June 13, 2024 02:30 pm - 05:30 pm 1T01875 - T.E. Computer Science & Enginering (Artificial Intelligence & Machine Learning) (Choice Based) (R-2019 'C' Scheme) SEMESTER - V / 48895 - Department Optional Course - 1: Statistics for Artificial Intelligence & Data Science
QP CODE: 10056538

Duration: 3hrs **[Max Marks:80]**

N.B. 1. Question No. 1 is compulsory.
2. Attempt any three questions out of remaining five.
3. All questions carry equal marks
4. Assume Suitable data, if required and state it clearly.

1 Attempt any four: 20

(a) Define Confidence Interval?

(b) In a certain property investment company with an international presence, workers have a mean hourly wage of $12 with a population standard deviation of $3. Given a sample size of 30, estimate and interpret the SE of the sample mean.

(c) What is hypothesis testing? Explain type I and type II errors?

(d) What do you mean by correlation and regression? Explain with example.

(e) What is analysis of variance? Explain its usage.

2 (a) X is a normally distributed variable with mean $\mu = 30$ and standard deviation $\sigma = 4$. Find 10

a) $P(x < 40)$
b) $P(x > 21)$
c) $P(30 < x < 35)$

(b) . Some vehicles pass through a junction on a busy road at an average rate of 300 per hour. 10

a. Find out the probability that none passes in a given minute.
b. What is the expected number of passing in two minutes?
c. Find the probability that this expected number found above actually pass through in a given two-minute period.

3 (a) For a certain type of computers, the length of time between charges of the battery is normally distributed with a mean of 50 hours and a standard deviation of 15 hours. John owns one of these computers and wants to know the probability that the length of time will be between 50 and 70 hours. 10

(b) The average score on a test is 80 with a standard deviation of 10. With a new teaching curriculum introduced it is believed that this score will change. On random testing, the score of 38 students, the mean was found to be 88. With a 0.05 significance level, is there any evidence to support this claim? 10

4 a) Explain QQ plots in detail. Show how scatterplots explores relationships between variables. 10

b) Given four samples A, B, C, D. Solve using one-way ANOVA to identify any difference between samples. 10

Observation	A	B	C	D
1	8	12	18	13
2	10	11	12	9
3	12	9	16	12
4	8	14	6	16
5	7	4	8	15

5 a) What is F-Test? If the F statistic as 2.38 and the degrees of freedom obtained by him were 8 and 3. Find out the F value from the F Table and determine whether we can reject the null hypothesis at 5% level of significance (one-tailed test). 10

b) Find the simple linear regression equation that fits the given data and coefficient of determination: 10

X	Y
2	69
9	98
5	82
5	77
3	71
7	84

6 a) Explain Binomial distribution in detail. 10

Bottles of water have a label stating that the volume is 12 oz. A consumer group suspects the bottles are under-filled and plans to conduct a test. What would a Type I error in this situation mean?

b) Write short notes on (any two) 10

1. Chi-square distribution.
2. Weibull distribution.
3. Stem & Leaf Plot
4. Box Plot

****

1T01865 - T.E. Computer Science & Enginering (Data Science) (Choice Based) (R-2019-20'C' Scheme) SEMESTER - V / 48885 - Department Optional Course - 1: Statistics for Artificial Intelligence & Data Science

QP CODE: 10014523 **[Time: 3 Hours]** DATE: 02/12/2022 **[Marks:80]**

N.B. 1. Question No. 1 is compulsory.
2. Attempt any three questions out of remaining five.
3. All questions carry equal marks
4. Assume Suitable data, if required and state it clearly.

Q.1 Attempt any four: **20**

a) Find the standard deviation of the average temperatures recorded over a five-day period last winter: 19, 21, 18, 24, 12?

b) X is a normally distributed variable with mean $\mu = 30$ and standard deviation $\sigma = 4$. Find:
i) $P(x < 40)$, ii) $P(30 < x < 35)$?

c) Discuss Boot strapping vs. re-sampling

d) The school principal wants to test if it is true what teachers say – that high school juniors use the computer an average 3.2 hours a day. What are our null and alternative hypotheses?

e) What do you mean by correlation and regression? Explain with example

Q.2 a) Find the value of the correlation coefficient from the data given in the following table: **10**

SUBJECT	AGE (X)	GLUCOSE LEVEL(Y)
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

b) **10**

Explain briefly why ANOVA is used? Solve using One-way ANOVA

OBSERVATIONS	A	B	C
1	25	31	24
2	30	39	30
3	36	38	28
4	38	42	25
5	31	35	28

method:

Q.3 a) Explain type I & type 2 error in detail. **10**
(ii) What is the use of scatter plot and box plot?

b) In a manufacturing unit, four teams of operators were randomly selected and sent to four different facilities for machining techniques training. After the training, the supervisor conducted the exam and recorded the test scores. At 95% confidence level does the scores are same in all four facilities? **10**
(Hint: Use Kruskal–Wallis test)

