Noida Institute of Engineering and Technology, Greater Noida

Statistical Technique-II

Module 2

Subject Name: Statistics and Probability

Subject Code: BAS0303N

Dr. Lokesh Chaudhary
Department of Mathematics

B Tech 4th Sem

Syllabus

Module 1: (Statistical Techniques-I)

Introduction: Measures of central tendency: Mean, Median, Mode, Standard deviation, Quartile deviation, Moment, Skewness, Kurtosis.

Module 2: (Statistical Techniques-II)

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.

Module 3: (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).

Syllabus

Module 4: (Expectations and Probability Distribution)

Expectations of single Random Variable, Mean, Variance, Moment Generating Function, Binomial, Poisson, Normal, Exponential distribution.

Module 5: (Hypothesis Tests and Control Charts)

Testing a Hypothesis, Null hypothesis, Alternative hypothesis, Level of significance, Confidence limits, Test of significance of difference of means, Z-test, t-test and Chi-square test, F-test, One way ANOVA.

Statistical Quality Control (SQC), Control Charts, Control Charts for variables (Mean and Range Charts), Control Charts for Variables (p, np and C charts).

Branch Wise Application

- Data Analysis
- Artificial intelligence
- Network and Traffic modeling

Course Objectives

- The objective of this course is to familiarize the students with concepts of Probability and statistical techniques. It aims to equip the students with adequate Knowledge of statistics that will enable them in formulating Problems and solving problems analytically.

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.

Course Outcomes (CO2)

CO1: Apply the concept of moments, skewness and kurtosis in relevant field.

CO2: Apply the concept of correlation, regression and curve fitting with real world problems.

CO3: Apply the concept of probability and random variable.

CO4: Apply the concept of Mathematical Expectations and Probability Distribution in real life problems.

CO5: Apply the concept of hypothesis testing and statistical quality control to create control charts.

Prerequisite

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

Module Content(CO2)

- 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

Module Objectives (CO2)

- 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 (CO2)

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 (CO2)

- **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 (CO2)

Fitting a Straight Line: Let $(x_i, y_i), i = 1,2, \ldots . n$ 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 (CO2)

Q1. 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 (CO2)

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 $y = 0.72 + 1.33x$

Curve Fitting (CO2)

➢ 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 $a = antilog \ A and \ B = \frac{B}{\log_{10} e}$

Curve Fitting (CO2)

➢ 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 (CO2)

Q1. Fit the following Equation, $y = ax^b$, to the following data:

x	1	2	3	4	5
y	0.5	1.7	3.4	5.7	8.4

Sol: The given equation is:

$y = ax^b$ (1)

then the normal equations are

$\log y = \log a + b \log x$

Let $Y = \log y$, $A = \log a$, $X = \log x$

$\sum Y = nA + b \sum X$ (2)

$\sum XY = A \sum X + b \sum X^2$ (3)

n=5

Curve Fitting (CO2)

x	y	$X = log\ x$	$Y = log\ y$	X^2	XY
1	0.5	0	-0.301	0	0
2	1.7	0.301	0.226	0.0694	0.0906
3	3.5	0.477	0.534	0.2536	0.2276
4	5.7	0.602	0.753	0.4551	0.3625
5	8.4	0.699	0.922	0.6460	0.4886
	Total	2.079	2.141	1.424	1.169

By Normal Equations

$2.141 = 5A + 2.079b$

$1.169 = 2.079A + 1.424b$

By solving these equations: $A = -0.334$ then $a = 0.46$ and $b = 1.75$

Hence the fitting is: $y = 0.46x^{1.75}$

Daily Quiz (CO2)

Q1. Fit a second degree parabola to the following data-

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

Q2. Find the best values of a & b for $y = ae^{bx}$ by the method of least squares to the following data:

x	0	5	8	12	20
y	3.0	1.5	1.0	0.55	0.18

Recap (CO2)

- ✓ Moments
- ✓ Relation between $v_r \; and \; \mu_r$
- ✓ Relation between $\mu_r \; and \; \mu'_r$
- ✓ Skewness & kurtosis
- ✓ Curve fitting

Topic objective (CO2)

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 (CO2)

- **Correlation:** In a bivariate distribution we are interested to find out i f 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 (CO2)

➢ **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 (CO2)

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

$$Cov(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$$

Correlation (CO2)

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 (CO2)

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

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

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 (CO2)

Q.Find the coefficient of correlation between the values of x *and* 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 (CO2)

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 (CO2)

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 (CO2)

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 (CO2)

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

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 (CO2)

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 (CO2)

$Mid\ value$	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 (CO2)

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 (CO2)

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:

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

Rank Correlation (CO2)

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 (CO2)

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 (CO2)

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

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 (CO2)

- 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 $k+1, k+2, ..., k+m$.

