Basic statistics July 2015

(3 Hours) [Total Marks: 60

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 two sections should be written on separate answer-books.

SECTION I

1. (a) The following data set gives the number of years for which 24 workers have been (07)
      with their current employers.
      15, 12, 9, 10, 5, 12, 3, 7, 16, 13, 11, 4, 3, 8, 7, 14, 11, 8, 4, 13, 2, 18, 6, 9
      Construct stem and leaf display for the data. Calculate mean, variance, Q_1, Q_2, Q_3
      and quartile deviation for the data.

    (b) A, B, C each fire one shot at a target. The probabilities of their hitting the target (03)
      are respectively given by 0.3, 0.25 and 0.1. If one bullet is found in the target
      find the probability that it came from A's gun.

2. (a) A card is chosen at random from a deck of 52 cards. Let A be the event that the (03)
      card is queen and B the event that it is diamond. Are A and B independent
      events?

    (b) On average a household receives 6 telemarketing phone calls per week. Find the (04)
      probability that a randomly selected households receives
        i)   Exactly 5 calls during a given week.
        ii)   At most 4 calls during a given week.
        iii) At least 2 calls during a given week.

    (c) The probability that a farmer is in debt is 0.8. What is the probability that 3 (03)
      randomly selected farmers are all in debt?

3. (a) Suppose the r.v. X and Y have following joint p.m.f. (04)

Y \ X	0	1	2
1	0.1	0.1	0
2	0.1	0.1	0.2
3	0.2	0.1	0.1

        (i)   Find marginal p.m.f. of X and Y.
        (ii) Find conditional p.m.f. of Y given X=2.
        (iii) Are X and Y independent? Why or why not?
        (iv) Find E(X) and E(Y)

    (b) Suppose that X has following p.d.f. (03)
      $$f(x) = K\,e^{-2k} \quad , x > 0$$
      $$\quad\quad = 0 \quad\quad , \text{otherwise}$$
        (i)   Find k.
        (ii) Find m.g.f. of X and hence find E(X).

    (c) Find p.g.f. of the binomial r.v. with parameters n=10 and p=0.3. (03)

[TURN OVER

SECTION II

4. (a) Define unbiased estimator. Show that sample variance is an unbiased estimator of population variance. (05)

   (b) 1, 0.9, 1.5, 0.8, 0.96, 1.01, 1.3, 0.80 is a random sample from the following distribution $f(x,\theta) = \frac{1}{\theta}e^{-x/\theta}$; $x \geq 0, \theta > 0$. Obtain moment estimator of θ. (05)

5. (a) $x_1, x_2, \ldots, x_n$ is a random sample from $N(0, \sigma^2)$. Obtain MLE of σ^2. Also obtain MLE of P(X>1). (06)

   (b) The following table shows the number of persons in a random sample of 210 listed according to the day of the week on which they prefer to do their grocery shopping. (04)

Day	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
No. of Persons	9	17	12	26	35	69	41

   Using $\alpha = 0.05$, test the null hypothesis that the proportion of persons who prefer to do their shopping on a particular day is same for all days of the week.

6. (a) Following table gives the blood pressures of seven adults before and after the completion of a special dietary plan. (05)

Before	210	180	195	220	231	199	224
After	193	186	186	223	220	183	233

   Let μ_d: mean of the population difference.
   Tests the hypothesis $H_0: \mu_d = 0$ against $H_1: \mu_d > 0$ with $\alpha = 0.05$ (d=before-after)

   (b) Consider the data on income distribution for households in two cities A and B. Test the hypothesis that distribution of households with regard to income levels is similar for both cities. Take $\alpha = 0.05$. (05)

	A	B
High income	70	34
Medium income	80	40
Low income	100	76

****

QP Code : 18475

(3 Hours)   [Total Marks: 60

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 sections should be written on separate answer-books.

SECTION I

1.   Explain overview of syndicated services.   (10)

2.   Explain stratified random sampling with merits and limitations.   (10)

3.   Design a questionnaire for employee satisfaction survey in an organization.   (10)

SECTION II

4. (a) How do we use factor analysis in attribute based perceptual mapping?   (05)
        What is the advantage of attribute based of perceptual mapping over non-attribute based method?

