M.Sc. Examination, 2022

Semester-I

Statistics

Course: MSC-11

(Linear Models and Distribution Theory)

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). Show that the general solution of the system of homogeneous equations $\boldsymbol{Ax} = \boldsymbol{0}$ can be expressed as

$$\tilde{\boldsymbol{x}} = (I - H)\boldsymbol{z}$$

Where $\boldsymbol{z}$ is any arbitrary vector, $\boldsymbol{H} = \boldsymbol{S}^{-}\boldsymbol{S}, \boldsymbol{S} = \boldsymbol{X}'\boldsymbol{X}$

How will you modify this result for the non-homogeneous consistent equations $\boldsymbol{Ax} = \boldsymbol{u}$?

(b) Let $\boldsymbol{l}_1'\boldsymbol{\beta}$ and $\boldsymbol{l}_2'\boldsymbol{\beta}$ are two estimable functions and $\boldsymbol{l}_1'\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{l}_2'\widehat{\boldsymbol{\beta}}$ be their least square estimates respectively. Find **Var**$(\boldsymbol{l}_1'\widehat{\boldsymbol{\beta}})$ and **Cov**$(\boldsymbol{l}_1'\widehat{\boldsymbol{\beta}}, \boldsymbol{l}_2'\widehat{\boldsymbol{\beta}})$.

6+4

2. (a). Define estimation space and error space. Show that the covariance between any linear function belonging to the error space and any BLUE is zero.

(b). Consider the following linear model:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \qquad i = 1,2; j = 1,2,3$$

Is $\tau_1 - \tau_2$ estimable?

7+3

3. (a) Show that all linear parametric functions, for the linear model $\boldsymbol{y} = \boldsymbol{x\beta} + \boldsymbol{\epsilon}$, in $\boldsymbol{\beta}$ are estimable if and only if X has full rank.

(b). Show that for the liner model $\boldsymbol{y} = \boldsymbol{x\beta} + \boldsymbol{\epsilon}$ a necessary and sufficient condition for a linear parametric function $\boldsymbol{\lambda}'\boldsymbol{\beta}$ to be estimable is that $\boldsymbol{\lambda}'$ is a linear combination of rows of $\boldsymbol{x}'\boldsymbol{x}$.

5+5

4. (a) Show that the conditional minimum of the sum of squares of the residuals $(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$, in the model $\boldsymbol{y} = \boldsymbol{X\beta} + \epsilon, \boldsymbol{E}(\epsilon) = 0, V(\epsilon) = \sigma^2 I$, subject to m conditions $\boldsymbol{\Lambda\beta} = \boldsymbol{d}$, where $\boldsymbol{\Lambda\beta}$ are estimable and rank$(\boldsymbol{\Lambda}) = m$, exceeds the error sum of squares (SSE).

(b) Show that $\frac{SSR(\beta)}{\sigma^2}$ follows a non-central χ^2 distribution.

6+4

5. (a) What is general linear hypothesis? When it is called "testable"?

(b) Consider the model $y_i = \theta_i + \epsilon_i, i = 1(1)n$

Where the parameters θ_i are subject to the restriction $\sum_{i=1}^{n} \theta_i = 0$

Write down a test procedure to test the following hypothesis

$H_0: \theta_i = \theta_j \qquad (i \neq j), i, j = 1(1)n$ 3+7

4. (a) State and prove Cochran's theorem on quadratic forms. Give an example where this theorem can be used.

(b) If $\boldsymbol{M} \sim \boldsymbol{W}_p(\boldsymbol{\Sigma}, m)$ and $\boldsymbol{B}$ is a $p \times q$ matrix, then show that

$$\boldsymbol{B}'\boldsymbol{M}\boldsymbol{B} \sim W_q(\boldsymbol{B}'\boldsymbol{\Sigma}\boldsymbol{B}, m)$$ 7+3

5. (a) If $\boldsymbol{M} \sim W_p(\boldsymbol{\Sigma}, m), m > p$ then show that the ratio

$\frac{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{M}^{-1}\boldsymbol{a}}$ has the χ^2_{m-p+1} distribution for any fixed p-vector $\boldsymbol{a}$

(b). Prove that the BLUE of any linear combination of estimable parametric function is the linear combination of their BLUEs. 7+3

M.Sc Semester I Examination, 2022

Statistics

MSC-15(Practical)

Time: Four Hours | Full Marks: 40

One may use computer, if necessary

1. Following data represent a random sample of size from the Cauchy population with the probability density function $f(x,\theta) = \frac{1}{\pi} \cdot \frac{1}{1+(x-\theta)^2}; -\infty < x, \theta < \infty$. Find out the MLE of θ. The observations are 3.8807, 2.9957, 5.2043, 4.9893, 2.7468, 4.9557, 4.9367, 3.9649, and 3.1674. 7

2. Consider the problem of point estimation of θ in $N(\theta,1)$. Given that θ belongs to [-1, 1]. On the basis of a sample of size n, the following estimator has been defined.

$$T = -1 \text{ if } \bar{X} < -1$$
$$= \bar{X} \text{ if } -1 \le \bar{X} \le 1$$
$$= 1 \text{ if } \bar{X} > 1,$$

$\bar{X}$ being sample mean. Assuming (i) squared error loss and (ii) absolute error loss draw the risk curve of $\bar{X}$ and T over the range $\theta \in [-1,1]$ on the same graph paper and comment. Take n=15. 8

3. (i) Suppose that the two observations on each of three treatments are as follows.

