Indian Statistical Institute

PG Diploma in Business Analytics 1st Year : 2015 2016
Mid-Semester Examination
Subject: Stochastic Processes and applications
Date: 07 /09/2015 Time: 3 hours Marks : 100

Note: Notations used are as explained in the class.
Answer Group-A and Group-B on separate answer scripts.

Group-A

1. You have to choose 2 persons (distinct) from a group of 5 persons so that each person has equal probability of being chosen. You have a coin (may be biased) and you may toss it as many times as you like. Describe a procedure, with proofs, to make such a selection. [20]

2. Consider the tea tasting example. There are two methods of preparing tea: A and B. A lady claims that she can distinguish the methods by taking couple of sips. in 90% cases. 5 cups of tea are prepared by method A and 5 cups by method B. These 10 cups are now put into a random order and given to her to taste. It is decided that her claim will be accepted if she can correctly classify at least 8 cups. What would be Type-I and Type-II errors? [15]

3. (a) State Bayes Theorem.
   (b) X, Y, Z are binary variables: X and Y are independent. X is 0 or 1 with probability 0.5 each. Y is 0 with probability 0.6 and 1 with probability 0.4. $Z = X + Y$ modulo 2. Compute $\text{Prob}(Y = 1/Z = 0)$. [5 + 10]

Group-B

1. (a) Define a **probability** function.
   (b) Let P be a probability function and $C_1, C_2, \cdots$ form a partition on the sample space S. Prove that
   $$P(A) = \Sigma_{i=1}^{\infty} P(A \cap C_i)$$
   [5 + 5]

2. (a) Define **discrete** and **continuous** random variables.
   (b) Let X and Y be two random variables defined on the same sample space. Prove the following
      i. $E(XY) = E(X)E(Y) + Cov(X, Y)$.
      ii. $V(X + Y) = V(X) + V(Y) + 2Cov(X, Y)$.
   [6 + 6 + 6]

3. A typesetter, on the average, makes one error in every 500 words typest. A typical page contains 300 words. Assuming binomial setting find the probabilty that there will be no more than two errors in five pages. Also, find the probability using the poisson approximation. [9]

4. If $X \sim N(\mu, \sigma^2)$, then prove that the random variable $Z = \frac{X-\mu}{\sigma}$ has a $N(0, 1)$ distribution. Calculate the expected value of X. [8 + 5]

INDIAN STATISTICAL INSTITUTE

PGDBA 2015-2016, I Semester

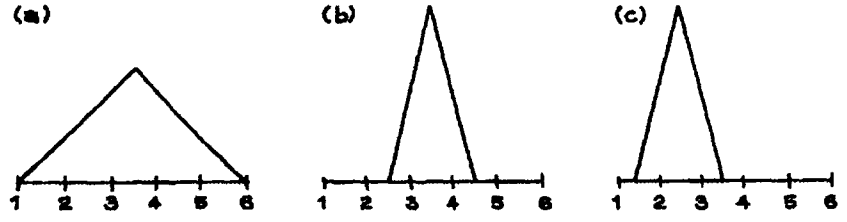
Statistical Structures in Data

Mid-semester examination

Maximum marks: 75 | 8 September 2015 | Maximum time: 3 hours

This examination is closed book, closed notes. Non-programmable calculators are allowed. The entire question paper is for 85 marks. The maximum you can score is 75.

1. Below are sketches of histograms for three large data sets. Match the sketch with the description. Some descriptions will be left over. Give your reasoning in each case. [5]



(i) ave ≈ 3·5, SD ≈ 1 (iii) Ave ≈ 3·5, SD ≈ 2 (v) ave ≈ 2·5, SD ≈ 0·5

(ii) ave ≈ 3·5, SD ≈ 0·5. (iv) ave ≈ 2·5, SD ≈ 1 (vi) ave ≈ 4·5, SD ≈ 0·5

2. In a particular examination, men generally scored better than women. The average score obtained by the men was 580, while the average score obtained by the women was 460. The standard deviation for either group was 120. The numbers of men and women examinees were about the same. The histograms are reasonably well approximated by appropriate normal distributions.

   a. What percentage of men and women would have obtained scores above 700?