  (b) Bank has done discriminant analysis where objective is to find relation (10)
        between whether customer will recommend the bank to others? Bank has
        used discriminant analysis where two categories were considered viz. 'will
        recommend' and 'will not recommend'. There wcre 40 customers who said
        that they will not recommend and 324 customers said they will
        recommend. Following is the output obtained after using discriminant
        analysis. Interpret the output.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.483[a]	100.0	100.0	.571

a. First 1 canonical discriminant functions were used in the analysis.

[TURN OVER

BB-Con. 12288-14.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Q6_b.Performance of investments with this provider	.812	83.594	1	362	.000
Q6_c.Fees or commissions charged	.927	28.302	1	362	.000
Q6_d.Depth of products and services to meet the range of your investment needs	.876	51.138	1	362	.000
Q6_e.Ability to resolve problems	.884	47.361	1	362	.000
Q6_f.Online services offered	.981	7.018	1	362	.008
Q6_g.Multiple providers' products to choose from	.937	24.255	1	362	.000
Q6_h.Quality of advice	.891	44.125	1	362	.000
Q6_i.Knowledge of representatives or advisors you deal with	.886	46.522	1	362	.000
Q6_j.Representative knowing your overall situation and needs	.918	32.173	1	362	.000
Q6_k.Access to other professional resources	.943	21.766	1	362	.000
Q6_l.Degree to which my provider knows me	.929	27.541	1	362	.000
Q6_m.Quality of service	.709	148.474	1	362	.000

BB-Con. 12288-14.   **[TURN OVER**

Canonical Discriminant Function Coefficients	
	Function
	1
Q6_b.Performance of investments with this provider	.271
Q6_c.Fees or commissions charged	.005
Q6_d.Depth of products and services to meet the range of your investment needs	.029
Q6_e.Ability to resolve problems	.079
Q6_f.Online services offered	-.015
Q6_g.Multiple providers' products to choose from	.015
Q6_h.Quality of advice	.017
Q6_i.Knowledge of representatives or advisors you deal with	.100
Q6_j.Representative knowing your overall situation and needs	-.151
Q6_k.Access to other professional resources	.034
Q6_l.Degree to which my provider knows me	-.040
Q6_m.Quality of service	.708
(Constant)	-4.519

Unstandardized coefficients

Functions at Group Centroids	
code	Function
	1
.00	-1.972
1.00	.243

Canonical Discriminant Function Coefficients	
	Function
	1
Q6_b.Performance of investments with this provider	.271
Q6_c.Fees or commissions charged	.005
Q6_d.Depth of products and services to meet the range of your investment needs	.029
Q6_e.Ability to resolve problems	.079
Q6_f.Online services offered	-.015
Q6_g.Multiple providers' products to choose from	.015
Q6_h.Quality of advice	.017
Q6_i.Knowledge of representatives or advisors you deal with	.100
Q6_j.Representative knowing your overall situation and needs	-.151
Q6_k.Access to other professional resources	.034
Q6_l.Degree to which my provider knows me	-.040
Q6_m.Quality of service	.708
(Constant)	-4.519

Unstandardized canonical discriminant functions evaluated at group means

5. (a) How do we find out whether a variable in factor analysis should be retained in analysis?   (05)

(b) Following is the output of conjoint analysis done for washing machines. What conclusions can company draw from it.   (10)

CODES

Capacity	X1	X2
5 ltr	1	0
6 ltr	0	1
10 ltr	0	0

BB-Con. 12288-14.   **[TURN OVER**

Horsepower	X3	X4
1 Hp	1	0
1.5 HP	0	1
2 HP	0	0

Size	X5	X6
100 cu ft	1	0
120 cu ft	0	1
150 cu ft	0	0

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	5.959	.489		12.178	.000
	x1	.722	.453	.305	1.594	.127
	x2	.989	.453	.418	2.183	.041
	x3	1.078	.453	.455	2.379	.027
	x4	.467	.453	.197	1.030	.315
	x5	-.311	.453	-.131	-.687	.500
	x6	-1.044	.453	-.441	-2.305	.032

a. Dependent Variable: score

****

(3 Hours) [Total Marks: 60

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 two sections should be written on separate answer-books.

