Questions for Applied Multivariate Statistical Modeling
Prof J Maiti, ISE, IIT Kharagpur

1. Define random variable. Give two examples. Give a mathematical expression for random variable.
2. What is multivariate situation? Give two examples. Define a mathematical expression for a variate.
3. What do you mean by modeling? Explain with examples the types of modeling approaches.
4. Define data. Classify data as per scale of measurement. Give examples for each case.

5. If $X \sim N_P(\mu, \Sigma)$ and $X_{n\times p}$ is n multivariate observations on p variables, give the mathematical expression for the following:
   a) Variance-covariance matrix
   b) Correlation matrix
   c) Mahalanobis distance
   d) Distribution of linear combination of the components of X
   e) Distribution of q linear combination of the components of X

6. Define sampling distribution. If $X \sim N_P(\mu, \Sigma)$, what will be the sampling distribution of $\overline{X}$ under the null hypothesis $H_0 : \mu = \mu_0$?
7. Let $X \sim N_P(\mu, \Sigma)$. The multivariate descriptive statistics for $X_{100\times 3}$ is as follows:

$$\overline{X} = \begin{bmatrix} 10 \\ 15 \\ 5 \end{bmatrix} \text{ and } S = \begin{pmatrix} 100 & & \\ 25 & 150 & \\ 10 & 20 & 50 \end{pmatrix}$$

Obtain Bonferroni simultaneous confidence intervals for μ_1, μ_2 and μ_3.

8. A quality control engineer wants to select the best vendor amongst the locally available vendors. A preliminary analysis of defects data screen out most of the vendors and two vendors remain competitive. But the quality control engineer wants to be sure about the best vendor or otherwise will go for both the vendors.

a) Convert the above-mentioned practical problem into a statistical problem (give details of the procedure you would adopt).
b) What method do you adopt to select the best vendor and how do you do that?

9. A population is characterized by a variable X defined as N(10, 25). A sample of size 25 is drawn at random from the population. What is the pdf of $\overline{X}$?

10. The covariance matrix of $X = N_2(\mu, \Sigma)$ is given below. Obtain its correlation matrix.

$$\Sigma = \begin{pmatrix} 9 & 6 \\ 6 & 25 \end{pmatrix}$$

11. Define statistical distance. A bivariate normal variable $X = N_2(\mu, \Sigma)$ has mean vector μ^T = (3, 5) and covariance matrix Σ as given in question 10. Obtain the equation for statistical distance.
12. Define Hotelling's T^2. State with two examples the uses of Hotelling's T^2.
13. X is $N_2(\mu, \Sigma)$. Obtain its bivariate normal density function (pdf).
14. Show that for p = 2, Mahalanobis $D^2 = (X - \mu)' \Sigma^{-1} (X - \mu)$ is an equation of an ellipse. If the variables are independent and have equal variance, what will be the shape of Mahalanobis D^2?
15. A quality control engineer obtains the following observations of a process with two variables (i = 4 and p=2):

$$X = \begin{pmatrix} 2 & 12 \\ 8 & 9 \\ 6 & 9 \\ 8 & 10 \end{pmatrix}$$

Specify the distribution of Hotelling T^2.

16. The following questions refer to the situation given below.
Let X be $N_2(\mu, \Sigma)$. A random sample of size (n=4) is collected and shown below:

$$X_{4\times 2} = \begin{pmatrix} 2 & 12 \\ 8 & 9 \\ 6 & 9 \\ 8 & 10 \end{pmatrix}$$

- Obtain multivariate descriptive statistics.
- Write out the bivariate normal density [assuming sample statistics of (i) as population parameters].
- Specify the T^2-distribution for the situation given.
- Test H_0: $\mu' = [7, 11]$ at the $\alpha = 0.05$ level. What conclusion do you reach?
- Construct 95% simultaneous confidence intervals for the population means.

