angirnagar University

Part IV B. Sc (Honors) Examination – 2023

Course No.: Stat- 408

Course Name: Stochastic Process

me: 02 Hours 30 Minutes

Mark: 35

Answer any Three of the following questions. Each question carries equal marks.

Q1. **(a)** Define Stochastic Process. What are the different types of this process explain with examples? [4]

(b) Explain Gaussian Process and Brownian Motion with properties. Also, mention some of the Gaussian Process's applications. [4]

(c) Consider the process $[X(t), t \in T]$ whose probability distribution, under a certain condition, is given by [11/3]

$$\Pr[X(t) = n] = \frac{at^{n-1}}{(1+at)^{n+1}} \quad , \quad n = 1,2, \ldots$$
$$= \frac{at}{1+at} \quad , \quad n = 0.$$

Test the stationarity of the process.

12:30

Q2. **(a)** Define Markov Process, Recurrent, and Transient State of a Markov Chain. State and prove the First Entrance Decomposition Formula. [4]

(b) Write down the properties of a communicate state. Explain in detail "How to find the higher order transition probabilities using Chapman-Kolmogorov equation?". [4]

(c) Let us consider the following data represents the daily average temperature for twenty-five consecutive days in Dhaka districts. Where, today's temperature depends on yesterday's temperature, not on the past. The temperature was defined by three states such as {0, 1, 2}, where 0: 26-29° C, 1: 30-33° C, and 2: 34-37° C. The data is as follows: [11/3]

1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 2, 1, 1, 1, 1, 2, 2, 2, 1, 1, 0

(i) Construct the transition probability matrix.

(ii) Draw the transition probability diagram.

(iii) What is the probability of a temperature of 26-29° C on Tuesday given that it was 30-33° C on Monday?

Q3. **(a)** Define the Counting Process along with its properties. What are the main assumption of this process explain in detail? [4]

(b) Write down the properties of a Poisson Process. Suppose $[N(t), t \geq 0]$ be a Poisson Process, then show that the autocorrelation coefficient between $N(t)$ and $N(t+s)$ is $\sqrt{\frac{t}{t+s}}$. [4]

(c) In good years, storms occur according to a Poisson process with rate 2 per unit time, while in other years they occur according to a Poisson process with rate 4 per unit time. Suppose next year will be a good year with probability 0.4. Let $N(t)$ denote the number of storms during the first t time units of next year. [11/3]

(i) Find $P\{N(t)=n\}$.

(ii) Is $\{N(t)\}$ a Poisson process? → rate of chg in rotation

(iii) Does $\{N(t)\}$ have stationary increments? Why or why not?

(iv) Does it have independent increments? Why or why not?

(v) If next year starts off with three storms by time $t = 2$, what is the conditional probability it is a good year? 0.82

P[good| Mis

Q4. **(a)** Derive the distribution of Renewal process. Under usual notations, show that Renewal process uniquely determines the distribution function. Suppose the distribution of interarrival time X_n is given by $f(x) = \lambda e^{-\lambda x}; x > 0$. Find the mean value function of the renewal process. **[4]**

(b) Under usual notations, show that the average renewal rate by time t converges with probability 1 to $\frac{1}{\mu}$ as $t \to \infty$ i.e $\lim_{t\to\infty}\left\{\frac{N(t)}{t} \to \frac{1}{\mu}\right\} \xrightarrow{W.P} 1$. **[4]**

(c) Suppose a Bluetooth headphone works on a battery. As soon as the battery is down (i.e., the charge level reaches 10%), it is recharged immediately. If X represents the lifetime of the battery (in hours) in a single charge and is distributed uniformly over the interval (1, 20), then at what rate does the device needs to be changed? **[11/3]**

Q5. **(a)** Define birth and death process with an example. Also, discuss a linear growth model with immigration.

(b) For a birth and death process let $\lambda_n = n\lambda + \theta\ (n \geq 0)$ and $\mu_n = n\mu\ (n \geq 1)$. Show that average number of people in the process at time t is $M(t) = n + \theta t$, when $\lambda = \mu$

In a birth and death process, each individual is assumed to give birth at an exponential rate of 10 per year and die at an exponential rate of 10 per year. Also, there is no increase in the population due to immigration. Explain the situation when the population size is 120. Also find the expected population size after 12 years.

(c) Define pure birth process with an example. For a pure birth process with rate λ, show that $P_{ii}(t) = e^{-\lambda t}$. **[11/3]**

Good Luck

Cov(N(t)

E(N(t))

E(N

mination – 2023

se No.: Stat- 405

Course Name: Data Mining

Time: 04 Hours

Mark: 70

Answer any Five of the following questions. Each question carries equal marks.

Q1. (a) What is meant by data mining? What are the differences between Supervised and Unsupervised Learning? Discuss these with a suitable example. [4]

(b) What is meant by KDD process? Identify and describe the phases in the KDD process. How does KDD differ from data mining? [5]

(c) What is meant by EM and jackknife estimators? Why is it important in statistical inference? Given the following set of values {1, 3, 9, 15, 20}, determine the jackknife estimate for both the mean and standard deviation of the mean. [5]

Q2. (a) What is meant by kNN? Briefly explain the different steps of kNN to classify the object. How can find the optimal numbers of k for kNN? [5]

(b) Apply the kNN algorithm to classify the item with information Sepal Length: 6.6, Sepal Width: 2.9, Petal Length: 5.6, and Petal Width: 1.2 based on the following training data for k=3. [5]

Table 1: Iris Data for kNN

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5	3.6	1.4	0.2	setosa
5.8	4	1.2	0.2	setosa
4.8	3.4	1.9	0.2	setosa
4.9	3.1	1.5	0.2	setosa
5.1	3.8	1.9	0.4	setosa
6.5	2.8	4.6	1.5	versicolor
5.6	2.9	3.6	1.3	versicolor
6.4	2.9	4.3	1.3	versicolor
5.4	3.0	4.5	1.5	versicolor
5.6	2.7	4.2	1.3	versicolor
6.5	3.0	5.8	2.2	virginica
5.8	2.8	5.1	2.4	virginica
6.7	3.3	5.7	2.1	virginica
6.1	2.6	5.6	1.4	virginica
6.7	3.3	5.7	2.5	virginica

$R_x(t)$ oRL

(c) Write down the different steps of the Random Forest algorithm for classification. [4]

Q3. (a) What do you mean by ANN? How does it differ from the perceptron algorithm? Discuss the basic structure of ANN. [4]

(b) Discuss the different steps in developing an artificial neural network. [5]

(c) How do you estimate the weight of ANN? Discuss the backpropagation algorithm. [5]

Q4. (a) What is meant by k-medoids clustering? Write down the different steps of the k-medoids clustering algorithm. What are its different advantages from other clustering techniques? [6]

(b) Apply Self Organizing Map (SOM) to cluster the A, B, C, and D data points for an iteration. Assume that the initial learning rate is 0.5 and the number of clusters to be formed is 2. [8]

Table 2: Data Point for SOM

i	A	B	C	D
1	1	0	1	1
2	1	1	0	1
3	1	1	0	0
4	0	0	0	1

$$\text{Weight matrix}, W = \begin{bmatrix} 0.2 & 0.9 \\ 0.4 & 0.7 \\ 0.6 & 0.5 \\ 0.8 & 0.3 \end{bmatrix}$$

semi supervised ml algo based to image

Q5. **(a)** What is meant by Computer vision and CNN? Explain the different steps of CNN. [5]

(b) Consider a grayscale image and a convolutional filter represented as follows: [5]

$$\text{Image} = \begin{bmatrix} 6 & 0 & 9 & 2 & 1 & 5 \\ 5 & 1 & 8 & 3 & 5 & 8 \\ 4 & 6 & 5 & 5 & 8 & 9 \\ 3 & 2 & 7 & 3 & 7 & 6 \\ 2 & 3 & 8 & 5 & 2 & 5 \\ 1 & 4 & 4 & 4 & 5 & 5 \end{bmatrix} \quad \text{and} \quad \text{Filter} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

accordin -5+4-8 -6

activation fun. (RELU)

6-5+4-8-6

Find Max Pooling and Average Pooling Feature Map using a 2x2 window with stride 2 after convolution with stride 1 and ReLU.

(c) Discuss the different types of Convolution Kernels used in CNN. [4]

Q6. **(a)** What is meant by Support Vector Machine? What are the different types of kernels used in support vector machine (SVM)? [5]

(b) Write down the different steps of support vector machine (SVM) for classification. [5]

(c) Discuss the Minimal Cost-Complexity Pruning Algorithm to prune a decision tree. [4]

Q7. **(a)** What is meant by CART analysis? How does it differ from the usual decision tree? How do you apply CART in data mining? Discuss the algorithm of CART analysis. [4]

(b) Discuss the advantages and disadvantages of ID3, C4.5 and C5.0. Are they improvement of Decision tree? How? [5]

(c) State the Naive Bayes Classifier. Classify the Height in the following Table 3 into two categories based on the median value of height – less than or equal to the median height (S), and greater than the median height (T). Consider **Output2**, [5]

(i) Compute **Information Gain** for gender and height.

