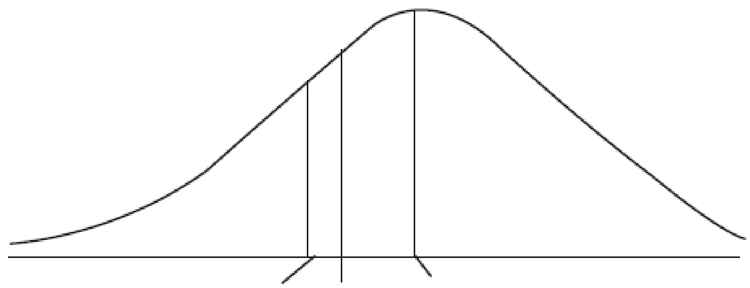
It is clear from the above diagram, in a positively skewed distribution, the value of the mean is maximum and that of the mode is least, the median lies in between the two. In the positively skewed distribution, the frequencies are spread out over a greater range of values on the right hand side than they are on the left hand side.

b) Positively skewed distribution:



It is clear from the above diagram, in a positively skewed distribution, the value of the mean is maximum and that of the mode is least, the median lies in between the two. In the positively skewed distribution the frequencies are spread out over a greater range of values on the right hand side than they are on the left hand side.

c) Negatively skewed distribution:



It is clear from the above diagram, in a negatively skewed distribution, the value of the mode is maximum and that of the mean is least. The median lies in between the two. In the negatively skewed distribution the frequencies are spread out over a greater range of values on the left hand side than they are on the right hand side.

Measures of skewness

The important measures of skewness are

(i) Karl –Pearason"s coefficient of skewness

(ii) Bowley"s coefficient of skewness

(iii) Measure of skewness based on moments

Karl – Pearson' s Coefficient of skewness:

According to Karl – Pearson, the absolute measure of skewness = ***mean - mode***. This measure is not suitable for making valid comparison of the skewness in two or more distributions because the unit of measurement may be different in different series. To avoid this difficulty, use relative measure of skewness called Karl–Pearson" s coefficient of skewness given by:

$$karl - person' scofficient\ of\ skewness = \frac{Mean - Mode}{S.D}$$

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

$$co\ efficient\ of\ skewness = \frac{3(Mean - Median)}{S.D}$$

Example 1:

Calculate Karl–Pearson's coefficient of skewness for the following data.

25, 15, 23, 40, 27, 25, 23, 25, 20

Solution:

Computation of Mean and Standard deviation

Size X	Deviation from A=25 X-A	d^2
25	0	0
15	-10	100
23	-2	4
40	15	225
27	2	4
25	0	0
23	-2	4
25	0	0
20	-5	25
N = 9	$\sum d = -2$	$\sum d^2 = 362$

$$\text{Mean}=A \pm \frac{\sum d}{n}$$

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

$$\text{Mean}=25 - 0.22$$

$$\text{Mean}=24.78$$

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

$$\sigma = \sqrt{\frac{362}{9} - \left(\frac{-2}{9}\right)^2}$$

$$\sigma = \sqrt{40.22 - 0.05}$$

$$\sigma = \sqrt{40.17}$$

σ = 6.3

Mode = 25, as this size of item repeats 3 times

Karl – Pearson's coefficient of skewness

$karl - person' scofficient\ of\ skewness = \frac{Mean - Mode}{S.D}$

$= \frac{24.78 - 25}{6.3} = \frac{-0.22}{6.3}$

$karl - person' scofficient\ of\ skewness =$-0.03

Example 2:

Find the coefficient of skewness from the data given below

Size :	3	4	5	6	7	8	9	10
Frequency :	7	10	14	35	102	136	43	8

Solution :

Size	Frequency (f)	d=X-A d=X-6	d^2	fd	fd^2
3	7	-3	9	-21	63
4	10	-2	4	-20	40
5	14	-1	1	-14	14
6	35	0	0	0	0
7	102	1	1	102	102
8	136	2	4	272	544
9	43	3	9	129	387
10	8	4	16	32	128
	N = 355			**$\sum fd = 480$**	**$\sum fd^2 = 1278$**

$\bar{X} = A \pm \frac{\Sigma fd}{N}$

$\bar{X} = 6 + \frac{480}{355}$

$\bar{X} = 6 + 1.35$

$\bar{X} = 7.35$

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

$\sigma = \sqrt{\frac{1278}{355} - \left(\frac{480}{355}\right)^2}$

$\sigma = \sqrt{3.6 - 1.82}$

$\sigma = \sqrt{1.78}$

$\sigma = 1.33$

Bowley's Coefficient of skewness:

In Karl – Pearson's method of measuring skewness the whole of the series is needed. Prof. Bowley has suggested a formula based on relative position of quartiles. In a symmetrical distribution, the quartiles are equidistant from the value of the median; i.e., Median – Q_1 = Q_3 – Median. But in a skewed distribution, the quartiles will not be equidistant from the median. Hence Bowley has suggested the following formula:

$$\text{Bowley's Coefficient of skewness (sk)} = \frac{Q_3 + Q_1 - 2\,\text{median}}{Q_3 - Q_1}$$

Example 4:

Find the Bowley's coefficient of skewness for the following series.

2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22

Solution:

The given data in order

Position of Q_1 is $= \frac{N+1}{4}$ *th item*

$= \frac{11+1}{4}\ th\ item = \frac{12}{4}\ th\ item$

= size of 3th item = 6

Position of Q3 is = $3\frac{N+1}{4}\ th\ item$

$= 3\frac{11+1}{4}\ th\ item$

= size of 9_{th} item

= 18

Median (m) = $\frac{N+1}{2}\ th\ item$

Median (m) = $\frac{11+1}{2}\ th\ item$

= size of 6_{th} item

=12

Bowley' s Coefficient of skewness (sk) $= \frac{Q_3 + Q_1 - 2\text{ median}}{Q_3 - Q_1}$

Bowley' s Coefficient of skewness (sk) $= \frac{18+6-2(12)}{18-6}$

Bowley' s Coefficient of skewness (sk) $= \frac{24 - 24}{18 - 6}$

Bowley' s Coefficient of skewness (sk) $= \frac{0}{12} = 0$

Since sk = 0, the given series is a symmetrical data.

UNIT-II

CORRELATION

Introduction

The term correlation is used by a common man without knowing that he is making use of the term correlation. For example when parents advice their children to work hard so that they may get good marks, they are correlating good marks with hard work.

The study related to the characteristics of only variable such as height, weight, ages, marks, wages, etc., is known as univariate analysis. The statistical Analysis related to the study of the relationship between two variables is known as Bi-Variate Analysis. Sometimes the variables may be inter-related. In health sciences we study the relationship between blood pressure and age, consumption level of some nutrient and weight gain, total income and medical expenditure, etc., The nature and strength of relationship may be examined by correlation and Regression analysis.

Thus, Correlation refers to the relationship of two variables or more. (e-g) relation between height of father and son, yield and rainfall, wage and price index, share and debentures etc.

Correlation is statistical Analysis which measures and analyses the degree or extent to which the two variables fluctuate with reference to each other. The word relationship is important. It indicates that there is some connection between the variables. It measures the closeness of the relationship. Correlation does not indicate cause and effect relationship. Price and supply, income and expenditure are correlated.

Definitions

1. Correlation Analysis attempts to determine the degree of relationship between variables-
Ya-Kun-Chou.

2. Correlation is an analysis of the co-variation between two or more variables.**- A.M. Tuttle.**

Correlation expresses the inter-dependence of two sets of variables upon each other. One variable may be called as (subject) independent and the other relative variable (dependent). Relative variable is measured in terms of subject.

Uses of correlation

1. It is used in physical and social sciences.
2. It is useful for economists to study the relationship between variables like price, quantity etc. Businessmen estimates costs, sales, price etc. using correlation.
3. It is helpful in measuring the degree of relationship between the variables like income and

expenditure, price and supply, supply and demand etc.

