|  | **PES University, Bengaluru**  (Established under Karnataka Act No. 16 of 2013) | | **UE20CS902** |
| --- | --- | --- | --- |
| **March 2024: END SEMESTER ASSESSMENT (ESA)**  **M TECH DATA SCIENCE AND MACHINE LEARNING\_ SEMESTER I**  **UE20CS902 – Statistical Methods for Decision Making** | | | |
| Time: 3 Hrs | | Answer All Questions | Max Marks: 100 |

| INSTRUCTIONS |
| --- |
| * All questions are compulsory. * Section A should be handwritten in the answer script provided. * Section B and C are coding questions which have to be answered in the system. * Assume the level of significance as 0.05 * Assume data is normally distributed and of equal variance * For all the questions state the hypothesis statements and business inferences. Each of these carries 2 marks accounting to 4 marks. |

**Section A -20 Marks**

| 1 | a) | Calculate the interquartile range (IQR) for the following dataset: 4, 7, 12, 5, 15, 9, 2. Find the 25th percentile (Q1) and the 75th percentile (Q3) as part of your solution. | 2 |
| --- | --- | --- | --- |
| b) | In a marketing survey, what is the difference between collecting data on customer age (in years) and customer satisfaction level (on a scale of 1-5)? | 2 |
| c) | A bag contains 5 red marbles, 3 blue marbles, and 2 green marbles. If one marble is drawn at random, what is the probability that it is not blue? | 2 |
| d) | The average number of customers arriving at a bank during a 15-minute interval is 3. What is the probability of exactly 5 customers arriving in the next 15-minute interval? | 2 |
| e) | In a drug trial to test the efficacy, what does a Type I error represent? | 2 |
| 2 | a) | In a quality control setting, what does a Type II error represent? | 2 |
|  | b) | What should the typical sample size be, to apply the central limit theorem. What happens for highly skewed distributions? | 2 |
|  | c) | A marketing manager wants to determine if there is a difference in website conversion rates among three different landing page designs. The company randomly assigns website visitors to one of the three landing page designs (A, B, or C) and tracks whether or not they make a purchase (conversion). What is the name of the statistical test to perform? | 2 |
|  | d) | A new medical test is developed to detect a relatively rare disease. Consider the following information:  Disease Prevalence: 1% of the population has the disease.  Test Accuracy:  If a person has the disease, the test correctly identifies it (true positive) 95% of the time.  If a person does not have the disease, the test incorrectly indicates the disease is present (false positive) 3% of the time. Given the following events,  A: The person has the disease.  B: The test result is positive.  Fill in the blanks:  **P(A) =**  **P(A') =**  **P(B|A) =**  **P(B|A') =** | 2 |
|  | e) | A group of test scores has a mean of 75 and a standard deviation of 8. If 7 points are added to every score in the group, what will happen to the mean and standard deviation? | 2 |
| **SECTION B – 40 MARKS** | | | |
| 3 | a) | Determine if the average weight of a product sample differs significantly from the specified target weight of 15 units. Refer notebook. | 8 |
| b) | Investigate if there's a significant difference in customer satisfaction scores between two product versions.  version\_a\_scores = np.array([78, 83, 76, 81, 85])  version\_b\_scores = np.array([80, 88, 91, 84, 79]) | 8 |
|  | c) | Three different groups of salaries depicting three departments are given. Check if there is a significant difference in the average salaries between the groups. Refer notebook for sample data. | 8 |
|  | d) | Check if the experience and the corresponding salaries are correlated.  experience = np.array([5, 8, 12, 6, 10])  salary = np.array([15, 24, 30, 18, 27]) | 8 |
|  | e) | Suppose you have election data from a region as follows:   * 60% of the population identifies as Democrat. * 40% of the population identifies as Republican. * Among Democrats, 90% voted in the recent election. * Among Republicans, 80% voted in the recent election.   If a randomly selected person from this region voted, what is the probability that they are a Democrat ? | 8 |
| **SECTION C – 40 MARKS** | | | |
| 4 | a) | A candy company claims that their bags of assorted candy contain 20% of each of the following flavors: cherry, lemon, orange, grape, and strawberry. You purchase a random bag and count the following:  Cherry: 25, Lemon: 32, Orange: 29, Grape: 18 , Strawberry: 26  Does the data provide enough evidence to suggest that the company's distribution claim is inaccurate? | 10 |
| b) | The average height of a certain population is 68 inches with a standard deviation of 3 inches. Assuming heights are normally distributed, what is the probability that a randomly selected individual will be taller than 72 inches? | 5 |
| c) | A market research company wants to conduct a survey to estimate the proportion of customers who are likely to recommend their product to others. They desire a margin of error of no more than 5% with a 95% confidence level. Based on a small pilot study, they estimate the population proportion to be around 0.6. What is the minimum required sample size for their survey? | 5 |
| 5 |  | Consider the given dataset (student\_data.csv) and answer the following question |  |
| a) | Do female and male students differ on average in their final course grades (G3)? | 10 |
| b) | Is there a difference in the proportion of students who achieve a high final grade (G3 >= 15) between those with at least one parent with higher education (Medu or Fedu >=3) and those without? | 10 |