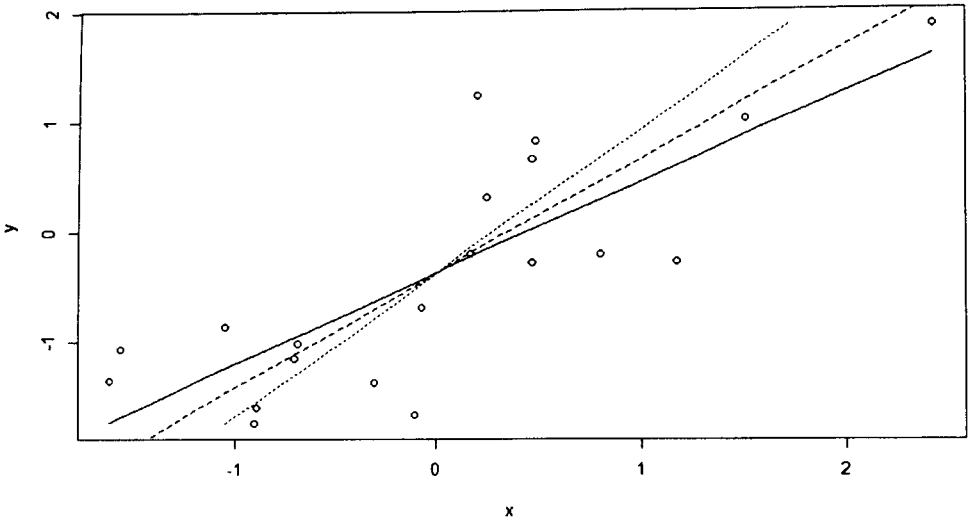
   b. What is the standard deviation of scores of the entire group of candidates?

   c. By using either the computations of part a or the computations of part b, and after justifying your choice, calculate the percentage of all candidates that would have obtained scores above 700.

[2+4+4=10]

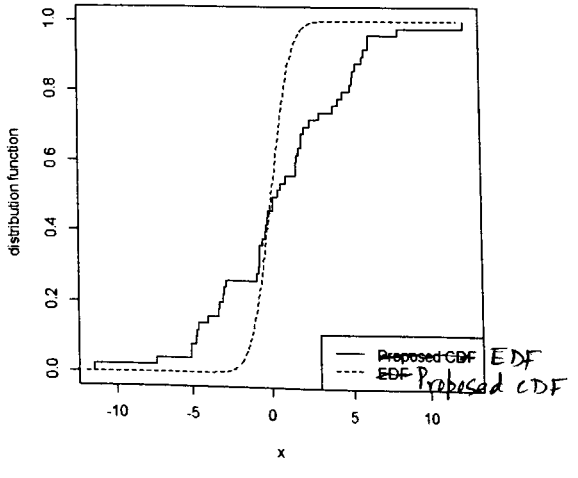
3. A large, representative sample of Americans was studied, in the Health and Nutrition Examination Survey conducted by the Public Health Service. The percentage of respondents who were left-handed decreased steadily with age, from 10% at 20 years to 4% at 70. Does the data show that many people change from left-handed to right-handed as they get older? Explain. [5]

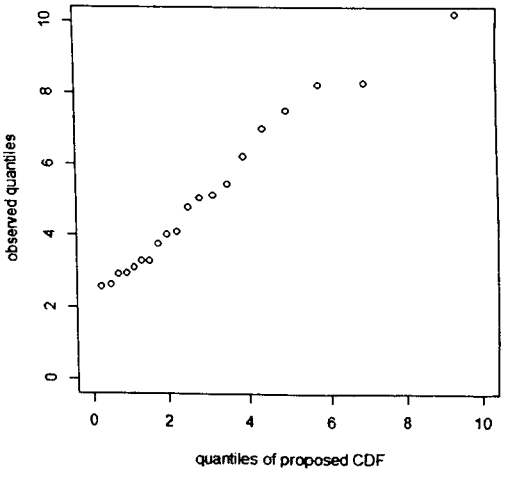
4. The three lines in the scatter plot represent the SD line, the least squares fitted line and the least squares fitted line obtained by interchanging the variables. Identify which line represents which, with reasons. [5]



5. Let $\hat{F}_X$ be the empirical distribution function (EDF) of a random sample of size n from a continuous cumulative distribution function (CDF) F_X. Show that the variance of $\hat{F}_X(x)$ for any x in the range $(-\infty, \infty)$ is $\frac{1}{n}F_X(x)\big(1 - F_X(x)\big)$. [5]

6. The overlaid plots of the EDF of a data set (sample size 50) and a proposed CDF are shown in the left panel below. The QQ plot of another data set (sample size 20) for a proposed CDF is shown in the right panel.





a. Draw free-hand sketch of the QQ plot corresponding to the left panel. Label the axes clearly.

