M.Sc. (Honours) Examination, 2023
Semester-II
Statistics
MSC-21-Inference-II
Time: 3 hrs Full Marks:40

Answer any **four** questions of the following.

1. (a) Let X be a random variable having density function $f \in \{f_0, f_1\}$ when $f_0(x) = 1;\ 0 < x < 1$ and $f_1(x) = \frac{1}{3};\ 0 < x < 3$. For testing $H_0 : f = f_0$ against $H_1 : f = f_1$, based on a single observation find out the power of most powerful test when α (size)= .05.

   (b) Propose a level α MP test for $H_0 : X \sim N(0, 1/2)$ against $H_1 : X \sim Cauchy(0, 1)$ based on a single observation.

   (c) Define a α similar test.

4+4+2

2. (a) For a nonparametric test of median (μ) being zero under H_0, let $X_\alpha, \alpha = 1(1)n$ be a continuous random variable for $\alpha = 1(1)n$. R^+_α is the rank of $|X_\alpha|$. Further define an indicator variable Z_α such that

$$Z_\alpha = \begin{cases} 1 \ if \ X_\alpha > \mu \\ 0 \ if \ X_\alpha < \mu \end{cases}$$

   Show that under H_0, R^+_α and Z_α are independently distributed.

   (b) Define a test having Neyman structure with respect to a sufficient statistic T.

   (c) Suppose X and Y be the independent Poisson random variables with parameters λ and μ respectively. Propose a UMP test for $H_0 : \mu \leq \lambda$ against $H_1 : \mu > \lambda$.

3+3+4

3. (a) Let X_1, X_2 and X_3 be collected from $U(\theta, \theta + 2)$. Does this family have Monotone likelihood ratio property? Also construct a UMP test for $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

   (b) Establish that Kolmogorov-Smirnov one sample test statistic is distribution free.

5+5

4. (a) Let $X_1, X_2, \cdots, X_n$ be a random sample from $f(x, \lambda) = 2\lambda e^{-\lambda x^2} x;\ x > 0$. Construct a UMP test for testing $H_0 : \lambda = 1$ against $H_1 : \lambda > 1$.

   (b) For an exponential family with single parameter θ and sufficient statistic $T(x)$, to construct a UMPU test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, show that $E_{\theta_0}[T\phi(T)] = \alpha E_{\theta_0}(T)$, α being the level of significance.

5+5

5. (a) Define one sample linear rank statistic associated with one sample location test.

(b) Find its mean and variance.

(c) Show that one sample Wilcoxon signed rank test statistic is a particular case of it.

5+5

6. (a) Suppose $F_X()$ and $F_Y()$ be the probability distribution functions of X and Y respectively. Discuss a nonparametric test in order to check X is stochastically larger than Y, clearly stating the null and alternative hypothesis.

(b) Let P_0,P_1, P_2 be the probability distributions assigning to the integers 1, 2, 3, 4, 5 the following probabilities:

	1	2	3	4	5
P_0	0.03	0.02	0.02	0.01	0.92
P_1	0.06	0.05	0.08	0.02	0.79
P_2	0.09	0.05	0.12	0	0.74

Determine whether there exists a UMP test for $H_0 : P = P_0$ against $H_1 : P \neq P_0$ at α(size)= .05.

5+5

M.Sc. Semester II Examination 2023
Subject: Statistics
Paper: MSC 22
Applied Multivariate Analysis

Full Marks: 40 Time: 3 hours

Answer any four of the following six questions of equal marks.
(Notations carry usual meanings)

1. (a) Derive Principal Components of a p- variate random vector. Cite two real-life applications of principal component analysis.

   (b) Prove or disprove- ”Factor Model solution always exists”.

   6+4

2. (a) Derive an optimum rule for discriminating two populations.

   (b) Let $f_1(x) = (1 - |x|)$ for $|x| \leq 1$, & $f_1(x) = 0$ for other values of x and $f_2(x) = (1 - |x - 0.5|)$ for $-0.5 \leq x \leq 0.5$ & $f_2(x) = 0$ for other values of x. Sketch the two densities. Also identify the classification regions to discriminate the two populations (Assume prior probabilities and cost of misclassifications are requal.)

