மனோன்மணியம்சுந்தரனார் பல்கலைக்கழகம்

MANONMANIAM SUNDARANAR UNIVERSITY

TIRUNELVELI – 12

தொலைநிலை தொடர்கல்வி இயக்கம்

DIRECTORATE OF DISTANCE & CONTINUING EDUCATION

BBA – FIRST YEAR

BUSINESS STATISTICS

MSU/DD&CE/2021/UG/(B.B.A.)

BUSINESS STATISTICS

Course Objective: To familiarize the students with various statistical Data Analysis tools that can be used for effective decision making. Emphasis will be on the application of the concepts learned.

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 framework only) (18 hrs)

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. (18 hrs)

UNIT III: REGRESSION ANALYSIS: Regression Analysis – Regression Vs Correlation, Linear Regression, Regression lines, Standard error of estimates. (18 hrs)

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. (18 hrs)

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 (18 hrs)

References:

1. Statistics Theory and Practice – R.S.N. Pillai and V. Bagavathi

UNIT – I

MEASURES 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 framework only) (18 hrs)

Measures of central tendency are a typical value of the entire group or data. It describes the characteristics of the entire mass of data. It reduces the complexity of data and makes them to compare. Human mind is incapable of remembering the entire mass of unwieldy data. So a simple figure is used to describe the series which must be a representative number. It is generally called, "a measure of central tendency or the average".

A central tendency is a central or typical value for a probability distribution. It may also be called a center or location of the distribution. Colloquially, measures of central tendency are often called averages. The term central tendency dates from the late 1920s. If a large volume of data is summarized and given is one simple term. Then it is called as the 'Central Value' or an 'average'. In other words an average is a single value that represents group of values.

Characteristics of an Ideal Measures:

A measure of central tendency is a typical value around which other figures congregate. Average condenses a frequency distribution in one figure. According to the statisticians, an average will be termed good or efficient if possesses the following characteristics:

- It should be rigidly defined. It means that the definition should be so clear that the interpretation of the definition does not differ from person to person.
- It should be easy to understand and simple to calculate.
- It should be such that it can be easily determined.
- The average of a variable should be based on all the values of the variable. This means that in the formula for average all the values of the variable should be incorporated.
- The value of average should not change significantly along with the change in sample. This means that the values of the averages of different samples of the same size drawn from the same population should have small variations.

- It should be amenable to algebraic treatment.
- It should be unduly affected by extreme values. i.e, the formula for average should be such that it does not show large due to the presence of one or two very large or very small values of the variable.
- It should be properly defined, preferably by a mathematical formula, so that different individuals working with the same data should get the same answer unless there are mistakes in calculations.
- It should be based on all the observations so that if we change the value of any observation, the value of the average should also be changed.
- It should not be unduly affected by extremely large or extremely small values.
- It should be capable of algebraic manipulation. By this we mean that if we are given the average heights for different groups, then the average should be such that we can find the combined average of all groups taken together.
- It should have quality of sampling stability. That is, it should not be affected by the fluctuations of sampling. For example, if we take ten or twelve samples of twenty students' each and find the average height for each sample, we should get approximately the same average height for each sample.

MEAN:

Mean is one of the types of averages. Mean is further divided into three kinds, which are the arithmetic mean, the geometric mean and the harmonic mean. These kinds are explained as follows;

i) Arithmetic Mean: Simple Arithmetic Average:

A. Individual Observation: Direct Method:

The arithmetic mean is most commonly used average. It is generally referred as the average or simply mean. The arithmetic mean or simply mean is defined as the value obtained by dividing the sum of values by their number or quantity. It is denoted as $\bar{X}$ (read as X-bar). Therefore, the mean for the values $X_1, X_2, X_3, \ldots\ldots, X_n$ shall be denoted by $\bar{X}$. Following is the mathematical representation for the formula for the arithmetic mean or simply, the mean.

$$\bar{X} = \frac{X1+X2+X3+\cdots+Xn}{N} = \frac{\Sigma X}{N}$$

Where, $\bar{X}$ = Arithmetic Mean; Σx = Sum of all the values of the variables i.e., $X_1 + X_2 + X_3 + \ldots + Xn$

N = Number of observations.

Illustration 1: Calculate mean from the following data:

Roll Numbers	1	2	3	4	5	6	7	8	9	10
Marks	40	50	55	78	58	60	73	35	43	48

Solution: Calculation of mean

Roll Numbers	**Marks (x)**
1	40
2	50
3	55
4	78
5	58
6	60
7	73
8	35
9	43
10	48
N = 10	ΣX = 540

$$\bar{X} = \frac{\Sigma X}{N}$$

$$= \frac{540}{10}$$

= 54 marks.

Short cut method:

The arithmetic mean can also be calculated by short cut method. This method reduces the amount of calculation. Formula for calculation

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

Where, $\bar{X}$ = Arithmetic Mean; A = Assumed mean; Σd = Sum of the deviations; N = Number of items.

Illustration 2: (Solving the previous problem)

Roll Numbers	Marks (X)	d = X - A
1	40	-10
2	50	0
3	55	5
4	78	28
5	58	8
6	60	10
7	73	23
8	35	-15
9	43	-7
10	48	-2
N = 10		**Σd = 40**

Let the assumed mean, A = 50

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

$$= 50 + \frac{40}{10}$$

= 54 marks.

B. Discrete Series: Direct Method:

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

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

Where, $\bar{X}$ = Arithmetic Mean; Σfx = the sum of products; N = Total frequency.

Illustration 3: Calculate mean from the following data:

Value	1	2	3	4	5	6	7	8	9	10
Frequency	21	30	28	40	26	34	40	9	15	57

Solution: Calculation of Mean

x	**f**	**Fx**
1	21	21
2	30	60
3	28	84
4	40	160
5	26	130
6	34	204
7	40	280
8	9	72
9	15	135
10	57	570
	$\Sigma f = N = 300$	**$\Sigma fx = 1716$**

$$\bar{X} = \frac{\Sigma fx}{N} = \frac{1716}{300} = 5.72$$

Short cut Method: Formula:

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

Where, $\bar{X}$ = Arithmetic Mean; A = Assumed mean; Σfd = Sum of total deviations; N = Total frequency.

Illustration: 4 (Solving the previous problem)

X	F	d = X - A	fd
1	21	-4	-84
2	30	-3	-90
3	28	-2	-56
4	40	-1	-40
5	26	0	0
6	34	1	34
7	40	2	80
8	9	3	27
9	15	4	60
10	57	5	285
	Σf = N = 300		**Σfd** = + 216

Let the assumed mean, A = 5

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

$$\bar{X} = 5 + \frac{216}{300} = 5.72$$

C. Continuous Series

In continuous frequency distribution, the value of each individual frequency distribution is unknown. Therefore an assumption is made to make them precise or on the assumption that the

frequency of the class intervals is concentrated at the centre that the midpoint of each class interval has to be found out. In continuous frequency distribution, the mean can be calculated by any of the following methods:

1. Direct Method
2. Short cut method
3. Step Deviation Method

1. Direct Method: The formula is $\bar{X} = \frac{\Sigma fm}{N}$

Where, $\bar{X}$ = Arithmetic Mean; Σfm = Sum of the product of f & m; N = Total frequency.

Illustration 5: From the following find out the mean:

Class Interval	0 – 10	10 – 20	20 – 30	30 – 40	40 - 50
Frequency	6	5	8	15	7

Solution: Calculation of Mean

Class Interval	**Mid Point (m)**	**Frequency (f)**	**fm**
0 – 10	$\frac{0+10}{2} = 5$	6	30
10 – 20	$\frac{10+20}{2} = 15$	5	75
20 – 30	$\frac{20+30}{2} = 25$	8	200
30 – 40	$\frac{30+40}{2} = 35$	15	525
40 - 50	$\frac{40+50}{2} = 45$	7	315
		Σf = N = 41	**Σfm = 1145**

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

$$= \frac{1145}{41} = 27.93$$

2. Short cut method: Formula:

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

Where, $\bar{X}$ = Arithmetic Mean; A = Assumed mean; Σfd = Sum of total deviations; N = Total frequency.

Illustration: 6 (Solving the previous problem)

Class Interval	M	d = m - A	F	fd
0 – 10	$\frac{0+10}{2} = 5$	5 – 25 = -20	6	-120
10 – 20	$\frac{10+20}{2} = 15$	15 – 25 = - 10	5	-50
20 – 30	$\frac{20+30}{2} = 25$	25 – 25 = 0	8	0
30 – 40	$\frac{30+40}{2} = 35$	35 – 25 = 10	15	150
40 - 50	$\frac{40+50}{2} = 45$	45 – 25 = 20	7	140
			Σf = N = 41	**Σfd** = +120

d = m – A; here A = 25

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

$$= 25 + \frac{120}{41}$$

$$= 25 + 2.93$$

$$= 27.93$$

3. Step Deviation Method

Formula:

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

Where, $\bar{X}$ = Arithmetic Mean; A = Assumed mean; $\Sigma f d'$ = Sum of total deviations;

N = Total frequency; C = Common Factor

Illustration: 7 (Solving the previous problem)

Class Interval	Mid Point (m)	Frequency (f)	d = m - A	$d' = \frac{m-A}{C}$	fd’
0 – 10	$\frac{0+10}{2} = 5$	6	5 – 25 = -20	-2	-12
10 – 20	$\frac{10+20}{2} = 15$	5	15 – 25 = -10	-1	-5
20 – 30	$\frac{20+30}{2} = 25$	8	25 – 25 = 0	0	0
30 – 40	$\frac{30+40}{2} = 35$	15	35 – 25 = 10	1	15
40 - 50	$\frac{40+50}{2} = 45$	7	45 – 25 = 20	2	14
		$\Sigma f = N = 41$			**$\Sigma fd' = +12$**

Here A = 25; C = 10

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

$$= 25 + \frac{12}{41} \times 10$$

$$= 25 + \frac{120}{41}$$

$$= 25 + 2.93$$

$$= 27.93$$

MEDIAN:

Median is the value of item that goes to divide the series into equal parts. It may be defined as the value of that item which divides the series into equal parts, one half containing values greater that it and the other half containing values less than it. Therefore, the series has to be arranged in ascending or descending order, before finding the median. If the items of a series are arranged in

ascending or descending order of magnitude, the item which falls in the middle of it is called median. Hence it is the "middle most" or "most central" value of a set of number.

Calculation of Median – Individual Series:

Illustration 1: Find out the median of the following items. X: 10, 15, 9, 25, 19.

Solution: Computation of Median

S. No.	Size of ascending order	Size of descending order
1	9	25
2	10	19
3	15	15
4	19	10
5	25	9

$$\text{Median} = \text{Size of } \frac{(N+1)_{\text{th}}}{2} \text{ item}$$

$$= \text{Size of } \frac{(5+1)_{\text{th}}}{2} \text{ item}$$

$$= 3^{\text{rd}} \text{ item} = 15.$$

Illustration 2: Find out the median of the following items. X: 8, 10, 5, 9, 12, 11.

Solution: Computation of Median

S. No.	X
1	5
2	8
3	9
4	10
5	11
6	12

$$\text{Median} = \text{Size of } \frac{(N+1)_{\text{th}}}{2} \text{ item}$$

$$= \text{Size of } \frac{(6+1)_{\text{th}}}{2} \text{ item}$$

= Size of 3.5^{th} item

$$= \text{Size of } \frac{(3\text{rd item} + 4\text{th item})}{2}$$

$$= \frac{9+10}{2} = 9.5$$

Calculation of Median – Discrete Series

Illustration 3: Locate median from the following:

Size of shoes	5	5.5	6	6.5	7	7.5	8
Frequency	10	16	28	15	30	40	34

Solution: Computation of Median

Size of shoes	F	c.f
5	10	10
5.5	16	26
6	28	54
6.5	15	69
7	30	99
7.5	40	139
8	34	173

$$\text{Median} = \text{Size of } \frac{(N+1)_{th}}{2} \text{ item}$$

$$= \text{Size of } \frac{(173+1)_{th}}{2} \text{ item}$$

= Size of 87^{th} item

= 7

Median – Continuous Series

Illustration 4: Calculate the median of the following table:

Marks	10 – 25	25 - 40	40 – 55	55 - 70	70 – 85	85 - 100
Frequency	6	20	44	26	3	1

Solution: Computation of Median

x	F	c.f
10 – 25	6	6
25 – 40	20	26
40 – 55	44	70
55 – 70	26	96
70 – 85	3	99
85 - 100	1	100

$$\text{Median} = L + \frac{\frac{N}{2} - cf}{f} \times i$$

$$\frac{N}{2} = \frac{100}{2} = 50;$$

$$L = 40;\ f = 44;\ cf = 26;\ i = 15$$

$$\text{Median} = 40 + \frac{50-26}{44} \times 15$$

$$= 40 + 8.18$$

$$= 48.\ 18 \text{ marks}$$

Merits:

1. It is easy to compute and understand.
2. It eliminates the effect of extreme items.
3. The value of median can be located graphically.
4. It is amenable to further algebraic process as it is used in the measurement of dispersion.
5. It can be computed even if the items at the extremes are unknown.

Demerits:

1. For calculating median, it is necessary to arrange the data; other averages do not need any arrangement.
2. Typical representative of the observations cannot be computed if the distribution of item is irregular.
3. It is affected more by fluctuation of sampling than the arithmetic mean.

MODE:

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

Mode or the modal value is defined as the value of the variable which occur more number of times or most frequently in a distribution.

Types of Mode:

i) Unimodal:

If there is only one mode in series, it is called unimodal.

Eg., 10, 15, 20, 25, 18, 12, 15 (Mode is 15)

ii) Bi – modal:

If there are two modes in the series, it is called bi - modal.

Eg., 20, 25, 30, 30, 15, 10, 25 (Modes are 25, 30)

iii) Tri – modal:

If there are three modes in the series, it is called Tri - modal.

Eg., 60, 40, 85, 30, 85, 45, 80, 80, 55, 50, 60 (Modes are 60, 80, 85)

iv) Multi – modal:

If there are more than three modes in the series it is called multi-modal.

Merits:

1. It can be easily ascertained without much mathematical calculation.

2. It is not essential to know all the items in a series to compute mode.
1. Open – end classes do not disturb the position of the mode.
2. Its values can be ascertained graphically as well as empirically.
3. It may be very well applied to qualitative as well quantitative data.
4. It is not affected by extreme values as in the average.

Demerits:

1. The mode becomes less useful as an average which the distribution is bi-modal.
2. It is not suitable for further mathematical treatment.
3. It is stable only when the sample is large.
4. Mode is influenced by magnitude of the class-intervals.

Mode - Individual Series

Illustration : 1. Calculate the mode from the following data of the marks obtain by 10 students.

Serial No.	1	2	3	4	5	6	7	8	9	1
Marks obtained	60	77	74	62	77	77	70	68	65	80

Solution:

Marks obtained by 10 students 60, 77, 74, 62, 77, 77, 70, 68, 65, and 80.

Here 77 is repeated three times.

∴The Mode mark is 77.

DISCRETE SERIES:

A grouping Table has six columns

Column 1: In column 1 rite the actual frequencies and mark the highest frequency.

Column 2: Frequencies are grouped in twos, adding frequencies of items 1 and 2; 3 and 4; 5 and 6; and so on.

Column 3: Leave the first frequency and then add the remaining in twos.

Column 4: Group of frequencies in threes.

Column 5: Leave the first frequency and group the remaining in threes.

Column 6: Leave the first two frequencies and then group the remaining the threes.

The maximum frequencies in all six columns are marked with a circle and an analysis table is prepared as follows:

1. Put column number on the left – hand side
2. Put the various probable values of mode on the right – hand side.
3. Enter the highest marked frequencies by means of a bar in the relevant box corresponding to the values they represent.

Illustration: 2. Calculate the mode from the following:

Size	10	11	12	13	14	15	16	17	18
Frequency	10	12	15	19	20	8	4	3	2

Solution: Grouping Table

Size	**Frequency**					
	1	**2**	**3**	**4**	**5**	**6**
10	10					
		22				
11	12			37		
			27			
12	15				46	
		34				
13	19					54

14	20	28	39	47		
15	8		12		32	
16	4	7				15
17	3		5	9		
18	2					

Analysis Table

Column No.	Size of item containing maximum frequency				
	11	**12**	**13**	**14**	**15**
1				1	
2		1	1		
3			1	1	
4			1	1	1
5	1	1	1		
6		1	1	1	
	1	3	5	4	1

The mode is 13, as the size of item repeats 5 times. But through inspection, we say the mode is 14, because the size 14 occurs 20 times. But this wrong decision is revealed by analysis table.

Calculation of Mode – Continuous Series

$$\mathbf{Z = L_1 + \frac{f1-f0}{2f1-f0-f2} \times i}$$

Where,

Z = Mode; L_1 = Lower limit of the modal class; f_1 = Frequency of the modal; f_0 = Frequency of the class preceding the modal class; f_2 = Frequency of the class succeeding the modal class; i = Class interval;

Illustration: 3.

Calculate the mode from the following:

Size of item	Frequency
0 – 5	20
5 – 10	24
10 – 15	32
15 – 20	28
20 – 25	20
25 – 30	16
30 – 35	34
35 – 40	10
40 – 45	8

Solution:

Grouping Table

Size of item	Frequency					
	1	**2**	**3**	**4**	**5**	**6**
0 – 5	20					
		44		76		
5 – 10	24					
			56			
10 – 15	32				84	
		60				80
15 – 20	28					
			48			
20 – 25	20			64		
		36				
25 – 30	16				70	

30 – 35	34	44	50			60
35 – 40	10		18	52		
40 - 45	8					

Analysis Table

Column No.	**Size of item containing maximum frequency**					
	0 – 5	**5 – 10**	**10 – 15**	**15 – 20**	**20 - 25**	**30 - 35**
1						1
2			1	1		
3		1	1			
4	1	1	1			
5		1	1	1		
6			1	1	1	
	1	3	**5**	3	1	1

$$Z = L_1 + \frac{\mathbf{f1-f0}}{\mathbf{2f1-f0-f2}} \text{ x i}$$

$L_1 = 10$; $f_1 = 32$; $f_0 = 24$; $f_2 = 28$; i = 5

$$Z = 10 + \frac{32-24}{2\,x\,32-24-28} \text{ x } 5$$

$$= 10 + \frac{40}{12} = 10 + 3.33$$

∴ The Mode is 13. 33

APPLICATION IN BUSINESS DECISIONS:

The field of statistics has numerous applications in business. Because of technological advancements, large amounts of data are generated by business these days. These data are now being used to make decisions. These better decisions we make help us improve the running of a department, a company, or the entire economy.

"Statistics is extensively used to enhance Business performance through Analytics"

- **Marketing:** As per Philip Kotler and Gary Armstrong marketing " identifies customer needs and wants , determine which target markets the organisations can serve best, and designs appropriate products, services and Programs to serve these markets"

Marketing is all about creating and growing customers profitably. Statistics is used in almost every aspect of creating and growing customers profitably. Statistics is extensively used in making decisions regarding how to sell products to customers. Also, intelligent use of statistics helps managers to design marketing campaigns targeted at the potential customers. Marketing research is the systematic and objective gathering, recording and analysis of data about aspects related to marketing. IMRB international, TNS India, RNB Research, The Nielson, Hansa Research and Ipsos Indica Research are some of the popular market research companies in India. Web analytics is about the tracking of online behaviour of potential customers and studying the behaviour of browsers to various websites.