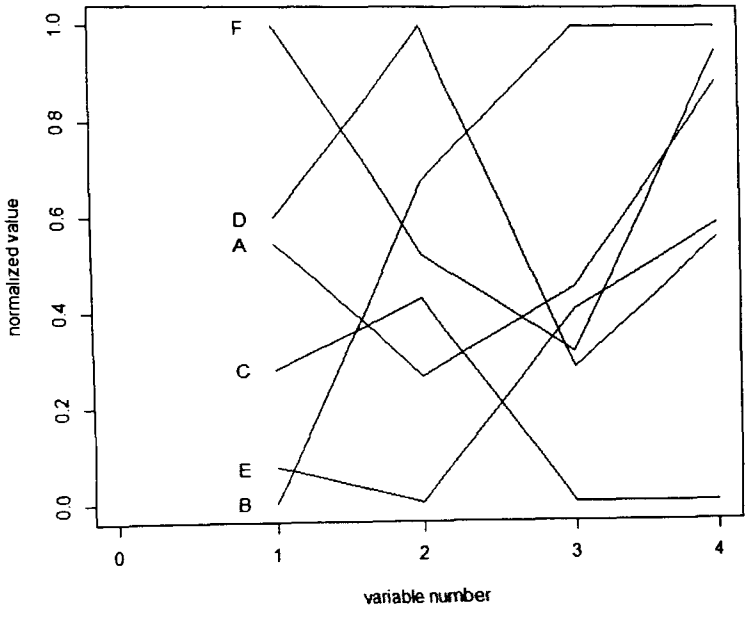
b. Draw free-hand sketch of the PP plot corresponding to the right panel. Label the axes clearly.

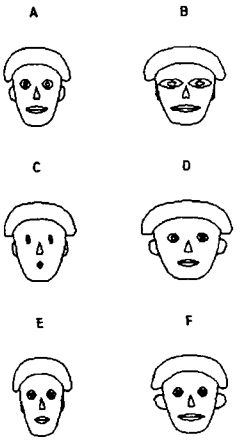
c. The proposed distribution for the first data set (sample size 50) is the standard normal distribution. Will the Shapiro-Wilks test of goodness of fit produce a small p-value or a large p-value? Explain.

d. Will the Kolmogorov-Smirnov test of goodness of fit for the first data set produce a small p-value or a large p-value? Explain.

[7+8+3+2=20]

7. The line plots and Chernoff faces for six cases (marked 'A' to 'F') are shown below.





a. Identify the features of the Chernoff faces that correspond to the variable numbers shown in the line plot.

b. Draw star plots of cases 'A' and 'B'.

[6+4=10]

8. A data set, compiled from 50 countries, contains the percentage of population having annual income below the first quartile of the income distribution of the country (X_1), the percentage of population between the first and the third quartiles (X_2) and the percentage of population above the third quartile (X_3). Note that $X_1 + X_2 + X_3 = 100$. Because of this interdependence of the variables, it is enough to consider only two variables. Analyst A considers X_1 and X_2, while Analyst B considers X_2 and X_3. Will the distance between a pair of countries found by the two analysts be the same, when one uses (i) the Euclidean distance, (ii) the Euclidean distance for standardised data, (iii) the Mahalanobis distance? Explain. [5]

9. If the triplet (X_1, X_2, X_3) has the multinomial distribution with parameters n, p_1, p_2 and p_3 with $p_1 + p_2 + p_3 = 1$, what is the conditional distribution of X_2 given $X_1 = x$? Explain. [7]

10. Consider the random variables U and V, such that U has the uniform distribution over the interval [0,1], and $V = U$.

    a. What are the marginal CDF's of U and V?

    b. Show that the joint CDF of U and V is given by

    $$F(u, v) = \min(u, v) \text{ for } 0 \leq u, v \leq 1.$$

    c. Use the above facts to explain the circumstances when two random variables X and Y having continuous marginal CDF's F and G would have a bivariate distribution with copula given by $C(u, v) = \min(u, v)$ for $0 \leq u, v \leq 1$.

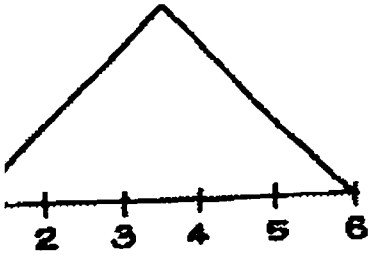
    d. Outline a scheme to generate a sample from the above bivariate distribution.

    e. Sketch the scatter plot of $\hat{G}(Y_i)$ vs. $\hat{F}(X_i)$, where $\hat{F}$ and $\hat{G}$ are the EDF's of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, and $(X_1, Y_1), \ldots, (X_n, Y_n)$ are samples generated as in part d.

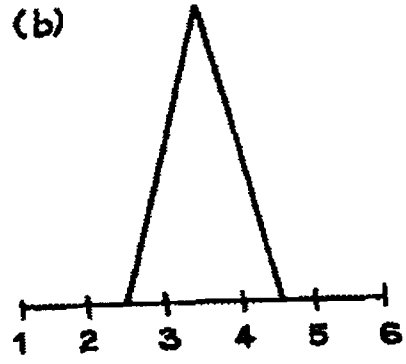
    f. What will be the sample correlation of the paired 'data' $\left(\hat{F}(X_1), \hat{G}(Y_1)\right), \ldots, \left(\hat{F}(X_n), \hat{G}(Y_n)\right)$?