   5+5

3. (a) Write down hierarchical clustering algorithm. Differentiate among single, complete and average linkage clustering methods.

   (b) Compare and contrast canonical correlation analysis (CCA) with multiple regression analysis. Under what circumstances would one choose CCA over multiple regression, and vice versa?

   5+5

4. (a) Find the principal components and proportion of total system variance explained by each when the covariance matrix is given by

$$\begin{pmatrix} \sigma^2 & \rho\sigma^2 & 0 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ 0 & \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

   where $-\frac{1}{\sqrt{2}} < \rho < \frac{1}{\sqrt{2}}$.

   (b) What is factor rotation and why is that performed? Explain how factor rotation does not change the factor model representation.

   5+5

5. (a) How does MANOVA differ from the univariate ANOVA (Analysis of Variance) technique? Explain the key advantages of using MANOVA when dealing with multiple dependent variables.

   (b) Interpretation of MANOVA results is essential for understanding the relationships between the independent and dependent variables. Explain how to interpret significant MANOVA findings, including the significance of individual dependent variables.

   4+6

6. (a) Explain how wouldyou obtain solution of a factor model with reasons.

   (b) Write a short note on K-means clustering method.

   5+5

M.Sc. Examination, 2023
Semester-II
Statistics
Course: MSC-23
(Regression Techniques)
Time: 3 Hours **Full Marks: 40**

Questions are of value as indicated in the margin.
Notations have their usual meanings

Answer any four questions.

1. Consider the simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with $E(\epsilon_i) = 0$, $var(\epsilon_i) = \sigma^2$, and and ϵ_i's are uncorrelated, $i \in \{1, 2, \cdots, n\}$.

(a) Show that $SSR = \hat{\beta}_1 S_{xy} = \hat{\beta}_1^2 S_{xx}$.

(b) Show that $E(SSR) = \sigma^2 + \beta_1^2 S_{xx}$ and $E(MSE) = \sigma^2$.

(c) Consider the maximum-likelihood estimator $\tilde{\sigma}^2$ of σ^2. Find the bias in $\tilde{\sigma}^2$.

(d) Prove that the maximum value of R^2 is less than 1 if the data contain repeated (different) observations on y at the same value of x.

$2 + 3 + 3 + 2$

2. (a) Consider the multiple regression model $\mathbf{y} = \mathbf{X}\beta + \epsilon$. Find the expression for the least-squares estimator $\hat{\beta}$ of β. Show that the least-squares estimator can be written as $\hat{\beta} = \beta + \mathbf{R}\varepsilon$, where $\mathbf{R} = (\mathbf{X^T X})^{\mathbf{-1}}\mathbf{X^T}$.

(b) Consider a correctly specified regression model with p terms, including the intercept. Make the usual assumptions about ϵ. Prove that $\sum_{i=1}^{n} Var(\hat{y}_i) = p\sigma^2$.

$5 + 5$

3. (a) Write a short note on PRESS Residual and PRESS statistic.

(b) Diagnose if the following statement is True/False with suitable explanation(s).

> A studentized residual (r_i) is just a deleted residual d_i divided by its estimated standard deviation $s(d_i)$ (first formula). This turns out to be equivalent to the ordinary residual divided by a factor that includes the mean squared error based on the estimated model with the i^{th} observation deleted, $MSE(i)$, and the leverage, h_{ii} (second formula). In other words, $r_i = \dfrac{d_i}{s(d_i)} = \dfrac{e_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}}$.

Hence (or otherwise), explain that the studentized residual for a given data point depends not only on the ordinary residual but also on the size of the mean squared error (MSE) and the leverage.

$5 + 5$

4. (a) (i) Briefly describe the principle of Logistic Regression and Probit Regression.

(ii) Among two logit models, how do you determine which model is better? Justify.

(b) (i) Describe a formal test for Lack of Fit under the suitable assumption(s).

(ii) Write a short note on influential points and leverage points.

