Basic Statistics
~~Jutpas~~ Jan. 2016

Department of Statistics
University of Mumbai
PGDASS 2015-2016
Paper I: Basic Statistics

Date: 4th January 2016 **Total Marks: 60**

Time: 11.00 a.m. to 2.00 p.m.

N.B. : 1) All questions are compulsory.
2) Simple nonprogrammable calculators are allowed.
3) Figures to the right indicate full marks.

1. (a) The following data give the time (in minutes) that each of 20 students took to complete a statistics test. (04)

55	49	53	59	38	56	39	58	47	53
58	42	37	43	44	55	51	46	45	47

(i) Construct a stem-and-leaf display for these data.
(ii) Determine the mean, median , quartiles and mode of this data.

(b) A card is chosen at random from a deck of 52 cards. Let A be the event that the card is king and B the event that it is a heart . Are A and B independent? (03)

(c) The probability that any given person is allergic to a certain drug is 0.03. What is the probability that none of three randomly selected persons is allergic to this drug? Assume that all three persons are independent. (03)

2. (a) An unbiased coin is tossed until a head is appeared. If X denotes the number of tosses required to get a head find the MGF and PGF of X. (03)

(b) Suppose that X has following p.d.f. (04)

$f(x) = k\, e^{-5x}$, $x > 0$,

= 0 , *otherwise*

(i) Find k .
(ii) Find $E(X)$ and $V(X)$.
(iii) What is the name of this distribution?

(c) On average a household receives 9.5 telemarketing phone calls per week. Find the probability that on a particular day a household receives exactly five calls. (03)

3. (a) The table below shows a bivariate probability distribution for two discrete random variables X and Y. (05)

X \ Y	1	2
0	0.15	0.05
1	0.20	0.15
2	0.25	0.20

(i) Obtain marginal p.m.f. of X and Y.

(ii) Find conditional p.m.f. of X given Y=2.
$E(X|Y=2)$ and $V(X|Y=2)$

(iii) Are X and Y independent?

(b) Suppose X and Y have joint density function given by (05)

$$f(x,y) = \frac{1+xy}{4} \quad , \quad |X|<1, |Y|<1$$

$$= 0 \quad , \quad otherwise$$

(i) Examine whether X and Y are independent random variables.

(ii) Obtain $E(X)$ and $E(Y)$.

4. (a) Suppose $x_1, x_2, ..., x_n$ is a random sample from Poisson distribution $P(\lambda)$. Obtain $\hat{\lambda}$ M.L.E. of λ. Is it unbiased? (06)

(b) Obtain M.L.E. of $P(X=1)$. Also Obtain asymptotic distribution of $\hat{\lambda}$. (04)

5. (a) Yields of 10 strawberry plants in an uniformity trials are given below. Test the hypothesis that average yield of strawberry plant is 205 against the alternative that it is not equal to 205 at 5% level of significance. (05)

Plant No	1	2	3	4	5	6	7	8	9	10
Yield in gms	239	176	235	217	234	216	318	190	181	225

(b) Following data gives BMR of a group of patients. Assume that the data follows $N(\mu,\sigma^2)$ distribution. (05)

0.79 0.76 0.61 0.80 0.68 0.43

Test the hypothesis that $H_0:\sigma^2=0.10$ against the alternative that $H_1:\sigma^2\neq 0.10$ at 5% level of significance. Also obtain 95% Confidence Interval for σ^2.

6. (a) Following table gives for a sample of married women , level of education and marriage adjustment score: (05)

Level of Education	Marriage Adjustment Score			
	Very low	Low	High	Very High
College	24	97	62	58
High School	22	28	30	41
Middle School	32	10	11	20

Can it be concluded from above data that the level of education is associated with degree of adjustment in marriage?

(b) In a small town, a sample of 200 individuals was selected. It contained 113 literates. Obtain 95% Confidence interval for proportion of literates in the population. (05)

****

Department of Statistics
University of Mumbai
PGDASS 2015-2016
Paper II: Marketing Research

Date: 06/01/2016 **Total Marks: 60**

Time: 11.00 a.m. to 2.00 p.m.