SECTION I

1. (a) For simple regression model $y = \beta_0 + \beta_1 X + e$ prove that regression $s.s. = \hat{\beta}_1^2 \Sigma(x_i - \bar{x})^2$.   (02)

   (b) Quarterly sales of pizzas (y) for 10 restaurants located near collage campuses and college student population (x) gives the following data. Assume regression model given in (a).   (06)
$\bar{x} = 14, \bar{y} = 130, \hat{\beta}_1 = 5, \Sigma(y_i - \bar{y})^2 = 15730, \Sigma(x_i - \bar{x})^2 = 568.$
      (i) Find $\hat{\beta}_0$. Write estimated regression equation and estimate sales when student population is 15.
      (ii) Write analysis of variance table and estimate σ^2.

   (c) Explain significance of graphs using Anscobe's Quartet.   (02)

2. (a) A multiple linear regression problem has six predictors and 30 observations. Regression and total sum of squares are respectively 3147.97 and 4296.97. Find multiple correlation coefficient. Write test procedure for testing the hypothesis that all predictor variables have no explanatory power. Find the value of test statistic for the given problem.   (05)

   (b) Why is transformation of variables necessary?   (02)

   (c) If the following assumptions do not hold, state in each case, what the problem. Problem is called.   (03)
      (i) Independence of predictor variables.
      (ii) Constancy of error variance.
      (iii) Independence of error.

3. (a) A study of regression of Y on X_1, X_2, X_3 gives the following values of VIF's and eigen values of correlation matrix of X_1, X_2 and X_3. Comment on the collinearity of the data based on   (04)
      (i) VIF values   (ii) Eigen values
$VIF_1 = 469.7 \quad VIF_2 = 1.0 \quad VIF_3 = 469.4 \quad \lambda_1 = 1.999 \quad \lambda_2 = 0.998 \quad \lambda_3 = 0.003.$

   (b) Describe Mellows C_P criteria.   (02)

   (c) How do you model 0 or 1 dependent variable using logistic regression? The 95% confidence intervals for odds ratio corresponding to each of the three predictor variables are given below. Conclude about effect of each variable.   (04)
$X_1: (0.73, 2.59) \quad X_2: (0.97, 1.48) \quad X_3: (1.87, 2.50).$

[TURN OVER

BB-Con: 12285-14.

SECTION II

4   An agricultural officer wants to study the effect of a factor Quantity of fertilizer with   (10) four levels and four different seasons on the yield (in standard bags) of specific crop. Analyze the data by using the following information by applying required model.

Quality fertilizer	Total		Season	Total		Raw s.s.= 13534
1	123		1	127		$N = 16$
2	122		2	112		$F_{3,9,0.05} = 3.86$
3	81		3	115		
4	126		4	98		

5.   Explain how you can check normality of residuals using Q-Q plot, also explain   (10) Anderson-Darling test for normality.

6.   The results of a chemical experiment involving three factors A, B and C are analyzed   (10) using SPSS. Following is the output given by the software. Answer the following questions;
    i.   Write the sum of squares for interaction effect of A and C. Is it significant?
    ii.   Write total sum of squares.
    iii.   Write the ANOVA table by ignoring non significant interactions.
    iv.   How many levels are there for factor B?
    v.   Write hypotheses and conclusion.
    vi.   Write the model.