(ii) Compute **Gain Ratio** for gender and height. Comments on your findings

For CNN, we rectified linear unit as activation fun. Through ReLU

Table 3: Data for Classification

Name	Gender	Height	Output1	Output2
Kristina	F	1.60 m	Short	Medium
Jim	M	2.00 m	Tall	Medium
Maggie	F	1.90 m	Medium	Tall
Martha	F	1.80 m	Medium	Tall
Stephanie	F	1.71 m	Short	Medium
Bob	M	1.86 m	Medium	Medium
Kathy	F	1.60 m	Short	Medium
Dave	M	1.70 m	Short	Medium
Worth	M	2.20 m	Tall	Tall
Steven	M	2.10 m	Tall	Tall
Debbie	F	1.80 m	Medium	Medium
Todd	M	1.95 m	Medium	Medium
Kim	F	1.90 m	Medium	Tall
Any	F	1.80 m	Medium	Medium
Wynette	F	1.75 m	Medium	Medium

-15 14 -8 0
12 -6 1 6
-11 10 -12 9
-4 13 3 -

For. retrieval extraction natural language pro Data mining

Q8. **(a)** What is meant by text mining? What are the different types of Text mining techniques? What is lexicon-based sentiment analysis? [6]

(b) Explain the methods to Compute Sentiment Scores, Lemmatization, Tokenization, Sentiment score, and VADER. [5]

(c) Explain the term Web Mining and W

Department of Statistics and Data Science
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2023
Course No.: Stat- 401
Course Name: Statistical Inference II

Time: 2 Hours 30 Minutes

Answer any Three of the following questions. Each question carries equal marks. **Mark: 35**

Q1. **(a)** Explain how the Jackknife is used to estimate the bias and variance of a parameter estimate. Compare the Jackknife to the Leave-One-Out Cross-Validation (LOOCV) approach in terms of computational requirements and applications. [4]

4.6

(b) Differentiate between the non-parametric and parametric Bootstrap approaches. Discuss the assumptions required for the validity of Bootstrap confidence intervals. [3.67]

(c) You have a dataset $X = \{x_1, x_2, \dots, x_n\}$ and the parameter of interest is the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ [4]

(i) Derive the Jackknife estimate of the bias of $\bar{X}$.

(ii) Compute the Jackknife estimate of variance for n=5 data points $X = \{2,4,6,8,10\}$.

Q2. **(a)** Define robust statistics and discuss their importance in data analysis. Why and when do you need to have robust estimators? [4]

5

(b) Discuss the impact of outliers on traditional statistical methods such as the mean and standard deviation, and explain how robust statistics mitigate these effects. [3.67]

(c) Explain the concept of *M*-estimator with its properties. Show that the maximum likelihood estimator is a special case of an *M*-estimator. [4]

Q3. **(a)** What are the different ways of evaluating the test of hypothesis? Explain each of the metrics with importance and typical range. [4]

1+4+9

14

(b) Define the concept of the most powerful (MP) test. State the Neyman-Pearson Lemma and explain its importance in hypothesis testing. Explain the concept of a uniformly most powerful (UMP) test. How does it differ from a MP test? [4]

10

(c) Let $X \sim Poisson(\lambda)$. You want to test, [3.67]

$$H_0: \lambda = 2$$
$$H_1: \lambda = 3$$

M_m^2 $\quad \beta(x-\mu)$

(i) Using the Neyman-Pearson Lemma, derive the most powerful test for this hypothesis.

(ii) Specify the critical region of the test for a significance level $\alpha = 0.05$.

S $\sqrt{\frac{n-1}{n}(\hat{\theta}-\bar{\hat{\theta}})^2}$

Q4. **(a)** Define randomized test, non-randomized test and Bayes test. Define critical function and power function and hence establish the relationship between them. Show that risk function is a linear function of the power function. [4]

~~effi~~ generizl—

(b) Let $X_1, X_2, \dots\dots\dots, X_n$ be a random sample from $f(x) = \theta e^{-\theta x}$, where x > 0 and $\theta > 0$; with the help of generalized likelihood ratio test, test $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. [4]

(c) If $X \sim N(\theta, \sigma^2)$ and σ^2 is known then test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. [3.67]

Q5. **(a)** What is sequential test of hypothesis? Discuss about sequential probability ratio test (SPRT). [5]

(b) Let k_0 and k_1 has error sizes α and β; show that k_0 and k_1 could be approximated by $k_0' = \frac{\alpha}{1-\beta}$ and $k_1' = \frac{1-\alpha}{\beta}$ and hence show that if α' and β' are the error sizes of the SPRT defined by k_0' and k_1' then $\alpha' + \beta' \leq \alpha + \beta$. [6.67]

Good Luck

Department of Statistics and Data Science
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2023
Course No.: Stat- 402
Course Name: Bayesian Inference

Time: 02 Hours 30 Minutes | Mark: 35

Answer any Three of the following questions. Each question carries equal marks.

Q1. (a) Explain Bayes' Theorem and its importance in Bayesian Inference. [3]

(b) Discuss how the concept of prior and posterior probability is applied in real-world scenarios, such as spam email filtering or medical diagnostics. [4]

C

(c) What does it mean for a prior to be "informative" and "non-informative"? Give examples. [4.67]

unifrm
prior =

Q2. (a) Let y be the number of heads in n coin flips, whose probability of heads is θ. If prior distribution for θ is uniform on the range $[0,1]$, derive the prior predictive distribution for y, $\Pr(y=k)=\int_0^1 \Pr(y=k|\theta)\,d\theta$. For each $k=0,1,\dots,n$. [3.5]

(b) Suppose you assign a $Beta(\alpha,\beta)$, and then you observe y heads out of n coin flips. Show algebraically that the posterior mean of θ always lies between the prior mean, $\frac{\alpha}{\alpha+\beta}$ and the observed relative frequency of heads, $\frac{y}{n}$. [3.5]

10

f(θ)

f(θ|)

(c) Suppose you have a Beta (4, 4) prior distribution on the probability θ that a coin will yield a 'head' when spun in a specified manner. The coin is independently spun 10 times, and 'heads' appear fewer than 3 times. Calculate the exact posterior density along with its posterior mean. [4.67]

Q3. (a) Define Bayes factor and explain how it differs from a p-value in hypothesis testing. [2]

(b) Show that Bayes factor for testing $H_0{:}\,\theta=\theta_0$ vs. $H_1{:}\,\theta=\theta_1$ is given by the ratio of the marginal likelihoods. [3]

(c) Suppose $X \sim N(\theta,4)$. Assume that H_0, H_1 are equally likely. [6.67]
To test $H_0{:}\,\theta=1\ v.s.\ H_1{:}\,\theta=2$,

(i) Compute the Bayes factor in favor of H_1 for $X=1.5$. Interpret the result.

(ii) If $n=5, \bar{x}=1.2$, the prior for θ is $N(1,2)$, calculate posterior odds in favor of H_0.

Q4. (a) Define the posterior distribution. Explain how it is derived in a Bayesian framework and describe the differences between Bayesian and frequentist approaches to parameter estimation. [3]

(b) Suppose a machine's failure time follows an exponential distribution with rate λ. Given prior knowledge that $\lambda \sim Gamma(\alpha,\beta)$ and observed data $t_1,t_2,\dots,t_n$, derive the posterior distribution of λ. [4]

(c) Define the quadratic loss function, absolute error loss function, and the LINEX loss function. Derive the Bayes estimate of θ under the quadratic loss function when $X \sim N(\theta,\sigma^2)$. [4.67]

$k+1+n-k+1$

Q5. **(a)** What is Markov chain simulation? What are the key steps involved in the Metropolis algorithm? **[4.67]**

(b) Define jumping distribution. What properties make a jumping distribution effective? Mention some common choices for jumping distribution. N, poin, Uni **[3.5]**

(c) What is the Gibbs sampler, and how does it work in Bayesian inference? How does the Gibbs sampler generate samples from a joint posterior distribution? **[3.5]**

Good Luck

Department of Statistics and Data Science
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2023
Course No.: Stat- 403
Course Name: Multivariate Analysis

Time: 04 Hours

Mark: 70

Answer any Five of the following questions. Each question carries equal marks.

Q1. (a) What is multivariate analysis? What are the main differences between dependence and interdependence techniques in multivariate analysis? [3]

(b) Define multivariate normal distribution with its likelihood function. Suppose a random vector $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \hline \mathbf{X}_2 \end{bmatrix}$ be distributed as $N_p(\boldsymbol{\mu}, \Sigma)$ with mean vector $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \hline \boldsymbol{\mu}_2 \end{bmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right]$, and $|\boldsymbol{\Sigma}_{22}| > 0$. Then find the conditional distribution $\mathbf{X}_1$ given that $\mathbf{X}_2 = \mathbf{x}_2$. [5]

(c) If $\mathbf{X}_1, \mathbf{X}_2$ and $\mathbf{X}_3$ are jointly normal with quadratic form [6]

$$Q = 3x_1^2 + 6x_2^2 + 2x_3^2 + 4x_1x_2 + 6x_1x_3 + 2x_2x_3 + 3x_1 - 2x_2 - x_3 .$$

(i) Find the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

(ii) Find the conditional distribution of $\mathbf{X}_1$ given $2\mathbf{X}_2 + 3\mathbf{X}_3$.

