Statistical Technique 1

Unit: 1

Subject Name: Mathematics-IV

Subject Code: AAS0402

Dr. Kunti Mishra
NIET, Gr Noida
Department of Mathematics

B Tech 4th Sem

Brief Introduction of Faculty

Dr. Kunti Mishra
Assistant Professor
Department of Mathematics

Qualifications :
M.Sc.(Maths), M. Tech.(Gold Medalist) in Applied and Computational Mathematics, Ph.D

Ph.D. Thesis : Some Investigations in Fractal Theory
Total Number of Research Papers:15
Area of Interests: Fixed Point Theory, Fractals
Teaching Experience: 9 years

Evaluation Scheme

NOIDA INSTITUTE OF ENGINEERING & TECHNOLOGY, GREATER NOIDA
(An Autonomous Institute)

B. TECH (CSE)
EVALUATION SCHEME
SEMESTER-IV

Sl. No.	Subject Codes	Subject Name	Periods			Evaluation Scheme				End Semester		Total	Credit
			L	T	P	CT	TA	TOTAL	PS	TE	PE		
1	AAS0402	Engineering Mathematics-IV	3	1	0	30	20	50		100		150	4
2	AASL0401	Technical Communication	2	1	0	30	20	50		100		150	3
3	ACSE0405	Microprocessor	3	0	0	30	20	50		100		150	3
4	ACSE0403A	Operating Systems	3	0	0	30	20	50		100		150	3
5	ACSE0404	Theory of Automata and Formal Languages	3	0	0	30	20	50		100		150	3
6	ACSE0401	Design and Analysis of Algorithm	3	1	0	30	20	50		100		150	4
7	ACSE0455	Microprocessor Lab	0	0	2				25		25	50	1
8	ACSE0453A	Operating Systems Lab	0	0	2				25		25	50	1
9	ACSE0451	Design and Analysis of Algorithm Lab	0	0	2				25		25	50	1
10	ACSE0459	Mini Project using Open Technology	0	0	2				50			50	1
11	ANC0402 / ANC0401	Environmental Science*/ Cyber Security*(Non Credit)	2	0	0	30	20	50		50		100	0
12		MOOCs** (For B.Tech. Hons. Degree)											
		GRAND TOTAL										**1100**	**24**

****List of MOOCs (Coursera) Based Recommended Courses for Second Year (Semester-IV) B. Tech Students**

S. No.	Subject Code	Course Name	University / Industry Partner Name	No of Hours	Credits
1	AMC0046	Algorithmic Toolbox	University of California San Diego	24	1.5
2	AMC0031	Data Structures	University of California San Diego	25	2

Syllabus

Unit-I (Statistical Techniques-I)

Introduction: Measures of central tendency: Mean, Median, Mode, Moment, Skewness, Kurtosis, Curve Fitting ,Method of least squares, Fitting of straight lines, Fitting of second degree parabola, Exponential curves ,Correlation and Rank correlation, Linear regression, nonlinear regression and multiple linear regression

Unit-II (Statistical Techniques-II)

Testing a Hypothesis, Null hypothesis, Alternative hypothesis, Level of significance, Confidence limits, p-value, Test of significance of difference of means, Z-test, t-test and Chi-square test, F-test, ANOVA: One way and Two way. Statistical Quality Control (SQC), Control Charts, Control Charts for variables (Mean and Range Charts), Control Charts for Variables (p, np and C charts).

Syllabus

Unit III (Probability and Random Variable)

Random Variable: Definition of a Random Variable, Discrete Random Variable, Continuous Random Variable, Probability mass function, Probability Density Function, Distribution functions.

Multiple Random Variables: Joint density and distribution Function, Properties of Joint Distribution function, Marginal density Functions, Conditional Distribution and Density, Statistical Independence, Central Limit Theorem (Proof not expected).

Unit IV (Expectations and Probability Distribution)

Operation on One Random Variable – Expectations: Introduction, Expected Value of a Random Variable, Mean, Variance, Moment Generating Function, Binomial, Poisson, Normal, Exponential distribution.

Syllabus

Unit V (Wavelets and applications and Aptitude-IV)

Wavelet Transform, wavelet series. Basic wavelets (Haar/Shannon/Daubechies), orthogonal wavelets, multi-resolution analysis, reconstruction of wavelets and applications.

Number System, Permutation & Combination, Probability, Function, Data Interpretation, Syllogism.

Branch Wise Application

- Data Analysis
- Artificial intelligence
- Network and Traffic modeling

Course Objectives

- The objective of this course is to familiarize the students with statistical techniques. It aims to present the students with standard concepts and tools at an intermediate to superior level that will provide them well towards undertaking a variety of problems in the discipline.

  The students will learn:

- Understand the concept of correlation, moments, skewness and kurtosis and curve fitting.
- Apply the concept of hypothesis testing and statistical quality control to create control charts.
- Remember the concept of probability to evaluate probability distributions.
- Understand the concept of Mathematical Expectations and Probability Distribution.
- Remember the concept of Wavelet Transform and Solve the problems of Number System, Permutation & Combination, Probability, Function, Data Interpretation, Syllogism.

Course Outcomes

CO1: Understand the concept of correlation, moments, skewness and kurtosis and curve fitting.

CO2: Apply the concept of hypothesis testing and statistical quality control to create control charts.

CO3: Remember the concept of probability to evaluate probability distributions

CO4: Understand the concept of Mathematical Expectations and Probability Distribution

CO2: Remember the concept of Wavelet Transform and Solve the problems of Number System, Permutation & Combination, Probability, Function, Data Interpretation, Syllogism.

Program Outcomes

S.No	Program Outcomes (POs)
PO 1	Engineering Knowledge
PO 2	Problem Analysis
PO 3	Design/Development of Solutions
PO 4	Conduct Investigations of Complex Problems
PO 5	Modern Tool Usage
PO 6	The Engineer & Society
PO 7	Environment and Sustainability
PO 8	Ethics
PO 9	Individual & Team Work
PO 10	Communication
PO 11	Project Management & Finance
PO 12	Lifelong Learning

PSOs

PSO	Program Specific Outcomes(PSOs)
PSO1	The ability to identify, analyze real world problems and design their ethical solutions using artificial intelligence, robotics, virtual/augmented reality, data analytics, block chain technology, and cloud computing
PSO2	The ability to design and develop the hardware sensor devices and related interfacing software systems for solving complex engineering problems.
PSO3	The ability to understand inter disciplinary computing techniques and to apply them in the design of advanced computing.
PSO4	The ability to conduct investigation of complex problem with the help of technical, managerial, leadership qualities, and modern engineering tools provided by industry sponsored laboratories.

CO-PO Mapping(CO1)

Sr. No	Course Outcome	PO1	PO 2	PO 3	PO4	PO 5	PO 6	PO 7	PO 8	PO 9	PO10	PO11	PO12
1	CO1	H	H	H	H	L	L	L	L	L	L	L	M
2	CO2	H	H	H	H	L	L	L	L	L	L	M	M
3	CO3	H	H	H	H	L	L	L	L	L	L	M	M
4	CO4	H	H	H	H	L	L	L	L	L	L	L	M
5	CO5	H	H	H	H	L	L	L	L	L	L	M	M

*L= Low *M= Medium *H= High

CO-PSO Mapping(CO2)

CO	PSO1	PSO2	PSO3	PSO4
CO.1	H	L	M	L
CO.2	L	M	L	M
CO.3	M	M	M	M
CO.4	H	M	M	M
CO.5	H	M	M	M

*L= Low *M= Medium *H= High

Program Educational Objectives(PEOs)

PEO-1: To have an excellent scientific and engineering breadth so as to comprehend, analyze, design and provide sustainable solutions for real-life problems using state-of-the-art technologies.

PEO-2: To have a successful career in industries, to pursue higher studies or to support entrepreneurial endeavors and to face the global challenges.

PEO-3: To have an effective communication skills, professional attitude, ethical values and a desire to learn specific knowledge in emerging trends, technologies for research, innovation and product development and contribution to society.

PEO-4: To have life-long learning for up-skilling and re-skilling for successful professional career as engineer, scientist, entrepreneur and bureaucrat for betterment of society.

Result Analysis

Branch	Semester	Sections	No. of enrolled Students	No. Passed Students	% Passed
CS	IV	A	67	65	97%
IOT	IV	A	49	45	91.83%

End Semester Question Paper Template

Link: 100 Marks Question Paper Template.docx

Prerequisite and Recap (CO1)

- Knowledge of Maths 1 B.Tech.
- Knowledge of Maths 2 B.Tech.
- Knowledge of Permutation and Combination.

Brief Introduction about the Subject with Videos

- We will discuss properties of complex function (limits, continuity, differentiability, Analyticity and integration)
- In 3rd module we will discuss application of partial differential equations
- In 4th module we will discuss numerical methods for solving algebraic equations, system of linear equations, definite integral and 1st order ordinary differential equation.
- In 5th module we will discuss aptitude part.
- https://youtu.be/iUhwCfz18os
- https://youtu.be/ly4S0oi3Yz8
- https://youtu.be/f8XzF9_2ijs

Unit Content

- Introduction
- Measures of central tendency: Mean, Median, Mode.
- Moment
- Skewness
- Kurtosis
- Curve Fitting
- Method of least squares
- Fitting of straight lines
- Fitting of second degree parabola
- Exponential curves
- Correlation and Rank correlation,
- Linear regression
- Nonlinear regression
- Multiple linear regression

Unit Objectives(CO1)

- The objective of this course is to familiarize the engineers with concept of Statistical techniques.
- It aims to show case the students with standard concepts and tools from B. Tech to deal with advanced level of mathematics and applications that would be essential for their disciplines.

Topic objectives (CO1)

Measures of central tendency

- **To present a brief picture of data-** It helps in giving a brief description of the main feature of the entire data.
- **Essential for comparison-** It helps in reducing the data to a single value which is used for doing comparative studies.
- **Helps in decision making-** Most of the companies use measuring central tendency to plan and develop their businesses economy.
- **Formulation of policies-** Many governments rely on this medium while forming any policies.