N.B. : 1) All questions are compulsory.
2) Simple nonprogrammable calculators are allowed.
3) Figures to the right indicate full marks.

SECTION I

Note: Solve any two questions.

1. What is the objective of Exploratory Research and what are the techniques to collect data in it? (15)

2. What are the scales one can use to understand consumers' attitude? (15)

3. What are various techniques to isolate the effects of confounding variables in causal research? (15)

SECTION II

4. The survey is conducted among McD customers to understand what attracts them to visit the store. The regression is run between frequency of visit and the four attributes. From the output below, explain which attributes are important in driving the no. of visits to McD store. (05)

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	2.48	0.15		17.05	0.00		
	They use the freshest ingredients	0.10	0.03	-0.03	6.86	0.00	0.67	1.49
	It is my favorite restaurant brand	0.29	0.03	0.34	8.64	0.00	0.19	5.10
	They have great tasting food	0.24	0.04	-0.05	7.12	0.00	0.73	1.38
	You can eat inexpensively there	0.05	0.03	0.06	2.38	0.02	0.81	1.23

4 - 5
5 - 10
6 - 5 7 - 10

a Dependent Variable: FREQ

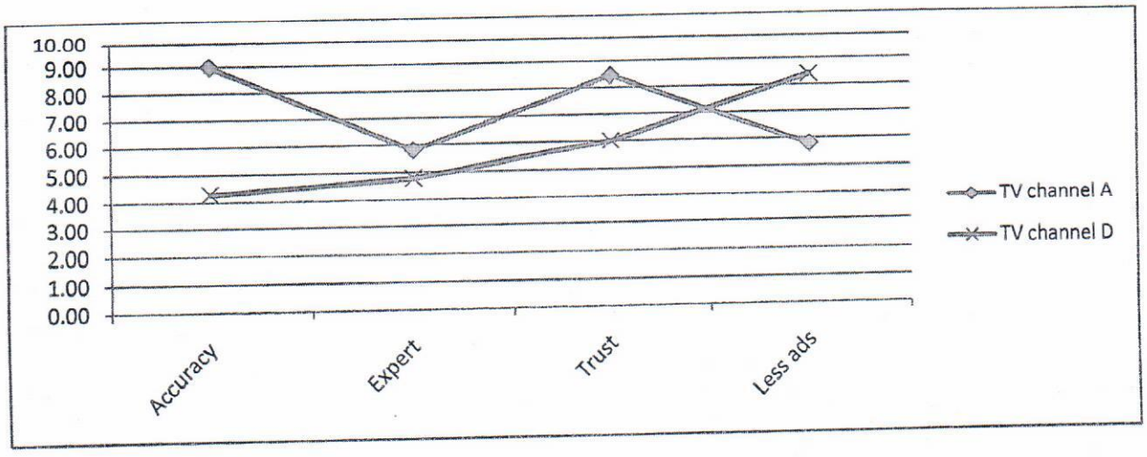
5. Listed are 8 questions taken from a survey of insurance agency to understand consumers' perception about insurance. Factor analysis is run on these 8 attributes and 3 factors are extracted. Un-rotated component matrix is given below. Looking at this loading matrix, suggest whether there is need of rotation? Also, explain difference between Varimax and Quartimax rotation. (10)

Component Matrix(a)

	Component		
	1	2	3
Having insurance means I don't have to worry about money in the future	0.7743	0.0630	0.0566
Having insurance makes me feel empowered	0.7543	0.2557	0.0576
Insurance helps me feel in control of my finances	0.7503	0.2537	0.1784
Insurance allows me to ensure the security of my family's future	0.3253	0.7895	0.2044
Insurance helps me protect my family	0.3510	0.7661	0.2308
Insurance gives my family financial security	0.5694	0.5775	0.1171
I want to protect my way of life should something happen to me	0.3576	0.1802	0.8164
I don't want to be a burden to my family when I die	0.0569	0.2005	0.7470

Extraction Method: Principal Component Analysis.

