BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

Work Integrated Learning Programmes Division

Cluster Programme - M. Tech in AI & ML

I Semester , 2023 – 24(APRIL,2024)

Comprehensive Examination (REGULAR)

Q.1.Consider the following joint probability distribution. **[7 Marks]**

Y / X	-1	-2	3
0	k	2k	2k
1	2k	k	k
2	k	2k	k
3	3k	3k	k

Then find

i) k value

ii) Marginal probability distributions of X,Y

iii) P(X < 2, Y< -3)

iv) P(X < 2 / Y < -3)

v) Are X and Y are independent? Validate.

Sol: i) $\sum f(x,y) = 1$

$\Rightarrow k + 2k + 2k + \cdots + 3k + k = 1$ → 1m

$20k = 1 \Rightarrow k = 1/20$

ii)Marginal probability distributions of X,Y

Y / X	-1	-2	-3	P(x)
0	k	2k	2k	5k=5\20=1\4
1	2k	k	k	4k=1\5
2	k	2k	k	4k=1\5
3	3k	3k	k	7k=7\20
P(y)	7k=7\20	8k=8\20	5k=5\20=1\4	

Marginal probability distribution of Y

Marginal distribution of X → 2m

iii) P(X < 2, Y< -3) $= P(0,-1) + P(0,-2) + P(1,-1) + P(1,-2)$ → 1m

0 , 1 ; -1, -2 $= k + 2k + 2k + k = 6k = 6/20 = 3/10$

iv) P(X < 2 / Y < -3)

$P(X<2|Y<-3) = \frac{P(X<2 \text{ and } Y<-3)}{P(Y<-3)} = \frac{P(X=0,1, \text{ and } Y=-1,-2)}{P(X=-1) + P(Y=-2)} = \frac{P(0,-1)+P(0,-2)+P(1,-1)+P(1,-2)}{7k+8k}$ 2 marks

$= \frac{k+2k+2k+k}{15k} = \frac{6k}{15k} = \frac{2}{5}$

v)Are X and Y are independent? Validate: $P(0,-1) = P(0)\,P(-1)$

$k \neq (5k)(7k)$, ∴ not independent. 1m

Q.2. National Highway Authorities of India (NHAI) proposes to automate the process of quality check in place of existing manual random check before the financial clearance of the bills for the work done by the agencies. NHAI wants to use Deep learning methods in the automated process. The following data gives the number of instances the methods correctly / incorrectly identified the quality of the work as proposed during bidding in DPR (Detailed project report)

Formulate a suitable hypothesis and validate it at 1% Level of significance. [**7 Marks]**

	Correctly identified the quality of the work as per the DPR	**Incorrectly identified the quality of the work as per the DPR**
Manual check	200	220
Automated check	300	280
Total	500	500

Answer:

It is to test

H0: Deep learning method does not significantly influence on correctly identifying the quality checking of the work

(or) There is no significant difference between manual and automated checks in identifying correctly while doing the quality checking of work

(Vs)

H1: Deep learning method significantly influences on correctly identifying the quality checking of the work

(or) There is significant difference between manual and automated checks in identifying correctly while doing the quality checking of work

(2 marks)

Observed frequency (O)	**Correctly identified the quality of the work as per the DPR**	**Incorrectly identified the quality of the work as per the DPR**	**Row total**
Manual check	200	220	420
Automated check	300	280	580
Column Total	500	500	1000=N

Expected frequency (E)= Row total x Column total / N

O	E	(O-E)^2/E
200	210	0.47619048
300	290	0.34482759
220	210	0.47619048
280	290	0.34482759
Total =1000	1000	1.64203612

The calculated Chi-square = SUM[(O-E)^2/E] =1.64203612

(3 Marks)

The critical Chi-square value at 1% los and (r-1)(c-1) = (2-1)(2-1) df is 6.635

Since, 1.64203612 < 6.635 i.e., Calculated value < Critical value

And/ or P = P(Chi-square (1 df) >1.642) = 0.2 > alpha =0.01 then we fail to reject H0 and we conclude that Deep learning method does not significantly influence on correctly identifying the quality checking of the work on the basis of given sample data.

(or) There is no significant difference between manual and automated checks in identifying correctly while doing the quality checking of work

(2 Marks)

(OR)

Let P1 = Proportion of correct quality checking by manual method

P2 = Proportion of correct quality checking by automated method

To test

H0: P1 = P2 (vs) H1: P1<P2 or H1: P1≠ P2

(2 marks)

p1 = 200/420 = 0.476, p2 = 300/580 = 0.517

P = (200+300)/(420+580) = 0.5

Z = (p1-p2)/Sqrt(PQ(1/n1+1/n2)) = -1.32 and |z| = 1.32

(3 marks)

At 1% los, the critical values for one tales and two tailed tests respectively 2.33 and 2.58

Since 1.32<2.33 and 2.58 i.e., calculated value < critical value then Ho is accepted and concluded that P1 = P2 i.e., there is no significant difference between proportions of correct quality checking by manual method and automated method.

