COURSE CODE AI235AT

USN

RV COLLEGE OF ENGINEERING
Autonomous Institution affiliated to VTU
III Semester B.E. April 2024 Examinations
DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Statistics for Data Science

(2022 SCHEME)

Time: 03 Hours ***Maximum Marks: 100***

Instructions to candidates:

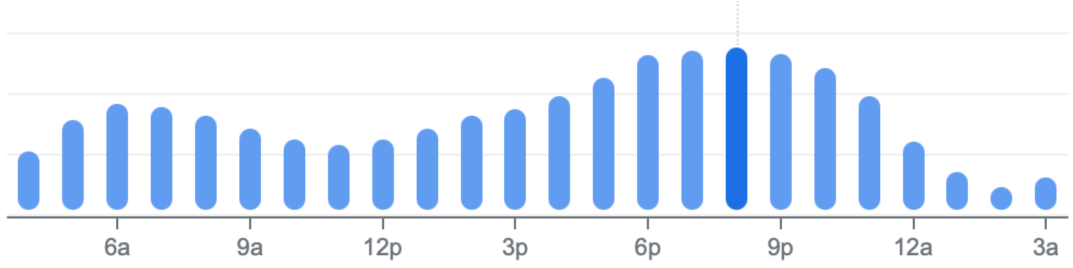
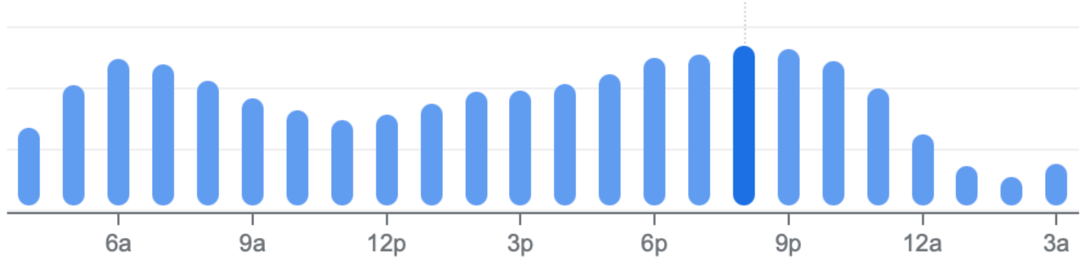
1. Answer all questions from Part A. Part A should be answered in the first three pages of the answer book only.
2. Answer FIVE full questions from Part B. In Part B, question number 2 is compulsory. Answer any one full question from 3 and 4, 5 and 6, 7 and 8, and 9 and 10.

PART-A (Objective type for one or two marks)
(True & false and match the following questions are not permitted)

1	1.1	Identify among the following, the one which has maximum variance and the one has minimum variance. Justify your choice. ***Dataset 1:*** Age of students pursuing 4th semester undergraduate engineering degree in all engineering colleges across India. ***Dataset 2:*** The marks scored by students in JEE exam.	2
	1.2	Illustrate a left skewed distribution with the help of a box plot	2
	1.3	You are looking forward to review a newly released movie. In order to do this, you visit one of the most famous theatres in Bangalore for the first day first show screening. You start randomly choose people to take viewers' opinion about the movie as they exit the movie hall at the end of the movie. Identify the issue with this kind of sampling and the consequences of the same.	2
	1.4	What is the difference between the terms sample and population? Give examples.	2
	1.5	You are given a diagonal matrix of size 100x100, with 50 positive diagonal elements and 50 negative values. Can this matrix be a valid covariance matrix? Justify your answer.	2
	1.6	State the Central Limit Theorem.	2
	1.7	What is a confidence interval of a statistical estimate?	2
	1.8	A random sample is drawn from a population of unknown standard deviation. Construct a 99% confidence interval for the population mean given. $N = 49$, $\overline{X}$=17.1, $\sigma_{\overline{X}}$= 2.1.	2
	1.9	We want to test whether the mean GPA of students from RVCE is 7.0 on a 10 point scale. Write the null hypothesis and alternate hypothesis for the above.	2
	1.10	What is the difference between two-sided and one-sided hypothesis tests?	2

PART-B (Maximum subdivisions is limited to 4 in each question)

		UNIT-I	
2	a	Assert or Reject the following statements with proper justification: (a) **Statement 1:** The average value for any data hides the data diversity information. (b) **Statement 2:** Median of the data depends on the outliers.	4 + 4

