Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Hons.) Examination 2017
Course Title: *Statistical Inference II*
Course No. Stat 401

Time: 4 Hours   Full Marks: 70

[Answer any FIVE questions. All questions carry equal marks.]

1. a) Discuss and derive Bhattacharyya lower bound.

   b) Prove that minimum variance estimator is unique.

2. a) What is equivariance estimator? If σ be equivariant for estimating θ with loss function $L(\theta, d) = \rho(d - \theta)$, then show that the bias, risk and variance of σ are all constant.

   b) Define vector of parameter and Wilk's generalized variance.

   c) What are Pitman estimator for location and Pitman estimator for scale? If a sample $X_1, X_2, \ldots, X_n$ is taken from uniform distribution over the interval $\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$, where θ is a location parameter, then find the Pitman estimator of θ.

3. a) What is meant by sufficient statistic? State when a sufficient statistic becomes a complete sufficient statistic. Show that a complete sufficient statistic is a minimal sufficient statistic.

   b) Suppose that a random sample is drawn from a distribution with parameter θ having a prior distribution $\pi(\theta)$ and that T is a sufficient statistic for θ. Show that the posterior distribution of θ given the random sample is identical to the posterior distribution of θ given T.

   c) Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, where both the parameters are unknown. Obtain a sufficient statistic for $\theta = (\mu, \sigma^2)$. Show that it is a complete statistic.

4. a) Define posterior Bayes estimator, Bayes loss and Bayes risk, conjugate prior and Jeffreys' non-informative prior.

   b) Discuss different ways of constructing conjugate prior. How can you check the existence of conjugate prior?

5. a) Explain the concept of median and modal unbiased in estimation. Discuss the situation when these estimators are appreciable.

   b) Let $X_1, \ldots, X_{2n+1}$ be independent random variables with a common density function: $f(x;\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$; $x > 0$, $\theta > 0$. Obtain the median unbiased estimate of θ.

   Let $Y_1 = \frac{x_{(n)}}{\ln n}$ and $Y_2 = \frac{\sum_{i=1}^{n} x_i}{n-1}$, where $x_{(n)} = Max X_i , i \le i \le n, n > 1$. Show that Y_1 and Y_2 both are modal unbiased estimate of θ.

6. a) Define Bayes test, randomized test and non-randomized test. Also, 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.

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

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

7. a) Discuss Bayesian hypothesis testing procedure. What are prior and posterior odds?

b) Discuss Bayesian analysis of contingency table.

c) Consider $X_i ; i = 1,2$ independent random variables follow $B(n_i, P_i)$. Test $H_0 : P_1 = P_2 = P_0$ versus $P_1 \neq P_2$ using an independent uniform prior, where $P_0 = \frac{1}{2}$, And hence compare with the classical method.

8. a) Explain the concepts, applications and justification of resampling methods. State, in brief, the different types of resampling.

b) Explain the procedure for jackknife point and interval estimations. Why jackknife method of resampling is also known as the leave-one-out method?

c) A random sample of 6 observations from a given population resulted in the following data: 5.2, 5.7, 4.6, 6.5, 8.6 and 2.9.

i) Find a jackknife point estimate of the population mean μ.

ii) Construct a 95% jackknife confidence interval for the population mean μ.

Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Hons.) Examination 2017
Course Title: ***Multivariate Analysis and Categorical Data Analysis***
Course No. 402

Time: 4 Hours   Full Marks: 70

[Answer any FIVE questions. All questions carry equal marks.]

1. (a) What do you mean by multivariate analysis? Explain with suitable real life example. What is multivariate normal distribution? State the properties of multivariate normal distribution.

(b) Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are iid $N_p(0, \Sigma)$. Find the maximum likelihood estimate (MLE) of Σ. Show that it is an unbiased estimator of Σ.

(c) Let $\mathbf{X}$ be $N_3(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu}' = [3 \quad 1 \quad 4]$ and $\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$.

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

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

2. (a) How can you assess the assumption of univariate and multivariate normality? Discuss Chi-square plot of assessing bivariate normality? Discuss the steps of detecting outliers of higher dimension of multivariate data set.

(b) If $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be independent observations from a population with mean $\boldsymbol{\mu}$ and finite covariance $\boldsymbol{\Sigma}$, then show that $n(\overline{\mathbf{X}} - \boldsymbol{\mu})' \mathbf{S}^{-1} (\overline{\mathbf{X}} - \boldsymbol{\mu})$ is approximately distributed as χ_p^2 for large $n - p$.

(c) The following data consists of two different measures of stiffness, and on each of $n = 3$ boards. Here, x_1 = sending a shock wave done the board, x_2 = measurements when vibrating the board.

Table 1 : Two measurements of stiffness with standardized values

x_1	1889	2403	2119
x_2	1651	2048	1700

Explain the procedure of assessing bivariate normality in the context of the above problem.

3. (a) What is a Hotteling T^2 statistics? Discuss the use of Hotteling T^2 in multivariate analysis.

(b) Derive the distribution of Hotteling T^2 along with its central probability distribution as an extension of the univariate t-distribution. Show that T^2 is invariant under changes in the units of measurements.

(c) A wildlife ecologist measured x_1 = tail length (in millimeters) and x_2 = wing length (in millimeters) for a sample of $n = 6$ female hook-billed kites (a bird in the family Accipitridae). These data are displayed in the following.

x_1 (tail length)	191	197	208	180	180	196
x_2 (wing length)	284	285	288	273	276	288

Department of Statistics

Jahangirnagar University

Part IV B. Sc. (Hons.) Examination 2017

Course Title: Design of Experiment and Analysis of Variance

Course No. Stat 403

e: 4 Hours Full Marks: 70

[Answer any FIVE questions. All questions carry equal marks.]

) What is design of experiment? What are the criteria needed for a good experiment? Examine how far these criteria are satisfied by basic experimental design.

) What do you mean by contrast, orthogonal contrast and mutually orthogonal contrast? Describe the procedures of testing the following hypotheses in completely randomized design with k treatments.

i) $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$

ii) $H_0: \mu_1 = \mu_2$

) Define Latin square design. Make a comparative study among different types of Latin square design. In performing a 2×2 Latin square design what problems do you face to analyze the data? How do you overcome these problems? Explain clearly.

) An experimenter wishes to compare k treatments with k^2 experimental units where fertility variation occurs along two perpendicular directions. Suggest a suitable design to carry out the experiment. Again, if there is one missing observation, then how do you proceed to analyze the data of this design? Estimate the efficiency of this design in comparison to other basic designs.

) For the model: $y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ij}$; $i = 1(1)p$, $j = 1(1)q$, with usual assumptions, conditions, variances and covariances and the distributions of the estimated values of the parameters of the model and if $(\alpha\beta)_{ij} = 0$, then show that

$$z = \frac{\sum_i \sum_j \widehat{(\alpha\beta)}_{ij} \hat{\alpha}_i \hat{\beta}_j}{\sqrt{\sum_i \hat{\alpha}_i^2 \sum_j \hat{\beta}_j^2}}$$

is distributed as normal with mean 0 and variance σ^2.

) Distinguish between additive and non-additive models. How do you test the additivity of a model in case of two-way classified data?

) What is multiple comparison test? Explain the necessity of this test in design of experiment. Discuss different multiple comparison tests.

) What is Graeco-Latin square design? How does it differ from Latin square design? Perform the analysis of data of this design.

) What do you mean by analysis of covariance? How does it differ from analysis of variance? Set up a linear model for ANCOVA of the data of randomized block design with two concomitant variables with necessary assumptions. How do you analyze the data of this experiment?

) How do you justify the usefulness of concomitant variables of the ANCOVA design? Display the ANCOVA table.

Define factorial experiment. How does it differ from single factor experiment? Distinguish between symmetrical and asymmetrical factorial experiments.

Interpret main and interaction effects of a factorial experiment with factors: A, B and C each at 2 levels with the procedure of analyzing data.

7. a) "Confounding is used to reduce the block size, not to reduce the size of the experiment." Ju the statement.

b) Explain the terms: total confounding, partial confounding and simultaneous confounding examples.

c) Display the layout of 2^5 factorial experiment in which interaction ABC and ADE confounded simultaneously and hence discuss the method of analysis of data.

8. a) Define an incomplete block design. When balanced incomplete block design (BIBD) symmetrical balanced incomplete block design (SBIBD) are obtained? For a SBIBD with parameters, show that $(r-\lambda)$ must be a perfect square.

b) Construct a SBIBD with parameters: $b = v = 4\lambda + 3$, $r = k = 2\lambda + 1$ and $\lambda = 1$. Discuss procedure of ANOVA with recovery of intra-block information.

Best of Luck

Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Hons.) Examination 2017
Course Title: *Sample Survey*
Course No. Stat-404

Time: 4 Hours Full Marks: 70

[Answer any FIVE questions. All questions carry equal marks.]

1. **(a)** Describe which situations warrant the use of varying probability sampling in place of simple random sampling?

   (b) Describe Lahiri's method for selecting a varying probabilities with replacement (WR) sample. Why and when, is this method preferred over cumulative total method?

   (c) What is Horvitz-Thompson (H-T) estimator? Discuss its merits and demerits. Find the variance of H-T estimator for the population mean and find its unbiased estimator.

2. **(a)** Distinguish between separate ratio estimator and combined ratio estimator.

   (b) If the total sample size n is large and the simple random sampling without replacement is done in each stratum independently, then show that the combined ratio estimator is a consistent estimator and find its sampling variance.

   (c) Explain the conditions under which the ratio estimator is a best linear unbiased estimator.

3. **(a)** Explain with example the difference between cluster and strata. Discuss the different types of cluster sampling. Differentiate among cluster sampling, simple random sampling and stratified random sampling.

   (b) Explain intra-cluster correlation coefficient and its effect in variance. State how it can be determined. A simple random sample of n clusters each containing M elements is drawn from N clusters using without replacement sampling in the population. Then show that the sample mean per element $\bar{\bar{y}}$ is an unbiased estimate. The mean per element in the population $\bar{\bar{y}}$ with variance $V(\bar{\bar{y}})=\frac{1-f}{n}\frac{NM-1}{M^2(N-1)}s^2[1+(M-1)\rho]$, under usual notations.

   (c) Determine the optimal sample size for estimating population mean for single stage cluster sampling of equal cluster size.

4. **(a)** Describe the concept of double sampling. In what kind of situations does the use of double sampling become necessary? What are the negative features of double sampling? Discuss.

   (b) Find the estimator of population mean in case of double sampling for regression method of estimation.

   (c) Find its bias and estimator of mean square error.

5. **(a)** Write the assumptions for the estimation of mobile population using capture-recapture principle. What is direct sampling and how it differs from inverse sampling?

   (b) Describe the negative hypergeometric model for estimating mobile populations. Obtain the Bailey's unbiased estimator of population size and find its unbiased estimator of variance of the estimator of population size.

(c) Show that the inverse sampling method is little more efficient than the direct method.

6. (a) Discuss the concept of inclusion probabilities. State its importance in sampling. What are 1st and 2nd order inclusion probabilities?

(b) For any sampling design $P(\cdot)$, show that $E_p[n(s)] = \sum_{i=1}^{N} \pi_i$, and hence prove that, for $P[n(s) = n] = 1$ and all values of s;

(i) $n = \sum_{i=1}^{N} \pi_i$,

(ii) $\sum_{i=1}^{N} \pi_i [\pi_i \pi_j - \pi_{ij}] = \pi_i (1 - \pi_i)$ under usual notations.

(c) For the following distribution of coconut trees,

Plot Number	1	2	3	4	5	6	7
Number of Coconut Trees	16	25	40	10	18	20	15

a sample is drawn using the following sampling design

$$P(s) = \begin{cases} \frac{1}{30} & \text{if } n(s) = 2 \\ \frac{1}{40} & \text{if } n(s) = 3 \\ 0 & \text{otherwise.} \end{cases}$$

Estimate the total number of trees in the population using the Horvitz-Thompson estimator assuming the set {2, 6, 7} is selected as sample and also estimate the variance of the estimate.

7. (a) Distinguish between two-phase sampling and two-stage sampling. And state how two-phase sampling is better than any other sampling.

(b) Under two-phase PPS sampling, find that unbiased estimator of population mean $\bar{y}$ and unbiased estimator of population variance of this estimator.

(c) Obtain the optimum sizes of the first and second sampling using simple linear cost function.

8. (a) Discuss the basic reasons for viewing non-response as a problem. What kind of surveys appears prone to high non-response rates? Distinguish with example between element non-response and item non-response.

(b) What is measurement error? How does this error arise? And how it can be controlled? How are sampling error and measurement error related to the total error in a survey?

(c) What is randomized response method? Under what circumstances would it be necessary to employ this method? Describe the method for the case when it is desired to estimate the proportion P of individuals who belong to some sensitive category by means of a survey with personal interview.

- Best of Luck -

Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Hons.) Examination 2017
Course Title: ***Data Mining***
Course No. Stat-405

Time: **2.5** Hours　　　　　　　　　　Full Marks: 35

[Answer any THREE questions. All questions carry equal marks.]

1. **(a)** What do you mean be the data mining? How does data mining access of a database differ from its traditional counterpart?

(b) Distinguish between predictive and descriptive data mining models. What are the different tasks of each model? Discuss one task from each model.

(c) Give an overview of development of data mining. What are the important implementation issues associated with data mining? Discuss any three of these issues.

2. **(a)** What do you mean by Decision Tree? Write down the algorithm to generate the Decision Tree.

(b) What do you mean by Fuzzy set, Fuzzy logic and Membership function? Discuss different types of membership function with example. What are the different types of operations on fuzzy sets? Explain with example.

(c) Given the following set of values {2, 5, 7, 9, 17, 20, 27, 28, 30 and 31}, determine the jackknife estimate for both the mean and standard deviation of the mean.

3. **(a)** How does the data warehouse differ from a database? How are they similar? Explain.

(b) The Scores of six students on two courses (X_1, X_2) are available, as shown in the following table:

Student ID	A	B	C	D	E	F
X_1	3	4	2	5	1	4
X_2	2	1	5	2	6	2

(i) Plot the data in a scatter diagram. How many groups would you say there are? And what are their members?

(ii) Apply the furthest neighbor method and the squared Euclidean distance as a measure of dissimilarity. Use a dendrogram to arrive at the number of groups and their membership.

(iii) Apply the K-means method, assuming that the data belong to two groups and that one of these groups consists of A and E. Compare the result with above results (i) and (ii).

(c) Why is naïve Bayesian classification called 'naïve'? Briefly outline the major ideas of naïve Bayesian classification.

4. **(a)** What do you mean by artificial neural networks and biological neural networks? What is an activation function used in neural network? Illustrate with an example.

(b) Discuss the following activation function used in neural network:
(i) Threshold,　(ii) Sigmoid, and　(iii) Gaussian.

(c) Discuss the different steps in developing an artificial neural network.

5. **(a)** Explain with example classification, confusion matrix, ROC? Describe the basic methods used to solve the classification problem. What are the different classification techniques?

(b) What do you mean by attribute selection measures in classification? Explain the following terms: Information gain, Gain, Gain ratio, and Gini index.

(c) The following table consists of training data from a database. Let *Y* be the class label attribute. Obtain Gini Index – the attribute selection measure in classification.

X1	*X2*	*X3*	*X4*	*Y*
N	F	Y	M	N
Y	F	Y	L	Y
N	E	M	M	Y
N	F	S	M	Y
Y	F	S	L	Y
Y	E	Y	M	Y
N	F	Y	H	N
N	E	Y	H	N
N	F	M	H	Y
N	E	S	M	N
Y	E	S	L	N
Y	E	M	L	Y
Y	F	M	H	Y
Y	F	S	M	Y

- Best of Luck -

Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Hons.) Examination 2017
Course Title: Actuarial Statistics
Course No. Stat 406

Time: 2.5 Hours Full Marks: 35

[Answer any THREE questions. All questions carry equal marks.]

1. a) Define actuarial science along with its relationship with life insurance. Briefly explain the principal and accumulated value.

b) Define simple and compound interest with suitable example. Which one do you think better for developing countries? Briefly explain the discount.

c) The Kelly family buys a new house for 93500 on May 1, 1986. How much was this house worth in May1, 1982 if the real estate prices have risen at a compound interest rate of 8% per year during that period?

2. a) A trust fund is to be built by means of deposits of amount 5000 at the end of each year, with a terminal deposit, as small as possible, at the end of the final year. The purpose of this fund is to establish monthly payments of amount 300 into perpetuity, the first payment coming one month after the final deposit. If the rate of interest is 12% per year convertible quarterly, find the number of deposits required and the size of the final deposit.

b) A loan of 25000 is to be repaid by annual payments at the end of each year for the next 20 years. During the first five years the payments are k per year; during the second five years the payments are $2k$ per year; during the third five years the payments are $3k$ per year and during the fourth five years the payments are $4k$ years. If $i = 0.12$, find k.

c) Deposits of 1000 are placed into a fund at the end of each year for the next 25 years. Five years after the last deposit, annual payments commence and continue forever. If $i = 0.09$, find the amount of each payment.

3. a) Explain bond and book value with examples and formulate the general equation for bond and book value.

b) Define amortization and sinking fund with suitable examples. Why do you use such techniques?

c) A bond of 500, redeemable at par after 5 years, pays interest at 13% per year convertible semiannually. Find the price to yield an investor

i) 8% effective per half-year;

ii) 16% effective per year.

d) A corporation decides to issue 15-year bonds with face amount of 1000 each. If interest payments are to be made at the rate of 10% convertible semiannually and if George is happy with a yield of 8% convertible semiannually, what should he pay for one of these bonds?

4. a) Define life table. Briefly explain the components of life tables.

b) Explain, both mathematically and verbally, why the following are true.

i) ${}_{m+n}p_x = ({}_np_x)({}_mp_{x+n})$

ii) $q_x + (p_x)(q_{x+1}) + ({}_2p_x)(q_{x+2}) + \ldots = 1$

c) Define along with the general formula:

i) analytical function;

ii) terminal age of population and

iii) force of mortality.

d) A scientist studies the mortality patterns of Golden-Winged Warblers. She establishes the following probabilities of deaths: $q_0 = 0.4$, $q_1 = 0.2$, $q_2 = 0.3$, $q_3 = 0.7$ and $q_4 = 1$. Starting with $l_0 = 1000$, construct a mortality table.

6. (a) What is Bongaarts aggregate model for target setting in fertility?

(b) Derive the expression to determine the prevalence of contraception rate in a target year.

(c) What do you mean by proximate determinant of fertility? What are the different proximate determinants of fertility? Describe the four indices of fertility.

7. (a) Why life table is necessary? Discuss different types of life table used in demography. Which one is the most suitable for Bangladesh and why?

(b) Distinguish between fecundity and fecundability. Obtain the mean and variance o fecundability.

(c) What is census coverage? Derive the expression for estimating the census coverag error.

8. (a) What is family planning? Write down the necessity of family planning for polic implementation.

(b) What is clinical and non-clinical method of contraception? Explain the abov methods of contraception.

(c) Explain the meaning of demographic window and demographic dividend.

- Best of Luck -

Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Hons.) Examination 2017
Course Title: ***Research Methodology***
Course No. Stat-408

Time: **2.5** Hours Full Marks: 35

[Answer any THREE questions. All questions carry equal marks.]

1. **(a)** What is research method? How does it differ from research methodology? What is research design? What are the essences of a research design?

   (b) What is proposition? Compare it with hypothesis. What are the functions of hypothesis? What are the steps of determination of sample size? Why is it important?

   (c) What is SWOT analysis? What are the purposes of this analysis? In which research field this analysis is more useful. What is Focus Group Discussion (FGD) technique? Is FGD an exploratory research? Justify your answer.

2. **(a)** Define reliability and distinguish it from validity. Does reliability ensure validity?

   (b) What are the various techniques of measuring reliability? Distinguish between test retest method and parallel forms method of measuring reliability.

   (c) Discuss Kuder-Richardson formula 20 and 21 for estimating reliability of a test. In what respects formula 20 differs from formula 21. The mean score of a 50-item test is 40 with a variance of 75. Estimate the reliability of this test.

3. **(a)** What do you mean by data collection? What are the pre-requisites for data collection?

   (b) Make a comparative study between qualitative and quantitative data collection techniques.

   (c) What do you mean by data preparation process? What are the different steps of data preparation process? A questionnaire returned from the field may be unacceptable for several reasons. What are those reasons?

4. **(a)** Describe the difference between absolute precision and relative precision when estimating a population mean. How do the degree of confidence and the degree of precision differ?

   (b) Describe the procedure for determining the sample size necessary to estimate a population mean, given the degree of precision and confidence and a known population variance. After the sample is selected, how is the confidence interval generated?

   (c) The management of a local restaurant wants to determine the average monthly amount spent by households in restaurants. Some households in the target market do not spend anything at all, whereas other households spend as much as $300 per month. Management wants to be 95 percent confidence of the findings and does not want the error to exceed plus or minus $5. What sample size should be used to determine the average monthly household expenditure?

5. **(a)** Describe a commonly used format for writing a scientific research report. Describe the following parts of report: Title page, table of contents, executive summary, problem definition, research design, data analysis, conclusions, and recommendations.

   (b) Why is the 'limitations of caveats' section included in the report?

   (c) Discuss the importance of objectivity in writing a research report.

- Best of Luck -

Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Hons.) Examination 2017
Course Title: ***Research Methodology***
Course No. Stat-408

Time: **2.5** Hours Full Marks: 35

[Answer any THREE questions. All questions carry equal marks.]

1. **(a)** What is research method? How does it differ from research methodology? What is research design? What are the essences of a research design?

   (b) What is proposition? Compare it with hypothesis. What are the functions of hypothesis? What are the steps of determination of sample size? Why is it important?

   (c) What is SWOT analysis? What are the purposes of this analysis? In which research field this analysis is more useful. What is Focus Group Discussion (FGD) technique? Is FGD an exploratory research? Justify your answer.

2. **(a)** Define reliability and distinguish it from validity. Does reliability ensure validity?

   (b) What are the various techniques of measuring reliability? Distinguish between test retest method and parallel forms method of measuring reliability.

   (c) Discuss Kuder-Richardson formula 20 and 21 for estimating reliability of a test. In what respects formula 20 differs from formula 21. The mean score of a 50-item test is 40 with a variance of 75. Estimate the reliability of this test.

3. **(a)** What do you mean by data collection? What are the pre-requisites for data collection?

   (b) Make a comparative study between qualitative and quantitative data collection techniques.

   (c) What do you mean by data preparation process? What are the different steps of data preparation process? A questionnaire returned from the field may be unacceptable for several reasons. What are those reasons?

4. **(a)** Describe the difference between absolute precision and relative precision when estimating a population mean. How do the degree of confidence and the degree of precision differ?

   (b) Describe the procedure for determining the sample size necessary to estimate a population mean, given the degree of precision and confidence and a known population variance. After the sample is selected, how is the confidence interval generated?

   (c) The management of a local restaurant wants to determine the average monthly amount spent by households in restaurants. Some households in the target market do not spend anything at all, whereas other households spend as much as $300 per month. Management wants to be 95 percent confidence of the findings and does not want the error to exceed plus or minus $5. What sample size should be used to determine the average monthly household expenditure?

