21AI41

USN | 1 | R | V | 2 | 1 | A | I | 0 | 5 | 7 |

RV COLLEGE OF ENGINEERING®

(An Autonomous Institution Affiliated to VTU)

IV Semester B. E. Examinations Oct/Nov-2023

Artificial Intelligence and Machine Learning

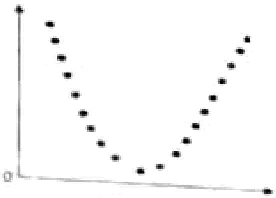
STATISTICS FOR DATA SCIENCE

Time: 03 Hours *Maximum Marks: 100*

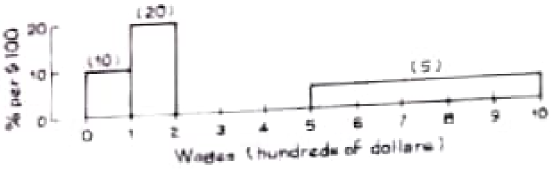
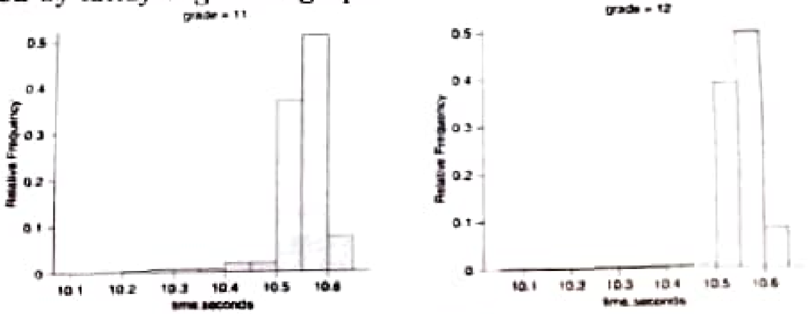
Instructions to candidates:

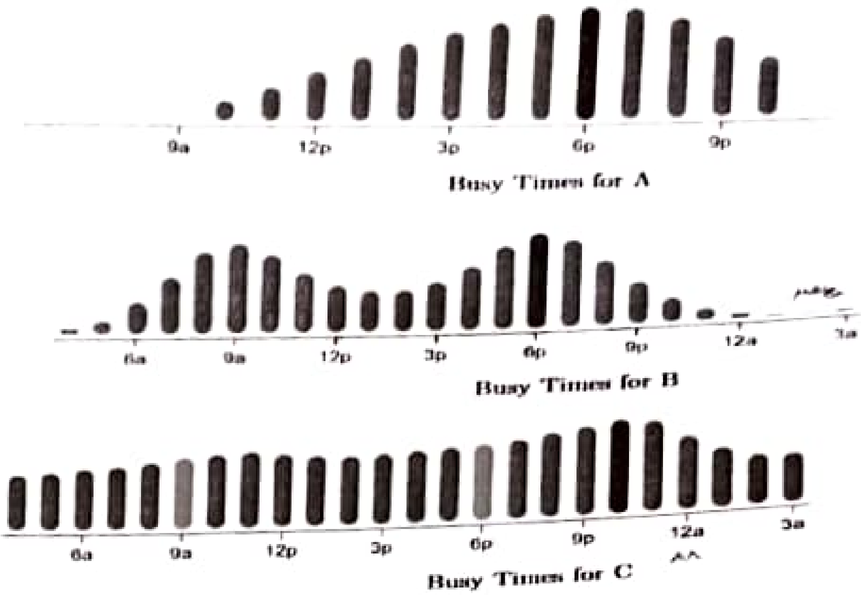
1. Answer all questions from Part A. Part A questions should be answered in first three pages of the answer book only.
2. Answer FIVE full questions from Part B. In Part B question number 2 is compulsory. Answer any ONE full question from 3 and 4, 5 and 6, 7 and 8, 9 and 10.

PART-A

1.1	Suppose we have a set of data in which there are certain data values beyond 3 standard deviations. What can we conclude about the spread of the data?	02
1.2	With respect to the scatterplot given in Fig 1.2 below, assert or reject the following statement with proper justification. Statement: The linear correlation coefficient corresponding to the scatterplot is 0.  Fig 1.2	02
1.3	Choose a household in Bangalore at random and let X denote the number of pets they own. Given below is the probability distribution. Find the missing probability value, denoted as p.	02
1.4	What details about data does a boxplot give?	02
1.5	Suppose X and Y are two random variables such that the sum of X and its corresponding Y is always 100. With this information, can you conclude on the value of correlation coefficient between X and Y? If yes, what is the value? If no, what more information do you need to obtain the correlation coefficient value?	02
1.6	State the law of large numbers.	02
1.7	The following are the number of steps walked in each of the last 7*days*. 6822, 5333, 7420, 7432, 6252, 7005, 6752 Assuming that the daily number of steps can be thought of as being independent realizations from a normal distribution, give a prediction interval that with 95*percent* confidence will contain the number of steps that will be walked tomorrow.	02
1.8	Given the following sample data: 2, 3, 4, 5, 6, 8, 9, 10, 12, 15, obtain the 20% trimmed mean.	02
1.9	For what value of c will the function correspond to a probability density function of continuous random variable X? $f_X(x) = \begin{cases} c(1-x) & 0 \le x \le 1 \\ 0 & otherwise \end{cases}$	02
1.10	Suppose the null hypothesis, $H0$, is: The victim of an automobile accident is alive when he arrives at the emergency room of a hospital. Write Type-1 and Type-2 errors.	02