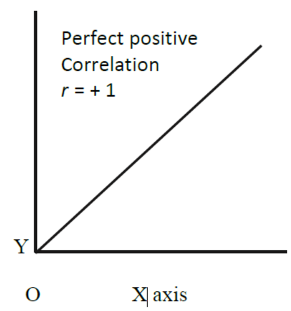
4. Sampling error can be calculated.

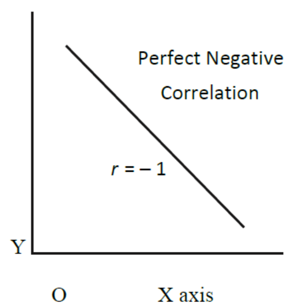
5. It is the basis for the concept of regression.

Scatter Diagram:

It is the simplest method of studying the relationship between two variables diagrammatically. One variable is represented along the horizontal axis and the second variable along the vertical axis. For each pair of observations of two variables, we put a dot in the plane. There are as many dots in the plane as the number of paired observations of two variables. The direction of dots shows the scatter or concentration of various points. This will show the type of correlation.

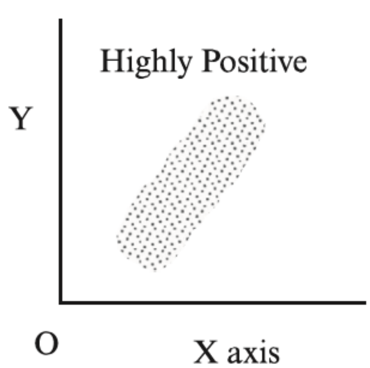
1. If all the plotted points form a straight line from lower left-hand corner to the upper right hand corner then there is Perfect positive correlation. We denote this as $r = +1$

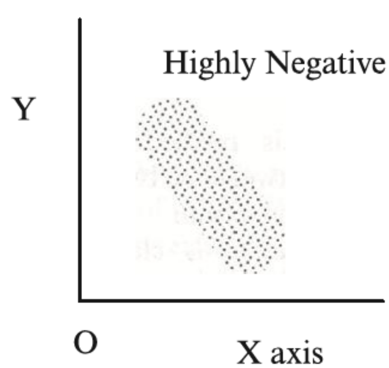




2. If all the plotted dots lie on a straight line falling from upper left-hand corner to lower right-hand corner, there is a perfect negative correlation between the two variables. In this case the coefficient of correlation takes the value $r = -1$.

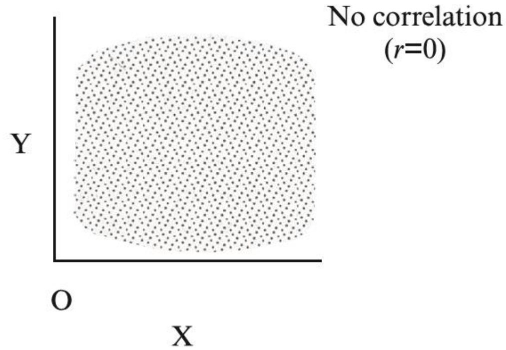
3. If the plotted points in the plane form a band and they show a rising trend from the lower left-hand corner to the upper right-hand corner the two variables are highly positively correlated.





1. If the points fall in a narrow band from the upper left-hand corner to the lower right-hand corner, there will be a high degree of negative correlation.

2. If the plotted points in the plane are spread all over the diagram there is no correlation between the two variables.



Merits:

1. It is a simplest and attractive method of finding the nature of correlation between the two variables.
2. It is a non-mathematical method of studying correlation. It is easy to understand.
3. It is not affected by extreme items.
4. It is the first step in finding out the relation between the two variables.
5. We can have a rough idea at a glance whether it is a positive correlation or negative correlation.

Demerits:

By this method we cannot get the exact degree or correlation between the two variables.

Types of Correlation

Correlation is classified into various types. The most important ones are

i) Positive and negative.

ii) Linear and non-linear.

iii) Partial and total.

iv) Simple and Multiple.

Positive and Negative Correlation:

It depends upon the direction of change of the variables. If the two variables tend to move together in the same direction (i.e.) an increase in the value of one variable is accompanied by an increase in the value of the other, (or) a decrease in the value of one variable is accompanied by a decrease in the value of other, then the correlation is called positive or direct correlation. Price and supply, height and weight, yield and rainfall, are some examples of positive correlation.

If the two variables tend to move together in opposite directions so that increase (or) decrease in the value of one variable is accompanied by a decrease or increase in the value of the other variable, then the correlation is called negative (or) inverse correlation. Price and demand, yield of crop and price, are examples of negative correlation.

Linear and Non-linear correlation:

If the ratio of change between the two variables is a constant then there will be linear correlation between them.

Consider the following.

X	2	4	6	8	10	12
Y	3	6	9	12	15	18

Here the ratio of change between the two variables is the same. If we plot these points on a graph, we get a straight line.

If the amount of change in one variable does not bear a constant ratio of the amount of change in the other. Then the relation is called Curvi-linear (or) non-linear correlation. The graph will be a curve.

Simple and Multiple correlation:

When we study only two variables, the relationship is simple correlation. For example, quantity of money and price level, demand and price. But in a multiple correlation we study more than two variables simultaneously. The relationship of price, demand and supply of a commodity are an example for multiple correlation.

Partial and total correlation:

The study of two variables excluding some other variable is called **Partial correlation**. For example, we study price and demand eliminating supply side. In total correlation all facts are taken into account.

Computation of correlation:

When there exists some relationship between two variables, we have to measure the degree of relationship. This measure is called the measure of correlation (or) correlation coefficient and it is denoted by **'r'** .

Algebraic or Mathematical Methods: Some of the methods of calculation of correlation coefficient are based on algebraic or mathematical treatment. The value of the coefficient of correlation by these formulae too remains between ±1. Following are the main mathematical methods –

1. Karl Pearson's Covariance method
2. Rank Correlation method

Co-variation:

The covariation between the variables x and y is defined as

$$\text{Cov}(x,y) = \frac{\sum(X-\bar{X})\,(Y-\bar{Y})}{N}$$

where $\bar{X}$, $\bar{Y}$ are respectively means of x and y and 'n' is the number of pairs of observations.

Karl Pearson's coefficient of correlation

Karl Pearson, a great biometrician and statistician, suggested a mathematical method for measuring the magnitude of linear relationship between the two variables. It is most widely used method in practice and it is known as pearsonian coefficient of correlation. It is denoted by '*r*'. The formula for calculating 'r' is

(i) $r = \frac{COV(x.y)}{\sigma_x.\sigma_y}$ where σ_x, σ_y are S.D of x and y

(ii) $r = \frac{\sum xy}{n\ \sigma_x\ \ \sigma_y}$

(iii) $r = \frac{\sum XY}{\sqrt{\sum X^2 . \sum Y^2}}$ $X = x - \bar{x}, Y = y - \bar{y}$

When the deviations are taken from the actual mean we can apply any one of these methods. Simple formula is the third one.

The third formula is easy to calculate, and it is not necessary to calculate the standard deviations of x and y series respectively.

Steps:

1. Find the mean of the two series x and y.
2. Take deviations of the two series from x and y.

$X = x - \bar{\mathbf{X}}, Y = y -$

3. Square the deviations and get the total, of the respective squares of deviations of x and y and denote by ΣX_2, ΣY_2 respectively.
4. Multiply the deviations of x and y and get the total and Divide by n. This is covariance.
5. Substitute the values in the formula.

$$r = \frac{COV(x.y)}{\sigma_x.\sigma_y} = \frac{\sum(x-\bar{x})\ \ (y-y)/n}{\sqrt{\frac{\sum(x-\bar{x})}{n}} \cdot \sqrt{\frac{\sum(y-\bar{y})}{n}}}$$

The above formula is simplified as follows

$$r = \frac{\sum XY}{\sqrt{\sum X^2 . \sum Y^2}} \qquad X = x - \bar{x}, Y = y - \bar{y}$$

Example 1:

Find Karl Pearson's coefficient of correlation from the following data between height of father (x) and son (y).