5. **(a)** Describe a commonly used format for writing a scientific research report. Describe the following parts of report: Title page, table of contents, executive summary, problem definition, research design, data analysis, conclusions, and recommendations.

   (b) Why is the 'limitations of caveats' section included in the report?

   (c) Discuss the importance of objectivity in writing a research report.

- Best of Luck -

Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Hons.) Examination 2017
Course Title: *Environmental Statistics*
Course No. Stat 409

Time: **2.5** Hours

Full Marks: 35

[Answer any THREE questions. All questions carry equal marks.]

1. a) What is an environment? What are the different components of environment? Discuss each of them.

b) What is environmental pollution? What are the different types of environmental pollution? What are the different sources of environmental pollution? How do they affect on humans?

2. a) What do you mean by dilution of pollutants? How can you apply the theory of successive random dilution (SRD) to common environmental phenomenon? Discuss in case of air and water quality.

b) Discuss the continuous mass balance model to measure the time varying concentration of pollutants.

3. What is diffusion and dispersion of pollutants? Discuss the Wedge Machine method to find the distribution of pollutants release from a source with respect to space and time.

4. What is composite sampling? Why composite sampling is important for environmental problems? What are the different types of composite sampling? Discuss each of them.

5. a) Discuss the statistical theory of rollback.

b) How can you predict concentrations of pollutants after a particular source or group of sources of pollutants are controlled?

c) If the random variable representing the source is correlated with the random variable representing dilution-diffusion phenomena, then show that the assumed correlation between the source and the dilution-diffusion phenomena in the past control state will be the same as in the pre-control state.

Best of Luck

Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Hons.) Examination 2017
Course Title: *Statistical Data Analysis VIII*
Course No. Stat Lab 410
Group C: Experimental Design

Time: 3.0 Hours

Full Marks: 08

[Answer the following questions.]

1. The following table gives the layout and the results of a 2^3 factorial experiment laid out in 4 blocks. The purpose of the experiment was to determine the effect of different kinds of fertilizers Nitrogen (N), Potassium (K) and Phosphate (P) on the yields of potato crop.

2^3 factorial experiment in 4 blocks

Block-1	Block-2	Block-3	Block-4
$nk = 391$	$kp = 507$	$p = 423$	$np = 461$
$kp = 492$	$p = 424$	$(1) = 187$	$nk = 372$
$p = 412$	$k = 372$	$np = 424$	$n = 203$
$np = 473$	$nk = 406$	$kp = 523$	$p = 424$
$(1) = 201$	$n = 189$	$nk = 434$	$k = 402$
$k = 365$	$nkp = 549$	$k = 379$	$(1) = 231$
$n = 206$	$np = 438$	$n = 228$	$nkp = 537$
$nkp = 550$	$(1) = 206$	$nkp = 571$	$kp = 535$

a) Analyze the data and set up the ANOVA table.

b) Test the hypothesis of main effects and interaction effects of different fertilizers at 5% level of significance and comment.

2. The following table gives the results of an experiment for comparing 7 treatments in 7 blocks of 3 units each, there thus being 3 replications of each treatment. Analyze the data and test the significance of treatment effects at 5% level of significance. Which treatment has the highest effect?

Treatment	Block						
	1	2	3	4	5	6	7
1	50	42	91	-	-	-	-
2	-	-	118	94	94	-	-
3	76	-	-	64	-	80	-
4	-	-	72	-	-	53	31
5	44	-	-	-	65	-	54
6	-	102	-	-	119	92	-
7	-	38	-	38	-	-	37

Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Hons.) Practical Examination 2017
Course Title: ***Statistical Data Analysis IX (Group-B: Data Mining)***
Course No. STAT-LAB-411

Time: 1.5 Hours **Full Marks: 8**

1. Given a data set named "Q1.csv", which contains data on the passengers on a Ship and the record of their survival when the ship wrecked. The attributes in the dataset are:
    Class: "1st" "2nd" "3rd" "Crew"
    Sex: "Male" "Female"
    Age: "Child" "Adult"
    Survived: "No" "Yes"

Table : Q1.csv

Class	Sex	Age	Survived
3rd	Female	Adult	Yes
1st	Male	Child	Yes
Crew	Female	Child	Yes
⋮	⋮	⋮	⋮
2nd	Male	Child	No
3rd	Male	Adult	Yes

We are interested in the association between the attributes. Find all association rules with support 0.2 and confidence 0.5. Hence Find
    i) the summary of the quality measure- support, confidence, and Lift.
    ii) the redundant association rules.
    iii) the association rule where right hand side contains "Survived".
    iv) the association rule where left hand side contains two attributes right hand side contains "Survived".
    v) Draw the scatter plots for the rules with the help of confidence and support, and lift.

2. Fit an ANN model to predict Rainfall (RAN) based on Temperature (TEM), Dew Point Temperature (DPT), Wind Speed (WIS), Humidity (HUM), and Sea Level Pressure (SLP). The following table named "Q2.csv" list monthly data on the Rainfall (RAN), Temperature (TEM), Dew Point Temperature (DPT), Wind Speed (WIS), Humidity (HUM), and Sea level Pressure (SLP) for the period January 1964 to December 2015 of Sylhet.

Table: Q2.csv

TEM	DPT	RAN	WIS	HUM	SLP
16.7	10.8	25	4.6	72.19	1014.7
20	11.2	25	4.6	65.86	1012.7
⋮	⋮	⋮	⋮	⋮	⋮
23.5	17.5	22	2.1	73.77	1013.4
19.1	14.6	5	2.4	76.77	1015.6

Estimate the parameters with model summary and draw the network architecture. Also estimate and show the different accuracy measures for the model.

3. The Euclidean distances between twelve the objects presented in the following matrix. Apply the single linkage, complete linkage, average linkage, and ward linkage to cluster the objects. Compare your results and comment on results.

$$D_{ij}=\begin{array}{c|ccccccccccc}
 & V1 & V2 & V3 & V4 & V5 & V6 & V7 & V8 & V9 & V10 & V \\
\hline
V1 & 0 & & & & & & & & & & \\
V2 & 13.1 & 0 & & & & & & & & & \\
V3 & 28.6 & 18.4 & 0 & & & & & & & & \\
V4 & 45.4 & 35.3 & 18.9 & 0 & & & & & & & \\
V5 & 66.8 & 55.6 & 39.2 & 26.7 & 0 & & & & & & \\
V6 & 87.3 & 76 & 59.9 & 46.1 & 21.8 & 0 & & & & & \\
V7 & 86.2 & 75.3 & 59.6 & 47.5 & 26.8 & 20.9 & 0 & & & & \\
V8 & 81.8 & 70.6 & 54.2 & 40.5 & 16.9 & 10.7 & 20.9 & 0 & & & \\
V9 & 60.3 & 49.7 & 34.8 & 27.1 & 20.8 & 36 & 38.6 & 31.6 & 0 & & \\
V10 & 35.8 & 26.8 & 17.5 & 24.7 & 39.4 & 58.9 & 59.1 & 53.8 & 27.2 & 0 & \\
V11 & 19.6 & 17.2 & 23.6 & 38.6 & 57.8 & 77.6 & 77.2. & 72.6 & 52.5 & 30.5 & 0 \\
V12 & 19.2 & 24.8 & 36.9 & 50.6 & 73.8 & 93.8 & 92.5 & 88.4 & 67.5 & 43.9 & 29
\end{array}$$

4. Randomly construct a data set of 100 observations according to the regression $Y_i = 5 + 2x_i + \varepsilon_i$, where $x_i = 1,2,\ldots,100$, and the errors are distributed with $\varepsilon_i \sim N$ Bootstrap the least-squares regression of Y on x using random resampling (c $r = 1{,}000$ bootstrap samples). In this case, plot the bootstrap distribution of th coefficient, and calculate the bootstrap estimate for this coefficient.

- Good Luck -

Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Hons.) Practical Tutorial Examination 2017
Course Title: *Statistical Data Analysis VIII (Group-A)*
Course No. STAT-LAB-410

Time: 1.5 Hours **Full Marks: 20**

1. Table 1 is from a survey conducted by the Wright State University School of Medicine and the United Health Services in Dayton, Ohio. The survey asked students in their final year of a high school near Dayton, Ohio whether they had ever used alcohol, cigarettes, or marijuana. Denote the variables in this 2 × 2 × 2 table by A for alcohol use, C for cigarette use, and M for marijuana use.

Table 1: Alcohol (A), Cigarette (C), and Marijuana (M). Use for High School Seniors

Alcohol Use	Cigarette Use	Marijuana Use	
		Yes	No
Yes	Yes	911	530
	No	44	451
No	Yes	8	43
	No	10	279

(a) Write the R-codes to construct the partial table of M, and the A-C and A-M marginal tables.

(b) Fit loglinear models (A, C, M), (A, AC), (A, CM), (AC, CM), (MA, CM), and (AC, AM, CM). Report Likelihood Ratio (LR) statistic, degrees of freedom (df) and p-values, and comment on the quality of fit.

(c) Calculate the fitted values of the best model. Interpret the results.

(d) Calculate the Marginal and Partial odds ratios for A-C and A-M.

2. At the start of a study to determine whether exercise or dietary supplements would slow bone loss in older women, an investigator measured the mineral content of bones by photon absorptiometry. Measurements were recorded for three bones on the dominant and nondominant sides and are shown in Table 2. This data is available in the file **F:/ multivar5th/T1-8.dat**.

Table 2: Mineral Content in Bones

Subject number	Dominant radius (x_1)	Radius (x_2)	Dominant humerus (x_3)	Humerus (x_4)	Dominant ulna (x_5)	Ulna (x_6)
1	1.103	1.052	2.139	2.238	0.873	0.872
2	0.842	0.859	1.873	1.741	0.590	0.744
⋮	⋮	⋮	⋮	⋮	⋮	⋮
25	0.915	0.936	1.971	1.869	0.869	0.868

(a) Examine the multivariate normality of the observations on six different variables of the mineral content of three bones on the dominant and nondominant sides of older women.

(b) Evaluate T^2 of the six variables $(x_1, x_2, \ldots, x_6)$ for testing $H_0 : \boldsymbol{\mu}' = [0.80 \quad 0.80 \quad 1.70 \quad 1.70 \quad 0.70 \quad 0.70]$. Hence, find out the sampling distribution of T^2.

(c) Construct the sample covariance matrix S for the above data matrix. Hence, determine the sample principal components and their variances for the covariance matrix S.

(d) Compute the proportion of total variance explained by the first two principal components obtained in Part (c). Interpret your result.

(e) Calculate the Euclidean distances between six different variables of the miner content of three bones on the dominant and nondominant sides of older wome Cluster the six variables using the single linage and complete linkage hierarchic methods. Draw the dendrograms and compare the results.

3. A random sample of $n = 70$ families will be surveyed to determine the association bet certain 'demographic' variables and certain 'consumption' variables. Let *criterio* contains $X_1^{(1)}$ = annual frequency of dining a restaurant, and $X_2^{(1)}$ = annual frequen attending movies; and *predictor set* contains $X_1^{(2)}$ = age of head household, $X_2^{(2)}$ = a family income, and $X_3^{(2)}$ = educational level of head of household.

Suppose 70 observations on the preceding variables give the sample correlation matrix

$$\mathbf{R} = \left[\begin{array}{c:c} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \hdashline \mathbf{R}_{21} & \mathbf{R}_{22} \end{array}\right] = \left[\begin{array}{cc:ccc} 1.0 & & & & \\ .80 & 1.0 & & & \\ \hdashline .26 & .33 & 1.0 & & \\ .67 & .59 & .37 & 1.0 & \\ .34 & .34 & .21 & .35 & 1.0 \end{array}\right]$$

(a) Find all the sample canonical correlations and the sample canonical variates.

(b) Stating any assumptions you make, test the hypotheses $H_0 : \sum_{12} = \mathbf{\rho}_{12} = \mathbf{0}$ at the level of significance. If H_0 is rejected, test for the significance (α=.05) of th canonical correlation.

(c) Do the demographic variables have something to say about the consumption vari Do the consumption variables provide much information about the demog variables?

- **Good Luck** -

Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Hons.) Practical Examination 2017
Course Title: ***Statistical Data Analysis IX (Group-B: Data Mining)***
Course No. STAT-LAB-411

Time: 1.5 Hours **Full Marks: 8**

1. Given a data set named "Q1.csv", which contains data on the passengers on a Ship and the record of their survival when the ship wrecked. The attributes in the dataset are:
   Class: "1st" "2nd" "3rd" "Crew"
   Sex: "Male" "Female"
   Age: "Child" "Adult"
   Survived: "No" "Yes"

Table : Q1.csv

Class	Sex	Age	Survived
3rd	Female	Adult	Yes
1st	Male	Child	Yes
Crew	Female	Child	Yes
⋮	⋮	⋮	⋮
2nd	Male	Child	No
3rd	Male	Adult	Yes

We are interested in the association between the attributes. Find all association rules with support 0.2 and confidence 0.5. Hence Find

i) the summary of the quality measure- support, confidence, and Lift.
ii) the redundant association rules.
iii) the association rule where right hand side contains "Survived".
iv) the association rule where left hand side contains two attributes right hand side contains "Survived".
v) Draw the scatter plots for the rules with the help of confidence and support, and lift.

2. Fit an ANN model to predict Rainfall (RAN) based on Temperature (TEM), Dew Point Temperature (DPT), Wind Speed (WIS), Humidity (HUM), and Sea Level Pressure (SLP). The following table named "Q2.csv" list monthly data on the Rainfall (RAN), Temperature (TEM), Dew Point Temperature (DPT), Wind Speed (WIS), Humidity (HUM), and Sea level Pressure (SLP) for the period January 1964 to December 2015 of Sylhet.

Table: Q2.csv

TEM	DPT	RAN	WIS	HUM	SLP
16.7	10.8	25	4.6	72.19	1014.7
20	11.2	25	4.6	65.86	1012.7
⋮	⋮	⋮	⋮	⋮	⋮
23.5	17.5	22	2.1	73.77	1013.4
19.1	14.6	5	2.4	76.77	1015.6

Estimate the parameters with model summary and draw the network architecture. Also estimate and show the different accuracy measures for the model.

3. The Euclidean distances between twelve the objects presented in the following matrix. Apply the single linkage, complete linkage, average linkage, and ward linkage to cluster the objects. Compare your results and comment on results.

$$D_{ij} = \begin{array}{c|ccccccccccc}
 & V1 & V2 & V3 & V4 & V5 & V6 & V7 & V8 & V9 & V10 & V \\
\hline
V1 & 0 & & & & & & & & & & \\
V2 & 13.1 & 0 & & & & & & & & & \\
V3 & 28.6 & 18.4 & 0 & & & & & & & & \\
V4 & 45.4 & 35.3 & 18.9 & 0 & & & & & & & \\
V5 & 66.8 & 55.6 & 39.2 & 26.7 & 0 & & & & & & \\
V6 & 87.3 & 76 & 59.9 & 46.1 & 21.8 & 0 & & & & & \\
V7 & 86.2 & 75.3 & 59.6 & 47.5 & 26.8 & 20.9 & 0 & & & & \\
V8 & 81.8 & 70.6 & 54.2 & 40.5 & 16.9 & 10.7 & 20.9 & 0 & & & \\
V9 & 60.3 & 49.7 & 34.8 & 27.1 & 20.8 & 36 & 38.6 & 31.6 & 0 & & \\
V10 & 35.8 & 26.8 & 17.5 & 24.7 & 39.4 & 58.9 & 59.1 & 53.8 & 27.2 & 0 & \\
V11 & 19.6 & 17.2 & 23.6 & 38.6 & 57.8 & 77.6 & 77.2. & 72.6 & 52.5 & 30.5 & 0 \\
V12 & 19.2 & 24.8 & 36.9 & 50.6 & 73.8 & 93.8 & 92.5 & 88.4 & 67.5 & 43.9 & 29.0
\end{array}$$

4. Randomly construct a data set of 100 observations according to the regression $Y_i = 5 + 2x_i + \varepsilon_i$, where $x_i = 1,2,\ldots,100$, and the errors are distributed with $\varepsilon_i \sim N(0,$ Bootstrap the least-squares regression of Y on x using random resampling (dra $r = 1{,}000$ bootstrap samples). In this case, plot the bootstrap distribution of the coefficient, and calculate the bootstrap estimate for this coefficient.

- **Good Luck** -

Department of Statistics
Jahangirnagar University
Savar, Dhaka

Part IV B.Sc. (Hons.) Practical Examination, 2017

Course No.: STAT-LAB 411

Course Title: Statistical Data Analysis IX (Group C: Environmental Statistics)

Time: 30 minutes **Marks: 8**

Answer the following question.

Suppose that the statistical part of the ozone air quality standard is modified so that the expected number of exceedances per year is 0.5 or less. If M denotes the number of exceedances in any 3-year period,

i. What is the that "more than one exceedance" will occur in a given year?

ii. What is the probability distribution of M when the standard is just attained and E[M] =0.5?

iii. compute the probabilities associated with M = 0, 1, 2, or 3.

iv. What is the probability that 3 or fewer exceddedances will occur in any 3-year period.

Department of Statistics
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2018
Course No. : Stat – 401
Course Name: Statistical Inference II

Time: 4 Hours **Mark: 70**

Answer any Five from the following questions. Each question carries equal Marks.

Q1. **(a)** Define with suitable example modal unbiased estimator and Median unbiased estimator.

(b) What is Jackknife method and what is bootstrap method? If X ~ Binomial (n,θ) find an unbiased estimator of θ^2.

Q2. **(a)** Define efficiency of on estimator. Show with an example that in terms of MSE an MLE may be worthless.

(b) State and prove Chapman, Robbins, and Kiefer Inequality.

(c) What is Pitman estimator for the scale parameter? If $x_1, x_2, \ldots, x_n$ be a random sample from the density $f(x;\mu)=\frac{1}{\mu}\exp\left(-\frac{x}{\mu}\right)I_{(0,\infty)}^{(x)}$. Then find the Pitman estimator for the scale parameter μ.

Q3. **(a)** What is Ellipsoid of concentration and Wilk's generalized variance?

(b) What is location invariant estimator and scale invariant estimator? Show that sample variance is not a location invariant but is a scale invariant estimator.

(c) Discuss about equivariance estimator. Show that the bias, risk, and variance of an equivariant estimator may be independent of the population parameter.

Q4. **(a)** What is data reduction technique? What are the techniques available for data reduction? Why sufficient statistic is called a data reduction technique?

(b) What do you mean by complete statistics of a family of density function? What is the difference between complete sufficient statistic and sufficient statistic? Show that a complete sufficient statistic is minimal sufficient statistic.

(c) Let $x_1, x_2, \cdots, x_n$ be independent random variable each with $P(\lambda);\ \lambda > 0$. Find a minimal sufficient statistic for λ and check if it is complete or not.

Q5. **(a)** Define randomize test and non-randomized test, power function and critical function, loss function and risk function, and most powerful test and Bayes test. Simple likelihood ratio test and generalized likelihood-ratio test.

(b) Discuss Bayesian testing procedure, Wald-type test and score test.

Q6. **(a)** What is SPRT? Write down the different steps of a SPRT.

(b) State and prove Wald's equation. Determine the expected sample size (i) if the null hypothesis is true (ii) if alternative hypothesis is true. Give an example.

(c) Determine the O.C. function of a SPRT.

Q7. (a) Define prior and posterior pdf, Bayes risk and loss function.

(b) What is Conjugate prior and Jeffery's non-informative prior? How can you const conjugate prior?

(c) Let, $X_1,\ldots,X_n$ be a random sample drawn from a Poisson distribution with parame where, θ is unknown. Assuming $G(\beta,\alpha)$ is a prior distribution of θ. Find $100(1-$ Bayesian interval for θ.

Q8. (a) Describe the Bayesian approach of finding confidence interval. What is the difference between classical and Bayesian approach of finding confidence interval?

(b) What is fiducial interval? Why it is necessary? Suppose that $X_1, X_2, \cdots, X_n$ be a ra sample from $N(\theta,1)$. Find the fiducial probability of θ. Also obtain the $100(1-$ fiducial interval for θ.

Good Luck

Department of Statistics
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2018
Course No. : Stat – 402
Course Name: Multivariate Analysis and Categorical Data Analysis

Time: 4 Hours **Mark: 70**

Answer any Five from the following questions. Each question carries equal Marks.

Q1. (a) Define Multivariate normal distribution and Wishart distribution. Write down the properties of the Wishart Distribution. Also write down some application of multivariate normal distribution.

(b) Let $X = \begin{bmatrix} X_1 \\ \cdots \\ X_2 \end{bmatrix}$ be distributed as $N_p(\mu, \Sigma)$ with $\mu = \begin{bmatrix} \mu_1 \\ \cdots \\ \mu_2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \Sigma_{11} & \vdots & \Sigma_{12} \\ \cdots & \cdots & \cdots \\ \Sigma_{21} & \vdots & \Sigma_{22} \end{bmatrix}$, and $|\Sigma_{22}| > 0$. Then under usual notations show that the conditional distribution of X_1, given $X_2 = x_2$, is normal and has

$$\text{Mean} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \quad \text{and} \quad \text{Covariance} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

(c) Consider the annual rates of return (including dividends) on the Dow-Jones industrial average for the years 1996–2005. These data, multiplied by 100, are –

-0.6, 3.1, 25.3, -16.2, -7.1, -6.2, 16.1, 25.2, 22.6, 26.0.

Use these 10 observations to complete the following.

(i) Construct a Q–Q plot. Do the data seem to be normally distributed? Explain.

(ii) Carry out a test of normality based on the correlation coefficient. Let the significance level be 0.10. (Corresponding to *n=10* and significance level 0.10 the $r_Q = 0.9351$.)

Q2. (a) Explain the use of distances in multivariate analysis. Which distance is more preferable in statistics and why?

(b) Define Hotteling T^2 statistics with its applications. Let $X_1, X_2, \ldots, X_p$ be a random sample from a $N_p(\mu, \Sigma)$, then show that T^2 can be expressed as a function of Wilk's Lambda.

(c) What do you mean by the confidence region of μ, where $X \sim N_p(\mu, \Sigma)$. Construct the 95% confidence region of μ for $n = 40$ pairs of observations having

$$\bar{x} = \begin{bmatrix} .567 \\ .603 \end{bmatrix} \text{ and } S = \begin{bmatrix} .014 & .012 \\ .012 & .015 \end{bmatrix}.$$

Hence, check whether $\mu' = [.560 \quad .580]$ is in that confidence region.

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

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

(c) Let $n_1 = 28$, $n_2 = 28$, $\bar{x}_1 = [.15 \quad -.23 \quad -.32]$, $\bar{x}_2 = [.14 \quad .18 \quad .25]$ and

covariance matrix $S_p = \begin{bmatrix} 0.88 & 0.36 & 0.23 \\ 0.36 & 0.77 & 0.20 \\ 0.23 & 0.20 & 0.55 \end{bmatrix}$.

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

Q4. (a) Define the multivariate multiple linear regression model along with its assumptions. does it differ from the multiple linear regression model?