Treatments		
t_1	t_2	t_3
8	5	12
6	3	14

Assuming the linear model to be $y = X\beta + \epsilon$, where $\beta' = (\mu, t_1, t_2, t_3)$, find the BLU estimates of the following functions:

a. μ

b. $t_1 - t_2$

c. $2\mu + t_1 + t_2$

d. $\mu + \frac{t_1+t_2+t_3}{3}$

e. $t_1 + t_2$

(ii) Would you like to reject the null hypothesis $t_1 - t_2 = 7$ at 5% level?

8+6

4. Consider the following data on perspiration in 10 healthy females measured in terms of sweating rate (X_1) along with $X_2 = Na\ content$ and $X_3 = Ka\ content$.

No.	X_1	X_2	X_3
1	3.7	48.5	9.3
2	5.7	65.1	8
3	3.8	47.2	10.9
4	3.2	53.2	12
5	3.1	55.5	9.7
6	4.6	36.1	7.9
7	2.4	24.8	14
8	7.2	33.1	7.6
9	6.7	47.4	8.5
10	5.4	54.1	11.3

Assuming $\mathbf{X} \sim N(\mu, \Sigma)$, perform a test procedure to test the following hypothesis:

$$H_0: \mu = \begin{pmatrix} 4 \\ 50 \\ 10 \end{pmatrix}$$

4

5. Practical Note Book and Viva-Voce 5

****

M.Sc. Examination, 2022

Semester-I

Statistics

Course: MSC-12

(Real Analysis and Measure Theory)

Time: 3 Hours Full Marks: 40

Questions are of value as indicated in the margin.

Notations have their usual meanings

NOTE: There are total 6 questions. Answer any 4 questions.

1. (a) Let F be the distribution function on $\Re$ given by:

$$F(x) = \begin{cases} 0, & x < -1 \\ 1+x, & -1 \leq x < 0 \\ 2+x^2, & 0 \leq x < 2 \\ 9, & x \geq 2 \end{cases}$$

Let μ be the Lebesgue-Stieltjes measure corresponding to F, compute the measure of each of the following sets (with explanations).

(i) $\{2\}$

(ii) $\{x : \ |x| + 2x^2 > 1\}$

(iii) $(-1, 0] \cup (1, 2)$

(b) Let $\Omega = \Re^2$. Define A_n as the interior of the circle with radius 1 and center $\left(\frac{(-1)^n}{n}, 0\right)$. Find $\limsup_n A_n$ and $\liminf_n A_n$ with suitable explanations.

[(1 + 2 + 2) + 5 = 10]

2. (a) State Monotone convergence theorem and Dominated convergence theorem. Define Characteristic function. If $X_1, X_2, \cdots, X_n$ are n independent random variables, then show that

$$\phi_{X_1+X_2+\cdots+X_n}(\omega) = \phi_{X_1}(\omega)\phi_{X_2}(\omega)\cdots\phi_{X_n}(\omega).$$

(b) Define open ball and interior point in $\Re^n$. Define open and close sets in n-dimensional Euclidian space $\Re^n$. Prove that, union of finite number of open sets is open. Prove/disprove that the union of infinite number of open sets is always open.

[5 + 5 = 10]

3. (a) Define almost sure convergence, convergence in probability and convergence in distribution. State Borel-Cantelli Lemma. Suppose $f_n \to f$ pointwise. Can we say that $\int f_n d\mu \to \int f d\mu$? Explain with suitable example.

(b) Let μ be a non-negative, finitely additive set function on field F. If $A_1, A_2, \cdots$ are disjoint sets in F and $\bigcup_{n=1}^{\infty} A_n \in F$, show that

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) \geq \sum_{n=1}^{\infty} \mu(A_n)$$

[5 + 5 = 10]

4. (a) State Fatou's lemma, Levy continuity theorem,, and Heine-Borel Theorem.