X	64	65	66	67	68	69	70
Y	66	67	65	68	70	68	72

Comment on the result.

X	**Y**	$X = x - \bar{X}$ **X=x-67**	X^2	**Y=y- $\bar{Y}$** **Y=y-68**	Y^2	**XY**
64	66	-3	9	-2	4	6
65	67	-2	4	-1	1	2
66	65	-1	1	-3	9	3
67	68	0	0	0	0	0
68	70	1	1	2	4	2
69	68	2	4	0	0	0
70	72	3	9	4	16	12
Σx=469	**Σy=476**	**0**	ΣX^2**=28**	**0**	ΣY^2**=34**	**Σ XY =25**

$$\bar{x} = \frac{\Sigma x}{n}$$
$$\bar{x} = \frac{469}{7}$$
$$\bar{x} = 67$$

$$\bar{y} = \frac{\Sigma y}{n}$$
$$\bar{y} = \frac{476}{7}$$
$$\bar{y} = 68$$

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \cdot \Sigma Y^2}}$$
$$r = \frac{25}{\sqrt{28*34}}$$
$$r = \frac{25}{\sqrt{952}}$$
$$r = \frac{25}{30.85} = 0.81$$

Since r = + 0.81, the variables are highly positively correlated. (i.e.) Tall fathers have tall sons.

Example 2

From the following data compute the co-efficient of correlation between X and Y:

	X Series	Y Series
No. of items	15	15
Arithmetic Mean	25	18
Square of deviations from mean	136	138

Summation of product of deviations of X and Y series from their respective Arithmetic Mean is 122X

Solution:

Denoting deviations of X and Y from the arithmetic means by x and y respectively the given data are

$\sum X^2 = 136; \quad \sum Y^2 = 138$

We apply Karl Pearson's method

$$r = \frac{\sum XY}{\sqrt{\sum X^2 . \sum Y^2}}$$

$$r = \frac{122}{\sqrt{136 \times 138}}$$

$$r = \frac{122}{\sqrt{18768}}$$

$$r = \frac{122}{137}$$

r = 0.89

Example 3. The following table gives the soil temperature and the germination time at various places. Calculate the co-efficient of correlation and interpret the value.

Temperature	57	42	40	38	42	45	42	44	40	46	44	43
Germination Time	10	26	30	41	29	27	27	19	18	19	31	29

Take 44 and 26 as assumed means

Solution:

We assume temperature as x and germination time as Y.

X	(X-44) dx	dx^2	Y	(Y-26) dy	dy^2	Dxdy
57	13	169	10	-16	256	-208
42	-2	4	26	0	0	0
40	-4	16	30	+4	16	-16
38	-6	36	41	+15	225	-90
42	-2	4	29	+3	9	-6
45	+1	1	27	+1	1	+1
42	-2	4	27	+1	1	-2
44	0	0	19	-7	49	0
40	-4	16	18	-8	64	+32
46	+3	4	19	-7	49	-14
44	0	0	31	+5	25	0
43	-1	1	29	+3	9	-3
N = 12	$\sum dx = -5$	$\sum dx^2 = 255$		$\sum dy = -6$	$\sum dy^2 = 704$	$\sum dxdy = -306$

$$r = \frac{\sum d_x d_y - \frac{(\sum d_x)(\sum d_y)}{N}}{\sqrt{\sum d_x{}^2 - \frac{(\sum d_x)^2}{N}} \cdot \sqrt{\sum d_y{}^2 - \frac{(\sum d_y)^2}{N}}}$$

$$r = \frac{-306 - \frac{(-5)(-6)}{12}}{\sqrt{255 - \frac{(-5)^2}{12}} \cdot \sqrt{704 - \frac{(-6)^2}{12}}}$$

$$r = \frac{-306 - \frac{30}{12}}{\sqrt{255 - \frac{25}{12}} \cdot \sqrt{704 - \frac{36}{12}}}$$

$$r = \frac{-306 - 2.5}{\sqrt{255 - 2.1} \cdot \sqrt{704 - 3}}$$

$$r = \frac{-308.5}{\sqrt{252.9}\ \sqrt{701}}$$

$$r = \frac{-308.5}{15.90 \times 26.47}$$

$$r = \frac{-308.5}{420.873}$$

$$r = -0.733$$

Rank Correlation

It is studied when no assumption about the parameters of the population is made. This method is based on ranks. It is useful to study the qualitative measure of attributes like honesty, colour, beauty, intelligence, character, morality etc. The individuals in the group can be arranged in order and there on, obtaining for each individual a number showing his/her rank in the group.

This method was developed by ***Edward Spearman*** in 1904. It is defined as –

$$r = 1 - \frac{6\sum D^2}{n^3 - n}$$

r = rank correlation coefficient.

Note: Some authors use the symbol ρ for rank correlation.

ΣD^2 = sum of squares of differences between the pairs of ranks.

n = number of pairs of observations.

The value of r lies between –1 and +1. If r = +1, there is complete agreement in order of ranks and the direction of ranks is also same. If r = -1, then there is complete disagreement in order of ranks and they are in opposite directions.

Computation for tied observations: There may be two or more items having equal values. In such case the same rank is to be given. The ranking is said to be tied. In such circumstances an
average rank is to be given to each individual item. For example, if the value so is repeated twice at the 5th rank, the common rank to be assigned to each item is $\frac{5+6}{2}$ = 5.5 which is the average of 5 and 6 given as 5.5, appeared twice.

If the ranks are tied, it is required to apply a correction factor which is $\frac{1}{12}$ (m3 – m). A slightly different formula is used when there is more than one item having the same value.
The formula is

$$r = 1 - \frac{6\,[\sum D^2 + \frac{1}{12}(m^3 - 3) + \frac{1}{12}(m^3 - 3) + \cdots]}{n^3 - n}$$

Where "m" is the number of items whose ranks are common and should be repeated as many times as there are tied observations.

Example 4:

In a marketing survey the price of tea and coffee in a town based on quality was found as shown below. Could you find any relation between tea and coffee price?

Price of tea	88	90	95	70	60	75	50
Price of coffee	120	134	150	115	110	140	100

Solution:

Price of tea	**Rank**	**Price of coffee**	**Rank**	**D**	D^2
88	3	120	4	1	1
90	2	134	3	1	1
95	1	150	1	0	0
70	5	115	5	0	0
60	6	110	6	0	0
75	4	140	2	2	4
50	7	100	7	0	0
					$\Sigma D^2=6$

$$r = 1 - \frac{6\sum D^2}{n^3 - n}$$

$$r = 1 - \frac{6 \times 6}{7^3 - 7}$$

$$r = 1 - \frac{36}{343 - 7}$$

$$r = 1 - \frac{36}{336}$$

$$r = 1 - 0.1071$$

$$r = 0.8929$$

Example 5:

In an evaluation of answer script the following marks are awarded by the examiners.

1st	88	95	70	60	50	80	75	85
2nd	84	90	88	55	48	85	82	72

Do you agree the evaluation by the two examiners is fair?

Solution :

X	**R1**	**y**	**R2**	**D**	**D^2**
88	2	84	4	2	4
95	1	90	1	0	0
70	6	88	2	4	16
60	7	55	7	0	0
50	8	48	8	0	0
80	4	85	3	1	1
85	3	75	6	3	9
					$\Sigma D^2=30$

$$r = 1 - \frac{6\sum D^2}{n^3 - n}$$

$$r = 1 - \frac{6 \times 30}{8^3 - 8}$$

$$r = 1 - \frac{180}{512 - 8}$$

$$r = 1 - \frac{180}{504}$$

$$r = 1 - 0.357$$

$$r = 0.643$$

Example 6:

Rank Correlation for tied observations. Following are the marks obtained by 10 students in a class in two tests.

