Jahangirnagar University
Department of Statistics and Data Science
Part IV B.Sc. (Honors) Examination – 2024
Course No.: Stat – 401
Course Name: Statistical Inference I
1st Tutorial -10.08.2025

Time: 60 minutes **Marks: 10**

N.B. Answer all the following questions.

1. Why do you need to apply to resampling? Explain it.
2. Consider the following set of observations {4, 5, 6, 7, 8}. Calculate the Jackknife mean. Also show that Jackknife mean is an unbiased estimate of the population mean.
3. Define the Cramer-Rao Inequality.
4. How can you find the Cramer-Rao Inequality from Bhattacharyya lower bound?

Good Luck!

STAT-402

1st Tutorial

Time: 45 Minutes | Marks: 10

Answer one from question 1 and 2

1. Consider the example of Stroke study: Monitoring safety of a thrombolytic drug administered for ischemic stroke. Suppose now that the ECASS 3 DSMB is asked to examine the interim results of rt-PA treated stroke patients with an emphasis on SICH and in the light of the results of the ECASS 2 study. For ease of exposition, we assume that in the ECASS 2 study 100 patients were treated with rt-PA and 8 patients suffered from SICH. Assume also that at the first interim analysis of the ECASS 3 study, 50 patients were treated with rt-PA and that SICH occurred in 10 patients. The DSMB wishes to obtain an accurate picture of the risk for SICH. 3+3

   a) Find the posterior distribution of θ (probability of suffering from symptomatic intercerebral hemorrhage (SICH)) after observing the data from the first ECASS 3 interim analysis.

   b) If the posterior distribution of θ is a Beta(19, 133)-distribution, calculate the posterior Mean, mode, and variance.

Or

2. a) Derive posterior distribution when sample data comes from binomial distribution. 3+3

   b) Also find the expression for mean, mode, variance from posterior distribution.

Answer any one from question 3 and 4.

3. Define two types of credible interval. Compare and interpret confidence interval and credible interval with a hypothetical example. 2+2

Or

4. Describe the use of credible interval and contour probability in Bayesian hypothesis testing. Compare Bayesian hypothesis testing with frequentist approach with example. 2+2

ECASS- European Cooperative Acute Stroke Study, DSMB - Data and Safety Monitoring Board (DSME)

Sept 02, 202

STAT-402 (2nd Tutorial)

Bayesian Inference (40 Minutes)

1.	(i) Derive the joint posterior of the normal likelihood with both the mean and variance parameters unknown. (ii) Also obtain the conditional posterior for the mean and variance. (iii) obtain the marginal distribution for the parameters. (iv) Find the the posterior predictive distribution of a new observation. (v)show that Joint posterior can be written as multiplication of marginal with conditional posterior. Make necessary assumptions and choose suitable prior for the parameters. Hintz: $y = \{y_1, y_2, \ldots, y_n\} \sim$ i.i.d. from $N(\mu, \sigma^2)$. $p(\mu \mid \sigma^2, y)$ is $N(y, \sigma^2/n)$, $p(\mu \mid y)$ is a $t_{n-1}(y, s^2/n)$ $p(\sigma^2 \mid y)$ is a Inv $-\chi^2(n-1, s^2)$ $p(\mu, \sigma^2 \mid y) = p(\mu \mid \sigma^2, y) \times p(\sigma^2 \mid y) = N(y, \sigma^2/n) \times$ Inv-$\chi^2(n-1, s^2)$	5
2.	Consider that you are given a coin, which is either fair or loaded (come up heads 70% of the time). Suppose you have a prior idea that it has a 60% probability that the coin is loaded. You flip it five times, get two heads and three tails and now figure out what's your posterior probability that this is the loaded coin. **(i)**Show the steps of Metropolis-Hastings algorithm to estimate posterior probabilities. **(ii)**Draw the diagram for this Markov chain with two states and show that this method gives us the theoretical right answer , i.e.,$\pi_i P = \pi_i$, where, P is the transition probability of the Markov chain with stationary distribution π_i.	4
3.	short notes, draw diagram if applicable: (i) Conditional conjugate prior (ii) Jeffry's prior (iii) Traceplot (iv) Burn in period.	1

Department of Statistics and Data Science, Jahangirnagar University
Part IV B.Sc. (Hons.) 1st Tutorial Examination 2024
Course Title: Multivariate Analysis, Course No. STAT-403

Time: 60 minutes **Full marks: 20**

3. Given the random vector $\mathbf{X}=\begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}$ be distributed as $N_p(\boldsymbol{\mu},\Sigma)$ with mean vector $\boldsymbol{\mu}=\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}$ and covariance matrix $\Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$, and $|\Sigma_{22}|>0$. Then find the conditional distribution $\mathbf{X}_1$ given that $\mathbf{X}_2=\mathbf{x}_2$.

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

- **If modulus(Last_two_digits of your Class Roll No, 3) = 0:**

  Partition $\mathbf{X}$ as $\mathbf{X}=\begin{bmatrix}X_1\\ X_2\\ X_3\end{bmatrix}=\begin{bmatrix}\mathbf{X}^{(1)}\\ \mathbf{X}^{(2)}\end{bmatrix}$. Find the distribution of $\mathbf{X}^{(1)}$ given $\mathbf{X}^{(2)}$.