Use of Statistics is indispensable in forecasting sales, market share and demand for various types of Industrial products.

Factor analysis, conjoint analysis and multidimensional scaling are invaluable tools which are based on statistical concepts, for designing of products and services based on customer response.

- **Finance:** Uncertainty is the hallmark of the financial world. All financial decisions are based on "Expectation" that is best analysed with the help of the theory of probability and statistical techniques. Probability and statistics are used extensively in designing of new insurance policies and in fixing of premiums for insurance policies. Statistical tools and

technique are used for analysing risk and quantifying risk, also used in valuation of derivative instruments, comparing return on investment in two or more instruments or companies. Beta of a stock or equity is a statistical tool for comparing volatility, and is highly useful for selection of portfolio of stocks. The most sophisticated traders in today's stock markets are those who trade in "derivatives" i.e financial instruments whose underlying price depends on the price of some other asset.

- **Economics:** Statistical data and methods render valuable assistance in the proper understanding of the economic problem and the formulation of economic policies. Most economic phenomena and indicators can be quantified and dealt with statistically sound logic. In fact, Statistics got so much integrated with Economics that it led to development of a new subject called Econometrics which basically deals with economics issues involving use of Statistics.

- **Operations:** The field of operations is about transforming various resources into product and services in the place, quantity, cost, quality and time as required by the customers. Statistics plays a very useful role at the input stage through sampling inspection and inventory management, in the process stage through statistical quality control and six sigma method, and in the output stage through sampling inspection. The term Six Sigma quality refers to situation where there is only 3.4 defects per million opportunities.

- **Human Resource Management or Development:** Human Resource departments are inter alia entrusted with the responsibility of evaluating the performance, developing rating systems, evolving compensatory reward and training system, etc. All these functions involve designing forms, collecting, storing, retrieval and analysis of a mass of data. All these functions can be performed efficiently and effectively with the help of statistics.

- **Information Systems:** Information Technology (IT) and statistics both have similar systematic approach in problem solving. IT uses statistics in various areas like, optimisation of server time, assessing performance of a program by finding time taken as well as resources used by the program. It is also used in testing of the software.

- **Data Mining:** Data Mining is used in almost all fields of business.

In Marketing, Data mining can be used for market analysis and management, target marketing, CRM, market basket analysis, cross selling, market segmentation, customer profiling and managing web based marketing, etc. In Risk analysis and management, it is used for forecasting, customer retention, quality control, competitive analysis and detection of unusual patterns.

In Finance, it is used in corporate planning and risk evaluation, financial planning and asset evaluation, cash flow analysis and prediction, contingent claim analysis to evaluate assets, cross sectional and time series analysis, customer credit rating, detecting of money laundering and other financial crimes.

In Operations, it is used for resource planning, for summarising and comparing the resources and spending.

In Retail industry, it is used to identify customer behaviours, patterns and trends as also for designing more effective goods transportation and distribution policies, etc.

Decision Situation and Corresponding Statistically Techniques

Area	*Decision Situation*	*Statistical Techniques Applicable*
Marketing	• Assessment / Forecast of Demand for the Product or a Service • Customer Profiling Market Research	Time series Correlation and Regression Cluster Analysis Conjoint Analysis Multidimensional Scaling
Retail Management	Identifying Customer Buying Behaviours and patterns	Correlation and Regression Cluster Analysis Conjoint Analysis
Finance and Banking	Evaluation of Investment Volatility of Stocks Predicting EPS Derivatives	Regression Analysis, Decision Analysis Beta analysis
Insurance	Determining the Premium	Profitability, Time series and Regression Analysis
Operations	Controlling and Improving Production Process and Quality Inventory Management	Statistical Quality Control Six Sigma Sampling inspection ABC Analysis
HRD	Performance Appraisal and Reward System	Normal Distribution Percentiles

Source: Statistics for management by T N Srivastava and Shailaja Rego, Published by the Tata McGraw- Hill Publishing Company Limited.

DISPERSION

Dispersion is studied to have an idea of the homogeneity or heterogeneity of the distribution. Measures of dispersion are the measures of scatter or spread about an average. Measures of dispersion are called the averages of the second order.

Methods of Measuring Dispersion:

There are various methods of studying variation or dispersion important methods studying dispersion are as follows:

1. Range
2. Inter - quartile range
3. Mean Deviation
4. Standard Deviation
5. Lorenz curve

1. Range

Range is the simplest and crudest measure of dispersion. It is a rough measure of dispersion. It is the difference between the highest and the lowest value in the distribution.

$$\textbf{Range} = L - S$$

Where, L = Largest Value; S = Smallest Value.

The Relative measure of range is called as the Co – efficient of Range.

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

Illustration 1:

Find the range of weights of 7 students from the following.

27, 30, 35, 36, 38, 40, 43

Solution:

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

$$\text{Here } L = 43;\ S = 27$$

$$\therefore \text{Range} = 43 - 27 = 16$$

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

$$= \frac{43-27}{43+27} = \frac{16}{70} = 0.23$$

Practical utility of Range

1. It is used in industries for the statistical quality control of the manufactured product.
2. It is used to study the variations such as stock, shares and other commodities.
3. It facilitates the use of other statistical measures.

Advantages

1. It is the simplest method
2. It is easy to understand and the easiest to compute.
3. It takes minimum time to calculate and accurate.

Disadvantages

1. Range is completely dependent on the two extreme values.
2. It is subject to fluctuations of considerable magnitude from sample to sample.
3. Range cannot tell us anything about the character of the distribution.

2. Quartile Deviation (Q.D):

Quartile deviation is an absolute measure of dispersion. Co-efficient of quartile deviation is known as relative measure of dispersion.

In the series, four quartiles are there. By eliminating the lowest items (25%) and the highest items (25%) of a series we can obtain a measure of dispersion and can find out the half of the distance between the first and the third quartiles. That is, [Q3 (third quartiles) – Q1 (first quartiles). The inter-quartile range is reduced to the form of the semi – inter quartile range (or) quartile deviation by dividing it by 2.

$$\text{Inter quartile range} = Q_3 - Q_1$$

$$\text{Inter quartile range or Quartile deviation} = \frac{\mathbf{Q3-Q1}}{\mathbf{2}}$$

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

Quartile Deviation – Individual Series

Illustration 2: Find out the value of Quartile Deviation and its coefficient from the following data:

Roll No.	1	2	3	4	5	6	7
Marks	20	28	40	30	50	60	52

Solution: Calculation of Q.D.

Marks arranged in ascending order:	20	28	30	40	50	52	60

$$Q_1 = \text{Size of } \frac{\boldsymbol{N+1}}{\mathbf{4}}^{\text{th}} \text{ item}$$

$$Q_1 = \text{size of } \frac{7+1}{4}^{\text{th}} \text{ item}$$

$$= \text{size of } \frac{8}{4}^{\text{th}} \text{ item}$$

$$= \text{size of } 2^{\text{nd}} \text{ item}$$

$$= 28$$

$$Q_3 = \text{Size of } 3\left(\frac{\boldsymbol{N+1}}{\mathbf{4}}\right)^{\text{th}} \text{ item}$$

Q_3 = Size of 3($\frac{7+1}{4}$)th item

= Size of 3($\frac{8}{4}$)th item

= size of $\frac{24}{4}$ th item

= size of 6th item

= 52

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

= $\frac{52-28}{2}$

= $\frac{24}{2}$

= 12

Coefficient of Q.D = $\frac{\mathbf{Q3-Q1}}{\mathbf{Q3+Q1}}$

= $\frac{52-28}{52+28}$

=$\frac{24}{80}$

= 0.3

Quartile Deviation – Discrete Series

Illustration 3: Find out the value of Quartile Deviation and its coefficient from the following data:

Age in years	20	30	40	50	60	70	80
No. of members	3	61	132	153	140	51	3

Solution:

Calculation of Q.D.

x	F	c.f.
20	3	3
30	61	64
40	132	196
50	153	349
60	140	489
70	51	540
80	3	543

Q_1 = Value of $\frac{\boldsymbol{N+1}}{\boldsymbol{4}}^{th}$ item

Q_1 = value of $\frac{543+1}{4}^{th}$ item = value of $\frac{544}{4}^{th}$ item

= value of 136^{th} item = 40 years

Q_3 = Value of $3(\frac{\boldsymbol{N+1}}{\boldsymbol{4}})^{th}$ item

Q_3 = value of $3(\frac{543+1}{4})^{th}$ item = value of $3(\frac{544}{4})^{th}$ item

= value of $3(136)^{th}$ item = value of 408^{th} item = 60 years

Q. D. = $\frac{\boldsymbol{Q3-Q1}}{\boldsymbol{2}} = \frac{60-40}{2} = \frac{20}{2}$ = 10 years

Coefficient of Q.D = $\frac{\boldsymbol{Q3-Q1}}{\boldsymbol{Q3+Q1}}$

$= \frac{60-40}{60+40} = \frac{20}{100}$ = 0.2

Quartile Deviation – Continuous Series

Illustration 4: Find out the value of Quartile Deviation and its coefficient from the following data:

Wages (Rs.)	30 – 32	32 – 34	34 - 36	36 - 38	38 - 40	40 - 42	42 - 44
Labourers	12	18	16	14	12	8	6

Solution: Calculation of Q.D.

Wages (x)	Labourers (f)	c.f.
30 – 32	12	12
32 – 34	18	30
34 – 36	16	46
36 – 38	14	60
38 – 40	12	72
40 – 42	8	80
42 – 44	6	86

Q_1 = size of $\frac{N}{4}^{th}$ item

= size of $\frac{86}{4}^{th}$ item

= 21.5^{th} item

∴ Q_1 lies in the group 32 - 34

$$Q_1 = L + \frac{\frac{N}{4} - cf}{f} \times i$$

$$= 32 + \frac{21.5-12}{18} \times 2 = 32 + \frac{9.5}{18} \times 2$$

$$Q_1 = 32 + \frac{19}{18} = 32 + 1.06 = 33.06$$

Q_3 = size of $\frac{3N}{4}^{th}$ item

= size of $\frac{3 \; X \; 86}{4}^{th}$ item

= 64.5 th item

∴ Q_3 lies in the group 38 – 40.

$$Q_3 = L + \frac{\frac{3N}{4} - cf}{f} \times i$$

$$= 38 + \frac{64.5-60}{12} \text{ x } 2 = 38 + \frac{4.5}{12} \text{ x } 2$$

$$Q_3 = 38 + \frac{9}{12} = 32 + 0.75 = 38.75$$

$$\text{Q. D.} = \frac{\mathbf{Q3-Q1}}{\mathbf{2}} = \frac{38.75-33.06}{2} = \frac{5.69}{2} = 2.85$$

$$\text{Coefficient of Q.D} = \frac{\mathbf{Q3-Q1}}{\mathbf{Q3+Q1}} = \frac{38.75-33.06}{38.75+33.06} = \frac{5.69}{71.81} = 0.08$$

Merits:

1. It is simple to calculate.
2. It is easy to understand.
3. Risk of excrement item variation is eliminated, as it depends upon the central 50 percent items.

Demerits

1. Items below Q1 and above Q3 are ignored.
2. It is not capable of further mathematical treatment.
3. It is affected much by the fluctuations of sampling.
4. It is not calculated from a computed average, but from a positional average.

3. Mean Deviation:

The mean deviation is also known as the average deviation. It is the average difference between the items in a distribution computed from the mean, median or mode of that series counting all such deviation as positive. Median is preferred to the average because the sum of deviation of items from median is minimum when signs are ignored. But, the arithmetic mean is more frequently used in calculating the value of average deviation. Hence, it is commonly called Mean deviation.

Mean Deviation – Individual Series

$$\text{M. D. (mean or median or mode)} = \frac{\sum |\boldsymbol{D}|}{\boldsymbol{N}}$$

Coefficient of Mean Deviation: $\frac{\textbf{Mean Deviation}}{\textbf{Mean or median or mode}}$

Illustration 5: Calculate mean deviation from mean and median for the following data:

100	150	200	250	360	490	500	600	671

Solution: Calculation of Mean Deviation

X	**$\lvert D \rvert = X - \bar{X}$; X – 369**	**$\lvert D \rvert$ = X – median; X - 360**
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
$\sum X = 3321$	**$\sum \lvert D \rvert = 1570$**	**$\sum \lvert D \rvert = 1561$**

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

$= \frac{3321}{9} = 369$

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

= Value of $\frac{(9+1)}{2}$th item

= Value of 5th item = 360

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

$= \frac{1570}{9} = 174.44$

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

$= \frac{1561}{9} = 173.44$

Coefficient of M.D. = $\frac{\text{M.D.}}{\bar{X}}$

$= \frac{174.44}{369} = 0.47$

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

$= \frac{173.44}{360} = 0.48$

Mean Deviation – Discrete Series

$$\textbf{M. D.} = \frac{\sum f|D|}{N}$$

Illustration 6: Calculate mean deviation from mean from the following data:

X	2	4	6	8	10
F	1	4	6	4	1

Solution: Calculation of Mean Deviation

x	F	fx	$\|D\| = x - \bar{X}$	$f\|D\|$
2	1	2	4	4
4	4	16	2	8
6	6	36	0	0
8	4	32	2	8
10	1	10	4	4
	N = ∑f = 16	**∑fx = 96**		$\sum f\|D\| = 24$

Mean $\bar{X} = \frac{\sum fX}{N} = \frac{96}{16} = 6$

M.D. from mean = $\frac{\sum f|D|}{N} = \frac{24}{16} = 1.5$

Coefficient of M.D. = $\frac{\text{M.D.}}{\bar{X}} = \frac{1.5}{6} = 0.25$

Mean Deviation – Continuous Series

$$\textbf{M. D.} = \frac{\sum f|D|}{N}$$

Illustration 7:

Calculate mean deviation from mean from the following data:

Class interval	2 - 4	4 - 6	6 - 8	8 - 10
Frequency	3	4	2	1

Solution:

Calculation of Mean Deviation

x	**M**	**f**	**Fm**	$\lvert D \rvert = m - \bar{X}$	$f \lvert D \rvert$
2 – 4	3	3	9	2.2	6.6
4 – 6	5	4	20	0.2	0.8
6 – 8	7	2	14	1.8	3.6
8 - 10	9	1	9	3.8	3.8
		N = ∑f = 10	**∑fm = 52**		$\sum f \lvert D \rvert = 14.8$

$$\text{Mean } \bar{X} = \frac{\sum fm}{N} = \frac{52}{10} = 5.2$$

$$\text{M.D. from mean} = \frac{\sum f \lvert D \rvert}{N} = \frac{14.8}{10} = 1.48$$

$$\text{Coefficient of M.D.} = \frac{\text{M.D.}}{\bar{X}} = \frac{1.48}{5.2} = 0.29$$

Merits

1. It is clear and easy to understand.
2. It is based on each and every item of the data.
3. It can be calculated from any measure of central tendency and as such is flexible too.
4. It is not disturbed by the values of extreme items as in the case of range.

Demerits:

1. It is not suitable for further mathematical processing.
2. It is rarely used in sociological studies.

4. Standard Deviation

Karl Pearson introduced the concept of Standard deviation in 1893. Standard deviation is the square root of the means of the squared deviation from the arithmetic mean. So, it is called as Root - Mean Square Deviation or Mean Error or Mean Square Error. The Standard deviation is denoted by the small Greek letter 'σ' (read as sigma)

Standard Deviation – Individual Observation

Deviation taken from Actual Mean

$$\boldsymbol{\sigma} = \sqrt{\frac{\sum \mathbf{x^2}}{\mathbf{N}}} \text{ or } \boldsymbol{\sigma} = \sqrt{\frac{\sum (\mathbf{X} - \bar{\mathbf{X}})^2}{\mathbf{N}}} \text{ or } \boldsymbol{\sigma} = \sqrt{\frac{\sum \mathbf{X^2}}{\mathbf{N}} - \left(\frac{\sum \mathbf{X}}{\mathbf{N}}\right)^2}$$

Illustration 8: Calculate the standard deviation from the following data;

14, 22, 9, 15, 20, 17, 12, 11

Solution: Calculation of standard deviation from actual mean

Values (X)	$\mathbf{X^2}$	$\mathbf{X - \bar{X}}$; **(X - 15)**	$\mathbf{(X - \bar{X})^2}$
14	196	-1	1
22	484	7	49
9	81	-6	36
15	225	0	0
20	400	5	25
17	289	2	4
12	144	-3	9
11	121	-4	16
$\mathbf{\sum X = 120}$	$\mathbf{\sum X^2 = 1940}$		$\mathbf{\sum (X - \bar{X})^2 = 140}$

$$\mathbf{N = 8;\ \bar{X} = \frac{\sum X}{N} = \frac{120}{8} = 15}$$

$$\sigma = \sqrt{\frac{\sum x^2}{N}} \text{ or } \sigma = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

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

$$= \sqrt{17.5}$$

$$= 4.18$$

Alternatively:

We can find out standard deviation by using variables directly, i.e., no deviation is found out.

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

$$= \sqrt{\frac{1940}{8} - \left(\frac{120}{8}\right)^2}$$

$$= \sqrt{242.5 - 225}$$

$$= \sqrt{17.5}$$

$$= \mathbf{4.18}$$

Deviation taken from Assumed Mean

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

Where d = X – A

Illustration 9:

Calculate the standard deviation from the following data;

30, 43, 45, 55, 68, 69, 75.

