Reg. No.																

Question Paper Code	13170

B.E. / B.Tech. - DEGREE EXAMINATIONS, NOV / DEC 2024

Third Semester

Computer Science and Business Systems

20BSMA305 - COMPUTATIONAL STATISTICS

Regulations - 2020

(Use of Statistical Tables is permitted)

Duration: 3 Hours | Max. Marks: 100

PART - A (MCQ) (20 × 1 = 20 Marks)

Answer ALL Questions

Marks K–Level CO

1. For a bivariate normal distribution, if X and Y are independent, then their correlation coefficient ρ is — *1 K1 CO1*
(a) 1 (b) -1 (c) 0 (d) Cannot be determined
2. In a multivariate normal distribution with p variables, how many parameters need to be estimated for the covariance matrix — *1 K2 CO1*
(a) p (b) p^2 (c) $\frac{p(p+1)}{2}$ (d) $\frac{p(p-1)}{2}$
3. If X and Y follow a bivariate normal distribution with means μx and μy, then the conditional expectation $E(Y|X=x)$ is: — *1 K1 CO1*
(a) A constant (b) A linear function of x
(c) A quadratic function of x (d) An exponential function of x
4. The maximum likelihood estimator of the mean vector μ in a multivariate normal distribution is: — *1 K1 CO1*
(a) The sample median (b) The sample mode
(c) The sample mean vector (d) The first observation
5. In linear regression model $Y=\beta_0+\beta_1 z_1+\beta_2 z_2+\cdots.+\beta_r z_r+\varepsilon$, what does ε refer? — *1 K1 CO2*
(a) The sample size (b) Error
(c) Regression Coefficient (d) Mean
6. Which statistic is commonly used to detect multi-collinearity in a regression model? — *1 K1 CO2*
(a) Durbin-Watson statistic (b) Variance Inflation Factor (VIF)
(c) Cook's Distance (d) Jarque-Bera statistic
7. For multiple regression model SST=200, SSE=50. The multiple coefficient of determination is ---------. — *1 K2 CO2*
(a) 0.25 (b) 4.00 (c) 0.75 (d) 0.92
8. A measure of goodness of fit for the estimated regression equation is the — *1 K1 CO2*
(a) multiple coefficient of determination (b) mean square due to error
(c) mean square due to regression (d) p-value
9. What is the multivariate test statistic for MANOVA? — *1 K1 CO3*
(a) Wilk's Lambda (b) Hotelling's Trace (c) Pillai's Trace (d) Roy's Largest Root
10. In a Multivariate Analysis of Variance (MANOVA), what does the test assess? — *1 K1 CO3*
(a) Differences in variances among multiple groups
(b) Differences in the mean vectors of several groups
(c) The relationship between two categorical variables
(d) Differences in single-variable means
11. The apparent error rate for the confusion matrix $\begin{pmatrix}3 & 2\\ 2 & 3\end{pmatrix}$ for the population π_1 and π_2 with 5 samples each is — *1 K2 CO3*
(a) 0.5 (b) 0.4 (c) 0.2 (d) 0.6

12. Which of the following is not the part of the exploratory factor analysis process? *1 K1 CO3*
(a) Extracting factors
(b) Determining the number of factors before the analysis
(c) Rotating the factors
(d) Refining and Interpreting the factors

13. Which of the following are recommended applications of PCA? *1 K1 CO4*
(a) Data Visualisation (b) As a replacement for linear regression
(c) Data compression (d) None of the above

14. What do factor scores help identify in factor analysis? *1 K1 CO4*
(a) Outliers in the dataset (b) The significance of each variable
(c) The relationship between factors (d) The total sample variance explained

15. What is the importance of the cumulative explained variance plot in PCA? *1 K1 CO4*
(a) It shows the distribution of eigenvalues in the dataset
(b) It indicates a sudden drop in variance explained
(c) It helps identify the optimal number of principal components
(d) It calculates the maximum eigen value for retention

16. What is a latent variable**?** *1 K1 CO4*
(a) It is a variable that cannot be measured directly.
(b) It is another name for a factor.
(c) Latent variables represent clusters of variables that correlate highly with each other.
(d) All of these are correct.