- **If modulus(Last_two_digits of your Class Roll No, 3) = 1:**

  Find the conditional distribution of $\frac{X_1+3X_3}{2}$ given $\mathbf{X}_2$.

- **If modulus(Last_two_digits of your Class Roll No, 3) = 2:**

  Find the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_1-2\mathbf{X}_2+\mathbf{X}_3$.

Department of Statistics and Data Science, Jahangirnagar University
Part IV B.Sc. (Hons.) 2nd Tutorial Examination 2024
Course Title: Multivariate Analysis
Course No. STAT-403

Time: 60 minutes **Full marks: 20**

1. Let $X_1, X_2, \ldots, X_n$ be a random sample from a $N_p(\mu, \Sigma)$, ~~show that T^2 can be expressed~~ as a ~~function of Wilk's Lamda~~. Also, express the sampling distribution of T^2.
2. Provide an example of how to conduct a profile analysis to determine whether the profiles are ***level*** (students with **odd Class Roll numbers**).
   Or, Provide an example of how to conduct a profile analysis to determine whether the profiles are ***coincident*** (students with **even Class Roll numbers**).

Department of Statistics and Data Science, Jahangirnagar University

3rd Tutorial Examination 2024

Course Title: Multivariate Analysis, Course No.: STAT-403

Time: 1 hours **Full marks: 20**

1. The mtcars dataset in R, which contains data on 32 car models and 11 variables such as

mpg	Miles/(US) gallon (fuel consumption)
cyl	Number of cylinders
disp	Displacement (cubic inches)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	1/4 mile time (seconds)
vs	Engine shape (0 = V-shaped, 1 = straight)
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

The following are the results of Principal Component Analysis (PCA) in R

```
Importance of components:
                          PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9   PC10  PC11
Standard deviation     2.378 1.370 0.994 0.838 0.691 0.597 0.511 0.442 0.373 0.287 0.215
Proportion of Variance 0.514 0.170 0.089 0.063 0.043 0.032 0.023 0.017 0.012 0.007 0.004
Cumulative Proportion  0.514 0.684 0.774 0.838 0.881 0.914 0.938 0.955 0.968 0.976 1.00

Loadings (first 7 components):

        PC1     PC2     PC3     PC4     PC5     PC6     PC7
mpg  -0.3625  0.0161 -0.2257 -0.0225 -0.1028 -0.1087  0.3677
cyl   0.3739  0.0437 -0.1753 -0.0025 -0.0584  0.1685  0.0572
disp  0.3681 -0.0493 -0.0614  0.2566 -0.3939 -0.3361  0.2143
hp    0.3300  0.2487  0.1400 -0.0676 -0.5400  0.0714 -0.0014
drat -0.2941  0.2746  0.1611  0.8548 -0.0773  0.2444  0.0211
wt    0.3461 -0.1430  0.3418  0.2458  0.0750 -0.4649 -0.0206
qsec -0.2004 -0.4633  0.4031  0.0680  0.1646 -0.3304  0.0500
vs   -0.3065 -0.2316  0.4288 -0.2148 -0.5995  0.1940 -0.2657
am   -0.2349  0.4294 -0.2057 -0.0304 -0.0897 -0.5708 -0.5873
gear -0.2069  0.4623  0.2857 -0.2646 -0.0483 -0.2435  0.6050
carb  0.2140  0.4135  0.5285 -0.1267  0.3613  0.1835 -0.1746
```

(a) Draw the scree plot and decide how many principal components to retain in this study?

(b) List all the adequate principal components for this study.

(c) If the variances of the original variables are:

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
0.3632	3.1895	15360.799	4700.866	0.2858	0.9573	3.1931	0.2540	0.2489	0.5443	2.6088

Find and interpret the maximum correlation between the 1st principal components and the original variable.

2. Use the Bartlett test to find the number of significant canonical correlations for the following studies

(a) with $n = 456$, $p = 4$, $q = 5$ $\hat{\rho}_1^* = 0.67$, $\hat{\rho}_2^* = 0.23$, $\hat{\rho}_3^* = 0.07$, and $\hat{\rho}_4^* = 0.03$.

(b) with $n = 744$, $p = 6$, $q = 4$ $\hat{\rho}_1^* = 0.46$, $\hat{\rho}_2^* = 0.38$, $\hat{\rho}_3^* = 0.12$, and $\hat{\rho}_4^* = 0.04$.

3. (a) Define an orthogonal factor model, detailing its components and key assumptions.

(b) Table 1 gives the data on five different kinds of overtime hours for the Madison-Wisconsin police department. Each observation represents a total for 12 pay periods, or about half a year.

Table 1: Five types of overtime hours for the Madison-Wisconsin Police Department

Legal Appearances Hours (x_1)	Extraordinary Event Hours (x_2)	Holdover Hours (x_3)	Compensatory overtime allowed (x_4)	Meeting Hours (x_5)
3387	2200	1181	14861	236
...	...	...	...	...

If the Madison-Wisconsin police department's overtime data does not follow multivariate normality, then which method we can use to estimate the factor loadings for this study? Why?