(b) Suppose the least square estimates $\hat{\boldsymbol{\beta}} = \left[\hat{\beta}_{(1)} \vdots \hat{\beta}_{(2)} \vdots \cdots \vdots \hat{\beta}_{(m)}\right]$ determined und multivariate multiple regression model with the error matrix $\boldsymbol{\varepsilon}$ and the design ma having full rank $(\mathbf{Z}) = r+1 < n$, then show that

(i) $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ (ii) $\text{Cov}\left(\hat{\beta}_{(i)}, \hat{\beta}_{(k)}\right) = \sigma_{ik}(\mathbf{Z}'\mathbf{Z})^{-1}$, $i, k = 1,$.

(iii) $E\left(\frac{1}{n-r-1}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}\right) = \boldsymbol{\Sigma}$ (iv) $\text{Cov}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\varepsilon}}\right) = 0$.

(c) Describe the likelihood ratio method to test the null hypothesis, $H_0 : \boldsymbol{\beta}_{(2)} = 0$, $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_{(1)} \\ \boldsymbol{\beta}_{(2)} \end{bmatrix}$, $\boldsymbol{\beta}_{(2)}$ is a $((q+1) \times m)$ matrix and $\boldsymbol{\beta}_{(2)}$ is a $((r-q) \times m)$ matrix.

Q5. (a) What is meant by Principal Component Analysis? What are the assumptions of prin component analysis? Write down some applications of principal component ana Describe the procedure to find the principal components for covariance matrices special structures.

(b) Let $X' = [X_1, X_2, \cdots, X_p]$ have covariance matrix Σ, with eigenvalue – eigenv pairs $(\lambda_1, e_1), (\lambda_2, e_2), \cdots, (\lambda_p, e_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p \geq 0$.

Let $Y_1 = e_1'X, Y_2 = e_2'X, \cdots, Y = e_p'X$ be the principal components. Then under notations show that $\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} \text{var}(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \text{var}(X_i$

(c) Describe the procedure to find the number of important principal components. Fin principal components and the proportion of total population variance explained by when the covariance matrix is –

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma^2\rho & 0 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ 0 & \sigma^2\rho & \sigma^2 \end{bmatrix}; \quad -\frac{1}{2} < \rho < \frac{1}{2}$$

Q6. (a) Define an orthogonal factor model with its components and assumptions. Is the relation between this model to regression model? Explain.

(b) State and prove the covariances structure for the orthogonal factor model. How cou check the adequacy of an orthogonal factor model? How could you calculate the proportion of variance of the i^{th} variable contributed by the m common factors?

(c) The following R output is obtained for conducting the factor analysis with 5 variables and $m = 2$ common factors

```
Call:
factanal(x = x, factors = 2, method = "mle", scale = T, center = T)
   Uniquenesses:
   V1    V2    V3    V4    V5
   0.497 0.252 0.474 0.610 0.176
   Loadings:                          Factor1 Factor2
   Factor1 Factor2                    SS loadings       1.671    1.321
   V1 0.601    0.378                  Proportion Var    0.334    0.264
   V2 0.849    0.165                  Cumulative Var    0.334    0.598
   V3 0.643    0.336
   V4 0.365    0.507
    V5 0.207    0.884

   Test of the hypothesis that 2 factors are sufficient.
   The chi square statistic is 0.58 on 1 degree of freedom.
   The p-value is 0.448
```

Find the followings:

(i) Find the matrix of specific variances. Hence, define the most significant variable which fit neatly into this factors model.

(ii) Find the estimated factor loadings and communalities. What proportion of the total population variance is explained by the first common factors? And by the 2nd common factor.

(iii) Check whether the 2 factors are adequate for this model?

Q7. (a) What is meant by cluster analysis? Define Different Dissimilarity Measures used in cluster analysis. Suppose we measure two variables X_1 and X_2 for three items A, B, and C and found:

Items	x_1	x_2
A	4	1
B	3	2
C	2	3

Use Ward's method to cluster the three items, and construct the dendrogram showing the values of error sum of squares at which the mergers take place.

(b) Consider the two data sets $x_1 = \begin{bmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{bmatrix}$ and $x_2 = \begin{bmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{bmatrix}$ For which $\bar{x}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$, $\bar{x}_2 = \begin{bmatrix} 5 \\ 8 \end{bmatrix}$ and

$S_{pooled} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$, Calculate linear discrimination function. Classify the observation

$x_0' = \begin{bmatrix} 2 & 7 \end{bmatrix}$ as population Π_1 or population Π_2 using any suitable rule.

(c) Discuss Fisher's linear discrimination function for two multivariate normal population. Also describe allocation rule based on this function.

Q8. **(a)** Define major types of categorical data with examples. What will be the approp measures of association for nominal-by-nominal, ordinal-by-ordinal and nominal-by-orc variables? Define them.

(b) Define relative risk, odds ratio, sensitivity, specificity of a test, Wald score, Likelihood test, and Wald confidence interval.

An important characteristic of glaucoma, an eye disease, is the presence of classical v field loss. Tonometry is a common form of glaucoma screening, whereas, for exampl eye is classified as positive if it has an intraocular pressure of 21 mmHg or higher single reading. Given the data shown in the following table

	Test Result		
Field Loss	Positive	Negative	Total
Yes	13	**2X**	
No	41	45	**86**
Total	54		

where *X* is the last two digits of your Class Roll No.

(i) Calculate the Predictive value positive, sensitivity and specificity of the test. He comments on each measure.

(ii) Find the relative risk (RR) and odds ratio (OR) of glaucoma patients.

(c) When does one should use Loglinear model instead of using Logit model? De Loglinear model of three-way table. Discuss the method of estimation of expecte frequency for a $(2 \times 2 \times 2)$ contingency table using a Loglinear model.

Good Luck

Department of Statistics
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2018
Course No. : Stat – 403
Course Name: Design of Experiments and Analysis of Variance

Time: 4 Hours **Mark: 70**

Answer any Five from the following questions. Each question carries equal Marks.

Q1. **(a)** Define design of experiment, Distinguish between absolute experiment and comparative experiment. How does design of experiment differ from sampling design?

(b) Describe the principles of replication, randomization and local control in experimental design and discuss how these principles are applied to give valid interpretation of data. Discuss the characteristics of a good experiment.

Q2. **(a)** In an agricultural experiment an experimenter wants to compare p varieties of a crop and he has p^2 experimental units to perform the experiment.

Suggest suitable designs to carry the experiment with justification in each of the following situations (i) p^2 – experimental plots are homogeneous. (ii) Fertility variation occurs along two perpendicular directions.

Set up the usual model with necessary assumptions and carry out the analysis of data briefly in each case.

(b) What is multiple comparison test? Describe Duncan's multiple range test for comparing treatment means.

Q3. **(a)** What do you mean by orthogonality of a design? Show that orthogonality of a RBD is lost if one or more observations are missing.

(b) In conducting a field experiment with 4 varieties of crops, an experimenter observed that the soil fertility varies. The varietes of crops are replaced 5 times. Suggest an appropriate design with the layout plan. Discuss the analysis procedure of the data obtain and set up the ANOVA table and test the hypothesis of variety effects of the crops.

Q4. **(a)** What is a split-plot design? How does it differ from randomized block design? Discuss he situations where split-plot design is useful.

(b) If p number of main-plot treatments and q number of sub-plot treatments is conducted in a split-plot design, give the layout of this design and set up linear model for this design with necessary assumptions. Is this design a non-orthogonal design? Justify your answer. Estimate the variance of two kinds of error of this design. Display the ANOVA table and estimate the efficiency of this design in comparison to RBD

Q5. **(a)** Discuss factorial experiment and single factor experiment with examples. Also compare them.

(b) Explain Yate's algorithm for calculating different component sum squares in a 2^3 factorial experiment in a RBD with 3 blocks. Also present the ANOVA table and comments.

Q6. **(a)** What do you mean by confounding? What are the necessities of confounding in fact experiment? Define total and partial confounding with example.

(b) Construct a 2^4 factorial experiment where ABCD and AC are partially confound Discuss the procedure of analyzing the data obtained form such design. Also set up ANOVA table.

Q7. **(a)** What is covariance analysis? Distinguish among analysis of variance, regression ana and covariance analysis. What is concomitant variable? What purpose is serve concomitant variances?

(b) Describe the analysis of covariance for the data of a completely randomized design two concomitant variables. Determine the estimates of the parameters of the m Discuss how the equality of adjusted treatment effects can be tested of this d analysis. Display the ANCOVA table.

Q8. **(a)** What is varietal trial? Discuss the utility of varietal trials in design of experiment. D incomplete block design. Explain when an incomplete block design is said to be bala incomplete block design (BIBD). For a BIB design with usual parameters show that b

(b) Describe the procedure of analysis of data obtained from a BIB design. Construct a design having parameters $v = 9, b = 12, r = 4, k = 3, \lambda = 1$ with the help of orthog latin square design.

Good Luck

Department of Statistics
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2018
Course No. : Stat – 404
Course Name: Sample Survey

Time: 4 Hours **Mark: 70**

Answer any Five from the following questions. Each question carries equal Marks.

Q1. **(a)** Explain the concept of inclusion probabilities. Define the first order and joint order inclusion probabilities. Discuss the importance of inclusion probabilities in sampling.

(b) Let S be a finite population containing 5 units. Calculate all first and second order inclusion probabilities under the following sampling design

$$P(s)=\begin{cases} 0.2 & for\ s=\{2,3,4\},\ s=\{2,5\} \\ 0.3 & for\ s=\{1,3,5\},\ \ s=\{1,4\} \\ 0, & otherwise \end{cases}$$

Also list all possible values of $n(s)$ under the above sampling design and verify the relation $E_p[n(s)]=\sum_{i=1}^{N}\pi_i$.

(c) Suppose there are six firms in a small town. These firms make up the population of this example. Also let us suppose that four of the six firms (A, B, C, F) are known to be located on the east side of the town and the remaining two (D and E) are located on the west side of the town. Calculate inclusion probabilities in case of stratified sampling if you draw a simple random sample without replacement of size 1 from each group.

Q2. **(a)** What is regression method of estimation? Under what circumstances would you recommend the method? State the properties of regression estimator. Compare and contrast regression method and ratio method of estimation.

(b) State the logic behind designing the linear regression method of estimation. What are the criteria to choose among regression, ratio and mean per unit estimates for estimating population total?

(c) Show that in simple random sampling in which b is a pre-assigned constant quantity, the linear regression estimator $\bar{y}_{LR}=\bar{y}+b\left(\bar{X}-\bar{x}\right)$ is unbiased. Obtain the minimum variance of $\bar{y}_{LR}$.

Q3. **(a)** Define the ratio estimator of population mean $\bar{Y}$. Find its approximate bias and approximate mean square error of estimator of $\bar{Y}$. Explain when ratio estimator is unbiased.

(b) Show that ratio estimator of population mean is more efficient than the usual SRS based respective estimators if $\rho>\frac{C_x}{2C_y}$ and the regression line passes through or nearby through the origin.

Q4. **(a)** Define two-stage cluster sampling and identify the situations where it is to be preferre over usual SRS and stratified sampling.

(b) Find the unbiased estimator of population mean for a two-stage sampling design when th total number of units in the population is not known. What will be its variance and ho will you estimate it?

Q5. **(a)** Describe the different assumptions for estimating population using capture-recaptu principle.

(b) Explain the concept of direct sampling and inverse sampling with special reference t wildlife population.

(c) Discuss the negative hypergeometric model and negative binomial model for estimatin population size of mobile population. Show that Bailey's estimator of population size unbiased. Find its variance and unbiased estimator of variance.

Q6. **(a)** Discuss the concept of sampling on two occasions for the estimation of current populati mean. What are the different types of estimators available in this case? Discuss briefly.

(b) Find the best linear estimator of the mean for the second occasion with its sampli variance. Also, find the minimum value of this variance.

Q7. **(a)** Define the following concepts with example: (i) unit non-response and item non-respo (ii) response bias and response variance.

(b) What are the sources of non-response errors and state how these can be controlled.

(c) Describe the Hansen-Hurwitz technique to reduce the non-response bias in mail surve Under this method, find the unbiased estimator of population mean, find its variance a estimator of its variance.

Q8. **(a)** How does multi-stage sampling differ from multi-phage sampling? Illustrate w examples the use of multi-stage sampling and multi-phage sampling. Discuss application of double sampling in regression and ratio estimation.

(b) Obtain an unbiased estimator of the population mean in case of double sampling. A find the sampling variance of this unbiased estimator.

Good Luck

Department of Statistics
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2018
Course No. : Stat – 405
Course Name: Data Mining

Time: 2.5 Hours **Mark: 35**

Answer any Three from the following questions. Each question carries equal Marks.

Q1. **(a)** What do you mean by data mining? How does data mining access of a database differs from its traditional counterpart?

(b) Given the following sets of values X={1, 3, 9, 15, 20, 7, 11, 17} and Y={4, 7, 11, 17, 19, 14, 19, 25}, determine the jackknife estimate for both the mean and standard deviation of the mean of each variable. Also find Jackknife estimate of correlation coefficients. Comment on your results.

(c) What is meant by the term machine learning? Also describe the different types of machine learning.

Q2. **(a)** What do you mean by attribute selection measures in classification? Explain the following terms Information gain, Gain, Gain ratio, Gini index.

(b) Why naïve Bayesian classification is called "naïve"? Briefly outline the major ideas and steps of naïve Bayesian classification?

(c) The following table consists of training data from a data base. Let Y be the class label attribute. Predicting the class label using naïve Bayesian classification for the following tuple – $X = (X1=Y, X2=F, X3=Y, X4=M)$. Comment on your results.

X1	*X2*	*X3*	*X4*	*Y*
N	F	Y	M	N
Y	F	Y	L	Y
N	E	M	M	Y
N	F	S	M	Y
Y	F	S	L	Y
Y	E	Y	M	Y
N	F	Y	H	N
N	E	Y	H	N
N	F	M	H	Y
N	E	S	M	N
Y	E	S	L	N
Y	E	M	L	Y
Y	F	M	H	Y
Y	F	S	M	Y

Q3. **(a)** What do you mean by clustering? What are the different types of clustering methods? Describe the different distance measures used in clustering. Also define the different similarity measures used in data mining.

(b) Consider the following distance matrix cluster the five object using Farthest- Neighbor and Minimal spanning tree method. Comment on your results.

$$D_{ij} = \{d_{ij}\} = \begin{array}{c|ccccc} & A & B & C & D & E \\ \hline A & 0 & & & & \\ B & 43 & 0 & & & \\ C & 52 & 17 & 0 & & \\ D & 177 & 136 & 127 & 0 & \\ E & 8 & 39 & 47 & 171 & 0 \end{array}$$

(c) Describe the K-means clustering algorithm for multidimensional datasets.

Q4. **(a)** What do you mean by association rules, potentially large item sets, candidates, candidates

item set? What are the different types of algorithm to finding frequent itemset? M
comparison any of the two algorithms.

(b) The *AllElectronics* transaction data base has nine transactions. Let $min\_sup, s = 2$, find all frequent itemsets using Apriori.

TID	*List of item_IDs*
T100	I1, I2, I3
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

(c) The following contingency table summarizes supermarket transaction data, where dogs refers to the transactions containing hot dogs, $\overline{hot\ dogs}$ refers to the transactions do not contain hot dogs, hamburgers refers to the transactions containing hamburgers, $\overline{hamburgers}$ refers to the transactions that do not contain hamburgers.

	hot dogs	$\overline{hot\ dogs}$	$\sum_{row}$
hamburgers	2000	500	2500
$\overline{hamburgers}$	1000	1500	2500
$\sum_{column}$	3000	2000	5000

(i) Suppose that the association rule "*hot dogs⇒hamburgers*" is mined. Give minimum support threshold of 25% and a minimum confidence threshold of 5 is this association rule strong?

(ii) Based on the given data, is the purchase of hot dogs independent of the purchas hamburgers? If not, what kind of correlation relationship exists between the two

(iii) Compare the use of the *all confidence, max confidence, Kulczynski*, and *co* measures with *lift* and *correlation* on the given data.

Q5. (a) Distinguish between (i) parametric and non-parametric models (ii) predictive descriptive data mining models. In what sense non-parametric techniques are m appropriate for data mining application. Give some example of non-parametric techniq used in data mining.

(b) Distinguish between fuzzy logic and boolean logic. What are the different types classical sets? What do you mean by fuzzy set? Give an example.

(c) The following data give the time (in seconds) that each of 32 students selected from university waited in line at their bookstore to pay for their textbooks in the beginning the Fall 2018 semester: 24, 27, 29, 29, 35, 8, 53, 1, 23, 45, 17, 29, 28, 25, 45, 24, 26, 40, 1, 37, 25, 28, 28, 3, 30, 25, 26, 4, 29, 53, and 60. Prepare a box-and-whisker p Are these data skewed in any direction? Find the mild and extreme outliers. Also find t portion of the outliers in each category from the data set. Comment on your results.

Good Luck

Department of Statistics

Jahangirnagar University

Part IV B. Sc (Honors) Examination – 2018

Course No. : Stat – 406

Course Name: Actuarial Statistics

Time: 2.5 Hours **Mark: 35**

Answer any Three from the following questions. Each question carries equal Marks.

Q1. **(a)** What nominal rate of interest compounded monthly is equivalent to a nominal rate of interest of 7.5% convertible semiannually?

(b) John deposits \$100 at the end of 20 years into a fund earning an annual effective interest rate of 7%

Mary makes 20 deposits into a fund at the end of each year for 20 years. The first 10 deposits are \$100 each, while the last 10 deposits are 100+X each. The fund earns an annual effective interest rate of 8% during first 10 years and 6% annual effective interest thereafter. At the end of 20 years, the amount of John's equals the amount in Mary's fund. Calculate X.

(c) Jim deposits X into a saving account at time 0, which pays interest at a nominal rate of i compounded semiannually. Mike deposits $2X$ into a different saving account at time 0, which pays simple interest at an annual rate of i. Jim an Mike earn the same amount of interest during last 6 months of the 8th year. Calculate i.

Q2. **(a)** Consider a 1000 par-value 10-year bond with semiannual 5% coupons. Assume this bond can be redeemed at par at any of the last 3 coupon dates. Find the price which will guarantee an investor a yield rate of 2% par half-year.

(b) Find the price of a 1000 par-value 10-year bond which has semiannual coupons of 5 the first half-year, 10 the second half-year, ..., 100 the last half-year, bought to yield 10% effective per year.

(c) Find the price of a 1000 par-value 10-year bond with coupons at 11% convertible semiannually, and for which the yield rate is 5% per half-year for the first 4 years and 6% half-year for the last 6 years.

(d) A 4% semiannual coupon \$100 bond maturity in 15 years is callable on any coupon date after the 11th. If called on the 11th through 20th coupon date, the redemption value would be \$110. If called on the 21st through 30th coupon date, the redemption value would be at par. Find the price that would ensure an investor a minimum yield of 2% per annum compounded semiannually.

Q3. **(a)** What do you mean by annuities? Discuss its usefulness in actuarial science. Also discuss in brief, the increasing annuity and decreasing annuity.

(b) Prove each of the following identities along with a verbal interpretation.

(i) $\ddot{a}_{\overline{n}|} = a_{\overline{n}|} + 1 - v^n$

(ii) $\ddot{s}_{\overline{n}|} = s_{\overline{n}|} - 1 + (1+i)^n$

(c) Elroy takes out a 5000 loan to buy a car. No payments are due for the first 8 months, b beginning with the end of the 9th month, he must make 60 equal monthly payments. $i=0.18$, find (i) the amount of each payment; (ii) the amount of each payment i9f there no payment free period,(i.e., if the first payment is due in one month and the remaining are made on a monthly basis, thereafter).

Q4. (a) Briefly explain amortization, bond and book value with example and formulate the gene equation for bond and book value.

(b) In addition to the notation already introduced in (a), let $k=\frac{P-C}{C}$ and $g=\frac{Fr}{c}$, De Makeham's Formula, which is $P=Cv^n+\frac{g}{i}(C-Cv^n)$.

(c) A 1000 loan is repaid by annual payments of 150, plus a smaller final payment. If $i=0$ and the first payment is made one year after the time of the loan, construct an amortiza schedule for the loan.

Q5. (a) A life insurance company determines that the probability of surviving for ten years is 0.9, 0.6 and 0.4 for the person aged 40, 50, 60 and 70 respectively. Determine each of following:

(i) The probability that a 40-year-old lives to age 80

(ii) The probability that a 50-year-old dies between age 70 and 80

(b) 20% of those who die between ages 25 and 75 die before age 50. The probability person aged 25 dying before age 50 is 0.20. Find the probability that a person age 50 die after age 75.

(c) Rose is 50 years old and purchases a 20-year term insurance with face value \$1,00 Determine the price of the insurance if $i=0.15$ and $p_x=0.94$ for all x.

(d) Given $l_x=\sqrt{100-x}$ Determine each of the following:

(i) The probability that a 36-year-old reaches age 64

(ii) Henry and Henrietta are both 19 years old. Find the probability that Henry l at least 17 years and Henrietta lives at most 44 years, and at least one of t survives for 32 years.

Good Luck

Department of Statistics
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2018
Course No. : Stat – 407
Course Name: Demography

Time: 4 Hours **Mark: 70**

Answer any Five from the following questions. Each question carries equal Marks.

Q1. **(a)** What do you mean by demographic transition? Explain the different parts of demographic transition theory.

(b) Which phase of demographic transition is fitted for Bangladesh? Explain the reasons elaborately.

(c) What is Bangladesh Demographic and Health Survey (BDHS)? Write down the necessity of demographic health survey. Compare the value of some demographic indicators of BDHS 2011 and 2014.

Q2. **(a)** What is population policy? What are the elements of population policy? What are the purposes of it?

(b) What do you mean by Health Policy? What are the objectives of it? Give some health infrastructure information.

(c) Discuss the health consequence of domestic violence and its implication on children's life.

Q3. **(a)** Discuss the meaning of population aging. Explain the indicators of aging. Define care index, economically active population and aging index.

(b) Explain the determinant of population aging. Describe the implications on health care of elderly.

(c) Describe the promising sector of elderly people in Bangladesh. Also explain the health situation of elderly in Bangladesh.

Q4. **(a)** Explain the meaning of concentration ratio. Describe the technique for estimating the completeness of death registration from intercensal cohort survivalship data.

(b) What is population projection? Write down the importance of population projection in policy implications.

(c) What are the different methods of population projection? Which method is appropriate for population projection and why? What are the important factors for the population projection? Explain the different way of tackling population momentum.

Q5. **(a)** Write down the importance of family planning? What do you mean by use effectiveness of the method? Write down the use effectiveness of different methods from BDHS 2014 data. Define natural fertility, potential fertility and observed fertility.

(b) What do you mean by fecundity and fecundability? Find the mean and variance fecundability.

(c) Derive the W. Brass technique of evaluating birth and death registration using age distribution and child survivorship data? Write down the assumptions and limitations the technique?

Q6. **(a)** What are the different measures of urbanization? Explain the positive consequences urbanization.