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

Tied Correlation (CO2)

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 (CO2)

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 (CO2)

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

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

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

Daily Quiz (CO2)

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

Q2. Given the following pairs of values:

Capital Employed (Rs. In Crore)	1	2	3	4	5	7	8	9	11	12
Profit (Rs. In Lakhs)	3	5	4	7	9	8	10	11	12	14

Do you think that there is any correlation between profits and capital employed? Is it positive or negative? Is it high or low?

Recap (CO2)

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

Topic objectives (CO2)

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 (CO2)

❑ 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 (CO2)

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 (CO2)

➢ **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 (CO2)

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.

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

$\sum xy = a\sum x + b\sum x^2$ ... .(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 (CO2)

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

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}.

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

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

Linear Regression (CO2)

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

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

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 (CO2)

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= $\sqrt{\frac{r\sigma_y}{\sigma_x} \times \frac{r\sigma_x}{\sigma_y}} = \sqrt{r^2} = r$ =coefficient of correlation.

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

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 (CO2)

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 (CO2)

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

$(\sigma_x - \sigma_y)^2 > 0$ 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

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

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 (CO2)

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 (CO2)

$tan\theta = \frac{1-r^2}{r} . \frac{\sigma_x \sigma_y}{{\sigma_x}^2+{\sigma_y}^2}$, where r, σ_x, σ_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 (CO2)

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

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 (CO2)

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 (CO2)

$1000y = 768x - 3608$

$y = 0.768x - 3.608$

$b_{yx} = 0.768$

$r\frac{\sigma_x}{\sigma_y} = \frac{6}{5}$ ……..(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 (CO2)

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 (CO2)

- **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 (CO2)

➢ 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 $a, b\ and\ c$ then we get linear function $y = a + bx + cz$ is called the regression plan.

Multiple Linear Regression(CO2)

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(CO2)

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

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 (CO2)

x	z	y	x^2	z^2	xy	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
$\sum x = 10$	$\sum z = 6$	$\sum y = 84$	$\sum x^2 = 30$	$\sum z^2 = 14$	$\sum xy = 240$	$\sum zx = 20$	$\sum yz = 156$

Daily Quiz (CO2)

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.

Q2 Find the two regression equation of X on Y and Y on X from the following data:

X:	10	12	16	11	15	14	20	22
Y:	15	18	23	14	20	17	25	28

Weekly Assignment (CO2)

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

Weekly Assignment (CO2)

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

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 1000)	7	10	12	14	17	24

MCQ (CO2)

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 (CO2)

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 (CO2)

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 (CO2)

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 (CO2)

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 (CO2)

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 (CO2)

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 (CO2)

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

End Semester Question Paper (CO2)

Printed Page:- 06 **Subject Code:- BAS0303**

Roll. No:

NOIDA INSTITUTE OF ENGINEERING AND TECHNOLOGY, GREATER NOIDA
(An Autonomous Institute Affiliated to AKTU, Lucknow)
B.Tech
SEM: III - THEORY EXAMINATION (2024- 2025)
Subject: Statistics & Probability

Time: 3 Hours **Max. Marks: 100**

General Instructions:

IMP: *Verify that you have received the question paper with the correct course, code, branch etc.*

1. *This Question paper comprises of* ***three Sections -A, B, & C.*** *It consists of Multiple Choice Questions (MCQ's) & Subjective type questions.*

2. *Maximum marks for each question are indicated on right -hand side of each question.*

3. *Illustrate your answers with neat sketches wherever necessary.*

4. *Assume suitable data if necessary.*

5. *Preferably, write the answers in sequential order.*

6. *No sheet should be left blank. Any written material after a blank sheet will not be evaluated/checked.*

SECTION-A 20

1. Attempt all parts:-

1-a. Wheat crops badly damaged on account of rains is: (CO1, K2) 1

(a) Cyclical movement
(b) Random movement
(c) Secular trend
(d) Seasonal movement

End Semester Question Paper (CO2)

1-b. Let the average of three numbers be 16. If two of the numbers are 8 and 12, then the remaining number is..... (CO1, K3) 1