6. The survey is conducted among the TV news channels. Following snake chart explains the performance of two channels on four important factors. What inference can you draw from below table? (05)



6. Following is the output of cluster analysis among soft drink consumers. What conclusion company should draw based on it? (10)

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
As a treat/reward	56.45	3	0.466561	1365	120.9916	0.00
Goes well with food/snack	2.43	3	0.331935	1365	7.320703	0.00
Good to serve to guests	6.45	3	0.346102	1365	18.63611	0.00
Has new flavours/tastes	31.01	3	0.315791	1365	98.19793	0.00
Helps me relax/unwind	23.34	3	0.352838	1365	66.14942	0.00
Is fun to consume	3.27	3	0.355425	1365	9.200247	0.00
Re-hydrates	5.15	3	0.363315	1365	14.17505	0.00
To aid digestion	2.23	3	0.3343	1365	6.67065	0.00
To quench thirst	10.33	3	0.283188	1365	36.47749	0.00
To reduce stress	3.1	3	0.314654	1365	9.85208	0.00
Wanted something tasty	5.19	3	0.30584	1365	16.96966	0.00

Number of Cases in each Cluster

Cluster	1	642.000
	2	156.000
	3	494.000
	4	77.000
Valid		1369.00
		0
Missing		.000

****

Department of Statistics
University of Mumbai
PGDASS 2015-2016
Paper III: Regression and Linear Models

Date: 8th January 2016 **Total Marks: 60**

Time: 11.00 a.m. to 2.00 p.m.

N.B. : 1) All questions are compulsory.
2) Simple nonprogrammable calculators are allowed.
3) Figures to the right indicate full marks.

1. (a) A test is run on a given process to study effect of independent variable x on y. Twenty observations are taken and the following results are obtained. (08)

$\bar{x} = 5.0$, $\Sigma(x_i - \bar{x}) = 160.0$, $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 80.0$
$\bar{y} = 3.0$, $\Sigma(y_i - \bar{y})^2 = 83.2$
Assume simple regression model $y = \beta_0 + \beta_1 x + e$.
(i) Calculate fitted regression equation.
(ii) Prepare analysis of variance table.
(iii) Determine correlation coefficient.

(b) How do you determine whether response variable and predictor variable are correlated? (02)

2. (a) If there is one response variable, p predictor variables and n observations, show the data set in tabular form. (02)

(b) In a multiple regression with two predictor variables and single response variable, data on ten observations gives Error S.S.=0.32513, Total S.S.=2.28010. Find Multiple correlation coefficient . How do you test significance of predictor variables? (05)

(c) What is hat matrix? Express fitted values in terms of elements of hat matrix and observed responses. (03)

3. (a) How do you model 0 or 1 dependent variable using logistic regression? How do you infer about effects of predictor variables on odds ratio using confidence intervals? (05)

(b) Explain how principal components analysis can be used to detect multicollinearity. Check the presence of multicollinearity if the eigen values of correlation matrix are: $\lambda_1 = 2.952, \lambda_2 = 0.040, \lambda_3 = 0.008$. (05)

4. A simple experiment is conducted to study the differences in the strain reading (the response) of their different heads on each of the five different machines. Since each head was observed on only one machine, the heads were nested within machines, giving twenty heads in totals. Four observations were taken on each head. An incomplete Analysis of variance table obtained in MINITAB is given below. Complete the table and answer the following questions, (10)

i) Write the statistical model.
ii) Write the assumptions of the model.
iii) Write the hypotheses.
iv) Identify the significant factors.
v) Write the conclusion.
vi) Rewrite the ANOVA table by ignoring in-significant factor.

