USN 1RV22AI007

RV COLLEGE OF ENGINEERING®

(An Autonomous Institution Affiliated to VTU)

III Semester B. E. Examinations April/May-2024

Artificial Intelligence and Machine Learning

STATISTICS FOR DATA SCIENCE

Time: 03 Hours | *Maximum Marks: 100*

Instructions to candidates:

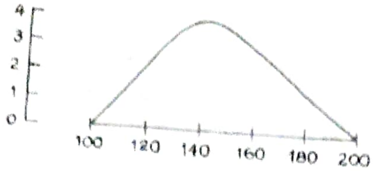
1. Answer all questions from Part A. Part A questions should be answered in first three pages of the answer book only.
2. Answer FIVE full questions from Part B. In Part B question number 2 is compulsory. Answer any one full question from 3 and 4, 5 and 6, 7 and 8, 9 and 10.
3. Students are allowed to use formula handbook of statistics for the data science.

PART-A

			M	BT	CO
1	1.1	If the correlation between two random variables is close to 1, What can we conclude about the two random variables?	02	2	2
	1.2	A scatter plot is unsuitable when the data points are extremely large. Logically reason out in not more than a sentence if the previous statement is correct.	02	2	2
	1.3	If X is a random variable normally distributed with mean 5 and variance 25, what is probability of the random variable X taking a value 5?	02	1	2
	1.4	State Central limit theorem.	02	1	2
	1.5	Can the following matrix be a valid covariance matrix? Justify your answer in a sentence. $\begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix}$	02	2	2
	1.6	If X and Y are independent random variables with variance $Var[X] = 5$ and $Var[Y] = 3$, find the variance of the random variable $Z = -2X + 4Y - 3$.	02	1	2
	1.7	How do we define a large sample when we carry out a large sample estimation of a population proportion, p?	02	1	2
	1.8	Define confidence interval.	02	1	2
	1.9	Define type-1 error in statistical hypothesis testing.	02	1	2
	1.10	What is the difference between a parameter and a statistic?	02	1	2

PART-B

2	a	Facebook data indicate that 50% of Facebook users have 100 or more friends, and that average friend count of users is 190. Based on this information, which of the following shapes best describes the distribution of the number of friends of Facebook users? Justify your choice in not more than one or two sentences. i) Left skewed ii) Right Skewed iii) Approximately symmetric.	04	3	1
	b	Is the interquartile range affected by outliers? Justify your answer in not more than 2 sentences.	04	2	4

	c	Given below is a histogram plotted by someone on a density scale. For all practical purposes, assume that the histogram is a triangle with height = 4 units. Do not worry about what are x-axis and y-axis. Is there anything wrong with the histogram? Justify your answer.  	04	3	2
	2d	Under what circumstances does trimmed mean make sense over the actual Mean? Answer in not more than 2 sentences.	04	2	4
3	a	On average, travelling between two university campuses in a city via shuttle bus takes 28 minutes with a standard deviation of 5 minutes. In a given week, a bus transported passengers 40 times. What is the probability that the average transport time, i.e., the average for 40 trips, was more than 30 minutes? Assume the meantime is measured to nearest minute.	06	3	1
	b	The IQs of 600 applicants to a certain college are approximately normally distributed, with a mean of 115 and a standard deviation of 12. If the college requires an IQ of at least 95, how many of these students will be rejected on this basis of IQ, regardless of their other qualifications? Note that IQs are recorded to the nearest integers.	06	3	3
	c	Airline companies are interested in the consistency of the number of babies on each flight so that they have adequate safety equipment. Suppose an airline conducts a survey. It chooses a specific long weekend, Which includes a Friday and next Monday as holidays, and it surveys six flights from Bangalore to Srinagar to determine the number of babies on the flights. The result of that study determines the amount of safety equipment needed .List exactly 3 things wrong with the way the survey was conducted.	04	3	4
		OR			
4	a	A multiple-choice quiz has 200 questions, each with 4 possible answers, of which only 1 is correct. What is the probability that sheer guesswork yields from 25 to 30 correct answers for 80 of the 200 problems the student does not know? Hint: Use a continuity correction factor of 0.5	06	3	1
	b	Assume SAT scores are normally distributed with a mean of 1518 and a standard deviation of 325. i) If one SAT score is randomly selected, find the probability between 1440 and 1480. ii) If 16 SAT scores are randomly selected, find the probability that they have a mean between 1440 and 1480.	06	3	3
	c	Suppose a newspaper company surveys to estimate whether or not the city's residents favor the new toll road. They take a sample of 100 subscribers and sends them a questionnaire that asks, "Do you think having to pay money to drive on the highway is fair?" And after analyzing the results from the 13 people who replied, the newspaper reports that 74% of the city's residents oppose the new toll road. List exactly 4 problems with this survey in not more than 4 sentences (one sentence per problem).	04	3	4