Measures of Central Tendency (CO1)

❑ Measures of Central Tendency or Averages:

Definition : According to Prof. Bowley: Averages are "statistical constants which enable us to comprehend in a single effort the significance of the whole."

Types of Measures of Central Tendency: There are five types of measures of central tendency

- ➢ Arithmetic Mean or Simple Mean
- ➢ Median
- ➢ Mode
- ➢ Geometric Mean
- ➢ Harmonic Mean

➢ <u>Arithmetic Mean</u>

Definition

Arithmetic mean of a set of observations is their *sum divided by the number of observations*, *e.g.*, the arithmetic mean $\bar{x}$ of n observations $x_1, x_2, ..., x_n$ is given by:

$$\bar{x} = \frac{x_{1+}x_{2+} \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_n$$

❖ **In case of the frequency distribution** $x_i|f_i, i = 1, 2, ..., n$, where f_i is the frequency of the variable x_i,

$$\bar{x} = \frac{f_1x_1 + f_2\,x_2 + \cdots + f_nx_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum_{i=1}^{n} f_ix_i}{\sum_{i=1}^{n} f_i} = \frac{1}{N}\sum_{i=1}^{n} f_ix_i \text{ , where } \sum_{i=1}^{n} f_i$$

$$= N$$

Arithmetic Mean(CO1)

In case of grouped or continuous frequency distribution, *x* is taken as the mid-value of the corresponding class.

Example: Find the arithmetic mean of the following frequency distribution:

X:	1	2	3	4	5	6	7
f:	5	9	12	17	14	10	6

Solution:

Computation of mean

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_n x_n}{f_1 + f_2 + \cdots + f_n}$$

$$= \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{1}{N} \sum_{i=1}^{n} f_i x_i$$

$$\text{where } \sum_{i=1}^{n} f_i = N$$

Arithmetic Mean(CO1)

x	f	fx
1	5	5
2	9	18
3	12	36
4	17	68
5	14	70
6	10	60
7	6	42
Total	73	299

By using formula $\sum_{i=1}^{n} f_i = N = 73, \sum_{i=1}^{n} f_i x_i = 299$

$$Mean = \frac{1}{N}\sum_{i=1}^{n} f_i x_i = \frac{299}{73} = 4.09$$

Daily Quiz (CO1)

Example: Calculate the mean for the following frequency distribution:

Class interval	**0-8**	**8-16**	**16-24**	**24-32**	**32-40**	**40-48**
Frequency	8	7	16	24	15	7

Solution: Arithmetic mean =25.404

Example: The average salary of male employees in a farm was Rs. 5,200 and that of females was Rs. 4,200. The mean salary of all the employees was Rs. 5,000.Find the percentage of male and female employees.

➢ Median:

Definition: Median of a distribution is the value of the variable which divides it into two equal parts.

It is the value such that the number of observations above it is equal to the number of observations below it. The median is thus a *positional average.*

❖ Ungrouped Data:

- If the number of observations is odd then median is the middle value after the values have been arranged in ascending or descending order of magnitude.
- In case of even number of observations, there are two middle terms and median is obtained by taking the arithmetic mean of middle terms.

Median(CO1)

Example

1. Median of Values 25, 20, 15, 35, 18. Median: 20
2. Median of Values 8, 20, 50, 25, 15, 30. Median: 22.5

❖ **Discrete Frequency Distribution**

In this case median is obtained by considering the cumulative frequencies. The steps involved

i. Find $\frac{N}{2}$, where $N = \sum_{i=1}^{n} f_i$

ii. See the cumulative frequency (c.f.) just greater than $\frac{N}{2}$.

iii. corresponding value of x is median.

Median(CO1)

Example: Obtain the median for the following frequency distribution:

x:	1	2	3	4	5	6	7	8	9
f:	8	10	11	16	20	25	15	9	6

Solution:

i. Find $\frac{N}{2} = \frac{8+10+11+16+20+25+15+9+6}{2} = \frac{120}{2} = 60,$

where $N = \sum_{i=1}^{n} f_i$

i. See the cumulative frequency (c.f.) just greater than $\frac{N}{2}$.

ii. corresponding value of x is median.

Median(CO1)

x	f	c.f.
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	105
8	9	114
9	6	120
Total	120=N	

Here N = 120, The cumulative frequency just greater than $\frac{N}{2}$ is 65 and the 2 value of x corresponding to 65 is 5. Therefore, median is 5.

Median(CO1)

❖ **Continuous Frequency Distribution**

In this case, the class corresponding to the c.f. just greate $\frac{N}{2}$ is called the median class and the value of median is obtained by the formula:

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

where

- l is the lower limit of the class,
- f is the frequency of the median class,
- h is the magnitude of the median class,
- c is the c.f. of the class preceding the median class,
- $N = \sum_{i=1}^{n} f_i$

Daily Quiz(CO1)

Example : find the median wages of the following distribution.

Wages	No. of workers
2000-3000	3
3000-4000	5
4000-5000	20
5000-6000	10
6000-7000	5

Solution: The median wage is Rs. 4,675.

Mode(CO1)

- **Mode:**
- Mode is the value which occurs most frequently in a set of observations and around which the other items of the set cluster densely.
- It is the point of maximum frequency or the point of greatest density.
- In other words the mode or modal value of the distribution is that value of the variate for which frequency is maximum.

Calculation of Mode

- **In case of discrete distribution:** Mode is the value of x corresponding to maximum frequency but in any one (or more)of the following cases.

Mode(CO1)

i. If the maximum frequency is repeated.

ii. If the maximum frequency occurs in the very beginning or at the end of distribution .

iii. If there are irregularities in the distribution, the value of mode is determined by the method of grouping.

❖ **In case of continuous frequency distribution:** mode is given by the formula

$$\text{Mode} = l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$$

where l is the lower limit, h the width and f_m the frequency of the model class f_1 and f_2 are the frequencies of the classes preceding and succeeding the modal class respectively. While applying the above formula it is necessary to see that the class intervals are of the same size.

Mode(CO1)

❖ **For a symmetrical distribution, mean, median and mode coincide.**

When mode is ill defined ,where the method of grouping also fails

its value can be ascertained by the formula

Mode=3Median-2Mean

This measure is called the empirical mode.

Q. Calculate the mode from the following frequency distribution**.**

Size(x)	4	5	6	7	8	9	10	11	12	13
Freqency (f)	2	5	8	9	12	14	14	15	11	13

Solution: Method of Grouping :

Mode(CO1)

$Size(x)$	1	2	3	4	5	6
4	2	7				
5	5		13			
6	8	17		15		
7	9		21		22	29
8	12	26		35		
9	14		28		40	43
10	14	29		40		
11	15		26		39	
12	11	24				
13	13					

Mode(CO1)

Since the item 10 occurs maximum number of times i.e.5times,hence the mode is 10.

Columns	*Size of item having max. frequ*
1 max.15	11
2max 29	10, 11
3 max 28	9, 10
4 max 40	10, 11, 12
5 max 40	8 9 10
6 max 43	9 10 11

Mode(CO1)

Q. Find the mode of the following:

Marks	0-5	6-10	11-15	16-20	21-25
No.of candidates	7	10	16	32	24
Marks	26-30	31-35	36-40	41-45	
No.of candidates	18	10	5	1	

Solution: Here the greatest frequency 32 lies in the class 16-20.Hence modal class is 16-20.But the actual limits of this class are 15.5-20.5.

$$l = 15.5, f_m = 32, f_1 = 16, f_2 = 24, h = 5$$

Mode(CO1)

$$\text{Mode}= l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$$

$$= 15.5 + \frac{32 - 16}{64 - 16 - 24} \times 5$$

$$= 15.5 + \frac{16}{24} \times 5$$

$$= 15.5 + \frac{10}{3}$$

$$= 18.83\ marks$$

Daily Quiz(CO1)

Q.1 Calculate the mean, median and mode of the following data-

Wages (in Rs)	0-20	20-40	40-60	60-80	80-100	100-120	120-140
No. of Workers	6	8	10	12	6	5	3

Recap(CO1)

- ✓ Measures of central tendency
- ✓ Mean
- ✓ Mode
- ✓ Median

Moments

- In mathematical statistics it involve a basic calculation. These calculations can be used to find a probability distribution's mean, variance, and skewness.

Moments (CO1)

- **Moments**: The moment of a distribution are the arithmetic means of the various powers of the deviations of items from some given number.
- Moments about mean (central moment)
- Moments about any arbitrary number (Raw Moment)
- Moments about origin

Central Moments (CO1)

- **Moment about mean (central moment):**

- **For an Individual Series :** If $x_1, x_{2,.....}x_n$ are the values of the variable under consideration , the r^{th} moment μ_r about mean $\overline{x}$ is defined as

Moment about mean $\mu_r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^r}{n}$; r = 0,1,2, … .

- **For a frequency Distribution:** If $x_{1,}x_{2,....,}x_n$ are the values of a variable x with the corresponding frequencies $f_1, f_2, \ldots, f_n$ respectively then r^{th} moment μ_r about the mean $\bar{x}$ is defined as

Central Moments (CO1)

$$\mu_r = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^r}{N}; \mathrm{r} = 0,1,2 \ldots.$$

where $N = \sum_{i=1}^{n} f_i$

in particular $\mu_0 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^0 = \frac{1}{N}\sum_{i=1}^{n} f_i = \frac{N}{N} = 1$

Note. In case of a frequency distribution with class intervals, the values of x are the midpoints of the intervals.

Example 1. Find the first four moments for the following individual series.

Solution: Calculation of Moments

x	3	6	8	10	18

Central Moments (CO1)

$S.No.$	x	$x-\overline{x}$	$(x-\overline{x})^2$	$(x-\overline{x})^3$	$(x-\overline{x})^4$
1	3	-6	36	-216	1296
2	6	-3	9	-27	81
3	8	-1	1	-1	1
4	10	1	1	1	1
5	18	9	81	729	6561
$n=5$	$\sum x = 45$	$\sum(x-\bar{x})=0$	$\sum(x-x)^2$ =128	$\sum(x-\bar{x})^3$ =486	$\sum(x-\bar{x})^4$ =7940

Central Moments (CO1)

For any distribution, $\mu_0 = 1$

$$\mu_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

For any distribution, $\mu_1 = 0$, for r=2,

$$\mu_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{128}{5} = 25.6$$

Therefore for any distribution , μ_2 coincides with the variance of the distribution.