(iii) Are $\mathbf{X}_1$ and $2\mathbf{X}_2 + 3\mathbf{X}_3$ independently distributed? Explain.

AM

Q2. (a) What is Hotelling T^2 statistic? Show that, T^2 is a generalization of univariate t_{n-1}^2. Also, show that T^2 is invariant under Linear Transformation. [5]

(b) Let $X_1, X_2, \cdots, X_n$ be a random sample from $N_p(\mu, \Sigma)$ population. Derive the likelihood ratio test for testing $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$. And hence show that T^2 can be obtained from likelihood ratio statistics. [6]

(c) (i) Evaluate T^2 for testing $H_0: \mu_0' = [7, 10]$ using data, $X = \begin{bmatrix} 3 & 11 \\ 8 & 9 \\ 6 & 10 \\ 7 & 9 \end{bmatrix}$. [3]

(ii) Specify the distribution of T^2 for the situation in c(i).

(iii) using c(i) and c(ii), test H_0 at $\alpha = 0.05$ level. What conclusion do you reach?

2×3

Q3. (a) Describe the one-way multivariate analysis of variance (MANOVA) with its assumptions to compare several mean vectors arranged according to treatment levels. [5]

(b) How could the profile analysis be useful for managing several specific possibilities in the question of equality of mean vectors? Explain. [5]

(c) Let $n_1 = 28$, $n_2 = 28$, $\bar{\mathbf{x}}_1 = [.15 \quad -.23 \quad -.32]$, $\bar{\mathbf{x}}_2 = [.14 \quad .18 \quad .25]$ and pooled covariance matrix $S_p = \begin{bmatrix} 0.88 & 0.36 & 0.23 \\ 0.36 & 0.77 & 0.20 \\ 0.23 & 0.20 & 0.55 \end{bmatrix}$. [4]

Test for the level profiles, assuming that the profiles are coincident. Use $\alpha = 0.05$.

Q4. **(a)** Define principal component analysis (PCA) with its objectives. How does PCA handle high-dimensional data? Describe the Likelihood Ratio test based on Lawley's procedure to test the adequacy of PCA for any study. **[4]**

(b) Suppose the random vector $\mathbf{X}' = \left[\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p\right]$ have $\text{Cov}(\mathbf{X}) = \Sigma$ with the eigenvalue-eigenvector pairs $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Hence, show that the i^th^ principal component is $Y_i = \mathbf{e}_i'\mathbf{X},\ i = 1, \ldots, p$. **[5]**

6

(c) In a study of size and shape relationships for painted turtles, Jolicoeur and Mosimann (1960) measured carapace length, width, and height. They performed a principal component analysis using logarithms of the dimensions of 24 male turtles. Following are the results of PCA **[5]**

```
Importance of components:
                           PC1      PC2       PC3
Standard deviation      0.002262 0.00042 4.683e-19
Proportion of Variance 0.966680 0.03332 0.000e+00
Cumulative Proportion  0.966680 1.00000 1.000e+00
                  PC1          PC2          PC3
Length   -0.7616419 -0.08114189 -0.6428979
Width    -0.4550873 -0.63930114  0.6198303
Height   -0.4612996  0.76466336  0.4499919
```

(i) List all the principal components with their variances.

(ii) Find the number of retained principal components for this study. Hence, explain your findings.

(iii) If the variances of length, width, and height are 0.0111, 0.0064, and 0.0060, respectively, then find and interpret the maximum correlation between the 1^st^ principal components and the original variable.

Q5. **(a)** What do you mean by canonical correlation analysis? Under what circumstances would you select canonical correlation analysis instead of multiple regression as the appropriate statistical technique? **[4]**

(b) Derive the canonical variates and hence calculate the canonical correlation for $X^{(1)}$ an $X^{(2)}$ set of variables having $E\left(X^{(1)}\right) = \mu^{(1)};\ E\left(X^{(2)}\right) = \mu^{(2)};$ $COV\left(X^{(1)}\right) = \Sigma_{11};\ COV\left(X^{(2)}\right) = \Sigma_{22},\ COV\left(X^{(1)}, X^{(2)}\right) = \Sigma_{12} = \Sigma_{21}'$ **[7]**

(c) Let $X_j = \begin{pmatrix} X_j^{(1)} \\ \cdots\cdots \\ X_j^{(2)} \end{pmatrix};\ j = 1, 2, \cdots, n.$ be a random sample from an $N_{p+q}(\mu, \Sigma)$, **[3]**

$$\Sigma = \begin{bmatrix} \underset{(p\times p)}{\Sigma_{11}} & \vdots & \underset{(p\times q)}{\Sigma_{12}} \\ \cdots\cdots & \cdots & \cdots\cdots \\ \underset{(q\times p)}{\Sigma_{21}} & \vdots & \underset{(q\times q)}{\Sigma_{22}} \end{bmatrix}$$

. Test the hypothesis $H_0 : \Sigma_{12} = 0_{(p\times q)}$.

Q6. **(a)** What do you mean by factor and factor analysis? Give example. What are the purposes of factor analysis? **[3]**

(b) What is the Orthogonal Factor Model? Show that for the orthogonal factor model, the variance of observable random variables can be expressed in terms of commonality and specific variance. Also, explain the estimation procedure of the orthogonal factor model. **[6]**

(c) Suppose the m common factor model holds. Discuss the testing procedure for the adequacy of the m common factor model. **[5]**

(a) Define discrimination analysis. Derive the expression for the minimum expected cost of the misclassification rule to separate objects into two populations. [5]

(b) How does Fisher's linear discriminant function differ from other discriminant functions? [5] Find the Fisher's linear discriminant function from the following output. Hence, develop the allocation rule to classify the data into gasoline or diesel trucks using $D^2 = \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)' \mathbf{S}_p^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right) = -5.690493.$

```
Call:
lda(x[, 4] ~ x[, 1] + x[, 2] + x[, 3])
Prior probabilities of groups:
   diesel          gasoline
0.3898305  0.6101695
Group means:
              x[, 1]                    x[, 2]                    x[, 3]
diesel       10.10565            10.76217              18.167826
gasoline     12.21861             8.11250               9.590278
Coefficients of linear discriminants:
              LD1
x[, 1]  0.13374629
x[, 2] -0.07030203
x[, 3] -0.16739189
```

$\frac{d}{b+c}$ ✓

$\frac{d}{b+c}$

$$\begin{array}{c|cc} & 1 & 0 \\ \hline 1 & a & b \\ 0 & c & \boxed{d} \end{array}$$

If among 110 trucks of diesel 70 are correctly classified and among 54 gasoline trucks; 40 trucks are misclassified. Then construct the confusion matrix and hence, calculate the apparent error rate (APER). What conclusion will you make about this classification?

(c) Classify a new observation $\mathbf{x}_0$ into one of two known populations π_1, or π_2 using the minimum expected cost of misclassification (ECM) rule given the following *misclassification costs, prior probabilities and density values.* [4]

		True Populations	
		π_1	π_2
Classify as:	π_1	0	12
	π_2	25	0
Prior Probabilities		0.65	0.35
Densities at $\mathbf{x}_0$		0.64	0.77

1 0 2 0.5 1 0 1 0 1 0

Can we allocate the same population, if we assume prior probabilities are equal?

(a) Define cluster analysis. Discuss the different steps in agglomerative hierarchical and nonhierarchical cluster methods. What is a dendrogram? Explain with an example. [5]

(b) Suppose four individuals possess the following characteristics. Use "*Ratio of matches to mismatches with 1-1 matches excluded*" similarity coefficient to find the homogeneous clusters of individuals: [5]

Individual	Salary	Weight	Job Position
1	48000 Tk	64 kg	Assistant Manager
2	96000 Tk	67 kg	Manager
3	32000 Tk	73 kg	Assistant Manager
4	95000 Tk	50 kg	Manager

Create

Define four binary variables X_1, X_2, and X_3 as

$$X_1 = \begin{cases} 1 & \text{Salary} \geq 60\text{k} \\ 0 & \text{Salary} < 60\text{k} \end{cases}, \quad X_2 = \begin{cases} 1 & \text{Weight} \geq 65 \\ 0 & \text{Weight} < 65 \end{cases}, \text{ and } X_3 = \begin{cases} 1 & \text{Manager} \\ 0 & \text{Otherwise} \end{cases}.$$

How are heteroscedasticity and multicollinearity controlled in regression modeling using cluster analysis? Give examples to illustrate. [4]

Department of Statistics and Data Science
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2023
Course No.: Stat- 404
Course Name: Design and Analysis of Experiment II

A' A'' B' A'.B' A''B' B'' A'B'' A''B''

Time: 4 Hours **Mark: 70**

Answer any Five of the following questions. Each question carries equal marks.