Solution:

Calculation of standard deviation from assumed mean

X	d = X – A= X - 55	d^2
30	-25	625
43	-12	144
45	-10	100
55	0	0
68	13	169
69	14	196
75	20	400
N = 7	**$\sum d = 0$**	**$\sum d^2 = 1634$**

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

$$= \sqrt{\frac{1634}{7} - \left(\frac{0}{7}\right)^2}$$

$$= \sqrt{233.429}$$

$$= 15.28$$

Standard Deviation – Discrete Series: Actual Mean Method:

$$\sigma = \sqrt{\frac{\sum fd^2}{N}}$$

Illustration 10:

Calculate the standard deviation from the following data;

Marks	10	20	30	40	50	60
No. of students	8	12	20	10	7	3

Solution:

Calculation of standard deviation (from actual mean)

x	F	Fx	$d = x - \bar{X}$ $x - 30.8$	d^2	fd^2
10	8	80	-20.8	432.64	3461.12
20	12	240	-10.8	116.64	1399.68
30	20	600	-0.8	0.64	12.80
40	10	400	9.2	84.64	846.40
50	7	350	19.2	368.64	2580.48
60	3	180	29.2	852.64	2557.92
	$N = \sum f = 60$	$\sum fx = 1850$			$\sum fd2 = 10858.40$

$$\text{Mean: } \bar{X} = \frac{\sum fx}{N}$$

$$= \frac{1850}{60}$$

$$= 30.8$$

$$\text{Standard Deviation: } \sigma = \sqrt{\frac{\sum fd^2}{N}}$$

$$= \sqrt{\frac{10858.40}{60}}$$

$$= 13.45$$

Assumed Mean Method:

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

Where d = X – A

Illustration 11: (Solving the previous problem)

Solution:

Calculation of standard deviation (from assumed mean)

x	f	d = x -30	d^2	fd	fd^2
10	8	-20	400	-160	3200
20	12	-10	100	-120	1200
30	20	0	0	0	0
40	10	10	100	100	1000
50	7	20	400	140	2800
60	3	30	900	90	2700
	$\mathbf{N = \sum f = 60}$			$\sum$fd = 50	$\sum \mathbf{f}d^2$ = 10900

$$\boldsymbol{\sigma} = \sqrt{\frac{\sum \mathbf{fd^2}}{\mathbf{N}} - (\frac{\sum \mathbf{fd}}{\mathbf{N}})^{\mathbf{2}}}$$

$$= \sqrt{\frac{\mathbf{10900}}{\mathbf{60}} - (\frac{\mathbf{50}}{\mathbf{60}})^{\mathbf{2}}}$$

$$= \sqrt{181.67 - 0.69}$$

$$= \sqrt{180.98}$$

$$\mathbf{= 13.45}$$

Step Deviation Method

$$\boldsymbol{\sigma} = \sqrt{\frac{\sum \mathbf{fd^2}}{\mathbf{N}} - (\frac{\sum \mathbf{fd}}{\mathbf{N}})^{\mathbf{2}}} \text{ x C}$$

Where d' = $\frac{X - A}{C}$; C = Common Factor

Illustration 12:

(Solving the previous problem)

Solution:

Calculation of standard deviation (from step deviation)

x	f	$d' = \frac{X-30}{10}$	d'^2	fd'	fd'^2
10	8	-2	4	-16	32
20	12	-1	1	-12	12
30	20	0	0	0	0
40	10	1	1	10	10
50	7	2	4	14	28
60	3	3	9	9	27
	$N = \sum f = 60$			$\sum fd' = 5$	$\sum fd'^2 = 109$

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

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

$$= \sqrt{1.817 - 0.0069} \times 10$$

$$= \sqrt{1.81} \times 10 = 1.345 \times 10$$

$$\sigma = 13.45$$

Standard Deviation – Continuous Series

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

Where $d = \frac{m - A}{C}$; C = Common Factor

Illustration13:

Compute the standard deviation from the following data:

Class	0 - 10	10 - 20	20 - 30	30 - 40	40-50
Frequency	5	8	15	16	6

Solution:

Computation of standard deviation

x	M	F	$d = \frac{m-25}{10}$	d^2	fd	fd^2
0 - 10	5	5	-2	4	-10	20
10 - 20	15	8	-1	1	-8	8
20 - 30	25	15	0	0	0	0
30 - 40	35	16	1	1	16	16
40 - 50	45	6	2	4	12	24
		$N = \sum f = 50$			$\sum fd = 10$	$\sum fd^2 = 68$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \text{ x C}$$

$$\sigma = \sqrt{\frac{68}{50} - \left(\frac{10}{50}\right)^2} \text{ x } 10 = \sqrt{1.36 - (0.2)^2} \text{ x } 10$$

$$= \sqrt{1.36 - 0.04} \text{ x } 10 = \sqrt{1.32} \text{ x } 10$$

$$= 1.1489 \text{ x } 10 = \mathbf{11.49}$$

Merits:

1. It is rigidly defined determinate.

2. It is based on all the observations of a series.

3. It is less affected by fluctuations of sampling and hence stable.

4. It is amenable to algebraic treatment and is less affected by fluctuations of sampling most other measures of dispersion.

5. The standard deviation is more appropriate mathematically than the mean deviation, since the negative signs are removed by squaring the deviations rather than by ignoring

Demerits:

1. It lacks wide popularity as it is often difficult to compute, when big numbers are involved, the process of squaring and extracting root becomes tedious.

2. It attaches more weight to extreme items by squaring them.

3. It is difficult to calculate accurately when a grouped frequency distribution has extreme groups with no definite range.

Uses:

1. Standard deviation is the best measure of dispersion.

2. It is widely used in statistics because it possesses most of the characteristics of an ideal measure of dispersion.

3. It is widely used in sampling theory and by biologists.

4. It is used in coefficient of correlation and in the study of symmetrical frequency distribution.

Co - efficient of variation (Relative Standard Deviation)

The Standard deviation is an absolute measure of dispersion. The corresponding relative measure is known as the co - efficient of variation. It is used to compare the variability of two or more than two series. The series for which co-efficient or variation is more is said to be more variable or conversely less consistent, less uniform less table or less homogeneous.

Variance:

Square of standard deviation is called variance.

$$\text{Variance} = \sigma^2; \sigma = \sqrt{\textbf{Variance}}$$

$$\text{Co – efficient of standard deviation} = \frac{\boldsymbol{\sigma}}{\bar{\mathbf{x}}}$$

$$\text{Co – efficient of variation (C.V.)} = \frac{\boldsymbol{\sigma}}{\bar{\mathbf{x}}} \text{ x } 100$$

Illustration 14: The following are the runs scored by two batsmen A and B in ten innings:

A	101	27	0	36	82	45	7	13	65	14
B	97	12	40	96	13	8	85	8	56	15

Who is the more consistent batsman?

Solution: Calculation of Co-efficient of Variation

Batsman A			Batsman B		
Runs Scored X	**dx = X - $\bar{X}$**	**dx^2**	**Runs Scored Y**	**dx = Y - $\bar{Y}$**	**dy^2**
101	62	3844	97	54	2916
27	-12	144	12	-31	961
0	-39	1521	40	-3	9
36	-3	9	96	53	2809
82	43	1849	13	-30	900
45	6	36	8	-35	1225
7	-32	1024	85	42	1764
13	-26	676	8	-35	1225
65	26	676	56	13	169
14	-25	625	15	-28	784
$\sum X = 390$		**$\sum dx^2 = 10404$**	**$\sum Y = 430$**		**$\sum dy^2 = 12762$**

Batsman A

$$\bar{X} = \frac{\sum X}{N} = \frac{390}{10} = 39$$

$$\sigma_X = \sqrt{\frac{\sum dx^2}{N}} = \sqrt{\frac{10404}{10}} = 32.26$$

$$C.V. = \frac{\sigma}{\bar{X}} \times 100$$

$$= \frac{32.26}{39} \times 100$$

$$= 82.72\%$$

Batsman B

$$\bar{Y} = \frac{\sum Y}{N} = \frac{430}{10} = 43$$

$$\sigma_Y = \sqrt{\frac{\sum dy^2}{N}} = \sqrt{\frac{12762}{10}} = 35.72$$

$$C.V. = \frac{\sigma}{\bar{Y}} \times 100$$

$$= \frac{35.72}{43} \times 100$$

$$= 83.07\%$$

Batsman A is more consistent in his batting, because the co-efficient of variation of runs is less for him.

Moments:

"Moment is a familiar mechanical term for the measure of a force with reference to its tendency to produce rotation. The strength of this tendency depends, obviously, upon the amount of the force and the distance from the origin of the point at which the force is exerted."

Moment is a term generally used in physics, mechanics and refers to the turning effect or rotating effect of a force. When it is applied in statistics, it describes the various' characteristics of frequency distribution, viz., central tendency, dispersion, skewness and kurtosis. Moments can be defined as the arithmetic mean of various powers of deviations taken from the mean of a distribution.

Moments about Mean or Central Moments:

For a frequency distribution, the first moment about the arithmetic mean is defined as the mean of deviations of items taken from their arithmetic mean. The first four moments about arithmetic mean or central moments are defined below:

	Individual series	**Discrete series**
First Moment about the Mean;μ_1	$\frac{\sum(X-\bar{X})}{N}=\frac{\sum d}{N}=0$	$\frac{\sum f(X-\bar{X})}{N}=\frac{\sum fd}{N}$
Second Moment about the Mean μ_2	$\frac{\sum(X-\bar{X})2}{N}=\frac{\sum d2}{N}=\sigma^2$	$\frac{\sum f(X-\bar{X})2}{N}=\frac{\sum fd2}{N}=\sigma^2$
Third Moment about the Mean μ_3	$\frac{\sum(X-\bar{X})3}{N}=\frac{\sum d3}{N}$	$\frac{\sum f(X-\bar{X})3}{N}=\frac{\sum fd3}{N}$
Fourth Moment about the Mean μ_4	$\frac{\sum(X-\bar{X})4}{N}=\frac{\sum d4}{N}$	$\frac{\sum f(X-\bar{X})4}{N}=\frac{\sum fd4}{N}$

μ is a Greek letter, pronounced as 'mu'.

The sum of deviation of items from arithmetic mean is always zero. Therefore, μ_1 would always be zero.

$$\mu_2=\sigma^2$$

$$\sigma=\sqrt{\mu_2}$$

Two important constants of a distribution are calculated from μ_2, μ_3 and μ_4. They are:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \qquad \beta_2 = \frac{\mu_4}{\mu_2^2}$$

β_1 measures skewness and β_2 kurtosis

Properties of Central Moments:

- The first moment about mean is always zero. i.e $\mu_1 = 0$
- The second moment about mean measures variance. That is,

  $\mu_2 = \sigma^2$ or $\sigma = \pm\sqrt{\mu_2}$.
- The third moment about mean measures skewness:
  - If $\mu_3 > 0$, the given distribution is positively skewed.
  - If $\mu_3 < 0$, the given distribution is negatively skewed.
  - If $\mu_3 = 0$, the given distribution is symmetrical.
- All odd moments (in a symmetrical distribution) are zero

  $\mu_1 = \mu_3 = \mu_5 = \mu_7$ .....
- The fourth moment about mean measures kurtosis. $\beta_1 = \frac{\mu_4}{\mu_2^2}$

Skewness:

A distribution which is not symmetrical is called a skewed distribution and in such distributions, the Mean, the Median and the Mode will not coincide, but the values are pulled apart. If the curve has a longer tail towards the right, it is said to be positively skewed. If the curve has a longer tail towards the left, it is said to be negatively skewed.

Test of Skewness:

The absence of asymmetry or skewness can be stated under the following conditions. In other words, when a distribution is symmetric, the following conditions are satisfied :

a) The value of the Mean, the Median and the Mode coincide. (The values are equal).
b) Q3 - Median = Median - Q1.
c) The sum of positive deviations are equal to the sum of negative deviations.
d) The frequencies on either side of the mode are equal.
e) If plotted on a graph paper and folded at the centre of the curve (ordinate), the two halves are equal. (be shaped curve).

Similarly, a skewed distribution will have the following characteristics :

a. X ≠ Median ≠ Z
b. Q3 -Median ≠ Median - Q1 .
c. The sum of positive deviations ≠ the sum of negative deviations.
d. The frequencies on either side of the mode are unequal.
e. The plotted graph, folded at the ordinate will have unequal halves.

Measures of Skewness:

The measures of asymmetry are usually called measures of skewness. Measures of skewness indicate not only the extent of skewness (in numerical expressions), but also the direction, the manner in which the deviations are distributed. These measures can be absolute or relative. The absolute measures are also known as measures of skewness. The relative measures are known as the coefficient of skewness. The absolute measure tells us the extent of asymmetry, whether it is positive or negative.

Symbolically,

Absolute skewness = Mean - Mode

= + Positive skewness:

= - Negative skewness:

- If the value of the Mean is greater than the Mode, the skewness is positive.
- If the value of the Mode is greater than the Mean, the skewness is negative.
- Greater the amount of skewness (negative or positive), the more the tendency towards asymmetry.
- The absolute measure of skewness will not be the proper measure for comparison, and hence in each series a relative measure or coefficient of skewness will have to be computed.

Objective of Skewness:

❖ Measures of skewness tell us the direction and extent of asymmetry in a series, and the permit us to compare two or more series with regard to these.

❖ Measures of skewness give an idea about the nature of variation of the items about the central value.

Kurtosis:

Kurtosis is a statistical measure used to describe the degree to which scores cluster in the tails or the peak of a frequency distribution. The peak is the tallest part of the distribution, and the tails are the ends of the distribution.

The expression 'Kurtosis' is used to describe the peakedness of a curve. The three measures - central tendency, dispersion and skewness, describe the characteristic of frequency distribution.

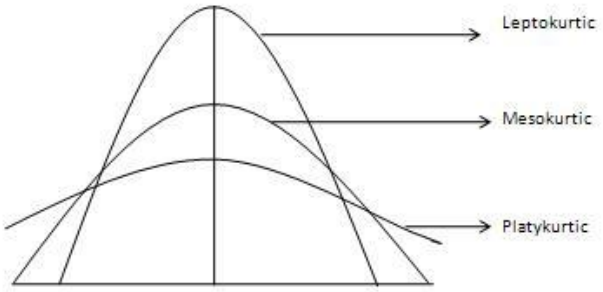
Measures of Kurtosis:

The measures of kurtosis of a frequency distribution are based upon the fourth moment about the mean of the distribution. Symbolically,

$$\beta_2 = \mu^4/\mu_2^2$$

μ^4 - 4^{th} Moment; μ_2^2 - 2^{nd} Moment

- If $\beta_2 = 3$, the distribution is said to be normal, and the curve is Mesokurtic.
- If $\beta_2 > 3$, the distribution is said to be more peaked, and the curve is Leptokurtic.
- If $\beta_2 < 3$, the distribution is said to be flat topped, and the curve is Platykurtic.



- **Mesokurtic:**

  Distributions that are moderate in breadth and curves with a medium peaked height.

- **Leptokurtic:**

  More values in the distribution tails and more values close to the mean (i.e. sharply peaked with heavy tails)

- **Platykurtic:**

  Fewer values in the tails and fewer values close to the mean (i.e. the curve has a flat peak and has more dispersed scores with lighter tails).

*****

Important Questions:

I. Fill in the blanks :

1. In a symmetrical distribution mean **is equal to** median **is equal to** mode.

2. Median is same as **second** quartile.

3. Median is a **positional** average, dividing the series when arranged as an array into **two equal** parts.

4. Median and mode are called **positional** averages.

5. The **quartile one and three** mark off the limits within which the middle 50% of the items lie.

6. **Median or Mode** can be calculated from a frequency distribution with open and classes.

7. Median is the average suited for **Open-End** classes.

8. Quartile deviation is **0.6745** of standard deviation.

9. Standard deviation is always **less** than range.

10. Variance is **square** of standard deviation.

11. Quartile deviation is **absolute** measure of dispersion.

II. Choose the correct answer:

1. Which average is affected most by extreme observations?

(a) Mode (b) Median **(C) Geometric Mean** (d) Arithmetic mean

2. Which of the following is the most unstable average?

(a) **Mode** (b) Median (c) Geometric mean (d) Harmonic mean

3. For dealing with qualitative data the best average is:

(a) Arithmetic mean (b) Geometric mean (c) Harmonic mean (d) **Median**

4. The sum of deviations taken from arithmetic mean is:

(a) Minimum (b) **Zero** (c) Maximum (d) Equal

5. The sum of squares of deviations from arithmetic mean is:

(a) Zero (b) Maximum (c) **Minimum** (d) Equal

6. When calculating the average growth of economy, the correct mean to use is?

(a) Weighted mean (b) **Geometric Mean** (c) Arithmetic mean (d) Harmonic Mean

7. When an observation in the data is zero, then its geometric mean is?

(a) Negative **(b) Zero** (c) Positive (d) Cannot be calculated.

8. The best measure of central tendency is:

(a) Arithmetic Mean (b) Geometric mean (c) Harmonic mean (d) Weighted mean

9. The point of intersection of the 'less than 'and 'more than ' ogives corresponds to:

(a) Mean **(b) Median** (c) Geometric mean (d) Mode

10. Sum of absolute deviations about median is :

(a) The Least (b) The greatest (c) Zero (d) Equal

11. The sum of squares of deviations is least when measured from:

(a) Median (b) Zero **(c) Mean** (d) Mode

12. The appropriate measure whenever the extreme items are to be disregarded and when the distribution contains indefinite classes at the end is :

(a) Median (b) Mode **(c) Quartile Deviation** (d) Mean Deviation

13. The quartile deviation includes the:

(a) First 50% (b) Last 50% **(c) Central 50%** (d) Atleast 50%

14. Which of the following is a relative measure of dispersion:

(a) Variance **(b) Coefficient of Variance** (c) Standard Deviation (d) Mean Deviation

15. The square of the variance of a distribution is the :

(a) Median (b) Mean (c) Mode **(d) None of these.**

III. Theoretical Questions:

1. Explain what is meant by dispersion. What are the methods of computing dispersion? What is the practical utility of such measures?

2. What is meant by a measure of dispersion? State the different methods of measuring it.

3. Why is that standard deviation is considered to be the most popular measure of dispersion?

4. What is coefficient of variation? What purpose does it serve? Also distinguish between ' variance' and 'coefficient of variation'.

5. Define coefficient of variation? In what situation would you prefer this as a measure of dispersion?

6. Define mean deviation. How does it differ from standard deviation?

7. Dispersion is known as the second average of the second order. Discuss.

8. What are the properties of a good measure of variation?

10. What are the requisites of a good measure of dispersion?

9. What are quartiles? How are they used for measuring dispersion?

10. How do you define coefficient of variation and what are its uses?

11. What are the objectives of measuring dispersion of a frequency distribution? Explain.

12. "Average and measures of dispersion are useful in understanding a frequency distribution". Elucidate the statement giving illustrations.

Practical Problems:

Measures of Central Tendency:

1. The monthly income of 10 families of a certain locality is given in rupees as below. Calculate the arithmetic average.

Families	A	B	C	D	E	F	G	H	I	J
Income in rupees	85	70	10	75	500	8	42	250	40	36

(Mean = Rs. 111. 60)

2. The coins are tossed 1024 times. The theoretical frequencies of 10 heads to 0 head are given below. Calculate the mean number of heads per tossing.

No. of heads	0	1	2	3	4	5	6	7	8	9	10
Frequency	1	10	45	120	210	252	210	120	45	10	1

(Mean = Rs. 5)

3. Find mean from the following frequency distribution:

Class Interval	15 – 25	25 – 35	35 – 45	45 – 55	55 – 65	65 - 75
Frequency	4	11	19	14	0	2

(Mean = Rs. 40. 2)

4. The following are the marks scored by 7 students; find out the median marks:

Roll Numbers	1	2	3	4	5	6	7
Marks	45	32	18	57	65	28	46

(Median marks = 45)

5. Find out the median from the following:

57	58	61	42	38	65	72	66

(Median = 59.5)

6. Find the median

X	13	14	15	16	17	18	19	20	21	22	23	24	25
F	37	162	343	390	256	433	161	355	65	85	49	46	40

(Median = 18)

7. Find the median:

Wages Rs.	60 – 70	50 – 60	40 – 50	30 – 40	20 - 30
No. of labourers	5	10	20	5	3

(Median = 46.75)

8. 10 persons have the following income:

Rs.	850	750	600	825	850	725	600	850	640	530

(Mode = 850)

9. Calculate the mode from the following series:

Size	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Frequency	40	48	52	57	60	63	57	55	50	52	41	57	63	52	48	40

(Mode = 9)

10. Find the mode:

Size	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
Frequency	5	7	12	18	16	10	5

(Mode = 37.5)

11. Calculate mean, median and mode form the following frequency distribution of marks at a test in statistics:

Marks	5	10	15	20	25	30	40	45	50
No. of students	20	43	75	76	72	45	9	8	50

(Mean = 22. 16; Median = 20; mode = 20)

12. Calculate the mean, median and mode for the following data.

Profits per shop	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 - 60
No. of shops	12	18	27	20	17	6

(Mean = 28; Median = 27.4; mode = 25.62)

Dispersion:

1. Calculate Range, Q.D, M.D (from mean), S.D and C.V of the marks obtained by 10 students given below:

50	55	57	49	54	61	64	59	59	56

(Range = 15, Q.D = 2.75, M. D = 3.6, S.D = 4.43 and C.V = 7.85%)

2. Compute Q.D from the following data:

Height in inches	58	59	60	61	62	63	64	65	66
No. of students	15	20	32	35	33	22	20	10	8

(Q.D = 1.5)

3. Compute quartile deviation from the following data.

x:	4 – 8	8 – 12	12 - 16	16 - 20	20 - 24	24 - 28	28 - 32	32 – 36	36 - 40
f:	6	10	18	30	15	12	10	6	2

(Q.D = 5.21)

4. Calculate mean deviation (from mean) from the following data:

x:	2	4	6	8	10

f:	1	4	6	4	1

(M.D = 1.5)

5. Calculate the mean deviation from the following data.

x:	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30	30 - 35	35 - 40
f:	449	705	507	281	109	52	16	4

(M.D = 5.25)

6. Calculate the S.D of the following:

Size of the item	6	7	8	9	10	11	12
Frequency	3	6	9	13	8	5	4

(S.D = 1.6)

7. Compute standard deviation of the following data:

Wages(Rs. Per day)	1 – 3	3 – 5	5 – 7	7 – 9	9 – 11
No. of workmen	15	18	27	10	6

(S.D = 2.33)

8. Two cricketers scored the following runs in the several innings. Find who is a better run getter and who more consistent player is?