Similarly, $\mu_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 = \frac{486}{5} = 97.2$

$\mu_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4 = \frac{7940}{5} = 1588$

Central Moments (CO1)

Now $\bar{x} = \frac{\sum x}{n} = \frac{45}{5} = 9$

$\mu_1 = \frac{\sum(x-\bar{x})}{n} = \frac{0}{5} = 0,$

$\mu_2 = \frac{\sum(x-\bar{x})^2}{n} = \frac{128}{5} = 25.6,$

$\mu_3 = \frac{\sum(x-\bar{x})^3}{n} = \frac{486}{5} = 97.2,$

$\mu_4 = \frac{\sum(x-\bar{x})^4}{n} = \frac{7940}{5} = 1588,$

Central Moments (CO1)

For any distribution, $\mu_0 = 1$ for r=1

$$\mu_1 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x}) = \frac{1}{N}\sum_{i=1}^{n} f_i x_i - \bar{x}\left[\frac{1}{N}\sum_{i=1}^{n} f_i\right] = \bar{x} - \bar{x} = 0$$

For any distribution, $\mu_1 = 0$, for r=2,

$$\mu_2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2 = (S.D)^2 = Variance$$

Therefore for any distribution , μ_2 coincides with the variance of the distribution.

Similarly, $\mu_3 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^3$

$\mu_4 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^4$ and so on.

Central Moments (CO1)

- **Example** $\mu_{1,}\mu_{2,}\mu_{3,}\mu_4$ for the following frequency distribution.

Marks	5-15	15-25	25-35	35-45	45-55	55-65
No.of students	10	20	25	20	15	10

- **Sol. Calculation of Moments**

Central Moments (CO1)

Marks	No.of Students(f)	Mid-Point (x)	fx	$x-\bar{x}=x-34$	$f(x-\bar{x})$	$f(x-\bar{x})^2$	$f(x-\bar{x})^3$	$f(x-\bar{x})^4$
5-15	10	10	100	-24	-240	5760	-138240	3317760
15-25	20	20	400	-14	-280	3920	-54880	768320
25-35	25	30	750	-4	-100	400	-1600	6400
35-45	20	40	800	6	120	720	4320	25920
45-55	15	50	750	16	240	3840	61440	983040
55-65	10	60	600	26	260	6760	175760	4569760
	N=100		$\sum fx$=3400		$\sum f(x-\bar{x})$=0	$\sum f(x-\bar{x})^2$=21400	$f(x-\bar{x})^3$=46800	$f(x-\bar{x})^4$=9671200

Central Moments (CO1)

$$\bar{x} = \frac{\sum fx}{N} = \frac{3400}{100} = 34$$

$$\mu_1 = \frac{\sum f(x-\bar{x})}{N} = \frac{0}{100} = 0$$

$$\mu_2 = \frac{\sum f(x-\bar{x})^2}{N} = \frac{21400}{100} = 214$$

$$\mu_3 = \frac{\sum f(x-\bar{x})^3}{N} = \frac{46800}{100} = 468$$

$$\mu_4 = \frac{\sum f(x-\bar{x})^4}{N} = \frac{9671200}{100} = 96712$$

Raw Moments (CO1)

- **Moments about an arbitary number(Raw Moments):**
- If $x_1, x_2, x_3, \ldots\ldots, x_n$ are the values of a variable x with the corresponding frequencies $f_1, f_2, f_{3,\ldots\ldots} f_n$ respectively then r^{th} moment $\mu_r{}'$ about the number $x = A$ is defined as

$$\mu'_r = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^r; r = 0,1,2, \ldots$$

Where, $N = \sum_{i=1}^{n} fi$

For $r = 0, \mu'_0 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^0 = 1$

Raw Moments (CO1)

For $r = 1, \mu'_1 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A) = \frac{1}{N}\sum_{i=1}^{n} f_i x_i - \frac{A}{N}\sum_{i=1}^{n} fi = \bar{x} - A$

For $r = 2, \mu'_2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^2$

For $r = 3, \mu'_3 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^3$ and so on.

In Calculation work, if we find that there is some common factor h(>1) in values of $x - A$,we can ease our calculation work by defining $u = \frac{x-A}{h}$.

In that case , we have

$$\mu'_r = \frac{1}{N}\left(\sum_{i=1}^{n} f_i {u_i}^r\right) h^r; r = 0,1,2, \dots.$$

Moments about the origin (CO1)

- **Moments about the Origin:**

If $x_1, x_2, \dots \dots, x_n$ be the values of a variable x with corresponding frequencies $f_1, f_2, \dots \dots, f_n$ respectively then r^{th} moment about the origin v_r is defined as

$$v_r = \frac{1}{N}\sum_{i=1}^{n} f_i {x_i}^r \text{ ; r} = 0,1,2, \dots .$$

Where, $N = \sum_{i=1}^{n} f_i$

For $r = 0, v_0 = \frac{1}{N}\sum_{i=1}^{n} f_i {x_i}^0 = \frac{N}{N} = 1$

For $r = 1, v_1 = \frac{1}{N}\sum_{i=1}^{n} f_i x_i = \bar{x}$

For $r = 2, v_2 = \frac{1}{N}\sum_{i=1}^{n} f_i {x_i}^2$ and so on.

Relations (CO1)

relations:

$\mu_1 = 0$

$\mu_2 = {\mu_2}' - {\mu_1}'^2$

$\mu_3 = {\mu_3}' - 3{\mu_2}'{\mu_1}' + 2{\mu_1}'^3$

$\mu_4 = \mu_4' - 4\mu_3'{\mu_1}' + 6{\mu_2}'{\mu_1}'^2 - 3{\mu_1}'^4$

- **Relation Between $\boldsymbol{v_r}$ and $\boldsymbol{\mu_r}$:**

$$v_1 = \bar{x}$$

$$v_2 = \mu_2 + \bar{x}^2$$

$$v_3 = \mu_3 + 3\mu_2\bar{x} + \bar{x}^3$$

$$v_4 = \mu_4 + 4\mu_3\bar{x} + 6\mu_2\bar{x}^2 + \bar{x}^4$$

Karl Pearson's Coefficients(CO1)

❖ Karl Pearson's $\boldsymbol{\beta}, \boldsymbol{\gamma}$ Coefficients:

Karl Pearson defined the following four coefficients based upon the first four moments of a frequency distribution about it mean:

$$\beta_1 = \frac{{\mu_3}^2}{{\mu_2}^3} \qquad \beta_2 = \frac{\mu_4}{{\mu_2}^2} \qquad (\beta - \text{coefficients})$$

$$\gamma_1 = +\sqrt{\beta_1} \qquad \gamma_2 = \beta_2 - 3 \qquad (\gamma - \text{coefficients})$$

The practical use of this coefficients is to measure the skewness and kurtosis of a frequency distribution .These coefficients are pure numbers independent of units of measurement.

Karl Pearson's Coefficients(CO1)

Example1 : The first three moments of a distribution about the value "2" of the variable are 1,16 and -40.Show that the mean is 3,variance is 15 and $\mu_3 = -86$.

Solution: We have A=2,$\mu'_1 = 1$, $\mu'_2 = 16$ and $\mu'_3 = -40$

We have that $\mu'_1 = \bar{x} - A \Longrightarrow \bar{x} = \mu'_1 + A = 1 + 2 = 3$

Variance=$\mu_2 = \mu'_2 - {\mu'_1}^2 = 16 - (1)^2 = 15$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2{\mu'_1}^3 = -40 - 3(16)(1) + 2(1)^3$$
$$= -40 - 48 + 2 = -86.$$

Karl Pearson's Coefficients(CO1)

Example 2: The first moments of a distribution about the value "35" are $-1.8, 240, -1020$ *and* 144000. Find the values of $\mu_1, \mu_2, \mu_3, \mu_4$.

Solution: $\mu_1 = 0$

$\mu_2 = \mu'_2 - {\mu_1}'^2 = 240 - (-1.8)^2 = 236.76$

$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2{\mu'_1}^3$

$= -1020 - 3(240)(-1.8) + 2(-1.8)^3 = 264.36$

$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2{\mu'}^2_1 - 3{\mu'}^4_1$

$= 144000 - 4(-1020)(-1.8) + 6(240)(-1.8)^{2-}3(-1.84)^4$

$= 141290.11.$

Karl Pearson's Coefficients(CO1)

Example 3: Calculate the variance and third central moment from the following data.

x_i	0	1	2	3	4	5	6	7	8
F_i	1	9	26	59	72	52	29	7	1

Solution: Calculation of Moments

x	f	$u=\frac{x-A}{h}, A=4, h=1$	f_u	fu^2	fu^3
0	1	-4	-4	16	-64
1	9	-3	-27	81	-243
2	26	-2	-52	104	-208
3	59	-1	-59	59	-59
4	72	0	0	0	0

Karl Pearson's Coefficients(CO1)

5	52	1	52	52	52
6	29	2	58	116	232
7	7	3	21	63	189
8	1	4	4	16	64
			$\sum fu = -7$	$\sum fu^2 = 507$	$\sum fu^3 = -37$

$$\mu'_1 = \left(\frac{\sum fu}{N}\right) \mathrm{h} = \frac{-7}{256} = -0.02734$$

$$\mu'_2 = \left(\frac{\sum fu^2}{N}\right) h^2 = \frac{507}{256} = 1.9805$$

Karl Pearson's Coefficients(CO1)

$$\mu'_3 = \left(\frac{\sum f u^3}{N}\right) h^3 = \frac{-37}{256} = -0.1445$$

Moments about Mean:

$\mu_1 = 0$

$\mu_2 = \mu'_2 - {\mu'_1}^2 = 1.9805 - (-.02734)^2 = 1.97975$

Variance=1.97975

Also $\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2{\mu_1}'^3$

$= (-0.1445) - 3(1.9805)(-0.02734) + 2(-0.02734)^3$

=0.0178997

Third central moment= 0.0178997.