2^4

Q1. (a) Set up a mathematical model for ANCOVA in RBD with one concomitant variable and discuss the analysis procedure of such data. [6]

(i) Justify the impact of the concomitant variable. 6

(ii) Check whether there are any differences in the effects of different levels of treatment.

6×2 ~~10~~ 12

(b) In agricultural research station an experiment is conducted to study the productivity of 2 varieties of potato using nitrogen fertilizer. The agricultural plots for cultivation are found homogeneous in respect of fertility. The potato varieties are randomly allocated to different plots. But the amount of fertilizer used (*x* kg/plot) in different plots are not same. The production of potato (*y kg*) in different plots along with amount of fertilizer used are given below: [8]

6×3 = 18

Plots	Potato 1		Potato 2	
	y	*x*	*y*	*x*
1	45	2	55	5
2	46	4	54	4
3	44	3	50	6

(i) Write down the appropriate model for this data. Justify the reason for your choice.

(ii) Complete the ANCOVA table.

(iii) Test whether the impact of the concomitant variable is homogeneous or not for all varieties of potato.

$y_{ij} = \mu + \alpha_i + \gamma_j + \epsilon_{ij}$

$f(x_{ij})$

$\frac{SS(reg)}{1} \Big/ \frac{SS(\text{error})}{p-q-1}$

6-2-1

Q2. (a) What is meant by factorial experiment? Write down the advantages and disadvantages of factorial experiment compared to single factor experiment. Discuss different types of factorial design with examples. [7]

(b) Explain Yates algorithm to calculate different component sum squares in a 2^4 factorial experiment in a CRD with 8 replications of each treatment combination. Also, present the ANOVA table and test whether the treatment combination is similar to each other. [7]

9×7-1

Q3. (a) Construct the yates table to calculate different component sum squares in a 3^2 factorial experiment in a RBD with 7 blocks. Also, present the ANOVA table. [6]

(b) An experiment was conducted using three heterogeneous plots to see the effect of nitrogen *N* and irrigation *I* on the yield of a certain variety of rice. [8]

(1+7) 3

$$N = \begin{cases} 0 & \text{for } 30 \ kg/ha \\ 1 & \text{for } 60 \ kg/ha \\ 2 & \text{for } 90 \ kg/ha \end{cases}, \quad I = \begin{cases} 0 & \text{for low level} \\ 1 & \text{for moderate level} \\ 2 & \text{for heigh level} \end{cases}$$

Possible treatment combinations and respective yields are given in the next table:

Plot					
I			II		
00 24	10 32	20 30	10 28	00 24	20 34
01 46	11 30	21 44	01 36	11 36	21 44
02 23	12 24	22 21	02 24	12 22	22 21

E = 6m²

(i) Write down the name of the design for this situation. Justify your answer.

(ii) Write the mathematical model for this data.

(iii) Estimate the components of ANOVA table. Construct the ANOVA table.

(iv) Test which the treatment combinations are similar.

Q4. (a) Explain the concept of blocking and confounding with suitable examples. Discuss the importance of blocking and confounding in experimental design. [4] 2

(b) What are the different types of confounding? Explain them with examples. Display the layout of 2^4 factorial experiment where ABCD and AB are partially confounding. [4] 96

(c) How do you analyze data obtained from such design in part (b) to test the important hypotheses? Also, set up the ANOVA table for this design. [6]

Q5. (a) Define Lattice design and Youden square design. Make a comparative study between these two designs. [6]

(b) Explain the concept of repeated measure design. Write down the scope of this design. How does this design work? Describe the benefits of repeated measures designs. [8]

Q6. (a) Describe the procedure of intra-block analysis of data obtained from a BIB design. Construct a layout plan for a BIB design having parameters b=v=13, r=k=4, λ=1. [7]

12

(b) Define an incomplete block design. For a symmetric BIBD, prove that [7]

$$\lambda = r_{ii'} \text{ where } r_{ii'} = \sum_{i=1}^{b} n_{ij} n_{i'j};\ \ i \neq i' = 1(1)b \text{ and } \lambda = \sum_{i=1}^{b} n_{ij} n_{ij'};\ \ j \neq j' = 1(1)v \text{ and}$$

also prove that, $(r-\lambda)$ is perfect square.

Q7. (a) What is split plot design? Give an example Discuss the scopes and applications of this design. [4]

(b) What do you mean by whole plot treatment error and subplot treatment error? Why do they occur? Also, discuss the importance of these two error structures for testing important hypotheses. [4]

(c) Discuss the procedure of analyzing data obtained from a split-plot design with two factors. Prepare ANOVA table. [6]

Q8. (a) Define nested design with example. Display the layout of two-stage nested design. [4]

(b) Discuss the importance of nested design. What would have happened if we incorrectly analyse the two-stage nested design as a two factor factorial experiment? [4]

(c) Discuss the procedure of analyzing data obtained from a two-stage nested design with two factors. Prepare ANOVA table. [6]

Good Luck

Department of Statistics and Data Science

Jahangirnagar University

Part IV B. Sc (Honors) Examination – 2023

Course No.: Stat- 405

Course Name: Data Mining

Time: 04 Hours **Mark: 70**

Answer any Five of the following questions. Each question carries equal marks.

Q1. **(a)** What is meant by data mining? What are the differences between Supervised and Unsupervised Learning? Discuss these with a suitable example. [4]

(b) What is meant by KDD process? Identify and describe the phases in the KDD process. How does KDD differ from data mining? [5]

(c) What is meant by EM and jackknife estimators? Why is it important in statistical inference? Given the following set of values {1, 3, 9, 15, 20}, determine the jackknife estimate for both the mean and standard deviation of the mean. [5]

Q2. **(a)** What is meant by kNN? Briefly explain the different steps of kNN to classify the object. How can find the optimal numbers of k for kNN? [5]

(b) Apply the kNN algorithm to classify the item with information Sepal Length: 6.6, Sepal Width: 2.9, Petal Length: 5.6, and Petal Width: 1.2 based on the following training data for k=3. [5]

Table 1: Iris Data for kNN

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5	3.6	1.4	0.2	setosa
5.8	4	1.2	0.2	setosa
4.8	3.4	1.9	0.2	setosa
4.9	3.1	1.5	0.2	setosa
5.1	3.8	1.9	0.4	setosa
6.5	2.8	4.6	1.5	versicolor
5.6	2.9	3.6	1.3	versicolor
6.4	2.9	4.3	1.3	versicolor
5.4	3.0	4.5	1.5	versicolor
5.6	2.7	4.2	1.3	versicolor
6.5	3.0	5.8	2.2	virginica
5.8	2.8	5.1	2.4	virginica
6.7	3.3	5.7	2.1	virginica
6.1	2.6	5.6	1.4	virginica
6.7	3.3	5.7	2.5	virginica

(c) Write down the different steps of the Random Forest algorithm for classification. [4]

Q3. **(a)** What do you mean by ANN? How does it differ from the perceptron algorithm? Discuss the basic structure of ANN. [4]

(b) Discuss the different steps in developing an artificial neural network. [5]

(c) How do you estimate the weight of ANN? Discuss the backpropagation algorithm. [5]

Q4. **(a)** What is meant by k-medoids clustering? Write down the different steps of the k-medoids clustering algorithm. What are its different advantages from other clustering techniques? [6]

(b) Apply Self Organizing Map (SOM) to cluster the A, B, C, and D data points for an iteration. Assume that the initial learning rate is 0.5 and the number of clusters to be formed is 2. [8]

Table 2: Data Point for SOM

i	A	B	C	D
1	1	0	1	1
2	1	1	0	1
3	1	1	0	0
4	0	0	0	1

$$\text{Weight matrix}, W = \begin{bmatrix} 0.2 & 0.9 \\ 0.4 & 0.7 \\ 0.6 & 0.5 \\ 0.8 & 0.3 \end{bmatrix}$$

semi supervised ML algo based for image classification

vision conv

Q5. (a) What is meant by Computer vision and CNN? Explain the different steps of CNN. [5]

(b) Consider a grayscale image and a convolutional filter represented as follows: [5]

-5+4-8
-6

$$\text{Image} = \begin{bmatrix} 6 & 0 & 9 & 2 & 1 & 5 \\ 5 & 1 & 8 & 3 & 5 & 8 \\ 4 & 6 & 5 & 5 & 8 & 9 \\ 3 & 2 & 7 & 3 & 7 & 6 \\ 2 & 3 & 8 & 5 & 2 & 5 \\ 1 & 4 & 4 & 4 & 5 & 5 \end{bmatrix} \text{ and } \text{Filter} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

activation funt. (RELU)

(-5+4-8-6
atom

Find Max Pooling and Average Pooling Feature Map using a 2x2 window with stride 2 after convolution with stride 1 and ReLU.