17. Which clustering algorithm is based on the concept of centroids? *1 K1 CO5*
(a) K-Means (b) DBSCAN (c) Agglomerative (d) Mean-Shift

18. Which clustering algorithm is based on the concept of minimizing the within-cluster variance? *1 K1 CO5*
(a) K-Means (b) DBSCAN (c) Agglomerative (d) Mean-Shift

19. What is needed for k-means clustering? *1 K1 CO5*
(a) Defined distance metric (b) Number of clusters
(c) Initial guess to cluster centroids (d) All of these

20. Which of the following is a goal of clustering algorithms? *1 K1 CO5*
(a) Classification (b) Regression
(c) Dimensionality reduction (d) Grouping similar data points together

PART - B (10 × 2 = 20 Marks)

Answer ALL Questions

21. Find the distribution of X_2 given $X_1 = x_1$ and $X_3 = x_3$, when $\mu =$ (-3,1,4) and $\Sigma = \begin{pmatrix} 5 & -2 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$. *2 K1 CO1*

22. A bivariate normal distribution has the following parameter *2 K2 CO1*
$\mu_x = 2, \quad \mu_y = 3, \quad \sigma_x = 2, \quad \sigma_y = 3, \quad \rho = 0.5$
Find $E[Y|X = 4]$ and $Var[Y|X = 4]$.

23. Write the null and alternative hypotheses for the Durbin-Watson test. *2 K1 CO2*

24. What is multi-collinearity? How does it affect regression estimates? *2 K2 CO2*

25. What is the test statistic for the variables $p = 2$ and population $g \geq 2$ in MANOVA? *2 K1 CO3*

26. The density function associated with populations π_1 and π_2 are given by $f_1(x_0) = 0.6$ and $f_2(x_0) = 0.4$ with probabilities $p_1 = 0.3$ and $p_1 = 0.7$. The cost of miss classification is given as $C(1/2) = 4, C(2/1) = 9$. How will you classify the new observation? *2 K2 CO3*

27. What is the formula for proportion of total population variance due to k^{th} principle component? *2 K2 CO4*

28. Convert the covariance matrix $\Sigma = \begin{pmatrix} 1 & 16 \\ 16 & 225 \end{pmatrix}$ to correlation matrix. *2 K2 CO4*

29. Define Cluster Analysis. *2 K2 CO5*

30. Define k-mean clustering. *2 K1 CO5*

PART - C (6 × 10 = 60 Marks)
Answer ALL Questions

31. a) Find the maximum likelihood estimates of 2×1 mean vector μ and 2×2 covariance matrix Σ based on the random sample X=$\begin{pmatrix} 3 & 6 \\ 4 & 4 \\ 5 & 7 \\ 7 & 7 \end{pmatrix}$ from a bivariate normal population. *10 K2 CO1*

OR

b) Let X = $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ be a normal random vector with the following mean vector and covariance matrix $\mu = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, Cov = $\begin{pmatrix} 4 & 1 \\ 1 & 1 \end{pmatrix}$.Let also A = $\begin{pmatrix} 2 & 1 \\ -1 & 1 \\ 1 & 3 \end{pmatrix}$, b = $\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$, Y= $\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$ = AX+b *10 K3 CO1*

(a) Find P($X_2 > 0$)
(b) Find the expected value vector of Y, $\mu_y = E_y$
(c) Find the covariance matrix of Y
(d) Find P($Y_2 \leq 2$)

32. a) Fit a Multivariate Straight line regression model for the following data *10 K3 CO2*

z_1	0	1	2	3	4
y_1	1	4	3	8	9
y_2	-1	-1	2	3	2

OR

b) If the response variables take the value Y= $\begin{pmatrix} 10 & 100 \\ 12 & 110 \\ 11 & 105 \end{pmatrix}$ and the design matrix *10 K3 CO2*

X = $\begin{pmatrix} 1 & 9 & 62 \\ 1 & 8 & 58 \\ 1 & 7 & 64 \end{pmatrix}$ takes the value.