(b) Determine the radius of convergence and interval of convergence for the following power series:

$$\sum_{n=1}^{\infty} \frac{2^n}{n}(4x-8)^n$$

[5 + 5 = 10]

5. (a) For any characteristic function $\phi(t)$, explain why $|\phi(t)| \leq 1$? Consider the random variable X that has a standard Cauchy distribution. Show that the moment generating function does not exist for this random variable on any real interval with positive length. Does the characteristic function exist for X?

(b) If $X \sim Exponential(\lambda)$, show that

$$\phi_X(\omega) = \frac{\lambda}{\lambda - j\omega}.$$

[5 + 5 = 10]

6. (a) Define adherence point and accumulation point of a set $S \subset \Re^n$. If A is open and B is close, explain why $A - B$ is open.

(b) State Bolzano-Weirstrass theorem and illustrate with a suitable example.

(c) Define Closure of a set. Show that A set S is closed iff $S = \bar{S}$. Define Derived set. Show that, a set is closed iff it contains all its accumulation points.

[3 + 3 + 4 = 10]

M.Sc. Examination, 2022

Semester-I

Statistics

Course: MSC-13

(Statistical Inference-I)

Time: Three Hours **Full Marks: 40**

Questions are of value as indicated in the margin

Notations have their usual meanings

1. Answer any FOUR questions of the following. 4x5=20

(a) Let $(X_1, X_2, ..., X_n)$ be a random sample from R $(0, \theta)$ distribution with unknown parameter θ. Write $X_{(n)} = \max(X_1, X_2, ..., X_n)$. Then, if $L(\theta)$ denotes the likelihood function, show that, for $\theta > 0, L(X_{(n)}) \geq L(\theta)$, and make your comment.

(b) Suppose $X_i; i = 1(1)n$ follows Bernoulli with parameter θ. The prior distribution of θ is Beta with parameters α and β. Find Bayes estimate of θ under squared error loss.

(c) Let $(X_1, X_2, ..., X_n)$ be a random sample of size n drawn from Bernoulli population with parameter π. Find an UMVUE of $g(\pi) = 1 + n\pi + \frac{n(n-1)}{2}\pi^2$.

(d) Describe Kendall's τ and derive its association with U-statistic.

(e) Describe squared error loss function, absolute error loss function and all-or-nothing loss function. What are the Bayes estimates in these cases? Comment on Bayes estimate of mean of normal distribution.

(f) Define U-statistic of degree m. Show that U-statistic is an unbiased estimator of population variance.

2. Answer any TWO questions of the following.

(a) Describe maximum likelihood method of parameter estimation. State its properties. Suppose $(X_1, X_2, ..., X_n)$ is a random sample of size n from a distribution with cdf

$$F(x \mid \alpha, \beta) = 0 \quad if\ x < 0$$

$$= \left(\frac{x}{\beta}\right)^{\alpha} \quad if\ 0 < x \leq \beta$$

$$= 1 \quad if\ x > \beta$$

where the parameters $\alpha(> 0)$ and $\beta(> 0)$ are unknown. Obtain the maximum likelihood estimators of α and β. 3+3+4

(b) Find the mean of U-statistic. Derive the limiting form of variance of it. State the result regarding the asymptotic distribution of U-statistic with conditions, if any. 3+4+3

(c) Derive Bayes estimate of a real parametric function $\gamma(\theta)$ under squared error loss. Let $X_1, X_2, ..., X_n$ be samples drawn from a normal distribution with mean μ and variance σ^2, where μ is unknown. The prior distribution of μ is assumed to be normal with mean γ and variance η^2. Derive the Bayesian point estimate of μ under the quadratic (squared error) loss function. 4+6

$\binom{n}{y}\theta^{y}(1-\theta)^{n-y}$

$P(\Sigma x_i > k) = \alpha.$

M.Sc. Examination 2022
Semester I
Statistics
Course: MSC-14
(Sample Survey)

Full Marks: 40 Time: 3 Hrs.

Answer any four questions.

1. (a) Explain the rationale behind the randomized response technique.

   (b) Describe how you will estimate the proportions of the people of West Bengal belonging to different political parties by unrelated questionnaire method, where the unrelated question has multiple answers.

   4+6

2. (a) Explain how the double sampling technique can be applied for the ratio estimation of the population total.

   (b) Find the approximate expressions for the bias and the MSE of the estimator of the population total under this sampling scheme.

   (c) Also find the condition under which the scheme outperforms simple random sampling.

   2+6+2

3. (a) Derive the expressions of the first and second order inclusions probabilities in PPSWR (N, n) and SRSWR (N, n) designs.

   (b) Also find the expected values of the effective sample sizes in these two cases separately. You should derive the necessary results.

   4+6

4. (a) Distinguish between informative and non-informative sampling designs.

   (b) Define homogeneously linear unbiased estimator (HLUE). Prove that among the class of HLUE s, no one exists with uniformly minimum variance.

   2+(2+6)

5. (a) What do you mean by two stage sampling? Compare its relative merits and demerits with that of simple random sampling.