$(2\frac{1}{2} + 2\frac{1}{2}) + (2\frac{1}{2} + 2\frac{1}{2})$

CONTINUED

5. (a) Explain how the following problematic scenario (s) can be handled in light of Ridge regression.

(i) The Least Squared (LS) estimate depends upon $(X'X)^{-1}$, we would have problems in computing β_{LS} if $X'X$ were singular or nearly singular.

(ii) In above case(s), a small changes to the elements of X lead to large changes in $(X'X)^{-1}$, and the least squared estimator β_{LS} may provide a good fit to the training data, but it may not fit sufficiently well to the test data.

(b) Define the Ridge estimator $\hat{\beta}_{\mathbf{R}}$. Obtain the mean squared error of the Ridge estimator. Justify the following statement:

Ridge estimate will not necessarily provide the best " fit " to the data.

$(2\frac{1}{2} + 2\frac{1}{2}) + 5$

6. (a) Compare between Least-Squared estimation in linear regression vs. Least-Squared estimation in nonlinear regression. Illustrate how the linearization can be done by a Taylor series expansion of the nonlinear Regression function, followed by the iteration method of parameter estimation.

(b) Write a short note on R-estimators as Robust Regression for Linear Model.

(OR)

Write a short note on M-estimators as Robust Regression for Linear Model.

$5 + 5$

M.Sc. Examination 2023
Semester-II
Statistics
Course: MSC-24 (Design of Experiments)
Full Marks: 40 **Time: 3 Hours**

(Answer any four questions.)

1. (a) What is a split-plot design? Write down the underlying model, hypotheses and the detailed analysis procedure of the design.

   (b) Find the efficiency of split-plot design with respect to a randomized block design.

   7+3

2. Consider a randomized block design with v blocks and v treatments. Augment treatment i with block $i(i = 1, 2, \dots v)$.

   (a) Is the resultant design connected?

   (b) Is it orthogonal?

   (c) Is $\tau_2 - 2\tau_3 + \tau_4$ estimable? If so, find the expression of its BLUE and its standard error. You should simplify your answer as much as possible.

   3+3+4

3. Construct BIBD s with the following parameters. You should properly state the results you use in each case.

   (a) $v = 13, b = 26, r = 6, k = 3, \lambda = 1$

   (b) $v = 15, b = 15, r = 7, k = 7, \lambda = 2$

   (c) $v = 9, b = 12, r = 4, k = 3, \lambda = 1$

   3+4+3

4. (a) For a symmetric BIBD with parameters (v, k, λ), show that any two blocks have exactly λ treatments in common.

   (b) In addition, if v is even, then prove that $(k - \lambda)$ is a perfect square.

   (c) Define the efficiency factor of a BIBD. Prove that it is less than 1.

   4+3+3

5. Derive the inter-block estimate of the contrast $\boldsymbol{p}'\boldsymbol{\tau}$ of the treatment effects and the standard error of the estimate. Show that the estimator is uncorrelated with the intra-block estimator.

   7+3

6. (a) Construct the layout of a $(3^3, 3^2)$ experiment confounding ABC^2, BC^2.

(b) Consider the 3^3 experiment conducted in 2 replications in blocks of 3^2 plots. The following information is given below.

Replicate 1: 100, 112, 202, 211, 220, 121, 010, 021, 002
Replicate 2: 001, 102, 012, 110, 200, 121, 222, 211, 020

Identify the confounded effects.

(c) Write down the ANOVA table of a 3^2 factorial experiment.

4+3+3

M.Sc. Semester II Examination 2023
Subject: Statistics
Paper: MSC 25
(Practical on Applied Multivariate and Inference II)

Full Marks: 40 Time: 4 hours

Answer all the following questions.
(Notations carry usual meanings)

1. Let X and Y be independent distributed Poisson with λ and Poisson with μ respectively. Construct a UMP test for testing $H_0 : \mu \leq \lambda$ against $H_1 : \mu \geq \lambda$ against the alternative $\lambda = .4$ and $\mu = .5$ at level of significance 0.1.