[2+3+3+2+2+1=13]

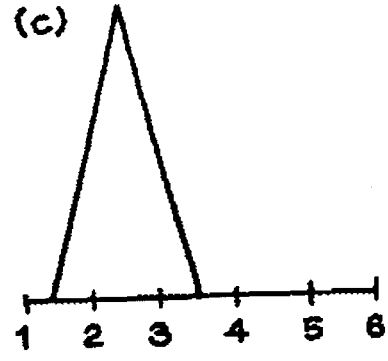
(i) ave ≈ 3.5, SD ≈ 1
(ii) ave ≈ 3.5, SD ≈ 0.5
(iii) ave ≈ 3.5, SD ≈ 2
(iv) ave ≈ 2.5, SD ≈ 1
(v) ave ≈ 2.5, SD ≈ 0.5
(vi) ave ≈ 4.5, SD ≈ 0.5



(b)



(c)



INDIAN STATISTICAL INSTITUTE
Mid Semester Examination

Course Name: PGDBA 2015 – 2017
Subject Name: Inference (BAISI3)
Date: 9th September, 2015 Maximum Marks: 100 Duration: 3 hours
Notes: Answer all questions. The paper carries 110 marks but the maximum you can score is 100.

1. Answer the following
   a. Inference has four steps in the context of business analytics. Explain the steps briefly. [8]
   b. Explain the meaning of summarization of data with an example. Name two measures of variation (spread), with brief explanations. The returns of stocks are known to be random variables. Suppose A and B are two different stocks and suppose you have collected data on their returns for a given period of time (suppose you observe the return over a period of one month). Suppose you have constructed histograms of the returns of the stocks and have computed the usual statistics. You found that the average return of both stocks over one month is the same and even the spreads are similar. However, returns from stock A are negatively skewed while the returns from stock A have a positive skew. Which stock will you prefer and why? [4 + 2 X 2 + 5 = 13]
   c. Briefly explain the difference between explanatory and predictive analytics with an example for each. [6]
2. Answer the following
   a. What is measurement? Explain briefly the different scales of measurement with examples. [2 X 5 = 10]
   b. While solving a business analytic problem, we may choose to measure the response variable in a manner such that it contains less information than are available in the system. Look at the following examples and identify how you could change the response variable such that it contains more information than what has been specified in the description given below. Identify the response variables and their scales of measurement clearly.
      i. A company wants to understand the reasons behind employees quitting the company so that appropriate actions may be initiated to improve retention. They plan to carry out dependency analyses to establish relationship between attrition in a given period – a binary variable and various other factors that may impact the attrition. [4]
      ii. Manufacturers of commercial explosives try to control the amount of explosive put into shells. Packing more explosive than necessary is expensive as well as unsafe and packing less explosive will lead to lower release of energy leading to ineffective blasts. Explosive manufacturers often count the average number of shells being packed for a given quantity of explosive. They attempt to find the relationship between this number and various other controllable factors believed to contribute to the packing density of explosives. [4]
   c. Suppose you have been asked to carry out marketing analytics. You are analysing bids made to customers and you want to provide a visual comparison of the impact of various factors on the propensity to win bids. Suppose the data were collected and summarized as a k X 2 contingency table (2 way tables with k rows and 2 columns. Each row represents a factor that may impact the outcome of the bid, i.e. win or loss, and each column gives the number of bids won or lost).
      i. Draw a sample table
      ii. Suggest a suitable visualization scheme with explanation. [2 + 5 = 7]

d. What is a mean function plot? Can the mean function plot be considered to be a plot of E(Y / X) for various values of X? Explain briefly. [2 + 3 = 5]

3. Answer the following
   a. What do we mean by the term random sample? Give two examples of drawing samples in the context of business analytics such that the samples are unlikely to be random. In this context discuss briefly about the concepts of sampling and non-sampling fluctuations. [2 + 2 X 3 + 4 = 12]
   b. Suppose we have a finite population of size N. Suppose samples are being drawn one at a time from this population without replacement. Let X_i denote the random variable representing the i^{th} draw, i = 1, 2,...N. Show that the rendom variables $X_1, X_2, ..., X_N$ are not independent but are identically distributed. [10]
4. Answer the following
   a. What is a consistent estimator? What is an unbiased estimator? [3 + 3 = 6]
   b. Explain how MSE may be used to compare the 'goodness' of estimates. [5]
   c. Explain briefly the concept of likelihood function with an example. [5]
   d. A large machine shop having over a thousand similar machines experiences machine failures from time to time. In order to ensure high uptime of all machines, a number of mechanics need to be employed to attend to machines that broke down. The company wants to estimate the chance of five or more failures per day. In order to estimate this chance, the company has collected data on failures on a daily basis for a number of days – say k days. Suppose the observed number of failures were $x_1, x_2,...,x_k$. Explain how this data may be used to estimate the probability of the stated event. Notice that 5 or more failures may be an unlikely event and may not be observed even in a reasonably large sample. State your assumptions clearly. [15]