1st Tutorial STAT-404

Q1. (a) Explain Yates algorithm to calculate different sum squares of components in a 2^3 factorial experiment on 6 graphs of each treatment combination, the graphs of which are internally homogeneous and externally heterogeneous. In addition, present the ANOVA table and test whether the treatment combinations are statistically similar to each other. 6

(b) Construct the Yates table to calculate different sum squares of components in a factorial experiment 3^2 with 4 replications of each treatment. Also, present the ANOVA table. 6

(b) An experiment was conducted to evaluate the effect of zinc (Z) and potassium (K) fertilizers on the yield of a certain variety of maize. Both Z and K were applied at 2 levels (0 and 1). 8

$$N = \begin{cases} 0 & \text{for low level} \\ 1 & \text{for High level} \end{cases} \qquad Z = \begin{cases} 0 & \text{for low level} \\ 1 & \text{for High level} \end{cases}$$

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

Replications			
I		II	
00	10	10	00
24	32	28	24
01	11	01	11
46	30	36	36

i. Write the mathematical model for this data. Justify your answer.

ii. Estimate the components of the ANOVA table. Construct the ANOVA table.

iii. Test whether the treatment combinations are similar.

STAT-404 [2nd Tutorial]

Q1. (a) Define asymmetrical factorial design. Estimate the yield as well as the effect for 2×3 factorial design. 5

(b) Obtain sum squares of different factors of 2×3 factorial design. Construct the ANOVA table. 5

Q2. (a) Define Split plot design. Discuss how does it differ from confounded design. 5

(b) Describe different steps of analyzing the data of a split plot design. Estimate the ANOVA table. 5

Q3. (a) What is confounding and what is its necessity? Discuss different types of confounding with examples. 5

(b) Discuss the block consisting of 2^4-factorial experiment if ABCD and AD interactions are simultaneously confounded in the same replication. Discuss the procedure of analysis of data to construct ANOVA table. 5

STAT-404 (3rd Tutorial)

Q1. (a) *At a horticultural research center, an experiment is conducted to study the productivity of two varieties of tomato using phosphate fertilizer. The agricultural plots for cultivation are found to be homogeneous concerning soil moisture. The tomato varieties are randomly allocated to different plots. However, the amount of fertilizer applied (z kg/plot) varies across the plots. The production of tomatoes (y kg/plot) of different plots, along with the amount of fertilizer used, is given below:* 8

Plots	Tomato 1		Tomato 2	
	y	z	y	z
1	62	3	70	5
2	65	5	72	4
3	60	4	68	6

i. *Write down the appropriate model for this data. Justify the reason of your choice.*

ii. *Complete the ANCOVA table.*

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

(b) When balanced incomplete block design (BIBD) obtained? For a BIBD with usual parameters, show that: 6

i. $bk = vr$

ii. $b \geq v + r - k$

(c) Construct a layout plan for a BIB design having parameters $v = b = 12, r = k =$ 6

Set B1

Department of Statistics and Data Science
Jahangirnagar University
Part IV B. Sc (Honors) First Tutorial Examination – 2024
Course Code: Stat- 405
Course Title: Data Mining

Marks: 20

Time: 01 Hour

Answer the following questions. Each question carries equal marks.

Q1. Find the mean, μ, for the data that follows the normal distribution where the known data are {1.5, 2, 8, 10, 9} with two missing items. Here $n = 7$ and $k = 5$. Suppose the initial guess value $\mu_0 = [X_0 \bmod 3]$, where X_0 = your class roll. Consider two decimal places for detailed calculation. **[5]**

Q2. The inputs of ANN be $x_1 = 0.5$, $x_2 = 0.9$, $x_3 = 0.3$ at node 1, node 2, and node 3 respectively; the weights between input nodes to two nodes (node 4 and node 5) of the hidden layer are $w_{14}^{(h)} = 0.3, w_{15}^{(h)} = 0.2$, $w_{24}^{(h)} = 0.5$, $w_{25}^{(h)} = 0.5, w_{34}^{(h)} = 0.7, w_{35}^{(h)} = 0.4$; and the weights between nodes of the hidden layer to the node (node 6) of the output layer are $w_{46}^{(o)} = 0.3, w_{56}^{(o)} = 0.5$. Also, the biases are $w_{04}^{(b)} = 0.3, w_{05}^{(b)} = 0.2$ at hidden layer; and $w_{06}^{(b)} = 0.4$ at output layer. The learning rate and target are 0.9 and 1.0 respectively. Find the updated weights after the first iteration. Consider three decimal places for detailed calculation. **[5]**

Q3. Apply Self-Organizing Map (SOM) to cluster the A, B, C, and D data points for an iteration. Assume that the initial learning rate is 0.8 and the number of clusters to be formed is 2. Also, obtain the learning rate after the first iteration. Consider three decimal places for detailed calculation. **[5]**

Table 1: Data Point for SOM

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

Consider the initial weight matrix

$$W = \begin{bmatrix} 0.2 & 0.7 \\ 0.3 & 0.6 \\ 0.4 & 0.4 \\ 0.5 & 0.9 \end{bmatrix}$$

Q4. Consider the distance matrix from a data set and the cluster solution. Compute the Dunn index for $k=2$ and $k=3$ to find the optimal number of clusters. Comment on your results. **[5]**

	A	B	C	D	E
A	0				
B	60	0			
C	69	34	0		
D	194	153	144	0	
E	24	56	64	188	0

$k=3$		
Cluster	No of Items	Items
Cluster 1	2	{A, E}
Cluster 2	2	{B, C}
Cluster 3	1	{D}

$k=4$		
Cluster	No of Items	Items
Cluster 1	2	{A, E}
Cluster 2	1	{B}
Cluster 3	1	{C}
Cluster 3	1	{D}

Department of Statistics and Data Science
Jahangirnagar University
Part IV B. Sc (Honors) Second Tutorial Examination – 2024
Course Code: Stat - 405
Course Title: Data Mining

Set SDS

Time: 01 Hour **Marks: 20**

Answer the following questions. Each question carries equal marks.