Daily Quiz(CO1)

Q1. The first four moments of a distribution are 3, 10.5,40.5,168.Comment upon the nature of the distribution.

Q2. For a distribution, the mean is 10,variance is 16,γ_1 is 1 and β_2 is 4. Find the first four moment about origin.

Recap(CO1)

- ✓ Measures of central tendency
- ✓ Moment

Topic objective (CO1)

Skewness

- It tells us whether the distribution is normal or not
- It gives us an idea about the nature and degree of concentration of observations about the mean
- The empirical relation of mean, median and mode are based on a moderately skewed distribution

❑ Skewness:

- It means *lack of symmetry*.
- It *gives us an idea about the shape of the curve* which we can draw with the help of the given data.
- A distribution is said to be skewed if—

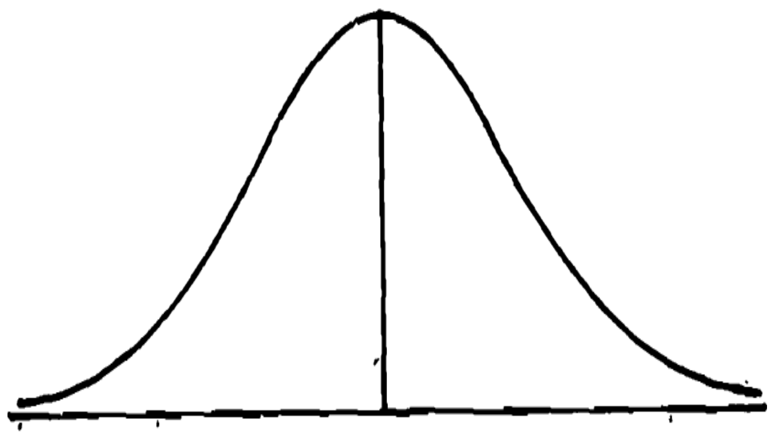
Mean, median and mode fall at different points, i.e.,

Mean ≠ Median ≠ Mode;

- Quartiles are not equidistant from median; and
- The curve drawn with the help of the given data is not symmetrical but stretched more to one side than to the other.

Symmetrical Distribution

A symmetric distribution is a type of distribution where the left side of the distribution mirrors the right side. In a symmetric distribution, the mean, mode and median all fall at the same point.



$\bar{x}$ (Mean) = M_0 = M_d

Measures of Skewness:

The measures of skewness are:

- $S_k = M - M_d$,
- $S_k = M - M_o$,
- $S_k = (Q_3 - M_d) - (M_d - Q_1)$,

where M is the mean, M_d, the median, M_o, the mode, Q_1, the first quartile deviation and Q_3, the third quartile deviation of the distribution.

These are the absolute measures of skewness.

- **Coefficients of Skewness:** For comparing two series we do not calculate these absolute measures but we calculate the relative measures called the *coefficients of skewness* which are pure numbers independent of units of measurement.

The following are the *coefficients of skewness:*

- Prof. Karl Pearson's Coefficient of Skewness,
- Prof. Bowley's Coefficient of Skewness,
- Coefficient of Skewness based upon Moments.

Prof. Karl Pearson's Coefficient of Skewness:

Definition

- It is defined as:

$$SK_p = \frac{A.M. - Mode}{S.D} = \frac{3(M - M_d)}{\sigma}$$

where σ is the standard deviation of the distribution. If mode is ill-

$Mode$=3Median-2mean

defined, then using the empirical relation,

$M_o = 3M_d - 2M$, for a moderately asymmetrical distribution, we have

- From above two formulas, we observe that $S_k = 0$ if $M = M_o = M_d$.
- Hence for a symmetrical distribution, mean, median and mode coincide.
- Skewness is positive if $M > M_o$ or $M > M_d$, and negative if $M < M_o$ or $M < M_d$.
- Limits are: $|S_k| \leq 3$ or $-3 \leq S_k \leq 3$.
- However, in practice, these limits are rarely attained.

Skewness(CO1)

Coefficient of Skewness based upon Moments

Definition

It is defined as: $\gamma_1 = \frac{\mu_3}{\sqrt{{\mu_2}^3}}$

where γ_1 are Pearson's Coefficients and defined as:

$S_k = 0$, if either $\beta_1 = 0$ or $\beta_2 = -3$. Thus $S_k = 0$, if and only if $\beta_1 = 0$.

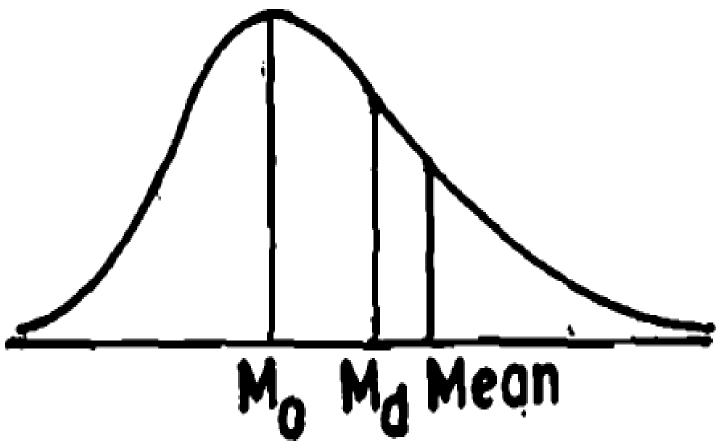
Thus for a symmetrical distribution $\beta_1 = 0$.

In this respect β_1 is taken as a *measure of skewness*.

Skewness(CO1)

- The coefficient of skewness based upon moments is to be regarded as without sign.
- The Pearson's and Bowley's coefficients of skewness can be positive as well as negative.

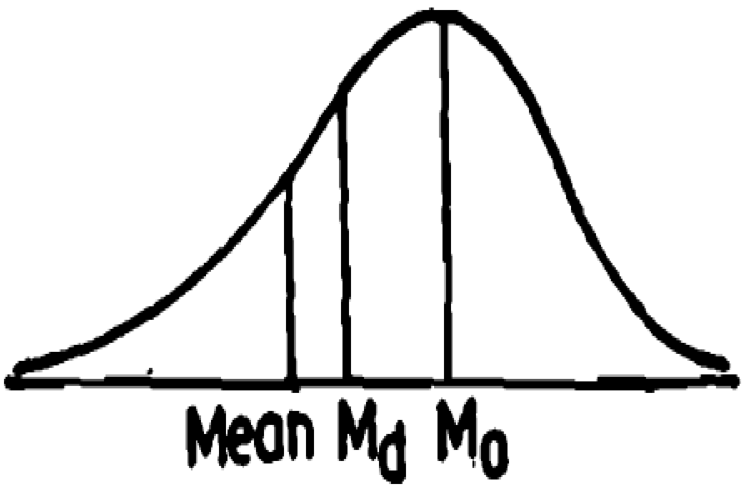
❖ **Positively Skewed Distribution:** The skewness is positive if the larger tail of the distribution lies towards the higher values of the variate (the right),i.e., if the curve drawn with the help of the given data is stretched more to the right than to the left.



Skewness(CO1)

❖ **Negatively Skewed Distribution:**

The skewness is negative if the larger tail of the distribution lies towards the lower values of the variate (the left), i.e., if the curve drawn with the help of the given data is stretched more to the left than to the right.



Skewness(CO1)

Pearson's $\boldsymbol{\beta_1}$ and $\boldsymbol{\gamma_1}$ Coefficients:

$$\boldsymbol{\gamma_1 = \sqrt{\beta_1} = \pm\frac{\mu_3}{\sqrt{{\mu_2}^3}}}$$

Q1. Karl Pearson coefficient of skewness of a distribution is 0.32, its standard deviation is 6.5 and mean is 29.6. find the mode of the distribution.

Solution: Given that $SK_p = 0.32$, σ=6.5 mean =29.6

$$SK_p = \frac{A.M.-Mode}{S.D} = \frac{3(M - M_d)}{\sigma}$$

$$0.32 = \frac{29.6 - Mode}{6.5} \Longrightarrow Mode = 27.52$$

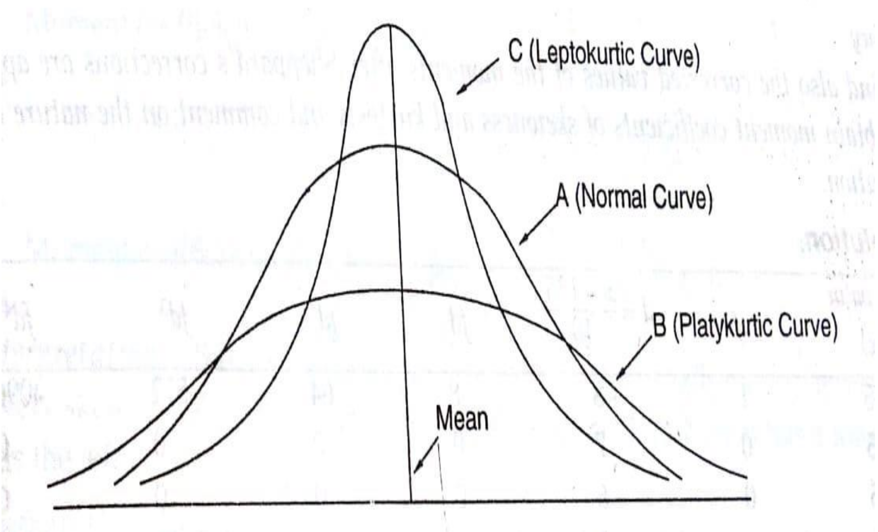
Topic objective (CO1)

Kurtosis

- Describe the concepts of kurtosis
- Explain the different measures of kurtosis
- Explain how kurtosis describe the shape of a distribution.

❑ Kurtosis

- If we know the measures of central tendency, dispersion and skewness, we still cannot form a complete idea about the distribution. Let us consider the figure in which all the three curves
- *A*, *B*, *and C* are symmetrical about the mean and have the same range.