Analysis of variance table:

Source	DF	SS	MS	F	P
machine	-	45.0750	11.2688	-	0.670
head	15	-	18.8583	1.762	0.063
Error	-	642.0000	-		
Total	79	969.9500			

5. Explain by specifying statistical model, hypotheses to be tested and assumptions in the two-way analysis covariance model. What will happen if coefficient of covariance in the model is not significant? (10)

6. The following table gives the observed values of response variable for three levels of a factor A. (10)

Factor levels	A_1	A_2	A_3	A_4	A_5
Response values:	25, 30, 26, 36	02, 10, 23, 15	10, 20, 19, 04	07, 10, 15, 18	05, 02, 07

Use Tukey's test to identify the factor level which gives maximum response. $q_{0.05}(5,14) = 4.41$.

****

Department of Statistics
University of Mumbai
PGDASS 2015-2016
Paper IV: Decision making and Forecasting

Date: 12th January 2016 **Total Marks: 60**

Time: 11.00 a.m. to 2.00 p.m.

N.B. : 1) All questions are compulsory.
2) Simple nonprogrammable calculators are allowed.
3) Figures to the right indicate full marks.

SECTION I

1. a. Explain the circumstances under which only qualitative methods have to be used for forecasting. (10)

   b. Describe the steps involved in Delphi method of forecasting and illustrate its uses with the help of any real life situation known to you.

2. Which of the following five statements are correct? Mention your opinion as 'CORRECT' or 'WRONG' against each statement and give detailed reasons to support your opinion (10)

   i. Forecast of aggregates is less reliable than the aggregate of Forecast.

   ii. Leading Indicator model is ideal when cause factor bears a lagged relationship with effect factor

   iii. In classical Decomposition Model of Time Series, it is safer to employ Additive model than the Multiplicative Model .

   iv. Structured decisions are taken in situations that are unclear or one-shot.

   v. Multi-colinearity will not affect the ability of the model to predict.

3. a. Describe the process of "Six Thinking Hats Technique" of Decision Making and explain its rationale. (10)

   b. Apply this technique to a software development company which wants to make long term decisions about undertaking unrelated diversification projects.

SECTION II

4. Explain the term 'inventory' and give reasons for carrying inventories. (10)
Define the following :

a. Set up cost
b. Holding cost
c. Shortage cost

5. For a fixed order quantity system find out the E.O.Q , optimum buffer stock, normal lead time consumption and reorder level with the following data; (10)

Annual consumption in units	= 5,000
Cost in Rs. Of one unit	= 1.5
Set up cost in Rs. per Production Run	= 13
Holding cost in Rs.per unit	= 0.22

Past lead times = 15 days, 25 days, 13 days, 14 days, 30 days & 17 days.

SECTION III

6. Guide Auto Ltd. manufactures and sells three products to the automobile industry. All the products must pass through a machining process, the capacity of which is limited to 20,000 hours per annum both by equipment design and government regulation. (10)

The following additional information is available:

Particulars	Product X	Product Y	Product Z
Selling Price Rs./unit	1,900	2,400	4,000
Variable Cost Rs./unit	700	1,200	2,800
Machining requirement Hrs/units	3	2	1
Maximum Possible sales – Units	10,000	2,000	1,000

Required:Statement showing the best possible production mix, which would provide the maximum profit for Guide Auto Ltd. together with supporting workings.

****

(1) V Six Sigma and statistical Process control Jan. 2016

N.B.: (1) All questions are compulsory.
(2) Use of Simple calculator is allowed
(3) Figures to the right indicate full marks.
(4) Answer to the both section should be written in separate answerbooks.

SECTION I

Instruction:

1. Use of Statistical Software and Excel along with six sigma excel template is allowed.
2. Write down null and alternative hypothesis for the test with name/s of the test/s used, basis of test and conclusion along with output of session window.