(c) Discuss the different types of Convolution Kernels used in CNN. [4]

peci. pio

Q6. (a) What is meant by Support Vector Machine? What are the different types of kernels used in support vector machine (SVM)? [5]

(b) Write down the different steps of support vector machine (SVM) for classification. [5]

(c) Discuss the Minimal Cost-Complexity Pruning Algorithm to prune a decision tree. [4]

pice pei.

Q7. (a) What is meant by CART analysis? How does it differ from the usual decision tree? How do you apply CART in data mining? Discuss the algorithm of CART analysis. [4]

(b) Discuss the advantages and disadvantages of ID3, C4.5 and C5.0. Are they improvement of Decision tree? How? [5]

complex [illegible]

(c) State the Naive Bayes Classifier. Classify the Height in the following Table 3 into two categories based on the median value of height – less than or equal to the median height (S), and greater than the median height (T). Consider **Output2**, [5]

(i) Compute **Information Gain** for gender and height.

(ii) Compute **Gain Ratio** for gender and height. Comments on your findings

Table 3: Data for Classification

Name	Gender	Height	Output1	Output2
Kristina	F	1.60 m	Short	Medium
Jim	M	2.00 m	Tall	Medium
Maggie	F	1.90 m	Medium	Tall
Martha	F	1.80 m	Medium	Tall
Stephanie	F	1.71 m	Short	Medium
Bob	M	1.86 m	Medium	Medium
Kathy	F	1.60 m	Short	Medium
Dave	M	1.70 m	Short	Medium
Worth	M	2.20 m	Tall	Tall
Steven	M	2.10 m	Tall	Tall
Debbie	F	1.80 m	Medium	Medium
Todd	M	1.95 m	Medium	Medium
Kim	F	1.90 m	Medium	Tall
Any	F	1.80 m	Medium	Medium
Wynette	F	1.75 m	Medium	Medium

for CNN, we rectified linear unit as activation fun. Threshold ReLU

$$\begin{bmatrix} -15 & 14 & -8 & 0 \\ 12 & -6 & 1 & 6 \\ -11 & 10 & -12 & 9 \\ -4 & 13 & 3 & -11 \end{bmatrix}$$

for retrieval
extraction
natural language process
Data mining

Q8. (a) What is meant by text mining? What are the different types of Text mining techniques? What is lexicon-based sentiment analysis? [6]

(b) Explain the methods to Compute Sentiment Scores, Lemmatization, Tokenization, Sentiment score, and VADER. [5]

(c) Explain the term Web Mining and Web mining taxonomy. [3]

Good Luck

Department of Statistics and Data Science
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2023
Course No.: Stat- 406
Course Name: Actuarial Statistics

Time: 02 Hours 30 Minutes

Mark: 35

Answer any Three of the following questions. Each question carries equal marks.

Q1. **(a)** Explain actuarial science and write down the important uses of actuarial statistics especially in the context of Bangladesh. **[11/3]**

8|7

(b) Define accumulation and amount of function. Briefly discuss simple and compound interest. Which one do you think better for developing countries like Bangladesh? **[4]**

(c) Define present value and discount. 1000 is to be accumulated by January 1, 2023, at a compound rate of discount of 9% per year. **[4]**

(i) Find the present value on January 1, 2020.

(ii) Find the value of i corresponding to d, where i is the interest rate and d is the discount rate.

$\frac{1000}{(1-0.09)^{-3}}$

753.571

(ii) 0.0989

Q2. **(a)** What is meant by nominal rate of interest 18% per year convertible monthly? Under usual notations, prove that **[5]**

$$\text{(i) } i = \left[1 + \frac{i^{(m)}}{m}\right]^m - 1 \qquad \text{(ii) } d = iv \qquad \text{(iii) } v = \frac{1}{1+i}$$

(b) Find the effective rate of interest which is equivalent to a nominal rate of interest 12% *per year convertible monthly?* **[3]**

(c) *The force of interest,* δ_t *is:* **[11/3]**

$$\delta_t = \begin{bmatrix} 0.1, & t & \leq 5 \\ 0.02t\,, & 6 < t & \leq 7 \\ 0.05, & t & \geq 8 \end{bmatrix}$$

Calculate the present value of \$100 payable at time 8.

Q3. **(a)** What is meant by Annuity? Mention some practical applications of annuities. **[11/3]**

(b) Under usual notations, show algebraically that **[5]**

3|9

(i) $a_{\overline{\infty|}} = \frac{1}{i}$

(ii) $a_{\overline{m+n|}} = a_{\overline{m|}} + v^m a_{\overline{n|}}$

(iii) $s_{\overline{m+n|}} = s_{\overline{m|}} + (1+i)^m s_{\overline{n|}}$

$\frac{(1+i)^n \cdot 1}{}$ $\frac{1-v^n}{i}$

$(1+i)^{1/12}$ $(1+\frac{i}{12})^{12}$

(c) Elroy takes out a \$5,000 loan to buy a car. No payments are due for the first 8 months, but beginning with the end of the 9th month, he must make 60 equal monthly payments. If i= 0.18, find the amount of each payment. **[3]**

134.07

$= 5000(1+j)^8$

Q4. **(a)** What is a loan amortization schedule? Consider a loan which is being repaid by equal annual payments of 1 for n years. Construct an amortization schedule. **[4]**

(b) What is a sinking fund? What is the major benefit of sinking funds? **[11/3]**

(c) A loan of L is to be repaid by sinking fund method over n years. Find the (equal) periodic payment of the sinking fund. **[4]**

$\frac{1-v^n}{i}$

$(1+j)^{c}$

Q5. **(a)** What is meant by premiums? What are the different types of premiums? Give some examples of gross premiums. [11/3]

(b) Draw a cash flow diagram of an n-year endowment insurance of face value 1 on (x). Derive an expression to find net single premium. [4]

(c) Given the following data: [4]

Mortality: Illustrative life table

x	l_x	d_x
~~26~~ 25	8,640,861	77,426
~~27~~ 26	8,563,435	83,527
~~28~~ 27	8,479,908	90,082
~~29~~ 28	8,389,826	

and $i = 0.05$. Calculate the net single premium of 3-year endowment insurance on (25) of 10,000.

$P(1+i)^t$

21 $P(1-d)^t$

Good Luck

Department of Statistics and Data Science
Jahangirnagar University
Part IV B. Sc (Honors) Examination - 2023
Course No.: Stat- 407
Course Name: Mathematical Demography

Time: 02 Hours 30 Minutes

Mark: 58

Answer any Five of the following questions. Each question carries equal marks.

Q1. (a) Define Age and sex composition with examples. How can you measure it? Write down them. Measure Age and sex composition using three different types of pyramids graphically and interpret them. [4]

(b) What is Age heaping? In which situations do you need to apply the Myers Index instead of the Whipplex Index? Describe the Whipplex Index shortly. [4]

(c) What is the difference between Cohort Diagram and Lexis Diagram? Draw a Lexis Diagram and explain it. [11/3]

Q2. (a) Define stable, stationary, and quasi-stable population. When does a stable population become a stationary population? [11/3]

(b) Prove that the death rate in a stationary population is the reciprocal to the life expectancy of birth. [4]

(c) Derive the Lotka's model for stable population theory. [4]

Q3. (a) Define cohort's component methods of population projection. In which situations will it be applicable rather than mathematical methods? Which method is the best for Bangladesh and why? [4]

(b) Write down the names of different mathematical methods. Explain them briefly and graphically. [11/3]

(c) Calculate the projected population for year 2030 using linear projection, geometric projection, and exponential projection and comments on them. [4]