Kurtosis (CO1)

Definition: Kurtosis is also known as *Convexity of the Frequency Curve* due to Prof. Karl Pearson.

- It *enables us to have an idea about the flatness or peakness* of the frequency curve.
- It is measure by the coefficient β_2 or its derivation γ_2 given as:

$$\beta_2 = \frac{\mu_4}{{\mu_2}^2}$$

- Curve of the type *A* which is *neither flat nor peaked* is called the *normal curve or mesokurtic curve* and for such curve $\beta_2 = 3$, i.e., $\gamma_2 = 0$.
- Curve of the type *B* which is *flatter than the normal curve* is known as *platycurtic curve* and for such curve $\beta_2 < 3$, i.e., $\gamma_2 < 0$.

Kurtosis (CO1)

Curve of the type *C* which is *more peaked than the normal curve* is called *leptokurtic curve* and for such curve $\beta_2 > 3$, i.e., $\gamma_2 > 0$.

Q2. For a distribution, the mean is 10, variance is 16, γ_1 is +1 and β_2 is 4. Comment about the nature of distribution. Also find third central moment.

Solution $1 = \pm\frac{\mu_3}{\sqrt{4096}} \Rightarrow$ $\boldsymbol{\mu_3}$**=64,** $\boldsymbol{\mu_2}$**=16,**

$$4 = \frac{\mu_4}{256} \Rightarrow \mu_4 = 1024$$

Since $\gamma_1 = +1$, the distribution is moderately positively skewed, i.e,
if we draw the curve of the given distribution, it will have longer tail towards the right.
Further, since $\beta_2 = 4 > 3$, the distribution is leptokurtic, i.e.,
it will be sightly more peaked than the normal curve.

Kurtosis (CO1)

Example 3 The first four moment about the working mean 28.5 of a distribution are 0.294,7.144,42.409 and 454.98. Calculate the first four moment about mean. Also evaluate β_1 and β_2 and comment upon the skewness and kurtosis of the distribution.

Solution: $\mu'_1 = .294, \mu'_2 = 7.144, \mu'_3 = 42.409, \mu'_4 = 454.98$ Moment about mean

$\mu_1 = 0,$

$\mu_2 = \mu'_2 - {\mu_1}'^2 = 7.0576.$

$\mu_3 = \mu'_3 - 3\mu'_2{\mu_1}' + 2{\mu_1}'^3 = 36.1588,$

$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2{\mu_1}'^2 - 3{\mu_1}'^4 = 408.7896$

Kurtosis (CO 1)

$$\beta_1 = \frac{{\mu^2}_3}{{\mu_2}^3} = 3.7193,$$

$$\beta_2 = \frac{\mu_4}{{\mu_2}^2} = 8.207$$

Skewness :β_1 is positive

$\gamma\ _1 = 1.9285$ so distribution is positivley skewed.

Kurtosis: $\beta_2 = 8.207 > 3$ so distribution is leptokutic.

Daily Quiz(CO1)

Q1. Find all four central moments and Discuss Skewness and Kurtosis for the following distribution-

Range of Expenditures	**2-4**	**4-6**	**6-8**	**8-10**	**10-12**
No. of families	38	292	389	212	69

Weekly Assignment(CO1)

Q1. The First four moments of a distribution about $x = 4$ are $1, 4, 10, and\ 45$. Find the first four moments about mean. Discuss the Skewness and Kurtosis and also comment upon the nature of the distribution.

Q2. Define the Mode and calculate Mode for the distribution of monthly rent Paid by Libraries in Karnataka

Monthly rent	500-1000	1000-1500	1500-2000	2000-2500	2500-3000	3000 & above
No.of Library	5	10	8	16	14	12

Q3. Write Short Note on

i. Range ii. Inter quartile range iii. Mean deviation iv. Standard deviation v. Variance

Q 4. Explain the measures of dispersion and also find the range & Coefficient of Range for the following data: 20, 35, 25, 30, 15.

Recap(CO1)

- ✓ Moments
- ✓ Relation between v_r and μ_r
- ✓ Relation between μ_r and μ'_r
- ✓ Skewness
- ✓ Kurtosis

Topic objectives(CO1)

Curve Fitting

- The objective of curve fitting is to find the parameters of a mathematical model that describes a set of data in a way that minimizes the difference between the model and the data.

Curve Fitting (CO1)

- **Curve Fitting :**Curve fitting means an exact relationship between two variables by algebraic equation. It enables us to represent the relationship between two variables by simple algebraic expressions e.g. polynomials, exponential or logarithmic functions. .It is also used to estimate the values of one variable corresponding to the specified values of other variables.

- **Method of Least Squares:** Method of least squares provides a unique set of values to the constants and hence suggests a curve of best fit to the given data.

Curve Fitting (CO1)

- **Fitting a Straight Line:** Let $(x_i, y_i), i = 1,2, \dots . n$ be n sets of observations of related data and

$$y = a.1 + b.x \qquad (1)$$

Normal equations

$$\sum y = na + b\sum x \qquad (2)$$

$$\sum xy = a\sum x + b\sum x^2 \qquad (3)$$

If n is odd then, $u = \frac{x-(middle\ term)}{interval(h)}$

If n is even then, $u = \frac{x-(mean\ of\ two\ middle\ terms)}{\frac{1}{2}(interval)}$

Curve Fitting (CO1)

Q.Fit a straight line to the following data by least square method.

x	0	1	2	3	4
y	1	1.8	3.3	4.5	6.3

Sol. Let the straight line obtained from the given data be

$y = a.1 + bx$ (1)

then the normal equations are

$\sum y = ma + b\sum x$ (2)

$\sum xy = a\sum x + b\sum x^2$ (3) m=5

Curve Fitting (CO1)

x	y	xy	x^2
0	1	0	0
1	1.8	1.8	1
2	3.3	6.6	4
3	4.5	13.5	9
4	6.3	25.2	16
$\sum x = 10$	$\sum y = 16.9$	$\sum xy = 47.1$	$\sum x^2 = 30$

$$\sum xy = a\sum x + b\sum x^2 \Rightarrow 47.1 = 10a + 30b$$

Solving we get $a = 0.72, b = 1.33$

Required lines is $y = 0.72 + 1.33x$

➢ Fitting of an Exponential Curve

Let $y = ae^{bx}$

Taking logarithm on both sides, we get

$\log_{10} y = \log_{10} a + bx \log_{10} e$

$Y = A + BX$

Where $Y = \log_{10} y$, $A = \log_{10} a$, $B = b \log_{10} e$, $X = x$

The normal equation for (1) are

$$\sum Y = nA + B\sum X \ and \ \sum XY = A\sum X + B\sum X^2$$

Solving these, we get A and B.

Then $a = antilog\ Aand\ B = \frac{B}{\log_{10} e}$

- **FITTING OF THE CURVE**

Let $y = ax^b$

Taking logarithm on both sides, we get

$\log_{10} y = \log_{10} a + b \log_{10} x$

$Y = A + BX$

Where $Y = \log_{10} y$, $A = \log_{10} a$, $B = b$, $X = \log_{10} x$

The normal equation to (1) are

$$\sum Y = nA + B \sum X \ \ and \ \ \sum XY = A \sum X + B \sum X^2$$

Which results A and B on solving and $a = antilog\ A, b = B$.

Curve Fitting (CO1)

Example Use the method of least squares to the fit the curve: $y = \frac{c_0}{x} + c_1\sqrt{x}$ to the following table of values:

X	0.1	0.2	0.4	0.5	1	2
Y	21	11	7	6	5	6

- Solution: Let given curve is $\boldsymbol{y = \frac{c_0}{x} + c_1\sqrt{x}}$

Normal equations are

$$\sum \frac{y}{x} = c_0 \sum \frac{1}{x^2} + c_1 \sum \frac{1}{\sqrt{x}}$$

$$\sum y\sqrt{x} = c_0 \sum \frac{1}{\sqrt{x}} + c_1 \sum x\,.$$

Curve Fitting (CO1)

x	y	$\frac{y}{x}$	$y\sqrt{x}$	$\frac{1}{\sqrt{x}}$	$\frac{1}{x^2}$
0.1	21	210	6.64078	3.16228	100
0.2	11	55	4.91935	2.23607	25
0.4	7	17.5	4.42719	1.58114	6.25
0.5	6	12	4.24264	1.41421	4
1	5	5	5	1	1
2	6	3	8.48528	0.70711	0.25
4.2		302.5	33.7152 4	10.1008 1	136.5

$$302.5 = 136.5c_0 + 10.10081c_1$$

$$33,71524 = 10.10081c_0 + 4.2c_1$$

so we have

$c_0 = 1.97327, c_1 = 3.28182$

Hence the curve is

$$\boldsymbol{y} = \frac{1.97327}{\boldsymbol{x}} + 3.28182\sqrt{\boldsymbol{x}}$$

Daily Quiz(CO1)

Q Fit a second degree parabola to the following data-

x	0	1	2	3	4
f	1	0	3	10	21

Recap(CO1)

- ✓ Moments
- ✓ Relation between v_r and μ_r
- ✓ Relation between μ_r and μ'_r
- ✓ Skewness & kurtosis
- ✓ Curve fitting

Topic objective (CO1)

Correlation

- Identify the direction and strength of a correlation between two factors.
- Compute and interpret the Pearson correlation coefficient and test for significance.
- Compute and interpret the coefficient of determination.
- Compute and interpret the Spearman correlation coefficient and test for significance.

Correlation(CO1)

- **Correlation:** In a bivariate distribution we are interested to find out if there is any correlation between the two variables under study.

- If the change in one variable affects a change in the other variable, the variables are said to be correlated.

❖ **Positive Correlation**

- If the two variables deviate in the same direction, i.e., if the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be *direct or positive.*

- For example, the correlation between **(i)** the heights and weights of a group of persons, and **(ii)** the income and expenditure; is positive.

Correlation(CO1)

- **Negative Correlation:**

- If the two variables deviate in the opposite directions, i.e., if increase (or decrease) in one results in corresponding decrease (or increase) in the other, correlation is said to be *diverse or negative.*

  - For example, the correlation between **(i)** the price and demand of a commodity, and **(ii)** the volume and pressure of a perfect gas; is negative.