1. Productivity in terms of handling the customer in min per customer for two cashiers of Andheri Branch is given below. Who is better & consistent? What Target should be assigned? How many observations we must gather to decide their base lines to estimate in 1 hour with 95% Conf. level (10)

Cashier A	14	18	21	12	15	23	20	18	15	16	24	45	48	49
Cashier B	23	21	43	12	18	34	21	27	14	21	16	14	21	19

2. Three diff. product samples distributed to 3 diff. QA Inspectors to check quality level. Comment on their measurement system. Also explain in detail the output session window & graph window. (05)

Sample No	1	1	1	1	1	1	2	2	2
QA Inspector	1	2	3	1	2	3	1	2	3
Trial No.	1	1	1	2	2	2	1	1	1
Quality Level	150	132	157	141	142	165	110	100	125
Sample No	2	2	2	3	3	3	3	3	3
QA Inspector	1	2	3	1	2	3	1	2	3
Trial No.	2	2	2	1	1	1	2	2	2
Quality Level	119	110	118	195	220	205	198	115	200

3 Explain in detail (05)
a. Measurement System Analysis
b. DOE Terms & Minitab method of Solving

4 Following 3 factors determine the crop yield. Conduct DOE & identify the factors affecting yield. (10)

Run Order	Factor 1	Factor 2	Factor 3	Crop Yield Kg/ Acre
1	3	40	100	92.300
2	1	40	100	96.020
3	1	40	80	96.670
4	1	40	100	96.020
5	3	30	100	74.010
6	1	30	80	76.000
7	3	40	80	74.000
8	3	40	100	94.000
9	1	30	100	76.000
10	3	40	80	94.000
11	3	30	100	75.000
12	1	40	80	97.000
13	3	30	80	73.000
14	3	30	80	74.000
15	1	30	80	76.000
16	1	30	100	75.000

SECTION II

5. (a) Students have to do a project as part of their final semester. A group of 6 students decided to visit a manufacturing unit and collect their quality data for analysis and use it for the project. It involved following activities. (05)

Srl	Activity	Time required in Days
1	Permission of the managing director of the manufacturing unit	8
2	Booking of ticket for travel	10
3	Collection of data at the unit	2
4	Analysis of data set 1 with 2 students	3
5	Analysis of data set 2 with 2 students	5
6	Analysis of dataset 3 with 2 students	4
7	Final analysis based on analysis of data set 1, 2 and 3	2
8	Writing of project Report	3
9	Printing 3 copies of the report by professional printer	2
10	Submission of the project	1

1. Draw activity diagram for these activities.
2. Find critical path
3. If the project submission date is 1st July then when should travel tickets be booked latest?

(b) Suggest appropriate QC tool that can be used for following problems (05)

Srl	Problem
1	A quality department team use to plan and agree with all departments the tasks for implementing a new quality management system. It encounters dependencies of various tasks and arrives at minimum time required for implementation.
2	University is facing an issue of poor attendance in class and decides to involve professors from various departments to find out root cause of this and implement new policy which will benefit students in their learning.
3	A washing powder has different efficiencies at achieving 'softness' and 'stain removal' in garments made of acrylic, polyester, wool and various fiber mixtures. If similar affects are found in a group of fibers, then changing the powder ingredients may affect the whole group in a similar way.
4	BMC studding whether there is any relationship between increase in property rates across various zones of city and increase in slums.
5	A firm of consulting engineers wants to ensure that all eventualities and their relations are covered in an investigation report into the laying of a new cross-country gas pipeline.

6. (a) A sample survey was done of people who attended a lecture on investments. It has been entered in following excel sheet. Using this sheet answer following questions. (05)