   (b) Suggest an unbiased estimator of the population total under this sampling scheme and find its variance.

   (c) Describe how can you obtain the optimal sizes of the first and the second stage samples subject to a given cost. (You may assume that each first stage unit contains the same number of second stage units).

   3+3+4

6. (a) Define Desh Raj's estimator. Prove that it is unbiased for the population total under PPSWOR. What is the drawback of this estimator and how can you overcome it?

   (b) Define Midzuno's sampling scheme. Find the first order inclusion probabilities under this scheme.

   (2+2+2)+(2+2)

M.Sc. Examination 2022
Semester I
Statistics (Practical)
Course: MSC-16

Full Marks: 40 **Time: 4 Hours**

1. At an experimental station; there are 80 fields sown with wheat. Each field was divided into 16 plots of equal size ($1/16^{th}$ hectare). Out of 80 fields, 8 were selected by SRSWOR. From each selected field, 4 plots were chosen by SRSWOR. The yields in kg/plot are given below:

Selected Field	1	2	3	4	5	6	7	8
Plots	Yield							
1	4.32	4.16	3.06	4.00	4.12	4.08	5.16	4.20
2	4.84	4.36	4.24	4.84	4.68	3.96	4.24	4.66
3	3.96	3.50	4.76	4.32	3.46	3.42	4.96	3.64
4	4.04	5.00	3.12	3.72	4.02	3.08	3.84	5.00

(i) Estimate the yield of wheat per hectare for the experimental station along with its standard error.

(ii) How can one estimate obtained from a simple random sample of 32 plots be compared with the estimate obtained in (i)?

(iii) Obtain optimum n and m under cost function $5n + 2mn = 100$, where n and m respectively stand for the number of first stage units drawn and the number of second stage units drawn from each sampled FSU by SRSWOR. 5+3+4

2. Consider the following population data.

Unit Number	x-value	y-value
1	7	32
2	11	41
3	4	25
4	10	67

Compare the performance of Desh Raj's estimator with Horvitz-Thompson estimator assuming a PPSWOR sample of size 2 is drawn. 8

3. A survey on 32 household was conducted and Warner's randomized response technique (Related Question method with $\pi = 0.61$) was applied among the heads of the households to ask about the habit of underpaying the income tax. The actual amount of income tax underpaid is given in the following table.

Household		Response	Amount underpaid
Serial No.	Size	Yes(1)/No(0)	
1	3	1	2300
2	2	1	17000
3	5	0	5568
4	1	1	1304
5	3	0	0
6	2	1	0
7	4	0	711
8	3	0	1203
9	2	1	9874
10	4	1	2200
11	4	1	0
12	7	0	12000
13	2	1	1807
14	3	1	1400
15	4	0	708
16	4	1	1500
17	5	1	0
18	2	0	1100
19	1	0	1825
20	4	1	0
21	5	1	1407
22	3	1	342
23	2	1	645
24	4	1	0
25	3	0	713
26	5	0	1822
27	2	1	0
28	3	1	1623
29	5	1	1108
30	6	0	365
31	2	1	0
32	3	1	1409

(i) Take a sample of 4 households using Rao-Hartley-Cochran's sampling scheme.

(ii) Estimate the total amount of income tax underpaid by these 32 households under the above scheme. Also provide an unbiased variance estimate.

(iii) Use the sample to estimate the proportion underpaying the income tax.

(iv) Now take a sample of 4 distinct households using Lahiri's method. Provide an estimate of the total amount of income tax underpaid under the sampling scheme.

5+3+2+5

4. Practical Note Book and Viva-Voce. 5

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 ~~would~~ do you 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 $\mathbf{R} = (\mathbf{X^T X})^{-1}\mathbf{X^T}$.
   (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 $i (i = 1, 2, \ldots v)$.

   (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 <u>all</u> 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~~ 8-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.

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	...	16	30	25	...
2	5	10	18	...	...
3	7	28	...	...	35
4	10	...	...	20	52
5	...	...	24	11	40
6	12	24	29	...	...

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)

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, 2024

Semester-III

Statistics

Course: MSC-31

Stochastic Process

Time: 3 hrs Full Marks:40

Answer all questions.

1. Answer any **five** from the following. 5×2

   (a) Write down the state space and index set of a Brownian stochastic process.

   (b) For a discrete time Markov process with state space $S = \{0, 1\}$ with $p_{00} = 1, p_{01} = 0, p_{10} = 0.5, p_{11} = 0.5$. Does there exist any unique steady state probability? Justify your answer.

   (c) Consider a Markov chain with state space $\{0, 1, 2\}$ and transition matrix $\begin{pmatrix} \frac{1}{4} & \frac{5}{8} & \frac{1}{8} \\ \frac{3}{4} & 0 & \frac{1}{4} \\ \frac{1}{2} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}$. Find $lim_{n\to\infty} p_{12}^{(n)}$.