A:	42	17	83	59	72	76	64	45	40	32
B:	28	70	31	0	59	108	82	14	3	95

(C.V (A) = 38% and C.V (B) = 75.6%)

9. Two brands of types are tested with the following results:

Life („000 miles)	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45	45 – 50
Brand X	8	15	12	18	13	9
Brand Y	6	20	32	30	12	0

(C.V (X) = 21.82% and C.V (Y) = 16.11%)

*****

UNIT - II

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. (18 hrs)

Meaning:

Correlation refers to the relationship of two or more variables. For example, there exists some relationship between the height of a mother and the height of a daughter, sales and cost and so on. Hence, it should be noted that the detection and analysis of correlation between two statistical variables requires relationship of some sort which associates the observation in pairs, one of each pair being a value of the two variables. The word 'relationship' is of important and indicates that there is some connection between the variables under observation. Thus, the association of any two variates is known as correlation.

Significance:

Correlation is useful in physical and social sciences. We can study the uses of correlation in business and economics. The following are the significance of study of correlation:

- Correlation is very useful to economics to study the relationship between variables, like price and quantity demanded. To the businessmen, it helps to estimate costs, sales, price and other related variables.
- Some variables show some kind of relationship; correlation analysis helps in measuring the degree of relationship between the variables like supply and demand, price and supply, income and expenditure, etc.
- The relation between variables can be verified and tested for significance, with the help of the correlation analysis. The effect of correlation is to reduce the range of uncertainty of our prediction.
- The coefficient of correlation is a relative measure and we can compare the relationship between variables which are expressed in different units.
- Sampling error can also be calculated.
- Correlation is the basis for the concept of regression and ratio of variation.

Correlation and Causation:

Correlation analysis deals with the association or co-variation between two or more variables and helps to determine the degree of relationship between two or more variables. But correlation does not indicate a cause and effect relationship between two variables. It explains only co-variation. The high degree of correlation between two variables may exist due to anyone or a combination of the following reasons.

1. Pure Chance:

Especially in a small sample, the correlation is due to pure chance. If we select a small sample from bivariate distribution, it may show a high degree of correlation, but in the universe there is no relationship between the variables. A high degree of mathematical correlation can be obtained even at the time when there is no relationship between the variables. For example, the comparison of the production of shoes with the agricultural production which have no relationship, may have a relationship. But if a relationship is formed, it may be only a chance of coincidence, and such type of correlation is called nonsensical or spurious correlation. Another example is the relationship between cars produced and the children born in a country.

Here this covariation may be due to chance and there is no logical basis for relationship. A measure of correlation may be arrived at on the basis of covariation, but it may be nonsensical or without meaning. That is, there may be correlation between two variables when the two variables do not operate in the same physical or social systems, when they have nothing to do with each other. And, again, such types of correlation is known as spurious or nonsensical correlation.

2. Both variables are influenced by some other variables :

A high degree of correlation between the variables may be due to some cause or different causes effecting each of these variables. For example, a high degree of correlation may exist between the yield per acre of paddy or wheat due to the effect of rainfall and other factors like fertilizers used, favourable weather condition, etc. But none of the two variables is the cause of the other. It is difficult to explain which is the cause and which is the effect; they may not have caused each other; nor one caused the other. But there is an outside influence.

3. Mutual Dependence :

In this, the variables affect each other. The subject and relative variable are to be judged for the circumstances. For example, the production of jute and rainfall. Rainfall is the subject and jute production is relative. The effect of the rainfall is directly related to the jute production.

Types of Correlation:

Correlation is classified into many types but the important are:

1. Positive and Negative Correlation
2. Simple and Multiple Correlations
3. Partial and Total Correlation
4. Linear and Non-linear Correlation.

1. Positive and Negative Correlation :

The correlation is said to be positive when the values of two variables move in the same direction, so that an increase in the value of one variable is accompanied by an increase in the value of the other variable or a decrease in the value of one variable is followed by a decrease in the value of the other variable. Example: Height and weight, rainfall and yield of crops, etc.,

The correlation is said to be negative when the values of two variables move in opposite direction, so that an increase or decrease in the values of one variable is followed by a decrease or increase in the value of the other. Example: Price and demand, yield of crops and price, etc.,

2. Simple and multiple Correlation :

When we study only two variables, the relationship is described as simple correlation; Example: The study of price and demand of an article.

When more than two variables are studied simultaneously, the correlation is said to be multiple correlation. Example: the relationship of price, demand and supply of a commodity.

3. Partial and total Correlation:

Partial correlation coefficient provides a measure of relationship between a dependent variable and a particular independent variable when all other variables involved are kept constant. i.e., when the effect of all other variables are removed.

Example: When we study the relationship between the yield of rice per acre and both the amount of rainfall and the amount of fertilizers used. In these relationship if we limit our correlation analysis to yield and rainfall. It becomes a problem relating to simple correlation.

4. Linear and Non-linear Correlation :

The correlation is said to be linear, if the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable.

The correlation is non-linear, if the amount of change in one variable does not bear a constant ratio to the amount of change in the other related variable.

Methods of Studying Correlation:

The following correlation methods are used to find out the relationship between two variables.

A. Graphic Method :
- **i.** Scatter diagram (or) Scattergram method.
- **ii.** Simple Graph or Correlogram method.

B. Mathematical Method :
- **i.** Karl Pearson's Coefficient of Correlation.
- **ii.** Spearman's Rank Correlation of Coefficient
- **iii.** Coefficient of Concurrent Deviation
- **iv.** Method of Least Squares.

C. Graphic Method

i. Scatter Diagram Method of Correlation:

This is the simplest method of finding out whether there is any relationship present between two variables by plotting the values on a chart, known as scatter diagram. In this method, the given data are plotted on a graph paper in the form of dots. X variables are plotted on the horizontal axis and Y variables on the vertical axis. Thus we have the dots and we can know the scatter or concentration of various points.

If the plotted dots fall in a narrow band and the dots are rising from the lower left hand corner to the upper right-hand corner it is called high degree of positive correlation.

If the plotted dots fall in a narrow band from the upper left hand corner to the lower right hand corner it is called a high degree of negative correlation.

If the plotted dots line scattered all over the diagram, there is no correlation between the two variables.

Merits:

1. It is easy to plot even by beginner.
2. It is simple to understand.
3. Abnormal values in a sample can be easily detected.
4. Values of some dependent variables can be found out.

Demerits:

1. Degree of correlation cannot be predicted.
2. It gives only a rough idea.
3. The method is useful only when number of terms is small.

ii. Simple Graph Method of Correlation:

In this method separate curves are drawn for separate series on a graph paper. By examining the direction and closeness of the two curves we can infer whether or not variables are related. If both the curves are moving in the same direction correlation is said to be positive. On the other hand, if the curves are moving in the opposite directions is said to be negative.

Merits:

1. It is easy to plot
2. Simple to understand
3. Abnormal values can easily be deducted.

Demerits:

1. This method is useless when number of terms is very big.
2. Degree of correlation cannot be predicted.

B. Mathematical Method:

i. Karl Pearson's Coefficient of Correlation:

Karl Pearson, a great biometrician and statistician, introduced a mathematical method for measuring the magnitude of linear relationship between two variables. This method is most widely used in practice. This method is known as Pearsonian Coefficient of Correlation. It is denoted by the symbol $'r'$; the formula for calculating Pearsonian r is:

(i) r = $\frac{Covariance\ of\ xy}{\sigma x \times \sigma y}$, (ii) r = $\frac{\Sigma xy}{N\sigma x \times \sigma y}$, (iii) r = $\frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$

$$x = (X - \overline{X}), y = (Y - \overline{Y})$$

$$\sigma x = Standard\ deviation\ of\ series\ x$$

$$\sigma y = Standard\ deviation\ of\ series\ y$$

The value of the coefficient of correlation shall always lie between **+1 and -1**.

When *r* = +1, then there is perfect positive correlation between the variables.

When *r* = -1, then there is perfect negative correlation between the variables.

When *r* = 0, then there is no relationship between the variables.

Illustration 1: Calculate Karl Pearson coefficient of correlation from the following data:

X	100	101	102	102	100	99	97	98	96	95
Y	98	99	99	97	95	92	95	94	90	91

Solution: Calculation of coefficient of correlation

X	x **= X -** $\bar{X}$ **=X – 99**	x^2	**Y**	**y = Y -** $\bar{Y}$ **=Y - 95**	y^2	Xy
100	1	1	98	3	9	3
101	2	4	99	4	16	8
102	3	9	99	4	16	12
102	3	9	97	2	4	6
100	1	1	95	0	0	0
99	0	0	92	-3	9	0
97	-2	4	95	0	0	0
98	-1	1	94	-1	1	1
96	-3	9	90	-5	25	15
95	-4	16	91	-4	16	16
∑ X = 990		**∑** x^2 **= 54**	**∑ Y = 950**		**∑** y^2 **= 96**	**∑ xy = 61**

$\bar{X} = \frac{\Sigma X}{N} = \frac{990}{10} = 99$; $\bar{Y} = \frac{\Sigma Y}{N} = \frac{950}{10} = 95$;

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

$$= \frac{61}{\sqrt{54 \times 96}} = \frac{61}{\sqrt{5184}}$$

$$= \frac{61}{72}$$

$$= + 0.85$$

Illustration 2: Calculate Karl Pearson coefficient of correlation from the following data:

X:	6	2	10	4	8
Y:	9	11	5	8	7

Solution: Calculation of coefficient of correlation

X	**X^2**	**Y**	**Y^2**	**XY**
6	36	9	81	54
2	4	11	121	22
10	100	5	25	50
4	16	8	64	32
8	64	7	49	56
$\sum X = 30$	**$\sum X^2 = 220$**	**$\sum Y = 40$**	**$\sum Y^2 = 340$**	**$\sum XY = 214$**

$$r = \frac{N\Sigma XY - (\Sigma X.\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2} \times \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{(5 \times 214) - (30 \times 40)}{\sqrt{5 \times 220 - (30)^2} \times \sqrt{5 \times 340 - (40)^2}}$$

$$= \frac{1070 - 1200}{\sqrt{1100 - 900} \times \sqrt{1700 - 1600}}$$

$$r = \frac{-130}{\sqrt{200} \times \sqrt{100}} = \mathbf{-\ 0.92}$$

RANK CORRELATION CO-EFFICIENT

Spearman's Rank Correlation Co-efficient:

In 1904, a famous British psychologist Charles Edward Spearman found out the method of ascertaining the coefficient of correlation by ranks. This method is based on rank. Rank correlation is applicable only to individual observations. This measure is useful in dealing with qualitative characteristics such as intelligence, beauty, morality, character, etc.,

The formula for Spearman's rank correlation which is denoted by P is;

$$P = 1 - \frac{6\Sigma D^2}{N(N^2-1)}$$

or

$$P = 1 - \frac{6\Sigma D^2}{(N^3-N)}$$

Where, P = Rank co-efficient of correlation

D = Difference of the two ranks

$\sum D^2$ = Sum of squares of the difference of two ranks

N = Number of paired observations

Like the Karl Pearson's coefficient of correlation, the value of **P** lies between **+ 1** and **– 1**.

Where ranks are given

Illustration 3: Following are the rank obtained by 10 students in two subjects, Statistics and Mathematics. To what extent the knowledge of the students in the two subjects is related?

Statistics	1	2	3	4	5	6	7	8	9	10
Mathematics	2	4	1	5	3	9	7	10	6	8

Solution:

Calculation of Pearman's rank correlation coefficient

Rank of Statistics (*x*)	**Rank of Mathematics (*y*)**	**D = *x* – *y***	**D^2**
1	2	-1	1
2	4	-2	4
3	1	+2	4
4	5	-1	1

5	3	+2	4
6	9	-3	9
7	7	0	0
8	10	-2	4
9	6	+3	9
10	8	+2	4
N = 10			$\boldsymbol{\Sigma D^2 = 40}$

$$P = 1- \frac{6\Sigma D^2}{N(N^2-1)}$$

$$P = 1- \frac{6 \text{ x } 40}{10(10^2-1)}$$

$$= 1- \frac{240}{10(100-1)}$$

$$= 1- \frac{240}{990}$$

$$= 1 - 0.24$$

$$= + \mathbf{0.76}$$

Where Ranks are not given:

Illustration 4:

A random sample of 5 college students is selected and their grades in Mathematics and Statistics are found to be:

Mathematics	85	60	73	40	90
Statistics	93	75	65	50	80

Solution:

Calculation of Pearman's rank correlation coefficient

Mathematics (*x*)	Rank x	Statistics (*y*)	Rank y	D = $x - y$	D^2
85	2	93	1	+1	1
60	4	75	3	+1	1
73	3	65	4	-1	1
40	5	50	5	0	0
90	1	80	2	-1	1
					$\boldsymbol{\Sigma D^2 = 4}$

$$P = 1 - \frac{6\Sigma D^2}{N(N^2-1)}$$

$$= 1 - \frac{6 \times 4}{5(5^2-1)}$$

$$= 1 - \frac{24}{5(25-1)}$$

$$= 1 - \frac{24}{5(24)}$$

$$= 1 - \frac{1}{5} = 1 - 0.2$$

$$\mathbf{= + 0.8}$$

Equal or Repeated Ranks:

When two or more items have equal values, it is difficult to give ranks to them. In that case the items are given the average of the ranks they would have received, if they are not tied. A slightly different formula is used when there is more than one item having the same value.

$$\mathbf{P = 1 - 6\left\{\frac{\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \cdots}{N^3 - N}\right\}}$$

m = the number of items whose ranks are common

Illustration 5: From the following data calculate the rank correlation coefficient after making adjustment for tied ranks.

X	48	33	40	9	16	16	65	24	16	57
Y	13	13	24	6	15	4	20	9	6	19

Solution: Calculation of Pearman's rank correlation coefficient

X	Rank *x*	Y	Rank *y*	D =R(*x*) – R(*y*)	D^2
48	8	13	5.5	2.5	6.25
33	6	13	5.5	0.5	0.25
40	7	24	10	-3.0	9.00
9	1	6	2.5	-1.5	2.25
16	3	15	7	4.0	16.00
16	3	4	1	2.0	4.00
65	10	20	9	1.0	1.00
24	5	9	4	1.0	1.00
16	3	6	2.5	0.5	025
57	9	19	8	1.0	1.00
					$\boldsymbol{\Sigma D^2 = 41}$

$$\mathbf{P = 1 - 6\{\frac{\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)\ldots}{N^3 - N}\}}$$

$$= 1 - 6\{\frac{41 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\ldots}{10^3 - 10}\}$$

$$= 1 - \{\frac{6\ (41+2+0.5+05)}{990}\}$$

$$= 1 - \{\frac{264}{990}\} = 1 - 0.267$$

$$= \mathbf{+\ 0.\ 733}$$

Merits:

1. It is simple to understand and easier to apply.

2. It can be used to any type of data, qualitative or quantitative.

3. It is the only method that can be used where we are given the ranks and not the actual data.

4. Even where actual data are given, rank method can be applied for ascertaining correlation by assigning the ranks to each data.

Demerits:

1. This method is not useful to find out correlation in a grouped frequency distribution.

2. For large samples it is not convenient method. If the items exceed 30 the calculations become quite tedious and require a lot of time.

3. It is only an approximately calculated measure as actual values are not used for calculations.

Important Questions:

I. Fill in the blanks:

1. The coefficient of correlation is independent of change of **Scale** and **Origin**.

2. If r is more than six times **Probable Error** it is called significant.

3. The relationship between Three or more variables is studied with the help of **Multiple** correlation.

4. If r= 0.3,r 2 will be **0.09**

5. The coefficient of correlation is under -root of two **Regression Coefficients**.

II. Tick the correct answer :

1. The coefficient of correlation :

(a) has no limits

(b) can be less than 1

(c) can be more than 1

(d) **Varies Between +- 1**

2. The value of r2 for a particular situation is 0.81. What is coefficient of correlation :

(a) 0.81 (b) **0.9** (c) 0.09.

3. Which of the following is the highest range of r?

(a) 0 and 1 (b) -1 and 0 (c) **-1 and 1**

4. The coefficient of correlation is independent of :

(a) change of scale only

(b) change of origin only

(c) **Both Change of Scale and Origin**.

5. The coefficient of correlation :

(a) cannot be positive

(b) cannot be negative

(c) **Can be either Positive or Negative**.

III. Theoretical Questions:

1. What is meant by correlation? What are the properties of the coefficient of correlation?

2. (a) Distinguish coefficient of correlation from coefficient of variation.

(b) What is scatter diagram? How does it helps us in studying the correlation between two variables, in respect of both their nature and extent?

3. Define Karl Pearson's coefficient of correlation. What is it intended to measure?

4. Distinguish between:

(a) Positive and negative correlation.

(b) Linear and non-linear correlation,

(c) Simple, partial and multiple correlation.

5. What are the advantages of spearman's rank correlation over Karl Pearson's correlation coefficient? Explain the method of calculating Spearman's rank correlation coefficient.

6. Define, coefficient of concurrent Deviations' and comment on its usefulness.

7. What is 'spurious' or non-sensual correlation? Explain with example.

8. Define coefficient of correlation and mention it's important properties.

9. What are the methods of calculating coefficient of correlation?

10. Explain the assumption on which Karl Pearson coefficient of correlation, is based.

Practical Problems:

1. Calculate Pearson‟s coefficient of correlation between advertisement cost and sales from the following data:

Advertisement Cost(„000) Rs.	39	65	62	90	82	75	25	98	36	78
Sales („00,000)	47	53	58	86	62	68	60	91	51	84

(*r = + 0.78*)

2. Compute the coefficient of correlation of the following score of A and B.

A:	5	10	5	11	12	4	3	2	7	1
B:	1	6	2	8	5	1	4	6	5	2

(*r = + 0.58*)

3. Ten competitors in a voice contest are ranked by 3 judges in the following orders:

I Judge	1	6	5	10	3	2	4	9	7	8
II Judge	3	5	8	4	7	10	2	1	6	9
III Judge	6	4	9	8	1	2	3	10	5	7

Use the rank correlation to gauge which pair of judges have the nearest approach to common likings in voice. (I & II = *- 0.212; II & III = - 0.297; I & III = + 0.636*)

4. Calculate the rank correlation coefficient for the following table of marks of students in two subjects:

Major I	80	64	54	49	48	35	32	29	20	18	15	10
Major II	36	38	39	41	27	43	45	52	51	42	40	52

(*r = - 0.685*)

5. The following table gives the score obtained by 11 students in Mathematics and Statistics. Find the rank correlation coefficient.