6

2. An urn contains 8 marbles of which m are white and (10-m) are black. To test $H_0 : m = 5$ against $H_1 : m > 5$ one draws 4 marbles without replacement. The null hypothesis is rejected if the sample contains 2 or 3 while balls otherwise accepted. Construct a test and its size.

4

3. Let a random sample of 10 observations, viz., 2,-1.2, 5, 7,-1.2,4, 2.9, 3.9, 2.5,3 are selected from $N(\mu, 16)$. Construct a test for $H_0 : \mu > 3 or \mu < 2$ against $H_0 : 2 < \mu < 3$ for a level of significance .5.

5

4. Consider the following two features of few items.

Item	Feature 1	Feature 2
A	1.5	1.0
B	2.0	1.5
C	3.0	5.0
D	4.0	4.5
E	3.5	4.0
F	9.0	8.5
G	8.5	9.0
H	8.0	8.0
I	1.0	2.0

Construct either of hierarchical (show dendrogram) or K-means (maximum 3) clusters.

8

5. The effectiveness of advertising for two rival products (Brand X and Brand Y) was compared. market research at a local shopping center was carried out with the participants being shown adverts for two rival brands of coffee, which they then rated on the overall likelihood of them buying the product (out of 10, with 10 being ”‘definitely going to buy the product”’). Below is the chart of rating.

Brand X		Brand Y	
Participant	Rating	Participant	Rating
1	3	1	9
2	4	2	7
3	2	3	5
4	6	4	10
5	2	5	6
6	5	6	8

On the basis of this rating do you think these two brands of coffee are the same?

4

6. Suppose you are conducting a study to compare two different teaching methods, Method A and Method B, in terms of their effects on students’ score on two subjects: Math and Science. Let you have the following data for a sample of students.

Method A
Math Scores:[85, 92, 78, 88, 76]
Science Scores:[78, 86, 82, 75, 80]
Method B
Math Scores:[92, 88, 75, 90, 85]
Science Scores:[85, 80, 88, 82, 78]

Assuming equal sample sizes for both methods, perform a one-way MANOVA to determine if there are any significant differences between the teaching methods in terms of the combined Math and Science scores.

8

7. Viva-voce and Practical Notebook

5

M.Sc. Examination, 2023
Semester-II
Statistics
Course: MSC-26 (Practical)
Time: 4 Hours **Full Marks: 40**

1. Consider the following incomplete block design (table 1). The yields are given as the entries of the table 1. Perform a complete intra-block analysis. Also find an estimate of the error variance. You may use R software for the calculations. 12

Treatment Block	A	B	C	D	E
1	$\cdots$	16	30	25	$\cdots$
2	5	10	18	$\cdots$	$\cdots$
3	7	28	$\cdots$	$\cdots$	35
4	10	$\cdots$	$\cdots$	20	52
5	$\cdots$	$\cdots$	24	11	40
6	12	24	29	$\cdots$	$\cdots$

Table 1: incomplete block design (The yields are given as the entries of the table)

2. The following table (table 2) gives the plants and yields (in suitable units) of a manurial experiment involving two factors N and P each at 3 levels.
Test for the significance of main effects and 2-factor interaction effects. 6

Replicate 1:

Treatment	12	00	10	21	22	01	11	20	02
Yield	223	236	240	300	189	160	284	271	259

Replicate 2:

Treatment	01	20	12	00	22	11	02	21	10
Yield	269	233	266	213	226	240	282	209	293

Replicate 3:

Treatment	02	11	01	21	22	10	12	00	20
Yield	191	300	278	209	226	233	182	270	258

Table 2: plants and yields (in suitable units)

CONTINUED

3. The kinematic viscosity of a certain solvent system depends on the ratio of the two solvents and the temperature (Data table $b.10$ of Table 3). You may attach the data using the following R code:

```
library("MPV")
data(table.b10)
```

(a) Fit a multiple linear regression model relating the viscosity to the two regressors.

(b) Test for significance of the regression. What conclusions can you draw?

(c) Use t-tests to assess the contribution of each regressor to the model. Discuss your findings.

(d) Calculate R^2 and R^2_{Adj} for this model. Compare these values to the R^2 and R^2_{Adj} for the simple linear regression model relating the viscosity to temperature only.