(b) Under usual notation show that

$$\text{(i) } TR_e = r_u - r_r \quad \text{and} \quad \text{(ii) } TR_a = \frac{r_{u_1} - r_{r_1}}{1+nr_{r_1}}$$

(c) What do you mean by advocacy? Elaborately discus about analysis,. strategy mobilization.

Q7. **(a)** Define over-enumeration and distinguish it from under-enumeration. Which error is r prevalent in census count?

(b) Define coverage error, content error and enumeration error. What are causes that attributable to these errors?

(c) What is re-enumeration technique of adjusting demographic data? Describe a metho adjusting an age distribution suffered from under enumeration.

Q8. **(a)** What do you mean by "age heaping"? Why does heaping occur in age data? Discu brief the principal causes of age misstatement in Bangladesh.

(b) Indentify the mean difference between Whipple's method and Myer's method of dete and quantifying age reporting error. Discuss these methods in brief.

(c) Indicate the limits of these methods to understand the magnitude of errors present i data.

Good Luck

Department of Statistics
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2018
Course No. : Stat – 408
Course Name: Research Methodology

Time: 2.5 Hours **Mark: 35**

Answer any Three from the following questions. Each question carries equal Marks.

Q1. **(a)** What is research? What are its significance? Enumerate the characteristics of research.

(b) What is qualitative research? How does it differ from quantitative research? Give some typical examples of quantitative research.

(c) What do you mean by non-experimental research? Briefly describe the exploratory research and conclusive research.

Q2. **(a)** What do you mean by research by research process? What is research problem? *"The task of defining the research problem often follows a sequential pattern"* Explain.

(b) What do you understand by the term 'measurement' in research? Define measurement error. Compare and contrast random error and systematic error with examples.

(c) What do you mean by action research? What are the purposes of this research? Distinguish between action research and pure research.

Q3. **(a)** What is research design? What are the broad classifications of a research design? State the main objectives of conceiving a research design.

(b) What are the steps of determination of sample size? Why the determination of sample size important? How do you determination the sample size for estimating proportion?

(c) If you want to be 99% confident of estimating the population proportion within an error of ± 0.25, what sample size is needed? State the assumption you make.

Q4. **(a)** Explain how sampling and statistical inference are useful for any research work. Why probability sampling is generally preferred in comparison to non-probability sampling?

(b) Make a comparative study between qualitative data collection techniques.

(c) What is data preparation? What are the importances of data editing? Make a comparison between editing and coding. A questionnaire returned from the field may be unacceptable for several reasons. What are those reasons?

Q5. **(a)** What do you mean by validity and reliability in research? Discuss how you improve the validity and reliability of research?

(b) Describe the steps involved in estimating reliability by using split-half method.

(c) What are the different statistics associated with cross-tabulation? Suppose, you are interested to measure the association between CGPA of a student and time spend on different social Medias. What measures you apply?

Good Luck

Department of Statistics
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2018
Course No. : Stat – 406
Course Name: Actuarial Statistics

Time: 2.5 Hours **Mark: 35**

Answer any Three from the following questions. Each question carries equal Marks.

Q1. **(a)** What nominal rate of interest compounded monthly is equivalent to a nominal rate of interest of 7.5% convertible semiannually?

(b) John deposits $100 at the end of 20 years into a fund earning an annual effective interest rate of 7%

Mary makes 20 deposits into a fund at the end of each year for 20 years. The first 10 deposits are $100 each, while the last 10 deposits are 100+X each. The fund earns an annual effective interest rate of 8% during first 10 years and 6% annual effective interest thereafter. At the end of 20 years, the amount of John's equals the amount in Mary's fund. Calculate X.

(c) Jim deposits X into a saving account at time 0, which pays interest at a nominal rate of i compounded semiannually. Mike deposits $2X$ into a different saving account at time 0, which pays simple interest at an annual rate of i. Jim an Mike earn the same amount of interest during last 6 months of the 8th year. Calculate i.

Q2. **(a)** Consider a 1000 par-value 10-year bond with semiannual 5% coupons. Assume this bond can be redeemed at par at any of the last 3 coupon dates. Find the price which will guarantee an investor a yield rate of 2% par half-year.

(b) Find the price of a 1000 par-value 10-year bond which has semiannual coupons of 5 the first half-year, 10 the second half-year, ..., 100 the last half-year, bought to yield 10% effective per year.

(c) Find the price of a 1000 par-value 10-year bond with coupons at 11% convertible semiannually, and for which the yield rate is 5% per half-year for the first 4 years and 6% half-year for the last 6 years.

(d) A 4% semiannual coupon $100 bond maturity in 15 years is callable on any coupon date after the 11th. If called on the 11th through 20th coupon date, the redemption value would be $110. If called on the 21st through 30th coupon date, the redemption value would be at par. Find the price that would ensure an investor a minimum yield of 2% per annum compounded semiannually.

Q3. **(a)** What do you mean by annuities? Discuss its usefulness in actuarial science. Also discuss in brief, the increasing annuity and decreasing annuity.

(b) Prove each of the following identities along with a verbal interpretation.

(i) $\ddot{a}_{\overline{n|}} = a_{\overline{n|}} + 1 - v^n$

(ii) $\ddot{s}_{\overline{n|}} = s_{\overline{n|}} - 1 + (1+i)^n$

(c) Elroy takes out a 5000 loan to buy a car. No payments are due for the first 8 months, bu beginning with the end of the 9th month, he must make 60 equal monthly payments. I *i=0.18*, find (i) the amount of each payment; (ii) the amount of each payment i9f there i no payment free period,(i.e., if the first payment is due in one month and the remaining 59 are made on a monthly basis, thereafter).

Q4. (a) Briefly explain amortization, bond and book value with example and formulate the genera equation for bond and book value.

(b) In addition to the notation already introduced in (a), let $k = \frac{P-C}{C}$ and $g = \frac{Fr}{c}$, Deriv Makeham's Formula, which is $P = Cv^n + \frac{g}{i}(C - Cv^n)$.

(c) A 1000 loan is repaid by annual payments of 150, plus a smaller final payment. If $i = 0.1$ and the first payment is made one year after the time of the loan, construct an amortizatio schedule for the loan.

Q5. (a) A life insurance company determines that the probability of surviving for ten years is 0 0.9, 0.6 and 0.4 for the person aged 40, 50, 60 and 70 respectively. Determine each of t following:

(i) The probability that a 40-year-old lives to age 80

(ii) The probability that a 50-year-old dies between age 70 and 80

(b) 20% of those who die between ages 25 and 75 die before age 50. The probability of person aged 25 dying before age 50 is 0.20. Find the probability that a person age 50 w die after age 75.

(c) Rose is 50 years old and purchases a 20-year term insurance with face value $1,00,0 Determine the price of the insurance if $i = 0.15$ and $p_x = 0.94$ for all x.

(d) Given $l_x = \sqrt{100 - x}$ Determine each of the following:

(i) The probability that a 36-year-old reaches age 64

(ii) Henry and Henrietta are both 19 years old. Find the probability that Henry li at least 17 years and Henrietta lives at most 44 years, and at least one of t survives for 32 years.

Good Luck

Department of Statistics
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2018
Course No. : Stat – 409
Course Name: Environmental Statistics

Time: 2.5 Hours **Mark: 35**

Answer any Three from the following questions. Each question carries equal Marks.

Q1. **(a)** What do you mean by environment? What are the different components of environment? Discuss each of them.

(b) What is environmental pollution? What are the different sources of environmental pollution? How do they affect on humans?

(c) What do you mean by environmental statistics? "Many areas of statistical methodology and modeling find application in environmental problems"- what are those areas of statistics? Discuss them in brief.

Q2. **(a)** What do you mean by diffusion and dispersion of pollutants? What are the different models to describe them?

(b) Consider a wedge machine consisting of an array of wedges in uniform rows. What will be the probability distribution of particle arrivals for all the channels comprising row 4? If the expected number of particles arriving in the first channel for row 4 is 25, then what will be the probability distribution of number of arrivals in this channel?

(c) What is dilution? Briefly discuss the formation of successive deterministic dilution using appropriate example.

Q3. **(a)** Why stochastic process is used to describe the environmental problems? What happened when the assumptions of Bernouli process have been violated in describing environmental pollution?

(b) Suppose that the statistical part of the ozone air quality standard is modified so that the expected number of exceedances per year is 0.5 or less. If M denotes the number of exceedances in any 3-year period,

(i) What is the probability distribution of M when the standard is just attained?

(ii) Compute the probabilities associated with M = 0, 1, 2, or 3.

(iii) What is the probability that 3 or fewer exceedances will occur in any 3-year period?

(c) What is normal process? What are the assumptions of normal process? Develop a normal process.

Q4. **(a)** What is the Statistical Theory of Rollback? How can you predict the concentration of pollutants after source of pollutants is controlled?

(b) If the random variable representing the source is correlated with the random variable representing dilution-diffusion phenomena, and if the source strength is multiplied by the rollback factor r, then

(i) What changes are observed in the expected value, variance, and coefficient o variation of the final concentration in the post-control state compared to th concentration of pre-control state?

(ii) What changes occur to the covariance and correlation coefficient between th source and dilution-diffusion variables for the controlled concentratio compared to uncontrolled concentration?

(c) Discuss the Environmental Transport Model in Air and Water.

Q5. (a) What is composite sampling? Why composite sampling is important for environment problems?

(b) What are the different types of composite sampling? Discuss each of them.

(c) Suppose we are interested in estimating the number N of fish in a pond by capturing trawl net a random sample of them, tagging them, releasing them and recapturing a seco random sample. The initial sample was of size 300. The second sample, of size 20 contained 50 previously tagged specimens. Find an estimate of total number of fish in t pond.

Good Luck

Department of Statistics
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2018
Course No. : Stat – 408
Course Name: Research Methodology

Time: 2.5 Hours **Mark: 35**

Answer any Three from the following questions. Each question carries equal Marks.

Q1. **(a)** What is research? What are its significance? Enumerate the characteristics of research.

(b) What is qualitative research? How does it differ from quantitative research? Give some typical examples of quantitative research.

(c) What do you mean by non-experimental research? Briefly describe the exploratory research and conclusive research.

Q2. **(a)** What do you mean by research by research process? What is research problem? *"The task of defining the research problem often follows a sequential pattern"* Explain.

(b) What do you understand by the term 'measurement' in research? Define measurement error. Compare and contrast random error and systematic error with examples.

(c) What do you mean by action research? What are the purposes of this research? Distinguish between action research and pure research.

Q3. **(a)** What is research design? What are the broad classifications of a research design? State the main objectives of conceiving a research design.

(b) What are the steps of determination of sample size? Why the determination of sample size important? How do you determination the sample size for estimating proportion?

(c) If you want to be 99% confident of estimating the population proportion within an error of ± 0.25, what sample size is needed? State the assumption you make.

Q4. **(a)** Explain how sampling and statistical inference are useful for any research work. Why probability sampling is generally preferred in comparison to non-probability sampling?

(b) Make a comparative study between among qualitative data collection techniques.

(c) What is data preparation? What are the importances of data editing? Make a comparison between editing and coding. A questionnaire returned from the field may be unacceptable for several reasons. What are those reasons?

Q5. **(a)** What do you mean by validity and reliability in research? Discuss how you improve the validity and reliability of research?

(b) Describe the steps involved in estimating reliability by using split-half method.

(c) What are the different statistics associated with cross-tabulation? Suppose, you are interested to measure the association between CGPA of a student and time spend on different social Medias. What measures you apply?

Good Luck

Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Hons.) Examination 2020
Course Title: Statistical Inference II
Course No. Stat 401

Time: 4 Hours | Full Marks: 70

[Answer any FIVE questions. All questions carry equal marks.]

1. Let X be a random variable with Poisson distribution with parameter $\lambda > 0$. Let $X_1, X_2, \ldots, X_n$ be a random sample from X. Consider as a prior density a Gamma density with parameters α and β.

   a) Find the Bayes estimator for λ and calculate its bias, variance and mean squared error (using the quadratic loss function).

   b) Derive the maximum likelihood estimator (MLE) for λ and calculate its bias, variance and mean squared error.

   c) Plot for sample sizes $n = 5$, 10 and 20, the mean squared errors, as a function of λ, for the MLE and for the Bayes estimator, when using a $\Gamma(10.0.1)$ as a prior density. Comment on this plot. Explain the behavior of the plotted mean squared errors.

2. Let $X_1, X_2, \ldots, X_n$ be a random sample from X having probability density function:

$$f_X(x;\theta) = \frac{\exp\left(\frac{x}{2} - \frac{e^{\frac{x}{2}}}{\theta}\right)}{2\theta} \qquad \textit{for } x \in R$$

   where, $\theta > 0$ is an unknown parameter.

   a) Find the maximum likelihood estimator of θ. Based on the maximum likelihood estimator, construct an (approximate) $100 \times (1-\alpha)\%$ confidence interval for θ. Avoid making unnecessary approximations.

   b) Provide an estimator for $\ln P(X \geq 4)$ and establish the asymptotic normality for the estimator.

   c) Construct the uniformly most powerful test of size α for the testing problem

$$H_0 : \theta \leq 1 \qquad \textit{versus} \qquad H_1 : \theta > 1$$

3. a) Define Jackknife method and Bootstrap method. Jackknife method is a special case of Bootstrap method, justify.

   b) What is model unbiased estimator? Show that a model unbiased estimate may not be unique.

   c) Let $X_1, X_2, \ldots, X_n \sim \textit{binomial}(n, p)$. Find an unbiased estimator of p^2.

4. Let $\Theta = \{0,1\}$ and let X be a discrete random variable with the following probability distribution:

x	1	2	3	4	5	6	7	8
$f(x;0)$	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.86
$f(x;1)$	0.14	0.12	0.1	0.08	0.06	0.04	0.02	0.44

a) Use the theorem of Neyman-Pearson to find the most power test of size $\alpha = 0.05$ for the hypothesis $H_0: \theta = 0$ versus $H_1: \theta = 1$.

b) Calculate the type II error probability of the test.

c) Let $\Theta = \{0,1\}$ and let X be a random variable with density function $f(x;0) = 1$ and $f(x;1) = 3x^2$ for $x \in]0,1[$. Find a most powerful test of size $\alpha = 0.2$ for the testing problem $H_0: \theta = 0$ versus $H_1: \theta = 1$.

5. a) What is Bayesian estimator? What is posterior distribution and posterior Bayes estimator?

b) What is noninformative prior and conjugate prior? How can you check the existence of conjugate prior?

c) Define Linex loss function. What would be the Bayes estimate under the Linex loss function?

d) If $X \sim N(\theta, \sigma^2)$ where σ^2 is known and prior density of θ is $g(\theta) \propto cons\tan t$, then find the moment generating function and Bayes estimator of θ using Linex loss function.

6. Let $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ be a random sample from (Y, Z) where Y and Z are independent random variables, having an exponential distribution with parameters λ and μ respectively.

a) Find the MLE for (λ, μ).

b) Suppose now that we only observe: $X_i = \min(Y_i, Z_i)$ and $\Delta_i = I(Y_i \le Z_i)$. Based on the observations: $(X_1, \Delta_1), \ldots, (X_n, \Delta_n)$, find the MLE of (λ, μ).

c) What is meant by a complete statistic of a family of density function? What is the difference between complete sufficient statistic and sufficient statistic? Show that a complete sufficient statistic is minimal sufficient statistic.

7. a) What is location invariant estimator and scale invariant estimator? Show that sample variance is not a location invariant but is a scale invariant estimator.

b) State and prove Chapman, Robbins and Kiefer inequality.

c) Let $X_1, \ldots, X_n$ be a random sample from X with density function:

$$f(x;\theta) = \frac{1}{2}\theta^3 x^2 e^{-\theta x} \; ; x > 0 \, , \theta > 0$$

Give an approximate 90% confidence interval for θ. Calculate this interval if for a random sample of size $n = 50$, the observations have led to the maximum likelihood estimate $\hat{\theta}_n = 10$.

8. a) What is sequential testing procedure? What is SPRT? Show by an example that sequential sampling and testing procedure need a smaller number of sample observations than classical testing procedure to make decision on the hypothesis.

b) For a SPRT, how can you approximately determine k_0 and k_1 with error sizes α and β?

c) Let α' and β' be the error sizes of the SPRT defined by $k_0' = \frac{\alpha}{1-\beta}$ and $k_1' = \frac{1-\alpha}{\beta}$.

Show that $\alpha' + \beta' = \alpha + \beta$.

Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Hons.) Examination 2020
Course Title: Multivariate Analysis
Course No. Stat 402

Time: 4 Hours Full Marks: 70

[Answer any FIVE questions. All questions carry equal marks.]

1. a) What do you mean by multivariate analysis? List different importance of multivariate analysis.

b) Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are iid $N_p(0, \Sigma)$. Find the maximum likelihood estimate (MLE) of Σ. Show that it is an unbiased estimator of Σ.

c) Let $\mathbf{X}$ be $N_3(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu}' = \begin{bmatrix} 5 & 3 & 2 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 4 & 1 \\ 0 & 1 & 9 \end{bmatrix}$.

i) Are $\frac{1}{2}(\mathbf{X}_1 + \mathbf{X}_2)$ and $\mathbf{X}_3$ independently distributed? Explain.

ii) Find the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_1 + \mathbf{X}_2 + \frac{1}{3}\mathbf{X}_3$.

2. a) How can you assess the assumption of multivariate normality? Describe. Discuss the steps of detecting outliers in a higher dimension of the multivariate data set.

b) If $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be independent observations from a population with mean $\boldsymbol{\mu}$ and finite covariance $\boldsymbol{\Sigma}$, then show that $n(\overline{\mathbf{X}} - \boldsymbol{\mu})' \mathbf{S}^{-1}(\overline{\mathbf{X}} - \boldsymbol{\mu})$ is approximately distributed as χ_p^2 for large $n - p$.

c) What are the most common multivariate quality control charts? Define them with their uses. Construct a T^2-chart for data in Table 1. Use $\alpha = 0.01$. Hence, comment.

Table 1: Two measurements of stiffness with bending strength.

x_1	1232	1115	2205	1897
x_2	4175	6652	7612	10914

where, x_1 = stiffness and x_2 = bending strength are two measurements in pounds/(inches)^2 for a sample of 4 pieces of a particular grade of lumber.

3. a) Discuss the use of Hotteling T^2 in multivariate analysis.

b) Derive the distribution of Hotteling T^2 along with its central probability distribution as an extension of the univariate t – distribution. Show that T^2 is invariant under changes in the units of measurements.

c) A wildlife ecologist measured x_1 = tail length (in millimeters) and x_2 = wing length (in millimeters) for a sample of $n = 5$ female Hook-billed kites (a bird in the family Accipitridae). These data displays in the following.

x_1 (tail length)	191	197	180	180	196
x_2 (wing length)	284	285	273	276	288

Using the above data,

i) Evaluate T^2 for testing $H_0 : \boldsymbol{\mu}' = \begin{bmatrix} 190 & 275 \end{bmatrix}$.

ii) Hence, find out the sampling distribution of T^2.

4. a) Define the multivariate multiple regression model. What are advantages of the multivariate multiple regression model than its counterparts?

b) Suppose the maximum likelihood estimator of $\boldsymbol{\beta}=\left[\beta_{(1)}\vdots\beta_{(2)}\vdots\cdots\vdots\beta_{(m)}\right]$ is $\hat{\boldsymbol{\beta}}=(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$ determined under the multivariate multiple regression model with the errors $\boldsymbol{\varepsilon}$ have a normal distribution and the design matrix $\mathbf{Z}$ having full rank$(\mathbf{Z})=r+1$, $n\geq(r+1)+m$, then show that $\hat{\boldsymbol{\beta}}$ has a normal distribution with $E(\hat{\boldsymbol{\beta}})=\boldsymbol{\beta}$ and $\text{Cov}\left(\hat{\beta}_{(i)},\hat{\beta}_{(k)}\right)=\sigma_{ik}(\mathbf{Z}'\mathbf{Z})^{-1}$, $i,k=1,\ldots,m$.

c) Suppose 10 American companies' sales and profit are two responses regressed on the covariate asset, specify the mathematical form of the multivariate multiple regression model. How could you test $H_0:\boldsymbol{\beta}_{(2)}=0$ for that model?

5. a) What is principal component analysis? Write its importance in multivariate analysis.

b) What are the purposes of principal components? Show that the sum of variances of the original variables is equal to the sum of variances of the principal components.

c) Obtain the MLE of the 1st principal component and find its variance.

6. a) What is factor analysis? Make a comparative study of factor model and multivariate regression model.

b) Describe any method to estimate factor loading. Write the importance of factor rotator.

c) Derive the test statistic for assessing the adequacy of a covariance structure explained by some common factors. What are the strategies for factor analysis?

7. a) Define discrimination and classification with examples. Write some likely areas in which they may be employed successfully.

b) Develop Fisher's linear discrimination for several populations mentioning its important objectives.

c) With a suitable example, explain the meaning of canonical correlation analysis.

8. a) Define cluster analysis with its purpose. What are the different distances and similarity coefficients for clustering items? How does the similarity coefficient can be measured from distances?

b) Suppose four individuals possess the following characteristics. Use "No 0-0 matches in numerator or denominator" similarity coefficient to find the homogeneous clusters of individuals.

Group	Age	Height	Eye Color	Gender
Individual 1	48	64 inch	Black	Female
Individual 2	56	67 inch	Black	Male
Individual 3	32	73 inch	Brown	Male
Individual 4	35	62 inch	Black	Female

Define four binary variables X_1, X_2, X_3, and X_4 as

$$X_1=\begin{cases}1 & \text{Age}\geq 40\\ 0 & \text{Age}<40\end{cases},\quad X_2=\begin{cases}1 & \text{Height}\geq 65\\ 0 & \text{Height}<65\end{cases},\quad X_3=\begin{cases}1 & \text{Black eyes}\\ 0 & \text{Otherwise}\end{cases},\quad X_4=\begin{cases}1 & \text{Female}\\ 0 & \text{Male.}\end{cases}$$

c) What are the different clustering methods? Discuss the different steps in agglomerative hierarchical and nonhierarchical cluster methods. What is a dendrogram? Illustrate the dendrogram in Figure 1 for the eight variables index with 0, 1, ..., 7.

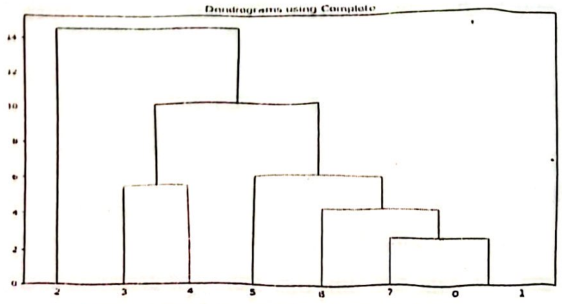


Figure 1: Dendrogram using Complete Linkage.

Department of Statistics
Jahangirnagar University
Part IV B.Sc. (Hons.) Examination 2020
Course Title: Design of Experiment
Course Code.: STAT-403

Time: 4 Hours **Full Marks: 70**

[Answer any FIVE questions. All questions carry equal marks.]