Mathematics	40	46	54	60	70	80	82	85	85	90	95
Statistics	45	45	50	43	40	75	55	72	65	42	70

(*r = + 0.36*)

6. Calculate the coefficient of concurrent deviation from the data given below:

Year	1976	1977	1978	1979	1980	1981	1982	1983	1984
Supply	160	164	172	182	166	170	178	192	186
Price	292	290	260	234	266	254	230	130	200

(*r = -1*)

*****

UNIT - III

REGRESSION ANALYSIS

Regression Analysis – Regression Vs Correlation, Linear Regression, Regression lines, Standard error of estimates. (18 hrs)

The statistical method which helps us to estimate the unknown value of one variable from the known value of the related variable is called Regression. The dictionary meaning of the word regression is 'return' or 'going back'. In 1877, Sir Francis Galton, first introduced the word 'regression'. The tendency to regression or going back was called by Galton as the 'Line of Regression'. The line describing the average relationship between two variables is known as the line of regression. The regression analysis confined to the study of only two variables at a time is termed as simple regression. The regression analysis for studying more than two variables at a time is known as multiple regressions.

Regression Vs Correlation:

S. No.	Regression	Correlation
1	It is a mathematical measure showing the average relationship between two variables.	It is the relationship between two or more variable, which vary in sympathy with the other in the same or the opposite direction.
2	Here x is a random variable and y is a fixed variable.	Both x and y are random variables.
3	In indicates the cause and effect relationship between the variables.	It finds out the degree of relationship between two variables.
4	It is the prediction of one value, in relationship to the other given value.	It is used for testing and verifying the relation between two variables.
5	It is an absolute figure.	It is a relative measure. The range of relationship lies between $\pm$ 1.
6	Here there is no such nonsense regression	There may be nonsense correlation between two variables.
7	It has wider application, as it studies linear and non-linear relationship between the variables.	It has limited application, because it is confined only to linear relationship between the variables.
8	It is widely used for further mathematical treatment.	It is not very useful for mathematical treatment.

9	It explains that the decrease in one variable is associated with the increase in the other variable.	If the coefficient of correlation is positive, then the two variables are positively correlated and vice - versa.
10	There is a functional relationship between the two variables so that we may identify between the independent and dependent variables.	It is immaterial whether X depends upon Y or Y depends upon X.

Linear Regression:

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable *causes* the other (for example, higher SAT scores do not *cause* higher college grades), but that there is some significant association between the two variables. A scatter plot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatter plot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

A linear regression line has an equation of the form $\boldsymbol{Y = a + bX}$, where $\boldsymbol{X}$ is the explanatory variable and $\boldsymbol{Y}$ is the dependent variable. The slope of the line is $\boldsymbol{b}$, and $\boldsymbol{a}$ is the intercept (the value of $\boldsymbol{y}$ when $\boldsymbol{x} = 0$).

Regression lines:

If we take two variables X and Y we have two regression lines:

i) Regression of X on Y and

ii) Regression of Y on X

The regression line of X on Y gives the most probable value of X for any given value of Y. The regression of Y on X gives the most probable value of Y for any given value of X. There are two regression lines in the case of two variables.

Regression Equations:

The algebraic expressions of the two regression lines are called regression equations.

Regression Equation of X on Y:

$$X_c = a + by$$

To determine the values of 'a' and 'b', the following two normal equations are to be solved simultaneously.

$$\sum X = Na + b\sum Y$$

$$\sum XY = a\sum Y + b\sum Y^2$$

Regression Equation of Y on X:

$$Y_c = a + bx$$

To determine the value of 'a' and 'b', the following two normal equations are to be solved simultaneously.

$$\sum Y = Na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

We can call these equations as normal equations.

Illustration 1: Determine the two regression equations of a straight line which best fits the data.

X	10	12	13	16	17	20	25
Y	10	22	24	27	29	33	37

Solution: Calculation of Regression

X	**X^2**	**Y**	**Y^2**	**XY**
0	100	10	100	100
12	144	22	484	264
13	169	24	576	312
16	256	27	729	432
17	289	29	841	493
20	400	33	1089	660
25	625	37	1369	925
$\sum X = 113$	**$\sum X^2 = 1983$**	**$\sum Y = 182$**	**$\sum Y^2 = 5188$**	**$\sum XY = 3186$**

Regression Equation of Y on X:

The two normal equations are:

$$\sum Y = Na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

Substituting the values,

$N = 7; \quad \sum X = 113; \quad \sum X^2 = 1983; \quad \sum Y = 182; \quad \sum XY = 3186;$

$$7a + 113b = 182 \qquad \ldots(1)$$

$$113a + 1983b = 3186 \qquad \ldots(2)$$

Multiplying (1), by 113,

$$791a + 12769b = 20566 \qquad \ldots(3)$$

Multiplying (2), by 7,

$$791a + 13881b = 22302 \quad \ldots(4)$$

Subtracting (4) from (3)

$$-1112\, b = -1736$$

$$b = \frac{-1736}{-1112} \Rightarrow \mathbf{b = 1.56}$$

Put b = 1.56 in (1) we get

$$7a + 113(1.56)b = 182$$

$$7a + 176.28 = 182 \Rightarrow 7a = 5.72$$

$$a = \frac{5.72}{7} \Rightarrow \mathbf{a = 0.82}$$

The equation of straight line is $Y_c = a + bX$

Put a = 0.82, b = 1.56

∴The equation of the required straight line is $Y_c = 0.82 + 1.56\, X$

This is called regression of y on x

Regression Equation of X on Y:

The two normal equations are:

$$\sum X = Na + b\sum Y$$

$$\sum XY = a\sum Y + b\sum Y^2$$

Substituting the values,

$$N = 7; \quad \sum X = 113; \quad \sum Y^2 = 5188; \quad \sum Y = 182; \quad \sum XY = 3186;$$

$7a + 182b = 113$ …(1)

$182a + 5188b = 3186$ …(2)

Multiplying (1), by 182,

$1274a + 33124b = 20566$ …(3)

Multiplying (2), by 7,

$1274a + 36316b = 22302$ …(4)

Subtracting (4) from (3)

$3192\ b = 1736$

$b = \frac{1736}{3192}$ => **b = 0.54**

Put b = 0.54 in (1) we get

$7a + 182(0.54) = 113$

$7a + 98.28 = 113$ => $7a = 14.72$

$a = \frac{14.72}{7}$ => **a = 2.1**

The equation of straight line is $X_c = a + bY$

Put a = 2.1, b = 0.54

∴The equation of the required straight line is $X_c = 2.1 + 0.54\ Y$

This is called regression of x on y

Deviation taken from Actual Means:

Regression equation of X on Y:

$$X - \bar{X} = r\frac{\sigma x}{\sigma y}(Y - \bar{Y})$$

Where, X = the value of *x* to be estimated for the given y value. $\bar{X}$ = Mean value of X variable. Y = the value of *y* given in the problem; $\bar{Y}$= Mean value of y variables.

$r\frac{\sigma x}{\sigma y} = \frac{\sum xy}{\sum y2}$ = Regression co - efficient of X on Y. x = X - $\bar{X}$; y = Y- $\bar{Y}$

Regression equation of Y on X:

$$Y - \bar{Y} = r\frac{\sigma x}{\sigma y}(X - \bar{X})$$

$r\frac{\sigma x}{\sigma y} = \frac{\sum xy}{\sum y2}$ = Regression co-efficient of Y on X.

Illustration 2: Find regression lines from the following data:

X	3	5	6	8	9	11
Y	2	3	4	6	5	10

And also estimate Y when X is 15.

Solution: Calculation of Regression Equations (by actual mean)

X	**x = X - $\bar{X}$**	**x^2**	**Y**	**y = Y − $\bar{Y}$**	**y^2**	**xy**
3	-4	16	2	-3	9	12
5	-2	4	3	-2	4	4
6	-1	1	4	-1	1	1
8	1	1	6	1	1	1
9	2	4	5	0	0	0
11	4	16	10	5	25	20
∑X = 42	**∑x = 0**	**∑ x^2 = 42**	**∑Y = 30**	**∑y = 0**	**∑ y^2 = 40**	**∑xy = 38**

$$\bar{X} = \frac{\Sigma X}{N} = \frac{42}{6} = 7 \qquad \bar{Y} = \frac{\Sigma Y}{N} = \frac{30}{6} = 5$$

Regression equation of X on Y:

$$X - \bar{X} = r\frac{\sigma x}{\sigma y}(Y - \bar{Y})$$

$$r\frac{\sigma x}{\sigma y} = \frac{\Sigma xy}{\Sigma y2} = \frac{38}{40} = 0.95$$

$$X - 7 = 0.95(Y - 5)$$

$$X - 7 = 0.95Y - 4.75$$

$$X = 0.95Y + 2.25$$

Regression equation of Y on X:

$$Y - \bar{Y} = r\frac{\sigma x}{\sigma y}(X - \bar{X})$$

$$r\frac{\sigma x}{\sigma y} = \frac{\Sigma xy}{\Sigma y2} = \frac{38}{42} = 0.90$$

$$Y - 5 = 0.90(X - 7)$$

$$Y - 5 = 0.90X - 6.30$$

$$Y = 0.90X - 1.30$$

When X is 15, Y will be, $Y = 0.90 \times 15 - 1.30$

$$= 13.5 - 1.30$$

$$Y = 14.8$$

Deviation taken from the Assumed Mean:

Regression equation of X on Y:

$$X - \bar{X} = r\frac{\sigma x}{\sigma y}(Y - \bar{Y})$$

Where, $r\frac{\sigma x}{\sigma y} = \frac{N\Sigma dxdy - (\Sigma dx)(\Sigma dy)}{N\Sigma dy^2 - (\Sigma dy)^2}$

dx = X - A; dy = Y - A; (A = assumed mean)

Regression equation of Y on X :

$$Y - \bar{Y} = r\frac{\sigma y}{\sigma x}(X - \bar{X})$$

$$r\frac{\sigma x}{\sigma y} = \frac{N\Sigma dxdy - (\Sigma dx)(\Sigma dy)}{N\Sigma dx^2 - (\Sigma dx)^2}$$

Illustration: 3. Find regression lines from the following data:

X	40	38	35	42	30
Y	30	35	40	36	29

Also calculate Karl Pearson's coefficient of correlation.

Solution: Calculation of Regression Equations (by assumed mean)

X	**d*x*** **= X – A**	**dx^2**	**Y**	**d*y* = Y – A**	**dy^2**	**d*x*.d*y***
40	5	25	30	0	0	0
38	3	9	35	5	25	15
35	0	0	40	10	100	0
42	7	49	36	6	36	42
30	-5	25	29	-1	1	5
	$\sum$ dx = 10	**$\sum dx^2$ =108**		**$\sum$ dy =1020**	**$\sum dy^2$ =162**	**$\sum$dx.dy = 62**

$\bar{X} = A \pm \frac{\Sigma dx}{N} = 35 \pm \frac{10}{5} = 37$

$\bar{Y} = A \pm \frac{\Sigma dy}{N} = 30 \pm \frac{20}{5} = 34$

Regression equation of X on Y

$X - \bar{X} = bxy\ (Y - \bar{Y})$

$bxy = \frac{N\Sigma dxdy - (\Sigma dx)(\Sigma dy)}{N\Sigma dy^2 - (\Sigma dy)^2}$

$= \frac{5 \text{ x } 62 - (10)(20)}{5 \text{ x } 162 - (20)^2}$

$= \frac{310 - 200}{810 - 400} = \frac{110}{410}$

$bxy = 0.27$

Regression equation of Y on X

$Y - \bar{Y} = byx\ (X - \bar{X})$

$byx = \frac{N\Sigma dxdy - (\Sigma dx)(\Sigma dy)}{N\Sigma dx^2 - (\Sigma dx)^2}$

$= \frac{5 \text{ x } 62 - (10)(20)}{5 \text{ x } 108 - (10)^2}$

$= \frac{310 - 200}{540 - 100} = \frac{110}{440}$

$byx = 0.25$

X – 37 = 0.27(Y – 34)

X – 37 = 0.27Y – 9.18

X = 0.27Y + 27.82

Y – 34 = 0.25(X – 37)

Y – 34 = 0.25X – 9.25

Y = 0.25X + 24.75

$$r = \sqrt{bxy \; x \; byx}$$

$$= \sqrt{2.2 \text{ x } 0.37}$$

$$= \sqrt{0.814}$$

$$r = \mathbf{0.9}$$

Illustration 4: Given the following data, calculate the expected value of Y when X = 12.

	X	**Y**
Arithmetic Mean ($\bar{X}$)	7.6	14.8
Standard Deviation (σ)	3.6	2.5
Coefficient of correlation*(r) = 0.99*		

Solution:

Regression of Y on X

$$Y - \bar{Y} = r \frac{\sigma y}{\sigma x} \; (X - \bar{X})$$

$$Y - 14.8 = 0.99 \text{ x } \frac{2.5}{3.6} (X - 7.6)$$

$$Y - 14.8 = 0.688 \; (X - 7.6)$$

$$Y - 14.8 = 0.688 \; X - 5.23$$

$$Y = 0.688 \; X - 5.23 + 14.8$$

$$Y = 0.688 \; X + 9.57$$

When X = 12 => Y = 0.688 (12) + 9.57 = 17.826

Hence the expected value of Y is 17.83.

Standard Error of Estimate :

We found it necessary to supplement an average for a series with a measure of dispersion or variation to show how representative the average is. The regression equations help us to predict the values of Y for values X or the value of X far values of Y . These are only estimations or predictions; but cannot be treated as a precise value. If we have a wide scatter or variation of the dots about the regression line, then it would be considered a poor representative of the relationship. The more closely the dots cluster around the line, the more representative it is and better the estimate based on the equation for this line. This variation about the line of average relationship can be measured in a manner analogous to the measuring of the variation of the items about an average. Thus, we use here a measure of variation similar to the standard deviation - the standard error of estimates. It is computed as is a standard, being also a square root of the mean of squared deviations. But the deviations here are not the deviations of the items from the arithmetic mean, they are rather the vertical distances of every dot from the line of average relationship.

It measures the scattering of the observations the regression line. It is calculated as follows :

Standard Error of X values from X_c $\left[S_{xy}\right] = \sqrt{\frac{\sum(X-X_c)^2}{N}}$

Standard Error of Y values from Y_c $\left[S_{yx}\right] = \sqrt{\frac{\sum(Y-Y_c)^2}{N}}$

Interpretation of Standard Error of Estimate :

1. Smaller the value, precision of the estimate is better.
2. Larger the value, lesser is correctness of the estimate.
3. If it is zero, there is no variation about the line and both the lines will coincide and correlation will be perfect.

Illustration:

Given the regression equation of Y on X as Y = 3 + 9X for the following data series, calculate (i) Standard error of estimate (ii) Explained variation in Y (iii) unexplained variation in Y.

X	1	2	3	4	5
Y	10	20	30	50	40

Solution :

X	Y	y = Y - $\overline{Y}$	y^2	y_c $[3+9X]$	$[Y-Y_c]$	$[Y-Y_c]^2$
1	10	-20	400	12	-2	4
2	20	-10	100	21	-1	1
3	30	0	0	30	0	0
4	50	20	400	39	11	121
5	40	10	100	48	-8	64
	$\sum Y = 150$	0	$\sum Y^2$ = 1000			$\sum [Y-Y_c]^2$ =190

$$\overline{Y} = \frac{\sum Y}{N} = \frac{150}{5} = 30$$

(i) Standard error of estimate $S_{yx} = \sqrt{\frac{\sum [Y-Y_c]^2}{N}} = \sqrt{\frac{190}{5}} = \sqrt{38} = 6.164$

(ii) Unexplained variation in Y = $\sum [Y-Y_c]^2 = 190$

(iii) Total variation $y^2 = 1000$

Explained variation = Total variation– Unexplained variation

$= y^2 - \sum [Y-Y_c]^2$

= 1000 - 190

= 810.

*****

Important Questions:

I. Fill in the blanks:

1.The variable, we are trying to predict, is called the **Dependent Variable**

2. Both the regression coefficients cannot **Exceed** one.

3. If both the regression coefficients are negative, the correlation coefficient would be **Negative**.

4. The regression analysis measures **Dependence** between X and Y.

5. When one regression coefficient is positive, the other would also be **Positive**.

6. The purpose of regression is to study **Dependence** between variables.

7. The under-root of two **Regression** coefficients gives us the value of correlation coefficient.

II. Tick the correct answer :

1. The greater the value of r:

(a) **The better are estimates, obtain through Regression Analysis**

(b) The worst are the estimates

(c) Really makesvno difference.

2. Where r is zero the regression lines cut each other making an angle of:

(a) 30° (b) 60° (c) 90° (d) **None of these**

3. The further the two regression lines cut each other :

(a) Greater will be degree of correlation

(b) **The Lessor will be Degree of Correlation**

(c) Does not matter.

4. The regression lines cut each other at the point of:

(a) **Average of X and Y**

(b) Average of X only

(c) Average of Y only.

5. When the two regression coincide, then r is :

(a) 0 (b) -1 **(c) +1** (d) 0.5

III. Theoretical Questions:

1. Explain the concepts of correlation and regression, bringing out the inter-relationship between them. Also state their numerical measures.

2. Explain clearly why there are usually two lines of regression. Point out the case when there is one line of regression. Illustrate your answer by diagram.

3. What is meant by 'regression'? Why should there be, in general, two lines of regression for each bivariate distribution? What do you think the coefficient of correlation between the variables would be if the two regression lines cut at right angles, and what if they coincide?

4. Distinguish clearly between 'correlation' and 'regression ' analysis.

5. Explain the concepts of regression and ratio of variation and State their utility in the field of economic enquiries.

6. What is regression? How is this concept useful to business forecasting?

7. Explain the meaning of regression coefficient and the regression lines.

8. What are the properties of the regression coefficients?

9. Explain the concept of regression. Why should there be in general two lines of regression for each bivariate frequency distribution.

Practical Problems:

1. Obtain the equations of lines of regression between the indices.

X:	78	77	85	88	87	82	81	77	76	83	97	93
Y:	84	82	82	85	89	90	88	92	83	89	98	99

(X = 0.79 Y + 13.82; Y = 0.59 X + 39.05)

2. Calculate the two regression equations of X on Y and Y on X from the data given below, taking deviations from actual means of X and Y.

Price (Rs.)	10	12	13	12	16	15
Amount demanded	40	38	43	45	37	43

Estimate the likely demand when the price is Rs. 20.

(X = -0.12 Y + 17.92; Y = -0.25 X + 44.25; Y = 39.25)

3. The correlation coefficient between two variable X and Y is $r = 0.6$. If $\sigma_x = 1.5$ $\sigma_y = 2.0$, = 10 and = 20, find the regression lines of Y on X and X on Y.