Q1. The decision tree model was built using X_1, X_2, X_3, and X_4 to predict the response of Y. The following table consists of training data from a database. Let Y be the class label attribute. Obtain the Gain Ratio for the attribute X_4 – the attribute selection measure in classification. [1.5x4=6]

X_1	X_2	X_3	X_4	Y
N	F	W	M	N
G	F	W	L	Z
N	D	M	M	Z
N	F	S	M	Z
G	F	S	L	Z
G	D	W	M	Z
N	F	W	H	N
N	D	W	H	N
N	F	M	H	Z
N	D	S	M	N
G	D	S	L	N
G	D	M	L	Z
G	F	M	H	Z
G	F	S	M	Z

Q2. The predicted probability and the true label are given in the following table. For each threshold, classify any instance with a predicted probability greater than or equal to the threshold as positive (1), and those below the threshold as negative (0). Construct the ROC curve and calculate the Area Under the ROC Curve (AUC). [4+4=8]

ID	Predicted Probability	True Label
1	0.899	1
2	0.827	0
3	0.772	1
4	0.734	1
5	0.660	0
6	0.546	1
7	0.412	0
8	0.360	0

Q3. The final partition matrix for different values of k is obtained for the Fuzzy C-means clustering. Find the optimal number of clusters to cluster the items A, B, C, D, E, F, and G. Comment on your results. [2x3+1=7]

Item	$k=2$	
A	0.945	0.055
B	0.938	0.062
C	0.191	0.809
D	0.163	0.837
E	0.597	0.403
F	0.648	0.352
G	0.214	0.786

Item	$k=3$		
A	0.026	0.967	0.007
B	0.033	0.958	0.009
C	0.994	0.004	0.002
D	0.963	0.021	0.015
E	0.340	0.464	0.196
F	0.333	0.504	0.162
G	0.004	0.002	0.995

Item	$k=4$			
A	0.000	1.000	0.000	0.000
B	0.000	1.000	0.000	0.000
C	0.990	0.003	0.002	0.005
D	0.991	0.003	0.002	0.004
E	0.001	0.002	0.001	0.996
F	0.002	0.002	0.001	0.995
G	0.000	0.000	1.000	0.000

Good Luck

Pa

r

Answe

Department of Statistics and Data Science
Jahangirnagar University
Part IV B. Sc (Honors) Third Tutorial Examination – 2024
Course Code: Stat - 405
Course Title: Data Mining

Set III-SDS

Marks: 20

Time: 01 Hour

Answer the following questions. Each question carries equal marks.

Q1. The following contingency table summarizes supermarket transaction data, where A refers to the transactions containing A, $\bar{A}$ refers to the transactions that do not contain $\bar{A}$, B refers to the transactions containing B, and $\bar{B}$ refers to the transactions that do not contain $\bar{B}$. [5]

Table 1: A 2x2 contingency table for items A and B

Item	B	$\bar{B}$	Total
A	4000	2000	6000
$\bar{A}$	1000	3000	4000
Total	5000	5000	10000

Find Yule's Q, J-Measures, Laplace, Interest, Conviction, Certainty Factor, Collective Strength, Klosgen, Kulc, and Imbalance Ratio.

Q2. Consider a grayscale image and a convolutional filter represented as follows: [5]

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

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

Q3. What is meant by Support Vector Machine? What are the different types of kernels used in support vector machine (SVM)? [5]

Q4. Define the following Clustering validity indices: (i) CH index, (ii) ~~CH index~~ Dunn Index, (iii) C-Index, (iv) Hartigan index, and (v) Gap index [5]

Good Luck

Department of Statistics and Data Science
Jahangirnagar University
Part IV B. Sc. (Honors) 1st Tutorial Examination, 2024
Course Title: Actuarial Statistics
Course No. STAT 406

Total marks: 10 **Total Time: 1 hr**

Answer the following questions.