- **Perfect Correlation:**

- Correlation is said to be perfect if the deviation in one variable is followed by a corresponding and proportional deviation in the other.

Correlation Coefficient:

- The correlation coefficient due to Karl Pearson is defined as a measure of intensity or degree of linear relationship between two variables.
- **Karl Pearson's Correlation Coefficient**
- Karl Pearson's correlation coefficient between two variables X and Y, is denoted by $r(X, Y)$ or r_{XY}, is a measure of *linear relationship* between them and is defined as:
- $r(X,Y) = \frac{Cov(x,y)}{\sigma_X \sigma_Y}$
- $f(x_i, y_i); i = 1, 2, ..., n$ is the bivariate distribution, then
- $Cov(X,Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$

Karl Pearson's Co –Efficient Of Correlation(or Product Moment Correlation Co-efficient)

Correlation co-efficient between two variable $x\ and\ y$, usually denoted by $r(x,y) or\ r_{xy}$ is a numerical measure of linear relationship between them and defined as

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum\ (y_i - \bar{y})^2}}$$

$$= \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2 . \frac{1}{n}\sum\ (y_i - \bar{y})^2}}$$

Correlation(CO1)

$$= \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$

$$r_{xy} = \frac{\sum(x - \bar{x})(y - \bar{y})}{n\sigma_x \sigma_y}$$

Or $r(x, y) = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$

Here n is the no. of pairs of values of $x \; and \; y$.

Note: Correlation co efficient is independent of change of origin and scale.

Let us define two new variables $u \; and \; v \; as$

$u = \frac{x-a}{h}, v = \frac{y-b}{k}$ where $a, b, h, k \; are \; constant \; then \; r_{xy} = r_{uv}$

Then $r(u, v) = \frac{n\sum uv - \sum u \sum v}{\sqrt{n\sum u^2 - (\sum u)^2}\sqrt{n\sum v^2 - (\sum v)^2}}$

Correlation(CO1)

Q.Find the coefficient of correlation between the values of $x \text{ } and \text{ } y$:

x	1	3	5	7	8	10
y	8	12	15	17	18	20

Sol. Here $n = 6$. The table is as follows.

x	y	x^2	y^2	xy
1	8	1	64	8
3	12	9	144	36
5	15	25	225	75
7	17	49	289	119
8	18	64	324	144
10	20	100	400	200
$\sum x = 34$	$\sum y = 90$	$\sum x^2 = 24$	$\sum y^2 = 14$	$\sum xy = 58$

Correlation(CO1)

Karl Pearson's coefficient of correlation is given by

$$r(x,y) = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

$$r(x,y) = \frac{(6\times 582)-(34\times 90)}{\sqrt{(6\times 248)-(34)^2}\sqrt{(6\times 1446)-(90)^2}} = 0.9879$$

Q. Find the co-efficient of correlation for the following table:

x	10	14	18	22	26	30
y	18	12	24	6	30	36

Solution: Let $u = \frac{x-22}{4}, v = \frac{y-24}{6}$

Correlation(CO1)

x	y	u	v	u^2	v^2	uv
10	18	-3	-1	9	1	3
14	12	-2	-2	4	4	4
18	24	-1	0	1	0	0
22	6	0	-3	0	9	0
26	30	1	1	1	1	1
30	36	2	2	4	4	4
Total		$\sum u = -3$	$\sum v = -3$	$\sum u^2 = 19$	$\sum v^2 = 19$	$\sum uv = 12$

Correlation(CO1)

Hence,n=6,$\bar{u} = \frac{1}{n}\sum u = \frac{1}{6}(-3) = -\frac{1}{2}$; $\bar{v} = \frac{1}{n}\sum v = \frac{1}{6}(-3) = -\frac{1}{2}$

Then $r_{uv} = \frac{n\sum uv - \sum u \sum v}{\sqrt{n\sum u^2 - (\sum u)^2}\sqrt{n\sum v^2 - (\sum v)^2}}$

$$= \frac{(6\times 12) - (-3)(-3)}{\sqrt{(6\times 19) - (-3)^2}\sqrt{(6\times 19) - (-3)^2}} = \frac{63}{\sqrt{105}\sqrt{105}} = 0.6$$

❖ **Calculation of co-efficient of correlation for a bivariate frequency distribution.**

- If the bivariate data on $x\ and\ y$ is presented on a two way correlation table and f is the frequency of a particular rectangle
- In the correlation table then

Correlation(CO1)

$$r_{xy} = \frac{\sum fxy - \frac{1}{n}\sum fx \sum fy}{\sqrt{\sum fx^2 - \frac{1}{n}(\sum fx)^2 \left[\sum fy^2 - \frac{1}{n}(\sum fy)^2\right]}}$$

Since change of origin and scale do not affect the co-efficient of correlation.$r_{xy} = r_{uv}$ where the new variables u, v are properly chosen.

Q. The following table given according to age the frequency of marks obtained by 100 students is an intelligence test:

Correlation(CO1)

Marks	18	19	20	21	total
10-20	4	2	2		8
20-30	5	4	6	4	19
30-40	6	8	10	11	35
40-50	4	4	6	8	22
50-60		2	4	4	10
60-70		2	3	1	6
Total	19	22	31	28	100

Calculate the coefficient of correlation between age and intelligence.

Solution: Age and intelligence be denoted by x *and* y respectively.

Correlation(CO1)

Mid value	x⟶ y↓	18	19	20	21	f	$u = \frac{y-45}{10}$	fu	fu^2	fuv
15	10-20	4	2	2		8	-3	-24	72	30
25	20-30	5	4	6	4	19	-2	-38	76	20
35	30-40	6	8	10	11	35	-1	-35	35	9
45	40-50	4	4	6	8	22	0	0	0	0
55	50-60		2	4	4	10	1	10	10	2
65	60-70		2	3	1	6	2	12	24	-2
	f	19	22	31	28	100	total	-75	217	59
	$v = x - 20$	-2	-1	0	1	Total				
	fv	-38	-22	0	28	-32				
	fv^2	76	22	0	28	126				
	fuv	56	16	0	-13	59				

Correlation(CO1)

Let us define two new variables $u\ and\ v\ as\ u = \frac{y-45}{10}, v = x - 20$

$$r_{xy} = r_{uv} = \frac{\sum fuv - \frac{1}{n}\sum fu \sum fv}{\sqrt{\sum fu^2 - \frac{1}{n}(\sum fu)^2\left[\sum fv^2 - \frac{1}{n}(\sum fv)^2\right]}}$$

$$= \frac{59 - \frac{1}{100}(-75)(-32)}{\sqrt{\left[217 - \frac{1}{100}(-75)^2\right]\left[126 - \frac{1}{100}(-32)^2\right]}} = \frac{59 - 24}{\sqrt{\frac{643}{4} \times \frac{2894}{25}}}$$

$$= 0.25$$

RANK CORRELATION:

Definition: Assuming that no two individuals are bracketed equal in either classification, each of the variables X and Y takes the values 1, 2, ...,n.

Hence, the rank correlation coefficient between A and B is denoted by r, and is given as:

$$\boldsymbol{r = 1 - \left[\frac{6\sum {D_i}^2}{n(n^2-1)}\right]}$$

Rank Correlation(CO1)

Question. Compute the rank correlation coefficient for the following data.

Person	A	B	C	D	E	F	G	H	I	J
Rank in maths	9	10	6	5	7	2	4	8	1	3
Rank in physics	1	2	3	4	5	6	7	8	9	10

Sol. Here the ranks are given and $n = 10$

Rank Correlation(CO1)

Person	R_1	R_2	D=$R_1 - R_2$	D^2
A	9	1	8	64
B	10	2	8	64
C	6	3	3	9
D	5	4	1	1
E	7	5	2	4
F	2	6	-4	16
G	4	7	-3	9
H	8	8	0	0
I	1	9	-8	64
J	3	10	-7	49
				$\sum D^2 = 280$

Rank Correlation(CO1)

$$r = 1 - \left[\frac{6\sum D^2}{n(n^2-1)}\right] = 1 - \left[\frac{6 \times 280}{10(100-1)}\right] = 1 - 1.697 = -0.697$$

Uses:

- It is used for finding correlation coefficient if we are dealing with qualitative characteristics which cannot be measured quantitatively but can be arranged serially.
- It can also be used where actual data are given.
- In case of extreme observations, Spearman's formula is preferred to Pearson's formula.

Limitations:

- It is not applicable in the case of bivariate frequency distribution.

Tied Correlation(CO1)

- For $n > 30$, this formula should not be used unless the ranks are given, since in the contrary case the calculations are quite time-consuming.

TIED RANKS: If some of the individuals receive the same rank in a ranking of merit, they are said to be tied.

- Let us suppose that m of the individuals, say, $(k+1)^{th}$, $(k+2)^{th}$, ..., $(k+m)^{th}$, are tied.

- Then each of these m individuals assigned a common rank, which is arithmetic mean of the ranks $k+1, k+2, \ldots, k+m$.

$$r = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12} m_1\left(m_1{}^2 - 1\right) + \frac{1}{12} m_2\left(m_2{}^2 - 1\right) + \cdots\right\}}{n(n^2-1)}$$

Tied Correlation(CO1)

Question: Obtain the rank correlation co-efficient for the following data:

x	68	64	75	50	64	80	75	40	55	64
y	62	58	68	45	81	60	68	48	50	70

Solution: Here marks are given so write down the ranks

Tied Correlation(CO1)

X	68	64	75	50	64	80	75	40	55	64	Total
Y	62	58	68	45	81	60	68	48	50	70	
Ranks in $X(x)$	4	6	2.5	9	6	1	2.5	10	8	6	
Ranks in Y(y)	5	7	3.5	10	1	6	3.5	9	8	2	
$D = x - y$	-1	-1	-1	-1	5	-5	-1	1	0	4	0
D^2	1	1	1	1	25	25	1	1	0	16	72

75 2 times

64 3 times

68 2 times

Tied Correlation(CO1)

$$r = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}m_1({m_1}^2 - 1) + \frac{1}{12}m_2({m_2}^2 - 1) + \frac{1}{12}m_3({m_3}^2 - 1)\right\}}{n(n^2 - 1)}$$

$$= 1 - \frac{6\left\{72 + \frac{1}{12}.2(2^2 - 1) + \frac{1}{12}.3(3^2 - 1) + \frac{1}{12}.2(2^2 - 1)\right\}}{10(10^2 - 1)}$$

$$= 1 - \left\{\frac{6 \times 75}{990}\right\} = \frac{6}{11} = 0.545$$

Daily Quiz(CO1)

Q1. Find the rank correlation coefficient for the following data:

x	23	27	28	28	29	30	31	33	35	36
y	18	20	22	27	21	29	27	29	28	29

Recap(CO1)

- ✓ Correlation
- ✓ Karl Pearson coefficient of correlation
- ✓ Rank Correlation
- ✓ Tied Rank

Topic objectives (CO1)

Regression

- Explanation of the variation in the dependent variable, based on the variation in independent variables and Predict the values of the dependent variable.