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination: 2015-16

Course Name: PGDBA

Subject Name: Fundamentals of Data Base Systems

Date: 10/09/2015 **Maximum Marks: 100** **Duration: 3 Hours**

Note: **Answer all questions to the point.**

Clearly mention all your assumptions for writing the answer.

Please ensure that all the answers are written in legible handwriting.

1. To design a suitable database for the GenX, chain of medical shops all over India. The company has provided the following information:
   - Patients who buy all the prescribed medicines from one of their shops are identified by a unique PatientId. For each patient, company tries to record their name, address, contact number and age for future reference.
   - Doctors, who prescribe medicine for these patients, are identified by a unique DoctorId. For each doctor, the name, highest degree, specialty, and address are important information for the company.
   - Every patient has a primary physician (the first prescription they submit to the medicine shop or the most frequently submitted Doctor's prescription). Every doctor has at least one patient.
   - Doctors prescribe medicine, along with a quantity, for their patients. A doctor may prescribe one or more medicines for several patients, and a patient may obtain prescriptions from several doctors.
   - Each prescription has a date and a quantity associated with it. Doctor may prescribe the same medicine for the same patient more than once if required. The company stores the last prescription submitted by each patient.
   - Each medical shop is identified by name and location. Also their address and phone numbers are important information. Each such shop has a shop manager.
   - For each medicine, the trade name and formula must be recorded. Medicines sold by a given pharmaceutical company are uniquely indentified by the trade name.
   - Pharmaceutical companies have long-term contracts with medical shops. A pharmaceutical company can have contract with several medical shops individually, and a medical shop can have contract with several pharmaceutical companies. For each contract, a start date, an end date, and the text of the contract are stored. Company keeps track of all the products of those pharmaceutical companies.

- Each medical shop sells several medicines and has a price for each of them through India.
- The company keeps track of all the employees working in all their shops. Each emplo is attached to one such shop.

a) Draw a suitable ER/EER diagram that captures the requirement of this company.
b) Write SQL statements to create the corresponding relations and capture as many of constraints as possible. If you cannot capture some constraints, explain why.
c) What will be the *changes in the diagram* for the following design requirem changes:
   i. Medicine may be sold at different prices by different medical shops?
   ii. The company stores the all the prescription submitted by each patient.
   iii. Patients may buy partial medicines.
   iv. Patients may visit any of these medical shops.
   v. Some of these medical shops are 24x7 open. Rests are 12x7. Duty of each sta in shift, each of 8hr duration. Company need to keep track of its emplo working in which shift.

20+20+20

2. Consider the following schema:
Suppliers (sid: integer, sname: string, address: string)
Parts (pid: integer, pname: string, color: string)
Catalog (sid: integer, pid: integer, cost: real, billing_date: date)
The *key fields are underlined*, and the *domain of each field is listed after the field name.*
Write the following queries in relational algebra or in SQL:

a) Find the sids of suppliers who supply every part.
b) Find the sids of suppliers who supply every RED coloured part.
c) Find the names of suppliers who supply only BLUE coloured part.
d) Find the sids of suppliers who supply some WHITE coloured or some GR coloured part, but not both.
e) Find the pids of those parts that were supplied by at least two different suppliers.
f) Find the pids of the most expensive parts supplied by suppliers named ABC.
g) Find, for each month, the part that was supplied maximum in quantity.
h) Find the pids of such parts supplied by every supplier at less than Rs. 200. (I supplier either does not supply the part or charges more than Rs. 200 for it, the p not selected.)

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination : 2015–16

Course : Post Graduate Diploma in Business Analytics (First Year)

Subject : Computing for Data Sciences : BAISI–4 for PGDBA–I

Date : 11 September 2015 | Maximum Marks : 90 | Duration : 3 Hours

Problem A [30]

1. Define *norm* on the n-dimensional vector space $\mathbb{R}^n$. Given a norm $\rho(\cdot)$ on $\mathbb{R}^n$, define a related notion of *distance* between any two vectors in $\mathbb{R}^n$, and state its properties. [2 + 3]