	A	B	C	D	E	F	G
1	Emp No	Name	DOB	Designation	Department	Salary	Bonus
2	101	X1	05-02-1970	Manager	Sales	120000	30000
3	102	X2	03-05-1976	Worker	Admin	20000	2000
4	103	X3	14-04-1986	Officer	Sales	80000	8000
5	104	X4	20-01-1970	Manager	Admin	100000	15000
6	105	X5	13-04-1971	Manager	HR	80000	12000
7	106	X6	07-08-1988	Worker	Production	30000	7500
8	107	X7	11-12-1981	Worker	Production	25000	5000
9	108	X8	27-10-1986	Worker	Production	15000	3750
10	109	X9	03-03-1981	Officer	Admin	75000	15000
11	110	X10	12-05-1976	Manager	Accounts	150000	15000
12	111	X11	25-04-1980	Officer	HR	55000	13750
13	112	X12	20-12-1990	Officer	HR	40000	10000
14	113	X13	23-01-1976	Worker	Sales	25000	6250
15	114	X14	10-12-1990	Officer	Production	35000	5250
16	115	X15	29-08-1977	Worker	HR	10000	1500
17	116	X16	31-12-1981	Worker	Sales	40000	6000
18	117	X17	06-11-1978	Worker	Production	15000	3000
19	118	X18	04-10-1988	Worker	Production	18000	4500
20	119	X19	01-09-1990	Officer	Production	45000	4500
21	120	X20	02-03-1975	Manager	Production	90000	13500

1. Write function to calculate number of Workers in the organisation.
2. Write function to calculate total expenditure on employees

3. Write function to calculate age of employee x9
4. Write function to calculate number of employees whose salary is greater than or equal to Rs 100000
5. Write function to calculate total salary paid to employees in the Production department

(b) A case study has two 2-level and three 3-level factors. Calculate degrees of freedom. Refer following table of standard orthogonal arrays and determine which orthogonal array can be used for this case study? What technique can be used to reduce number of experiments in this case? With this modified method, which orthogonal array can be used? Explain your answer and draw a table of experiment layout. (05)

Orthogonal Arrays	Number of Rows	Maximum Number of Factors	Maximum Number of columns at these levels			
			2	3	4	5
L4	4	3	3	-	-	-
L8	8	7	7	-	-	-
L9	9	4	-	4	-	-
L12	12	11	11	-	-	-
L16	16	15	15	-	-	-
L'16	16	5	-	-	5	-
L18	18	8	1	7	-	-
L25	25	6	-	-	-	6
L27	27	13	-	13	-	-
L32	32	31	31	-	-	-
L'32	32	10	1	-	9	-
L36	36	23	11	12	-	-
L'36	36	16	3	13	-	-
L50	50	12	1	-	-	11
L54	54	26	1	25	-	-
L64	64	63	63	-	-	-
L'64	64	21	-	-	21	-
L81	81	40	-	40	-	-

L9	**Factors**			
Expt. No	A	B	C	D
1	A1	B1	C1	D1
2	A1	B2	C2	D2
3	A1	B3	C3	D3
4	A2	B1	C2	D3
5	A2	B2	C3	D1
6	A2	B3	C1	D2
7	A3	B1	C3	D2
8	A3	B2	C1	D3
9	A3	B3	C2	D1

7. (a) An airlines company would like to analyze and prioritise the quality complaints received from its customers. The complaint data is as below: (05)

Type of complaint	Number
Baggage delay	23
Missed connections	15
Lost baggage	7
Poor cabin service	3
Ticketing error	2

Which QC tool can be used to represent this? Use appropriate QC tool, draw appropriate analysis and write your findings.

(b) Cold drink manufacturing company is assessing risks involved in it's launch of new product, which involves significant investment and would like to ensure that top risks are dealt with appropriate counter measures that will minimise the loss in case of risk occurrence. (05)

i. Following are the risks identified with it's corresponding probability of risk occurrence. Calculate risk exposure and find out the most serious risk.

Risk	Probability of occurring	Loss if risk occurs
Product recall situation	2%	80,000
Significant product rejection	0.1%	1,000,000
Competitive strike	10%	25,000

ii. Following are the countermeasure against the most serious risk with it's corresponding costs and new probabilities of reduced risks. Find out new risk exposure and risk reduction leverage.

Countermeasure	**Total Cost**	**New Risk Probability**	**New Total Loss**
Advertising Campaign	40,000	3%	5,000
Price Promotions	30,000	5%	10,000
Simultaneous Launch	10,000	8%	15,000

iii. Find out the cost effective countermeasure

****

VII Jan. 2016.