(e) Find a 99% C.I for the regression coefficient for temperature for both models in part d. Discuss any differences.

5

4. A study was performed to investigate new automobile purchases. A sample of 20 families was selected. Each family was surveyed to determine their oldest vehicle's age and total family income. A follow-up survey was conducted 6 months later to determine if they had actually purchased a new vehicle during that time period ($y = 1$ indicates yes, and $y = 0$ indicates no). You may attach the data in Table 3 using the following R code:

```
library("MPV")
data(p13.5)
```

(a) Fit a logistic regression model to the data and interpret the model coefficients β_1 and β_2.

(b) Does the model deviance indicate that the logistic regression model is adequate?

(c) What is the estimated probability that a family with an income of $45,000 and a car that is 5 years old will purchase a new vehicle in the next 6 months?

(d) Expand the linear predictor to include an interaction term. Is there any evidence that this term is required in the model?

(e) Find approximate 95% confidence intervals on the model parameters for the logistic regression model.

5

5. Hald cement data: The response variable y is the heat evolved in a cement mix. The four explanatory variables are ingredients of the mix, i.e., x_1: tricalcium aluminate, x_2: tricalcium silicate, x_3: tetracalcium alumino ferrite, x_4: dicalcium silicate. You may attach the data by using the following R code (or in Table 3):

```
library("BAS");
 data(Hald);
```

(a) Using the Hald cement data, find the eigenvector associated with the smallest eigenvalue of X^TX. Interpret the elements of this vector.

(b) What can you say about the source of multicollinearity in these data?

2+2

6. Analyse the chemical process data in Table b.5 (to be found in the "MPV" package of R, or in Table 3) for evidence of multicollinearity. Use the variance inflation factors and the condition number of X^TX.
You may attach the data by using the following R code:

```
library("MPV")
data(table.b5)
```

3

7. Practical Notebook & Viva voce 5

ATTACHMENT FOLLOWS

ATTACHMENT

	y	x1	x2	x3	x4
1	78.50	7.00	26.00	6.00	60.00
2	74.30	1.00	29.00	15.00	52.00
3	104.30	11.00	56.00	8.00	20.00
4	87.60	11.00	31.00	8.00	47.00
5	95.90	7.00	52.00	6.00	33.00
6	109.20	11.00	55.00	9.00	22.00
7	102.70	3.00	71.00	17.00	6.00
8	72.50	1.00	31.00	22.00	44.00
9	93.10	2.00	54.00	18.00	22.00
10	115.90	21.00	47.00	4.00	26.00
11	83.80	1.00	40.00	23.00	34.00
12	113.30	11.00	66.00	9.00	12.00
13	109.40	10.00	68.00	8.00	12.00

(a) Hald Dataset

	x1	x2	y
1	0.92	-10.00	3.13
2	0.92	0.00	2.43
3	0.92	10.00	1.94
4	0.92	20.00	1.59
5	0.92	30.00	1.32
6	0.92	40.00	1.13
7	0.92	50.00	0.97
8	0.92	60.00	0.85
9	0.92	70.00	0.75
10	0.92	80.00	0.67
11	0.75	-10.00	2.27
12	0.75	0.00	1.82
13	0.75	10.00	1.49
14	0.75	20.00	1.25
15	0.75	30.00	1.06
16	0.75	40.00	0.92
17	0.75	50.00	0.80
18	0.75	60.00	0.71
19	0.75	70.00	0.63
20	0.75	80.00	0.57
21	0.57	-10.00	1.59
22	0.57	0.00	1.32
23	0.57	10.00	1.12
24	0.57	20.00	0.96
25	0.57	30.00	0.83
26	0.57	40.00	0.73
27	0.57	50.00	0.65
28	0.57	60.00	0.58
29	0.57	70.00	0.52
30	0.57	80.00	0.47
31	0.36	-10.00	1.16
32	0.36	0.00	0.99
33	0.36	10.00	0.86
34	0.36	20.00	0.75
35	0.36	30.00	0.67
36	0.36	40.00	0.59
37	0.36	50.00	0.53
38	0.36	60.00	0.48
39	0.36	70.00	0.44
40	0.36	80.00	0.40