17. The following questions (**i – iv**) refer to the situation given below.
Let **X** be $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. A random sample of size (n=5) is collected and shown below:

$$X_{5\times 3} = \begin{pmatrix} 9 & 12 & 3 \\ 2 & 8 & 4 \\ 6 & 6 & 0 \\ 5 & 4 & 2 \\ 8 & 10 & 1 \end{pmatrix}$$

Compute

i) The mean vector $\overline{\mathbf{X}}$ for the data set
ii) The covariance matrix (of **X**) **S**
iii) The correlation matrix (of **X**) **R** [Hint – $\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$, where **D** is a diagonal matrix with diagonal elements of **S**]
iv) The eigenvalues of **S**

The following questions (v – vii) assume that the population covariance matrix $\sum$ equal S.

v) Define the bivariate normal density for X_1 and X_2

vi) Obtain the distribution of $\begin{pmatrix} X_1 - X_2 \\ X_2 - X_3 \end{pmatrix}$

vii) Which of the following variables are independent
   a. X_1 and X_2
   b. X_2 and X_3
   c. (X_1 , X_2) and X_3

18. Explain the following (to the point)
   - Bonferroni simultaneous confidence intervals
   - Wilk's Lambda
19. If $X \sim N_2 (\mu, \sum)$ and the random sample of size (n =3) is as below.

$$X_{3\times 2} = \begin{pmatrix} 6 & 9 \\ 10 & 6 \\ 8 & 3 \end{pmatrix}$$

   (i) Obtain the multivariate statistics for X.
   (ii) What is the sampling distribution of T^2 in this case?
   (iii) Obtain simultaneous confidence interval for μ(take $\alpha = 0.05$)
20. A manufacturer produces light bulbs of a particular type. The duration of continuous blowing in hours and intensity of light in lux are the two important quality characteristics for use. The manufacturer claims the following:
   (i) The average life of the bulbs produced is 2000 hours (continuous blowing), and
   (ii) The average intensity in lux is 100.

As a retailer of this bulb, how do you build a statistical model to test the manufacturer's claim?

21. Define the followings (to the point)
    1. Normality
    2. Homoscedasticity
    3. Linearity
    4. Uncorrelated error terms

22. A sample of 10 observations for $X \sim N_2(\mu, \sum)$ revealed the Mahalanobis D^2 as below:

    0.59 0.81 0.83 0.97 1.01 1.02 1.20 1.88 4.34 5.33

    The χ^2-quantiles are

    0.10 0.33 0.58 0.86 1.20 1.60 2.10 2.77 3.79 5.99

    Assess multivariate normality for the data.

23. The followings are sample data provided by a logistic company on the weight of six shipments, the distances moved, and the damage that was incurred:

Observation no.	1	2	3	4	5	6
Weight ('000 lbs)	5	4	2	1.5	3.4	5.2
Distance ('000 km)	2	2.5	1.3	2.8	1.0	1.9
Damage ('000 Rs)	20	15	10	11	16	23

    (i) Obtain a dependent model using multiple regressions.
    (ii) Obtain 95 % confidence interval for β_1 and β_2.
    (iii) What % of variance of the DV is explained by IVs? Calculate adjusted coefficient of determination and comment on the result.

24. The followings are sample statistics provided by a automobile company on the production, the logistic cost, and the sales volume that was incurred:

$$(X'X)^{-1} = \begin{pmatrix} 5 & 40 & 10 \\ 40 & 360 & 86 \\ 10 & 86 & 30 \end{pmatrix}$$

$$X'Y = \begin{pmatrix} 30 \\ 260 \\ 53 \end{pmatrix}$$

and ANOVA TABLE (incomplete)

Model	**SS**	**DF**	**MS**	**F**	**P**
Regression	**20.989**				
Residual	**9.011**				
Total	**30.000**				

i) Complete the ANOVA Table above.
ii) Obtain a dependent model using multiple regressions and mention the Y variable.
iii) Obtain 95 % confidence interval for β_1 and β_2.
iv) What % of variance of the DV is explained by IVs? Calculate adjusted coefficient of determination and comment on the result.