2. Let the ℓ^p norm of a vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ in $\mathbb{R}^n$ be defined as $||\mathbf{x}||_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$. Comment on the significance of the ℓ^1 and ℓ^2 norms of $\mathbf{x}$ in $\mathbb{R}^n$, in terms of the geometrical depiction of the unit vectors in $\mathbb{R}^n$. Is there any relation between the ℓ^1 and ℓ^2 norms of $\mathbf{x}$ and the statistical properties of the set of real numbers $\{x_1, x_2, \ldots, x_n\}$? [5 + 5]

3. Let an *inner product* on $\mathbb{R}^n$ be defined as the *dot product* of two vectors: $\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^n x_i y_i$, where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ and $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$. What is the geometrical significance of this inner product in $\mathbb{R}^n$? Is there any statistical significance of this inner product in connection with the sets of real numbers $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_n\}$? [2 + 3]

4. Suppose that you have an $n \times p$ matrix $\mathbf{X}$ representing a dataset, comprising of n independent observations along p features. Assume that the dataset is *centered*, that is, the mean of values along each column in $\mathbf{X}$ is zero. Comment on the statistical significance of the matrix $\mathbf{X}^T\mathbf{X}$ in terms of the features and observations in the dataset. [5]

5. What can you say about the dataset if the matrix $\mathbf{X}^T\mathbf{X}$ is diagonal? What can you say if the matrix $\mathbf{X}^T\mathbf{X}$ is block-diagonal, with k distinct blocks along the main diagonal? [2 + 3]

Problem B [30]

1. Describe the role of an $m \times n$ matrix $\mathbf{X}$ as a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$. Your description should include the conceptual notions of the fundamental subspaces – RowSpace, ColSpace and NullSpace of $\mathbf{X}$, as well as Rank of $\mathbf{X}$. [7]

2. Given the fundamental subspaces of an $m \times n$ matrix $\mathbf{X}$, how do you determine the following?
   (a) Whether the matrix is a 1-to-1 linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$;
   (b) Whether the matrix is an onto linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$;
   (c) Whether the matrix is an invertible linear map from $\mathbb{R}^n$ to $\mathbb{R}^n$. [3]

3. Suppose that the *full* Singular Value Decomposition of an $m \times n$ matrix $\mathbf{X}$ results in:

$$\mathbf{X} = \begin{bmatrix} | & & | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_r & \cdots & \mathbf{u}_m \\ | & & | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{bmatrix} \begin{bmatrix} | & & | & & | \\ \mathbf{v}_1 & \cdots & \mathbf{v}_r & \cdots & \mathbf{v}_n \\ | & & | & & | \end{bmatrix}^T$$

Represent this decomposition as $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, and comment on the dimension of each matrix this representation. Discuss the connection of these matrices with the fundamental subspac of $\mathbf{X}$. How can you determine the Rank of $\mathbf{X}$ given this SVD representation? [3 − 5 +

4. As per the above representation of the SVD of $\mathbf{X}$, determine the dimension and rank of ea of the matrices $\mathbf{Z}_i = \sigma_i \mathbf{u}_i \mathbf{v}_i^T$, where $1 \leq i \leq r$. Is there a way to reconstruct the origin matrix $\mathbf{X}$ given the matrices $\mathbf{Z}_i$ for $1 \leq i \leq r$? [3 +

5. Is there a way to reconstruct the original matrix $\mathbf{X}$ given the matrices $\mathbf{Z}_i$ for $1 \leq i \leq k$, whe k is strictly less than r? If so, provide such a construction. If not, provide an *approxima* reconstruction of $\mathbf{X}$ using the available matrices $\mathbf{Z}_i$ for $1 \leq i \leq k$, and comment on the quali of such an approximation. [2 +

Problem C

[1

Represent a book in the form of an $m \times n$ matrix $\mathbf{B}$, where m is the total number of sentences the book and n is the total number of distinct words in the book, such that the entry $\mathbf{B}[i, j]$ in t matrix represents the frequency of occurrence of the j-th word W_j in the i-th sentence S_i.

Importance of the words and sentences are denoted by *scores*. The score u_i of S_i is equal to the s of scores of the words in it, weighted by the frequencies of occurrence. The score v_j of W_j is eq to the sum of scores of the sentences it is contained in, weighted by the frequencies of occurrenc

$$u_i = \sum_{j=1}^{n} \mathbf{B}[i, j] \cdot v_j \quad \text{for } i = 1, 2, \ldots, m \qquad\qquad v_j = \sum_{i=1}^{m} \mathbf{B}[i, j] \cdot u_i \quad \text{for } j = 1, 2, \ldots, n$$

Devise an efficient strategy to identify 10 *keywords* (i.e., the most important words) from the bo

Problem D

[1

Suppose that you have a dataset where m individuals have reviewed a collection of n movies, have provided scores (between 0 to 9, say) for each one. Suppose that I have also watched reviewed some (not all) of these n movies, and you know my scores. Devise a strategy to sug movies for me, from within the same set of the n movies, which I have not watched, but I may

Answer ALL questions, respecting the order of sub-questions. Problems C and D will be considered for bonus m

Indian Statistical Institute

PG Diploma in Business Analytics 1st Year : 2015–2016
Semester Examination
Subject: Stochastic Processes and applications
Date: 23/11/2015 Time: 3 hours Marks : 100

Answer Group-A and Group-B on separate answer scripts.
Notations used are as explained in the class.