N.B. (1) All Questions are compulsory. (3 Hours)
(2) Figures to the right indicate marks.
(3) Calculators are allowed.

1. Describe application of following tests/statistical methods in clinical research studies (any 4 of 5 – 2.5 marks each) (10)
   a. Wilcoxon Signed rank Test
   b. Fisher's Exact Test
   c. ANCOVA
   d. Logistic Regression
   e. Multiplicity Adjustment

2. Describe information required for sample size calculation in a clinical trial. (10)

3. Answer following questions (5 marks each)
   a. Briefly describe use of Odds Ratio and Relative Risk in Clinical Trials (10)
   b. Briefly describe Application of Randomization and Blinding in Clinical Trials

4. Describe contents of a statistical analysis plan for a clinical trial. Also discuss considerations involved in writing / reviewing the statistical analysis Plan. (10)

5. Please read the following scenario and answer questions mentioned below. (10)
   A sponsor wants to demonstrate that their new drug "Novex" is better than existing available drug "Refrex" for treatment of pneumonia due to Methicillin Resistant Staphylococcus Aureus (MRSA).

   Hence they are planning to conduct a clinical trial where patients with MRSA pneumonia will be randomly assigned (1:1 proportion) to receive "Novex" or "Refrex" for 14 days. Primary endpoint of study is achievement of cure. Cure is defined as complete eradication of the infectious organism (MRSA) within 14 days after start of treatment. If organism persists, then treatment would be considered to have failed for respective subject.

   Sponsor has approached you for statistical support.

   A. What information will you need to calculate sample size for this study.[3 marks]
   B. How will you write the statistical hypothesis for this study [3 marks]
   C. Discuss statistical methods that will be used for testing sponsor's objectives in this study. [3 marks]

6. Please answer following multiple choice questions (by indicating on your answer sheet) question number followed by option that represents the correct answer (10)
   (e.g. Mention 6a : (i) for question 6a if option numbered (i) is the correct answer in your opinion) [2 Marks Each]

a. Which of the following statement about Data Step is not true?

(i) A WHERE statement is a valid statement inside a Data Step.

(ii) A data step can be executed without an input dataset in set statement.

(iii) A data step can use DO LOOPs to iteratively process observations.

(iv) A data step can never contain an OUTPUT statement.

b. Which of the following statement correctly checks condition "A" is greater than or equal to "B" and that both are not missing?

(i) Where A <= B > . ;

(ii) Where A >= B gt . ;

(iii) Where A >= B and not missing;

(iv) None of the above.

c. Which of the following is a syntactically correct PROC SQL statement?

(i) PROC SQL; OUTPUT * FROM A; QUIT;

(ii) PROC SQL; UPDATE * FROM A; QUIT;

(iii) PROC SQL; SELECT * FROM A; QUIT;

(iv) PROC SQL; CREATE * FROM A; QUIT;

d. Which of the following is a correct statement about SAS data step /procedures?

(i) Proc Freq can produce standard deviation but not odds ratio.

(ii) Proc TTEST cannot produce box plots.

(iii) Proc GLM can be used for Analysis of Covariance.

(iv) Proc Univariate cannot perform Test of normality.

e. If Data AA has 16 rows, which of the following SAS data step WILL NOT create data BB with 32 rows?

(i) DATA AA; SET BB BB; RUN;

(ii) DATA AA; SET BB; OUTPUT; OUTPUT; RUN;

(iii) DATA AA; SET BB; SET BB; RUN;

(iv) DATA AA; SET BB; DO i = 1 to 2; OUTPUT; END; RUN;

****

multivariate techniques Jan. 2016

N.B.: (1) All questions are compulsory.
(2) Use of Simple calculator is allowed
(3) Figures to the right indicate full marks.
(4) Answer to the both section should be written in separate answerbooks.