Q1.[7+3] a) A life insurance company issues an n-year endowment insurance policy to a life aged x with:

- Sum assured (face value): 1 (payable at the end of the year of death or at maturity if the insured survives n years)
- Level annual premium: P Payable at the beginning of each year for at most n years, as long as the insured is alive
- Interest rate: i per annum
- Mortality basis: Select life table l_x is given
- Expenses:
  - Per policy issue expense: e_0 (incurred at the start of the policy)
  - Per premium collection expense: e_p (incurred at the beginning of each year when premium is paid)
  - Per death claim expense: e_d (incurred at the end of year of death)
  - Per maturity claim expense: e_m(incurred at the end of year n, if the insured survives)

Using the principle of equivalence, derive the expression to calculate the annual gross premium per policy for this insurance.

b) Calculate the level premium P of a 3-year term insurance on (55) of 1000. Given that

- *Mortality: Illustrative life table*

x	l_x	d_x
55	8,640,861	77,426
56	8,563,435	83,527
57	8,479,908	90,082
58	8,389,826	

- P is paid at the at the beginning of each year for 3 years as long as the insured is alive
- $i = 0.05$

Department of Statistics and Data Science
Jahangirnagar University Examination, 2024
Part IV B. Sc. (Honors) 2nd Tutorial Examination, 2024
Course Title: Actuarial Statistics
Course No. ST[illegible] Total Time: 60 Minutes

Total marks: 10

Answer the following questions

---

a) Give some examples of life annuities?

b) An annuity on (x) provides k annually beginning at age $(x+n)$. Nothing is paid before n years. Annual premiums are payable for m $(< n)$ years beginning at x. Draw a cash flow diagram and obtain the expression for annual premium, assuming limiting age ω

c) A 60-year-old couple purchases a joint life annuity that pays $20,000 annually at the end of each year, ceasing at the second death.

Write the EPV in terms of joint survival probabilities

STAT-407

2nd Tutorial 4th Year-2024

1. Discuss the proximate determinants of fertility as proposed by Bongaarts. Describe the procedure of estimating fertility indices following Bongaarts.
2. Suppose total fecundity = 15, percent married = 60, average duration of postpartum infecundability = 5 months, contraceptive prevalence = 65%, and average contraceptive effectiveness = 95%. Using Bongaarts model, estimate the Total Fertility Rate (TFR).
3. Write down the different techniques used for estimating child mortality. Describe any one of these methods in detail.
4. Define the concept of fertility differentials. Discuss the major socioeconomic and demographic factors responsible for fertility differentials in populations along with its framework.

SJAT-407

2nd Tutorial 4th Year-2024

Q1. A clinical trial was conducted to assess the effectiveness of a new oral contraceptive. A total of 250 women participated in the study and used the method continuously for 18 months. During this period, 5 accidental pregnancies were reported.

1. Compute the Pearl Index. Interpret your result.
2. Suppose another method resulted in a Pearl Index of 1.2. Compare the two methods and comment on which one is more effective and why.

Q2. The following table presents data on:

Age of Mother (years)	15–19	20–24	25–29	30–34	35–39	40–44	45–49
Annual Births of Daughters	0.014	0.050	0.069	0.058	0.028	0.007	0.001
Survival Rate	4.955	4.941	4.926	4.903	4.875	4.842	4.800

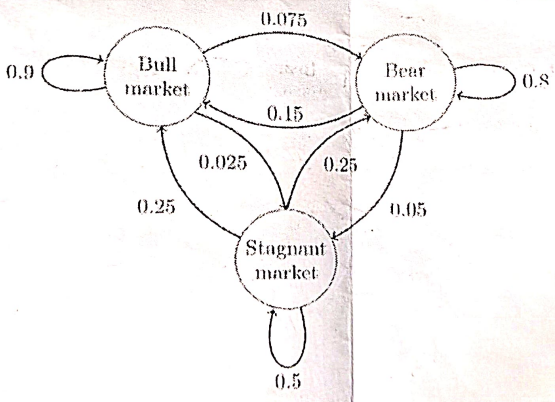
Each age group represents a 5-year interval within the reproductive span.

1. Estimate the average number of daughters a woman is expected to bear during her lifetime, both ignoring and considering female mortality.
2. Determine whether this population is replacing itself, growing, or declining.
3. Estimate the average age of mothers at childbearing.

Jahangirnagar University
Department of Statistics and Data Science
1st Tutorial, Course Title: Stochastic Process, Course Code: STAT-408

Marks: 10 **Time: 60 Minutes**

1. Explain the stationery process and the independent increment of a stochastic process. **1.5**
2. Consider the process $X(t) = A_1 + A_2 t$ where A_1 and A_2 are independent random variables with $E(A_i) = a_i, V(A_i) = \sigma_i^2 \; ; i = 1, 2$. Test the stationarity. **1.5**
3. Define Markov process and classification of Markov Chain. Show that, for any state j, $\sum_{n=0}^{\infty} P_{jj}^{(n)} = \frac{1}{1-F_{jj}}$ and if state j is recurrent then $\sum_{n=0}^{\infty} P_{jj}^{(n)} = \infty$. **3**
4. The states represent whether a hypothetical stock market is exhibiting a bull market, bear market, or stagnant market trend during a given week. According to the figure, a bull week is followed by another bull week 90% of the time, a bear week 7.5% of the time, and a stagnant week the other 2.5% of the time. Labeling the state space {1 = bull, 2 = bear, 3 = stagnant} construct the transition matrix for this example and using this matrix .



i. Specify the classes of the Markov chain and determine whether they are transient or recurrent. **1**
ii. Is the Markov chain irreducible and aperiodic? **1**
iii. Calculate the long-term fraction of weeks during which the market is stagnant, bull market, or bear market. **2**

Jahangirnagar University
Department of Statistics and Data Science
2nd Tutorial, Course Title: Stochastic Process, Course Code: STAT-408