(a) 28
(b) 18
(c) 12
(d) 30

1-c. One card is drawn from a standard pack of 52 plying cards. Find the probability that it is either a king or a queen. (CO2, K3) 1

(a) $\frac{1}{13}$
(b) $\frac{2}{13}$
(c) $\frac{3}{13}$
(d) None of these

1-d. If two events A and B are mutually exclusive, then the probability $P(A \cap B)$ is: (CO2, K2) 1

(a) 0
(b) 1
(c) P(A).P(B)
(d) None of these

End Semester Question Paper (CO2)

1-e. In binomial distribution probability of success in each trial remains _____.(CO3. K1) 1

(a) 0
(b) 1
(c) Constant
(d) Not defined

1-f. In Normal Distribution, Mean deviation about mean is (CO3, K1) 1

(a) σ
(b) $2\sigma/5$
(c) $4\sigma/5$
(d) $6\sigma/7$

1-g. The area of critical region depends on the size of.... (CO4, K1) 1

(a) Type I error
(b) Type II error
(c) Test statistics
(d) Sample

1-h. In conducting one way analysis of variance --- test statistics would be used. (CO4,K1) 1

(a) Z
(b) T
(c) Chi-Square
(d) F

End Semester Question Paper (CO2)

1-i. If $A=\{1,4,6\}$, $B=\{3,6\}$ and $C=\{3,4,6\}$ then $A \cap (B \cap C)$ is.... (CO5,K3) 1

(a) $\{3,4,5,6\}$

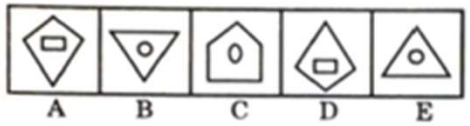
(b) $\{4,6\}$

(c) $\{1,4,6\}$

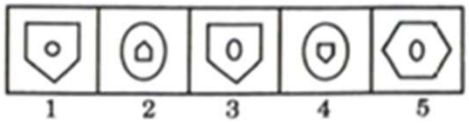
(d) $\{6\}$

1-j. Select a figure from amongst the answer figure which will continue the same series as stablished by the five-figure problem. (CO5,K2) 1

Problem Figure



Answer Figure



(a) 1

(b) 2

(c) 3

(d) 5

End Semester Question Paper (CO2)

2. Attempt all parts:-

2.a.	Prove that if two variables are independent, their correlation is zero but vice versa is not true. (CO1,K2)	2
2.b.	Find the probability of getting 53 Sundays in a leap year. (CO2,K3)	2
2.c.	A Binomial random variable X satisfies the relation 9P(X=4) =P(X=2) ,When n=6 . Find value of P(X=1) . (CO3, K3)	2
2.d.	Define an estimator in statistics. (CO4, K1)	2
2.e.	Check whether the function f : N→N is defined by $f(x)=x^2+12$ is one-one or not.(CO5, K3)	2

SECTION-B 30

3. Answer any five of the following:-

3-a.	The mean and standard deviation of the marks of 100 candidates was found to be 40 and 5.1, respectively. Later, it was discovered that a score of 40 was wrongly read as 50. Find out the correct mean and standard deviation respectively. (CO1,K3)	6
3-b.	Calculate Spearman's rank correlation coefficient from the following data: (CO1, K3)	6

End Semester Question Paper (CO2)

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

3-c. A random variable X has the following distribution, Find the value of K. Also find Mean and variance (CO2, K3) 6

x	-2	-1	0	1	2	3
P(x)	0.1	K	0.2	2K	0.3	K

3-d. A bag X contains 2 white and 3 red balls and another bag Y contains 4 white and 5 red balls. One ball is drawn at random from one of the bags and is found to be red. Find the probability that it was drawn from bag Y. (CO2, K3) 6

3.e. If 10% of the bolts produced by a machine are defective, determine the probability that out of 10 bolts choosen at random 6

i) 1

ii) None

iii) at most 2 bolts will be defective. (CO3,K3)

3.f. 6

A sample of 20 items has mean 42 units and S.D. 5 units. Test the hypothesis that it is a random sample from a normal population with mean 45 units. (If the tabular value at 5% LOS for 19 d. f. is 2.09). (CO4, K3)

3.g. How many different words can be formed using all the letters of the world ALLAHABAD 6

1. *When the vowels occupy the even position.*
2. Both L do not occur together. (CO5,K3)

End Semester Question Paper (CO2)

SECTION-C 50

4. Answer any one of the following:-

4-a. In a partially destroyed laboratory record of analysis of a correlation data, the following results only are legible: (CO1,K3) 10

Variance of x = 9; Regression equations: $8x - 10y + 66=0$, $40x - 18y = 214$.