   (d) Let $X(t)$ be a Poisson process with rate $\lambda = 2$ /min S_n is the waiting time till the n th event happens. Find the conditional expectation of waiting time of sixth event given that within first 5 minute three events have already occurred.

   (e) Customers arrive at a store with the Poisson flow having rate 10/hr. Each is either male or female with probability 1/2. Compute the probability that at least 5 men entered within 10 a.m. to 10.30 a.m.

   (f) Define an ergodic state.

2. Choose the most suitable word for the following multiple choice questions. 5×2

   (a) For a Wiener process

      i. measure of drift is double to measure of spread.
      ii. measure of drift is a function of time so is measure of spread.
      iii. measure of drift is function of time but measure of spread is function of time square.
      iv. measure of drift is equal to measure of spread.

   (b) Let π_j denote the long run proportion of time that the chain spends in state j where $\pi_j = 0$. Which of the following is false?

      i. No stationary distribution exists for j.
      ii. state j is null recurrent.
      iii. state j is positive recurrent.
      iv. If $i \longleftrightarrow j$ then i is null recurrent.

   (c) For a transition density matrix

      i. all off diagonal elements are positive
      ii. all diagonal elements are zero.
      iii. sum of elements of each row is greater than 0.
      iv. sum of the off diagonal elements is opposite in sign to the diagonal elements.

   (d) Let $W(t)$ be a standard Brownian motion. Then $P(W(1) + W(2) > 0)$ is

      i. 0
      ii. 0.5
      iii. 0.95
      iv. 1

(e) Two persons are catching fish independently with a Poisson flow at rate 2/hr. What is the expected amount of time that all of them will catch at least one fish?

i. 45 min

ii. 1 hr

iii. 30 min

iv. none of the above

3. Answer any **four** of the following. 4×5

(a) Deduce the difference equation on generalized birth and death process clearly stating the necessary assumptions.

(b) Show that if the intervals between successive occurrence of an event E are independently distributed with a common exponential distribution with mean $\frac{1}{\lambda}$, then flow of event E will follow a Poisson process.

(c) Prove that for a closed set, probability of moving out of the set is zero.

(d) Define renewal density. Show that renewal density is the derivative of renewal function.

(e) What do you mean by strict stationarity of a stochastic process? Show that strict stationarity does not imply covariance stationarity and vice versa.

M.Sc. Examination 2023

Semester III

Subject: Statistics

Paper: MSC-32 [Advanced Data Analysis Techniques]

Full Marks: 40

Time: 3 Hrs.

Answer any four questions (*Symbols have their usual meaning*)

1. a) Briefly discuss delta method for univariate case and its multivariate extension. Using this method or, otherwise approximate $Var(\frac{1}{\bar{X}})$ when $X_1, \cdots, X_n \overset{iid}{\sim} Poisson(\lambda)$.
   b) Why do we use odd's ratio? Find the asymptotic distribution of sample log-odd's ratio. (3+3)+(1+3)

2. a) What do you mean by prospective and retrospective study? Discuss advantages and disadvantages of both the studies.
   b) How does the Generalized Linear Model (GLM) differ from the Linear Model (LM), and in what situations would you choose to use a GLM over an LM when analyzing and modeling relationships in data?
   (1+1+4)+4

3. a) Let $f(y;\theta)$ be a exponential family with the following mathematical form: $f(y;\theta) = \exp\left[\frac{y\cdot\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$. Here, y is the observed value, θ is the natural parameter, ϕ is the dispersion parameter, $a(\phi)$ is the dispersion function, $b(\theta)$ is the cumulant generating function, and $c(y,\phi)$ is a normalization term. Show that, $V(Y) = a(\phi)b''(\theta)$. Hence find variance of the binomial distribution.
   b) Explain how to generate samples from $N(0, \sigma = 2)$ using the accept-reject method, with a $C(0,1)$ as the proposal distribution.
   (4+2)+4

4. a) Describe logistic regression model with some of it's application.
   b) Write down the detailed steps involved in fitting a logistic regression model. 3+7

5. a) Explain the basic rationale behind Gibbs sampling. Suppose $X \sim N(\theta, \sigma^2)$ with σ known, $\theta \sim C(\alpha, \beta^2)$ with α, β known. Then describe how Gibbs Sampling can be applied to simulate observations from the posterior of $\theta \mid X$.
   b) Describe the step-by-step process of using bootstrap in a linear regression model. (2+4)+4

6. a) Briefly write down the usage of EM algorithm with examples. Also discuss it's convergence.
   b) Write a short note on 'deviance' in the context of GLM. (3+3)+4