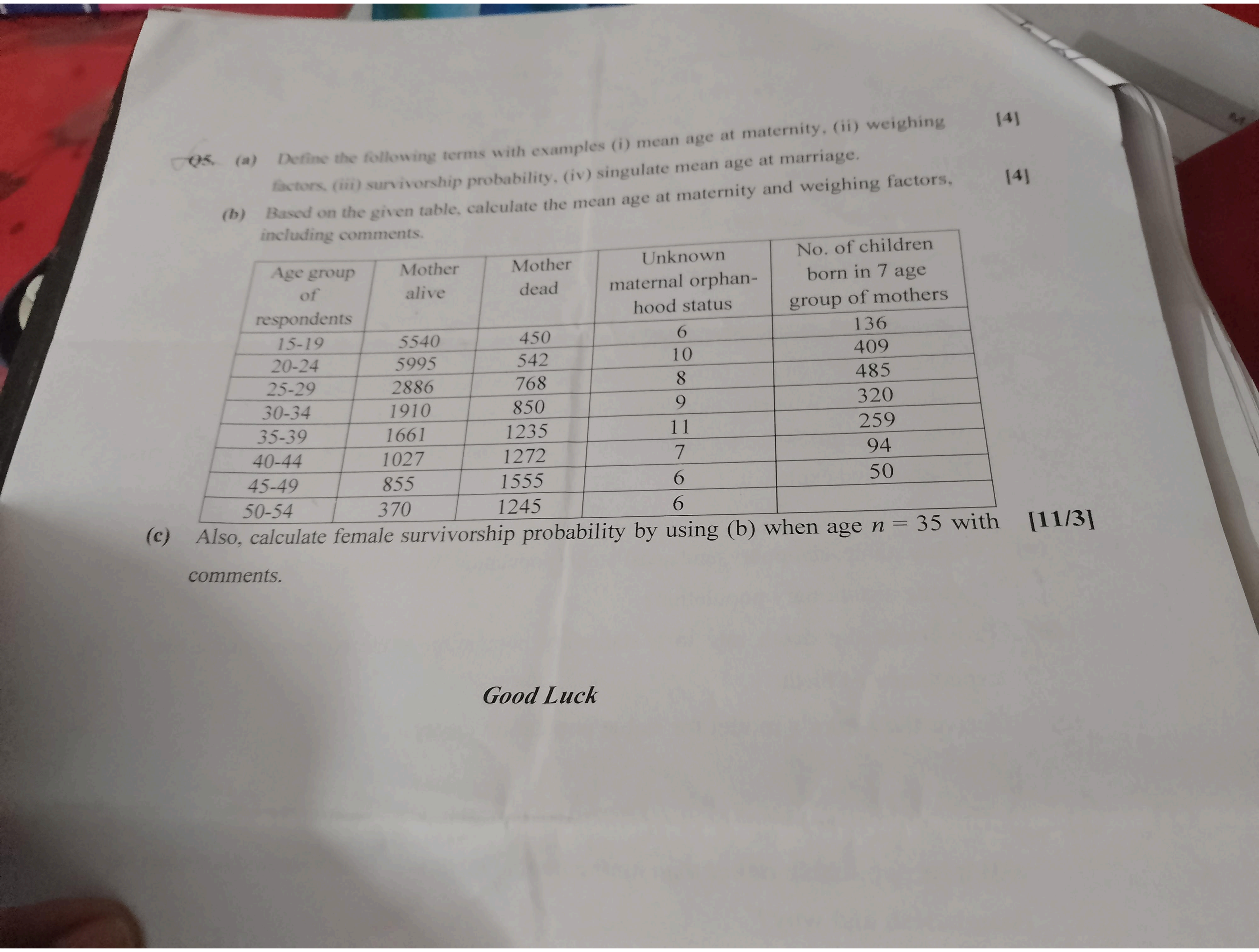
Year	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Population (Crore)	15.6	15.8	16.0	16.2	16.4	16.5	16.7	17.0	17.1	17.3
Growth Rate (%)	1.25	1.20	1.24	1.26	1.17	1.12	1.15	1.16	1.08	1.03

18.265
18.749

Q4. (a) What do you mean by fertility? What are the differentials in fertility? Display them by a conceptual frame work? [4]

(b) What are the proximate determinants of fertility as perceived by Bongaarts and Potter? Describe also the procedure of estimating the fertility indices by the Bongaarts model. [4]

(c) Suppose total fecundity rate is 15, percent of married 60, duration of postpartum infecundity is 5 months, contraceptive prevalence is 63%, and average effectiveness of contraceptives 95%. Find the total fertility rate. [11/3]



Q5. (a) Define the following terms with examples (i) mean age at maternity, (ii) weighing factors, (iii) survivorship probability, (iv) singulate mean age at marriage. [4]

(b) Based on the given table, calculate the mean age at maternity and weighing factors, including comments. [4]

Age group of respondents	Mother alive	Mother dead	Unknown maternal orphan-hood status	No. of children born in 7 age group of mothers
15-19	5540	450	6	136
20-24	5995	542	10	409
25-29	2886	768	8	485
30-34	1910	850	9	320
35-39	1661	1235	11	259
40-44	1027	1272	7	94
45-49	855	1555	6	50
50-54	370	1245	6	

(c) Also, calculate female survivorship probability by using (b) when age $n = 35$ with comments. [11/3]

Good Luck

Department of Statistics and Data Science
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2023
Course No.: Stat- 408
Course Name: Stochastic Process

Time: 02 Hours 30 Minutes

Answer any Three of the following questions. Each question carries equal marks. Mark: 35

Q1. (a) Define Stochastic Process. What are the different types of this process explain with examples? [4]

(b) Explain Gaussian Process and Brownian Motion with properties. Also, mention some of the Gaussian Process's applications. [4]

(c) Consider the process $[X(t), t \in T]$ whose probability distribution, under a certain condition, is given by [11/3]

$$\Pr[X(t) = n] = \frac{at^{n-1}}{(1+at)^{n+1}} \quad , \quad n = 1,2,\ldots$$
$$= \frac{at}{1+at} \quad , \quad n = 0.$$

12:30

Test the stationarity of the process.

Q2. (a) Define Markov Process, Recurrent, and Transient State of a Markov Chain. State and prove the First Entrance Decomposition Formula. [4]

(b) Write down the properties of a communicate state. Explain in detail "How to find the higher order transition probabilities using Chapman-Kolmogorov equation?". [4]

(c) Let us consider the following data represents the daily average temperature for twenty-five consecutive days in Dhaka districts. Where, today's temperature depends on yesterday's temperature, not on the past. The temperature was defined by three states such as $\{0, 1, 2\}$, where 0: 26-29° C, 1: 30-33° C, and 2: 34-37° C. The data is as follows: [11/3]

1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 2, 1, 1, 1, 1, 2, 2, 2, 1, 1, 0

(i) Construct the transition probability matrix.

(ii) Draw the transition probability diagram.

(iii) What is the probability of a temperature of 26-29° C on Tuesday given that it was 30-33° C on Monday?

Q3. (a) Define the Counting Process along with its properties. What are the main assumption of this process explain in detail? [4]

(b) Write down the properties of a Poisson Process. Suppose $[N(t), t \geq 0]$ be a Poisson Process, then show that the autocorrelation coefficient between $N(t)$ and $N(t+s)$ is $\sqrt{\frac{t}{t+s}}$. [4]

(c) In good years, storms occur according to a Poisson process with rate 2 per unit time, while in other years they occur according to a Poisson process with rate 4 per unit time. Suppose next year will be a good year with probability 0.4. Let $N(t)$ denote the number of storms during the first t time units of next year. [11/3]

(i) Find $P\{N(t)=n\}$.

(ii) Is $\{N(t)\}$ a Poisson process? → rate of chg in rotation

P[good | Mi...

(iii) Does $\{N(t)\}$ have stationary increments? Why or why not? ≈

(iv) Does it have independent increments? Why or why not?

(v) If next year starts off with three storms by time $t = 2$, what is the conditional probability it is a good year? 0.82

Q4. **(a)** Derive the distribution of Renewal process. Under usual notations, show that Renewal process uniquely determines the distribution function. Suppose the distribution of interarrival time X_n is given by $f(x) = \lambda e^{-\lambda x}; x > 0$. Find the mean value function of the renewal process. **[4]**

(b) Under usual notations, show that the average renewal rate by time t converges with probability 1 to $\frac{1}{\mu}$ as $t \to \infty$ *i.e* $\lim_{t\to\infty}\left\{\frac{N(t)}{t} \to \frac{1}{\mu}\right\} \xrightarrow{W.P} 1$. **[4]**

(c) Suppose a Bluetooth headphone works on a battery. As soon as the battery is down (i.e., the charge level reaches 10%), it is recharged immediately. If X represents the lifetime of the battery (in hours) in a single charge and is distributed uniformly over the interval (1, 20), then at what rate does the device needs to be changed? **[11/3]**

Q5. ***(a)*** *Define birth and death process with an example.* Also, discuss a linear growth model with immigration. **[4]**

(b) For a birth and death process let $\lambda_n = n\lambda + \theta\ (n\geq 0)$ and $\mu_n = n\mu\ (n\geq 1)$. Show that average number of people in the process at time t is $M(t) = n + \theta t$, when $\lambda = \mu$. **[4]**

In a birth and death process, each individual is assumed to give birth at an exponential rate of 10 per year and die at an exponential rate of 10 per year. Also, there is no increase in the population due to immigration. Explain the situation when the population size is 120. Also find the expected population size after 12 years.

(c) Define pure birth process with an example. For a pure birth process with rate λ, show that $P_{ii}(t) = e^{-\lambda t}$. **[11/3]**

Good Luck

Cov(N(t), N(t+s))

W.R E(N(t), N(t

≈ ft + (

≈ ft

E(N(

≈ f

Department of Statistics and Data Science
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2023
Course No.: Stat- 409
Course Name: Genetics and Bioinformatics

Time: 04 Hours **Mark: 70**

Answer any Five of the following questions. Each question carries equal marks.

Q1. (a) What do you mean by Bioinformatics, and what are the various goal of it? Why Bioinformatics in Statistics? [5]

(b) Describe the central dogma of life in the context of bioinformatics as well as molecular biology. Clarify the following concepts: [4]

i) Gene ii) Trait iii) Chromosome iv) Protein

(c) Consider the following base sequence of a given DNA segment: [5]

T A C A A T G A T C T G A C G A T T

(i) What is the sequence of the complementary DNA strand? 3' 5'

(ii) Given the original template strand, what would be the sequence of an RNA strand after transcription?