Q1. a) Distinguish between experimental design and sampling design. Explain clearly the basic principles of experimental design stating their uses in planning an experiment.

b) 5 treatments are randomly allocated to 25 homogeneous plots, each being equally replicated. Discuss the method of analyzing data suggesting an appropriate design. Set up the ANOVA table mentioning the test hypothesis and test statistics.

Q2. a) Discuss the layout plan of randomized block design (RBD). Estimate the parameters of its model and perform the analysis of variance. How do you test the hypothesis of treatment effect?

b) Mention the situation where multiple comparison tests can be applied. Discuss some multiple comparison tests known to you.

Q3. a) Distinguish between Latin Square Design (LSD) and Graeco-Latin Square Design. LSD cannot be treated as a complete layout. Justify.

b) State and explain the model of Graeco-LSD. Estimate the parameters of this model and set up the ANOVA table mentioning the test statistics for different hypotheses.

Q4. a) Define factorial experiment. How does it differ from single factor experiment? Distinguish between symmetrical and asymmetrical factorial experiments.

b) Interpret the main and interaction effects of a factorial experiment with factor: A, B, and C each at 2 levels with the procedure of analyzing data.

5. a) What is confounding and what are its necessity? Discuss different types of confounding with examples.

b) Discuss the block consists of 2^4-factorial experiment if $ABCD$ and AC interactions are simultaneously confounded in the same replication. Discuss the procedure of analysis of data to test the hypothesis.

6. a) What is split plot design? Discuss how does it differs from randomized block design and confounded design? Mention the situation where a split plot design is used?

b) Describe different steps of analyzing the data of a split plot design. Explain why there are two types of error in split plot design? Discuss how the important hypothesizes will be tested of this design?

7. a) Define an incomplete block design. When balanced incomplete block design (BIBD) and symmetrical balanced incomplete block design (SBIBD) are obtained? For a SBIBD with usual parameters, show that $(r - \lambda)$ must be a perfect square.

b) Construct a SBIBD with parameters: $b = v = 4\lambda + 3$, $r = k = 2\lambda + 1$ and $\lambda = 1$. Discuss the procedure of ANOVA with recovery of intra-block information.

Q8. a) What do you mean by analysis of covariance (ANCOVA)? Give an example where it is used?

b) Set up a mathematical model for ANCOVA in RBD with two concomitants variables and discus the analysis procedure of such data.

Best of Luck!

Department of Statistics
Jahangirnagar University
Part IV B.Sc. (Hons.) Examination 2020
Course Title: Sampling Technique II
Course Code.: STAT- 404

Time: 2.30 Hours **Full Marks: 35**

[Answer any THREE questions. All questions carry equal marks.]

Q1. a) Define the terms cluster and stratum with suitable example. Explain cluster sampling and enumerate its important features.

b) Explain the concept of primary sampling unit, secondary sampling unit and ultimate sampling unit with suitable example.

c) Give the expressions for the unbiased estimator of mean and the estimator of its variance when the total number of units in the population is known. How will you write the expressions for unbiased estimator of total?

Q2. a) Describe a situation where two stage sampling is appropriate.

b) Explain the differences between stratified sampling and two stage sampling.

c) Prove that the mean per second-stage unit in the sample is an unbiased estimate of the population mean for two-stage sampling with equal first-stage units. Also obtain the variance of the estimate of mean.

Q3. a) Under what circumstance successive sampling is used? Give an example.

b) What are the reasons of using sampling over two successive occasions?

c) Suggest a best linear estimator of a population characteristic (e. g. mean) on current occasion. Verify your proposal.

d) Find an optimum size of unmatched sample on second occasion.

Q4. a) What are the objectives of using double sampling plan? Explain, in brief.

b) Define ratio estimator of $\overline{Y}$ for double sampling plan. Find an approximate expression for bias and mean squared error of this estimator when the second phase sample is a subsample of the first phase sample.

c) Find the optimum size of the first and second phase samples.

Q5. a) What do you mean by inverse sampling? Describe the procedure of estimating population size by direct method and inverse sampling method. Show that the estimator of population size in each method is biased.

b) Define randomized response. Describe the Warner's randomized response technique for estimating the prevalence of sensitive attributes.

Best of Luck!

Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Hons.) Examination 2020
Course Title: Data Mining
Course No. Stat 405

Time: 2.5 Hours Full Marks: 35

[Answer any THREE questions. All questions carry equal marks.]

1. a) What is meant by KDD? Write down the different steps of Knowledge Discovery Process. What are the different types of Visualization Techniques to represent data in data mining?

b) Find the similarity matrix between the following three variables using the Dice, Cosine and Jaccard similarity measures and comment on your result.

X:	7.5	10.1	8.4	9.7	8.4	8.1	2.9	7.5	2.9	7.1
Y:	6.8	4.4	2.8	8.1	5.7	10.1	9.1	6.7	8.4	5.3
Z:	3.9	7.7	3.7	3.4	3.1	2.4	3.7	3.4	1.9	7.1

c) What is meant by machine learning? Briefly describe the different types of machine learning.

2. a) What is meant by ANN and bias node in ANN? What are the different steps to in developing an artificial neural network?

b) Suppose there are two items, {A, B} where $A \Rightarrow B$ has support of 15% and a confidence 60%. Because these values are high, a typical association rule algorithm probably would deduce this to be a valuable rule. However, if the probability of purchase item *B* is 70%, then we see that the probability of purchasing *B* has gone down, presumably because *A* was purchased. Find lift, chi-square, all_confidence, max_confidence, Kulczynski, cosine and imbalance ratio.

3. a) Explain fuzzy sets with an example. What is the difference between fuzzy sets and classical sets? Let

$$A = \{(0,0.9),(1,0.3),(2,0),(3,0.2),(4,0.6),(5,0.4),(6,0.3),(7,0.2),(8,0.1),(9,0)\}$$

and

$$B = \{(0,0.7),(1,0.5),(2,0.9),(3,0),(4,0.2),(5,0.5),(6,0.8),(7,1),(8,1),(9,0.7)\}$$

be two fuzzy sets. Find out the fuzzy intersection and union for the above two sets.

b) What do you mean by fuzzy logic and fuzzy membership function? Discuss different types of membership functions with an example.

c) For the following set of values $(1, 3, 9, 15, 20, 28$ and $31)$ determine the jackknife estimate for both the mean and standard deviation of the mean.

4. a) Define clustering. What are the different types of clustering methods? State the difference between clustering and classification.

b) What do you mean by Decision Tree (DT)? What are the advantage and disadvantages of DT? Describe the different issues of the DT algorithm.

c) What do you mean by attribute selection measures in classification?

5. a) What is meant by Decision Tree Induction? What are the different types of decision tree algorithm? Briefly outline the major steps of decision tree classification.

b) The following **R** outputs for naive Bayesian classification are found for the Titanic data to predict the class label of the attribute survived based on the attributes Class, Sex and Age. Classify the tuple *X*=(Class=3rd, Sex=Male, Age=Adult) from the following **R** output.

```
 Class      Sex          Age        Survived
1st :325   Female: 470   Adult:2092   No :1490
2nd :285   Male  :1731   Child: 109   Yes: 711
3rd :706
Crew:885
naiveBayes.default(x = X, y = Y, laplace = laplace)
A-priori probabilities:
Y
    No      Yes
0.676965 0.323035
Conditional probabilities:
   Class
Y          1st        2nd        3rd       Crew
 No  0.08187919 0.11208054 0.35436242 0.45167785
 Yes 0.28551336 0.16596343 0.25035162 0.29817159
   Sex
Y       Female      Male
 No  0.08456376 0.91543624
 Yes 0.48382560 0.51617440
   Age
Y       Adult      Child
 No  0.96510067 0.03489933
 Yes 0.91983122 0.08016878
```

Department of Statistics
Jahangirnagar University
Part IV B.Sc. (Hons.) Examination 2020
Course Title: Epidemiology and Biostatistics
Course Code.: STAT- 406

Time: 4 Hours **Full Marks: 70**

[Answer any FIVE questions. All questions carry equal marks.]

Q1. a) Define epidemiology. Discuss the different components of epidemiology.

b) Explain the meaning of causation of a disease? What are the aims and objectives of epidemiology?

c) What is epidemiological triads? Explain the terms with example: Endemic, Epidemic and Pandemic. Write some applications of epidemiology in public health.

Q2. a) Define observational and experimental epidemiology. What do you mean by exposure? Explain the cross-sectional study design with an example.

b) How cases and controls are selected in a case-control study? Outline the design of a case-control study and a cohort study. Make a comparison between them.

c) Define randomized control trials. Mention some methods which can be used to control confounding in the design of an epidemiological study?

Q3. a) What does incidence and prevalence of a disease represent? Differentiate between incidence and prevalence. Mention the factors for which they are increased by.

b) Define the following measures: (i) Relative Risk, (ii) Risk Difference, (iii) Population Attributable Risk, (iv) Attributable Risk. Calculate each measures for women who smoke and who have never smoked from the following table and interpret each results:

Table: Relationship between cigarette smoking and incidence rate of stroke in cohort of 1,18,539 women

Smoking Category	No. of cases of stroke	Person-years of observation (over 8 years)	Stroke incidence rate (per 1,00,000) person- years)
Never Smoked	70	395 594	17.7
Ex-smoker	65	232 712	27.9
Smoker	139	280 141	49.6
Total	274	908 447	30.2

c) Explain Rate, Ratio and Percentage with example. Is there any difference among them?

Q4. a) What do you mean by screening tests? Give some examples of screening tests.

b) Discuss with an example: false positive and false negative, sensitivity and specificity, predictive value positive and predictive value negative.

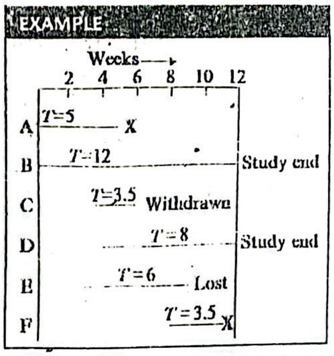
c) Define exponential regression model. Derive the likelihood function for this model. Also explain the estimation method and testing procedure for this model.

Q5. a) What is surveillance? Explain the necessity of surveillance system. Discuss active surveillance and passive surveillance. Also explain the elements of surveillance system.

b) What is a contagious disease? What is HIV/AIDs? What are the risk factors for HIV? What are the prevention and control strategies for this disease?

Q6. a) What is survival analysis? Give some examples of survival analysis. What are the go survival analysis?

b) What is censoring? Why censoring may occur? Explain different types of censoring fro following example:



c) Define survivor function and hazard function with their properties. Why hazard is a rate ra than a probability? Explain.

Q7. a) What is a Kaplan-Meier Survival curve? Given the following the survival time data (in ye for 21 participants.

17+, 6+, 19+, 9+, 20+, 10+, 11+, 13, 22, 16, 23, 6, 25+, 35+, 32+, 32+, 34+, 6, 7, 10,

Compute: **i)** $m_j, q_j, R(t_{(j)}), \hat{S}(t_{(j)})$, **ii)** Average survival time, **iii)** Average hazard rate.

b) How can you evaluate whether KM curves for two or more groups are statistically equiva Use the following results to compute the log-rank test statistic. What is your null hypoth and how is the test statistic distributed under this null hypothesis? What are your conclusi about the test?

$t_{(j)}$	m_{1j}	m_{2j}	n_{1j}	n_{2j}	e_{1j}	e_{2j}
1.4	0	1	25	25	0.500	0.500
1.6	0	1	25	24	0.510	0.490
1.8	1	1	25	23	1.042	1.958
2.2	1	0	24	22	0.522	0.478
2.4	0	1	23	22	0.511	0.489
2.5	1	0	23	21	0.523	0.477
2.6	1	0	22	21	0.516	0.484
2.8	0	1	21	21	0.500	0.500
2.9	0	1	21	20	0.512	0.488
3.0	1	0	21	19	0.525	0.475

c) State whether the following statements are true or false. If false, correct it.

i) The hazard function theoretically has no upper bound

ii) The risk set at six weeks is the set of individuals whose survival times are less than equal to six weeks.

iii) The measure of effect used in survival analysis is an odd ratio.

iv) A hazard rate of one per day is equivalent to seven per week.

v) In practice, the survivor function is usually graphed as a smoothed curve.

Q8. a) What is the Cox-proportional hazard model?

b) What is the Cox regression model? What are the assumptions of Cox regression model? Wh is the Cox Regression model used for?

c) How do you fit a Cox-proportional hazard model? What are the advantages of Cox-proportio hazard model vs logistic regression?

Best of Luck!

Department of Statistics

Jahangirnagar University

Part IV B.Sc. (Hons.) Examination 2020

Course Title: Advanced Demography

Course Code.: STAT- 407

Time: $2\frac{1}{2}$ Hours **Full Marks: 35**

[Answer any THREE questions. All questions carry equal marks.]

Q1. a) Define Demographic transition theory. In which stages, ZPG arises and why? Show graphically how different demographic indicators changed from one stage to another stage?

b) What is population explosion? In which stage population explosion arises and why? Write down the demographic indicated value of Bangladesh in recent time and explain in which stage Bangladesh lies based on indicated value?

c) What is population policy? Point out major objectives of population policy? How it can be applied in Bangladesh? What are the obstacles if a country wants to apply population policy?

Q2. a) Discuss the concept of population projection and forecasting. What are the important base information necessary for population projection? Discuss them with merits and demerits.

b) What are the different methods of population projection known to you? Which method of projection you will choose for Bangladesh and why?

c) How low, medium, and high population projection are carried out? What are the uses of population projection?

Q3. a) Define proximate determinants of fertility. Is there any relationship among proximate determinants of fertility and TFR? If they are related which factor more affected to TFR? How will you estimate the different indices?

b) If c_m= 0.750, c_c= 0.489, c_a= 0.959, c_i= 0.823 and TF = 15.3, then estimate TFR and explain the estimated value. Calculate the impact on fertility reduction using given indices and explain fertility inhibiting effects of the proximate determinants using calculating value of the fertility reduction.

c) Describe how will you project the future fertility using Bongaarts model. If TFR in 2007 is 2.7 and expected TFR in the year 2025 is 2.1, what will be the required contraceptive prevalence rate to achieve the replacement level fertility? Calculate it and explain it.

Q4. a) Describe briefly the computational procedure of estimating female adult mortality using maternal orphan hood status of respondents.

b) The given data is children ever born during the 12 months preceding the survey by age of mother at time of survey. Calculate mean age at maternity and weighting factors using given information.

Age group of mother at the time of the interview	15-19	20-24	25-29	30-34	35-39	40-44	45-49	Total
No. of children ever born in past year, B(i)	136	409	485	320	259	94	50	753

c) Estimate survivorship probabilities using widowhood data classified by age. The given data is proportions single ignoring those of unknown marital status. Calculate singulate mean age at marriage and explain it.

Age group	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54
Proportions single for female	0.8101	0.3918	0.1608	0.0936	0.0628	0.0325	0.0470	0.0323

Q5. a) What is migration? What are the important causes of migration? Define pull factor and pus factor. What are the positive and negative consequences of migration?

b) What is city size distribution? What is census coverage? Using the Levis diagram develop th census coverage model and interpret the different parameters.

Best of Luck!

Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Hons.) Examination 2020
Course Title: Stochastic Process
Course No. Stat 408

Time: 2.5 Hours Full Marks: 35

[Answer any THREE questions. All questions carry equal marks.]

1. a) Define Stochastic Process. Categorize Stochastic Process based on Time and Space with appropriate example.

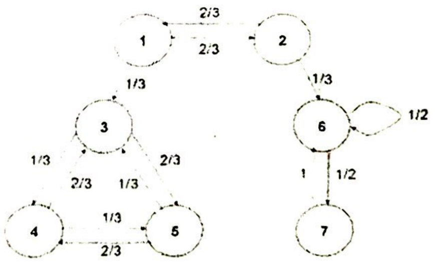
b) Suppose $[X(t), t\in T]$ be a stochastic process, where $\Pr[X(t)=n]=\frac{e^{-at}(at)^n}{n!}$; $n=0,1,2,\dots,a>0$. Is this process stationary?

c) Define Martingales Process. Let $\{Z_i\}; i=1,2,\cdots$ be a sequence of *i.i.d.* random variables with mean 0 and let $X_n=\sum_{i=1}^{n} Z_i$, then show that $\{X_n\}_{n=1}^{\infty}$ is a martingale.

2. a) Define Markov Chain. How Chapman-Kolmogorov equation used in computing transition probabilities?

b) Discuss absorbing state, class of states and communicate states with examples. Suppose state *i* communicate with state *j* and state *j* communicate with state *k*. Show that state *i* also communicates with state *k*.

c) Consider the following finite-state Markov Chain



i) Identify all the classes present in the chain and the states belonging to each class.

ii) Find the period of each class and determine whether the class is transient or recurrent.

3. a) When is a counting process said to be a Poisson Process? Find the distribution of number of arrivals in a Poisson process.

b) If $\{N(t), t\geq 0\}$ is Poisson Process and $s<t$, then prove that

$$\Pr\{N(s)=k \mid N(t)=n\}=\binom{n}{k}\left(\frac{s}{t}\right)^k\left(1-\frac{s}{t}\right)^{n-k}.$$

c) Suppose car passes Prantic Gate at a Poisson rate of 1 per minute. If 5% of the cars are Toyota, then:

i) What is the probability that at least one Toyota passes by during an hour?

ii) Given that 10 Toyotas had passed by an hour, what is the expected number of cars has passed by at that time.

iii) If 50 cars have passed by an hour, what is the probability that 5 of them are Toyotas.

4. a) Define Renewal Process. State and derive the distribution of Renewal Process.

b) What is Renewal function? Show that Renewal process uniquely determines the distribution function.

5. a) Define Queuing Process with Example. What are the common characteristics of a Queuing process? What a Queuing process measure?

b) Derive the distribution of Queueing Time and derive the Average amount of time that a customer spends for waiting in the queue.

c) The arrivals at a counter in a bank occur in accordance with the Poisson process at an average rate of 8 per hour. The duration of service of a customer has an exponential distribution with a mean of 6 minutes that is $\frac{6}{60}$ hours.

i) Find the probability that an arriving customer has to wait on arrival
ii) Find the probability that 4 customers are in the system
iii) Find the probability that an arriving customer has to spend less than 15 minutes in the bank.
iv) Also estimate the fraction of time that the counter is busy.

Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Hons.) Examination 2020
Course Title: Bioinformatics
Course No. Stat 409

Time: 2.5 Hours — Full Marks: 35

[Answer any THREE questions. All questions carry equal marks.]

1. a) What is the genetic information? Why deoxyribonucleic acid (DNA) is called the carrier of the genetic information? Explain. Why do individuals differ with regard to their genetic information?

   b) Define the following terms:

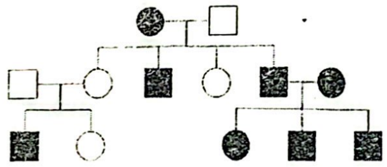
   i) Phenotype ii) Heredity iii) Mutation iv) Replication of DNA
   v) Genetic recombination vi) Single nucleotide polymorphism (SNP)

   c) What is a genotype for DNA sequencing data? With a suitable example, explain all the possible types of genotype for a diallelic locus?

2. a) Define the penetrance function. How would you apply this function to calculate the probabilities of being affected depending on the genotypes? Explain for a diallelic locus.

   b) What is an ancestry chart? Sketch an ancestry chart for 4 generations, where disease transmission is due to autosomal dominant fashion but showing the reduced penetrance in the 3rd generation. Also, write down the characteristics of your chart.

   c) Mention the characteristics of disease transmission pattern of the following ancestry chart.



3. a) Discuss the importance of biological databases in bioinformatics. What are the available biological databases? List any two Protein databases.

   b) What kind of information is stored in the EMBL-EBI databases? What other databases with this information exist? Find the major equivalents in the US and in Japan.

   c) What are the differences among Homology, Similarity and Identity in Bioinformatics?

4. a) What is Multiple Sequence Alignment (MSA)? What is the available algorithm of MSA? Write briefly.

   b) Why scoring matrix are essential for multiple sequence alignment? What is the difference between PAM and BLOSUM?

   c) How Markov Chains and Hidden Markov Models are used in Bioinformatics?

5. a) What is the Genome-Wide Association Studies (GWAS)? Mention the properties of a marker that are necessary to retain to perform a GWAS.

   b) What is genotyping? Explain the important features of the Illumina genotyping technology. How the missing genotypes are treated in GWAS?

   c) How would you present the inheritance patterns of a particular disease into probabilistic models? Explain for the all-possible patterns. Also, explain the testing procedure of genetic association for any model.

Department of Statistics
Jahangirnagar University
Part IV B.Sc. (Hons.) Examination 2020
Course Title: Categorical Data Analysis
Course Code.: STAT- 410

Time: 4 Hours **Full Marks: 70**

[Answer any FIVE questions. All questions carry equal marks.]

Q1. **a)** Define categorical variable. Explain the different distributions for categorical data.

b) Discuss the statistical inference for categorical data.

c) Explain the significance test about a binomial proportion.

Q2. **a)** For 2×2 contingency table with cell probabilities π_{ij} where $i, j = 1, 2$. Explain the independence hypothesis, theoretical odds ratio in terms of π_{ij} and empirical odds ratio in terms of n_{ij}. Does the odds ratio change if rows and column are interchanged?

b) During a study of car accidents, the Highway Safety Council found that 65% accidents occurred at night, 52% accidents were alcohol-related and 20% accidents occurred at day and other-related.

i. Conduct a test of statistical independence. Report p-value and interpret.

ii. Also, estimate and find a 95% confidence interval for the population odds ratio.

Q3. **a)** Describe the components of a generalized linear model (GLM).

b) Consider the Binomial (n, p) distribution with probability mass function

$$f(y/p) = \binom{n}{y} p^y (1-p)^{n-y}, \qquad y = 0, 1, \dots, n$$

Show that $f(y/p)$ can be written in the canonical GLM form and write down an expression for the canonical parameter θ in terms of p.

c) Given a single observation y from $Binomial\ (n, p)$, calculate the deviance, expressed as a function of $y, n\ and\ p$.

Q4. **a)** Briefly explain the role of the link function in a GLM. Show that the normal distribution $N(\mu, \sigma^2)$ can be written in GLM form.

b) Define the generalized linear mixed model (GLMM) and discuss the logistic GLMM for binary matched pairs.

Q5. In the 2×2 contingency tables below, the data relating smoking to cervical cancer in women, have been stratified by the number of sexual partners that a woman has had.

Zero or One Partners			
	smoker		
Cancer	Yes	No	Total
Yes	12	25	37
No	21	118	139
Total	33	143	176

Two or More Partners			
	smoker		
Cancer	Yes	No	Total
Yes	96	92	188
No	142	150	292
Total	238	242	480

a) Compute the Mantel-Haenszel estimate of the summary odds ratio. Interpret your result.

b) Can you give a 99% confidence interval for the estimated odds ratio? Interpret your results.

c) What do you conclude about the relationship between smoking and cervical cancer?