M.Sc. Examination, 2023
Semester-III
Statistics
Course: MSC-33(MSE-1)
(Operations Research and Optimization Techniques)

Time: Three Hours **Full Marks: 40**

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

Answer **any five** questions

1. What is a transportation problem? Is it considered to be a Linear Programming Problem? Show that a balanced transportation problem always has a feasible solution. 2+3+3

2. Briefly state the role of modeling in Operations Research. Mention different types of models and their solutions. 2+6

3. For the M/M/1 queuing system find the expected number of customers in the system in the steady state and also the expected queue length. Find the cumulative distribution function for the waiting time of a customer who has to wait in an M/M/1 queuing system. 4+4

4. What is two-person zero-sum game? Transform this game to a Linear Programming Problem. Prove that if mixed strategies be allowed, then there always exists a value of the game. 2+2+4

5. (a) Define an inventory. What are the advantages and disadvantages of having inventories?
(b) Suppose that Q^* is the optimal order quantity and K^* is the corresponding minimum annual variable cost. Show that if a value of $Q=(1+\alpha)Q^*$ is used, $\frac{K}{K^*}=1+\frac{\alpha^2}{2(1+\alpha)}$, where K is the annual variable cost corresponding to an order quantity Q. 4+4

6. (a) What is replacement Problem? Give some illustrations. Discuss replacement policy of equipments that deteriorates gradually with change in time value of money.
(b) What is preventive replacement? Find out criterion for optimal replacement time in such situation. 4+4

7. (a) Distinguish between deterministic and probabilistic models of inventory.
(b) For an inventory model, if P(r) denotes the probability of requiring r units, where r is a discrete variable, C_1 is the inventory holding cost per unit of time, C_2 is the shortage cost per unit per unit of time, then show that the stock level which minimizes the total expected cost is that value of S which satisfies the conditions:

$$\sum_{r=0}^{S-1} P(r) < \frac{C_2}{C_1+C_2} < \sum_{r=0}^{S} P(r).$$ 4+4

8. Write short notes on any two of the following: 4+4

(a) (s, S) inventory policy
(b) Duality problem in LPP
(c) Saddle point in game theory
(d) Congestion factor in Queuing model

M.Sc. Semester III Examination 2023

Subject: Statistics

Paper: MSC 34

(Time Series Analysis)

Full Marks: 40 Time: 3 hours

Answer any four of the following six questions of equal marks.

(Notations carry usual meanings)

1. (a) What is a time series? Explain the differences between time series data and cross-sectional data.
   (b) Describe with examples why time series analysis is important in real-world applications.

5+5

2. (a) Define stationarity and invertibility of a time series.
   (b) Discuss the stationarity and invertibility of the following model:

$$(1 - L)Y_t = (1 - 1.5L)\epsilon_t$$

where ϵ_t follows $WN(0, \sigma^2)$ process.

5+5

3. (a) Define auto-correlation function (ACF) of a time series. State and prove its properties.
   (b) Deduce the ACF of an MA(2) process.

5+5

4. (a) What is a periodogram? Write down the general interpretations of a periodogram.
   (b) Define the White-noise process and justify its name with the help of frequency domain analysis.

5+5

5. (a) Describes the steps involved in ARIMA modeling.
   (b) How do you select the appropriate orders (p, d, q) for an ARIMA model?

6+4

6. (a) How does the exponential smoothing differ from the simple moving average method?
   (b) Compare different exponential smoothing methods for time series forecasting.

3+7

M.Sc. Examination 2023
Semester III
Subject: Statistics (Practical)
Paper: MSC-35

Full Marks: 40 Time: 4 Hrs.

Answer all questions (*Symbols have their usual meaning*)

1. 200 boys and 100 girls were administered with intelligence test and their IQ's were determined in the following table. Write a program in R to find whether there is any evidence if sex-difference in intelligence. 5

Sex	<80	80–120	>120
Male	20	165	15
Female	12	77	11
Total	32	242	26

2. The table below presents the test-firing results for 21 surface-to air anti-air craft missiles at targets of varying speed. The result of each test is either a hit ($y = 1$) or a miss ($y = 0$). a) Fit a logistic regression model to the response variable y. Use a simple linear regression model as the structure for the linear predictor. b) Does the model deviance indicate that the logistic regression model from part a is adequate? c) Provide an interpretation of the parameter β_1 in this model. d) Expand the linear predictor to include a quadratic term in target speed. Is there any evidence that this quadratic term is required in the model? 10

Test	Target Speed (x)	y	Test	Target Speed (x)	y	Test	Target Speed (x)	y
1	400	0	8	470	0	15	280	1
2	220	1	9	480	0	16	210	1
3	490	0	10	310	1	17	300	1
4	210	1	11	240	1	18	470	1
5	500	0	12	490	0	19	230	0
6	270	0	13	430	0	20	430	0
7	200	1	14	330	1	21	460	0

