MASTER OF COMPUTER SCIENCE AND ENGINEERING
First Year, Odd Semester Examination, 2025
STATISTICS FOR DATA SCIENCE

Time- Three Hours **Full Marks-100**

All the answers under same "Section" number should be answered together

Section 1: Answer any "Four" questions [CO1]

1. 4 X 5 =20

a. The numbers of calls received in 245 successive one minute intervals at a telephone exchange are shown in the following frequency distribution. Evaluate the mean, median and mode.

No. of Calls	0	1	2	3	4	5	6	7	Total
Frequency	14	21	25	43	51	40	39	12	245

b. Draw histogram, frequency polygon and ogives (less than type) for the following frequency distribution:

Wages (Rs.)	500-590	600-690	700-790	800-890	900-990	1000-1090	1100-1190
No. of Agents	16	20	32	28	20	10	4

c. The ages of twenty husbands and wives are given below. Form a two way frequency table showing the relationship between the ages of husbands and wives with class intervals 20-25 and 25-30 etc.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Age of Husband	28	37	42	25	29	47	37	35	23	41
Age of Wife	23	30	40	26	25	41	35	25	21	38
	11	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
Age of Husband	27	39	23	33	36	32	22	29	38	48
Age of Wife	24	34	20	31	29	35	23	27	34	47

d. Use actual mean method and determine Karl Pearson's coefficient of correlation for the following data

X	10	12	14	16	18	20
Y	5	8	11	14	16	18

e. Find a suitable measure of skewness for the following distribution.

Annual Sales (Rs. '000)	0-20	20-50	50-100	100-250	250-500	500-1000
No. of Firms	20	50	69	30	25	19

f. A random variable has the following probability distribution. Find the expectation and standard deviation of random variable.

X	4	5	6	8
Probability	0.1	0.3	0.4	0.2

Section 2: Answer any "Four" questions [CO2][CO3]

2. a. What do you mean by skewness and kurtosis? How they are related to moments?
b. If the first four moments of a distribution about the value 3 are 2, 10, 40 and 218 respectively, find the moments about the origin and mean.

(3+2)+10=15

3. a. What do we mean by multiple linear regressions? Give an illustration with few sample data.
b. Fit a straight line by the method of least squares to the following data and estimate the probable number of days absent when age is 40 years.

Age	21	42	38	64	53	61	47
Absence (#days)	4	14	10	38	19	34	17

(3+2)+10=15

4. a. If two variables are independent, their correlation coefficient is zero. Is the converse true? Explain by means of an example.
b. Find the coefficient of correlation from the following data:

X	65	63	67	64	68	62	70	66
Y	68	66	68	65	69	66	68	65

(3+2)+10=15

5. a. What do we mean by rank correlation coefficient? Why do we prefer such a coefficient? What are the use cases?
   b. In a certain examination of 10 students obtained the following marks in ML and NLP. Find Spearman's rank correlation coefficient. How does it differ from Pearson's product-moment correlation coefficient?

Roll No.	1	2	3	4	5	6	7	8	9	10
ML	90	30	82	45	32	65	40	88	73	66
NLP	85	42	75	68	45	63	60	90	62	58

(2+1+2)+10=15

6. a. What are the advantages and disadvantages of stratified sampling? What are the error cases in test of significance?
   b. The mean yield of rice from a Jolagor district was 210 *lbs*. with standard deviation 10 *lbs*. per acre from a sample of 100 plots. In another Noyuna district, the mean yield was 220 *lbs* with S.D = 12 *lbs*. from a sample of 150 plots. Assuming that the standard deviation of yield in the entire state was 11 *lbs*., test whether there is any significant difference between the mean yield of crops in the two districts.

(3+2)+10=15

7. a. Write down some important properties of Poisson distribution. Which one is better between Cosine and Euclidean distances and why?
   b. A sample of 100 dry battery cells tested to find the length of life produced the following results: $x' = 12$ hours, $\sigma = 3$ hours. Assuming that the data are normally distributed, what percentage of battery cells are expected to have life (i) more than 15 hours, (ii) less than 6 hours, and (iii) between 10 and 14 hours?

Given	z	2.5	2	1	0.67
	Area	.4938	.4772	.3413	.2487

(3+2)+10=15

Section 3: Answer any "Five" questions [CO4][CO5]

8. 5 X 4 = 20

a) What is the significance of using Lorenz curve in statistics?
b) What are the limitations of using Range and Quartiles in statistics?
c) What is a Scatter Diagram? Explain how this can be used to indicate the degree and type of association between two variables.
d) What do you mean by explained and unexplained variations? How they are related to each other?
e) What do you mean by maximum likelihood estimate? How does it differ from other estimates?
f) Define "Coefficient of Variation". What are the special uses of this measure?
g) What is a semi-logarithmic graph and why do we use it? What are the differences between natural scale and log scale in case of graphical presentation of data?
h) Why do we use covariance? Explain with a suitable example.