Q6. a) What is a count response? Give an example. Identify the natural parameter and canonical link of Poisson responses and hence derive the Poisson log-linear model. What do you mean by over dispersion of a Poisson GLM?

b) What is an exponential dispersion family distribution? Identify dispersion and natural parameters. Show that Poisson distribution belongs to the exponential dispersion family.

c) Based on the above family distribution, derive the general expression of mean and variance function of random components and hence, derive the mean and variance function of a Poisson response.

Q7. Consider the data below obtained on car sales in Dhaka on a certain week.

Number sold	Total number	Gender	Income (> tk60,000)
23	40	M	Yes
12	18	M	No
36	51	F	Yes
7	21	F	No

i) Write down a regression model for estimating the probability that a certain type of person purchases a car. State all assumptions with appropriate notation.

ii) Using the notation in part (i), write down the log-likelihood for the regression coefficient vector, and the expression for the deviance.

iii) Under the model in (i) write down the expression for an estimate of the average number of cars sold to a male with income over tk 60, 000.

Q8. a) Describe how a generalized linear model may be used to relate response variables $Y_1, Y_2, \dots, Y_n$ to a set of explanatory variables $X_1, X_2, \dots, X_p$. Explain the log linear model of independence for two way contingency table with interaction effect.

b) Discuss the inference for log linear model for two way contingency table.

DEPARTMENT OF STATISTICS

Jahangirnagar University

BSc (Part IV) Examination - 2020

Course Name: Statistical Data Analysis IV

STAT-LAB-411
(Group A)
Time: 3 hours
Full Marks: 40

Answer ALL questions.

1. **(20 Marks)** Let X be a r.v. with Poisson distribution with parameter $\lambda > 0$. Let $X_1, \ldots, X_n$ be a random sample from X. Consider as a prior density a Gamma density with parameters α and β.

(a). Find the Bayes estimator for λ, and calculate its bias, variance and Mean Squared Error (using the quadratic loss function).

(b). Derive the Maximum Likelihood Estimator (MLE) for λ, and calculate its bias, variance and Mean Squared Error.

(c). Plot for sample sizes $n = 5, 10$ and 20, the Mean Squared Errors, as a function of λ, for the MLE and for the Bayes estimator, when using a $\Gamma(10, 0.1)$ as a prior density.
Comment on this plot. Explain the behaviour of the plotted Mean Squared Errors.

(d). Consider the sample size $n = 10$. Plot the Mean Squared Errors, as a function of λ, of the following estimators:

- the MLE estimator
- the Bayes estimator using $\Gamma(0.5, 0.1)$ as a prior density.
- the Bayes estimator using an Exponential density with parameter 0.002 as a prior density.
- the Bayes estimator using $\chi^2(6)$ as a prior density.

Comment on this plot.

(e)(e1). Simulate a sample of size $n = 50$ from a Poisson distribution with $\lambda = 4$. Based on this sample of observations calculate the realisations of the estimators $\widehat{\lambda}_n^{\text{MLE}}$ and two Bayes estimators $\widehat{\lambda}_{1n}^{\text{B}}$ and $\widehat{\lambda}_{2n}^{\text{B}}$ when using as prior density a Gamma density $\Gamma(9, 0.5)$ and $\Gamma(10, 1)$ respectively.

(e2). Repeat the previous action 500 times (Monte Carlo simulation), obtaining as such 500 realized values for the estimators $\widehat{\lambda}_n^{\text{MLE}}$ and $\widehat{\lambda}_{1n}^{\text{B}}$ and $\widehat{\lambda}_{2n}^{\text{B}}$ (i.e. 500 estimates for the parameter λ). Provide boxplots for each of the estimators, based on the 500 samples. Comment on your findings.

2. **(10 Marks)** Let us consider a Bayesian Poisson regression model for the frequency of number of death due to COVID-19 say Y with density

$$Y_i|\mu_i \sim \text{Poission}(\mu_i)$$

where, $Y_i = 0, 1, 2, \ldots$. Assuming $\log(\mu_i)$ can be expressed by the linear combination of X predictors, we can write

$$\text{data:} \quad Y_i|\beta_0, \beta_1, \beta_2 \sim \text{Poission}(\mu_i) \quad \text{with} \quad \log(\mu_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = X_i \boldsymbol{\beta},$$

where, $\boldsymbol{\beta}$'s represents the regression parameters and X_i's are the covariates as $X_i = \{1,$ temperature, humidity$\}$. The data are given in Table 1. Suppose the following prior distribution are placed on parameter where $\pi()$ indicates a prior distribution;

$$\pi(\beta_0) = \text{uniform}(0, 1) \quad \text{and} \quad \pi(\beta_1).\pi(\beta_2) = \text{normal}(0, \sigma^2).$$

Table 1: Data for the Q2.

death	temperature	humidity
3	30	52
1	31	28
6	33	39
3	33	41
4	31	48
5	32	39
7	32	29
4	31	33
10	32	39
15	32	27
9	33	37
7	33	27
10	30	37
9	30	49
10	32	30
7	33	34
4	35	34
9	36	23
5	36	38
8	35	23

(a). Find the realized value of $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)^t$ by using Gibbs Sampling Algorithm.

(b). Provide the 90% credible interval of the $\widehat{\beta}$.

(c). Is $\widehat{\beta}_i$ $(i = 0, 1, 2)$ significant? Interpret your results.

(d). How do check the convergence diagnostic of the Bayesian estimates. Mention your model diagnosis results.

3. **(10 Marks)** Real data are collected from 140 patients. The decisions involved whether to give the patient thrombolysis or not in the emergency room. There are Canadian guidelines as to whether to thrombolysis such patients or not, so each physician was evaluated as th whether they followed the guidelines for each subject they treated or not. For example, the first physician followed the guidelines in 19 out of 20 patients she/he treated. While each physician has their own rate of "success" (following the guidelines), it may be that overall, these rates may themselves have a distribution. The idea is to estimate both each physicians rate, as well as the rate of the "next" physician.The latter is accomplished by looking at the posterior distribution of the "extra" variable "y". The data are presented in Table 2. Note the hierarchical structure, the rates follow a distribution:

$$Y_i|\theta_i \sim \text{Binomial}(n_i, \theta_i).$$

Suppose the prior distribution of θ_i is

$$\text{logit}(\theta_i) \sim \text{normal}(\mu, \sigma^2),$$

with hyperprior

$$\mu \sim \text{normal}(0, 10^2) \text{ and } \sigma^2 \sim \Gamma(100, 100).$$

(a). Find the posterior mean and standard error of $\widehat{\mu}$.

Question 3 continues on the next page...

Table 2: Data for the Q3.

n	x
20	19
6	5
20	18
12	4
4	11
24	23
10	10
18	16
26	24

(b). Find the Bayesian estimate of the predictive distribution for rate of each group.

(c). Show the summary statistics of the Bayesian estimate of $w = \exp(y)/(1 + \exp(y))$, where $y \sim$ normal(μ, σ^2).

(d). Discuss your WinBUGS output and check the model diagnosis.

Department of Statistics
Jahangirnagar University
Part IV B.Sc. (Hons.) Examination 2020
Course Title: Statistical Data Analysis VIII
Course No.: STAT LAB - 411 (Group B)

Time: 1.30 Hours **Full Marks: 20**

Q1. Make a hypothetical data of split-plot design experiment in an agricultural study where the main-plot has 3 levels and sub-plot has 4 levels each replicated 4 times and then

i. Analyze the data and test the significance of main plot treatment.

ii. Also test the significance of the subplot treatment effects.

iii. Is whole plot and sub-plot interaction significant?

Q2. An experiment on a small grain was conducted with 2 variates $V_1$and V_2, 3 fertilizes f_1, f_2, f_3 in 4 blocks. The yield results are given below:

F	V	1	2	3	4
1	1	166	113	103	180
	2	192	208	171	196
2	1	145	231	168	216
	2	231	190	171	242
3	1	204	172	178	175
	2	227	144	186	230

i. Analyze the data.

ii. Partition the treatment SS into their orthogonal component each with 1 df.

iii. Test the significance of each component.

Q3. Make the layout and the hypothetical yield data of an experiment conducted in a BIBD where $b = 4,\ v = 4,\ r = 3,\ k = 3$, and $\lambda = 2$ and then

i. Analyze the data.

ii. Test the significance of difference among the treatment.

iii. Test the significance of difference between the adjusted means of 2nd treatment and 4th treatment.

Department of Statistics
Jahangirnagar University
Part IV B.Sc. (Hons.) Examination 2020
Course Title: Statistical Data Analysis IX
Course No.: STAT LAB - 412 (Group A)

Time: 1.30 Hours **Full Marks: 20**

The Madison, Wisconsin, police department regularly monitors many of its activities as part of an ongoing quality improvement program. Table 1 gives the data on five different kinds of overtime hours. Each observation represents a total for 12 pay periods, or about half a year. This data is available in the file **multivar5th/T5-8.dat**.

Table 1: Five types of overtime hours for the Madison, Wisconsin, Police department

Legal Appearances Hours	Extraordinary Event Hours	Holdover Hours	Compensatory Overtime Allowed	Meeting Hours
x_1	x_2	x_3	x_4	x_5
3387	2200	1181	14861	236
3109	875	3532	11367	310
⋮	⋮	⋮	⋮	⋮
3516	1223	1175	15078	161

1. Examine the multivariate normality of the observations on five types of overtime hours for the Madison, Wisconsin, Police department.

2. Evaluate T^2 of the five variables ($x_1, x_2, \ldots, x_5$) for testing $H_0 : \mu' = [3500 \quad 1400 \quad 2600 \quad 13500 \quad 800]$. Hence, find out the sampling distribution of T^2.

3. Construct the principal component analysis using the sample covariance matrix S for the above data matrix.

   (i). Determine the sample principal components and their variances for the covariance matrix S.

   (ii). Compute the proportion of total variance explained by the first two principal components. Interpret your result.

4. Conduct the factor analysis with 5 variables (five types of overtime hours for the Madison, Wisconsin, Police department) and $m = 2$ common factors.

   (i). Find the matrix of specific variances. Hence, define the most significant variable which fit neatly into our factors.

   (ii). Find the estimated factor loadings and communalities. Interpret the estimated factor loadings.

   (iii). What proportion of the total population variance is explained by the first common factors? And by the 2nd common factor.

   (iv). Check whether the 2 factors are adequate for our model?

5. Calculate the sample correlation matrix R for the above data matrix. Arrange the data into

two sets of variables, i.e., $\mathbf{X}^{(1)} = \{X_1, X_2, X_3\}$ and $\mathbf{X}^{(2)} = \{X_4, X_5\}$.

(i). Find all the sample canonical correlations and all the pairs of sample canonical
Hence, interpret the first sample canonical variates $\hat{U}_1$ and $\hat{V}_1$.

(ii). Let $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ be the sets of standardized variables corresponding to $\mathbf{X}^{(1)}$ anc
respectively. What proportion of the total sample variance of the first set
explained by the canonical variate $\hat{U}_1$? What proportion of the total sample varia
the $\mathbf{Z}^{(2)}$ set is explained by the canonical variate $\hat{V}_1$?

6. Calculate the Euclidean distances between five different variables of overtime hours. Clus
the five variables using the single linkage and complete linkage hierarchical methods. Dra
the dendrograms and compare the results.

- Good Luck -

Department of Statistics
Jahangirnagar University
Part IV B.Sc. (Hons.) Examination 2020
Course Title: Statistical Data Analysis X
Course No.: STAT LAB - 413 (Group A)

Time: 2 Hours **Full Marks: 21**

Q1. A television team is interested in estimating total number of votes cast up to 1 p.m. on the day of election in a rural constituency. The total number of villages in the constituency is 36, and the available frame showing total number of persons eligible to vote in each village is given in the following table.

Village #	Total votes	Village #	Total votes	Village #	Total votes	Village #	Total votes
1	900	10	576	19	341	28	681
2	1340	11	1083	20	649	29	990
3	860	12	1644	21	1366	30	1232
4	1716	13	871	22	1199	31	749
5	405	14	605	23	890	32	836
6	704	15	970	24	667	33	910
7	816	16	1413	25	1380	34	1060
8	1426	17	1136	26	571	35	1270
9	1113	18	870	27	1570	36	1710

a) Select a PPS with replacement, sample of six villages, taking size variable as the total number of votes in the village, using (1) cumulative total method, and (2) Lahiri's method.

b) Suppose that The number of votes which were casted up to 1 p.m. for the selected villages are given below :

578	780	698	517	780	1121

Estimate total number of votes which were casted up to 1 p.m. in the constituency. Also find (approximate) 95% confidence interval of total number of casted votes

Q2. A population of agriculture holders of ten villages for 1967 and 1971 is given. Select a sample or with n = 2, under RHC sampling scheme and estimate total number of agriculture holders for 1971. Find the estimate of sampling variance of this estimator.

Village #	No of Agriculture holders (1971)	No of Agriculture holders (1967)
1	60	56
2	55	50
3	60	45
4	70	60
5	75	62
6	65	65
7	50	51
8	60	55
9	65	53
10	80	70

Q3. A graduate student of statistics was asked to estimate average time per week for v the undergraduate students of a certain university view television (TV). The overall gra point average (OGPA) of the students was taken as the auxiliary variable. As th investigator found it difficult to record OGPA of all the 1964 undergraduate studen a first-phase sample of 150 students was selected. The OGPA of the students includ in this initial sample were recorded from their personal files in the Registrar's offic The average OGPA for the first-phase sample was computed as 2.87 (on 4.00 basi A subsample of 36 students was then selected from the first-phase sample. The studer selected in the subsample were contacted personally to find the total time for which th view television in a week.

Table 3 The OGPA (x) and number of hours per week (y) devoted to TV viewing

Student #	y	x	Student #	y	x	Student #	y	x
1	8	2.51	13	14	2.86	25	5	3.25
2	3	3.41	14	11	2.24	26	3	3.49
3	1	3.25	15	3	3.43	27	8	2.63
4	5	3.04	16	6	3.25	28	4	3.61
5	12	2.73	17	7	2.73	29	13	2.17
6	6	3.10	18	5	2.91	30	14	3.10
7	9	2.58	19	4	3.07	31	6	3.01
8	2	3.46	20	10	2.61	32	8	2.58
9	0	3.69	21	8	2.48	33	9	2.41
10	8	2.83	22	12	3.39	34	4	2.96
11	9	2.91	23	6	2.95	35	5	2.85
12	6	3.06	24	1	3.77	36	10	3.74

Estimate average time per week of viewing TV. Also, obtain 95% confidence interval of averag

Q4. The co-operative societies of a state provide loans to farmers. The society declares an individ as a defaulter if he/she does not repay the loan within the specified time limit. An investigator interested in estimating the average amount of loan, per society, standing against the defaulte The total number of co-operative societies in the state is 10126. However, the list of all t societies is not available at the state headquarter but the same is available block level. Therefore, it seems appropriate to use two-stage sampling for selecting sample of societies. Keeping in view the budget and time constraints, it was decided select 12 blocks from the total of 117 blocks and approximately 10 percent of the societ from each of the sample blocks. The information obtained from the selected societ is given in Table 4.

Table 4 Dues (in '000 rupees) standing against the defaulters

Block	M_i	m_i	Amount due from defaulters										
1	60	6	12.5	36.4	26.0	55.6	58.1	40.8					
2	102	10	57.4	16.8	20.3	70.1	34.6	22.6	44.9	28.4	175	33.7	
3	48	5	12.9	41.6	34.7	30.8	61.1						
4	113	11	28.7	82.4	37.3	41.9	24.7	36.6	39.3	49.6	26.0	76.8	51.6
5	92	9	44.8	42.9	51.7	28.8	36.4	40.1	61.6	47.8	77.4		
6	57	6	31.6	24.8	69.9	44.9	59.7	38.6					
7	82	8	49.6	36.9	27.3	63.6	73.0	44.9	87.1	61.2			
8	96	10	53.5	34.9	41.5	43.4	56.6	28.9	23.4	32.8	60.2	47.6	
9	53	5	41.7	549	33.9	27.9	46.3						
10	71	7	24.4	38.9	47.8	45.0	32.6	66.5	58.3				
11	77	8	42.9	37.3	30.8	51.9	60.1	34.6	28.4	38.3			
12	56	6	44.7	34.9	61.7	74.6	37.4	49.2					

Estimate the average amount, per society, standing against defaulters, and also comp 95% confidence interval for it.

xtensive damage has been caused to wheat crop by a storm in a certain area. The farmers of the fected area, consisting of 20 villages, approached the Government for compensation of this ss. In order to decide the amount of compensation, the administration needs to assess the total sidual yield of the crop in that area. Keeping the objective in view, select a WOR sample of five llages using Sen Midzuno method with initial selection probabilities proportional to the area der wheat crop. **Estimate the total yield of wheat along with its standard error**. The area der wheat crop (in hectares) for 20 villages of the population is given in the following table. e yield (in '000 quintals) are given for all villages for the sake of convenience, but in practice se will not be known in advance.

Village #	Area under wheat	Total yield (000)
1	9**	21.16
2	250	4.55
3	220	4.312
4	460	7.886
5	235	5.464
6	970	19.468
7	603	14.532
8	785	13.785
9	1425	30.068
10	2196	45.589
11	315	6.892
12	975	18.857
13	426	8.375
14	524	9.815
15	1245	28.099
16	1040	21.362
17	1550	32.116
18	1576	36.075
19	570	7.540
20	1140	25.223

st two digits of your examination roll number

- Good Luck -

Department of Statistics
Jahangirnagar University
Part IV B.Sc. (Hons.) Examination 2020
Course Title: Statistical Data Analysis X
Course No.: STAT LAB - 413 (Group B)

Time: 1.30 Hours **Full Marks: 21**

Q1. Let $\{X_n\}$ be a Markov chain with the state space $\{1,2,3\}$ and transition probability matrix 3

$$P = \begin{bmatrix} 0 & 0.4 & 0.6 \\ 0.25 & 0.75 & 0 \\ 0.4 & 0 & 0.6 \end{bmatrix}$$

Let the initial distribution be-

$p(0) = [p_1(0), p_2(0), p_3(0)] = [0.4, 0.2, 0.4]$

Calculate the following probabilities:

a. $P[X_1 = 2, X_3 = 2, X_3 = 1 | X_0 = 1]$

b. $P[X_1 = 2, X_3 = 2, X_3 = 1]$

c. $P[X_1 = 2, X_4 = 2, X_6 = 2]$

Q2. Consider the Markov Chain shown in Figure 1. 5

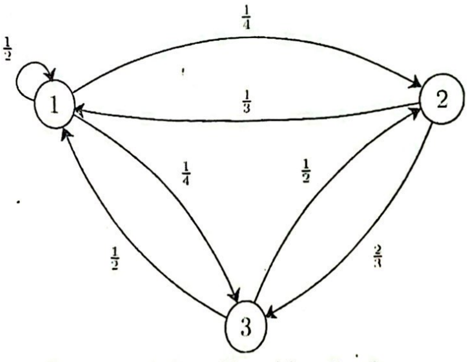


Figure 1. A State Transition Graph

a) Is this chain irreducible?
b) Is this chain aperiodic?
c) Find the stationary distribution for this chain
d) Is the stationary distribution a limiting distribution for the chain?
e) Construct the Transition Probability Matrix.

Q3. A taxi driver conducts his business in three different towns 1, 2, and 3. On any given day, when he is in town 1, the probability that the next passenger he picks up is going to a place in town 1 is 0.3, the probability that the next passenger he picks up is going to town 2 is 0.2, and the probability that the next passenger he picks up is going to town 3 is 0.5. When he is in town 2, the probability that the next passenger he picks up is going to town 1 is 0.1, the probability that the next passenger he picks up is going to town 2 is 0.8, and the probability that the next passenger he picks up is going to town 3 is 0.1. When he is in town 3, the probability that the next passenger he picks up is going to town 1 is 0.4, the probability that the next passenger he picks up is going to town 2 is 0.4, and the probability that the next passenger he picks up is going to town 3 is 0.2. 5

a) Determine the state-transition diagram for the process.
b) Give the transition probability matrix for the process.

c) What is the limiting-state probabilities?

d) Given that the taxi driver is currently in town 2 and is waiting to pick up his first customer for the day, what is the probability that the first time he picks up a passenger to town 2 is when he picks up his third passenger for the day?

e) Given that he is currently in town 2, what is the probability that his third passenger from now will be going to town 1?

Q4. Suppose that customers arrive at a counter in accordance with a Poisson process with mean rate of 2 per minute ($\lambda = 2/\text{minute}$). Then the interval between any two successive arrivals follows exponential distribution with mean $\frac{1}{\lambda}=\frac{1}{2}$ minute. What is the probability that the interval between two successive arrivals is

a) More that 1 minute 4

b) 4 minutes or less

c) Between 1 and 2 minutes

d) Find the expected time until the 9th customer.

Q5. The arrivals at a counter in a bank occur in accordance with the Poisson process at an average rate of 8 per hour. The duration of service of a customer has an exponential distribution with a mean of 6 minutes that is $\frac{6}{60}$ hours. 4

a) Find the probability that an arriving customer has to wait on arrival.

b) Find the probability that 4 customers are in the system..

c) Find the probability that an arriving customer has to spend less than 15 minutes in the bank.

d) Also estimate the fraction of time that the counter is busy.

- **Good Luck** -

Department of Statistics
Jahangirnagar University
Part IV B. Sc. (Honors) Practical Examination – 2020
Course No. : STAT LAB – 414 (Group A)
Course Name: Statistical Data Analysis XI

Answer the following questions.

Time: 03 Hours **Marks: 21**

Q1. The following data at "**Table 1:** Monthly normal maximum temperature of different observatories in Bangladesh" represent the monthly normal maximum temperature (°C) for different 34 observatories of the Bangladesh Meteorological Department (BMD). [Source: MET report, no. 08/2016, ISSN 2387-4201, Title: Climate of Bangladesh, Page - 19, Chapter 3]. You must use your Examination roll as a random seed point.

Table 1: Monthly normal maximum temperature of different observatories in Bangladesh.