3. Simulate n observations from the following mixture distribution $\alpha_1 N(\mu_1, \sigma_1^2) + \alpha_2 N(\mu_2, \sigma_2^2)$. Take the parameters of your choice and weights are such that i) Both are equal ii) First one is bigger than second one iii) First one is less than second one. Consider the statistics $T_n = \left[\frac{1}{n}\sum_{i=1}^{n} x_i^{\alpha}\right]^{1/\alpha}$, varying α as $-1.5, -1, -0.5, 0.5, 1, 1.5$. Find the bootstrap estimate of the standard error of the above estimate for all three case and draw the corresponding histogram. 10

4. Simulate a random sample from the density $f(x) = ce^{-x^2}$, $x \geq 0$. Using a proposal density $g(x) = e^{-x}$, $x \geq 0$. Also propose and justify an another $g(x)$ for simulation. 10

5. Practical note book and viva-voce. 5

M.Sc. Semester III Examination, 2023

Subject: Statistics
MSC-36(Practical)

Time: Four Hours

Full Marks: 40

Answer **all** questions. Candidates may use a computer for the Q. No. 5 and 6.

1. Maximize $z=5x_1+8x_2$

Such that $3x_1+2x_2\geq 3$

$x_1+4x_2\geq 10$

$x_1+x_2\leq 5$

$x_1\geq 0,\ x_2\geq 0$

Solve the problem graphically. 4

2. Solve the transportation problem

	DI	DII	DIII	Supply
OI	4	3	2	10
OII	1	5	0	13
OIII	3	8	5	12
Demand	8	5	4	

6

3. A self service store employs one cashier at its counter. Nine customers arrive on an average of every five minute while the cashier can serve 10 customers in every five minute. Assuming Poisson distribution for arrival rate and exponential distribution for service rate, find (i) average number of customers in the system, (ii) average number of customers in the queue and (iii) average waiting time a customer spends in the system. 4

4. A fish vendor sells fish at the rate of Rs.50 per kg on the day of the catch. He pays Rs.2 per kg of fish not sold on the day of the catch for cold storage. Fish one day old is sold at the rate of Rs.30 per kg and there is unlimited demand for it. The demand of fresh fish is known to follow a uniform distribution over the range from 30 to 50.

(a) Determine the optimum quantity of fish that should be procured by the vendor.

(b) Calculate the maximum profit. Assume that the cost of procurement is Rs.35 per kg. 4

5. Consider the production data of a factory as follows:

203, 211, 219, 230, 238, 250.

a) Forecast the series values using exponential smoothing method by taking α=0.32. Plot the original and the forecasted series on the same graph, also write down the MSE.

b) You might have noticed in the results of a) that the forecasts are not that good. Improve your forecast and show the original and the forecasted values on the same graph (You free to assume any parameter value you may need for this). Also mention the MSE in this case.

5

6. Consider the following observations:

1124.18, 1099.27, 1475.66, 1927.19, 1507.88, 1333.15, 1758.01, 1714.98, 1499.96, 1584.80, 1223.13, 1005.16, 1467.27, 1058.59, 1252.02, 1778.27, 1804.83, 1985.80, 1521.48, 1549.19, 2657.69, 1820.96, 1762.64, 1138.78, 1677.46, 2062.71, 1912.33, 2611.93, 2518.59, 2292.15, 2053.17, 3016.70, 2705.78, 2039.64, 2358.18, 1504.98, 2272.55, 1940.75, 2347.31, 3082.62, 3556.77, 2919.21, 2634.30, 2858.82, 2758.46, 2580.01, 2444.16, 2286.00, 2713.02, 2442.28, 3331.12, 3422.31, 3630.87, 3539.44, 3397.08, 3716.67, 3495.05, 3140.39, 3048.92, 2595.11.

a) Read the data as a time series data. Also show its Time Series plot.
b) Comment on the stationarity of the Time Series with reasons.
c) Fit an appropriate ARIMA model to the data and comment on your quality of fitting with reasons.
d) Forecast 15 observations using the above ARIMA fitting.
e) Plot the original and the forecasted time series using along with the error bands.

2+2+5+1+2

7. Practical Note Book and Viva-voce.

5

****

M.Sc. Examination, 2024
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) What do you understand by accelerated life test?
(d) Define availability at time t and limiting average availability of a system.
(e) Differentiate between system reliability and software reliability.
(f) For a proportional failure time model, derive the reliability function in term of the baseline reliability function.