(b) Table b.10

	x1	x2	y
1	45000.00	2.00	0.00
2	40000.00	4.00	0.00
3	60000.00	3.00	1.00
4	50000.00	2.00	1.00
5	55000.00	2.00	0.00
6	37000.00	5.00	1.00
7	31000.00	7.00	1.00
8	40000.00	4.00	1.00
9	75000.00	2.00	0.00
10	43000.00	9.00	1.00
11	50000.00	5.00	1.00
12	35000.00	7.00	1.00
13	65000.00	2.00	1.00
14	53000.00	2.00	0.00
15	48000.00	1.00	0.00
16	49000.00	2.00	0.00
17	37500.00	4.00	1.00
18	71000.00	1.00	0.00
19	34000.00	5.00	0.00
20	27000.00	6.00	0.00

(c) p13.5

	y	x1	x2	x3	x4	x5	x6	x7
1	36.98	5.10	400.00	51.37	4.24	1484.83	2227.25	2.06
2	13.74	26.40	400.00	72.33	30.87	289.94	434.90	1.33
3	10.08	23.80	400.00	71.44	33.01	320.79	481.19	0.97
4	8.53	46.40	400.00	79.15	44.61	164.76	247.14	0.62
5	36.42	7.00	450.00	80.47	33.84	1097.26	1645.89	0.22
6	26.59	12.60	450.00	89.90	41.26	605.06	907.59	0.76
7	19.07	18.90	450.00	91.48	41.88	405.37	608.05	1.71
8	5.96	30.20	450.00	98.60	70.79	253.70	380.55	3.93
9	15.52	53.80	450.00	98.05	66.82	142.27	213.40	1.97
10	56.61	5.60	400.00	55.69	8.92	1362.24	2043.36	5.08
11	26.72	15.10	400.00	66.29	17.98	507.65	761.48	0.60
12	20.80	20.30	400.00	58.94	17.79	377.60	566.40	0.90
13	6.99	48.40	400.00	74.74	33.94	158.05	237.08	0.63
14	45.93	5.80	425.00	63.71	11.95	130.66	1961.49	2.04
15	43.09	11.20	425.00	67.14	14.73	682.59	1023.89	1.57
16	15.79	27.90	425.00	77.65	34.49	274.20	411.30	2.38
17	21.60	5.10	450.00	67.22	14.48	1496.51	2244.77	0.32
18	35.19	11.70	450.00	81.48	29.69	652.43	978.64	0.44
19	26.14	16.70	450.00	83.88	26.33	458.42	687.62	8.82
20	8.60	24.80	450.00	89.38	37.98	312.25	468.38	0.02
21	11.63	24.90	450.00	79.77	25.66	307.08	460.62	1.72
22	9.59	39.50	450.00	87.93	22.36	193.61	290.42	1.88
23	4.42	29.00	450.00	79.50	31.52	155.96	233.95	1.43
24	38.89	5.50	460.00	72.73	17.86	1392.08	2088.12	1.35
25	11.19	11.50	450.00	77.88	25.20	663.09	994.63	1.61
26	75.62	5.20	470.00	75.50	8.66	1464.11	2196.17	4.78
27	36.03	10.60	470.00	83.15	22.39	720.07	1080.11	5.88

(d) Table b.5

Table 3: The datasets: (only for reference and are **NOT recommended to be typed manually** in R). Students may use the pre-installed R packages and R codes mentioned in the question to retrieve the datasets quickly and save time.

CONTINUED

M.Sc. Examination, 2023
Semester-IV
Statistics
Course: MSC-41(Reliability Analysis)
Time: Three Hours Full Marks: 40

Questions are of value as indicated in the margin
Notations have their usual meanings

Answer **Question No. 1** and **any four** of the remaining questions

1. Answer any ***four*** of the following questions: 2x4=8
(a)What do you mean by standby component of a system?
(b) Write down the UMVUE of reliability function, R(t) for exponentially distributed lifetime.
(c) State the "Lack of Memory" property.
(d) What are irrelevant and redundant components in a system?
(e) When is a distribution said to be NBU?
(f) What do you understand by accelerated life test?

