PES University, Bengaluru (Established under Karnataka Act No. 16 of 2013)	**UE20CS902**

AUG 2023: 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.

Section A -20 Marks

1	a)	Consider the following two samples: Sample 1: 10, 9, 8, 7, 8, 6, 10, 6 Sample 2: 10, 6, 10, 6, 8, 10, 8, 6 Compare the samples in terms of mean, standard deviation	2
	b)	Find the Z score for the value of X =40, where mean = 45, standard deviation is 5	2
	c)	Consider the following ANOVA table, fill in the missing values.	2

Source	Sum of Squares	Degree of Freedom	MS	F
Treatment	1000			
Error		3	166	
Total	1500	27		

	d)	What does ANOVA stand for? State its purpose . Why is it preferred to multiple paired t-tests ?	2
	e)	List the different types of sampling techniques (any 4 four).	2
2	a)	What are the measures of dispersion? Name them.	2
	b)	Explain the usefulness of central limit theorem in sampling distribution and inference.	2
	c)	Consider two dice are rolled simultaneously. Find the probability that the sum of the two numbers is at least 9.	2
	d)	Define p-value. What does a very low p-value indicate ?	2
	e)	Define Bayes theorem and provide with an example where Bayes theorem can be applied.	2

SECTION B – 40 MARKS

3	a)	Based on the given use case of Jim is a real estate agent, (details are given in notebook), predict what is the probability that Jim will make, no sales? One sale? two or more sales? In a 90-day period, what is the probability that Jim will make no sales?	8
	b)	The compressive strength of samples of cement can be modeled by a normal distribution with a mean of 6000 kilograms per square centimeter and a standard deviation of 100 kilograms per square centimeter. i. What is the probability that a sample's strength is less than 6250 Kg/cm2 (2mark) ii. What is the probability that a sample's strength is between 5800 and 5900 Kg/cm2 (2 mark) iii. What strength is exceeded by 95% of the samples (2 mark)	8
	c)	The production manager case study at the Yen Sewing Factory is provided in notebooks. Plot the	8

		acceptance and rejection regions to test the claim of the production manager with 99% confidence.	
		i. State the null hypothesis and the alternate hypothesis. (1 mark) ii. Which test is to be performed. (1 mark) iii. Compute test statistics, p value. (3 marks) iv. At the 0.01 significance level, can we conclude that water consumption has increased? Bobbin diameter is less than 18mm(1 mark)	
	d)	Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process. Specifically, catalyst 1 is currently in use, but catalyst 2 is acceptable. Since catalyst 2 is cheaper, it should be adopted, providing it does not change the process yield. A test is run in the pilot plant and results in the data shown table. Is there any difference in mean yields for an α =.05 and assume equal variances Number 1 2 3 4 5 6 7 8 Catlayst1 91.50 94.18 92.18 95.39 91.79 89.07 94.72 89.21 Catalyst2 89.19 90.95 90.46 93.21 97.19 97.04 91.07 92.75 i. State the hypothesis and type of test to be used (2 marks) ii. Test the hypothesis and conclude (4 marks)	8

	e)	A multinational company use case is provided in the notebook. Test the company's claim that the course was effective in developing soft skills with 90% confidence using the p-value technique.
		
		Read the data set Paired_data.xlsx
		  Consider the scores given in the file `paired_data.xlsx`
		  please consider the first two attributes
		
		1. Check if the attributes are normally distributed (2 marks)
		2. State the null hypothesis and alternate hypothesis (1 mark)
		3. Test the hypothesis and conclude. (3 marks)
			8

SECTION C – 40 MARKS

4	a)	Consider the admission.xlxs file and answer the following questions (Data description is given in notebook)
		  i. Provide a summary statistics of all the variables and explain which variables has high variability (3 mark)
		    List out the numerical and categorical features in the dataset (2 marks)
		    (Total- 5marks)
		  ii. Provide a histogram for the variable salary and Percentage of marks scored in MBA, based on histogram and cacluation of skewness and kurtosis what would you describe. (5 marks)
		  iii. Draw a boxplot for salary based on different MBA specialization, what do you observe in terms of outliers, and salary difference for specialization segments. (5 marks)
		  iv. Draw a scatter plot for all of the variables, what is your observation and conclusion for the relationship between students salary and the other numeric variables (5 marks)
			20

	b)	Consider the purchases.csv file and answer the following questions
		  i. Check whether the percentile in entrance test from the data follows normal distribution by using a proper plot and with suitable statistical test. (5 marks)
		  ii. Conduct a hypothesis to see whether there is a difference in terms of salary for students degree specialization of management and engineering. What do you conclude (consider level of significance 0.05) (5 marks)
		  iii. Conduct an hypothesis to prove whether the salary of a student from commenrce stream is higher than Science stream, do you agree or not. (consider level of significance 0.05) (5 marks)
		  iv. Conduct an ANOVA to check whether there is salary difference for students with diffetent MBA_specialization of Marketing & Finance , Marketing & HR, Operations & Management, Marketing & IB. State your inferences. (consider level of significance 0.05) (5 marks)
			20