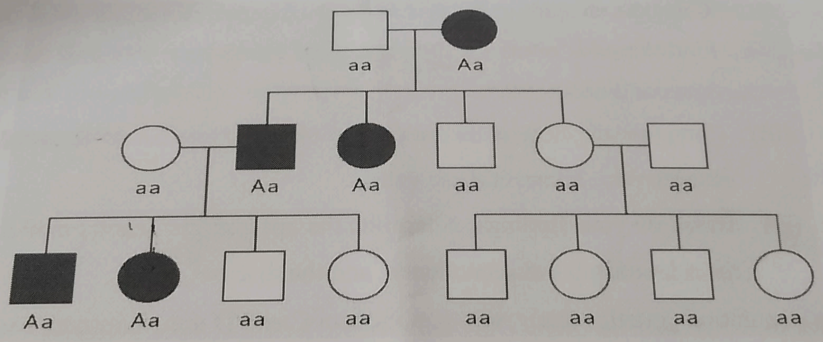
(iii) What would be the result of translation?

A→T

Q2. (a) What is a pedigree? Sketch an ancestry chart for 5 generations, where disease transmission is due to X-chromosomal dominant fashion but showing the reduced penetrance in the 3*rd* generation. [5]

(b) With suitable examples distinguish between, (i) autosomal dominant vs autosomal recessive inheritance pattern, (ii) dominant vs recessive allele. [4]

(c) [5]



Analyze the pedigree to identify the inheritance pattern and compare its characteristics with the expected traits of a specific Mendelian inheritance mode in the family.

d n
(i) even gen af
(ii) [illegible]
(ii) affected half for the child if father in aff

Q3. (a) Define penetrance function. What are the important assumptions for developing general probabilistic model in genetics? Create a table showing genotype penetrance for dominant, recessive, and codominant Mendelian inheritance patterns. [5]

(b) Design a comparative table illustrating the Genotypic Relative Risk (GRR) for the genotypes across four genetic risk models: Recessive, Dominant, Additive and Multiplicative under the assumption of phenocopies. Ensure the table highlights the progression of risk within each model and accounts for both genetic and environmental contributions. [4]

(c) How does the penetrance rate influence the Genetic Relative Risk (GRR) for a genetic disorder? For a disorder with a prevalence of 1 in 20,000, calculate the GRR when the penetrance rates are 70% and 85%. Compare the results and explain the impact of penetrance on GRR. [5]

Q4. (a) Define SNP, and how does it differ from general polymorphism? Can you explain genotyping and what are the various genotyping methods, and how to select the appropriate genotyping method? [4]

22

(b) What is Hardy-Weinberg Equilibrium (HWE), and what factors can disrupt it? Under the general condition of HWE, [6]

$$P(A_1) = p_{11} + \frac{1}{2}p_{12} \text{ and } P(A_2) = p_{22} + \frac{1}{2}p_{12}\text{, where } p + q = 1.$$

p^2+q^2+2pq

Find the relationship between genotype and allele frequency.

$(0.5)^2 + (0.1)^2 + 0.5$

>1

0.5 0.5

(c) If the frequency of the heterozygous genotype $f(Aa)$ is 0.50, what are the allele frequencies $f(A)$ and $f(a)$ assuming Hardy-Weinberg equilibrium? [4]

p q

LP, case po sta sta Antifact Ren-ion-int

Q5. (a) Define genetic association. What are the reasons for genetic association? Discuss the different study design in genetic association study. → case, cohort [4]

(b) A study investigates the association between **GeneA** and the risk of developing a certain disease. There are three genotypes: $\boldsymbol{aa, aA, and\ AA}$. The following table shows the number of individuals with and without the disease: [5]

$2pq = 0.50$

$pq = 0.25$

Genotype	Disease (Affected)	No Disease (Not Affected)	Total
aa	10	90	100
aA	40	60	100
AA	45	55	100

$p^2+q^2+2pq=1$

$p^2+q^2=0.5$

$(p^2-q^2)^2 = (p^2+q^2)^2 - 4pq$

(i) Calculate the odds of the disease for each genotype and interpret the results.

(ii) Find the odds ratios for aA vs. aa and AA vs. aa. Interpret these ratios in terms of disease risk.

$(0.5)^2+(0.5)^2$

$+0.5$

(iii) Compute the log odds ratios θ_1 and θ_2 for $aA\ vs. aa\ and\ AA\ vs. aa$, respectively. Interpret these values.

2

(iv) Based on your findings, determine the appropriate genetic model to describe the relationship between GeneA and the disease.

codominant

(c) In a population genetics study, two genetic loci C and D are examined, each with two alleles: C₁, C₂ and D₁, D₂. The observed haplotype frequencies are presented in the following contingency table: [5]

$2pq = 0.5$

$p^2+q^2+2pq=1$

	D₁	D₂
C₁	0.30	0.20
C₂	0.10	0.40

$p^2+q^2=0.5$

Using these frequencies, answer the following:

$(p+q)^2 - 2pq = 0.5$

(i) Calculate the marginal allele frequencies for C₁, C₂, D₁, and D₂. 0.5, 0.5, 0.4

(ii) Compute the linkage disequilibrium coefficient D using the provided haplotype frequencies. Interpret its value in terms of genetic association.

$2pq = 0.5$ $= 1$ $(p+q)^2 - 4 \times pq$

(iii) Calculate the normalized LD measure D' and the squared correlation coefficient r^2. Discuss the biological significance of these measures in understanding genetic linkage and association.

Q6. **(a)** What is the Genome-Wide Association Studies (GWAS)? Mention different imputation algorithms used in GWAS. Explain one of them. **[5]**

(b) Suppose, there is a gene named APOC3 in a given genome having 10 SNPs for n individuals. Discuss the steps of association testing in the context of a case-control setting. **[5]**

(c) What is a GWAS catalog? Explain the main features of such catalog. What types of information are stored here? **[4]**

Q7. **(a)** Make a comparative analysis among the available types of biological databases with suitable example. **[5]**

(b) Mention some database search algorithms used in Bioinformatical research. What are the existing repository guidance for the Nucleic acid sequence data according to the Journal of scientific data of NARURE? **[5]**

(c) List different features of the following Biological databases: **[4]**

(i) TREMBL, (ii) PIR, (iii) DDBJ, and (iv) SWISS PROT

Q8. **(a)** What are the primary machine learning techniques, Why ML in Bioinformatics and how are they applied in analyzing biological data? **[5]**

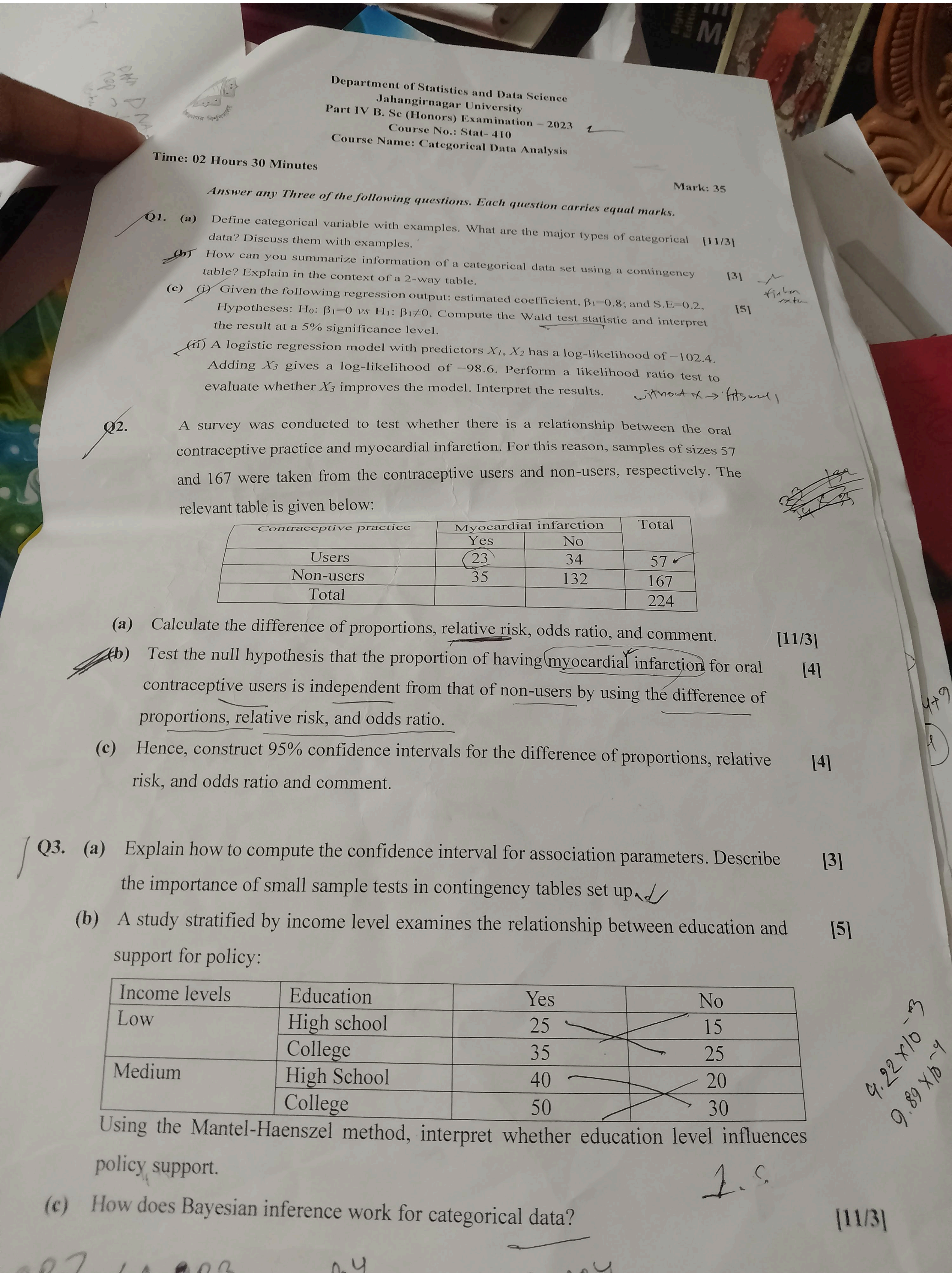
(b) *How does a Support Vector Machine (SVM) work, and what are its key applications in bioinformatics?* **[4]**