Group-A

1. Let the n-step transition matrix $\mathbf{P}(m, m+n) = (p_{ij}(m, m+n))$ be the matrix of n-step transition probabilities $p_{ij}(m, m+n) = P(X_{m+n} = j | X_m = i)$. Prove that

$$p_{ij}(m, m+n+r) = \sum_k p_{ik}(m, m+n) p_{kj}(m+n, m+n+r).$$

Also prove that $\mathbf{P}(m, m+n) = \mathbf{P}^n$, the nth power of $\mathbf{P}$. [8 + 5]

2. (a) Let $\mu_i^{(n)} = P(X_n = i)$ be the probabilty mass function of X_n, and we write $\boldsymbol{\mu}^{(n)}$ for the row vector with entries $(\mu_i^{(n)} : i \in S)$ where S is the state space. Prove that $\boldsymbol{\mu}^{(m+n)} = \boldsymbol{\mu}^{(m)} \mathbf{P}^n$.
   (b) Define **transient** and **persistent non-null** state.
   (c) Define **closed** and **irreducible** set of states. [5 + 4 + 4]

3. Let $S = \{1, 2, 3, 4, 5, 6\}$ and the transition matrix of a Markov chain is

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{3}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Find the states which are **persistent**. Are they **non-null**?
Find their **mean recurrence times**. [3 + 1 + 8]

4. (a) If $i \leftrightarrow j$ (states i and j intercommunicate) then prove that they have the same period.
   (b) Consider the transition matrices of two Markov chains

$$\mathbf{P_1} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{bmatrix}, \mathbf{P_2} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3} & 0 \\ 0 & \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Find their period. [8 + 4]

Group-B

1. You have to choose 2 persons (distinct) from a group of 4 persons so that each person has equal probability of being chosen. You have a coin (may be biased) and you may toss it as many times as you like. Describe a procedure, with proofs, to make such a selection. [10]

2. Consider the time series model $x_t = c + x_{t-1} + w_t$, $t = 1, 2, \cdots$, c is a constant, $x_0 = 0$ and w_t is white noise. What is its mean function? [5]

3. Consider the MA process where $v_t = \frac{1}{4}(w_{t-2} + w_{t-1} + w_t + w_{t+1})$, w_t is white noise. Find out the autocorrelation function of this series. [10]

4. Consider the time series $x_t = \beta_1 + \beta_2 t + w_t$, where β_1, β_2 are known constants and w_t is white noise. Determine

   (a) if x_t is stationary.

   (b) if $y_t = x_t - x_{t-1}$ is stationary. [5 + 5]

5. Consider the following time series $1, 3, 7, 13.5, 23, 36, 53, 74.5, 101, 133$. Compute the second order differences and suggest a model on this series. [15]

INDIAN STATISTICAL INSTITUTE
Semester Examination

Course Name: PGDBA 2015 – 2017
Subject Name: Inference (BAISI3)
Date: 24th November, 2015 Maximum Marks: 100 Duration: 3 hours
Notes: Answer all questions. The paper carries 122 marks but the maximum you can score is 100.

1. Answer the following
   a. High end automobiles have sensors that can keep track of driving behaviour. These data are used by automobile insurance companies to understand the driving habits of individual drivers. The drivers are then classified as risky or non-risky and the insurance premiums for the individual drivers are fixed accordingly. Consider the task of classifying a driver as risky or non-risky as a problem of hypothesis testing and explain under what conditions the insurance company will be committing a type I error and under what condition a type II error would be committed. State your null and alternative hypotheses clearly. [3 + 3 = 6]
   b. A particular gambling game consists of drawing 3 chips at random from a box with replacement. The chips are numbered 1, 2, 3, ..., 12 and the prize is given on the basis of the sum of the numbers on the chips drawn. The organizer of the game claims that there are equal number of chips for each denomination and hence all numbers are equally likely to be drawn. However, some insiders tell you that there are twice as many chips numbered 1, 2, and 3 compared to other chips and hence these chips are twice as likely to occur. Suppose you want to verify this claim by formulating this problem as a problem of hypothesis testing.
      i. Are H_0 and H_1 simple or composite? Explain. [4]
      ii. Suppose you decide to believe the claim of the organizer in case the sum of numbers of 3 chips drawn randomly with replacement happens to be 11 or larger, and you decide to believe the claim of the insider in case the sum happens to be less than 11. Suppose further that the null hypothesis represents the claim of the organizer.
         A. List the points in the critical region
         B. Find α
         C. What is the power?
         D. Is the test one tailed or two tailed? Explain briefly. [4 + 4 + 4 + 3 = 15]
2. Answer the following
   a. A cigarette manufacturing company believes that the moisture level of tobacco has a negative impact on the quality of cigarettes manufactured. In order to verify this claim, the company has collected samples of 1000 good cigarettes and 1000 bad cigarettes. It was observed that out of the good cigarettes only 143 cigarettes contained tobacco with high moisture and the rest had low moisture. For the bad cigarettes, 562 cigarettes had high moisture and the rest had low moisture.
      i. Does this data support the claim of the company? Explain. [4]
      ii. Suppose you want to estimate the relative risk of getting a bad cigarette as the moisture level changes from high to low. Do you need any additional assumptions / data? Explain briefly the method of estimating relative risk in this situation. [6]