Students	A	B	C	D	E	F	G	H	I	J
Test 1	70	68	67	55	60	60	75	63	60	72
Test 2	65	65	80	60	68	58	75	63	60	70

Calculate the rank correlation coefficient between the marks of two tests.

Solution:

Student	**Test 1**	**R1**	**Test 2**	**R2**	**D**	**D^2**
A	70	3	65	5.5	- 2.5	6.25
B	68	4	65	5.5	-1.5	2.25
C	67	5	80	1	4	16
D	55	10	60	8.5	1.5	2.25
E	60	8	68	4	4	16
F	60	8	58	10	-2	4
G	75	1	75	2	-1	1
H	63	6	63	7	-1	1
I	60	8	60	8.5	0.5	0.25
J	72	2	70	3	-1	1
						$\Sigma D^2=50$

60 is repeated 3 times in test 1

60, 65 is repeated twice in test 2

m=3; m=2; m=2

$$r = 1 - \frac{6\,[\sum D^2 + \frac{1}{12}(m^3 - 3) + \frac{1}{12}(m^3 - 3) + \cdots]}{n^3 - n}$$

$$r = 1 - \frac{6[50 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)]}{10^3 - 10}$$

$$r = 1 - \frac{6[50 + \frac{1}{12}(27 - 3) + \frac{1}{12}(8 - 2) + \frac{1}{12}(8 - 2)]}{1000 - 10}$$

$$r = 1 - \frac{6[50 + \frac{1}{12}(24) + \frac{1}{12}(6) + \frac{1}{12}(6)]}{990}$$

$$r = 1 - \frac{6[50 + \frac{24}{12} + \frac{6}{12} + \frac{6}{12}]}{990}$$

$$r = 1 - \frac{6[50 + 2 + 0.5 + 0.5]}{990}$$

$$r = 1 - \frac{6[53]}{990}$$

$$r = 1 - \frac{672}{990}$$

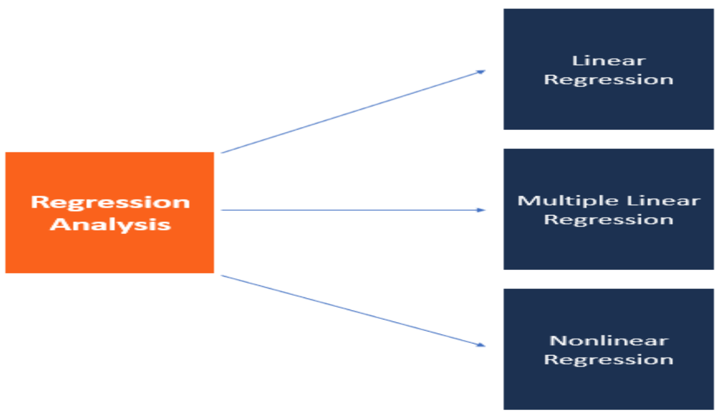
$$r = 1 - 0.678$$

$$r = 0.322$$

UNIT III

REGRESSION ANALYSIS

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them



Regression analysis includes several variations, such as linear, multiple linear, and nonlinear. The most common models are simple linear and multiple linear. Nonlinear regression analysis is commonly used for more complicated data sets in which the dependent and independent variables show a nonlinear relationship.

Regression Analysis – Linear Model Assumptions

Linear regression analysis is based on six fundamental assumptions:

1. The dependent and independent variables show a linear relationship between the slope and the intercept.
2. The independent variable is not random.
3. The value of the residual (error) is zero.
4. The value of the residual (error) is constant across all observations.
5. The value of the residual (error) is not correlated across all observations.
6. The residual (error) values follow the normal distribution.

Regression Analysis – Simple Linear Regression

Simple linear regression is a model that assesses the relationship between a dependent variable and an independent variable. The simple linear model is expressed using the following equation:

$Y = a + bX + \epsilon$

Where:

- **Y** – Dependent variable
- **X** – Independent (explanatory) variable
- **a** – Intercept
- **b** – Slope
- **ϵ** – Residual (error)

Regression Analysis – Multiple Linear Regression

Multiple linear regression analysis is essentially similar to the simple linear model, with the exception that multiple independent variables are used in the model. The mathematical representation of multiple linear regression is:

$Y = a + bX_1 + cX_2 + dX_3 + \epsilon$

Where:

- **Y** – Dependent variable
- **X_1, X_2, X_3** – Independent (explanatory) variables
- **a** – Intercept
- **b, c, d** – Slopes
- **ϵ** – Residual (error)

Multiple linear regression follows the same conditions as the simple linear model. However, since there are several independent variables in multiple linear analysis, there is another mandatory condition for the model:

- **Non-collinearity:** Independent variables should show a minimum correlation with each other. If the independent variables are highly correlated with each other, it will be difficult to assess the true relationships between the dependent and independent variables.

Distinguish Between Correlation and Regression.

Definition of Correlation -

The term correlation is a combination of two words 'Co' (together) and the relation between two quantities. Correlation is when it is observed that a change in a unit in one variable is retaliated by an equivalent change in another variable, i.e., direct or indirect, at the time of study of two variables. Or else the variables are said to be uncorrelated when the motion in one variable does not amount to any movement in a specific direction in another variable. It is a statistical technique that represents the strength of the linkage between variable pairs.

Correlation can be either negative or positive. If the two variables move in the same direction, i.e. an increase in one variable results in the corresponding increase in another variable, and vice versa, then the variables are considered to be positively correlated. For example, Investment and profit.

On the contrary, if the two variables move in different directions so that an increase in one variable leads to a decline in another variable and vice versa, this situation is known as a negative correlation. For example, Product price and demand.

Definition of Regression-

A statistical technique based on the average mathematical relationship between two or more variables is known as regression, to estimate the change in the metric dependent variable due to the change in one or more independent variables. It plays an important role in many human activities since it is a powerful and flexible tool that is used to forecast past, present, or future events based on past or present events. For example, The future profit of a business can be estimated on the basis of past records.

There are two variables x and y in a simple linear regression, wherein y depends on x or say that is influenced by x. Here y is called as a variable dependent, or criterion, and x is a variable independent or predictor. The line of regression y on x is expressed as below:

$Y = a + bx$

where, a = constant,

b = regression coefficient,

The a and b are the two regression parameters in this equation.

Difference between Correlation and Regression

BASIS FOR COMPARISON	CORRELATION	REGRESSION
Meaning	Correlation is a statistical measure that determines the association or co-relationship between two variables.	Regression describes how to numerically relate an independent variable to the dependent variable.
Usage	To represent a linear relationship between two variables.	To fit the best line and to estimate one variable based on another.
Dependent and Independent variables	No difference	Both variables are different.
Indicates	Correlation coefficient indicates the extent to which two variables move together.	Regression indicates the impact of a change of unit on the estimated variable (y) in the known variable (x).
Objective	To find a numerical value expressing the relationship between variables.	To estimate values of random variables on the basis of the values of fixed variables.