PART-B

	a	Reason out as to why the following is a bad measure of the variability of data. $\sum(x-\bar{x})$	04
	b	Explain why the sample standard deviation or variance is calculated using $(N-1)$ in the denominator, where N is the number of samples.	04
	c	For an exam given to a class, the students' scores ranged from 35 to 98 with a mean of 74. Which of the following is the most realistic value for the standard deviation: 10, 1, 12, 60? Clearly explain what is unrealistic about the other values.	04
	d	You are told that the average marks scored in a test by a class is 30 on 50, with the marks clustered between 24 to 36. List exactly 4 conclusions that you can draw from the given information.	04
3	a	In the histogram shown in Fig 3a, monthly wages for part-time employees is shown with the numbers in parenthesis corresponding to the densities. Nobody earned more than $1000 a month. However, the block over the class interval from $200 to $500 is missing. How tall must the block be?  Fig 3a	04
	b	It is estimated that 50% of emails are spam emails. Some software has been applied to filter these spam emails before they reach your inbox. A certain brand of software claims that it can detect 99% of spam emails and the probability for a false positive (a non-spam email detected as spam) is 5%. Now if an email is detected as spam, then what is the probability that it is in fact a non-spam email?	06
	c	Given below are two histograms (Refer Fig 3c) corresponding to the time (in seconds), it required for Grade 11 and Grade 12 athletes to complete a 100*metre* race. Write atleast 4 investigative questions that could be answered by analyzing these graphs.  Time taken to complete 100m race Fig 3c **OR**	0
4	a	You are given three charts shown in Fig 4a depicting the busy times, one corresponding to that of Pattanagere metro station and the other corresponding to Bangalore airport and third one unknown. However, the person who extracted these three charts from Google failed to indicate clearly which one corresponds to metro station and which one corresponds to airport and which one is unknown. Your task to identify among the three charts, the one that could be associated with a metro station and the other one that could be associated with the airport. Give 4 valid reasons for each case to support your claim. **Please do not make wild guess. Reason out logically.**	



Busy Times for A

Busy Times for B

Busy Times for C

Fig 4a

12

b Assert or reject the following statement with proper justification. For any two independent events, A and B, $P(A \cap B \mid B) = P(A)$ 04

5 a An automobile battery manufacturer claims that its midgrade battery has a mean life of $50 months$ with a standard deviation of $6 \, months$. Suppose the distribution of battery lives of this particular brand is approximately normal.

i) On the assumption that the manufacturer's claims are true, find the probability that a randomly selected battery of this type will last less than $48 \, months$. 0.3707

ii) On the same assumption, find the probability that the mean of a random sample of 36 such batteries will be less than $48 \, months$. 0.0228 08

b A person eats at the same restaurant every day. Suppose the time X between the moments that person enters the restaurant and the moment that food is served is normally distributed with mean $4.2 \, minutes$ and standard deviation $1.3 minutes$. Find the probability that when the person enters the restaurant today, it will be atleast $5 \, minutes$ until the food is served. 0.26[illegible] 08

OR

6 a An online retailer claims that 90% of all orders are shipped within $12 hours$ of being received. A consumer group placed 121 orders of different sizes and at different times of day; 102 orders were shipped within $12 hours$.

i) Compute the sample proportion of items shipped within $12 \, hours$.

ii) Confirm that the sample is large enough to assume that the sample proportion is normally distributed. Use $p = 0.90$, corresponding to the assumption that the retailer's claim is valid.

iii) Assuming the retailer's claim is true, find the probability that a sample of size 121 would produce a sample proportion as low as was observed in this sample.

iv) Based on the answer to part (iii), draw a conclusion about the retailers claim. 12

b Give the difference between stratified sampling and systematic random sampling. 04

IQ examination scores for sixth-graders are normally distributed with mean value 100 and standard deviation 14.2.

i) Obtain the standardized random variable Z. Why do we do this standardization? 04

ii) What is the probability a randomly chosen sixth-grader has a score greater than 130? 06

iii) What is the probability a randomly chosen sixth-grader has a score between 90 and 115? 06

OR

a The top 5% of applications as measured by a competitive exam (CE) scores will get scholarships. If CE is normally distributed with mean $\mu = 500$ and standard deviation $\sigma = 100$, how high should your competitive exam score be to get scholarship? 08

b The reliability of an electrical fuse is the probability that a fuse chosen at random from production will function under its designed conditions. A random sample of 1000 fuses was tested and $x = 27$ defectives were observed. Calculate the approximate probability of observing 27 or more defectives, assuming that the fuse reliability is 0.98.
Hint: Please read the question carefully. This is approximating binomial using normal distribution and therefore you have to use correction factor of 0.5. 08

a The price of a smart watch in a showroom is $Rs.$ 17900. A person bought five of the same watch on an online discount sale for the following prices.
15500, 17900, 17500, 17500, 16100
Assuming that the discount prices of the watch are normally distributed, determine whether there is sufficient evidence in the sample, at the 5% level of significance, to conclude that the average price of the watch is less than $Rs.$ 17,900 if purchased at an online discount sale. 10

b A coin is tossed 1050 times and lands on heads 500 times. Construct a 90% confidence interval for the probability $'p'$ of getting a head. 06

OR

a A company fills 80 $gram$ containers of moisturizer by a machine. The machine is set to dispense a mean of 81 $gram$ per container. Randomness in the process can shift the mean away from 81 and cause either underfill or overfill. In such a case, the machine is recalibrated. Regardless of the mean amount dispensed, the standard deviation of the amount dispensed always has value 2.2$gram$. A quality control engineer routinely selects 30 containers to check the amounts filled. With one such sample, the mean is $\bar{X} = 82$ $gram$ and the sample standard deviation is $\hat{\sigma} = 2.5 gram$. Find out if there is sufficient evidence in the sample to indicate, at the 1% level of significance, that the machine needs recalibration. 10

b Explain the concept of Type 1 error in hypothesis testing and provide an example scenario where it can occur. 06