	b	 ***Fig 2(a) Popular Time based on visit to Bangalore City Railway Station on Day 1***  ***Fig 2(b) Popular Time based on visit to Bangalore City Railway Station on Day 2*** You are given the popular times based on visit to Bangalore City Railway station on two specific days. Based on the given data, (a) Identify which of the above two figures can represent the distribution corresponding to Saturday. Justify your choice. (b) From both the figures, can we conclude that most people prefer traveling by night trains? Justify your answer.	4 + 4
		UNIT-II	
3	a	A random variable X has an unknown mean μ and standard deviation σ = 2. If the probability that X exceeds 7.5 is .8023, find the unknown mean μ.	6
	b	Suppose the mean length of time that a caller is placed on hold when telephoning a customer service center is 23.8 seconds, with standard deviation 4.6 seconds. Find the probability that the mean length of time on hold in a sample of 1,200 calls will be within 0.5 second of the population mean.	6
	c	Give two scenarios each as to when you will employ a) convenience sampling b) cluster sampling	4
		OR	
4	a	Scores on the common final exam given in a large enrollment multiple section course were normally distributed with mean 69.35 and standard deviation 12.93. The department has the rule that in order to receive an A in the course his score must be in the top 10% of all exam scores. Find the minimum exam score that meets this requirement.	6
	b	Suppose the mean number of days to germination of a variety of seed is 22, with standard deviation 2.3 days. Find the probability that the mean germination time of a sample of 160 seeds will be within 0.5 day of the population mean.	6
	c	Describe the terms (a) population, (b) sample (c) sampling distribution of the sample mean and (d) standard error.	4

UNIT-III			
5	a	List out the properties that a matrix has to satisfy in order to be a valid variance-covariance matrix	4
	b	Establish clearly if each of the following matrix can be a valid variance-covariance matrix or not. (I) $M_1 = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ (ii) $M_2 = \begin{pmatrix} 2 & 1 \\ 0 & -1 \end{pmatrix}$ (iii) $M_3 = \begin{pmatrix} 4 & 3 & 3 \\ 3 & 4 & 3 \end{pmatrix}$ $M_4 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	12
		OR	
6	a	Let (X, Y) be a pair of discrete random variables taking values (0, 0), (1, 1) and (2, 0) with equal probability. (i) Find E[XY] (ii) Find E[X] (iii) Find E[Y] (iv) Find covariance Cov(X, Y). (v) Which of the following is TRUE about X and Y. Justify your answer with logical reasons. **Statement 1**: X and Y are independent **Statement 2:** X and Y are uncorrelated.	10
	b	X and Y are random variables with E[X] = E [Y] = 0 such that X has standard deviation σ_X= 2 while Y has standard deviation σ_Y= 4. (a) For V = X - Y, what are the smallest and largest possible values of Var[V]? (b) For W = X-2Y, what are the smallest and largest possible values of Var [W]?	6

UNIT-IV			
7	a	A random sample of n 50 observations from a quantitative population produced $\overline{X}$=56.4 and $\sigma^2_{\overline{X}}$ =2.6. Give the best point estimate for the population mean μ, and calculate the margin of error.	8
	b	A random sample of 985 "likely" voters—those who are likely to vote in the upcoming election—were polled during a phone-athon conducted by the Republican Party. Of those surveyed, 592 indicated that they intended to vote for the Republican candidate in the upcoming election. Construct a 90% confidence interval for p, the proportion of likely voters in the population who intend to vote for the Republican candidate. Based on this information, can you conclude that the candidate will win the election?	8
		OR	
8	a	A government agency was charged by the legislature with estimating the length of time it takes citizens to fill out various forms. Two hundred randomly selected adults were timed as they filled out a particular form. The times required had mean 12.8 minutes with standard deviation 1.7 minutes. Construct a 90% confidence interval for the mean time taken for all adults to fill out this form.	5
	b	In a random sample of 250 employed people, 61 said that they bring work home with them at least occasionally. a. Give a point estimate of the proportion of all employed people who bring work home with them at least occasionally. b. Construct a 99% confidence interval for that proportion.	5
	c	Define the following terms: (i) Point Estimate (ii) Interval Estimate (iii) Confidence Interval	6

UNIT-V			
9	a	Define the two hypotheses used in binary hypothesis testing. Also, describe the two types of errors that may occur during hypothesis testing. Suppose you want to find out whether the new type of automobile fuel freshly released in the market is efficient or not. What would be the two hypotheses? And what type of test will you choose? i.e., one sided or two sided?	8
	b	Authors of a computer algebra system wish to compare the speed of a new computational algorithm to the currently implemented algorithm. They apply the new algorithm to 50 standard problems; it averages 8.16 seconds with standard deviation 0.17 second. The current algorithm averages 8.21 seconds on such problems. Test, at the 1% level of significance, the alternative hypothesis that the new algorithm has a lower average time than the current algorithm.	8
		OR	
10	a	A population has a mean is 25 and a standard deviation of five. The sample mean is 24, and the sample size is 27. What distribution should you use to perform a hypothesis test? In a population of fish, approximately 42% are female. A test is conducted to see if, in fact, the proportion is less. State the null and alternative hypotheses.	4 + 4
	b	You flip a coin and record whether it shows heads or tails. You know the probability of getting heads is 50%, but you think it is less for this particular coin. What type of test would you use? Given a sample of 50 tosses of the coin with sample mean (for number of heads) is 0.46, perform a hypothesis testing, at 1% level of significance, the alternative hypothesis.	8

Signature of Scrutinizer: Signature of Chairman

Name: Name