REGRESSION LINES

A regression line is a line that best describes the linear relationship between the two variables. It is expressed by means of an equation of the form:

$$y = a + bx$$

The regression equation of x on y is:

$$X = a + bY$$

The regression equation of Y on X is:

$$Y = a + bX$$

Example 1: Calculate the regression coefficient and obtain the lines of regression for the following data

X	1	2	3	4	5	6	7
Y	9	8	10	12	11	13	14

Solution:

X	Y	X^2	Y^2	XY
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
$\sum X=28$	$\sum Y=77$	$\sum X^2=140$	$\sum Y^2=875$	$\sum XY=334$

$$\bar{X} = \frac{\sum X}{N} = \frac{28}{7} = 4 \qquad \bar{Y} = \frac{\sum Y}{N} = \frac{77}{7} = 11$$

Regression Coefficient of X on Y

$$b_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum Y^2 - (\sum Y)^2}$$

$$= \frac{7(334) - (28)(77)}{7(875) - (77)^2}$$

$$= \frac{2338 - 2156}{6125 - 5929}$$

$$= \frac{182}{196}$$

$b_{XY} = 0.929$

Regression equation of X on Y

$X - X = b_{xy} (Y - Y)$

$X - 4 = 0.929 (Y - 11)$

$X - 4 = 0.929Y - 10.219$

The regression equation X on Y is **$X = 0.929Y - 6.219$**

Regression coefficient of Y on X

$$b_{yx} = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2}$$

$$= \frac{7(334) - (28)(77)}{7(140) - (28)^2}$$

$$= \frac{2338 - 2156}{980 - 784}$$

$$= \frac{182}{196}$$ **$b_{yx} = 0.929$**

Regression equation of Y on X

$\mathbf{Y - \overline{Y} = b_{yx} (X - \overline{X})}$

Y – 11 = 0.929 (X - 4) ________

Y = 0.929X – 3.716 + 11

The regression equation Y on X is **Y = 0.929X – 6.219**

Example 2: Consider the following set of points: {(-2 , -1) , (1 , 1) , (3 , 2)}

Find the least square regression line for the given data points.

x	y	x y	x^2
-2	-1	2	4
1	1	1	1
3	2	6	9
Σx = 2	Σy = 2	Σxy = 9	$\Sigma x^2 = 14$

$$a = \frac{(n\Sigma xy - \Sigma x\Sigma y)}{(n\Sigma x^2 - (\Sigma x)^2)}$$

$$= \frac{(3*9 - 2*2)}{(3*14 - 2^2)}$$

$$= 23/38$$

$$b = (1/n)(\Sigma y - a\ \Sigma x)$$

$$= (1/3)(2 - (23/38)*2)$$

$$= 5/19$$

Example 3: The sales of a company (in million dollars) for each year are shown in the table below.

x (year)	2005	2006	2007	2008	2009
y (sales)	12	19	29	37	45

a) Find the least square regression line y = a x + b.

b) Use the least squares regression line as a model to estimate the sales of the company in 2012

Solution: We first change the variable x into t such that t = x - 2005 and therefore t represents the number of years after 2005. Using t instead of x makes the numbers smaller and therefore manageable. The table of values becomes.

t (years after 2005)	0	1	2	3	4
y (sales)	12	19	29	37	45

We now use the table to calculate a and b included in the least regression line formula.

t	y	t y	t^2
0	12	0	0
1	19	19	1
2	29	58	4
3	37	111	9
4	45	180	16
Σx = 10	Σy = 142	Σxy = 368	$\Sigma x^2 = 30$

We now calculate a and b using the least square regression formulas for a and b.

$a = (n\Sigma t\, y - \Sigma t\Sigma y) / (n\Sigma t^2 - (\Sigma t)^2) = (5*368 - 10*142) / (5*30 - 10^2) = 8.4$

$b = (1/n)(\Sigma y - a\, \Sigma x) = (1/5)(142 - 8.4*10) = 11.6$

b) In 2012, t = 2012 - 2005 = 7

The estimated sales in 2012 are: y = 8.4 * 7 + 11.6 = 70.4 million dollars

UNIT IV

TIME SERIES

Meaning

- A time series is a data set that tracks a sample over time.
- In particular, a time series allows one to see what factors influence certain variables from period to period.
- Time series analysis can be useful to see how a given asset, security, or economic variable changes over time.
- Forecasting methods using time series are used in both fundamental and technical analysis.
- Although cross-sectional data is seen as the opposite of time series, the two are often used together in practice.

Definition

"A time series is a set of observation taken at specified times, usually at equal intervals"

"A time series may be defined as a collection of reading belonging to different time periods of some economic or composite variables"

-Ya-Lun-Chau

- Time series establish relation between **"cause"** and **"effects"**
- One variable is **"Time"** which is independent variable and the second is **"Data"** which is the dependent variable.

Importance of Time Series Analysis

As the basis of Time series analysis businessman can predict about the changes in economy. There are following points which clear about its importance:

1. Profit of experience
2. Safety from future
3. Utility studies
4. Sales Forecasting
5. Stock market analysis
6. Process and quality control

7. Budgetary analysis
8. Yield projections
9. Inventory studies
10. Economic forecasting
11. Risk analysis and evaluation of changes
12. Census analysis

Components of Time Series

The change which are being in time series, they are affected by economic, social, natural, industrial and political reasons. These reasons are called components of Time Series.

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

Secular Trend

The increase or decrease in the movements of a time series is called Secular trend. A time series data may show upward trend or download trend for a period of years and this may be due to factors like:

- Increase in population
- Change in technological progress
- Large scale shift in consumer demands

Seasonal variation

Seasonal variation is short term fluctuation in a time series which occur periodically in a year. This continues to repeat year after year.

- The major factors that are weather conditions and customs of people.
- More woolen cloths are sold in winter than in the season of summer.
- Each year more ice creams are sold in summer and very little in winter season.
- The sales in the departmental stores are more during festive seasons that in the normal days.

Cyclical variations

Cyclical variations are recurrent upward or downward movements in a time series but the period of cycle is greater than a year. Also these variations are not regular as seasonal variations.

A business cycle showing these oscillatory movements has to pass through four phases-prosperity, recession, depression and recovery. In a business, these four phases are completed by passing one to another in this order.

Irregular variation

Irregular variations are fluctuations in time series that are short in duration, erratic in nature and follow no regularity in the occurrence pattern. These variations are also referred to as residual variations since by definitions they represent what is left out in a time series after trend, cyclical and seasonal variations. Irregular fluctuations result due to the occurrence of unforeseen events like:

- Floods
- Earthquakes
- Wars
- Famines

MEASUREMENT OF SECULAR TREND

The following methods are used for calculation of trend

- Free Hand Curve or Graphic Method
- Semi-Average Method
- Moving Average Method
- Least Square Method

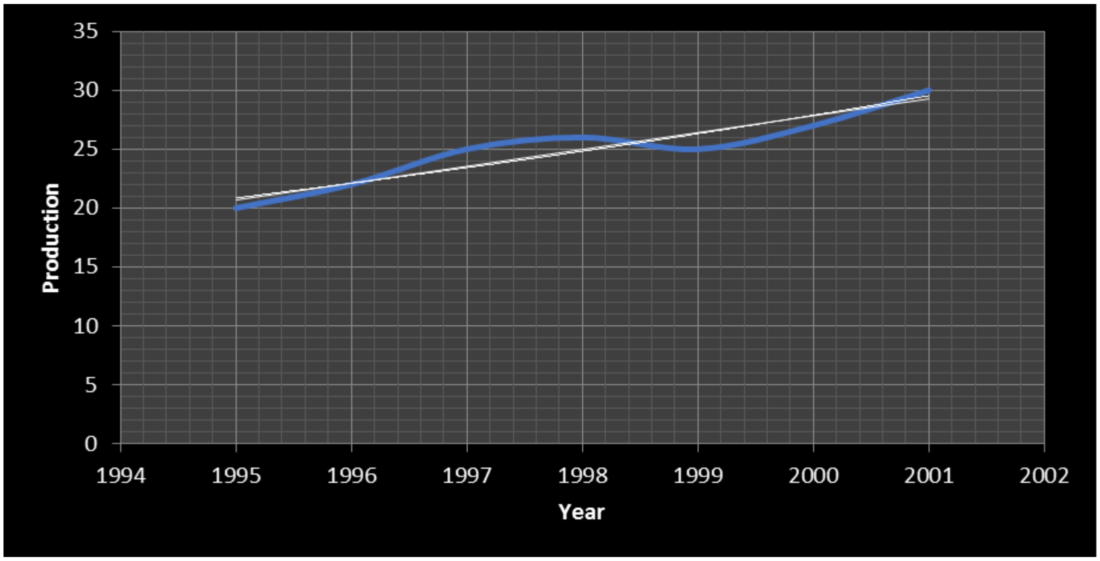
Free Hand Curve or Graphic Method

In this method the data is denoted on graph paper. We take **“Time”** on **‘x’** axis and **“Data”** on the **‘y’** axis. On graph there will be a point for every point of time. We make a smooth hand curve with help of these plotted points.