2. Briefly describe failure rate and reliability function. Establish the relationship between them. Show that, under suitable assumptions, failure rate of a series system is the sum of individual component failure rates. 2+3+3

3. Define Failure Rate Average. Prove that constancy of ratio of Failure Rate to Failure Rate Average is the characterization of the Weibull distribution. 2+6

4. (i) Explain Time censoring and Number censoring in life testing.
(ii) Obtain the maximum likelihood estimators of the scale and shape parameters of Weibull distribution using Number censoring. Hence obtain the estimate of reliability. 3+ (4+1)

5. Derive a bivariate exponential distribution from a fatal shock model. Does this BVED satisfy the lack of memory property? Work out the regression functions. 2+2+4

6. (i) Distinguish between IFR and DFR.
(ii) Show that in case of Weibull distribution failure rate depends on its shape parameter and the distribution can be used in all three phases of bathtub curve.
(iii) For a 2-out-of-3 system, let the reliability of each component be 'p'. Find the reliability of the system assuming the components are statistically independent. How does the system reliability behave in relation to p? 2+3+3

7. (i) Let $T_1, T_2, ..., T_n$ be independently and identically distributed random variables having reliability function given by

$$R(t) = 1 - \lambda t + o(t) \text{ as } t \to 0.$$

Show that $X_n = n.\min\left(T_1, T_2, ..., T_n\right)$ has asymptotically an exponential distribution as $n \to \infty$.
(ii) Define Mean Residual Life. Discuss how a life distribution is classified based on mean residual life. 4+4

M.Sc. Examination 2023
Semester-IV
Statistics
Course: MSC-42 (Bayesian Inference)

Time: 3 Hours **Full Marks: 40**

Questions are of value as indicated in the margin
Notations have their usual meanings

Answer **any four** questions

1. (a) What is a conjugate prior? Give an example.
(b)Show that for Poisson (θ), under conjugate prior set up, the posterior expectation of θ is a convex combination of the prior expectation and the sample average.
(c)For a new observation $\tilde{y}$, find the expression of the predictive mean and variance under (b).
2+5+3

2. (a) What is Gibbs sampling? Describe general properties of a Gibbs sampler.
(b)For a normal set up with semi-conjugate prior for the variance and normal prior for mean, describe the Gibbs sampling process in details. 5+5

3. Let $\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots \boldsymbol{Y_n} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Assuming normal prior for the mean vector $\boldsymbol{\theta}$ and Wishart prior for the dispersion matrix $\boldsymbol{\Sigma}$, find the posterior distributions of $\boldsymbol{\theta}|\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots \boldsymbol{Y_n}$ and $\boldsymbol{\Sigma}|\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots \boldsymbol{Y_n}$**.**
10

4. (a) Describe a Bayesian method to compare two groups.
(b) How will you generalize this method for multiple groups? Describe the process with the help of the hierarchical normal model. 5+5

5. (a) Describe the procedure for the Bayesian analysis of a regression model using semi-conjugate prior distributions.
(b) What is “g-prior”? How will you modify the analysis if you are using “g-prior” in (a)?
6+4

6. (a) Describe the role of Markov chain Monte Carlo (MCMC) in Bayesian analysis. How does it differ from Monte Carlo (MC) method? Give an example.
(b) How does the correlation of the MCMC samples affect posterior approximation?
(c) Show that, in general,
$$Var_{MCMC} \geq Var_{MC}$$
5+2+3

M.Sc. Semester IV Examination 2023

Subject: Statistics

Paper: MSC43/MSS 09

(Introductory Data Science and Statistical Machine Learning)

Time: Three Hours　　　　　　Full Marks: 40

Answer any **four** of six questions.

1. a) What do you mean by feature scaling and centering? Differentiate between them with examples.

   b) Briefly describe the ways of detecting outliers in a dataset and how you would analyze a dataset that contains outliers.

   5+5

2. a) What are the different advantages and disadvantages of the decision tree algorithm?

   b) Why should one prefer a (random) forest (collection of trees) to a single tree?