(X = 0.45 Y + 1; Y = 0.8 X + 12)

4. By using the following data. Find out the two lines regression and from them compute the Karl Pearson‟s Coefficient of Correlation:

$\sum X = 250$; $\sum Y = 300$; $\sum XY = 7900$; $\sum X^2 = 6500$; $\sum Y^2 = 10{,}000$; N -= 10.

(*r = - 0.8*)

5. Given the following data, estimate the marks in mathematics obtained by a student who has scored 60 marks in statistics.

Mean marks of mathematics	80
Mean marks of statistics	50
S.D marks in mathematics	15
S.D marks in statistics	10
Coefficient of correlation	0.4

(X = 0.6 Y + 50; X = 86.)

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. (18 hrs)

Time Series - Meaning and Significance:

An arrangement of statistical data in accordance with time of occurrence or in chronological order is called a time series. Thus when we observe numerical data at regular intervals of time the set of observation is known as time series. The regular intervals may be an hour, a day, a week a month, a year, a decade etc.,

Time series analysis can be useful to see how a given asset, security, or economic variable changes over time. It can also be used to examine how the changes associated with the chosen data point compare to shifts in other variables over the same time period.

Utility of Time Series:

Time series analysis is useful indifferent fields like economics, science, research work, etc because of the following reasons.

- It helps in understanding the past behavior.
- It helps in planning and forecasting the future operations.
- It facilitates comparison between data of one period with another period.
- It is useful not only to economists but also to the businessman.
- It helps in evaluating current accomplishments.
- It is useful for forecasting the trade cycles.

Components of Time Series.

The following are the components of time series.

i. Secular trend
ii. Seasonal variation

iii. Cyclical variations

iv. Irregular variation.

i) Secular Trend :

A secular or long-term trend refers to the movement of a series reflecting continuous growth or decline over a long period of time. There are many types of trend. Some trend rise upward and some trend fall downward.

The following are the types of trend.

a.Linear (or) Straight Line Trend.

b.Non-Linear (or) Curvilinear Trend

If the values of time series are plotted on a graph and if it forms a straight line then it is called Linear (or) Straight Line Trend.

If the values of time series are plotted on a graph an if it forms a curve then it is called Non-linear (or) Curvilinear Trend.

ii) **Seasonal variation:**

Seasonal variations are those periodic movements in business activities within the year, recurring periodically year after year. Generally, seasons variations appear at weekly, monthly or quarterly intervals.

Causes:

Seasonal variation may occur due to climate, weather conditions, customs, habits and traditions.

Uses:

Seasonal variation analysis is used to formulate correct policy decisions and of planning future operations. It is useful to businessmen, producers, agriculturist etc., to study effects of seasonal variations and to isolate them from the trend.

Cyclical variation:

According to Linecoin L. Chou, "Up and down movements are different from seasonal fluctuations, in that they extend over longer period of time – usually two or more years".

Business time series is influenced by the wave-like change of prosperity and depression. This up and down movement is known as cyclical variation. Cyclical variation analysis is helpful to businessmen for stabilizing the level of business activities and also useful to the economist for formulating suitable policies.

Irregular variation:

Irregular variations refer to such variations in business activities which do not repeat in a definite pattern. They are also called 'erratic' accidental or random variations which are generally non-recurring and unpredictable. Irregular variations of time series are either random or caused by some sporadic forces such as war, flood, revolution, etc. It is usually a short-term one, but it will affect all the components of time series.

Measurement of Secular Trend:

The following are the four methods which can be used for determining the trend.

1. **The free hand or Graphic method:**

   In this method we must plot the original data on the graph. Draw a smooth curve carefully which will show the direction of the trend.

2. **Semi –Averages Method:**

   In this method the original data are divided into two equal parts and averages are calculated for both the parts. These averages are called Semi –Averages. Trend line is drawn with the help of this average.

3. **Moving Average Method :**

In this method, the average value for a number of years or months or weeks is taken into an account and placing it at the centre of the time span (period of moving average) and it is the normal or trend value for the middle period.

4. Method of Least Squares :

It is a mathematical as well as an analytical method. Under this method a straight line trend can be fitted to the given time series of data.

Illustration 1: Fit a trend line for the following data by the free - hand method.

Year :	2000	2001	2002	2003	2004	2005	2006
Sales in units: (in '000s)	65	95	85	115	110	120	130

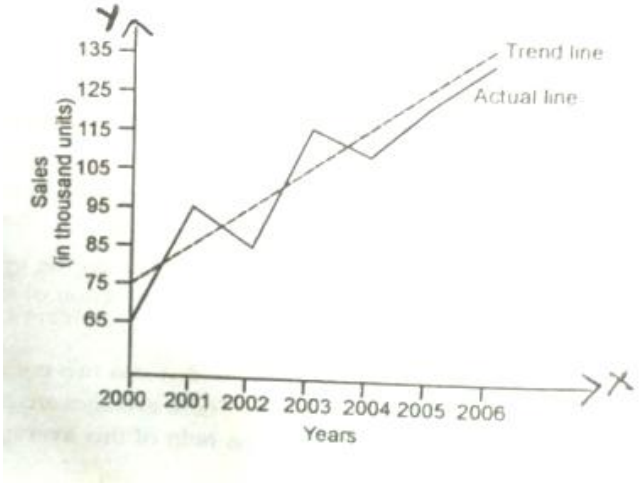


Illustration 2: Draw a trend line by the method of semi averages

Year : `	2001	2002	2003	2004	2005	2006
Sales in units : (in '000s)	60	77	82	120	116	130

Solution: Calculation of trend line using semi average method

Year	**Sales in units : (in '000s)**	
2001	60	
2002	77	219/3 = 73
2003	82	
2004	120	
2005	116	366/3 = 122
2006	130	

Note: If the even number of periods is given we can divide two equal parts. If the period is given in odd number of years, the value of the middle year is omitted.

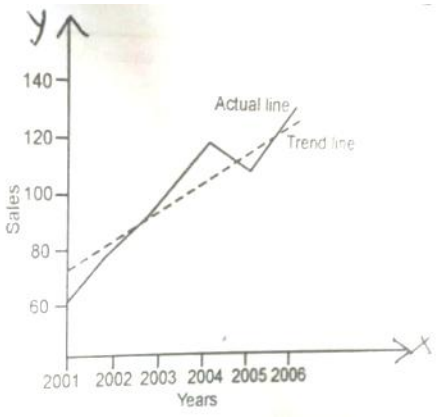
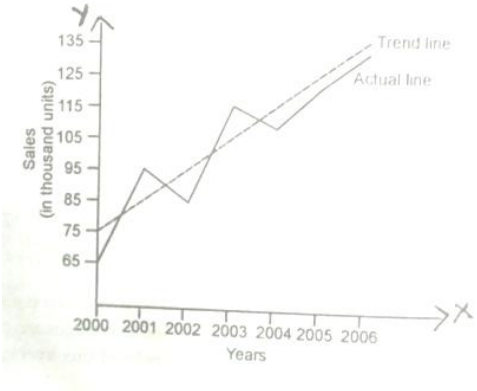


Illustration 3: Draw a trend line by the method of semi-averages

Year :	2000	2001	2002	2003	2004	2005	2006
Sales in units : (in '000s)	100	110	120	115	130	135	140

Solution: Calculation of Trend Line using Semi Average Method

Year	**Sales in units : (in '000s)**
2000	100
2001	330/3 = 110
2002	120
2003	115 Left out the middle item
2004	130
2005	405/3 = 135
2006	140



3. Moving average method:

In this method the average value for a number of years or months or weeks is taken into account. It is placed at the centre of the time – span. It is the normal or trend value for the middle period.

Calculation of moving averages:

The formula for calculating 3 yearly moving averages is

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

The formula for 4 yearly moving averages is

$$\frac{a+b+c+d}{4}, \frac{b+c+d+e}{4}, \frac{c+d+e+f}{4}$$

The formula for 5 yearly moving averages is

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

Steps to calculate 3 yearly moving averages:

1. Compute the value of first three years and place the three year total against the middle year.
2. Leave the first year's value and compute the value of next three years and place the total against the middle year ie., against 3rd year.
3. This process should be continued until the last year's value is taken for calculation.
4. The three yearly total values are divided by 3 and the value placed in the next column. This value is called the Trend value of moving average.

Steps for calculating of 4yearly moving average (or) Even period of moving average:

1. Computer the value of first four years and place the total value in between second and the third year.

2. Leave the first year's value and compute the value of next four years and place the total in between third and fourth year.

3. This process must be continued until the last year's value is taken for calculation.

4. Computer the first two four year totals and place the value against the middle of two four years totals.

5. Leave the first four years total and compute the next two four year's totals and place it against the middle of next four year.

6. This method must be continued until all the four year totals.

7. The two four yearly totals is divided by 8 and put the value in the next column. This value is called Trend value.

Illustration 4: Find the 3-yearly moving average from the following time series data.

Year :	1998	1999	2000	2001	2002	2003	2004	2005
Sales in units : (in tones)	30.1	45.4	39.3	41.4	42.2	46.4	46.6	49.2

Solution: Calculation of 3 yearly moving average method

Year	**Sales (in tons)**	**3 yearly moving total**	**3 yearly moving value**
1998	30.1	-	-
1999	45.4	114.8	38.27
2000	39.3	126.1	42.03
2001	41.4	122.9	40.97
2002	42.2	130.0	43.33
2003	46.4	135.2	45.07
2004	46.6	142.2	47.40
2005	49.2	-	-

Illustration 5: Calculate the trend value by using three yearly moving averages of the following data.

Year :	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Production (in'000 tons)	21	22	23	25	24	22	25	26	27	26

Solution:

Calculation of trend value by using 3 yearly moving averages

Year	Production (in '000 tons)	3 yearly moving total	3 yearly moving Average (Trend value)
1990	21	-	-
1991	22	66	22.00
1992	23	70	23.33
1993	25	72	24.00
1994	24	71	23.67
1995	22	71	23.67
1996	25	73	24.33
1997	26	78	26.00
1998	27	79	26.33
1999	26	-	-

Illustration 6: Calculate the 5 yearly moving averages from the following data:

Year :	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
No. of Students:	705	685	703	687	705	689	715	685	725	730

Solution: Calculation of 5 yearly moving average method

Year	No. of students	5 yearly moving total	Moving Value
1997	705		
1998	683		
1999	703	3485	697.0
2000	687	3469	693.8
2001	705	3499	699.8
2002	689	3481	696.2
2003	715	3519	703.8
2004	685	3544	708.8
2005	725		
2006	730		

Illustration 7: Calculate the four – yearly moving average for the following data

Year :	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
Production : (in '000 tons)	464	515	518	467	502	540	557	571	586	612

Solution:

Calculation of 4 yearly moving average method

Year	Production	4 Yearly Moving	Combined Total	Moving Average
1994	464			
1995	515			
		1964		
1996	518		3966	3966/8 = 495.75
		2002		
1997	467		4029	4029/8 = 503.63
		2027		
1998	502		4093	4093/8 = 511.63
		2066		
1999	540		4236	4236/8 = 529.50
		2170		
2000	557		4424	4424/8 = 553.00
		2254		
2001	571		4580	4580/8 = 572.50
		2326		
2002	586			
2003	612			

4. Methods of Least Square:

A Straight line trend can be fitted from the time series of given data. The trend line is called the line of best fit.

The sum of deviations of the actual values of Y and the trend value (Yc) is zero. And the sum of squares of deviations of actual value of Y and the trend value (Yc) is the least.

i.e. $(Y - Yc) = 0$ and $(Y - Yc)^2$ = least value

This method is called the method of least square (or) the line of best fit. The method of least square can be used in the liner and non-linear trend

$$Y_c = a + bx;$$

Y_c = required trend value; X = unit of time; a and be are constants

The constants a and b are calculated from the following two normal equations.

$$\mathbf{\Sigma Y = Na + b \Sigma X} \qquad \text{\_\_\_\_\_ } \textit{(1)}$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \qquad \text{\_\_\_\_\_(2)}$$

Where N = Number of years or months

When x = 0 (middle year taken as origin)

$$\Sigma X = 0$$

$$b\Sigma X = 0 \text{ and } a\Sigma X = 0$$

From (1), $\Sigma Y = Na + 0 \rightarrow a = \frac{\Sigma Y}{N}$ or $\bar{Y}$

From (2), $\Sigma XY = 0 + b\Sigma X^2 \rightarrow b = \frac{\Sigma XY}{\Sigma X^2}$

$$a = \frac{\Sigma Y}{N} \text{ and } b = \frac{\Sigma XY}{\Sigma Y^2}$$

Illustration 8: Compute the trend from the following by the method of least squares:

Year :	2000	2001	2002	2003	2004
Population (in Lakhs)	830	920	710	900	1690

Solution: Computation of trend values

Year	**Production (in lakhs) (Y)**	**Deviation from 2002 (X)**	**XY**	**X^2**
2000	830	2000-2002=-2	-1660	4
2001	920	2001-2002=-1	-920	1
2002	710	2002-2002=0	0	0
2003	900	2003-2002=1	900	1
2004	1690	2004-2002=2	3380	4
	$\Sigma y = 5050$	**$\Sigma x = 0$**	**$\Sigma xy = 1700$**	**$\Sigma x^2 = 10$**

Since $\sum x = 0$, $a = \frac{\Sigma y}{N} = 5050/5 = 1010$

$$b = \frac{\Sigma XY}{\Sigma X^2} = 1700 / 10 = 170$$

$$Y_c = a + bx = 1010 + 170\ x$$

$Y_c = 1010 + 170\ x$

When X = -2, $Y_{2000} = 1010 + 170(-2) = 1010 – 340 = 670$

When X = -1, $Y_{2001} = 1010 + 170\ (-1) = 1010 - 170 = 840$

When X = 0, $Y_{2002} = 1010+170(0) = 1010 – 0 = 1010$

When x = 1, $Y_{2003} = 1010 + 170(1) = 1010 + 170 = 1180$

When x = 2, $Y_{2004} = 1010 + 170\ (2) = 1010 + 340 = 1350$

Illustration 9: Fit a straight line trend by the method of least squares from the following data and estimate the earnings for the year 2008:

Year :	1999	2000	2001	2002	2003	2004	2005	2006
Earnings (in Lakhs)	38	40	65	72	69	60	87	95

Solution: Computation of trend values

Year	**Earnings (Rs. in lakhs) (Y)**	**Deviation from 2002.5 (X)**	**XY**	**X^2**
1999	38	1999 - 2002.5 = -3.5	-133.0	12.25
2000	40	2000 - 2002.5 = -2.5	-100.0	6.25
2001	65	2001 - 2002.5 = -1.5	-97.5	2.25
2002	72	2002 - 2002.5 = -0.5	-36.0	0.25
2003	69	2003 - 2002.5 = 0.5	34.5	0.25
2004	60	2004 - 2002.5 = 1.5	90.0	2.25
2005	87	2005 - 2002.5 = 2.5	217.5	6.25
2006	95	2006 - 2002.5 = 3.5	332.5	12.25
N = 8	**Σy = 526**	**Σx = 0**	**Σxy = 308**	**$\Sigma x^2 = 42$**

$$Y_c = a + bx$$

$$a = \frac{\Sigma y}{N} = 526/8 = 65.75$$

$$b = \frac{\Sigma XY}{\Sigma X^2} = 308 / 42 = 7.33$$

That is $Y_c = 65.75 + 7.33\ x$

In the year 2008, the value of x is 5.5 that is (2008 – 2002.5)

∴ The earnings for the year 2008 will be

$$Y_{2008} = 65.75 + 7.33(5.5) = 65.75 + 40.31 = 106.06$$

Parabolic Curve:

Parabolas are non-linear, they form into smooth curves. The shape of these curves depends upon the value of the constants a, b, c, etc. The general form of the equation of power series is $Y = a + bx + cx^2 + dx^3$.... The equation of this type does not represent a curve of strictly parabolic type but in common usage, the term parabolic curve is used to indicate curves obtained by equations of the above type. The trend equation in this case is:

$$Y_C = a + bX + cX^2$$

where,

a is the trend value at the time origin,

b is the slope at the origin, and

c establishes whether the curve is up or down and how much.

The values of a, b and c can be determined by solving the following three normal equations simultaneously :

$$\sum Y = Na + bX + cX^2 \qquad ...(1)$$

$$\sum XY = aX + bX^2 + cX^3 \qquad ...(2)$$

$$\sum X^2Y = aX^2 + bX^3 + cX^4 \qquad ...(3)$$

In solving the above second-degree equations much time and labour can be saved by taking the time origin in the middle of the series so that X=0. But if X=0, then the sum of any odd power of X, such as X3 is also zero. Therefore the three equations are reduced to;

$$\sum Y = Na + c\sum X^2 \qquad .....(1)$$

$$\sum XY = b\sum X^2 \qquad ..... (2)$$

$$\sum X^2Y = a\sum X^2 + c\sum X^4 \qquad ..... (3)$$

Solving equations (1) and (3) we obtain the values of a and c and the value of b is obtained by solving equations (2):

$$a = \frac{\sum Y - c\sum X2}{N}$$

$$b = \frac{\sum XY}{\sum X2}$$

$$c = \frac{\sum X2Y - a\sum X2}{\sum X4}$$

Illustration: Fit a parabola of the second degree to the data given below:

Year	2003	2004	2005	2006	2007
Sales ('000)	16	18	19	20	24

Solution : **Construction of Trend Value**

Year	**Y**	**Year - 2005 X**	**X^2**	**X^3**	**X^4**	**XY**	**X^2Y**	**Trend value**
2003	16	-2	4	-8	16	-32	64	16. 08
2004	18	-1	1	-1	1	-18	18	17. 46
2005	19	0	0	0	0	0	0	19. 12
2006	20	1	1	1	1	20	20	21. 06
2007	24	2	4	8	16	48	96	23. 28

N = 5	$\sum Y = 97$	$\sum X = 0$	$\sum X^2 = 10$	$\sum X^3 = 0$	$\sum X^4 = 34$	$\sum XY = 18$	$\sum X^2Y = 198$	

$$\mathbf{Y_C = a + bX + cX^2}$$

Since $\sum X = 0$

$$a = \frac{\sum Y - c\sum X2}{N}$$

$$a = \frac{97 - c(10)}{5} = \frac{97 - 10c}{5}$$

$$5a = 97 - 10c$$

$$5a + 10c = 97$$

$$b = \frac{\sum XY}{\sum x^2} = \frac{18}{10} = 1.8$$

$$c = \frac{\sum X^{2Y - a\sum x^2}}{\sum x^4} = \frac{198 - a(10)}{34}$$

$$= \frac{198 - 10a}{34} = 34c = 198 - 10a$$

$$34c + 10a = 198$$

Multiply equation (1) by 2 and subtract it from equation (2)

$$10a + 34c = 198$$

$$\underline{10a + 20c = 194}$$

$$\underline{14c = 4}$$

$$c = \frac{4}{14} = 0.29.$$

Substitute the value of c in equation (1)

$$5a + 10(0.29) = 97$$

$$5a = 97 - 2.9 = 94.1$$

$$a = \frac{94.1}{5} = 18.82$$

value of a = 18.82; b = 1.8 ; c= 0.29

Substitute the value in the equation $Y_C = a + bX + c_{x^2}$

$$y_{2003} = 18.82 + 1.8(-2) + 0.29(-2)^2$$

$$= 18.82 - 3.6 + 1.16$$

$$= 16.38$$

$$y_{2004} = 18.82 + 1.8(-1) + 0.29(-1)^2$$

$$= 18.82 - 1.8 + 0.29 = 17.31$$

$$y_{2005} = 18.82 + 1.8(0) + 0.29(-0)^2 = 18.82$$

$$y_{2006} = 18.82 + 1.8(1) + 0.29(1)^2 = 20.91$$

$$y_{2007} = 18.82 + 1.8(2) + 0.29(2)^2$$

$$= 18.82 + 3.6 + 1.16$$

$$= 23.58$$

Logarithmic Trend :

The straight-line arithmetic trend is used where the time series is found to be increasing and decreasing by equal absolute amounts in each time period. The logarithmic straight line is used as an expression of the secular movement when the series is increasing or decreasing by a constant percentage rather than a constant absolute amount. Such tendency is found in many economic and business data.