Tests of Between-Subjects Effects

Dependent Variable: results

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	72044.889[a]	17	4237.935	73.138	.000
Intercept	2285136.111	1	2285136.111	39436.673	.000
A	64445.389	2	32222.694	556.096	.000
B	2466.778	1	2466.778	42.571	.000
C	3312.389	2	1656.194	28.582	.000
A * B	406.056	2	203.028	3.504	.052
A * C	1083.111	4	270.778	4.673	.009
B * C	16.722	2	8.361	.144	.867
A * B * C	314.444	4	78.611	1.357	.288
Error	1043.000	18	57.944		
Total	2358224.000	36			
Corrected Total	73087.889	35			

a. R Squared = .986 (Adjusted R Squared = .972)

****

BB-Con: 12285-14.

QP Code : 18480

(3 Hours) [Total Marks: 60

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 sections should be written on separate answer-books.

SECTION I

1. (a) A ten year old packaged tours and travel company needs a quarter-wise sales (10) forecast for the next year. Prepare a step-by-step systematic action plan for this purpose.

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

   Statement Number 2.1: A quantitative model cannot be considered appropriate unless error analysis is carried out.

   Statement Number 2.2: Moving averages of period 3 (three) should effectively smoothen the quarterly sales figures of multivitamin tablets.

   Statement Number 2.3: Delphi method can be used to forecast future values of a cardinal variable.

   Statement Number 2.4: Leading indicator model of regression is the most appropriate technique to forecast the next year's rank of India in the international survey of quality of urban life.

   Statement Number 2.5: In time series analysis, it is safer to employ the multiplicative model than the additive model.

3. (a) Explain the difference between 'structured decisions' and 'unstructured (10) decisions'. Give one example of each.

   (b) What are the quantitative tools available for data-based decision making? Briefly describe each of them.

SECTION II

4. What are the key objectives of scientific inventory control? Explain each in (10) brief.

5. Define problem of Economic Order Quantity with known demand. The (10) shortages are not allowed. Rate of replenishment of inventory is instantaneous. Determine optimum production quantity for each production run of equal length stating the conditions and diagram.

BB-Con. 12290-14. **[TURN OVER**

SECTION III

6.   The following particulars are extracted from the records of a company   (10)

	Product A	Product B
Selling price (per unit)	Rs.100	Rs.120
Raw Material Required (per unit)	2 kg	3 kg
Raw Material Cost	Rs.10	Rs.15
Direct Labour Cost	Rs.15	Rs10
Direct Expenses	Rs.5	Rs.6
Other Variable Expenses	Rs.15	Rs.20
Fixed	Rs.52500	

Find the product mix which will yield maximum profit, when raw material is in scarcity (Key factor) and available only to the extent of 10000 kg.

Assume the maximum demand of each product as 3500 units

****

BB-Con. 12290-14.

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. Govt. Food Dept. collected data from two Noodles mfg. plant in terms of pesticide content in their product. in milligram. Kindly advise which product is safer & consistent? Which Company complies with Govt. Target of 25 mg ? How many samples Govt. must collect to check their base-line to estimate in 1mg using 95% confidence level? Calculate Zst & DPMO for both Plants?   (10)

Month	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar
Plant A	20	36	23	30	33	35	29	36	35	25	35	26
Plant B	31	46	28	40	45	39	38	54	34	26	48	39

2. Three diff. blood samples distributed to 3 diff. labs to check blood sugar level. If 3 labs are considered as 3 operators then 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
Lab Name	1	2	3	1	2	3	1	2	3
Trial No.	1	1	1	2	2	2	1	1	1
Blood Sugar	150	132	157	141	142	165	110	100	125

Sample No	2	2	2	3	3	3	3	3	3
Lab Name	1	2	3	1	2	3	1	2	3
Trial No.	2	2	2	1	1	1	2	2	2
Blood Sugar	119	110	118	195	220	205	198	115	200

3. Explain in detail   (05)
    a. Control Charts
    b. Six Sigma DMAIC Phases & steps

4. Following 3 factors determine the crop yield in Kg/Acre. Conduct DOE & identify the factors affecting Yield. Write-down regression equation & Using Solver decide levels for maximize yield.   (10)