5	a	You are told that the determinant of a 2×2 real symmetric matrix is 15.With this information, is it possible to conclude that the matrix is valid covariance matrix? Justify your choice in not more than 2 sentences.	06	3	3
	b	Two fair six-sided dice are rolled (one green and one orange), with outcomes X and Y respectively for the green and the orange. i) Compute the covariance of $X+Y$ and $X-Y$ ii) Are $X+Y$ and $X-Y$ independent? Show that they are, or that they are not(whichever is true)	10	3	3
		OR			
6	a	Let X, Y and Z are random variables. The covariance matrix of (X, Y, Z) is given below: $C_{X,Y,Z} = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 4 & -1 \\ 1 & -1 & 4 \end{pmatrix}$ Obtain the covariance matrix of $(X, X+Y, X+Z)$	10	3	3
	b	Establish $Var[X+Y] = Var[X] + Var[Y] + 2Cov[X+Y]$, where X and Y are random variables.	06	3	3
7	a	To estimate μ, the average sugar content of a newly marked oral rehydration salt (ORS) solution, 44 tetra-packs are randomly chosen, and their sugar contents determined. i) If the average sugar finding is 1.74 milligrams, what is a 95 percent confidence interval estimator of μ? ii) How large a sample is necessary for the length of the 95 percent confidence interval to be less than or equal to 0.3 milligrams? Assume that it is known from past experience that the standard deviation of the sugar content of an ORS solution is equal to 0.7 milligrams.	10	3	4
	b	Out of a random sample of 100 students at a university, 82 stated that they were not trained in cooking. Based on this, construct a 99 percent confidence interval estimate of p, the proportion of all the students at the university who are not trained in cooking.	06	3	4
		OR			
8	a	A poll conducted on students indicated 46 percent of the population was in favor of the course on Statistics for Data Science, with a margin of error of ±3 percent. What does this mean? Can we infer how many people were questioned? Assume 95% confidence interval.	10	3	4
	b	Two chemists working for a fast food company, have been producing a very popular sauce. Let's call them Jesse and Mr.White. Gus, their boss, is tired of Mr.White's negative attitude and is thinking about "firing" him and keeping only Jesse on payroll .The problem, however, is that Mr.White seems to produce a higher quality sauce whenever he is in charge of production compared to Jesse. Before making a final decision, Gus collected some data measuring the quality of different batches of sauce Mr>White and Jesse produced. The result, measured on a quality scale, are listed below:			

<table>
<tr><td></td><td></td><td><table><tr><th></th><th><i>Average</i></th><th><i>Standard Deviation</i></th><th><i>Sample size</i></th></tr><tr><td><i>Mr.White</i></td><td>97</td><td>1</td><td>7</td></tr><tr><td><i>Jesse</i></td><td>94</td><td>3</td><td>10</td></tr></table>Based on this data, can we know which is the better chemist is? Use 95% confidence interval to prove your claim.</td><td>06</td><td>3</td><td>4</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>9</td><td>a</td><td>The price of a popular tennis racket at a national chain store is $179. Portia bought five of the same racket at an online auction site for the following
Prices :155,179,175,175,161
Assuming that the auction prices of rackets are normally distributed, determine whether there is sufficient evidence in the sample, the 5% level of significance, to conclude that the average price of the racket is less than $179 if purchased at an online auction.</td><td>10</td><td>3</td><td>3</td></tr>
<tr><td></td><td>b</td><td>The daily yield for a local chemical plant has average 880 tons for the last several years. The quality control manager would like to know whether this average has changed in recent months. She randomly selected 50 days from the computer database and computes the average and standard deviation of the $n = 50$, yielding an average of 871 tons and a standard deviation of 21 tons, respectively. Write the null hypothesis, the alternative hypothesis and the test statistic.</td><td>06</td><td>3</td><td>3</td></tr>
<tr><td></td><td></td><td><b>OR</b></td><td></td><td></td><td></td></tr>
<tr><td>10</td><td>a</td><td>The average weekly earnings for female social workers is $670. Do men in the same positions have average weekly earnings that are higher than those for women? A random sample of $n = 40$ male social workers showed the sample mean to be $725 and the standard deviation to be $102.Test the appropriate hypothesis using $\alpha = 0.01$ and clearly state your inference.</td><td>10</td><td>3</td><td>3</td></tr>
<tr><td></td><td>b</td><td>An election was challenged in court because there was suspicion of fraudulent use of absentee ballots. The judge had to decide whether to overturn the election and remove the winner from office and would do so if he believed the hypothesis that fraud was involved. Although there were other possible reasons why the votes might differ from past elections, to simplify matters, suppose fraud was the only possible reason. Therefore, the judge must decide between:
$H0$:No fraud was involved
Ha:Fraud was involved
The “summary statistic” used to test these hypotheses was the difference between Party A and Party B votes in the absentee ballots. Expert witnesses presented evidence (based on the results from prior elections and votes by machine ballot)that the p-value testing these hypothesis is 0.06.
i) Describe what a Type 1 error would be and the consequences in this situation.
ii) Describe what a Type 2 error and its consequences would be in this situation</td><td>06</td><td>3</td><td>3</td></tr>
</table>