Station	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Period
Barisal	25.5	28.5	32.4	33.5	33.4	32	31.2	31.4	31.7	31.7	29.8	26.8	1981-2010
Bhola	25.6	28.5	31.9	33	32.9	31.7	30.8	31.2	31.4	31.7	29.8	26.9	1981-2010
Bogra	24.4	27.5	31.4	33.5	33.3	32.8	32.1	32.5	32.2	31.9	30.2	26.6	1981-2010
Chandpur	24.6	27.9	31.7	33.1	33.2	32.2	31.5	31.8	31.8	31.6	29.5	26.2	1981-2010
Chittagong	26	28.3	30.8	31.9	32.4	31.7	31	31.4	31.8	31.7	30	27.2	1981-2010
...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Sandwip	25.4	27.6	30.3	31.5	31.7	30.8	30.1	30.6	31	31.3	29.5	26.6	1981-2010
Satkhira	25.6	28.8	33	35.1	35.2	33.6	32.2	32.3	32.3	32.2	30.1	26.9	1981-2010
Sitakunda	26.6	28.9	31.4	32.3	32.5	31.4	30.6	31.3	31.8	32.1	30.4	27.8	1981-2010
Srimangal	25.1	28.1	31.6	32.9	32.2	32.1	32.1	32.5	32.2	31.5	29.3	26.6	1982-2010
Sayedpur	22.8	26.6	30.8	32.3	32.5	32.2	32.1	32.5	32.1	31.1	28.9	25.3	1991-2010
Sylhet	25.6	27.7	30.7	31	31.2	31.3	31.5	32.1	31.7	31.4	29.6	26.7	1981-2010
Tangail	23.9	27.5	31.7	33.9	33.4	32.7	31.9	32.2	32.1	31.7	29.4	26	1987-2010
Teknaf	27.4	29.1	31	32.2	32.3	30.6	29.9	30.2	30.9	31.5	30.3	28.2	1981-2010

Apply the *k*-means clustering algorithm to find the homogeneous regions. Before applying the *k*-means clustering algorithm find the optimal value of *k*. Present the results for the optimal value of *k* and comment on the results. You must use your examination roll as a seed point to draw the random sample. Also, find the cluster solutions for Ward linkage clustering algorithm.

Q2. The data set "**Table 2**: Data on Atmospheric Record of Mymensingh (Mymensingh.csv)" is the historical climate record on different atmospheric parameters of the location Mymensingh. Remove the missing values then apply the kNN algorithm to the atmospheric data from the region Mymensingh of Bangladesh to classify the rainfall (RAN) category [No Rain and Trace (NRT), Light Rain (LTR), Moderate and High Rain (MHR)] based on Temperature (TEM), Dew Point Temperature (DPT), Wind Speed (WIS), Humidity (HUM), and Sea Level Pressure (SLP). Find the optimal value of *k*. Use Seventy-five percent observation as training data and the rest of the data as test data. Hence, find the prediction accuracy rate and error rate for test data. You must use your Examination roll as a random seed point to draw the training data.

Table 2: Data on Atmospheric Record of Mymensingh (Mymensingh.csv)

ID	Year	Month	TEM	DPT	WIS	HUM	SLP	RAN
1	1960	1	16.9	11.3	2	73.39	1016	NRT
2	1960	2	21.4	12.6	1.7	66.34	1013	NRT
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
671	2015	11	23.1	18.7	1.7	81.73	1013.7	NRT
672	2015	12	18.3	14.9	1.8	82.68	1015.9	NRT

Q3. Fit an ANN model to predict the rainfall level based on data mentioned in **Q2 [Table 2**: Data o Atmospheric Record of Mymensingh (Mymensingh.csv)**]**. Briefly discuss your result Use your **examination roll number** as a seed point to draw training samples. Obtain the accuracy ra and error rate for training data and test data. Compare your results with the results of Q2.

Q4. Apply CART algorithm to predict the rainfall level based on Temperature (TEM), Dew Poi Temperature (DPT), Wind Speed (WIS), Humidity (HUM), and Sea Level Pressure (SLP) which a mentioned in **Q2 [Table 2**: Data on Atmospheric Record of Mymensingh (Mymensingh.csv)**]**. Brief discuss your results. Use your **examination roll number** as a seed point to draw training sampl Obtain the accuracy rate and error rate for training data and test data. Compare your results with t results of Q2 and Q3.

Q5. Given a data set named "**titanic.raw.rdata**", which contains data on the passengers on the Ship Tita and the record of their survival when the ship wrecked. The attributes in the dataset are:

Class: "1st" "2nd" "3rd" "Crew"

Sex: "Male" "Female"

Age: "Child" "Adult"

Survived: "No" "Yes"

Table 3: titanic.raw.rdata

Class	Sex	Age	Survived
3rd	Male	Child	No
.	.	.	.
3rd	Female	Child	No
.	.	.	.
Crew	Male	Adult	Yes
:	.	.	.
.	.	.	.
Crew	Female	Adult	Yes

The researchers are interested in the association between the attributes for the 96 percent samples of th data. Draw 96 percent of samples randomly from the data set and find all association rules w support=0.50, confidence=0.50. Use your **examination roll number** as a seed point to draw sampl Hence find

(a) the summary of the quality measure- support, confidence, and Lift.

(b) the redundant association rules.

(c) the association rule where right hand side contains "Survived".

(d) the association rule where left hand side contains two attributes right hand side contai "Survived".

(e) Draw the scatter plots for the rules with the help of confidence and support, and lift.

Finally, comment on your findings.

Best of Luck

Department of Statistics
Jahangirnagar University
Part IV B.Sc. (Hons.) Examination 2020
Course Title: Statistical Data Analysis XI
Course No.: STAT LAB - 414 (Group B)

Time: 02 Hours **Full Marks: 14**

Q1. Consider the treatment of patients (620) with endocarditis caused by *Staphylococcus aureus* (SA). 400 people took Standard Antibiotic Treatment and 120 people took New Drug. Among these 400 people, 248 people survived and among 120 people, 17 people died.

a) Construct the contingency table.
b) What are the odds of dying with the new drug as opposed to the standard antibiotic therapy protocol?
c) Calculate Relative Risk.
d) Interpret (b) and (c).

Q2. The data for this question contain remission times of 42 multiple leukemia patients in a clinical trial of a new treatment. The variables in the dataset are given below:
Variable 1: survival time (in weeks)
Variable 2: status (1 = in remission, 0 = relapse)
Variable 3: sex (1= female, 0 = male)
Variable 4: log WBC
Variable 5: Rx status (1 = placebo, 0 = treatment)
The computer results are provided below for several different Cox models that were fit to this dataset.

Model 1:						
Variable	**Coef.**	**Std. Err.**	**p>|z|**	**Haz. Ratio**	**[95% Conf. Interval]**	
Rx	0.894	0.3116	0.622	2.446	0.070	85.812
Sex	-1.012	0.752	0.178	0.363	0.083	1.585
log WBC	1.693	0.441	0.000	5.437	2.292	12.897
Rx × Sex	1.952	0.907	0.031	7.046	1.191	41.702
Rx × log WBC	-0.151	0.531	0.776	0.860	0.304	2.433
	Log likelihood = -69.515					
Model 2:						
Rx	0.405	0.561	0.470	1.500	0.499	4.507
Sex	-1.070	0.725	0.140	0.343	0.083	1.422
log WBC	1.610	0.332	0.000	5.004	2.610	9.592
Rx × Sex	2.013	0.883	0.023	7.483	1.325	42.261
	Log likelihood = -69.555					
Model 3:						
Rx	0.587	0.542	0.279	1.798	0.621	5.202
Sex	-1.073	0.701	0.126	0.342	0.087	1.353
Rx × Sex	1.906	0.815	0.019	6.726	1.362	33.213
	Log likelihood = -83.475					
Model 4:						
Rx	1.391	0.457	0.002	4.018	1.642	9.834
Sex	0.263	0.449	0.558	1.301	0.539	3.139
log WBC	1.594	0.330	0.000	4.922	2.578	9.397
	Log likelihood = -72.109					

a) Evaluate whether you would prefer model 1 or model 2. Explain your answer.
b) Using model 2, give an expression for the hazard ratio for the effect of the Rx variable adjusted for SEX and log WBC.

c) Using your answer in part (ii), compute the hazard ratio for the effect of Rx for males and for females separately.

d) By considering the potential confounding of log WBC, determine which models 2 and 3 you prefer. Explain.

e) Of the models provided which model do you consider to be best? Explain.

Q3. The following data are a sample from the 1967-1980 Evans Country Study. Survival times (in years) are given with 25 participants.

5.8, 2.9, 8.4, 8.3, 9.1, 4.2, 4.1, 1.8, 3.1, 11.4, 2.4, 1.4, 5.9, 1.6, 2.8, 4.9, 3.5, 6.5, 9.9, 3.6, 5.2, 8.8, 7.8, 4.7, 3.9

Compute:

a) $m_j, q_j, R(t_{(j)})$

b) KM survival probabilities

c) Average survival time

d) Average hazard rate

- **Good Luck** -

Department of Statistics
Jahangirnagar University
Part IV B. Sc (Honors) Examination – 2021
Course No.: STAT – 401
Course Title: Statistical Inference II

Time: 4 Hours **Marks: 70**

[Answer any FIVE from the following questions. All questions carry equal marks.]

Q1. (a) What do you mean by Best Critical Region (BCR)? Is there any connection among best critical region, most powerful test and Neymar Pearson Theorem?

(b) Under what circumstances is the Uniformly Most Powerful (UMP) test available? How can you overcome the situation if the UMP test does not exist?

(c) Suppose $X_1, X_2, \dots, X_n$ be a random sample from a Normal Population with mean μ and variance $\sigma^2 = 64$. Find the test with the best critical region, that is, find the Uniformly Most Powerful test with the sample size n = 25 and a significance level α = 0.10 to test the sample null hypothesis $H_0: \mu = 20$ against the composite alternative hypothesis $H_a: \mu > 20$.

Q2. (a) Make a comparative study among the three Likelihood based test:

(i) Likelihood Ratio (LR)
(ii) Wald
(iii) Lagrange Multiplier (LM)

Under what situation the three tests are equivalent? State the importance of gradient, Hessian and Information matrix in developing the tests.

(b) In a cross over trial comparing a new drug to a standard, π denotes the probability that the new one is judged better. It is desired to estimate π and test H_a: π = 0.50 against H_a: π ≠ 0.50. In 20 independent observations, the new drug is better each time.

(i) Find and sketch the likelihood function. Is it close to the quadratic shape that large sample normal approximations utilize?

(ii) Conduct a Wald test and construct a 90% Wald Confidence interval for π. Are these sensible?

(iii) Conduct Score Test. Construct a 95% Score Confidence interval. Interpret.

(iv) Conduct a LR test and construct a likelihood-based confidence interval. Interpret your results.

Q3. (a) What is data reduction technique? What are the available techniques for data reduction? Why sufficient statistic is called a data reduction technique?

(b) What do you mean by complete statistic of a family of density function? Differentiate between complete sufficient statistic and sufficient statistic. Show that a complete sufficient statistic is minimal sufficient statistic.

(c) Let $X_1, X_2, \dots, X_n$ be independent random variable each with P(λ); λ > 0. Find a minimal sufficient statistic for λ and check if it is complete or not.

Q4. **(a)** What are the different lower bounds are used for variance estimators? State the Wolfowitz regularity conditions used in Rao-Cramer lower Bound.

(b) Let $X_1, X_2, \dots, X_n$ be $NIID\ (\mu, \sigma^2)$. Show that the Wolfowitz regularity conditions hold to estimate $\sigma^2, \theta = (0, \alpha)$.

(c) Define Asymptotic efficiency of an unbiased Regular estimator. Let $X_1, X_2, \dots, X_n$ be random sample of size n from $\cdot\ N\ (\mu, \sigma^2)$. Show that $E(X_n)$ is an asymptotic efficient unbiased regular estimator for μ with variance σ^2/n.

Q5. **(a)** Distinguish between Model Unbiased and Median Unbiased Estimators. Are the two estimators unique? Give reason in favor of your answer.

(b) Let $X_1, X_2, \dots, X_n$ be n independent random variables each having Exponential density: $f(x) = {}^1/_\lambda e^{-\frac{1}{\lambda}x}$; $x > 0$

Show that the following two estimators are Modeal Unbiased estimates for λ.

$$y_1 = \frac{X_{(n)}}{\log n},\ X_{(n)} = \max_{i=1,2,\dots,n} X_i, y_2 = \frac{S_{(n)}}{n-1},\ S_{(n)} = \sum_{i=1}^{n} X_i$$

(c) Discuss the steps of estimating parameters by Newton-Rapson method to construct likelihood-based tests.

Q6. **(a)** What is Concentration Ellipsoid? How can it be utilized in linear estimator theory?

(b) Under Usual notations show that the Concentration Ellipsoid of a one-dimensional vector is simply the interval: $-\sigma \leq \varepsilon \leq \sigma$

where σ denotes the standard deviation of x. Also, show that the Concentration Ellipsoid of a two-dimensional random vector can be represented by form of a circle.

(c) Discuss the steps of determining Concentration Ellipsoid of a random vectors where the covariance matrix is singular.

Q7. **(a)** What is Bayesian estimator? What is posterior distribution and posterior Baye's estimator?

(b) What is non-informative prior and conjugate prior? How can you check the existence of conjugate prior?

(c) Discuss different ways of constructing conjugate prior. How can you check the existence of conjugate prior?

(d) If $X \sim (\theta, \sigma^2)$ where σ^2 is known and prior density of θ is $g(\theta) \infty constant$. Then find *mgf* and Bayes estimator of θ using Linux loss function.

Q8. **(a)** What is sequential testing procedure? What is SPRT? Show by an example that sequential sampling and testing procedure need less number of sample observations classical testing procedure to make decision on the hypothesis.

(b) For a SPRT how can you approximately determine k_0 and $k_1$with error sizes α and β.

(c) Let α' and β' be the error sizes of the SPRT defined by $k_0' = \alpha/(1-\beta)$ and $k_1' = (1-\alpha)/\beta$ .Show that $\alpha' + \beta' = \alpha + \beta$.

******Best of Luck******

covariance matrix $S_p = \begin{bmatrix} 0.88 & 0.36 & 0.23 \\ 0.36 & 0.77 & 0.20 \\ 0.23 & 0.20 & 0.55 \end{bmatrix}$.

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

Q4. **(a)** What is Principle Component Analysis (PCA)? Why do we need PCA? Explain the purpose of performing PCA with examples

(b) Show that principal components are uncorrelated with original variables and have the variances equal to the eigenvalue of variance-covariance matrix.

(c) Describe the advantages of using equal correlation structure assumption in PCA. Body weight (in grams) for $n = 150$ female were obtained immediately after the birth of their first four litters and their sample correlation matrix is

$$R = \begin{bmatrix} 1 & .7501 & .6329 & .6363 \\ .7501 & 1 & .6925 & .7386 \\ .6329 & .6925 & 1 & .6625 \\ .6363 & .7386 & .6625 & 1 \end{bmatrix}$$

Test whether the correlation matrix is equal correlation structure or not.

Q5. **(a)** Define an orthogonal factor model with its components and assumptions. How could you check the adequacy of an orthogonal factor model?

(b) What are the factor loadings? Describe two methods for estimation of factor loadings with mentioning their advantages and disadvantages.

(c) The following R output is obtained for conducting the factor analysis with 5 variables and $m = 2$ common factors

```
Call:
factanal(x = x, factors = 2, method = "mle", scale = T, center
= T)
 Uniquenesses:
    V1    V2    V3    V4    V5
 0.497 0.252 0.474 0.610 0.176
 Loadings:                                          Factor1 Factor2
    Factor1 Factor2                 SS loadings       1.671   1.321
 V1 0.601   0.378                   Proportion Var    0.334   0.264
 V2 0.849   0.165                   Cumulative Var    0.334   0.598
 V3 0.643   0.336
 V4 0.365   0.507
 V5 0.207   0.884
 Test of the hypothesis that 2 factors are sufficient.
 The chi square statistic is 0.58 on 1 degree of freedom.
 The p-value is 0.448
```

Find the followings:

(i) Find the matrix of specific variances. Hence, define the most significant variable which fit neatly into this factors model.

(ii) Find the estimated factor loadings and communalities. What proportion of the total population variance is explained by the first common factors?

(iii) Check whether the 2 factors are adequate for this model?

Q6. **(a)** Define canonical correlation analysis. How to check whether a canonical correlation analysis is adequate for your study?

(b) Suppose $z^{(1)}$ and $z^{(2)}$ are the values of the standardized set of variables $\mathbf{Z}^{(1)}$ with p variables and $\mathbf{Z}^{(2)}$ with q variables, have covariances $\mathbf{R}_{11}$ and $\mathbf{R}_{22}$, respectively, and

Q7. (a) What do you mean by analysis of covariance? How does it differ from analysis of variance?

(b) Set up a linear model and discuss the analysis of covariance for the data of Latin square design with two concomitant variables. Display the ANCOVA table.

Q8. (a) What is incomplete block design? Explain the situation when a design is said to be balanced incomplete, randomized incomplete, and symmetric incomplete.

(b) Describe intra and inter block analysis from a balanced incomplete block design? Also prepare ANOVA table for this design.

Best of Luck

Department of Statistics
Jahangirnagar University
Part IV B. Sc (Honors) Examination - 2021
Course Name: Sampling Techniques II
Course No.: Stat – 404

Answer any THREE from the following questions. Each question carries equal marks.

Time: 02 Hours 30 Minutes **Marks: 35**

Q1. (a) Explain the situations where probability proportion to size (PPS) sampling is used in practical cases.

(b) Describe Sen-Midzuno-Lahri method of selecting a sample by PPS sampling without replacement. Prove that (under usual notations) an unbiased estimator of population total is $\hat{Y}=\sum_{i=1}^{n}\frac{y_i}{\pi_i}$ with sampling variance $V\left(\hat{Y}\right)=\sum_{i=1}^{N}\frac{\left(1-\pi_i\right)}{\pi_i}y_i^2+2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{\left(\pi_{ij}-\pi_i\pi_j\right)}{\pi_i\pi_j}y_iy_j$.

Also, give an expression for the unbiased estimator of $V\left(\hat{Y}\right)$.

(c) Find the gain in efficiency due to sampling with unequal probabilities over simple random sampling with replacement (SRSWR).

Q2. (a) What do you mean by two-stage sampling? Why it is so called?

(b) In case of two-stage sampling, if the n first-stage units and the m second-stage units from each chosen first-stage units are selected by simple random sampling without replacement, then show under usual notations that $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$ is an unbiased estimator of the population mean and its variance is given by:

$$\text{var}\left(\bar{y}\right)=\left(\frac{N-n}{N}\right)\frac{S_b^2}{n}+\left(\frac{M-m}{M}\right)\frac{S_w^2}{mn}$$

where, $S_b^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{Y}_i-\bar{Y}\right)^2$, $S_w^2=\frac{1}{N}\sum_{i=1}^{N}S_i^2$ and

$$S_i^2=\frac{1}{M-1}\sum_{j=1}^{M}\left(y_{ij}-\bar{Y}_i\right)^2 \quad ; \quad i=1,2,\ldots,N$$

(c) Consider an appropriate cost function for the above sampling design and find the optimum sample sizes.

Q3. (a) Under what circumstance successive sampling is used? Give examples.

(b) What are the reasons of using sampling over two or more successive occasions?

(c) Suggest a best linear estimator of a population characteristic (e.g. mean) on current occasion. Verify your proposal.

(d) Obtain an optimum size of matched sample on second occasion.

Q4. (a) What are the objectives of using double sampling plan? Explain, in brief.

(b) What is double sampling plan? Define regression estimator of population mean. Find an approximate expression for bias and mean squared error of this estimator when the second phase sample is a subsample.

(c) Find the optimum size of first and second phase samples.

Q5. (a) What do you mean by inverse sampling? Describe the procedure of estimating popu size by direct method and inverse sampling method.

(b) Show that the estimator of population size in each method mentioned in 5(a) is biase

(c) Define randomized response. Describe the Warner's randomized response techniq estimating 'the prevalence of sensitive attributes'.

Best of Luck

$\text{Cov}(\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) = \mathbf{R}_{12}$. Let $\hat{\rho}_1^{*2} \geq \hat{\rho}_2^{*2} \geq \cdots \geq \hat{\rho}_p^{*2}$ be the p ordered eigenvalues of $\mathbf{R}_{11}^{-1/2}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1/2}$ with corresponding eigenvectors $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_p$ and $p \leq q$. Also, let $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \ldots, \hat{\mathbf{f}}_p$ be the eigenvectors of $\mathbf{R}_{22}^{-1/2}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}^{-1/2}$, where each $\hat{\mathbf{f}}_i$ is proportional to $\mathbf{R}_{22}^{-1/2}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\hat{\mathbf{e}}_i$. Then show that the k^{th} pair of canonical variates, $k = 1, 2, \ldots, p$,

$$\hat{\mathbf{U}}_k = \hat{\mathbf{e}}_k'\mathbf{R}_{11}^{-1/2}\mathbf{z}^{(1)} \text{ and } \hat{\mathbf{V}}_k = \hat{\mathbf{f}}_k'\mathbf{R}_{22}^{-1/2}\mathbf{z}^{(2)} \text{ maximizes } \text{Corr}(\hat{\mathbf{U}}_k, \hat{\mathbf{V}}_k) = \hat{\rho}_k^*.$$

(c) Illustrate the steps of interpreting the estimated sample canonical variates. How to evaluate whether the canonical variates are "good" summaries of their respective sets of variables?

Q7. (a) What is discrimination in multivariate analysis? How does it differ from classification? Explain with an example.

(b) Discuss how can you classify two multivariate normal populations with discrimination rule?

(c) The Salmon fishery is a valuable resource for both the USA and Canada. The fish carry information about their birthplace in the growth rings on their scales. Typically the rings associated with freshwater growth are smaller for the Alaskan-born than for the Canadian-born salmon. The following are the summary results of classifications of the diameter of the growth ring reasons, magnified 100 times, where,

X_1=diameter of rings for the 1st-year freshwater growth (hundreds of an inch),

X_2=diameter of rings for the 1st-year marine growth (hundreds of an inch)

Summary of Classification

Put into	Group	
True Group	1	2
1	44	1
2	6	49
Total	50	50

What conclusions will you draw about this classification?

Q8. (a) What do you mean by clustering? Explain the use of clustering in your own word with examples. What are drawbacks of this clustering?

(b) What is hierarchical clustering and non-hierarchical clustering? Write down the steps of hierarchical agglomerative clustering method. Briefly discuss different linkage method considered in hierarchical clustering.

(c) Consider the following matrix of distances

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 11 & 2 & 0 & \\ 5 & 3 & 4 & 0 \end{bmatrix} \end{array}$$

Cluster the four items using average linkage hierarchical procedure. Draw the dendrogram and comment on your results.

Best of Luck

Department of Statistics
Jahangirnagar University
Part IV B. Sc (Honors) Examination - 2021
Course Name: Data Mining
Course No.: Stat – 405

Answer any THREE from the following questions. Each question carries equal marks.

Time: 02 Hours 30 Minutes **Marks: 35**

Q1. **(a)** What is meant by data mining? Briefly discuss the data mining algorithm with its components.

(b) Define the term machine learning. What are the different types of machine learning techniques used in data mining? Explain.

(c) Describe the different steps of Knowledge Discovery in Database.

Q2. **(a)** How can find the optimal number of cluster for k-modes clustering? Explain.

(b) Apply k-prototype clustering algorithm to find the cluster solution for $k = 2$ for the following data on $X_1, X_2, \ldots, X_5$. Use the ID-4 and ID-9 as initial cluster prototypes.

ID	X_1	X_2	X_3	X_4	X_5
1	11	12.5	3.7	F	P
2	10.2	8.3	3.8	M	C
3	11.3	11.9	3.7	F	P
4	10	8.7	3.7	M	C
5	10.7	10	3.8	M	C
6	11.4	13	3.7	F	P
7	11.3	13	4	F	P
8	10.3	8	3.8	M	C
9	11.7	13.2	3.8	F	P
10	10.7	9	3.8	M	C
11	10	7.5	3.8	M	C

(c) Write down the algorithm of Random Forest for Regression or Classification.