(c) *What is the difference between artificial neural networks (ANNs) and deep learning models? How can neural networks contribute to the development of new predictive models in bioinformatics? Outline the methods and tools that are commonly used for integrating machine learning techniques in bioinformatics research.* **[5]**

Good Luck

A11 - P1q1

A1A2
A1A2
A2



Department of Statistics and Data Science
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2023
Course No.: Stat- 410
Course Name: Categorical Data Analysis

Time: 02 Hours 30 Minutes

Mark: 35

Answer any Three of the following questions. Each question carries equal marks.

Q1. **(a)** Define categorical variable with examples. What are the major types of categorical data? Discuss them with examples. **[11/3]**

(b) How can you summarize information of a categorical data set using a contingency table? Explain in the context of a 2-way table. **[3]**

(c) (i) Given the following regression output: estimated coefficient, $\beta_1=0.8$; and S.E=0.2, Hypotheses: H_0: $\beta_1=0$ *vs* H_1: $\beta_1\neq0$. Compute the Wald test statistic and interpret the result at a 5% significance level. **[5]**

(ii) A logistic regression model with predictors X_1, X_2 has a log-likelihood of −102.4. Adding X_3 gives a log-likelihood of −98.6. Perform a likelihood ratio test to evaluate whether X_3 improves the model. Interpret the results. without x → fits well

Q2. A survey was conducted to test whether there is a relationship between the oral contraceptive practice and myocardial infarction. For this reason, samples of sizes 57 and 167 were taken from the contraceptive users and non-users, respectively. The relevant table is given below:

Contraceptive practice	Myocardial infarction		Total
	Yes	No	
Users	23	34	57
Non-users	35	132	167
Total			224

(a) Calculate the difference of proportions, relative risk, odds ratio, and comment. **[11/3]**

(b) Test the null hypothesis that the proportion of having myocardial infarction for oral contraceptive users is independent from that of non-users by using the difference of proportions, relative risk, and odds ratio. **[4]**

(c) Hence, construct 95% confidence intervals for the difference of proportions, relative risk, and odds ratio and comment. **[4]**

4+9

Q3. **(a)** Explain how to compute the confidence interval for association parameters. Describe the importance of small sample tests in contingency tables set up. **[3]**

(b) A study stratified by income level examines the relationship between education and support for policy: **[5]**

Income levels	Education	Yes	No
Low	High school	25	15
	College	35	25
Medium	High School	40	20
	College	50	30

Using the Mantel-Haenszel method, interpret whether education level influences policy support.

9.22×10^{-3}

9.89×10^{-4}

1.5

(c) How does Bayesian inference work for categorical data? **[11/3]**

Q4. **(a)** What do you mean by generalized linear models (GLMs)? What are the components of this model? Discuss them. Briefly explain the role of link function in a generalized linear model. **[11/3]**

(b) Suppose, the response variable Y follows Bernoulli distribution with probability mass function: $f(y;n,p)=\binom{n}{y}p^y(1-p)^{1-y}$; $y=0,1,...,n$ **[3]**

Show that, $f(y;n,p)$ can be written in the canonical GLM form. Write down the expression for the canonical parameter.

(c) You are given the following dataset from a survey of urban households: Outcome Variable: Energy Usage (categorical: Low, Medium, High), Predictors: Household Size (continuous), Income Level (ordinal: Low, Medium, High), and Appliance Usage (count). **[5]**

multinor logistic

$P\left(\frac{y=1}{M=1}\right)=\beta_0 + \beta x_1$

(i) Propose an appropriate GLM to analyze this data. Justify your choice.

(ii) If you were to fit a multinomial logistic regression model, how would you interpret the coefficients for Income Level?

Q5. *A healthcare dataset considers the following study variables: the number of Doctor Visits (count), Patient Age, Chronic Illness Status (Yes/No), and Year of Observation.*

(a) Which GLM among the Logistic, Poisson, or negative binomial is appropriate for the variables mentioned above? Why? **[4]**

(b) If the data is longitudinal, explain how you would incorporate random effects or use GEE to analyze this data. **[11/3]**

(c) A GEE model predicts disease status (Yes/ No) over 5 visits, with two predictor variables: treatment, and baseline health. The correlation structure is exchangeable. Explain why GEE is appropriate for this scenario and how the correlation structure affects the analysis. **[4]**

Good Luck

Department of Statistics and Data Science
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2023
Course No.: Stat- 411
Course Name: Financial Statistics

02 Hours 30 Minutes **Mark: 35**

Answer any Three of the following questions. Each question carries equal marks.

Q1. **(a)** Explain the primary reasons why firms should engage in risk management. How do these reasons challenge traditional investment theories? **[3]**

(b) Give definition of market risk, liquidity risk, operational risk, credit risk, and business risk. **[11/3]**

(c) What are the key stylized facts about financial asset returns, and how do they influence the modeling of risk in financial market? **[5]**

Prove that the portfolio rate of return is $r_{PF,t+1} \equiv \sum_{i=1}^{n} w_i r_{i,t+1}$.

Q2. **(a)** Describe primary and secondary marketplaces using examples. Briefly discuss the efficient-market hypothesis. What does the stock market mean when you say "bullish" or "bearish"? **[3]**

(b) How to illustrate log returns versus simple returns. Suppose, if an asset's value increases from \$100 to \$120 in one year, what would the simple return and log return be? Discuss the importance of asset return analysis in investment decision making. **[11/3]**

OLS

(c) What is volatility? How to model volatility with ARCH for financial time series forecasting? **[5]**

Q3. **(a)** What is GARCH variance model? How does the Risk Metrics variance model differ from the GARCH model in forecasting daily volatility? Discuss the benefit of GARCH model over Risk Metrics model. Derive the variance of the daily returns k days ahead using GARCH model. **[4]**

$P_t - P_y$

(b) Obtain the maximum likelihood estimator for GARCH model. **[4]**

(c) How would you perform diagnostic checking for the estimated variance model? Explain the process. **[11/3]**

Q4. **(a)** What is a technical indicator? How to use resistance and support lines in trading. Describe the bullish and bearish divergences in the MACD (moving average convergence divergence). **[3]**

(b) How does the Relative Strength Index (RSI) operate? How to identify overbought and oversold conditions with RSI. Let's say a stock's price changes over 14 days as follows: **[11/3]**

Gains: 3, 2, 1, 4, 5, 3, 5, 6

Losses: 1, 2, 3, 2, 1, 2

(c) Describe briefly the Dow theory. How Elliott Waves and Dow theory work. Explain some of the common patterns of impulse and corrective waves. [5]

Q5. (a) Define portfolio variance. How to evaluate a single asset's and a portfolio's value at risk. [11/3]

(b) What is an option? Examine the basic characteristics of the option. Suppose, Trader X thinks that the share price of a Company – which is currently trading at $250 per share – will rise over the next month. Trader X finds a call option contract with a strike price of $265 and an expiration date of exactly two months' time. The contract covers 100 shares and has a premium (or price) of $5.60. [4]

(i) How much profit has Trader X made on this trade?

(ii) What would have happened if the Company had remained below than $265 per share strike price when the contract expired?

(c) Assume that it is January 6, 2025 and the stock is trading at $20.50. Consider the stock with a strike price of $20, expiring in 3 months. Using past stock prices, the volatility in the stock prices is estimated at 60%. The riskless rate is 4.63%. Calculate the value of the call option with Black-Scholes model. [4]

Good Luck

$P_{21} = 120$

$P_{y} = 100$

$\log\left(\frac{P_{21}}{P_{y}}\right) = \log P_1 - \log P_2$

[illegible] of