Example 1: Draw the trend line by graphic method and estimate the production in 2003.

Year	1995	1996	1997	1998	1999	2000	2001
Production	20	22	25	26	25	27	30

Solution: Year is represented in x axis. Production is represented as y axis.



Semi-Average Method

In this method the given data are divided in **two** parts, preferable with the **equal** number of years.

When there are **even** numbers of years, the middle most year and arithmetic mean of the observed values are **found out** for each half.

When there are **odd** numbers of years, the middle most year and the corresponding value are **omitted**.

Example 2: Find the trend line from the following data by Semi-Average Method

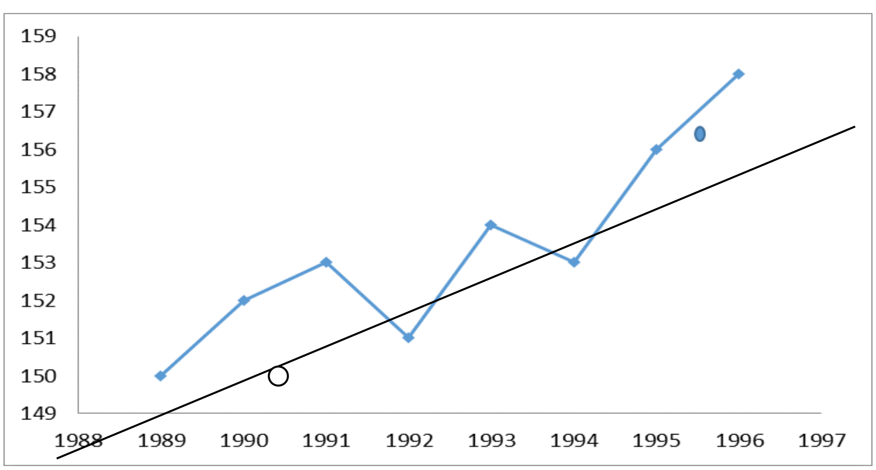
Year	1989	1990	1991	1992	1993	1994	1995	1996
Production	150	152	153	151	154	153	156	158

Solution: There are 8 trends. Now we distributed it in equal part. Now we calculated average mean for every part.

First Part = $\frac{150 + 152 + 153 + 151}{4} = 151.50$

Second Part = $\frac{154 + 153 + 156 + 158}{4} = 155.25$

<table>
<tr><th>Year</th><th>Production</th><th>Arithmetic mean</th></tr>
<tr><td>1989</td><td>150</td><td rowspan="4">151.50</td></tr>
<tr><td>1990</td><td>152</td></tr>
<tr><td>1991</td><td>153</td></tr>
<tr><td>1992</td><td>151</td></tr>
<tr><td>1993</td><td>154</td><td rowspan="4">155.25</td></tr>
<tr><td>1994</td><td>153</td></tr>
<tr><td>1995</td><td>156</td></tr>
<tr><td>1996</td><td>158</td></tr>
</table>



Moving Average Method

The method of moving averages is one of the most useful method of estimating trend. It is an algebraic method. Graph sheet is not used for calculating trend.

For a series, there is only one arithmetic mean; there are many moving averages. Moving totals are found and they are divided by appropriate number to get the moving average.

For Example: If we calculating Three year's moving average then according to this method;

$$= \frac{(1)+(2)+(3)}{3}, \quad \frac{(2)+(3)+(4)}{3}, \quad \frac{(3)+(4)+(5)}{3}, \ldots\ldots\ldots$$

Where (1),(2),(3)……..are the various years of time series

Example 3: Calculate 5 yearly moving average of number of students studying in Commerce College as shown by the following figure.

Year	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
No. of Students	332	311	357	392	402	405	410	427	405	438

Solution:

Year	**No. of Students**	**5 Yearly Moving Total**	**5 Yearly Moving Average**
1987	332	-	-
1988	311	-	-
1989	357	1794	358.8
1990	392	1867	373.4
1991	402	1966	393.2
1992	405	2036	407.2
1993	410	2049	409.8
1994	427	2085	417.0
1995	405	-	-
1996	438	-	-

Methods of least square

The trend project method fits a trend line to a series of historical data points and then projects the line into the future for medium-to-long range forecasts. Several mathematical trend equations can be developed (such as exponential and quadratic), depending upon movement of time-series data.

Reasons to study trend: A few reasons to study trends are as follows:

1. The study of trend allows us to describe a historical pattern so that we may evaluate the success of previous policy.
2. The study allows us to use trends as an aid in making intermediate and long-range forecasting projections in the future.
3. The study of trends helps us to isolate and then eliminate its influencing effects on the time-series model as a guide to short-run (one year or less) forecasting of general business cycle conditions.

Linear Trend Model

If we decide to develop a linear trend line by a precise statistical method, we can apply the least squares method. A least squares line is described in terms of its y-intercept (the height at which it intercepts the y-axis) and its slope (the angle of the line). If we can compute the y intercept and slope, we can express the line with the following equation

$$\hat{y} = a + bx$$

Where $\hat{y}$ = predicted value of the dependent variable

a = y-axis intercept

b = slope of the regression line (or the rate of change in y for a given change in x)

x = independent variable (which is time in this case)

Least squares is one of the most widely used methods of fitting trends to data because it yields what is mathematically described as a ‘line of best fit’. This trend line has the properties that

(i) The summation of all vertical deviations about it is zero, that is, $\Sigma(y - \hat{y}) = 0$,

(ii) The summation f all vertical deviations squared is a minimum, that is, $\Sigma(y - \hat{y})$ is least, and

(iii) The line goes through the mean values of variables x and y. For linear equations, it is found by the simultaneous solution for a and b of the two normal equations:

Σy = na + bΣx and Σxy = aΣx + bΣx2

Where the data can be coded so that Σx = 0, two terms in three equations drop out and we have

Σy = na and Σxy = bΣx2

Coding is easily done with time-series data. For coding the data, we choose the centre of the time period as x = 0 and have an equal number of plus and minus periods on each side of the trend line which sum to zero.

Alternately, we can also find the values of constants a and b for any regression line as:

$$B = \frac{\sum Xy - n\bar{x}\bar{y}}{\sum X^2 - n(\bar{x})^2} \quad \text{and } a = \bar{y} - b\bar{x}$$

Example 5: Below are given the figures of production (in thousand quintals) of a sugar factory:

Year	1992	1993	1994	1995	1996	1997	1998
Production	80	90	92	83	94	99	92

(a) Fit a straight-line trend to these figures.

(b) Plot these figures on a graph and show the trend line.

(c) Estimate the production in 2001.