Marks: 10 **Time: 60 Minutes**

Q1. Write down the properties of a Poisson Process. If $\{N(t), t \geq 0\}$ is Poisson process and $s < t$ then show that 2

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

Q2. Suppose the interarrival time X_n have gamma distribution with density $f(x) = \lambda e^{-\lambda x}$; $x > 0$. Find the distribution of Renewal process along with its mean function. 3

Q3. Define a Linear Growth Model with Immigration. Suppose $\{X(t); t \geq 0\}$ be a birth and death process with 3

$$\lambda_n = n\lambda + \theta \quad ; \quad n \geq 0$$
$$\text{and } \mu_n = n\mu \quad ; \quad n \geq 1$$

Show that $M(t) = \frac{\theta}{\lambda-\mu}\left[e^{(\lambda-\mu)t} - 1\right] + ne^{(\lambda-\mu)t}$.

Q4. Suppose that customers arrive in accordance with a Poisson process with mean rate of 2 per minute $\lambda = 2/minute$). Then the interval between any two successive arrival's follows exponential distribution with mean $\frac{1}{\lambda} = \frac{1}{2} minute$. What is the probability that the interval between two successive arrival is 2

a) More than 1 minute
b) 4 minute or less
c) Between 1 and 2 minutes
d) Find the expected time until the 9th customer.

Tutorial 1
Date: 30.06.2025

Jahangirnagar University
Savar, Dhaka-1342

Time: 1 Hour
Full marks: 20

Department of Statistics and Data Science

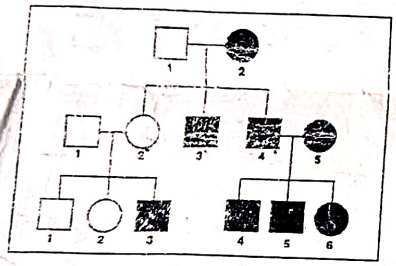
Subject: Statistical genetics and Bioinformatics

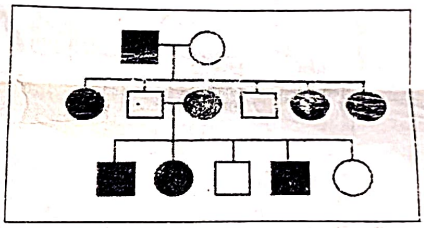
Course code: STAT-409

Answer the following questions. The mark and its breakdown are given on the right-hand side of each question.

Q1. Make a comparison among **(i)** DNA, **(ii)** Gene, and **(iii)** Chromosome. **4.5**

Q2. Make a comparative study between the following two **ancestry patterns.** **4.5**





Q3. What is the relation between a person's **genotype** and a **phenotype** that he/she has? Explain with a suitable example. **03**

Q4. Suppose, a **SNP** is diallelic ("A" and "T"). **03**

(i) What would be the possible **genotype combinations**?

(ii) If this SNP is in **HWE** law, compute the genotype frequencies if the **major allele** frequency is 0.75?

Q5. Define the general form of a **penetrance function**, and, hence modify it for all **inheritance patterns**? Apply these penetrance functions along with the concept of GRR for possible **genetic modeling**. **05**

Good Luck!

Tutorial 2
Date: 04.08.2025

Jahangirnagar University
Savar, Dhaka-1342

Time: 1 Hour
Full marks: 20

Department of Statistics and Data Science

Subject: Statistical genetics and Bioinformatics

Course code: STAT-409

Answer the following questions. The mark and its breakdown are given on the right-hand side of each question.

Q1. What is the purpose of **quality control** (QC) for SNPs in genetic analysis? Give an explanation of its various stages. 04

Q2. Explain the concept of **Hardy-Weinberg equilibrium (HWE)** in the context of a diallelic ("A" and "T") SNP. If both the parents are **heterozygous**, how would you calculate the probability that their offspring would be **(i)** homozygous, **(ii)** heterozygous, **(iii)** opposite homozygous? 04

Q3. Suppose, there are two diallelic ("A" and "T") SNPs that follow **HWE** rule. How would you calculate the strength of **LD** for this pair of SNPs? Explain in detail. 04

Q4. **(i)** If an **allele** is observed to be associated with a **disease**, how can it be interpreted? 4.5

(ii) List the different **sequencing platforms** used by **Illumina** Biotechnological Company.

(iii) Provide a list of **mandatory** data types based on **NARURE's** data repository guidance.

Q5. What is the GenBank? Mention its important features. 3.5

Good Luck!

Tutorial 3
Date: 13.10.2025

Jahangirnagar University
Savar, Dhaka-1342

Time: 1 Hour
Full marks: 20

Department of Statistics and Data Science

Subject: Statistical Genetics and Bioinformatics

Course code: STAT-409

Answer the following questions. The mark and its breakdown are given on the right-hand side of each question.

Q1. What is the Next Generation Sequencing **(NGS)**? Explain the specific applications of NGS in **(i)** Cancer research, **(ii)**Microbiology and infectious diseases. **04**

Q2. Define the concept of Short Tandem Repeat **(STR)**. Explain its applications *in genomic studies.* **04**

Q3. List the approaches used in **NGS** for analyzing the DNA sequences in genomic studies. Also, state the different steps of a **NGS workflow**. Write down the differences between **NGS** and **Sanger sequencing**. **04**

Q4. Suppose, there are two diallelic ("**A**" and "**T**") SNPs that follow **HWE** rule. Explain the detail steps for testing association in **GWAS**. **04**

Q5. List the different scope of **machine learning** methods in bioinformatics. How a machine learning finds its application to predict the **cardiovascular disease**? Explain. **04**

Good Luck!