Q3. **(a)** Why naïve Bayesian classification is called "naïve"? Briefly outline the major ideas and steps of naïve Bayesian classification?

(b) Apply kNN algorithm to classify the item with information Sepal Length: 5.8, Sepal Width: 3.1, Petal Length: 3.8, and Petal Width: 1.2 based on the following training data for k=3.

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	4.5	1.5	versicolor
5.6	2.7	4.2	1.3	versicolor
6.5	3	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) Explain the following terms Information gain, gain ratio, and Gini index, Tree Pruning, F_β, F_1 score, specificity, precision, recall.

Q4. **(a)** What is an association rule? Define support and confidence for an association rule with example. Discuss the Apriori algorithm of association rule.

(b) Apply *k*-modes clustering algorithm to divide the items into *k=3* clusters. You should use the items (ID-1), (ID-7) and (ID-8) for the initial clusters.

ID	1	2	3	4	5	6	7	8
X_1	A	C	B	D	C	D	B	D
X_2	M	Y	N	Z	M	Y	N	Z
X_3	R	W	W	W	R	W	R	R

(c) Using suitable example define clustering. Discuss the use of distance for measuring similarities and dissimilarities in clustering algorithm.

Q5. **(a)** What do you mean by web mining? Briefly explain the different data mining techniques with example?

(b) The following contingency table summarizes supermarket transaction data, where coffee refers to the transactions containing coffee, $\overline{coffee}$ refers to the transactions that do not contain coffee, milk refers to the transactions containing milk, and $\overline{milk}$ refers to the transactions that do not contain milk.

	milk	$\overline{milk}$	$\sum_{row}$
coffee	1000	100	1100
$\overline{coffee}$	10000	100000	110000
$\sum_{column}$	11000	100100	111100

(i) Based on the given data, is the purchase of coffee independent of the purchase of milk? If not, what kind of correlation relationship exists between the two?

(ii) Find the value of *all confidence, max confidence, Kulczynski*, and *cosine* measures with *lift, correlation, and imbalance ratio* for the given data. Comment on your result.

(c) Define the different types of Activation function. Write down the advantage and disadvantage of Neural Networks for classification.

Best of Luck

Department of Statistics
Jahangirnagar University
Part IV B. Sc (Honors) Examination - 2021
Course Name: Epidemiology and Biostatistics
Course No.: Stat – 406

Answer any FIVE from the following questions. Each question carries equal marks.

Time: 04 Hours **Marks: 70**

Q1. (a) Define Epidemiology. Discuss the different components of epidemiology.

(b) What are the aims and objectives of Epidemiology? Explain the scopes of Epidemiology.

(c) Define communicable and non-communicable disease. Explain Epidemiological triad with an example.

Q2. (a) What do you mean by epidemiological Study designs? What is case report? How does it differ from Case Series?

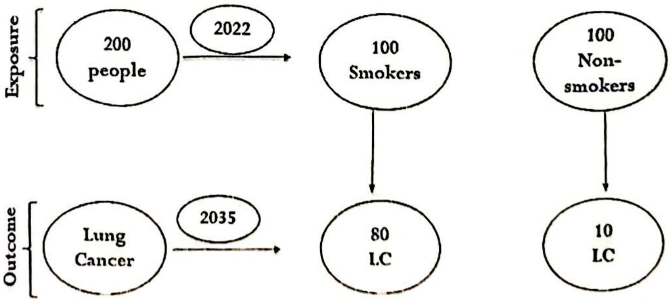
(b) What do you mean by Cohort Study? Describe the design of Cohort of Study. What are the advantages of this study?

(c) Data from a hypothetical Cohort Study relating birth weight with mortality at 5 years given below. Calculate the relative risk of mortality associated with low birth weight. Include a 95% confidence interval and interpret the findings.

Birth Weight	Mortality	
	Yes	No
≤ 2500	108	163
> 2500	117	268

Q3. (a) What do you mean by exposure of a disease? Mention different measures which are used to measure the disease frequency. What does incidence and prevalence of a disease represent?

(b) Define the following measures: (i) Odds Ratio, (ii) Relative Risk, (iii) Population Attributable Risk, (iv) Attributable Risk. Calculate each measures who smoke and who have never smoked from the following table and interpret each results:



Q4. (a) What is prevention? Explain different levels of prevention.

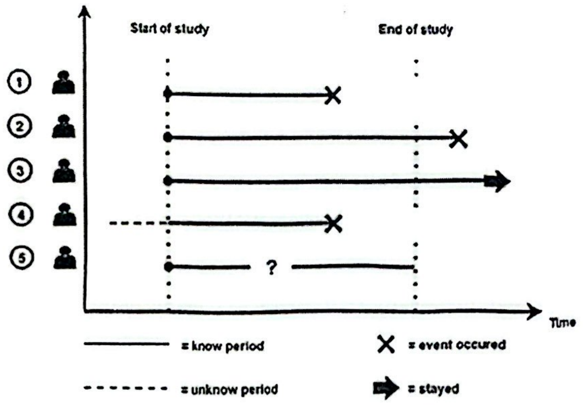
(b) How can you determine the presence or absence of a disease? Mention some tests which are common to diagnose a disease. What are the different types of such tests?

(c) Define sensitivity, specificity, predictive value positive, predictive value negative, and yield of the test. Determine each measures from the following table:

Test Results	True Characteristics in Population	
	Disease	No Disease
Positive	80	100
Negative	20	800

Q5. **(a)** Explain survival analysis with its terminologies and mention some examples. What are the goals of survival analysis?

(b) What is censored data? How can we work with censored data? Explain different types of censoring from the following example:



(c) Define survivor function and hazard function with their properties. Show some graphs of different hazard functions.

Q6. **(a)** Explain Kaplan-Meier Survival curve with its general features. Given the following the survival time data (in years) for 25 participants: 11.7, 10.0, 5.7, 9.8, 5.3, 3.5, 9.2, 2.5, 8.7, 9.2, 12.1+, 6.6, 2.2, 1.8, 10.2, 3.8, 3.0, 12.3 +, 5.4, 8.2, 12.2+, 2.6, 11.0, 10.7, 11.1.

Compute: (i) Risk Set, (ii) Survival Probability, (iii) Average survival time, (iv) Average hazard rate.

(b) Explain the Log-Rank Test for two groups. What are the alternative tests we can use to the log rank test?

(c) What do you conclude about whether or not the three survival curves are the same?

Group	Events Observed	Events Expected
Group 1:	50	26.30
Group 2:	47	55.17
Group 3:	31	46.53

Log-rank = chi2(2) = 29.18
P-value = Pr > chi2 = 0.0000
G = 3 groups; df = 2

Q7. **(a)** Define Cox-proportional hazard model. How do you fit this model?

(b) Define Logistic regression model. Explain with example, how the logistic regression model has become a popular tool for the analysis of Case-control studies.

(c) What is Cox-regression model? What are the purposes of this model?

Q8. **(a)** What do you mean by incidence? What are the main variants of incidence? Describe cumulative incidence and incidence density.

(b) What do you mean by Prevalence? Distinguish between Prevalence and incidence. Develop a $100(1-\alpha)\%$ confidence interval for prevalence rate.

(c) Define relative risk (RR). Interpret if $RR = 1$, $RR < 1$, and $RR > 1$. How do you distinguish RR from odds ratio (OR)?

Best of Luck

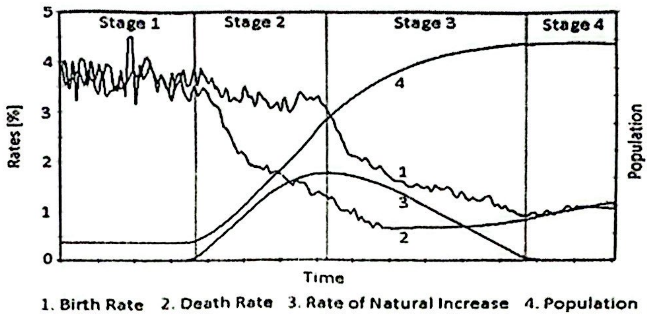
Department of Statistics
Jahangirnagar University
Part IV B. Sc (Honors) Examination - 2021
Course Name: Advanced Demography
Course No.: Stat – 407

Answer any THREE from the following questions. Each question carries equal marks.

Time: 02 Hours 30 Minutes **Marks: 35**

Q1. **(a)** What can you know from the demographic transition theory? Write down the recent demographic variables of Bangladesh and explain in which stage Bangladesh lies?

(b) Explain the different stages of demographic transition theory of below figure.



(c) Based on this figure (b), which stage is better and why? Also identify the problems of other stages.

Q2. **(a)** What is the basic difference between mathematical methods and cohort's component methods of population projection? Which one is best in respect of Bangladesh and why?

(b) Draw the curves of linear function, geometric function, exponential function and binary logistic function and explain them.

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

Year	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Population (Crore)	15.2	15.4	15.6	15.8	16.0	16.2	16.4	16.5	16.7	17.0
Growth Rate (%)	1.25	1.28	1.25	1.20	1.24	1.26	1.17	1.12	1.15	1.16

Q3. **(a)** What is the relationship between proximate determinants of variables and total fertility rate? Why proximate determinants of variables reduced to four variables, explain it with basic concept.

(b) If $TFR = 0.3$, $TMFR = 4.5$, $u = 0.59$, $e = 0.90$, $i = 0.60$, and $TA = 0.20$; comments of these values with the definition of abbreviated forms and estimate four proximate determinants of variables using these values.

(c) Also show impact on fertility reduction and identify the variable which is responsible more to reduce fertility.

Q4. **(a)** What do you mean by urbanization? Discuss the different process of urbanization.

(b) What is tempo of urbanization? Discuss the tempo of urbanization measurement if the annual average rate of change in percent population living in urban areas is assumed to be change arithmetically.

Q5. **(a)** What is migration? What are the important causes of migration? Define pull factor and push factor. What are the positive and negative consequences of migration?

(b) What is city size distribution? Estimate the value of Z and give an interpretation by considering a hypothetical example where Z is the constant of the relationship between the rank of any city and the size of the largest city. Under usual notations show that, $W = r_u - r$.

Best of Luck

Department of Statistics
Jahangirnagar University
Part IV B. Sc (Honors) Examination - 2021
Course Name: Stochastic Process
Course No.: Stat – 408

Answer any THREE from the following questions. Each question carries equal marks.

Time: 02 Hours 30 Minutes **Marks: 35**

Q1. **(a)** Define Random variable and Stochastic Process. Categorized Stochastic Process based on Time and Space with appropriate example.

(b) When a process is said to be evolutionary? Suppose $[X(t), t \in T]$ be a stochastic process where $\Pr[X(t)=n]=\frac{e^{-at}(at)^n}{n!}$; $n=0,1,2,\ \dots\ , a>0$. Is this process evolutionary?

(c) Let $\{X_n\}$ be a Markov chain with the state space $\{1,2,3\}$ and transition probability matrix

$$P = \begin{bmatrix} 0 & 0.4 & 0.6 \\ 0.25 & 0.75 & 0 \\ 0.4 & 0 & 0.6 \end{bmatrix}$$

Let the initial distribution be- $p(0)=[p_1(0), p_2(0), p_3(0)] = [0.4, 0.2, 0.4]$

Calculate the following probabilities:

(i) $P[X_1=2, X_3=2,\ X_3=1 | X_0=1]$

(ii) $P[X_1=2, X_3=2,\ X_3=1]$

(iii) $P[X_1=2, X_4=2,\ X_6=2]$

Q2. **(a)** Define Markov Chain. How can you use Chapman-Kolmogorov equation to compute transition probabilities?

(b) Discuss absorbing state, class of states, communicate states, recurrent and transient state of a Markov Chain with example. How would you calculate the mean recurrent time?

(c) Consider the Markov Chain shown in Figure 1.

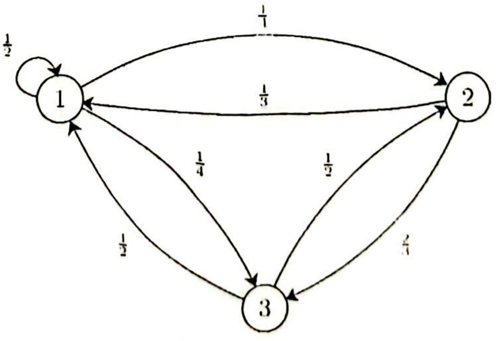


Figure 1. A State Transition Graph

(i) Is this chain irreducible?

(ii) Is this chain aperiodic?

(iii) Find the stationary distribution for this chain.

(iv) Is the stationary distribution a limiting distribution for the chain?

(v) Construct the Transition Probability Matrix.

Q3. **(a)** Write shortly on Counting process and Branching Process. How branching process is associated with stochastic process? What are its assumptions?

(b) If $\{N(t), t \geq 0\}$ is Poisson Process then prove that the autocorrelation coefficient between $N(t)$ and $N(t+s)$ is $\sqrt{\frac{t}{t+s}}$

(c) Define Compound Poisson process with real life example. Find the mean and variance of Compound Poisson process. Suppose that families migrate to an area at a Poisson rate of 2 per month and the number of people in each family is independent and takes on the values 1, 2, 3, 4 with respective probabilities, 0.4, 0.2, 0.2, 0.2 and 0.

(i) Estimate the expected number of individuals migrated to this area during one year?

(ii) What is the probability that more than 60 families migrated to the area within in last year?

Q4. **(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 9 and die at an exponential rate 9. Also there is no increase in the population due to immigration. Explain the situation when the population size is 100. Also find the expected population size after 10 years.

(c) Define pure birth process with an example. For a pure birth process with rate λ, under usual notations show that $P_{ii}(t) = e^{-\lambda t}$.

Q5. **(a)** Define *M/M/1* queuing model. Also derive the distribution of the *M/M/1* queuing model having infinite capacity.

(b) For an *M/M/1* queuing model with infinite capacity, find the average number of customers waiting in the queue.

(c) Suppose customers visit the website of an online store at random at an average rate of 7 per minute and each customer keep the system busy for an average of $^1/_{10}$ minutes. Find:

i). Server utilization rate for the online store and make your comments.

ii). Average number of customers visiting the store.

iii). What is the probability that a customer has to wait more than 3 minutes to get the desired service?

A=T
G≡C

Best of Luck

Department of Statistics
Jahangirnagar University
Part IV B. Sc (Honors) Examination - 2021
Course Name: Bioinformatics
Course No.: Stat – 409

Answer any THREE from the following questions. Each question carries equal marks.

Time: 02 Hours 30 Minutes **Marks: 35**

Q1. **(a)** How would you distinguish among chromosome, gane, and DNA? Explain the structure of a DNA segment. Also, discuss different properties of DNA.

(b) Why gene is called the basic unit of genetics? Under which conditions the nature of genetic information depends? How does two DNA segments differ? Explain with suitable example.

(c) Define the following terms: (i) the Central Dogma of life, (ii) crossing over, (iii) recombination between two chromosomal segments, (iv) recombination function, and (v) source of genetic variation.

Q2. **(a)** What is meant by a genetic marker? Mention its properties. How can you check the quality of the genetic markers for your analysis? Explain. Also, distinguish between genotyping and imputation in the context of a genetic association study.

(b) Distinguish between polymorphism and monomorphism of a genetic marker with an example. Suppose you have two SNPs with the following information: SNP1: major allele frequency=0.65 and SNP2: minor allele frequency=0.15, calculate the strength of heterogeneity for each SNP and hence comment.

(c) What is linkage disequilibrium (LD)? How would you measure it? What are the advantages if two SNPs are in LD? Explain with an example. Consider a diallelic marker with a major allele frequency $f(A)=0.55$. What would be the genotypes frequencies if the marker is in HWE?

Q3. **(a)** What is meant by the genetic association? Write down the importance of Genome-Wide Association Studies (GWAS) in the field of bioinformatics. Distinguish between: (i) orthologous gene and paralogous gene, (ii) local alignment and global alignment.

(b) Mention different genotyping methods known to you. Make a comparative analysis between HapMap and 1000 Genomes projects.

(c) Which genetic model is usually used in GWAS and why? How an association is tested in GWAS? Your explanation should include the ways of data handling, tabular presentation and the testing procedure.

Q4. **(a)** Data bank stores all gens and protein related information. Among them NCBI, EMBL, DDBJ, and PDB are well known Data bank.

(i) What are the purposes of establishing a Data Bank?

(ii) Write on each of the data bank above describing their primary functions.

(iii) Why researchers need to use BLAST?

(b) What is SWISS-PORT? How it can help in our research in Molecular Biology discipline.

(c) Suppose you have an amino acid sequence that represents a protein. What are the different types of analysis available now-a-days that may be apply on this protein data?

Q5. (a) What is multiple sequence alignment? Write an application of multiple sequence alignment.

(b) In your own words, write the process of center star method for multiple sequence alignment.

(c) Write the process of Clustalw method of MSI?

Best of Luck

Department of Statistics

Jahangirnagar University

Part IV B. Sc (Honors) Examination - 2021

Course No.: Stat-410

Course Name: Categorical Data Analysis

Time: 04 Hours Marks: 70

Answer any FIVE questions. All questions carry equal marks.

1. (a) What are the different types of response variables used in developing regression-type model? Give an overview of likelihood ratio, Wald and Score tests for testing parameters $H_0 : \pi = 0.3$ of the Binomial distribution of categorical responses.

(b) Consider a group of Americans who were classified according to their Race and their opinion about afterlife.

	Belief in Afterlife	
Gender	Yes	No or Undecided
Male	630	200
Female	70	50

(i). Construct a 90% confidence interval for the corresponding true relative risk and interpret the results.

(ii). Construct a 90% confidence interval for the odds ratio, and interpret.

(iii). Test whether one sex is more likely than the other to believe in afterlife, (1) using χ^2 test statistic, (2) using likelihood ratio test statistic under usual notations. Interpret your results

2. (a) What is meant by Generalized Linear Models (GLM)? Discuss the different components of a GLM. Describe the parameter estimation of GLM.

(b) What is a count response? Give an example. Identify the natural parameter and canonical link of Poisson responses and hence derive the Poisson log-linear model. What do you mean by over dispersion of a Poisson GLM?

(c) Explain the Deviance and Pearson Goodness-of-Fit Statistics for GLM. What are the importance of the Goodness-of-fit statistics?

3. (a) What is count response? What do you mean by over dispersion? Discuss it in terms of Poisson distribution.

(b) Describe Fisher's exact test for two-way contingency table.

(c) The results of a study comparing radiation therapy with surgery in treating cancer of the larynx is given below:

	Cancer Controlled	Cancer not Controlled
Surgery	29	2
Radiation Therapy	15	5

(i). Find and interpret the P-value for Fisher's exact test with (a) $H_a : \theta > 1$, and (b) $H_a : \theta \neq 1$.

(ii). Obtain and interpret the mid P-value for $H_a : \theta \neq 1$ and find the corresponding confidence interval based on mid P-values.

4. (a) What do you mean by multi-category logit model? Derive Baseline-category logit model for nominal responses.

(b) Obtain the estimate of response probabilities for the baseline category logit model. Derive cumulative logit model for ordinal responses.

Question 4 continues on the next page. .

(c) A model fit predicting preference for President in the US (Democrat, Republican, Independent) using annual income (in \$10,000 dollars) is $log\left(\frac{\hat{\pi_D}}{\hat{\pi_1}}\right) = 3.1 - 0.3x$ and $log\left(\frac{\hat{\pi_R}}{\hat{\pi_1}}\right) = 1.0 + 0.2x$.

i). State the prediction equation for $log\left(\frac{\hat{\pi_R}}{\hat{\pi_D}}\right)$. Interpret its slope.

ii). Find the range of x for which $\hat{\pi_R} > \hat{\pi_D}$.

iii). State the prediction equation for $\hat{\pi_1}$.

5. For baseball national league games during nine decades, the following table shows the percentage of times that the starting pitcher pitched a complete game.

Decade	Percent Complete	Decade	Percent Complete	Decade	Percent Complete
1900-1909	72.7	1930-1939	44.3	1960-1969	27.2
1910-1919	63.4	1940-1949	41.6	1970-1979	22.5
1920-1929	50	1950-1959	32.8	1980-1989	13.3

(a) Treating the number of games as the same in each decade, the maximum likelihood (ML) fit of the linear probability model is $\hat{\pi} = 0.7578 - 0.0695x$, where x = decade ($x = 1, 2, \ldots, 9$). Interpret 0.7578 and -0.0694.

(b) Substituting $x = 10, 11, 12$, predict the percentages of the complete games for the next three decades. Are these predictions plausible? Why?

(c) The ML fit with logistic regression is

$$\hat{\pi} = \frac{\exp(1.148 - 0.315x)}{1 + \exp(1.148 - 0.315x)}$$

Obtain $\hat{\pi}$ for $x = 10, 11, 12$. Are these more plausible?

6. (a) What is meant by matched pairs? Describe a suitable test for testing marginal homogeneity. Hence Derive the testing procedures.

(b) Define Cohen's Kappa coefficient. Explain how it measures the strength of agreement. Write down its disadvantages.

(c) A social survey asked subjects whether they believed in heaven and whether they believed in hell. The results are shown in the following table:

Believe	Believe in Hell	
in Heaven	Yes	No
Yes	855	150
No	2	180

i). Test the hypothesis that the population proportions answering yes were identical for heaven and hell.

ii). Find 90% confidence interval for the difference between the population proportions. Interpret.

iii). Estimate and interpret the odds ratio for a logistic model for the probability of a yes response as a function of the item (heaven or hell), using the marginal model.

7. (a) Define log-linear model. How to interpret the parameters of the log-linear model and how they can be used to explain joint and conditional associations among variables? In what logistic situation model and log-linear model is appropriate in categorical data analysis?

(b) The following table is taken from a report on the relationship between aspirin use (X) and myocardial infarction (Y) by the Physicians Health Study Research Group at Harvard Medical School. The Physicians Health Study was a five-year randomized study testing whether regular intake of aspirin reduces mortality from cardiovascular disease. Every other day, physicians participating in the study took either one aspirin tablet or a placebo. The study was blind - the physicians in the study did not know which type of pill they were taking.

	Myocardial Infarction	
Group	Yes	No
Placebo	189	10845
Aspirin	104	10933

(i). Fit the independence model and check the goodness of fit. Report $\{\widehat{\lambda_j^Y}\}$. Interpret the results of $\{\widehat{\lambda_1^T} - \widehat{\lambda_2^T}\}$.

(ii). For the saturated model report $\widehat{\lambda_{ij}^{XY}}$. Show how to interpret these estimates using an odds ratio.

8. (a) What is multilevel modeling? What are the benefits of multilevel modeling analysis? Distinguish between the marginal model and the conditional model with an example.

(b) Define GEE models. When is it applicable? How to fit the GEE models? Discuss the different working correlation structures of GEE models.

(c) What is meant by GLMM? Why is it important for cluster data? Formulate GLMM for the Binary data. Distinguish between GEE and mixed effect models.

Good Luck