❑ REGRESSION ANALYSIS:

- Regression measures the nature and extent of correlation .Regression is the estimation or prediction of unknown values of one variable from known values of another variable.

Difference between curve fitting and regression analysis: The only fundamental difference, if any between problems of curve fitting and regression is that in regression, any of the variables may be considered as independent or dependent while in curve fitting, one variable cannot be dependent.

Curve of regression and regression equation:

- If two variates x and y are correlated i.e., there exists an association or relationship between them, then the scatter diagram

Regression Analysis(CO1)

will be more or less concentrated round a curve. This curve is called the curve of regression and the relationship is said to be expressed by means of curvilinear regression.

- The mathematical equation of the regression curve is called regression equation**.**

Some following types of regression will discuss here:

- Linear Regression
- Non- linear Regression
- Multiple linear Regression

Linear Regression(CO1)

- **LINEAR REGRESSION**:
- When the point of the scatter diagram concentrated round a straight line, the regression is called linear and this straight line is known as the line of regression.
- Regression will be called non-linear if there exists a relationship other than a straight line between the variables under consideration.

Linear Regression(CO1)

LINES OF REGRESSION: A line of regression is the straight line which gives the best fit in the least square sense to the given frequency.

LINES OF REGRESSION:

Let $y = a + bx$ ----.(1)

be the equation of regression line of $y \; on \; x$.

$\sum y = na + b \sum x \; \ldots \ldots .(2)$

$\sum xy = a \sum x + b \sum x^2 \; \ldots \ldots .(3)$

Solving (2) and (3) for ‘a’ and ‘b’ we get.

$$b = \frac{\sum xy - \frac{1}{n} \sum x \sum y}{\sum x^2 - \frac{1}{n} (\sum x)^2} = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} \ldots ..(4)$$

Linear Regression(CO1)

$a = \frac{\sum y}{n} - b\frac{\sum x}{n} = \bar{y} - b\bar{x}$ … …(5)

Eqt.(5) given $\bar{y} = a + b\bar{x}$

Hence $y = a + bx$ line passes through point$(\bar{x}, \bar{y})$

Putting $a = \bar{y} - b\bar{x}$ in equation $y = a + bx$,we get

$y - \bar{y} = b(x - \bar{x})$………(6)

Eqt.(6) is called regression line of $y\ on\ x.'\ b'$ is called the regression coefficient of $y\ on\ x$ and is usually denoted by b_{yx}.

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$x = a + by$$
$$x - \bar{x} = b_{xy}(y - \bar{y})$$

Where b_{xy} is the regression coefficient of x on y and is given by

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2}$$

Or $b_{xy} = r\frac{\sigma_x}{\sigma_y}$ where the terms have their usual meanings.

USE OF REGRESSION ANALYSIS:

A) In the field of a business this tool of statistical analysis is widely used .Businessmen are interested in predicting future production, Consumption ,investment, prices, profits and sales etc.

B) In the field of economic planning and sociological studies, projections of population birth rates ,death and other similar variables are of great use.

Linear Regression(CO1)

Where $\bar{x}$ and $\bar{y}$ are mean values while

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

In eqt.(3),shifting the origin to$(\bar{x}, \bar{y})$, we get

$$\sum (x - \bar{x})(y - \bar{y}) = a\sum (x - \bar{x}) + b\sum (x - \bar{x})^2$$

$$\Rightarrow nr\sigma_x\sigma_y = a(0) + bn{\sigma_x}^2$$

$$\Rightarrow b = r\frac{\sigma_y}{\sigma_x}$$

Where r is the coefficient of correlation $\sigma_x and\ \sigma_y$ are the standard deviations of x and y series respectively.

Regression Analysis Properties(CO1)

Properties of Regression Coefficients:

Property 1. Correlation coefficient is the geometric mean between the regression coefficients.

Proof : The coefficients of regression are $\frac{r\sigma_y}{\sigma_x}$ and $\frac{r\sigma_x}{\sigma_y}$.

G.M. between them= $\sqrt{\frac{r\sigma_y}{\sigma_x} \times \frac{r\sigma_x}{\sigma_y}} = \sqrt{r^2} = \text{r} =$coefficient of correlation.

Property 2. If one of the regression coefficients is greater than unity, the other must be less than unity.

Proof. The two regression coefficients are $b_{yx} = \frac{r\sigma_y}{\sigma_x}$ and $b_{xy} = \frac{r\sigma_x}{\sigma_y}$.

Regression Analysis Properties(CO1)

Let $b_{yx} > 1$,then $\frac{1}{b_{yx}} < 1$

Since $b_{yx} . b_{xy} = r^2 \leq 1$

$$b_{xy} \leq \frac{1}{b_{yx}} < 1$$

Similarly if $b_{xy} > 1, then\ b_{yx} < 1.$

Property 3. Airthmetic mean of regression coefficient is greater than the Correlation coefficient.

Proof. We have to prove that

$$\frac{b_{yx} + b_{xy}}{2} > r$$

$$r\frac{\sigma_y}{\sigma_x} + r\frac{\sigma_x}{\sigma_y} > 2r$$

Regression Analysis Properties(CO1)

$\sigma_x{}^2 + \sigma_y{}^2 > 2\sigma_x\sigma_y$

$\left(\sigma_x - \sigma_y\right)^2 > 0$ which is true.

Property 4: Regression coefficients are independent of the origin but not of scale.

Proof. Let $u = \frac{x-a}{h}, v = \frac{y-b}{k}$, where a, b, h and k are constants

$$\mathrm{b_{yx}} = \frac{r\sigma_y}{\sigma_x} = \mathrm{r.}\frac{k\sigma_v}{h\sigma_u} = \frac{k}{h}\left(\frac{r\sigma_v}{\sigma_u}\right) = \frac{k}{h}b_{vu}$$

Similarly, $b_{xy} = \frac{h}{k}b_{uv}$,

Thus b_{yx} and b_{xy} are both independent of a and b but not of $h\ and\ k$.

Regression Analysis Properties(CO1)

Property 5: The correlation coefficient and the two regression coefficient have same sign.

Proof: Regression coefficient of $y\ on\ x = b_{yx} = r\frac{\sigma_y}{\sigma_x}$

Regression coefficient of x $on\ y = b_{xy} = r\frac{\sigma_x}{\sigma_y}$

Since σ_x and σ_y are both positive; b_{yx}, b_{xy} and r have same sign.

- **Angle Between Two Lines of Regression:**

If θ is the acute angle between the two regression lines in the case of two variables $x\ and\ y$,show that

Regression Analysis Properties(CO1)

$tan\theta = \frac{1-r^2}{r} . \frac{\sigma_x \sigma_y}{{\sigma_x}^2+{\sigma_y}^2}$, where $r, \sigma_{x,} \sigma_y$ have their usual meanings.

Explain the significance of the formula where $r = 0$ and $r = \pm 1$

Proof: Equations to the lines of regression of y on x and x on y are

$y - \bar{y} = \frac{r\sigma_y}{\sigma_x}(x - \bar{x})$ and $(x - \bar{x})= \frac{r\sigma_x}{\sigma_y}(y - \overline{y)}$

The slopes are $m_1 = \frac{r\sigma_y}{\sigma_x}$ and $m_2 = \frac{\sigma_y}{r\sigma_x}$

$$\tan\theta = \pm \frac{m_2 - m_1}{1+m_2 m_1} = \pm \frac{\frac{\sigma_y}{r\sigma_x} - \frac{r\sigma_y}{\sigma_x}}{1+\frac{{\sigma_y}^2}{{\sigma_x}^2}}$$

Regression Analysis Properties(CO1)

$$= \pm \frac{1-r^2}{r} \cdot \frac{\sigma_y}{\sigma_x} \cdot \frac{{\sigma_x}^2}{{\sigma_x}^2 + {\sigma_y}^2} = \pm \frac{1-r^2}{r} \cdot \frac{\sigma_x \sigma_y}{{\sigma_x}^2 + {\sigma_y}^2}$$

Since $r^2 \leq 1$ and σ_x, σ_y are positive.

tan$\theta = \frac{1-r^2}{r} \cdot \frac{\sigma_x \sigma_y}{{\sigma_x}^2+{\sigma_y}^2}$ Where $r = 0, \theta = \frac{\pi}{2}$ the two lines of regression are Perpendicular to each other. Hence the estimated value of y is the same for all values of x and vice versa.

When $r = \pm 1, tan\theta = 0$ so that $\theta = 0 \; or \; \pi$

Hence the lines of regression coincide and there is perfect correlation between the two variates $x \; and \; y$.

Linear Regression(CO1)

Q. The equation of two regression lines, obtained in a correlation analysis of 60 observations are:

$5x = 6y + 24$ and $1000y = 768x - 3608$.What is the correlation Coefficient ?Show that the ratio of coefficient of variability of x to $that$ of y is $\frac{5}{24}$.What is the ratio of variance of x and y?