Calculate $\hat{\beta}$
Calculate $\hat{\varepsilon}$
Fit a linear regression model.

33. a) Find the Fisher's discriminant for the following : *10 K2 CO3*

$$X_1 = \begin{pmatrix} -2 & 5 \\ 0 & 3 \\ -1 & 1 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 0 & 6 \\ 2 & 4 \\ 1 & 2 \end{pmatrix}, \quad X_3 = \begin{pmatrix} 1 & -2 \\ 0 & 0 \\ -1 & -4 \end{pmatrix}$$

OR

b) Construct a MANOVA table for the following *10 K2 CO3*

$$\begin{pmatrix} \begin{pmatrix} 9 \\ 3 \end{pmatrix} & \begin{pmatrix} 6 \\ 2 \end{pmatrix} & \begin{pmatrix} 9 \\ 7 \end{pmatrix} \\ \begin{pmatrix} 0 \\ 4 \end{pmatrix} & \begin{pmatrix} 2 \\ 0 \end{pmatrix} & \\ \begin{pmatrix} 3 \\ 8 \end{pmatrix} & \begin{pmatrix} 1 \\ 9 \end{pmatrix} & \begin{pmatrix} 2 \\ 7 \end{pmatrix} \end{pmatrix}$$

34. a) Let $\Sigma = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}$ determine the principal componentsY_1, Y_2, Y_3. *10 K3 CO4*

OR

b) Let $\rho = \begin{bmatrix} 1 & 0.63 & 0.45 \\ 0.63 & 1 & 0.35 \\ 0.45 & 0.35 & 1 \end{bmatrix}$ be the covariance matrix. Assuming an $m = 1$ factor model find the loading matrix L. *10 K3 CO4*

35. a) Construct the Principle components Y_1, Y_2 & Y_3 for the covariance matrix$\Sigma = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$. Also find Var($Y_1$) , Cov($Y_1$,$Y_2$) & show that $\sum_{i=1}^{3} \sigma_{ii} = \sum_{i=1}^{3} \lambda_i$ *10 K3 CO5*

OR

b) Suppose we measure two variables X_1 and X_2 for four items A,B,C and D. The data are as follows: *10 K3 CO5*

	Observations	
Item	X_1	X_2
A	5	4
B	1	-2
C	-1	1
D	3	1

Use K-means clustering technique to divide the items in to K=2 clusters. Start with the initial groups (AB) and (CD).

36. a) If X $\sim N_3(\mu,\ \Sigma)$, $\mu = \begin{pmatrix} 5 \\ 3 \\ 7 \end{pmatrix}$ *and* $\Sigma = \begin{pmatrix} 4 & -1 & 0 \\ -1 & 4 & 2 \\ 0 & 2 & 9 \end{pmatrix}$, then find *10 K2 CO4*

(i) P($X_1 > 6$)
(ii) P($5X_2 + 4X_3 > 70$)
(iii) P($4X_1 - 3X_2 + 5X_3 < 80$).

OR

b) Suppose R = $\begin{pmatrix} 1 & 0.632 & 0.511 & 0.511 & 0.155 \\ 0.632 & 1 & 0.574 & 0.322 & 0.213 \\ 0.511 & 0.574 & 1 & 0.183 & 0.146 \\ 0.115 & 0.322 & 0.183 & 1 & 0.683 \\ 0.155 & 0.213 & 0.146 & 0.683 & 1 \end{pmatrix}$ be the correlation matrix. The eigenvalues more than unity are 2.437 and 1.407. *10 K2 CO4*
F_1=$(0.732\ \ 0.831\ \ 0.726\ \ 0.605\ \ 0.563)^T$
F_2=$(-0.437\ \ -0.280\ \ -0.374\ \ 0.694\ \ 0.719)^T$ are the loading factors. Find the residual matrix.