Department of Statistics and Data Science
Jahangirnagar University
Tutorial Examination – 2024

Course Name: Categorical Data Analysis **Course Code: STAT-410**
Total marks: 10 **Time: 40 Minutes.**

Q1. Explain the intuition behind the Wald test, Lagrange Multiplier (LM) test, and Likelihood Ratio Test (LRT) in the context of hypothesis testing in regression.

Q2. Define the odds ratio (OR) and explain how it is interpreted in the context of a binary logistic regression model.

Q3. A researcher is studying choice of transport (Car, Bus, Bicycle) as a function of income. Formulate a multinomial logistic regression model for this problem and explain how the baseline category is chosen.

Q4. Discuss how the Negative Binomial regression addresses overdispersion. Illustrate with an applied example.

Jahangirnagar U

Course Name: Financial Statistics
Course Code: STAT- 411
Second Tutorial

Marks: 10
Time: 60 minutes
Date: 10/12/2025

1. What is Option Pricing? Explain call option and put option. What are the models do we use to capture option pricing? Write the equation of BSM call option price.
2. What is meant by copula modeling approach? Suppose we observe returns for two assets (A and B) across three observations, as shown below:

Observation	Asset A return	Asset B return
1	0.03	0.04
2	-0.02	0.01
3	0.06	-0.01

Using this dataset, complete the following tasks:

i) Compute the marginal probabilities for each asset using the empirical distribution.
ii) Apply the Gaussian Copula transformation by mapping marginal probabilities to the standard normal space.
iii) Calculate the Gaussian Copula density for Observation 1.

Course Name: Financial Statistics
Course Code: STAT- 411
First Tutorial

Marks: 10
Time: 60 minutes
Date: 24/09/2025

1. You have daily returns for a stock for 10 trading days:

$$[0.01, -0.02, 0.015, 0.03, -0.01, 0.02, -0.005, 0.015, -0.01, 0.005]$$

Using a GARCH(1,1) specification:

$$h_{t+1} = w + \alpha r_t^2 + \beta h_t$$

Where,

- $w = 1 \times 10^{-6}$,
- $\alpha = 0.10$,
- $\beta = 0.85$
- r_t = daily return
- h_t =sample variance of the 10 returns

a) Compute the sample (daily) standard deviation.
b) Compute the forecasted daily variance and volatility using recursive formula.
c) Draw a line plot of volatility over the 10 days to visualize clustering.
d) Interpret the (b) and (c).
e) What does it mean when we say "volatility is persistent"? Justify your result.

Jahangirnagar University
Department of Statistics and Data Science
Part IV B.Sc. (Honors) Practical Examination – 2024
Course No.: Stat Lab-412 (Group-A)
Course Name: Statistical Data Analysis VIII (Based on Statistical Inference)
Practical Tutorial -09.12.2025

Time: 60 minutes **Marks: 07**

Please use the "airquality" data by utilizing R software and answer the following questions with comments also in each question.

1. Estimate the jackknife standard error of the mean Ozone level.
2. Plot the Jackknife estimate of the mean Ozone level with interpret also. Examine whether any observation strongly influences the mean.
3. Generate 1000 bootstrap samples of the median Temperature.
4. Estimate the bootstrap standard error of the median.
5. Construct 95% confidence intervals using:
   (*i*) Percentile method
   (ii) Basic bootstrap method
6. Visualize the bootstrap distribution using a histogram.
7. Compare which resampling method is the best for this dataset?

Good Luck!

PRACTICAL Tutorial

Date: 27.11.2025

Jahangirnagar University
Savar, Dhaka-1342

Time: 1 Hour

Full marks: 10

Department of Statistics and Data Science

STAT-LAB-413 (Group B): Statistical Data Analysis IX

Related Subject: Statistical Genetics and Bioinformatics (STAT-409)

Answer the following question.

Q1. Suppose, the gene **ABL1** has 15 SNPs for 200 individuals in a case-control setting.

(a) Perform the quality control (QC) of SNPs using, **(i)** minor allele frequency (MAF), **(ii)** PIC, **(iii)** pairwise LD measures (r^2).

(b) Show the following measures in a plot, **(i)** the MAF values, **(ii)** PIC values, **(iii)** r^2.

(c) After QC, how many markers are to be used for further analysis? Justify about your scientific reduction of markers by interpretation of each plot.

Good Luck!