Solution: Regression line of x on y is

$$5x = 6y + 24$$

$$x = \frac{6}{5}y + \frac{24}{5}$$

$$b_{xy} = \frac{6}{5}$$

Regression line of y on x is

Linear Regression(CO1)

$1000y = 768x - 3608$
$y = 0.768x - 3.608$
$b_{yx} = 0.768$

$r\frac{\sigma_x}{\sigma_y}=\frac{6}{5}$……..(3)

$r\frac{\sigma_y}{\sigma_x}=0.768$….(4)

Multiply equations(3) and (4) we get
$r^2 = 0.9216 \Rightarrow r = 0.96$

Dividing (3) by (4) we get

$$\frac{{\sigma_x}^2}{{\sigma_y}^2} = \frac{6}{5} \times \frac{1}{0.768} = 1.5625$$

Linear Regression(CO1)

Taking square root, we get

$\frac{\sigma_x}{\sigma_y}$=$1.25 = \frac{5}{4}$

Since the regression lines pass through the point$(\bar{x}, \bar{y})$ we have

$5\bar{x} = 6\bar{y} + 24$

$1000\bar{y} = 768\bar{x} - 3608$

Solving the above equation $\bar{x} and \bar{y}$,we get $\bar{x}$=6, $\bar{y}$ =1

Coefficient of variability of $x = \frac{\sigma_x}{\bar{x}}$

Coefficient of variability of y $= \frac{\sigma_y}{\bar{y}}$

Required ratio=$\frac{\sigma_x}{\bar{x}} \times \frac{\bar{y}}{\sigma_y} = \frac{\bar{y}}{\bar{x}}\left(\frac{\sigma_x}{\sigma_y}\right) = \frac{1}{6} \times \frac{5}{4} = \frac{5}{24}$

Non-Linear Regression(CO1)

- **Non-linear Regression:**

Let $y = a.1 + bx + cx^2$

Be a second degree parabolic curve of regression of y on x.

$$\Rightarrow \sum y = na + b\sum x + c\sum x^2$$

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

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

Multiple Linear Regression(CO1)

- **Multiple Linear Regression:**

Where the dependent variable is a function of two or more linear or non linear independent variables. consider such a linear function as $y = a + bx + cz$

$$\sum y = ma + b\sum x + c\sum z$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum xz$$

$$\sum yz = a\sum z + b\sum xz + c\sum z^2$$

Solving the above equations we get values of $a, b \text{ and } c$ then we get linear function $y = a + bx + cz$ is called the regression plan.

Multiple Linear Regression(CO1)

Q. Obtain a regression plane by using multiple linear regression To fit the data given below.

x	1	2	3	4
y	12	18	24	30
z	0	1	2	3

Sol. Let $y = a + bx + cz$ *be the required regression plane where* a, b, c *are the constants to* be determined by following equations.

$$\sum y = ma + b\sum x + c\sum z$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum xz$$

Multiple Linear Regression(CO1)

$$\sum yz = a\sum z + b\sum xz + c\sum z^2$$

Here $m = 4$ Substitution yields,

$84 = 4a + 10b + 6c$

$240 = 10a + 30b + 20c$

$156 = 6a + 20b + 14c$

$a = 10, b = 2, c = 4$

Hence the required regression plane is

$y = 10 + 2x + 4z$

Multiple Linear Regression(CO1)

x	z	y	x^2	z^2	yx	zx	yz
1	0	12	1	0	12	0	0
2	1	18	4	1	36	2	18
3	2	24	9	4	72	6	48
4	3	30	16	9	120	12	90
$\sum x = 10$	$\sum z = 6$	$\sum y$=84	$\sum x^2$= 30	$\sum z^2$= 14	$\sum yx$=240	$\sum zx$=20	$\sum yz$=156

Daily Quiz(CO1)

Q1 Two lines of regression are given by $7x - 16y + 9 = 0$ and $-4x + 5y - 3 = 0$ and $var(x)$=16.Calculate

(i) the mean of x and y

(ii) variance of y

(iii) The correlation coefficient.

Weekly Assignment(CO1)

Q1. Fit a straight line trend by the method of least square to the following data:

Year	1979	1980	1981	1982	1983	1984
Production	5	7	9	10	12	17

Q2. From the following data calculate Karl Pearson's coefficient of skewness

Marks Less than	10	20	30	40	50	60	70
No. of students	10	30	60	110	150	180	200

Q3. Write regression equations of X on Y and of Y on X for the following data -

Weekly Assignment(CO1)

X	1	2	3	4	5
Y	2	4	5	3	6

Q4. Fit a straight line trend by the method of least squares to the following data: -

Year	2012	2013	2014	2015	2016	2017
Sales of T.V. sets (in’000)	7	10	12	14	17	24

Faculty Video Links, Youtube & NPTEL Video Links and Online Courses Details

Suggested Youtube/other Video Links:

https://youtu.be/wWenULjri40
https://youtu.be/mL9-WX7wLAo
https://youtu.be/nPsfqz9EljY
https://youtu.be/nqPS29IvnHk
https://youtu.be/aaQXMbpbNKw
https://youtu.be/wDXMYRPup0Y
https://youtu.be/m9a6rg0tNSM
https://youtu.be/Qy1YAKZDA7k
https://youtu.be/Qy1YAKZDA7k
https://youtu.be/s94k4H6AE54
https://youtu.be/lBB4stn3exM
https://youtu.be/0WejW9MiTGg
https://youtu.be/QAEZOhE13Wg
https://youtu.be/ddYNq1TxtM0
https://youtu.be/YciBHHeswBM
https://youtu.be/VCJdg7YBbAQ
https://youtu.be/VCJdg7YBbAQ
https://youtu.be/yhzJxftDgms

MCQ (CO1)

Q1. Which one is true

i. Correlation helps to determine the validity of a test.

ii. Correlation helps to determine the reliability of a test.

iii. Correlation indicates the nature of the relationship between two variables.

iv. All of the above

Q2. Which one is true

i. $If\ b_{xy} > 1, then\ b_{yx} < 1.$

ii. $\frac{b_{yx} + b_{xy}}{2} > r$

iii. $\frac{b_{yx} + b_{xy}}{4} > 2r$

iv. $If\ b_{yx} > 1, then\ b_{xy} < 1.$

MCQ (CO1)

Q3. Sum of squares of items 2430, mean is 7 N=12, find the variance.

i. 176.5

ii. 12.38

iii. 153.26

iv. 14

Q4. Calculate the standard variation of the following

9, 8, 6,5,8,6

i. 2

ii. 3

iii. 1.414

iv. 2.414

Glossary (CO1)

Q 1 An in complete distribution is given below:

x	10-20	20-30	30-40	40-50	50-60	60-70	70-80
f	12	30	X	65	Y	25	18

Given that median value is 46 and N=229

i. X

ii. Y

iii. Mean

iv. Mode

Pick the correct option from glossary

a. 45.82

b. 33.5

c. 46.07

d. 45

Glossary (CO1)

Q2. For the following:

i. Equation of line y on x

ii. Regression coefficient x on y

iii. Correlation coefficient

iv. Equation of line x on y

Pick the correct option from glossary

a. $(x - \bar{x}) = b_{xy}(y - \bar{y})$

b. r(x,y)

c. $(y - \bar{y}) = b_{yx}(x - \bar{x})$

d. b_{xy}

Old Question Papers

[First Sessional Set-1 (CSE,IT,CS,ECE,IOT).docx]

[Second Sessional Set-2 (CSE,IT,CS,ECE,IOT).docx]

[Maths IV PUT.docx]

[Maths IV final paper_2022.pdf]

Expected Questions for University Exam(CO1)

Q1 Obtain normal equation by method of least square to the curve $y = c_0 x + \frac{c_1}{\sqrt{x}}$ .Fit it to the following data:

x	**0.1**	**0.2**	**0.4**	**0.5**	**1**	**2**
y	21	11	7	6	5	6

Q2. Find the multiple linear regressions of x on y and z from the data relating to three variables:

x	**7**	**12**	**17**	**20**
y	4	7	9	12
z	1	2	5	8

Q3. If θ is the angle between the two line of regression.then express $\tan\theta$ in terms of correlation coefficient(r). Explain the significance when $r = 0$ and $r = \pm 1$.

Q4. Two lines of regression are given by $7x - 16y + 9 = 0$ and $-4x + 5y - 3 = 0$ and $var(x)$=16.Calculate-(i) the mean of x and y (ii) S.D. of y (iii) the correlation coefficient.

•

Expected Questions for University Exam(CO1)

Q5 An incomplete distribution of families according to their expenditure per week is given below. The median and mode for the distribution are Rs 25 and Rs 24 respectively. Calculate the missing frequencies.

Expenditure	0-10	10-20	20-30	30-40	40-50
No. of families	14	?	27	?	15

Q6. The first four moments of a distribution about 2 are 1,2.5,5.5 and 16 resp.Calculate the four moments about mean and about the origin.

Recap (CO1)

We discussed the following topics:

- ✓ Measures of central tendency – mean, median, mode
- ✓ Moment
- ✓ Skewness
- ✓ Kurtosis
- ✓ Curve fitting
- ✓ Least squares principles of curve fitting
- ✓ Correlation
- ✓ Regression analysis

References

Text Books

- Erwin Kreyszig, Advanced Engineering Mathematics, 9thEdition, John Wiley & Sons, 2006.
- P. G. Hoel, S. C. Port and C. J. Stone, Introduction to Probability Theory, Universal Book Stall, 2003(Reprint).
- S. Ross: A First Course in Probability, 6th Ed., Pearson Education India, 2002.
- W. Feller, An Introduction to Probability Theory and its Applications, Vol. 1, 3rd Ed., Wiley, 1968.

References

Reference Books

- B.S. Grewal, Higher Engineering Mathematics, Khanna Publishers, 35th Edition, 2000. 2.T.Veerarajan : Engineering Mathematics (for semester III), Tata McGraw-Hill, New Delhi.

- R.K. Jain and S.R.K. Iyenger: Advance Engineering Mathematics; Narosa Publishing House, New Delhi.

- J.N. Kapur: Mathematical Statistics; S. Chand & Sons Company Limited, New Delhi.

- D.N.Elhance,V. Elhance & B.M. Aggarwal: Fundamentals of Statistics; Kitab Mahal Distributers, New Delhi.

Thank You