   7+3

3. a) Explain why the linear regression technique is not applicable to model the target variable's yes/no or zero/one data. Describe a suitable technique to model such types of data with reasons.

   b) In comparison to other techniques what is/are the additional advantage(s) of the Logistic Regression technique in classification problems?

   6+4

4. What is Kernel-trick in Support Vector Machines (SVM)? Why is it important? Answer with an illustration. Cite examples of a few Kernel functions. Is SVM sensitive to feature scaling? 10

5. a) What do you understand by unstructured data? Give examples.

   b) Explain the concept of text analysis with examples. How does it differ from Natural Language Processing (NLP)? In this context also explain the terms 'Tokenization', 'Lammentization' and 'Stop words.

   3+7

[Turn Over]

6. a) Consider the following transactional data:

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Find the support of the item set {Milk, Diaper, Bread}. Take the rule {Milk, Diaper}→ {Beer} then find and interpret the Confidence and Lift of the rule.

b) Mention some advantages and disadvantages of Association Rule Mining.

6+4

M.Sc Semester IV Examination, 2023

Statistics

MSC-44(Practical)

Time: Four Hours Full Marks: 40

(One may use Computer, if necessary)

1. In order to estimate the mean burning time of a particular brand of bulb, 30 bulbs were left burning. The bulbs that failed are not replaced upon failure. The following burning time (in hours) were recorded: 20, 27, 52, 61, 110, 122, 150, 214, 232, 238, 371, 393, 426, 445, 472, 503, 526, 581, 627, 698, 805, 909, 976, 1001, 1016, 1033, 1086, 1192, 1322, 1581.
   Calculate the maximum likelihood estimate and minimum variance unbiased estimate of R(t) at t=300 hrs. 6

2. 800 electronic components were placed on life tests. The system is observed at 3 hrs, 6 hrs, 9 hrs…, 30 hrs. The number of failures is noted. Calculate the hazard rates, and plot it in diagram and comment. 5

Failure data for 800 electronic components

Time interval in hours	Number of Failures in interval
0-3	190
3-6	72
6-9	46
9-12	30
12-15	17
15-18	13
18-21	14
21-24	9
24-27	6
27-30	3

3. Consider the following data on the weight of poultry on different days of measure:
Days of measure (X): 8, 15, 22, 29, 36, 43, 50
Weight in grams (Y): 177, 236, 285, 350, 376, 401, 430
The regression equation is:
$$y = \mu + \beta(x - \bar{x}\ ) + e_{ij}$$
where $e_{ij} \sim N(0, \sigma^2)$
Perform a Bayesian linear regression of Y on X assuming informative and non-informative priors for the parameters involved in the above model. 6

4. Consider the following data on the wing length in millimeters of nine members of a species of midge (small, two-winged flies).
1.64, 1.70, 1.72, 1.74, 1.82, 1.82, 1.82, 1.90, 2.08
From these nine measurements we wish to make inference on the population mean θ. Studies from other populations suggest that wing lengths are typically around 1.9 mm and standard deviation should not be too far from 0.01.
Find the posterior distribution of θ and 95% credible interval for θ. Compare this credible interval with 95% confidence interval for θ and comment. 6

5. Classic Housing Bank wishes to automate its housing loan processing system. It has provided its previous applicants' data along with their loan status 'Yes' or 'No' i.e., approved or not approved. They consider most of the information (e.g., Gender, Education Level, Annual Income, Dependent members in the family, Loan amount, credit history, etc.) collected in the loan application form. Given their data, develop a prediction model for predicting an applicant's chance of getting the loan. Show the accuracy of your model in the appropriate form. Interpret your model.
Also, comment on the followings:
a. How does the loan approval chance vary if the applicant is a Male than a female?
b. How does the loan approval chance vary if the applicant has a bad credit history?
c. From the supplied new applicant's data what percentage of the applicants may get their loans approved?
(Data files will be provided during the exam)
Note: Credit History: 1 indicates the applicant has a credit history 12

6. Notebook & Viva-voce 5