What were (a) the mean values of x and y (b) the standard deviation of y (c) the coefficient correlation between x and y.

4-b. Fit a second degree parabola by the method of least squares to the following data . (CO1, K3) 10

X	0	1	2	3	4
Y	1	4	10	17	30

5. Answer any one of the following:-

5-a. State and prove Bayes theorem. 10

In a Neighbourhood, 90% children were falling sick due flu and 10% due to measles and no other disease. The probability of observing rashes for measles is 0.95 and for flu is 0.08. If a child develops rashes, find the child's probability of having flu. (CO2, K3)

End Semester Question Paper (CO2)

5-b. Let the two dimensional continuous random variable(X,Y) has joint PDF given by 10

$$f(x,y) = \begin{cases} 6x^2y\,, & 0<x<1,\ 0<y<1 \\ 0\,, & \text{elsewhere} \end{cases}$$

Find (i) $P(0<x<3/4\,,\ 1/3<y<2)$ (ii) $P(x+y<1)$. (CO2, K3)

6. Answer any one of the following:-

6-a. A manufacturer of envelopes knowns that the weight of the envelopes is normally distributed with mean 1.9gm and variance 0.01 square gm. Find how many envelopes weighing (i)2gm or more (ii)2.1gm or more, can be expected in a given packet of 1000envelopes? Given that the area under the standard curve **between $z=0$ and $z=1$ is 0.3413,** **between $z=0$ and $z=2$ is 0.4772.** (CO3,K3) 10

6-b. 10

Four coins were tossed 200 times. The number of tosses showing 0,1,2,3 and 4 heads were found to be as under.

Fit a binomial distribution to these observed results. Find the expected frequencies. (CO3, K3)

No. of Heads	0	1	2	3	4
No. of Tosses	15	35	90	40	20

End Semester Question Paper (CO2)

7. Answer any one of the following:-

7-a. 10

To test of significance of the variations of the retail prices in the commodity in three principal cities: Mumbai, Bangalore and Chennai. The four shops were chosen at random in each city and prices observed in INR were as follows:

Mumbai	16	8	12	14
Bangalore	14	10	10	6
Chennai	4	10	8	8

Do the data indicate that the prices in the three cities are significantly different? Given that the tabular value of F is 4.26 5% LOS with d.f. is (2,9). (CO4,K3)

7-b. From the following data, find whether hair color and gender are associated. 10

End Semester Question Paper (CO2)

Gender↓	Colour					
	Fair	Red	Medium	Dark	Black	Total
Boys	529	849	504	119	36	2100
Girls	544	677	451	97	14	1783
Total	1136	1526	955	216	50	3883

Given that the tabular value of $\chi 2$ is 9.488 at 5%LOS with d.f. 4. (CO4, K3)

8. Answer any one of the following:-

8-a. Solve the following: (CO5,K3) 10

1. What is the sum of all five-digit numbers formed by 2, 3, 4, 5, 6 without repetition?
2. What is the sum of all five-digit numbers formed by 2, 3, 4, 5, 6 with repetition?

End Semester Question Paper (CO2)

8-b. Study the following table and answer the questions based on it 10

Expenditures of a Company (in Lakh Rupees) per Annum Over the given Years.

Year	Item of Expenditure				
	Salary	Fuel and Transport	Bonus	Interest on Loans	Taxes
1998	288	98	3	23.4	83
1999	342	112	2.52	32.5	108
2000	324	101	3.84	41.6	74
2001	336	133	3.68	36.4	88
2002	420	142	3.96	49.4	98

i) Find the average amount of interest per year which the company had to pay during this period?

ii) The total amount of bonus paid by the company during the given period is approximately what percent of the total amount of salary paid during this period?
iii) Total expenditure on all these items in 1998 was approximately what percent of the total expenditure in 2002?
iv) The total expenditure of the company over these items during the year 2000 is?
v) The ratio between the total expenditure on Taxes for all the years and the total expenditure on Fuel and Transport for all the years respectively is approximately?
(CO5,K3)

References

Text Books

- N. P. Bali: A Textbook of Engineering Mathematics-IV, University Science Press.
- H. K. Dass: Introduction to Engineering Mathematics, S. Chand.
- 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

- P. G. Hoel, S. C. Port and C. J. Stone, Introduction to Probability Theory, Universal Book Stall, 2003(Reprint).
- 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