Run Order	Fertilizer Qty Kg/ Acre	Water Qty Lts/ acre/ day	Soil pre conditioning Period Days	Crop Yield Kg/ Acre
1	3	40	100	82.300
2	1	40	100	86.020
3	1	40	80	86.670
4	1	40	100	86.020
5	3	30	100	64.010
6	1	30	80	66.000
7	3	40	80	84.000
8	3	40	100	84.000
9	1	30	100	66.000
10	3	40	80	84.000
11	3	30	100	65.000
12	1	40	80	87.000
13	3	30	80	63.000
14	3	30	80	64.000
15	1	30	80	66.000
16	1	30	100	65.000

SECTION II

5. (a) A passenger flight needs certain activities to be done before take-off. The activities are listed in following table along with its predecessor wherever applicable.   (05)

Srl	Activity	Predecessor	Time required in Minutes
1	Bring flight to the terminal		10
2	Cleaning	1	30
3	Technical checking	1	20
4	Fuelling	3	40
5	Loading of baggage	2	20
6	Passenger boarding	2	20
7	Final checks	4,5,6	5
8	Door closing & Ready to takeoff	7	5

1. Draw activity diagram for these activities.
2. Find critical path
3. If the departure time of the aircraft is 10.20 AM, find out what time its cleaning required to be started?

(b) Draw flowchart for a process of delivering spare parts across the counter to mechanics in car repairs workshop.   (05)

1. Mechanic requests a part
2. Manager may approve request or reject request

2

3. Approved request goes to the store person
4. If the request is rejected, mechanic has been informed about it
5. Store person looks for availability of stock
6. If stock is available, part is kept ready for delivery at the counter
7. Mechanic has been informed by SMS
8. Mechanic collects the part
9. If part is not available, purchase order is raised to purchase part
10. Part is received from the supplier against purchase order
11. Part is registered in stock
12. Part delivery process followed from 5 to 7

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

	A	B	C	D	E	F	G
1	Srl No	Name	DOB	Gender	Education	Occupation	Income
2	1	X1	21-10-1985	Male	SSC	Business	120000
3	2	X2	03-05-1981	Female	Graduate	Home Maker	20000
4	3	X3	08-06-1971	Male	Post Graduate	Service	80000
5	4	X4	30-07-1979	Male	HSC	Service	100000
6	5	X5	25-10-1979	Male	HSC	Service	80000
7	6	X6	19-05-1985	Female	Graduate	Business	30000
8	7	X7	15-08-1972	Male	Graduate	Service	25000
9	8	X8	23-06-1975	Male	Graduate	Service	15000
10	9	X9	02-12-1980	Male	Post Graduate	Service	75000
11	10	X10	17-04-1992	Female	Post Graduate	Service	150000
12	11	X11	19-06-1971	Male	Graduate	Business	55000
13	12	X12	22-06-1990	Male	HSC	Service	40000
14	13	X13	08-07-1986	Male	Post Graduate	Service	25000
15	14	X14	13-01-1976	Male	Post Graduate	Service	35000
16	15	X15	19-09-1973	Female	SSC	Home Maker	10000
17	16	X16	29-03-1989	Male	Graduate	Service	40000
18	17	X17	05-08-1972	Male	PHD	Business	15000
19	18	X18	03-08-1976	Female	SSC	Home Maker	18000
20	19	X19	07-02-1974	Male	HSC	Service	45000
21	20	X20	05-08-1991	Female	Graduate	Retired	90000

1. Write function to calculate number of Female respondents.
2. Write function to calculate total income of all respondents
3. Write function to calculate age of last respondent
4. Write function to calculate number of respondents whose income is greater than or equal to Rs 100000
5. Write function to calculate total income of Post Graduates in the list

(b) TV news channel did a survey to find out number of people viewing the channel in the city between 6 PM and 12 PM. A data of 3 days is as follows. Which QC tool can be used to graphically represent the data? Use this QC tool and graphically represent the data. Describe your analysis in 1-2 lines. (05)

	No of Viewers		
Time	Day 1	Day 2	Day 3
6-7	2680	2019	1065
7-8	3047	3214	2584
8-9	4450	4387	3375
9-10	7310	7950	4548
10-11	5300	5136	2977

7. (a) Suggest appropriate QC tool that can be used for following problems (05)

Srl	Problem
1	A pharmaceutical company is studding effect of a medicine having 3 different strengths on different age groups
2	A class of students have various ideas about how to make learning more interesting. We need to group them and arrive at the most effective ideas.
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	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.
5	A funeral director maps out the tasks involved in a funeral, paying attention to how multiple funerals might overlap. By rearranging tasks, he enables one more funeral to be fitted into a day, thus increasing customer responsiveness along with business volume.