(2 marks)

Q.3. A sample of 30 milk carton provides a sample mean of 505 ml. The population standard deviation is believed to be 10 ml. Perform a hypothesis test, at the 0.05 level of significance, population mean 500 ml and to help determine whether the filling process should continue operating or be stopped and corrected.Use p – value also to validate the hypothesis.

[7 Marks]

Given: Sample size = 30 , Sample mean = 505 ml ,

Population mean = 500 ml , Population Standard deviation = 10 ml

Significance level 0.05

$H_0: \mu = 500$
$H_a: \mu \neq 500$ [1 Mark]

$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{505-500}{10/\sqrt{30}} = 2.74$ [2 Mark]

$\frac{\alpha}{2} = \frac{.05}{2} = .025$, Critical value$z_{\alpha/2} = 1.96$ [1 Mark]

As calculated z = 2.74 > Critical value =1.96, We reject H_0 [1 Mark]

There is sufficient statistical evidence to infer that the null hypothesis is not true.

P value approach:

As P(z > 2.74) =0.0031 [1 Mark]

p–value = 2(0.0031) = .0062

Thus we reject H0, as *p*–value = .0062 < significance level = .05 [1 Mark]

Q.4. For the following data production of wheat in tons, calculate the 3 year moving average. Also find the mean square error (MSE) and mean absolute deviation (MAD). **[7 Marks]**

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Production (tons)	86	63	45	58	43	57	98	100	120	150

Key answer:

Year	Production (tons)	3 yearly moving average	Error^2	Abs(Error)
2011	86			
2012	63	64.67	2.778	1.667
2013	45	55.33	106.778	10.333
2014	58	48.67	87.111	9.333
2015	43	52.67	93.444	9.667
2016	57	66.00	81.000	9.000
2017	98	85.00	169.000	13.000
2018	100	106.00	36.000	6.000
2019	120	123.33	11.111	3.333
2020	150			
Sum			587.222	62.333
MSE			73.403	7.792

Marks distribution:

3 yearly moving average – 3 marks

MSE – 2 marks

MAD – 2 marks

Q.5.A state government is studying the number of traffic fatalities (in humdreds) in the state resulting from drunken driving in the last months of 2023. Using simple exponential smoothing, with the smoothing factor 0.6, forecast the traffic fatalities for the 13th month (i.e. first month of 2024). **[6 Marks]**

Month	1	2	3	4	5	6	7	8	9	10	11	12
Traffic fatalities	28	30	28	28	27	24	23	26	21	27	22	35

Key answer: $F_{t+1} = \alpha Y_t + (1-\alpha)F_t$

Month	Traffic fatalities	$F_{t+1} = 0.6Y_t + 0.4F_t$
	α	**0.6**
1	28	
2	30	28
3	28	29
4	28	28
5	27	28
6	24	27
7	23	25
8	26	24
9	21	25
10	27	23
11	22	25
12	35	23
Forecast for 13th month = 30		

Marks distribution:

