AI235AT

USN | 1 | R | V | 2 | 3 | A | I | C | A | 7 |

RV COLLEGE OF ENGINEERING®

(An Autonomous Institution Affiliated to VTU)

III Semester B. E. Regular / Supplementary Examinations Feb/Mar-2025

Artificial Intelligence and Machine Learning

STATISTICS FOR DATA SCIENCE

Maximum Marks: 100

Time: 03 Hours

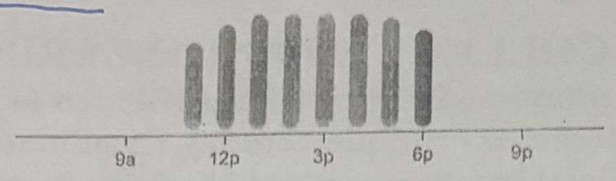
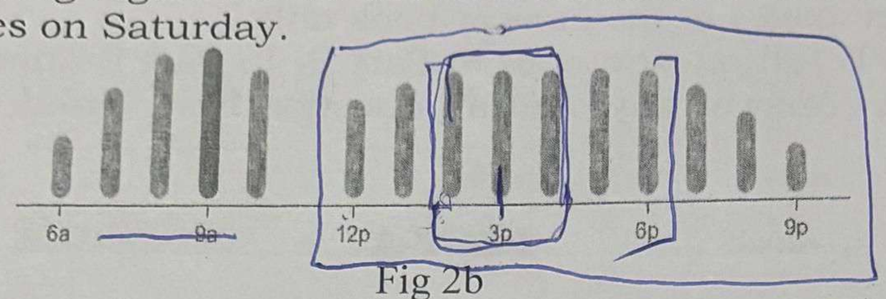
Instructions to candidates:

1. Answer all questions from Part A. Part A questions should be answered in first three pages of the answer book only.
2. Answer FIVE full questions from Part B. In Part B question number 2 is compulsory. Answer any one full question from 3 and 4, 5 and 6, 7 and 8, 9 and 10.

PART-A

			M	BT	CO
1	1.1	A sample of 99 weights has a mean of $24\,Kg$ and median of $24.5\,Kg$. Unfortunately, it was just discovered that a measurement was erroneously recorded as $30Kg$, which was actually $35\,Kg$. Which of the following statements are $TRUE$? Statement 1: Mean remains the same, median is increased Statement 2: Mean is increased, median remains the same Statement 3: We need more information to make any comment	02	2	1
	1.2	It is known that a Box-plot is a five number summary of a set of data. List the five numbers that every Box-plot summarizes.	02	2	2
	1.3	A sample of size $N = 25$ is drawn from an infinite population to estimate its mean μ. What is the distribution does the standardized sample mean follow?	02	2	3
	1.4	You want to sample student opinions about a proposed change in procedures for changing subjects for your electives. You hand out questionnaires to 100 students as they arrive for class at $7.30\,A.M.$ What is wrong with this sampling procedure?	02	3	3
	1.5	For what value of k, can the following be a valid joint PMF? $p(x,y) = \frac{k}{ky}$, $x = -1-2, y = 1,3$ and $p(x,y) = 0$ for all other values of x and y not listed above.	02	2	1
	1.6	Is the following statement is $TRUE$? Reason out your choice. Is X and Y are uncorrelated continuous random variables then the joint PDF $f_{x,y}(x,y) = f_x(x)f_y(y)$, where $f_X(x)$ and $f_Y(y)$ are the marginal PDFs of X and Y.	02	2	1
	1.7	Sketch and label a 82% confidence interval for a standard normal curve. You must clearly indicate the corresponding z values.	02	2	2
	1.8	How many times should the sample size be increased in order to cut the margin of error by 50%?	02	2	3
	1.9	A transport authority official says that 70% of men pass their drivers test in the first attempt, while 65% of women pass the test in the first attempt. It is of interest to know whether these proportions are equal. What are the null and alternate hypotheses?	02	2	3
	1.10	An automobile company develops a new types of seat belts, and it must be subject testing before receiving the necessary permission from the authorities. Suppose the null hypothesis is "the belt is unsafe". What is the Type II Error?	02	2	3