25. What is parameter testing? For the given data below, obtain a multiple regression equation. Test the parameters of the model and comment on the result.

$$X'X = \begin{pmatrix} 16 & 40 & 200 \\ 40 & 120 & 500 \\ 200 & 500 & 3000 \end{pmatrix}, \quad X'Y = \begin{bmatrix} 723 \\ 1963 \\ 8210 \end{bmatrix},$$

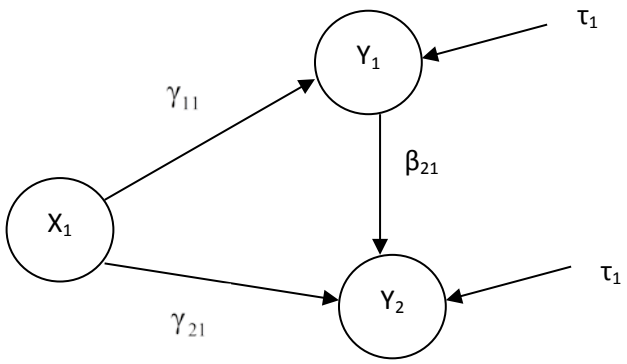
SSR= 2578
SSE= 234
n=16.

26. Define statistical distance. Name a few multivariate models that are based on statistical distance.
27. Evaluate Hotelling's T^2, for testing $Ho: \mu' = [7,11]$, using the data

$$X' = \begin{bmatrix} 2 & 8 & 6 & 8 \\ 12 & 9 & 9 & 10 \end{bmatrix}.$$

Specify the distribution of T^2 for this situation.Test Ho at $\alpha = 0.05$ level. What conclusion do you reach?

28. How do you assess casualty? Is path model a casual model? Enumerate the steps in path modeling.Consider the following path diagram. Obtain normal equations.



29. Define causal and correlation relationships .For the following causal relationships, develop path diagrams and structural equations**.**

i) $X_1, X_2 \longrightarrow Y_1$

ii) $X_1, X_2 \longrightarrow Y_1$

$X_2, Y_1 \longrightarrow Y_2$

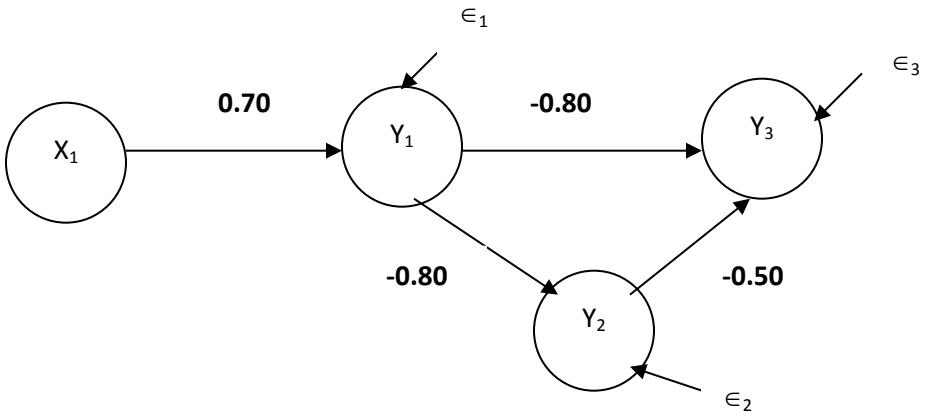
iii) $X_1, X_2 \longrightarrow Y_1$

$X_2, X_3, Y_1, Y_3 \longrightarrow Y_2$

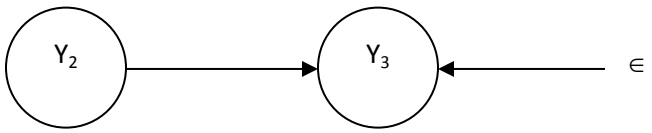
$Y_1, Y_2 \longrightarrow Y_3$

For question (ii), develop normal equations and complete correlation matrix in terms of parameters.

30. Consider the following path model with hypothesized model parameters (standardized).



i) Write out the structural equation for each of the endogenous variables.
ii) Determine the complete correlation matrix.
iii) Decompose the correlation between Y_3 and Y_2 into
    5. Correlation due to direct effects,
    6. Correlation due to indirect effects,
    7. Correlation due to common causes, and
    8. Correlation due to correlated causes,
iv) Suppose the above model is correct, but instead the researcher believed in and estimated the following model:



What conclusion would the researcher likely draw? Discuss the consequences of this misspecification.

31. The following data set is obtained from a study conducted to determine the effect of job stress on work safety perception. Two measurements (p=2), job stress and safety perception were taken on two groups of people with sample size of 19. The mean vectors and covariance matrices are as follows:

For group-1: X-bar = 140 S = 550 210
420 210 330

For group-2: X-bar = 150 S= 480 210
400 250 530

The management claims that there is no significant difference in job stress and safety perception between the two groups of employees. Does the data substantiate the claim at $\alpha = 0.01$?