Simple exponential smoothing – 5 marks

Forecast for 13 period (January 2024 – 1 marks

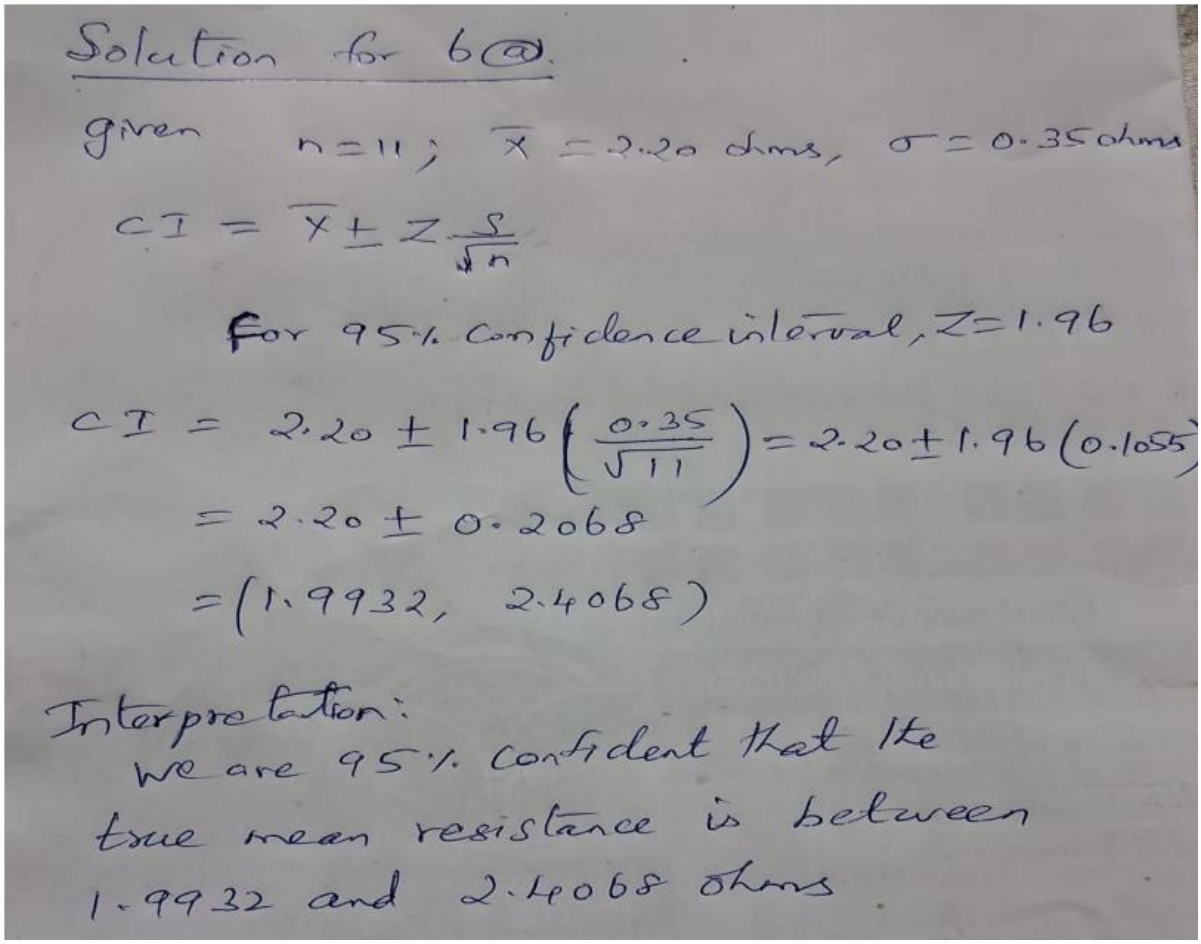
Q.6.a). A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms. Determine a 95% confidence interval for the true mean resistance of the population.

[3 Marks]

b).i).What is the conclusion if the coefficient of correlation is zero?

ii). Let X_1, X_2, X_3 are features and Y is the target variable. Coefficient of Correlation between them are $(Y,X_1) = 0.85$, $(Y,X_2) = 0.60$,$(Y,X_3) = 0.10$, $(X_1, X_2) = 0.95$, $(X_1,X_3) = 0.30$.Then how can you proceed further as a part of pre –processing before fitting linear Regression model. Discuss and validate your actions in this regard.

[3 Marks]



Ans: b(i): If the coefficient of correlation between two variables is zero, it implies that there is no linear relationship between these variables. However, it's important to note that a zero correlation coefficient does not necessarily mean there is no relationship at all between the variables—it just means that the relationship is not linear. **[1 mark]**

Ans b(ii): **[2 marks]**

To preprocess the data before fitting a linear regression model based on the given correlations between the features (X1, X2, X3) and the target variable (Y), as well as between the features themselves, follow these steps:

1). Check feature importance based on correlation with target Y

Given the correlation values, X1 (correlation of 0.85) is highly correlated with Y, followed by X2 (0.60) and X3 (0.10). This suggests that X1 is likely the most important feature for predicting Y.

2). Handle multicollinearity among inputs:

The correlation between X1 and X2 is very high (0.95), indicating strong multicollinearity. Similarly, the correlation between X1 and X3 is moderate (0.30), suggesting some level of correlation that needs to be managed.

3).Transformation or feature selection:

Given the high correlation between X1 and X2, consider keeping only one of these features based on their individual correlation with Y or use dimensionality reduction techniques like PCA to create orthogonal features.

In conclusion, these preprocessing steps, we can enhance the quality and interpretability of the linear regression model, making it more robust against issues such as multicollinearity, outliers, and feature scaling disparities. The specific actions taken should be validated through appropriate statistical analysis and visualization techniques to ensure the effectiveness of the model.

XXXXXXXXX