PART-B

2	a	The busy times shown in Fig. 2a correspond to that of an amusement park located on the outskirts of Bangalore.  Fig 2a Based on the busy times plot given, what inference can you make about the working hours of the amusement park? Justify your answer in one or two sentences.	04	3	2
	b	The following Fig.2b shows the *MTR* restaurant near Lal Bagh's busy times on Saturday.  Fig 2b Can we infer that the most preferred time for lunch at *MTR* is 3 *pm* on Saturdays? Justify your answer in one or two sentences.	04	3	2
	c	Find the correlation between the city and the number of private vehicles plying in the city. These numbers given are approximate only. *City* | *Bangalore* | *Chennai* | *Delhi* | *Kolkata* | *Mumbai* *Number of Vehicles (in Lakhs)* | 23.1 | 7.3 | 20.6 | 12 | 25	04	3	2
	d	Define median absolute deviation.	04	2	1
3	a	Suppose that in a population of voters in a certain region 38% are in favor of a particular bond issue. Nine hundred randomly selected voters are asked if they favor the bond issue. i) Verify that the sample proportion $\hat{p}$ computed from samples of size 900 meets the condition that its sampling distribution be approximately normal.	08	3	3
		ii) Find the probability that the sample proportion computed from a sample of size 900 will be within 5 percentage points of the true population proportion.	08	3	3
		OR			
4	a	Let $\underline{X}$ be the mean of a random sample of size 50 drawn from a population with a mean of 112 and a standard deviation of 40. i) Find the mean and standard deviation of $\underline{X}$.	02	3	4
		ii) Find the probability that $\underline{X}$ assumes a value between 110 and 114.	07	3	4
		iii) Find the probability that $\underline{X}$ assumes a value greater than 113.	07	3	4
5	a	Given the joint PDF of two continuous random variables Y_1, Y_2 $f_{y1,y2}(y_1, y_2) = \frac{16y_2}{y_1^3} \quad y_1 > 2, 0 < y_2 < 1.$ Prove or disprove the random variables are independent.	08	3	4
	b	Why cannot a matrix $A^{7\times5}$ be a valid covariance matrix?	04	3	4
	c	Two random variables which are uncorrelated *CAN* be *independent. Is the previous statement is true?* Reason out your answer.	04	3	4

			OR				
6	a		List the properties of a covariance matrix.	06	2	2	
	b		If X and Y are two random variables, obtain the $Var(X-Y)$ value.	06	2	2	
	c		State Central limit theorem.	04	2	2	
7	a		To estimate the proportion of students at a large college who are female, a random sample of 120 Students is selected. There are 69 female students in the sample. Construct a 90% confidence interval for the proportion of all students at the college who are female.	08	3	3	
	b		Find the minimum sample size necessary to construct a 99% confidence interval of μ with a margin of error $E = 0.2$. Assume that the population standard deviation is $\sigma = 1.3$.	08	3	3	
			OR				
8	a		The environmental Protection Agency (EPA) is connected about the amounts of PCB, a toxic chemical, in the milk of nursing mothers. In a sample of 20 women, the amounts (in parts per million) of PCB are as follows: $16, 0, 0, 2, 3, 6, 8, 2, 5, 0, 12, 10, 5, 7, 2, 3, 8, 17, 9, 1$ Use these data to obtain a i) 95 percent confidence interval ii) 99 percent confidence interval Of the average amount of PCB in the milk of nursing mothers.	08	3	3	
	b		A school district is trying to determine its student's reaction to proposed dress code. To do so, the school selected a random sample of 50 students and questioned them. If 20 are in favor of the proposal, then i) Estimate the proportion of all students who are in favor. ii) Estimate the standard error of the estimate.	08	3	3	
9	a		Historical data indicate that the mean acidity (pH) level of rain in a certain industrial region in Bangalore is 5.2. To see whether there has been any recent change in this value, the acidity levels of 12 rainstorms over the past year have been measured, with the following results: $6.1, 5.4, 4.8, 5.8, 6.6, 5.3, 6.1, 4.4, 3.9, 6.8, 6.5, 6.3$ Are these data strong enough, at the 5 percent significance level, for us to conclude that the acidity of the rain has changed from its historical value?	10	3	4	
	b		Define the following: i) Null hypothesis ii) Alternative Hypothesis iii) Rejection Region	06	3	4	
			OR				
10	a		A small component in an electronic device has two small holes where another tiny part is fitted. In the manufacturing process the average distance between the two holes must be tightly controlled at 0.02 mm, else may units would be defective and wasted. Many times throughout the day quality control engineers take a small sample of the components from the production line, measure the distance between the two holes, and make adjustments if needed.				

	Suppose at one time four units are taken and the distances are measures as $0.021, 0.019, 0.023, 0.020$ Determine, at the 1% level of significance, if there is sufficient evidence in the sample to conclude that an adjustment is needed. Assume the distances of interest are normally distributed.	10	3	4
b	Define the following: i) Type 1 error and ii) Type 2 error with examples.	06	3	4