(b) Let us consider a case study that has one 2-level factor (A) and three 3-level factors (B, C and D). Calculate degrees of freedom. Which technique can be used to fit this case study with L9 orthogonal array? Explain your answer and draw a table of Experiment layout. (Refer following L9 orthogonal array) (05)

4

Expt. No	Factors			
	A	B	C	D
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	2	1	2	3
5	2	2	3	1
6	2	3	1	2
7	3	1	3	2
8	3	2	1	3
9	3	3	2	1

****

(3 hours)

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

1.   Describe application of following tests/statistical methods in clinical   (10)
      research studies (any 4 of 5 – 2.5 marks each)
         a. Two Sample T test
         b. Chi Square Test
         c. Wilcoxon Rank Sum Test
         d. Logistic Regression
         e. Bonferroni Adjustment

2.   Describe information needed to calculate sample size for a clinical trial.   (10)

3.   Answer following questions (5 marks each)   (10)

    a.   What are different types of data collected in a clinical trial?

    b.   Describe application of Randomization and Blinding in clinical trials.

4.   Describe the role of a statistical analyst at various stages of a clinical trial   (10)
      project.

5.   Please read the following scenario and answer questions mentioned   (10)
      below.

      A sponsor wants to conduct a bioequivalence study of two topical (i.e.
      locally applied) anti-fungal ointment formulations. Formulations are
      called "Test" and "Reference". In this study sponsor intends to compare
      efficacy of "Test" Ointment to "Reference" Ointment. Test and Reference
      Ointments will be administered to separate groups of patients by
      randomly assigning patients to receive any one of the two ointments (Test
      or Reference) for 1 week. The Primary efficacy parameter in this study
      will be proportion of subjects who achieve "clinical cure" from fungal
      skin infection after application of ointment. Patient will be considered to
      have achieved "clinical cure" if all signs and symptoms of fungal
      infection subside after 1 week treatment with ointment. Regulatory
      authority will consider the two formulations (Test and Reference) to be
      bioequivalent, if difference in proportions of "Clinical Cure" with two
      formulations (Test – Standard) lies within the range of -0.2 to + 0.2 (or -
      20% to +20%). Sponsor has sought your help in providing statistical plan

and sample size calculation.

    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. [4 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 <u>not</u> <u>true</u>?

      (i) Data Step ends with a RUN statement.

      (ii) A data step can never 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 use IF-THEN-ELSE statements to execute conditional logic

    b. Which of the following is the <u>correct syntax</u> to subset a dataset ABCD in order to create a new dataset EFGH , and subsetting condition being, value of variable x is greater than value of variable y?

      (i)   Data EFGH;

           SET ABCD;

            Where x > y;

           Run;

      (ii) Data EFGH;

            SUBSET ABCD (on x >y);

            Output;

           Run;

(iii) Data EFGH;

      SET ABCD;

      ON x > y;

      Output; ·

      Run;

(iv) Data EFGH;

      SET ABCD;

      Output if x > y;

      Run;

c. Which of the following is a <u>correct</u> PROC SORT statement?

    (i) PROC SORT DATA = A;   OUT= B;   BY X, Y;   RUN;

    (ii) PROC SORT DATA = A   OUT=B;   BY X Y;   RUN;

    (iii) PROC SORT DATA = A   OUT=B;   BY; X Y; RUN;

    (iv) PROC SORT DATA = A;   OUT=B;   BY X, BY Y; RUN;

d. Which of the following is a <u>correct</u> statement about SAS procedures?