Solution:

Year	**Time period(x)**	**Production (y)**	**x^2**	**xy**	**Trend value $\bar{y}$**
1992	1	80	1	80	84
1993	2	90	4	180	86
1994	3	92	9	276	88
1995	4	83	16	332	90
1996	5	94	25	470	92
1997	6	99	36	594	94

1998	7	92	49	644	96
Total	**28**	**630**	**140**	**2576**	

$$\bar{x} = \frac{\sum x}{n} = \frac{28}{7} = 4, \quad \bar{y} = \frac{\sum y}{n} = \frac{630}{7} = 90$$

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2} = \frac{2576 - 7(4)(90)}{140 - 7(4)^2} = \frac{56}{28} = 2$$

$$a = \bar{y} - b\bar{x} = 90 - 2(4) = 82$$

Therefore, linear trend component for the production of sugar is:

$$\hat{y} = a + bx = 82 + 2x$$

The slope b = 2 indicates that over the past 7 years, the production of sugar had an average growth of about 2 thousand quintals per year.

Parabolic Trend Model

The curvilinear relationship for estimating the value of a dependent variable y from an independent variable x might take the form

$\hat{y} = a + bx + cx^2$

This trend line is called the parabola.

For a non-linear equation $\hat{y} = a + bx - cx^2$, the values of constants a, b, and c can be determined by solving three normal equations.

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

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$$

$$\Sigma x^2y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

When the data can be coded so that $\Sigma x = 0$ and $\Sigma x^3 = 0$, two term in the above expressions drop out and we have

$$\Sigma y = na + c\Sigma x^2$$

$\Sigma xy = b \Sigma x2$

$\Sigma x2y = a \Sigma x2 + c \Sigma x4$

To find the exact estimated value of the variable *y*, the values of constants *a*, *b*, and *c* need to be calculated. The values of these constants can be calculated by using the following shortest method:

Example 6 : The prices of a commodity during 1999-2004 are given below. Fit a parabola to these data. Estimate the price of the commodity for the year 2005.

Year	1999	2000	2001	2002	2003	2004
Price	100	107	128	140	181	192

Solution: To fit a parabola yˆ = a + bx + cx2, the calculations to determine the values of constants a, b, and c

Year	**Time Scale (x)**	**Price (y)**	**x^2**	**x^3**	**x^4**	**xy**	**x^2y**	**Trend values (y)**
1999	-2	100	4	-8	16	-200	400	97.72
2000	-1	107	1	-1	1	-107	-107	110.34
2001	0	128	0	0	0	0	0	126.68
2002	1	140	1	1	1	140	140	146.50
2003	2	181	4	8	16	362	724	169.88
2004	3	192	9	27	81	576	1728	196.82
Total	**3**	**848**	**19**	**27**	**115**	**771**	**3099**	**847.94**

(i) $\Sigma y = na - b \Sigma x + c \Sigma x2$

$848 = 6a + 3b + 19c$

(ii) $\Sigma xy = a \Sigma x + b \Sigma x2 + c \Sigma x3$

$771 = 3a + 19b + 27c$

(iii) $\Sigma x2y = a \Sigma x2 + b \Sigma x2 + c \Sigma x4$

3099 = 19a + 27b + 115c

Eliminating a from eqns. (i) and (ii), we get (iv) 694 = 35b + 35c

Eliminating a from eqns. (ii) and (iii), we get

(v) 5352 = 280b + 168c

Solving eqns. (iv) and (v) for b and c we get b =18.04 and c = 1.78.

Substituting values of b and c in eqn. (i), we get a = 126.68.

Hence, the required non-linear trend line becomes

y = 126.68 +18.04x + 1.78x2

UNIT V

INDEX NUMBERS

“Index numbers are statistical devices designed to measure the relative changes in the level of a certain phenomenon in two or more situations”. The phenomenon under consideration may be any field of quantitative measurements. It may refer to a single variable or a group of distinct but related variables. In Business and Economics, the phenomenon under consideration may be:

- The prices of a particular commodity like steel, gold, leather, etc. or a group of commodities like consumer goods, cereals, milk and milk products, cosmetics, etc.
- Volume of trade, factory production, industrial or agricultural production, imports or exports, stocks and shares, sales and profits of a business house and so on.
- The national income of a country, wage structure of workers in various sectors, bank deposits, foreign exchange reserves, cost of living of persons of a particular community, class or profession and so on.

The various situations requiring comparison may refer to either

- The changes occurring over a time, or
- The difference(s) between two or more places, or
- The variations between similar categories of objects/subjects, such as persons, groups of persons, organisations etc. or other characteristics such as income, profession, etc.

CHARACTERISTICS OF INDEX NUMBERS

1. **Index Numbers are specialized averages:** An average is a summary figure measuring the central tendency of the data, representing a group of figures. Index number has all these functions to perform. L R Connor states, "in its simplest form, it (index number) represents a special case of an average, generally a weighted average compiled from a sample of items judged to be representative of the whole". It is a special type of average – it averages variables having different units of measurement.

2. **Index Numbers are expressed in percentages:** Index numbers are expressed in terms of percentages so as to show the extent of change. However, percentage sign (%) is never used.
3. **Index Numbers measure changes not capable of direct measurement**: The technique of index numbers is utilized in measuring changes in magnitude, which are not capable of direct measurement. Such magnitudes do not exist in themselves. Examples of such magnitudes are 'price level', 'cost of living', 'business or economic activity' etc.
4. **Index Numbers are for comparison:** The index numbers by their nature are comparative. They compare changes taking place over time or between places or between like categories.

PROBLEMS OF CONSTRUCTING INDEX NUMBERS

The above discussion enables us to identify some of the important problems, which may be faced in the construction of index numbers:

1. **Choice of the Base Period:** Choice of the base period is a critical decision because of its importance in the construction of index numbers. A base period is the reference period for describing and comparing the changes in prices or quantities in a given period.

2. **Selection of Weights to be used:** It should be amply clear from the various indices discussed in the lesson that the choice of the system of weights, which may be used, is fairly large.

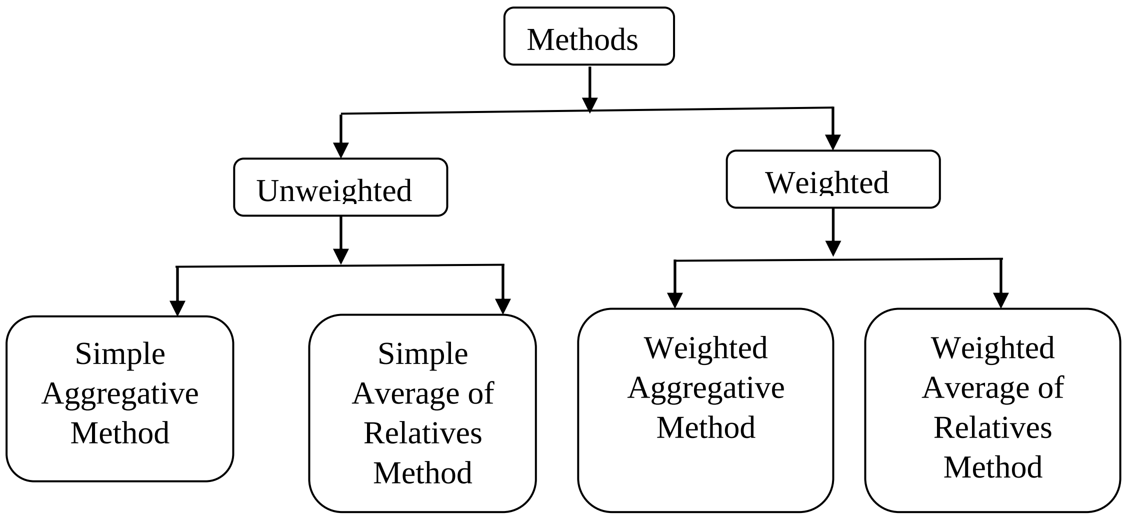
3. **Type of Average to be used:** What type of average should be used is a problem specific to simple average indices. Theoretically, one can use any of the several averages that we have, such as mean, median, mode, harmonic mean, and geometric mean.

4. **Choice of Index:** The problem of selection of an appropriate index arises because of availability of different types of indices giving different results when applied to the same data.