1. (a) Explain the use of Descriptive statistics in summarizing data having multivariate observations? (04)

(b) Explain briefly the spectral decomposition of a matrix. (02)

(c) Let $X_{4\times1}$ have mean vector (04)

$\mu = (1 \quad 5 \quad 2 \quad 4)'$ and variance-covariance matrix.

$$\Sigma = \begin{bmatrix} 2 & 0 & 3 & 0 \\ 0 & 1 & 5 & 1 \\ 3 & 5 & 4 & -1 \\ 0 & 1 & -1 & 1 \end{bmatrix}$$

Obtain the mean and variance of A X.
where $A = (2 \quad 1 \quad 3 \quad -1)$ Also obtain the correlation matrix ρ for X.

2. (a) Let $X_{3\times1}$ be a random vector with mean vector $\mu_{3\times1}$ and covariance matrix Σ. A random sample of size 4 is selected as (05)

$$\begin{bmatrix} 2 & 3 & 4 & 1 \\ 6 & 1 & 5 & 2 \\ 4 & 2 & 4 & 2 \end{bmatrix}$$

Obtain the unbiased estimates of μ and Σ. Obtain generalized sample variance and interpret.

(b) State the p.d.f. of the multivariate normal distribution and state some properties. (03)

(c) (02)

Let $X \sim N_3$ with $X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 5 & 2 & 0 \\ 2 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix}$

Discuss independence of

(i) X_1 and X_2

(ii) (X_1, X_2) and X_3.

3. (a) A random sample of size 5 is selected from $N_3(\mu, \Sigma)$ (03)

$$X=\begin{bmatrix}3 & 5 & 2 & 1 & 4\\ 6 & 3 & 7 & 4 & 5\\ 8 & 4 & 5 & 6 & 7\end{bmatrix}$$

Obtain MLE of μ and Σ.

(b) Explain the method of using Q-Q plots to test normality of given data. (04)

(c) (03)

Let $X \sim N_3(\mu, \Sigma)$ with $\mu=\begin{bmatrix}5\\2\\3\end{bmatrix}$ and $\Sigma=\begin{bmatrix}1 & 1 & 1\\1 & 3 & 2\\1 & 2 & 2\end{bmatrix}$

Let $X=\begin{bmatrix}X^{(1)}\\ \hline X^{(2)}\end{bmatrix}=\begin{bmatrix}X_1\\ \hline X_2\\X_3\end{bmatrix}$

Obtain the distribution of $X^{(1)}$ given $X^{(2)}=\begin{bmatrix}4\\5\end{bmatrix}$

4. (a) Explain how the Hotelling's T^2 statistic is used to test the equality of vector means from two independent multivariate normal populations with same variance-covariance matrix Σ. (05)

(b) Explain the regression method of modeling relationships between m response variables $Y_1, Y_2, ..., Y_m$ and single set of predictor variables $X_1, X_2, ..., X_r$. (05)

5. (a) Write down the steps in agglomerative hierarchical clustering algorithm for grouping N objects. (05)

(b) If $Y_1, Y_2, ..., Y_p$ denote the principal components of $\Sigma = V(X)$ where $X' = (X_1, X_2, ..., X_p)$. State (i) Y_2 (ii) $V(Y_2)$ (iii) cov (Y_1, Y_3) (iv) ρ_{Y_2,X_3} (v)proportion of variance explained by Y_1, Y_2, Y_3. (05)

6. (a) If π_1 and π_2 are two Bivariate populations where $\pi_1 \sim N_2(\mu_1, \Sigma)$ and $\pi_2 \sim N_2(\mu_2, \Sigma)$. Where $\mu_1=\begin{bmatrix}10\\15\end{bmatrix}$, $\mu_2=\begin{bmatrix}10\\25\end{bmatrix}$, $\Sigma=\begin{bmatrix}18 & 12\\12 & 32\end{bmatrix}$ (06)

(i) Compute the distance between π_1 and π_2.

(ii) Compute the linear discriminant function

(b) Explain the different methods to decide the number of factors in factor analysis. (02)

(c) Write a short note on MDS. (Multi Dimensional Scaling). (02)

****