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) Prove that hazard rate r(t) is increasing in t>0 iff $\overline{F}(t+x)/\overline{F}(t)$ is decreasing in t for each $x \geq 0$, where $\overline{F}(t)=1-F(t)$ and $F(t)$ is the distribution function of the random variable T.
(ii) When a distribution is said to belong to IFRA class? 5+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 $X_n = n.\min(T_1, T_2, ..., T_n)$ has asymptotically an exponential distribution as $n \to \infty$.
(ii) Explain type I and type II censoring used in life testing. Estimate the mean parameter of an exponential distribution using both types of censoring.

4+4

M.Sc. Examination 2024

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?

   (b) Show that for Bernoulli (θ), under conjugate prior set up, the posterior expectation of θ is a convex combination of the prior expectation and the sample average. Discuss the case when sample size $n \to \infty$.

   (c) How will you modify your results if uniform prior is considered for θ?

   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. (a) Let $\theta \sim Gamma(a, b)$ and $y_1, \cdots, y_n|\theta \sim Poisson(\theta)$. Obtain the posterior distribution of $\theta|y_1, \cdots, y_n$.

   (b) For a new observation $\tilde{y}$, find the expression of the predictive mean and variance under (a).

   6+4

4. (a) What is HPD region in Bayesian analysis?

   (b) Describe a Bayesian method to compare two groups.

   (c) Show that, in general,

   $Var_{MCMC} \geq Var_{MC}$ 2+5+3

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) What is Metropolis algorithm? How is it different from Metropolis-Hastings algorithm?

   (b) Show that Gibbs sampling is a special type of Metropolis-Hastings algorithm.

   6+4

M.Sc. Semester IV Examination 2024
Subject: Statistics
Paper: MSC 43/MSS 09
(Introductory Data Science and Statistical Machine Learning)

Full Marks: 40 Time: 3 hours

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

1. (a) What are the common techniques used in data preprocessing, and why are they essential before model training?
   (b) Describe with examples how would you handle missing values in a data set to conduct a useful analysis and model buildings.

   6+4

2. (a) What do you understand by suvervised machine learning? Describe its different types with examples.
   (b) What is ensemble learning, and why is it needed? What are the methods of ensembling different models?

   5+5

3. (a) Explain 'Boosting' with an illustrative example.
   (b) What is gradient descent algorithm? When is it used? Explain it by an example.

   5+5

4. (a) Define the terms: i) Entropy, ii) Information Gain and iii) Gini impurity
   (b) Explain how those are used in classification problems.

   3+7

5. (a) Mention a few challenges faced in Text Analysis. Also describe some practical applications of text analysis.
   (b) Define Support Vector Machine (SVM) and mention its objective. In this connection also explain the terms: Margin, Support Vectors, Regularization Parameter and Soft Margin Classification.

   4+6

6. Compare different types of decision tree algorithms with respect to their features, limitations and uses.

   10

M.Sc Semester IV Examination, 2024

Statistics

MSC-44(Practical)

Time: Three Hours Full Marks: 40

(One may use Computer, if necessary)

1. 1. Plot the failure rate function of (1) series system (2) parallel system, each consisting of two independent components, the first having a failure time distribution with i.e. $f_1(x) = 3\,exp\,(-3x)$ and the second having the i.e. $f_2(x) = 2.4\,x^2 exp\,(-0.8x^3)$ (take at least 6 points).
   Also find the Reliability of each system for a mission time 100 units. 6

2. 100 electronic tubes of a certain type were tested. The test terminated after the first 15 tubes blew. Time failures occur after the following hours:
   40,65,90,120,195,265,350,420,501,620,655,730,815,910,980.
   Estimate R(t) at t=500 hrs. 5

3. Suppose it is known that for a particular drug, the success rate (θ) varies from 0.2 to 0.6. The drug was administered to 20 (n) healthy volunteers and we observed 15 (r) success. Using this information find a prior distribution for the success rate θ. Find the predictive posterior of the number of success in next 40 trials. Find an estimate of probability of getting atleast 25 successes out of 40 trials. 7

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. 5

5. Upload the supplied data file in your system and conduct a thorough data analysis to build a predictive model. The data set have six variables namely, 'Student_Id', 'Study_Hours', 'Assignment_Score', 'Class_Attendance' (in %), 'Participation', 'Final_Exam_Score'.The answer report must include the following:
   a) Scrutinizing the data set and taking appropriate measures to handle data discrepancies.
   b) Visual analytics and interpretations of 'Study_Hours' and 'Final_Exam_Score'.

c) Is there any significant difference in 'Final_Exam_Score' between students with different levels of Participation (High, Medium, Low)? Answer with reasons.

d) Develop a predicting model to classify students' performance as 'Pass' or 'Fail' based on a threshold of 50 in the 'Final_Exam_Score' using at least two methods. Use 'Study_Hours', 'Assignment_Scores', 'Class_Attendance' and 'Participation' as predictors.

Compare the model performances for prediction using the confusion matrices. Associated R or Python code should be written on the script. 12

6. Notebook & Viva-voce 5