STAT-404 (Lab Tutorial)

1. In agricultural research station an experiment is conducted to study the productivity of balsam apples under nitrogen fertilizer. Four levels of nitrogen as urea are used in the experiment. The levels are 30 kg/ha, 60 kg/ha, 90 kg/ha and 120 kg/ha. The production of balsam apples in a plot of size 10' x 15' are recorded along with the number of plants per plot. Production or balsam apple (y_{ij} kg) along with number of plants (x_{ij})

N_1		N_2		N_3		N_4	
y	x	y	x	y	x	y	x
50.5	42	55.0	52	52.3	45	60.0	50
48.0	50	60.2	55	51.4	48	61.0	48
40.5	46	40.5	48	35.0	40	65.0	52
52.0	48	50.0	51	48.0	46	35.0	35
50.0	45	56.0	48	55.0	49	38.0	40

i. Write down the appropriate model for this data. Justify your choice.
ii. Complete the ANOVA/ANCOVA table.
iii. Conduct a test procedure to check whether the four nitrogen levels work similarly or not

2. A pharmaceutical company produced 2 varieties of protein, and these were given to guinea pigs of different ages, different body condition and in different regional temperature. The guinea pigs were classified into 2 classes according to body condition and 2 classes according to age groups. The body weight of each guinea pig was recorded after the applying treatment with 2 different temperatures. The increased body weights of each guinea pig are shown below:

Age Group	Body Condition	Temperature	Protein level	
			Low dose	High dose
Younger	Not So Good	Low	729	749
Younger	Good	Low	1075	1037
Younger	Not So Good	Low	642	669
Younger	Good	Low	633	550
Elder	Not So Good	Low	433	868
Elder	Good	Low	1063	1052
Elder	Not So Good	Low	635	650
Elder	Good	Low	601	604
Younger	Not So Good	High	729	749
Younger	Good	High	1075	1037
Younger	Not So Good	High	642	669
Younger	Good	High	633	550
Elder	Not So Good	High	433	868
Elder	Good	High	1063	1052
Elder	Not So Good	High	635	650
Elder	Good	High	601	604

i. Analyze the data when fourth order interaction as well as second order interaction of body condition and protein level are confounded and test the varieties of Age Group.
ii. Analyze the data when the third order interaction of body condition, temperature and protein level is confounded and test the varieties of Body condition.
iii. Test the varieties of temperature is similar or not.

Department of Statistics and Data Science
Jahangirnagar University
Part IV B. Sc. (Hons.) Practical Tutorial Examination 2024
Course Title: *Statistical Data Analysis IX (Group-A)*
Course No. STAT-LAB-413

Time: 2 Hours **Full Marks: 20**

The Madison, Wisconsin, police department regularly monitors many of its activities as part of an ongoing quality improvement program. Table 5.8 gives the data on five different kinds of overtime hours. Each observation represents a total for12 pay periods, or about half a year. This data is available in the file **multivar5th/T5-8.dat.s**

Table 1: Five types of overtime hours for the Madison, Wisconsin, Police department

Legal Appearances Hours x_1	Extraordinary Event Hours x_2	Holdover Hours x_3	Compensatory overtime allowed x_4	Meeting Hours 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 six variables $(x_1, x_2, \ldots, x_5)$ *for testing* $H_0 : \boldsymbol{\mu}' = [3500 \quad 1400 \quad 2600 \quad 13500 \quad 800]$. 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 variates. 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 $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, respectively. What proportion of the total sample variance of the first set $\mathbf{Z}^{(1)}$ is explained by the canonical variate $\hat{U}_1$? What proportion of the total sample variance of 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. Cluster the five variables using the single linage and complete linkage hierarchical methods. Draw the dendrograms and compare the results.

7. A public health researcher is designing a study to understand the factors influencing adolescent health. They plan to collect data on several variables, including:

- Physical Activity (hours per week)
- Screen Time (hours per day)
- Socioeconomic Status
- Body Mass Index (BMI)
- Academic Stress Score
- Blood Pressure
- Self-Esteem Score

The researcher is considering two different analysis plans:

Plan A: Use a single statistical model to understand how Physical Activity, Screen Time, and Socioeconomic Status (as independent variables) are collectively associated with the combination of BMI, Blood Pressure, and Self-Esteem Score (as dependent variables).

Plan B: Use a statistical model to understand how Physical Activity, Screen Time, and Socioeconomic Status (as independent variables) predict only the Self-Esteem Score (as the dependent variable).

(i) Which analysis plan (A or B) is an example of a multivariate analysis? Justify your answer by explaining the defining characteristic of a multivariate analysis.

(ii) Which analysis plan (A or B) is an example of a multivariable analysis? Justify your answer by explaining the defining characteristic of a multivariable analysis.

(iii) Explain why a researcher might choose the multivariate approach (Plan A) over creating three separate multivariable models for BMI, Blood Pressure, and Self-Esteem.

- Good Luck -

RETAKE Tutorial 3 Date: 24.11.2025	Jahangirnagar University Savar, Dhaka-1342	Time: 1 Hour Full marks: 20

Department of Statistics and Data Science

Subject: Statistical Genetics and Bioinformatics

Course code: STAT-409

Answer the following questions. The mark and its breakdown are given on the right-hand side of each question.

Q1. List some important resources of a **NGS** lab. How would you proceed to select the best sequencing (NGS) company for your research? Explain. **05**

Q2. List some applications of Short Tandem Repeat **(STR)**. Distinguish between a **NGS** and a **qPCR** methodology. **05**

Q3. What is the importance of multiple testing in a **GWAS**? How can you calculate the error criteria in a GWAS **multiple testing**? State some methods to control such errors. **05**

Q4. How is a **machine learning** method used in Single-Cell Analysis and Precision Medicine? Explain. **05**

Good Luck!