The equation of this curve is :

$$\text{Log } Y = \text{Log } a + X \text{ Log } b \text{ or } Y = ab^x$$

Normal equations are :

$$\sum (\text{Log } Y) = N \text{ Log } a + \sum (X) \text{ Log } b$$

$$\sum (X.\text{Log } Y) = \text{Log } a \sum (X) + \sum (X^2) \text{ Log } b$$

If middle year is taken as origin, then normal equations are:

$$\sum \text{Log } Y = N \text{ Log } a \qquad \therefore \text{ Log } a = \frac{\sum \text{Log } Y}{N}$$

$$\sum (X.\text{LogY}) = \sum (X^2) \text{ Log } b \qquad \therefore \text{ Log } b = \frac{\sum (X \text{ LogY})}{\sum (X2)}$$

They are to be converted into actual numbers to arrive at natural numbers. Plotted on a semi- logarithmic graphs, the curve will be straight line.

Steps :

1. Find the time deviation of each year from the middle year (X)
2. Square up these deviations (X^2)
3. Convert the original data into logarithms (Log Y) and multiply by X (X Log Y).
4. Find out Log a = $\frac{\sum \text{Log } Y}{N}$, and keep the value so obtained before the middle year.
5. Find out the value of growth by Log b = $\frac{\sum (X \text{ LogY})}{\sum (X2)}$ and value to be multiplied with each deviation and after adding or subtracting as is required, keep the value opposite to that year (Log Y).
6. Convert Log Y into natural numbers.

Illustration:

The sale of a company in thousands of rupees for the year 2001 to 2007 are in given below.

Year	2001	2002	2003	2004	2005	2006	2007
Sales	32	47	65	92	132	190	275

Estimate sales figures for the year 2008 using an equation of the form $Y= ab^X$, where X= years and Y= sales.

Solution:

Computation of Logarithmic Straight Line

Year	**Sales**	**X**	**Log Y**	**X^2**	**X Log Y**	**Log Y**	**A.L of Log Y**
2001	82	-3	1.5051	9	-4.5153	1.5084	32.24
2002	47	-2	1.6721	4	-3.3442	1.6624	45.96
2003	65	-1	1.8129	1	-1.8129	1.8164	65.52
2004	92	0	1.9638	0	0	1.9704	93.42
2005	132	1	2.1206	1	2.1206	2.1244	133.10
2006	190	2	2.2788	4	4.5576	2.2784	189.00
2007	275	3	2.4393	9	7.3179	2.4324	270.60
Total	833		$\sum LogY$ 13.7926	$\sum x^2 = 28$	$\sum X.LogY =$ 4.3237		

$$\text{Log a} = \frac{\sum LogY}{N} = \frac{13.7926}{7} = 1.9704$$

$$\text{Log b} = \frac{\sum (X.LogY)}{\sum (_{X^2})} = \frac{4.3237}{28} = 0.154$$

The trend equation is Log Y = 1.9704 + 0.154 X

The value of Log Y, when X = -3 = 1.9704 + 0.154 (-3)

= 1.9704 - 0462

= 1.5084

The value of Log Y, when X = -2 = 1.9704 + 0.154(-2)

= 1.9704 + 0.308

= 1.6624

The value of Log Y, when X = -1 = 1.9704 + 0.154(-1)

= 1.9704 - 0.154

= 1.8164

The value of Log Y, when X = 0 = 1.9704 + 0.154(0)

=1.9704

The value of Log Y, when X = 1 = 1.9704 + 0.154(1)

= 1.9704 + 0.154

=2.1244

The value of Log Y, when X = 2 = 1.9704 + 0.154(2)

= 1.9704 + 0.308

= 2.2784

The value of Log Y, when X=3 =1.9704 + 0.154(3)

= 1.9704 + 0.462

= 2.4324

The value of Log Y, when X = 4 i.e. 2003

= 1.9704 + 0.154(4)

= 1.9704+0.616

= 2.5864

Antilog of 2.5864 = 385.9

Forecasted sales for 2008 = **Rs. 3,85,900/-**

Important Questions:

I. Fill in the blanks :

1. An overall tendency of rise or fall in a time series is called the **Secular Trend**.

2. The one that is useful for forecasting in the short-term is the **Seasonal** component.

3. A time series consists of data arranged **Chronologically**

4. The additive model of a time series is expressed as **Y=T+S+C+I**

5. The equation of the Gompertz curve is of the form **Y= KA BX**

6. The line obtained by method of least squares is known as the line of **Best Fit**

II. Tick the correct answer:

1. The most widely used of measuring seasonal variation is:

(a) **Ratio to Moving Average Method**

(b) Ratio to Trend method

(c) Link Relative method.

2. Trend refers to a long-term tendency to:

(a) Decrease only

(b) **Either Increase or Degrease**

(c) Increase only

3. Seasonal variations repeat during a period of:

(a) **One Year** (b) Five year (c) Seven years

4. Cyclic fluctuations are caused by:

(a) Strikes and Lockouts (b) Floods (c) Wars (d) **None of These**

5. The trend is linear if:

(a) **The Growth Rate is Constant**

(b) Rate of growth is positive

(c) Growth is not constant.

6. The most important factors causing seasonal variations are:

(a) Growth of Population

(b) **Weather and Social Customs**

(c) Depression in business

7. If the trend is absent, seasonal indices are known by:

(a) Ratio to trend method

(b) Ratio to Moving average method

(C) Simple Average Method

8. Which of the following components is used for a short-term forecast?

(a) Cyclical (b) Trend (c) Seasonal **(d) None of these**

9. In time series analysis both trends and seasonal variations are studied because they:

(a) Describe past patterns

(b) Allow projections into the future

(c) Allow the Elimination of the Component from the Series.

III. Theoretical Questions:

1. What is a time Series? Distinguish between seasonal, cyclical and random fluctuations. Describe any method of eliminating their influence.

2. Distinguish between seasonal variation, cyclical variation and secular trend.

3. Give an account of the common components of time series data. Explain anyone method of obtaining long-term trend.

4. Explain the importance of time series Analysis in business forecasting.

5. What is a " Time series"? What is the main object of constructing a time Series? Explain fully the components of a time series.

6. Distinguish between trend, seasonal variations and cyclical fluctuations in a time Series. How can trend be isolated from fluctuations?

7. What is ' moving average' ? What are its uses in time series?

8. What do you mean by seasonal variation? Explain with a few examples the utility of such a study.

9. What are the component of time series? Show how a time series is built up from these components using an example?

10. What is meant by trend? How would you fit a straight line trend by the method of least squares?

11. Explain the utility of time series to a businessman and an economist. Also state the different components of fluctuations in a time Series.

12. What are the methods of isolating the trend components in a time Series? Compare their merits

13. Explain how you would deseasonalise a time series and state the assumptions you would be making.

14. What is secular trend? Explain any one method of measuring the trend of a time series.

15. Critically examine the different methods of measuring trend, pointing out their merits and demerits.

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 (18 hrs)

Meaning and Significance:

An index number is a number which is used to measure the level of a certain phenomenon as compared to the level of the same phenomenon at some standard period. It is a statistical device for comparing the general level of magnitude of a group of related variables in two or more situations. It is a number which indicates the change in magnitude.

The primary purposes of an index number are to provide a value useful for comparing magnitudes of aggregates of related variables to each other, and to measure the changes in these magnitudes over time.

Problems in the Construction of Index Numbers:

The following are some guidelines to be remembered when we construct index numbers

1. Purpose or Object :

The statistician must clearly determine the purpose for which the index numbers are to be constructed, because there is no all purpose index numbers. Every index number has got its own uses and limitations. For example, if we want to study the changes in the value of money, then we have to construct index numbers of wholesale price. If we want to study the changes in the cost of living of workers of any place, the cost of living index numbers of the workers must be constructed. Cost of living index numbers of workers in an industrial area and those of the workers of an agricultural area are different in respect of requirements. Therefore, it is very essential to define clearly the purposes of the index numbers and that too beforehand.

2. Selection of Base :

The base period of an index number is very important as it is used for the construction of index numbers. Every index number must have a base. One cannot say whether the price level has increased or decreased, unless one compares the price level of the current year with the price level of the previous year.

Thus when selecting a base period, the year must be recent and normal. A normal year is one which is free from economic and natural disturbances, widespread failure of rains, earthquakes, war, strikes, production crisis, etc. If such abnormal years are considered, then the index number will be a misleading one.

Therefore, the year to be selected as base year must be, normal year or a typical year and a recent year. In the fast changing world of today, the more recent the year the more representative will it be. The base may be of the following type.

(a) Fixed base :

The name reveals that the base year is a fixed one. The prices of a particular year, selected as a base period are treated as equal to 100. The changes in the prices of subsequent years are shown as the percentage of the base year.

(b) Average base :

Sometimes it is difficult to select an year as base through normality. Under such a critical position, the average of several years is considered better, as abnormalities can be reduced to great extent.

(c) Chain base :

In fixed base method, the base year once selected, remains fixed and all index numbers are based on the same base year. In this method, there is no fixed base year. It changes from year to year. When a comparison is desired from year to year, a system of chain base is used. It is the previous year that is taken as the base for the current year; and the change is calculated as a percentage of that year. For instance, for 2006, the base year is 2005; for 2005, the base year is

2004; for 2004, the base year is 2003 and so on. There is no question of normal year. Long period comparison cannot be made.

3. Selection of commodities :

(Selection of Regiman) If we study the price changes of one commodity, we have to include only one item. For instance, if we study the changes in production of cloth, then we may include the production of mill cloth, power loom cloth, handloom cloth, silk, khadi, etc.,and there is no problem. Another example, say index of retail price; we cannot include all commodities sold in retail. We include only the important commodities which are representative of tastes, habits and customers of the people. For the purpose of finding the cost of living index number, of low income group, we have to select only those items or commodities, which are mostly consumed by that group.

4. Source of data:

The price relating to the thing to be measured must be collected. If we want to study the changes in industrial production, we must collect the prices relating to the production of various goods of factories. The price may be collected from reliable sources. The prices of commodities are the raw materials for the construction of index numbers. The prices may be collected from the public sources or from standard commercial magazines. The collection of data price must be representative, comparable and accurate.

5. Selection of Averages:

One can ues any average. But in practice, the arithmetic average is used, because it is easy for computation; geometric mean and harmonic mean are difficult to calculate. But geometric mean is preferred because of the following characteristics (a) Geometric mean is the best measure and (b) It gives less weight to bigger items and more weight to smaller items.

6. Weighting:

All commodities are not equally important. The main purpose of an index number of prices, is to ascertain the changes in the price level. In case of simple average, all commodities will have

equal importance. But on actual practice, different groups of people will have different preferences on different commodities. For instance, when the price of rice or wheat is double and the price of sugar is halved, the people suffer much, because the price of rice which is essential has been doubled; but as regards the sugar it is not so important as rice or wheat. Therefore, to stress the importance, the system of weighting is adopted. John Griffin observes, "In simple terms weighting is designed to give component series an importance in proper relations to their real significance ".

Methods of constructing of index numbers

The various methods of construction of index numbers are as follows:

I. **Unweighted Index Numbers.**

1. Simple Aggregative method

This is the simplest method of constructing the index numbers. The prices of the different commodities of the current year are added and the total is divided by the sum of the prices of base year commodity and multiplied by 100:

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

Illustration 1:

Construct an index number for 2016 taking 2015 as base from the following data.

Commodity	Price in 2015 (Rs.)	Price in 2016 (Rs.)
A	50	60
B	40	80
C	70	110
D	90	70
E	50	40

Solution:

Calculation of index number

Commodity	Price in 2015 (Rs.)	Price in 2016 (Rs.)
A	50	60
B	40	80
C	70	110
D	90	70
E	50	40
	$\sum P_0 = 300$	**$\sum P_1 = 360$**

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

$= \frac{360}{300} \times 100$

$= 120$

This means that as compared to 2015, in 2016 there is a net increase in the prices of commodities to the extent of 20%.

2. Simple Average of Price relative method:

In this method, the price relative of each item is calculated separately and then averaged.

Arithmetic Mean of Price Relatives: $\mathbf{P_{01}} = \frac{\boldsymbol{\Sigma P}}{\boldsymbol{N}}$

Geometric Mean of Price Relatives: $\mathbf{P_{01}}$ **= antilog** $\frac{\sum \boldsymbol{logP}}{\boldsymbol{N}}$

Where,

$P = \frac{P_1}{P_0} \times 100$;

N = Number of items.

Illustration 2: Compute a price index for the following by using both arithmetic mean and geometric mean:

Commodity	A	B	C	D	E	F
Price in 2014 (Rs.)	20	30	10	25	40	50
Price in 2015 (Rs.)	25	30	15	35	45	55

Solution: Calculation for Price Index

Commodity	**Price in 2014 P_0**	**Price in 2015 P_0**	**Price Relative P $= \frac{P_1}{P_0}$ x 100**	**Log P**
A	20	25	125	2.0960
B	30	30	100	2.0000
C	10	15	150	2.1761
D	25	35	140	2.1461
E	40	45	112.5	2.0511
F	50	55	110	2.0414
N = 6			**∑P = 737.5**	**∑log P =12.5116**

Arithmetic Mean of Price Relatives: $\mathbf{P_{01}} = \frac{\Sigma P}{N}$

$= \frac{737.5}{6}$

$= 122.92$

Geometric Mean of Price Relatives: $\mathbf{P_{01}}$ **= Antilog** $\frac{\sum logP}{N}$

$=$Antilog $\frac{12.5116}{6}$

$=$ Antilog 2.0853

$= 121.7.$

II. Weighted Index Number

Weighted Aggregative Index Numbers: Under this method weights are assigned to various items and a number of formulae have been used. The following are some of the methods, generally used.

A. Laspeyre's method: In this method, the base year quantities are taken as weights:

$$P_{01\,(La)} = \frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \mathbf{x\ 100}$$

B. Paasche's method: In this method, current year quantities are taken as weights:

$$P_{01\,(Pa)} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \mathbf{x\ 100}$$

C. Bowley Dorfish method: This is an index number got by the arithmetic mean of Laspeyre's and Paasche's methods. This method takes into account both the current and the base periods.

$$P_{01\,(B)} = \frac{\mathbf{L+P}}{\mathbf{2}}$$

Where, L = Laspeyre's method; P = Paasche's method

D. Fisher's ideal Method: Fisher's price index number is given by the geometric mean of Laspeyre's and Paasche's formula;

$$P_{01\,(F)} = \sqrt{\mathbf{L\ x\ P}}$$

E. Marshal Edge worth Method: Under this method, the arithmetic mean of base year and current year quantities are taken as weights ($w = \frac{q_0 + q_1}{2}$)

$$P_{01\,(Ma)} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \text{ x } 100$$

F. Kelly's method: Kelly's method number uses quantities of some period (which is neither the base period nor the current period) as weights. This weight is kept constant for all periods. If we denote it by q, then Kelly's method number is given by

$$P_{01\,(K)} = \frac{\Sigma P_1 q}{\Sigma P_0 q} \mathbf{x\ 100}$$

G. Walsch's method: Under this method, the geometric mean of the base year and current year quantities as weights.

$$P_{01\,(Wa)} = \frac{\Sigma P_1 \sqrt{q_0 q_1}}{\Sigma P_0 \sqrt{q_0 q_1}} \text{ x } 100$$

Illustration 3: Calculate index number from the following data through a) Laspeyre's method b) Paasche's method c) Bowley's method d) Fisher's ideal formula method and e) Marshall Edge worth method:

	Base Year		Current Year	
	Kilo	**Rate (Rs.)**	**Kilo**	**Rate (Rs.)**
Bread	10	3	8	3.25
Meat	20	15	15	20
Tea	2	25	3	23

Solution: Calculation of Index Numbers

	Base Year		Current Year		p_1q_0	p_0q_0	p_1q_1	p_0q_1
	Kilo	**Rate**	**Kilo**	**Rate**				
	q_0	p_0	q_1	p_1				
Bread	10	3	8	3.25	32.50	30	26	24
Meat	20	15	15	20	400.00	300	300	225
Tea	2	25	3	23	46.00	50	69	75
Total					$\sum p_1q_0$ = **478.50**	$\sum p_0q_0$ = **380**	$\sum p_1q_1$ = **395**	$\sum p_0q_1$ = **324**

a)Laspeyre's method: $P_{01\,(La)} = \frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \mathbf{x\ 100} = \frac{478.50}{380} \text{ x } 100 = \mathbf{125.9}$

b)Paasche's method: $P_{01\,(Pa)} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \mathbf{x\ 100} = \frac{395}{324} \text{ x } 100 = \mathbf{121.9}$

c)Bowley Dorfish method: $P_{01\,(B)} = \frac{\mathbf{L+P}}{\mathbf{2}} = \frac{125.9+121.9}{2} = 123.9$

d)Fisher's ideal Method: $P_{01\,(F)} = \sqrt{\mathbf{L \text{ x } P}} = \sqrt{125.9 \text{ x } 121.9}$

$= \sqrt{15347.21} = \mathbf{123.9}$

e)Marshal Edge worth Method: $P_{01\,(Ma)} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \text{ x } 100$

$$= \frac{478.5}{380} + \frac{395}{324} \text{ x } 100 = \frac{873.5}{704} \text{ x } 100 \text{ = } 1.24 \text{ x } 100 = \textbf{124.}$$

2. Weighted average of price relative.

$$P_{01} = \frac{\sum PV}{\sum V} \text{ ; } P_{01} = \text{Antilog } \frac{\sum V \log P}{\sum V}$$

Illustration 4: Compute a price index for the following by using both arithmetic mean and geometric mean:

Item	Price in 1976	Price in 1977	Quantity in 1976
Wheat	2	2.50	40 kg
Sugar	3	3.25	20 kg
Milk	1.50	1.75	10 lit.