    (i) Proc NPAR1WAY cannot be used for performing Wilcoxon Rank sum Test.

    (ii) Data Step can combine two identical SAS datasets vertically.

    (iii) Proc sort can sort as well as transpose a SAS dataset

    (iv) Proc univariate cannot be used to perform tests of normality.

e. Which one of the following SAS procedure will <u>NOT display</u> output by default?

    (i)   PROC PRINT

    (ii) PROC FREQ

    (iii) PROC MEANS

    (iv) PROC SORT

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) Define Euclidean and statistical distance from a point $P(x_1, x_2, \ldots x_n)$ to the   (06) origin. Briefly discuss significance of statistical distance. State properties that should be satisfied by distance function.

  (b) The following matrix gives marks of 4 students in two exams.   (04)

$$\begin{bmatrix} 4 & 8 \\ 8 & 9 \\ 7 & 7 \\ 5 & 4 \end{bmatrix}$$

Obtain the variance covariance matrix of the marks. Also obtain the correlation matrix of the marks.

2. (a) Consider the following data matrix.   (06)

$$X = \begin{bmatrix} 3 & 6 & 6 \\ 2 & 3 & 1 \\ 7 & 12 & 8 \end{bmatrix}$$

Obtain generalized sample variance and interpret. Let $b = \begin{pmatrix} 1 & -1 & 2 \end{pmatrix}'$ and $c = \begin{pmatrix} 1 & 2 & 1 \end{pmatrix}'$. Obtain sample variance of $b'X$ and $c'X$. Also obtain the sample covariance between $b'X$ and $c'X$.

  (b) Let X_1, X_2, X_3 and X_4 be independent identically distributed random   (04) vectors with mean vector μ and covariance matrix Σ. Obtain the mean vector and covariance matrix of

     (i) $X_1 - 3X_2 + X_3 + X_4$

     (ii) $\sum_{i=1}^{3} X_{i+1} - X_i$

3. (a) Let $X \sim N_p(\mu, \Sigma)$ where $X = \begin{bmatrix} X^{(1)} \\ \hline X^{(2)} \end{bmatrix}$ $\mu = \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix}$ $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma 22 \end{bmatrix}$ (03)

State (i) conditional distribution of $X^{(1)}$ given $X^{(2)} = x^2$

     (ii) Distribution of $(X - \mu)' \Sigma^{-1}(X - \mu)$.

(b) Let X_1, X_2, X_3, X_4 and X_5 be mutually independent $N_3(\mu, \Sigma)$ where (04)

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

Let $Y_1 = \sum_{i=1}^{5} \frac{1}{5} X_i$ and $Y_2 = X_1 - X_2 + X_3 - X_4$.

Obtain the distribution of $Z = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$. Check the independence of Y_1 and Y_2.

(c) Explain the brief "Transformations to near normality" (03)

4. (a) Explain the method of using the T^2-statistic for paired comparisions of two multivariate means from multivariate normal distribution. (05)

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

5. (a) List similarity coefficients defined in terms of frequencies of matches and mismatches for any two items. The scores of individuals 1 and 2 on the $p=6$ variables are (05)

Individual		X_1	X_2	X_3	X_4	X_5	X_6
	1	0	0	0	1	1	1
	2	1	1	1	0	1	0

Compute the table showing the number of matches and mismatches. Also compute the similarity coefficients listed by you.

(b) Let Σ denote the covariance matrix of X where $X' = [X_1, X_2, ..., Xp]$ and Y_1, Y_2 denote the first two principal components of Σ. Obtain Var(Y_1), cov(Y_1, Y_2), correlation coefficient between Y_1 and X_3. (05)

6. (a) Write down the orthogonal factor model and explain the terms. (i) Communality (ii) specific variance. (05)

(b) Given the following data

$$\bar{X}_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad \bar{X}_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}. \quad S_{pooled} = \begin{bmatrix} 7.3 & -1.1 \\ -1.1 & 4.8 \end{bmatrix}$$ (05)

Compute

(i) the distance between two populations
(ii) the sample discriminant function.

****