32. Define principal components (PCs). For the following population covariance matrix obtain PCs. $\Sigma = \begin{pmatrix} 4 & 2 \\ 2 & 9 \end{pmatrix}$

33. Suppose the random variables X_1, X_2 and X_3 have the covariance matrix

$$\Sigma = \begin{pmatrix} 5 & -4 & 0 \\ -4 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

Obtain principal components. Prepare scree plot and determine the number of components to be retained. What are the correlation coefficients between the PCs and the *X* variables? What conclusions can you draw from this?

34. Determine the population principal components Y_1 and Y_2 for the covariance matrix given below and also calculate the proportion of the total population variance explained by the first principal component.

$$\Sigma = \begin{pmatrix} 5 & 2 \\ 3 & 2 \end{pmatrix}$$

35. How many principal components can be extracted when the population covariance matrix is

$$\Sigma = \begin{pmatrix} \sigma^2 & \sigma^2 p & 0 \\ \sigma^2 p & \sigma^2 & \sigma^2 p \\ 0 & \sigma^2 p & \sigma^2 \end{pmatrix}, \quad -\frac{1}{\sqrt{2}} < p < \frac{1}{\sqrt{2}}?$$

Find the proportion of population variance explained by each of the components.

36. What is factor analysis? Differentiate between confirmatory factor model and exploratory factor model.

37. What is factor rotation? Show that $\Sigma = L^* L^{*'} + \psi$, where $L^* = LT$ (L, L^*, T, Σ, ψ represent their usual meaning).

38. Define specific variance, commonality, factor loadings, and factor scores.

39. Show that after orthogonal rotation of factors, the estimated covariance matrix for the observed variables remains unchanged.

40. Suppose the random variables X_1, X_2 and X_3 have the covariance matrix

$$\Sigma = \begin{pmatrix} 5 & -4 & 0 \\ -4 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

Assume m = 1 factor model, calculate the loading matrix L and matrix of specific variances Ψ using principal component solution method.
Calculate communalities and interpret these quantities. What portion of the total variance is explained by the common factor?

41. The following correlation matrix was obtained from a secondary source of data.

Attribute (Variable)	X1	X2	X3	X4	X5
X1	1.00				
X2	0.02	1.00			
X3	0.96	0.13	1.00		
X4	0.42	0.71	0.50	1.00	
X5	0.01	0.85	0.11	0.79	1.00

The SPSS program run shows that the eigen-values for the 5 principal components are 2.85, 1.81, 0.20, 0.10 and 0.03.

- Construct a scree plot.
- How many factors should be retained?
- Compute correlation between factors and attributes.

42. For (41), the un-rotated two-factor solution is as below:

	Taste	Money	Flavor	Snack	Energy
Factor 1	0.56	0.78	0.64	0.94	0.80
Factor 2	0.82	-0.52	0.75	-0.10	-0.54

Obtain a rotated (orthogonal) factor solution (loadings) and name the factors.

43. The following data shows un-rotated factor loadings for two factors with five variables. Obtain rotated factor loadings (**must use graph paper**) and comment on the result.

	Unrotated factor loadings	
Variables	F_1	F_2
X_1	0.50	0.80
X_2	0.60	0.70
X_3	0.90	-0.25
X_4	0.80	-0.30
X_5	0.60	-0.50

44. Write down the general equations for structural equations modeling. Show that the variance-covariance matrix of the indicator variables in structural equations model is [the terms represent their usual meanings]:

$$\sum_{(YX(YX)')} = \begin{pmatrix} \Lambda_Y \Sigma_{\eta\eta'} \Lambda'_Y + \theta_\varepsilon & \Lambda_Y \Sigma_{\eta\xi'} \Lambda'_X \\ \Lambda_X \Sigma_{\xi\eta'} \Lambda'_Y & \Lambda_X \phi \Lambda'_X + \theta_\delta \end{pmatrix}$$

45. An engineer-in-charge of shop floor production is facing a problem of poor productivity and wants to optimize the performance of his workforce. He realized that working environment and workforce expertise have strong influence on workers' performance. On further enquiry, he established that workers' performance solely in term of productivity measure is not exhaustive and scrap ratio (for quality) and number of injuries (for safety) may well be other indicators of performance. The shop floor has problems of heat and humidity, vibrating equipment while well ventilated and they vary from season to season. The workforce varies in education, skill, and experience. Develop a suitable conceptual model to study this situation for identifying influencing factors and their contributions. Also mention the mathematical formulation of the model, if any.