Solution: Construction of Index

Item	p_0	p_1	q_0	$V = P_0q_0$	P = p1/p0 x 100	PV	Log P	Log PV
Wheat	2	2.50	40	80	125	10000	2.0969	167.7520
Sugar	3	3.25	20	60	108.33	6499.8	2.0346	122.0760
Milk	1.50	1.75	10	15	116.67	1750.05	2.0669	31.0065
Total				**∑V= 155**		**∑PV = 18249.85**		**∑ Log PV =320.8345**

$$P_{01} = \frac{\sum PV}{\sum V} = \frac{\mathbf{18249.85}}{\mathbf{155}} = 117.74$$

$$P_{01} = \text{Antilog } \frac{\sum V \log P}{\sum V}$$

$$= \text{Antilog } \frac{\mathbf{320.8345}}{\mathbf{155}}$$

$$= \text{Antilog } 2.0699$$

$$= 111.4$$

Test of consistency of Index Numbers

1. **Time Reversal Test**

$$P_{01} = \sqrt{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}} \; ; \; P_{10} = \sqrt{\frac{\Sigma P_0 q_1}{\Sigma P_1 q_1} \times \frac{\Sigma P_0 q_0}{\Sigma P_1 q_0}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \times \frac{\Sigma P_0 q_1}{\Sigma P_1 q_1} \times \frac{\Sigma P_0 q_0}{\Sigma P_1 q_0}}$$

$$= \sqrt{1} = 1$$

It shows that the Time Reversal Test is satisfied

2. **Factor Reversal Test**

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \; ; \; Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma p_0 q_0}{\Sigma q_0 p_1}}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$$

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

$$= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

Illustration 5:

Compute Index Number, using Fishers Ideal Formula and show that it satisfies time – reversal test and factor – reversal test.

Commodity	Quantity	Base Year Price	Quantity	Current Year Price
A	12	10	15	12
B	15	7	20	5
C	24	5	20	9
D	5	16	5	14

Solution:

Computation of Index Number

Commodity	q_0	p_0	q_1	p_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	12	10	15	12	144	120	180	150
B	15	7	20	5	75	105	100	140
C	24	5	20	9	216	120	180	100
D	5	16	5	14	70	80	70	80
					$\sum p_1q_0$ **= 505**	$\sum p_0q_0$ **= 425**	$\sum p_1q_1$ **= 530**	$\sum p_0q_1$ **= 470**

$$P_{01} = \sqrt{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}} \times 100$$

$$= \sqrt{\frac{505}{425} \times \frac{530}{470}} \times 100$$

$$= \sqrt{1.188 \times 1.128} \times 100 = 115.8$$

Test of consistency of Index Numbers

1. **Time Reversal Test**

Time Reversal Test is satisfied when $P_{01} \times P_{10} = \sqrt{1} = 1$

$$P_{01} \times P_{10} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}$$

$$= \sqrt{\mathbf{1}} = \mathbf{1}$$

$$P_{10} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}$$

$$= \sqrt{\frac{470}{530} \times \frac{425}{505}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{505}{425} \times \frac{530}{470} \times \frac{470}{530} \times \frac{425}{505}}$$

$$= \sqrt{1}$$

$$= \mathbf{1}$$

2. Factor Reversal Test

Factor Reversal Test is satisfied when $P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100;$$

$$Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma p_0 q_0}{\Sigma q_0 p_1}} \times 100$$

$$\mathbf{P_{01} \times Q_{01}} = \sqrt{\frac{505}{425} \times \frac{530}{470} \times \frac{470}{425} \times \frac{530}{505}}$$

$$= \frac{530}{425}$$

$$= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

∴ Hence the given data satisfies the time – reversal test and factor – reversal test.

Single commodity case

Link Relatives:

In this method, the relative for each year is found out from the prices of the preceding year. Here the base changes from year to year. The index numbers by using this method are called link relative index numbers or link relatives.

Chain Base method:

Chain index is one in which the figures for each year are first expressed as percentages of the preceding year. These percentages are then chained together by successive multiplication.

The following formula is used to find out chain index:

$$\text{Chain Index} = \frac{\text{Link relative of the current year x previous year chain index}}{100}$$

$$\text{Link relative} = \frac{\textbf{Current years price}}{\textbf{previous year price}} \text{ x } 100$$

Advantages of chain base index

1. It is a great advantage to economists and businessmen.
2. Without any recalculation we can omit old items and introduce new items.
3. Weights can be adjusted as frequently as possible.
4. It is free from seasonal variations.

Base shifting:

The change of reference base period is called shifting the base. For comparison, it is necessary to shift the base from one period to another period. We can express the series in terms of a recent period. It is called base shifting

$$\text{Index number} = \frac{\textbf{Current year's old index number}}{\textbf{New base year's old index number}} \text{ x } 100$$

Illustration:

Reconstruct the following Index Numbers by shifting base to: (i) 2005 and (ii) 2007.

Year:	2000	2001	2002	2003	2004	2005	2006	2007
Index No.'s:	120	150	160	180	200	200	210	240

Solution: Constructing of Index Numbers (Base Change)

S. No.	Year	Index No.'s	2005 (Base = 200)	2007 (Base = 240)
1	2000	120	$\frac{120}{200} * 100 = 60$	$\frac{120}{240} * 100 = 50$
2	2001	150	$\frac{150}{200} * 100 = 75$	$\frac{150}{240} * 100 = 62.5$
3	2002	160	$\frac{160}{200} * 100 = 80$	$\frac{160}{240} * 100 = 66.66$
4	2003	180	$\frac{180}{200} * 100 = 90$	$\frac{180}{240} * 100 = 75$
5	2004	200	$\frac{200}{200} * 100 = 100$	$\frac{200}{240} * 100 = 83.33$

6	**2005**	**200**	$\frac{200}{200} * 100 = 100$	$\frac{200}{240} * 100 = 83.33$
7	2006	210	$\frac{210}{200} * 100 = 105$	$\frac{210}{240} * 100 = 87.5$
8	**2007**	**240**	$\frac{240}{200} * 100 = 120$	$\frac{240}{240} * 100 = 100$

$$\frac{\textbf{Current year's old index number}}{\textbf{New base year's old index number}} \text{ x } 100$$

Splicing and Deflating Index Numbers:

Splicing two Index Number Series:

The statistical method connects an old index number series with a revised series in order to make the series continuous is called splicing.

The formula is :

Spliced Index Number = Index Number of Current Year x $\frac{\text{Old Index of New Base Year}}{100}$

Illustration:

Two sets of Indices, one with 1993 as base and the other with 2001 as base are given below:

(a) Year	Index No.'s	(b) Year	Index No.'s
1993	100		
1994	110		
1995	120	2001	100
1996	190	2002	105
1997	300	2003	90
1998	330	2004	95
1999	360	2005	102
2000	390	2006	110
2001	400	2007	96

The Index (a) with 1993 base was discontinued in 2001. You are required to Splice the second Index Number (b) with 2001 base to the first index number.

Solution: Splicing of Index Numbers

S. No.	Year	Index No.'s *(a) with 1993 as base*	Index No.'s *(b) with 2001 as base*	Index Numbers *(b) Spliced to with 1993 as base*
1	1993	100		

2	1994	110		
3	1995	120		
4	1996	190		
5	1997	300		
6	1998	330		
7	1999	360		
8	2000	390		
9	2001	**400**	100	100 x 400/100 = 400
10	2002		105	105 x 400/100 = 420
11	2003		90	90 x 400/100 = 360
12	2004		95	95 x 400/100 = 380
13	2005		102	102 x 400/100 = 480
14	2006		110	110 x 400/100 = 408
15	2007		96	96 x 400/100 = 384

$$\text{Spliced Index Number} = \text{Index Number of Current Year x } \frac{\text{Old Index of New Base Year}}{100}$$

Deflating Index Number:

The process of deflating or decreasing a figure with the help of index numbers, so as to allow for change in the price level is calling deflating. By this method a series of money wages or income can be corrected for price changes to find out the level of real wages or income.

The formula is:

$$\text{Money wage} = \frac{\text{Money Wage}}{\text{Price Index}} * 100$$

$$\text{Real Wage of Income Index Number} = \frac{\text{Index of Money Wage}}{\text{Price Index Number}} * 100 \quad \text{Or}$$

$$= \frac{\text{Real Wage of the Current Year}}{\text{Real Wage of the Base Year}} * 100$$

Illustration: Given the following data:

Year	Weekly Take Home Pay (Wages)	Consumer Price Index
2002	109.5	112.8
2003	112.2	118.2
2004	116.4	127.4
2005	125.08	138.2
2006	135.4	143.5
2007	138.1	149.3

i) What was the real average weekly wage for each year?

ii) In which year did the employee have the greatest buying power?

iii) What percentage increase in the weekly wages for the year 2007 is required, if any, to provide the same buying power that the employees enjoyed in the year in which they had the highest real wages?

Solution: Calculation of Real Wages

S. No.	Year	Weekly Take Home Pay (Wages)	Consumer Price Index	Real Wages
1	2002	109.5	112.8	$\frac{109.5}{112.8} * 100 = 97.07$
2	2003	112.2	118.2	$\frac{112.2}{118.2} * 100 = 92.92$
3	2004	116.4	127.4	$\frac{116.4}{127.4} * 100 = 91.37$
4	2005	125.08	138.2	$\frac{125.08}{138.2} * 100 = 90.51$
5	2006	135.4	143.5	$\frac{135.4}{143.5} * 100 = 94.36$
6	2007	138.1	149.3	$\frac{138.1}{149.8} * 100 = 92.19$

i) Real average weekly wage can be obtained by the formula:

$$\text{Real Wage} = \frac{\text{Money Wage}}{\text{Price Index}} * 100$$

ii) The employee had the greatest buying power in 2002 as the real wage was maximum in 2002.

iii) Absolute difference = 2002 Real Wages - 2007 Real Wages

= 97.07 - 92.19

= + 4.88

*****

I. Fill in the blanks:

1. Index numbers are **<u>Specialised</u>** averages.

2. Theoretically the best average in construction of index number is **<u>Geometric Mean</u>**

3. The two tests suggested by fisher which a good index should satisfy, are **Time Reversal Test and Factor Reversal Test**.

4. The base period should always be **Normal**

5.Index numbers are called **Economic Barometers** to measure changes in economic phenomena.

6. **Geometric Mean** is the most, appropriate average for constructing the index.

7. Quantity index reflects **Quantity** changes from one period to another.

8. Family budget method is a method to calculate **Consumer** price index.

II. Choose the correct answer :

1. Paasche index is based on:

(a) Base year quantities

(b) Current Year Quantities

(c) None of these.

2. The circular test is satisfied by:

(a) Simple aggregative index

(b) Passche's index

(c) Laspeyre 's

(d) Kelly's Index.

3. The best average in the construction of index number is :

(a) Median

(b) Geometric Mean

(c) Mode

(d) Arithmetic mean

4. The circular test is satisfied when:

(a) p12 *p23 *p31=0

(b) p 12* p23* p31=1

(c) p21*p 32* p31 =1

5. If one wants to measure changes in total monetary worth, then the right choice should be:

(a) A Quantity index

(b) A Price index

(c) A Value Index

6. Commodities which show considerable price fluctuations could be best measured by a:

(a) Quantity Index

(b) Value index

(c) Price index.

7. The aggregate price index that uses base year quantities as base is:

(a) Passche's index

(b) Fisher's index

(c) Laspeyre's Index

8. The two price indices P01 and P10 when multiplied, satisfy the following test:

(a) Circular Test

(b) Factor reversal test

(c) Time Reversal test.

III. Theoretical Questions:

1. "Index numbers are economic barometers". Explain the statement and explain what precautions should be taken in making use of published index numbers.

2. What is an Index number? Why are index numbers called " Economic barometers"?

3. Analyse the problems in the construction of index numbers and comment on the need for weighting.

4. What are the tests of a good index number? Define fisher' Ideal index number and show that it satisfies all these tests.

5. Discuss briefly the uses and limitations of index numbers of prices.

6. Distinguish between fixed-based and chain based index numbers. What are the points of importance in choosing the base in the determination of cost of living index numbers?

7. Explain how a cost of living index numbers is constructed and examine critically the formula used to construct such an index number. What purpose do such indices serve?

8. You are required to construct a cost of living index for the textile workers in Bombay. What information will you collect for the purpose? Explain the method of constructing the index?

9. Define:

(i) Laspeyres

(ii) Paasche's and

(iii) Fisher index number of prices.

10. Comment on the statement, " Laspeyre over estimates the price changes while paasche under estimates them in general, and hence Fisher provides a better estimate than both Laspeyre's and Paasche's".

11. What do you understand by price relatives and discuss the method of constructing index numbers based on them.

12. What are the chain Base Index Numbers? How are they constructed? What are their uses?

13. What do you understand by deflating of index numbers? Illustrate your answer.

14. What is base shifting? Why does it become necessary to shift the base of index numbers?

15. " An index number is a special type of average" . Discuss.

Practical Problems:

1. Construct an index number for 2016 taking 2015 as base from the following data.

Commodity	Price in 1984 (Rs.)	Price in 1985 (Rs.)
A	90	95
B	40	60
C	90	110
D	30	35

(Ans.: $P_{01} = 120$)

2. Compute a price index for the following by average of price relative method.

Commodity	A	B	C	D	E	F
Price in 2005 (Rs.)	10	25	30	20	40	50
Price in 2006 (Rs.)	15	30	40	30	50	55

($P_{01} = 131.39$)

3. Compute Index Number, using Fishers Ideal Formula and show that it satisfies time – reversal test and factor – reversal test.

Commodity	Quantity	Base Year Price	Quantity	Current Year Price
A	50	6	56	10
B	100	2	120	2
C	60	4	60	6

D	30	10	24	12
E	40	8	36	12

(P_{01} = 139.8)

4. Calculate a) Paasche‟s index, b) Laspeyre‟s index and c) Fisher‟s index numbers for the following data:

Commodity	**P_0**	**q_0**	**P_1**	**q_1**
A	12	20	15	25
B	10	8	16	10
C	15	2	12	1
D	60	1	65	1
E	3	2	10	1

(Ans. a) 130.13; b) 129.09; and c) 129.60)

5. Compute the trend from the following by the method of least squares:

Year :	1976	1977	1978	1979	1980	1981
Population (in Lakhs)	24	25	29	26	22	24

Estimate the likely production for 1984. (Ans. 23.119)

6. The following table gives the price per egg on the first of each month.

Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
1988	18	19	20	21	23	24	26	26	28	25	23	29
1989	18	21	21	21	22	24	25	26	27	24	23	21
1990	19	17	20	22	23	25	25	26	26	24	22	20

Calculate the seasonal variations for each of the months Jan. to Dec.

(Ans. 80.9; 83.9; 89.7; 94.2; 100; 107.5; 111.9; 114.9; 119.3; 107.5; 100.9; 89.8)

*****

Model Question

Reg. No. : .....

Code No. : 24037 E **Sub. Code : AMBA 11/ AMSL 11**

B.B.A. (CBCS) DEGREE EXAMINATION,

NOVEMBER 2020.

First Semester

Business Administration/Shipping and Logistics–Main

BUSINESS STATISTICS

(For those who joined in July 2020 onwards)

Time : Three hours Maximum : 75 marks

PART A — (10 × 1 = 10 marks)

Answer ALL questions.

Choose the correct answer :

1. Data collected from "Hindu" Newspaper is an example of
   (a) Primary data
   (b) Secondary data
   (c) Primary and Secondary data
   (d) None of these
2. Total angle of the pie-chart is
   (a) 45 (b) 90
   (c) 180 (d) 360
3. In chronological classification, data are classified on the basis of
   (a) Attributes (b) Class interval
   (c)Time (d) Locations

4. Which of the following is the most unstable average?
   (a) Mode (b) Median
   (c) Geometric Mean (d) Harmonic Mean
5. Range is ————————
   (a) Large value + Small value
   (b) Large value – Small value
   (c) Large value × Small value
   (d) Large value / Small value
6. Standard deviation is also called ————————
   (a) Root mean square deviation
   (b) Root Deviation
   (c) Sigma Square Deviation
   (d) Positive Mean deviation
7. The co-efficient of correlation
   (a) cannot be positive
   (b) cannot be negative
   (c) can be either positive or negative
   (d) none of these
8. The relationship between three or more variable is studies with help of ———————— correlation.
   (a) Positive
   (b) Negative
   (c) Linear
   (d) Multiple
9. The circular test is satisfied by
   (a) Simple aggregative index
   (b) Passche's Index
   (c) Laspeyre's index
   (d) Kelly's index
10. Seasonal variations respect during a period of
   (a) One year
   (b) Five year
   (c) Seven year
   (d) Three year

PART B — (5 × 5 = 25 marks)

Answer ALL questions, choosing either (a) or (b).

Each answer should not exceed 250 words.

11. (a) Discuss the functions of statistics.

or

(b) Describe the methods of collecting primary data.

12. (a) Explain the different types of classification.

or

(b) Calculate Harmonic Mean from the following data.

Size of Items	6	7	8	9	10	11
Frequency	4	6	9	5	2	8

13. (a) Calculate quartile deviation and its co-efficient of A's monthly earning for a year.

Month	1	2	3	4	5	6	7
Monthly earning	239	250	251	251	257	258	260

Month	8	9	10	11	12
Monthly earning	261	262	262	273	275

or

(b) Calculate the mean deviation from mean for the following data.

Class internal	2–4	4–6	6–8	8–10
Frequency	3	4	2	1

14. (a) Find out the co-efficient of correlation.

X	4	3	2	5	6
Y	1	2	3	5	4

or

(b) Given the following data, calculate the expected value of Y, when X = 12.

	X	Y
Average	7.6	14.8
Standard deviation	3.6	2.5

r = 0.99

15. (a)

Commodity	Price in 2014	Price in 2015
A	90	95

B	40	60
C	90	110
D	30	35

Construct an index number for 2015 taking 2014 as base.

or

(b) Explain the uses of time series.

PART C - (5 × 8 = 40 marks)

Answer ALL questions, choosing either (a) or (b)

Each answer should not exceed 600 words.

16. (a) Describe the scope and uses of statistics in business.

or

(b) Draw a multiple bar diagram for the following data.

Year	Sales (,000)	Gross profit (,000)	Net Profit (,000)
2012	100	30	10
2013	120	40	15
2014	130	45	25
2015	150	50	25

17. (a) Calculate the mode from the following series.

Size of Item	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45
Frequency	20	24	32	28	20	16	34	10	8

or

(b) Find out the Geometric mean

Yield of wheat	7.5-10.5	10.5-13.5	13.5-16.5	16.5-19.5	19.5-22.5	22.5-25.5	25.5-28.5
No. of	5	9	19	23	7	4	1

Farms

18. (a) Calculate standard deviation for the following data.

Class Interval	5-10	10-15	15-20	20-25	25-30
Frequency	6	5	15	10	14

or

(b) Prices of a particular commodity in five years in two cities are given below.

Price in city 'A'	Price in city 'B'
20	10
22	20
19	18
23	12
16	15

From the above data find the city which more stable prices.

19. (a) Explain the different types of correlation.

or

(b) Calculate co-efficient of correlation from the following data.

x :	12	9	8	10	11	13	7
y :	14	8	6	9	11	12	3

20. (a) Calculate Index number through Bowley's Method form the following data.

	2016		2017	
Commodity	Price	Quantity	Price	Quantity
A	10	3	8	3.25

B	20	15	15	20
C	2	25	3	23

or

(b) Calculate three yearly moving average of the following data.

Year:	2006	2007	2008	2009	2010
No. of Students:	15	18	17	20	23

Year:	2011	2012	2013	2014	2015
No. of Students:	25	29	33	36	40

Code No. : **24037 E**

LEARNING MATERIAL PREPARED BY:

Miss. K. MEENA, MBA, M.PHIL, PGDCA, (Ph.D).,

ASSISTANT PROFESSOR,

DEPARTMENT OF BUSINESS ADMINISTRATION,

GOVERNMENT ARTS AND SCIENCE COLLEGE FOR WOMEN,

SATHANKULAM – 628 704.

THOOTHUKUDI DISTRICT

E-MAIL ID: meenakmba@gmail.com.

*****