5. **Selection of Commodities:** Commodities to be included in the construction of an index should be carefully selected. Only those commodities deserve to be included in the construction of an index as would make it more representative.

6. **Data Collection:** Collection of data through a sample is the most important issue in the construction of index numbers. The data collected are the raw material of an index. Data quality is the basic factor that determines the usefulness of an index. The data have to be as accurate, reliable, comparable, representative, and adequate, as possible.

METHODS OF CONSTRUCTING INDEX NUMBER



Simple Aggregative Method

It consists in expressing the aggregative price of all commodities in the current year as a percentage of the aggregative price in the base year.

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

Where P_{01} = Current price Index number

Σp_1 = the total of commodity prices in the current year

Σp_0 = the total of same commodity prices in the base year.

Example 2: Construct the price index number for 2003, taking the year 2000 as base year

Commodity	Price in the year 2000	Price in the year 2003
A	60	80
B	50	60
C	70	100
D	120	160
E	100	150

Solution : Calculation of simple Aggregative index number for 2003 (against the year 2000)

Commodity	Price in 2000 (in Rs) p_0	Price in 2003 (in Rs.) p_1
A	60	80
B	50	60
C	70	100
D	120	160
E	100	150
Total	$\sum p_0 = 400$	$\sum p_1 = 550$

Here $\sum p_0 = 400$, $\sum p_1 = 550$

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

$$= \frac{550}{400} \times 100 = \frac{275}{2}$$

$$= \mathbf{137.5}$$

Example 2: Compute the index number for the years 2001, 2002, 2003 and 2004, taking2000 as base year, from the following data:

Year	**2000**	**2001**	**2002**	**2003**	**2004**
Price	120	144	168	204	216

Solution: Price relatives for different years are

2000	$\frac{120}{120} \times 100 = 100$
2001	$\frac{144}{120} \times 100 = 120$
2002	$\frac{168}{120} \times 100 = 140$
2003	$\frac{204}{120} \times 100 = 170$
2004	$\frac{216}{120} \times 100 = 180$

Simple Average of Price Relative Method

In this method, the price relative for all commodities is calculated and then their average is taken to calculate the index number.

$$P_{01} = \frac{\sum \frac{p_1}{P_0} \times 100}{N}$$

If A.M. is used as average where

P_{01} is the price index,

N is the number of items

P_0 is the price in the base year

P_1 of corresponding commodity in present year (for which index is to be calculated)

Example 3: Construct by simple average of price relative method the price index of 2004,taking 1999 as base year from the following data:

Commodity	**A**	**B**	**C**	**D**	**E**	**F**
Price (in 1999)	60	50	60	50	25	20
Price (in 2004)	80	60	72	75	$37\frac{1}{2}$	30

Solution:

Commodity	Price (in 1999) (in Rs.)[p_0]	Price (in 2004) (in Rs.) [p_1]	Price Relatives $\left(\frac{p_1}{P_0} \times 100\right)$
A	60	80	133.33
B	50	60	120.00
C	60	72	120.00
D	50	75	150.00
E	25	$37\frac{1}{2}$	150.00
F	20	30	150.00
			823.33

$$P_{01} = \frac{\Sigma \frac{p1}{P_0} \times 100}{N}$$

$$= \frac{823.3}{6}$$

$$= 137.22$$

Weighted Index Numbers

These are those index numbers in which rational weights are assigned to various chains in an explicit fashion.

Weighted Aggregative Index Numbers:

These index numbers are the simple aggregative type with the fundamental difference that weights are assigned to the various items included in the index.

- Dorbish and Bowley's Method
- Fisher's Ideal Method
- Marshall-Edgeworth Method
- Laspeyres Method
- Paasche Method
- Kelly's Method

Laspeyres Method

This method was devised by Laspeyres in 1871. In this method the weights are determined by quantities in the base.

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

Paasche's Method

This method was devised by German Statistician Paasche in 1874. The weights of current year are used as base year in constructing the Paasche's Index Number.

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

Dorbish and Bowley's Method

This method is a combination of Laspeyres and Paasche's methods. If we find out the arithmetic average of Laspeyres and Paasche's index we get the index suggested by Dorbish and Bowley.

$$P_{01} = \frac{\frac{\sum P_1q_0}{\sum P_0q_0} + \frac{\sum p_1q_1}{\sum p_0q_1}}{2} \times 100$$

Fisher's Ideal Method

Fisher's deal index number is the geometric mean of the Laspeyres and Paasche's index numbers.

$$P_{01} = \sqrt{\frac{\sum P_1q_0}{\sum P_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} \times 100$$

Marshall-Edgeworth Method

In this index the numerator consists of an aggregate of the current years price multiplied by the weights of both the base year as well as the current year.

$$P_{01} = \frac{\sum P_1q_0 + \sum p_1q_1}{\sum P_0q_0 + \sum p_0q_1} \times 100$$

Kelly's Method

Kelly thinks that a ratio of aggregates with selected weights (not necessarily of base year or current year) gives the base index number.

$$P_{01} = \frac{\sum P_1q}{\sum P_0q} \times 100$$

q refers to the quantities of the year which is selected as the base, it may be any year, either base year or current year.

Example 4: Given below are the price quantity data, with price quoted in Rs. per kg and production in qtls. Find 1) Laspeyres Index 2) Paasche's Index 3) Fisher's Ideal Index

	2002		2007	
Items	**Price**	**Production**	**Price**	**Production**
Beef	15	500	20	600
Mutton	18	590	23	640
Chicken	22	450	24	500

Solution:

Items	Price (p_0)	Production(q_0)	Price (p_1)	Production(q_1)	(p_1q_0)	(p_0q_0)	(p_1q_1)	(p_0q_1)
Beef	15	500	20	600	10000	7500	12000	9000
Mutton	18	590	23	640	13570	10620	14720	11520
Chicken	22	450	24	500	10800	9900	12000	11000
Total					**34370**	**28020**	**38720**	**31520**

1) **Laspeyres Index**

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

$$= \frac{34370}{28020} \times 100 = 122.66$$

2) **Paasche's Index**

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

$$= \frac{38720}{31520} \times 100 = 122.84$$

3) Fisher's Ideal Index

$$P_{01} = \sqrt{\frac{\sum P_1q_0}{\sum P_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} \times 100$$

$$= \sqrt{\frac{34370}{28020} \times \frac{38720}{31520}} \times 100 = 122.69$$

Weighted Average of Price Relative

In Weighted Average Price Relative, the price relatives for the current year are calculated on the basis of the base year price. These price relatives are multiplied by the respective weight of items. These products are added up and divided by the sum of weights.

$$P_{01} = \frac{\sum PV}{\sum V}$$

Where -

$$P = \frac{P_1}{} \text{X } 100$$

P_0

P = Price relative

V = Value weights = p_0q_0

VALUE INDEX NUMBER

Value is the product of price and quantity. A simple ratio is equal to the value of the current year divided by the value of base year. If the ratio is multiplied by 100 we get the value index number.

$$V = \frac{\sum p_1q_1}{\sum P_0q_0} \times 100$$

CHAIN INDEX NUMBER

When this method is used the comparisons are not made with fixed base, rather the base changes from year to year. For example, for 2007,2006 will be the base; for 2006,2005 will be the same and so on.

Chain index for current year

$$= \frac{\text{Average link relative of current year x Chain index of previous year}}{100}$$

Example 5: From the following data construct an index number by chain base method.

Price of commodity from 2006 to 2008.

YEAR	PRICE
2006	50
2007	60
2008	65

Solution:

Year	**Price**	**Link relative**	**Chain Index (Base 2006)**
2006	50	100	100
2007	60	$\frac{60}{50} \times 100 = 120$	$\frac{120 \times 100}{100} = 120$
2008	65	$\frac{65}{60} \times 100 = 108$	$\frac{108 \times 120}{100} = 129.60$