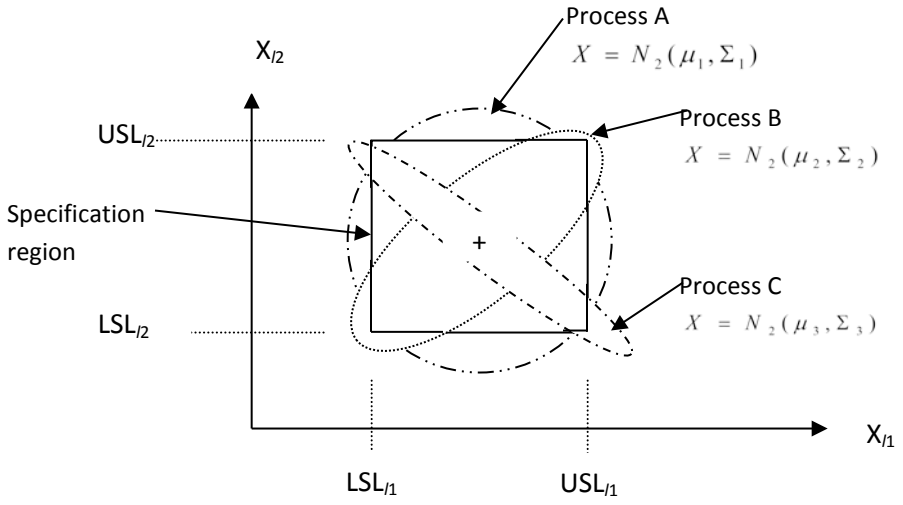
46. In testing overall fit of any multivariate models generally two types of fit measures are used. What are these? Define them. Give examples with respect to SEM.

47. As a manufacturer of heavy earth moving machinery, your company requires some parts of the machine to be purchased from different suppliers. Amongst the purchased items, 'boom assembly' is the most important part. The key characteristics of the 'boom assembly' are vertical (X_1) and horizontal reach (X_2), angle of rotation (X_3), and load bearing capacity (X_4). Consider the case as a vender selection problem and obtain statistical models for selecting the best vendor using:

    a) Pair-wise comparison of vendors, and
    b) Comparison amongst all the vendors simultaneously.

48. What do you mean by confidence region (CR)? The figure given below shows the performance status (99.73% CR) of three different processes (l = 1, 2, 3) characterized by

$X \sim N_2(\mu_l, \Sigma_l)$.



Based on the diagram, comment on the following:

(i) Comparison of mean vectors for the processes A, B and C,
(ii) Comparison of process variability for the processes A, B and C,
(iii) Comparison of correlation structure of the variables for the processes A, B and C, and
(iv) What proportion of products produced by each of the processes is nonconforming (defective)?

49. What ways MANOVA differs from ANOVA? What statistic is commonly used in testing hypothesis in MANOVA? Define it.

50. The following observations on two responses are collected for three treatments.

The observation vectors $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ are

Treatment 1: $\begin{bmatrix} 4 \\ 6 \end{bmatrix}, \begin{bmatrix} 3 \\ 7 \end{bmatrix}, \begin{bmatrix} 7 \\ 5 \end{bmatrix}, \begin{bmatrix} 4 \\ 8 \end{bmatrix}, \begin{bmatrix} 5 \\ 8 \end{bmatrix}$

Treatment 2: $\begin{bmatrix} 4 \\ 5 \end{bmatrix}, \begin{bmatrix} 2 \\ 7 \end{bmatrix}, \begin{bmatrix} 2 \\ 5 \end{bmatrix},$

Treatment 3: $\begin{bmatrix} 3 \\ 6 \end{bmatrix}, \begin{bmatrix} 6 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 3 \\ 4 \end{bmatrix}.$

Construct the one-way MANOVA Table.