Facility 1	Facility 2	Facility 3	Facility 4
88	77	71	52
82	76	56	65
86	84	64	68
87	59	51	81

Q.4 a) If the sample mean and expected mean value of the marks obtained by 15 students in a class test is 290 and 300 respectively. What is the t-score if the standard deviation of the marks is 50? **10**

b) Find out what is the relation between the GPA of a class of students and the number of hours of study and the height of the student **10**

GPA	Height	Study Hours
2.9	66	7
3.16	57	7
3.62	64.5	6
2	62	7
3.45	69.5	8
2.8	65	9
3.63	63	6
2.81	68	5
3.33	59.5	4
2.75	64	10
3.86	69	7

Q.5 a) A farmer is trying out a planting technique that he hopes will increase the yield on his pea plants. The average number of pods on one of his pea plants is 145 pods with a standard deviation of 100 pods. This year, after trying his new planting technique, he takes a random sample of his plants and finds the average number of pods to be 147. He wonders whether this is a statistically significant increase. What are his hypotheses and the test statistic? Use a 0.05 significance level. **10**

b) Find the simple linear regression equation that fits the given data and coefficient of determination: **10**

Hour	Temp
2	21
4	27
6	29
8	86
10	86
12	92

Q.6 a) An agent sells life insurance policies to five equally aged, healthy people. According to recent data, the probability of a person living in these conditions for 30 years or more is 2/3. Calculate the probability that after 30 years if **10**

i. All five people are still living.
ii. At least three people are still living.
iii. Exactly two people are still living. (Hint: Binomial Distribution)

b) Write short notes on (any two) **10**

i. Confidence Interval
ii. Central Limit Theorem
iii. Standard Error

——————————

1T01875 - T.E. Computer Science & Enginering (Artificial Intelligence & Machine Learning) (Choice Based) (R-2019 'C' Scheme) SEMESTER - V / 48895 - Department Optional Course - 1: Statistics for Artificial Intelligence & Data Science QP CODE: 10039067 DATE: 04/12/2023

Duration: 3hrs **[Max Marks:80]**

(1) Question No 1 is Compulsory.
(2) Attempt any three questions out of the remaining five.
(3) All questions carry equal marks.
(4) Assume suitable data, if required and state it clearly.

1 Attempt any **four** **[20]**

a) Write a short note on hypothesis testing.
b) What is Fisher's exact test?
c) Write a short note Simple Linear Regression
d) Write a short note on Random sampling
e) What is the empirical CDF function?

2 a) Construct a frequency distribution table for the following weights (in gm) of 30 oranges using the equal class intervals, one of them is 40-45 (45 not included). The weights are: 31, 41, 46, 33, 44, 51, 56, 63, 71, 71, 62, 63, 54, 53, 51, 43, 36, 38, 54, 56, 66, 71, 74, 75, 46, 47, 59, 60, 61, 63. **[10]**

(a) What is the class mark of the class intervals 50-55?
(b) What is the range of the above weights?
(c) How many class intervals are there?
(d) Which class interval has the lowest frequency?

b) What is the primary purpose of conducting a one-way ANOVA. Explain the key components of a one-way ANOVA, including the dependent variable, independent variable, and factors. **[10]**

3 a) Find the standard error of the estimate for the average number of children in a household in your city by using the data collected from a sample of households in your city. Then find a 95% confidence interval for the data. **[10]**

Household	No. of children
1	2
2	3
3	1
4	0
5	5
6	2
7	1
8	4

b) What is the concept of correlation in statistics, how is it different from regression? **[10]**

4 **a)** A radar unit is used to measure speeds of cars on a motorway. The speeds are normally distributed with a mean of 90 km/hr and a standard deviation of 10 km/hr. What is the probability that a car picked at random is travelling at more than 100 km/hr? **[10]**

b) Explain Numerical and Categorical data types with appropriate examples **[10]**

5 **a)** Duracell manufactures batteries that the CEO claims will last an average of 300 hours under normal use. A researcher randomly selected 20 batteries from the production line and tested these batteries. The tested batteries had a mean life span of 270 hours with a standard deviation of 50 hours. Do we have enough evidence to suggest that the claim of an average lifetime of 300 hours is false? **[10]**

b) Explain linear least square regression (LLSR) along with it's advantages and disadvantages. **[10]**

6 **a)** A farmer is trying out a planting technique that he hopes will increase the yield on his pea plants. The average number of pods on one of his pea plantsis 145 pods with a standard deviation of 100 pods. This year, after trying his new planting technique, he takes a random sample of his plants and finds theaverage number of pods to be 147. He wonders whether or not this is a statistically significant increase. What are his hypotheses and the test statistic? **[10]**

b) What is the Chi-Square Test in statistics, and in what kind of situations or research scenarios is it commonly used? **[10]**

****