b. In analytics studies, data are often missing. Explain the concepts of Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR) briefly with examples. [3 + 3 + 3 = 9]

c. A large software company conducts employee satisfaction surveys. In order to reduce attrition, the company takes specific actions to address the grievances of employees with lower levels of satisfaction. The company carries out a second round of survey to find the effectiveness of the actions. The surveys are conducted online and attempts are made to cover all employees. However, some employees may not respond and hence questions of missing data may arise. While the response was almost complete for the first survey, several people did not respond for the second survey. On close scrutiny it was noted that about 50% of the people whose satisfaction score was lower than the 25^{th} percentile, did not respond. However, totally about 70% people responded and it is considered to be a good rate of response. Can you ignore the missing data and go ahead computing the effectiveness of the measures taken by the company? If yes, why? If not, explain what kind of errors do you expect if you go ahead with the available data? [6]

d. Suppose an e-commerce organization wants to estimate the proportion of visitors of their web page who ends up buying their product. For this estimation, every visit to the web site is considered an opportunity for sales. Suppose the organization has kept track of the last n visits of their website and found that these visits have resulted in k sales ($k \leq n$).

   i. What random variable is being studied? What distribution is it likely to follow? What parameter would you estimate? [1 + 1 + 1 = 3]

   ii. Write down the likelihood function and briefly explain how the function will change in case the parameter to be estimated changes. [2 + 2 = 4]

3. Answer the following

a. Software maintenance organizations receive service requests from their customers. It may be assumed that the number of service requests received in an hour follows a Poisson distribution. Suppose the organization has collected data on the time between subsequent arrivals of service requests. Suppose the observed values are $X_1, X_2, \ldots, X_N$ hours respectively and there are reasons to believe that the inter arrival times are independent of each other. How would you use these data to estimate the probability that the number of service requests received in an hour is less than some number k? [10]

b. A company manufacturing chemicals uses reagent A for some of its processes and the process is known to have a mean yield of at most 25 units. A chemist claims that a new reagent B would increase the yield. While the average yield of processes when reagent B is used is not known, the variance is known to be 64 units. Suppose you have taken a random sample of size 5 and obtained an average yield of 27 units on the basis of this sample. Construct a set of one sided confidence intervals for the mean yield when reagent B is used under the assumption that the yield follows normal distribution, and comment about the validity of the claim made by the chemist. Note that $\Phi(1.64) = 0.95$, $\Phi(1.96) = 0.975$, $\Phi(2.33) = 0.99$ and $\Phi(2.35) = 0.995$. [10]

c. Many companies are involved in carrying out exit polls to predict results of elections. In this context

   i. What is the population? [2]

   ii. Suppose the company is trying to predict the results for one constituency having 4 candidates representing 4 different political parties. What is the random variable and what distribution is it expected to follow? What parameter are they trying to estimate? [3 + 3 = 6]

4. Answer the following
   a. What are different measurement scales? [2 X 4 = 8]
   b. What measurement scales are used in the following
      i. Telephone numbers
      ii. Calendar years
      iii. Monthly salary of individuals [3 X 1 = 3]
   c. Suppose the lawyer's association of a city maintains the list of all lawyers. Suppose further that every lawyer belongs to exactly one law firm. In order to select a law firm randomly, it was proposed to choose a lawyer at random from the list of lawyers and the firm the lawyer belongs to is then chosen. Is this a random sample? Explain briefly. [3]
   d. Briefly explain the difference between explanatory and predictive analytics with an example for each. [6]
   e. An organization maintains an inventory of spares. The company keeps track of the level of inventory and places orders as soon as the level of inventory comes below a particular value. The users demand spares from time to time. The number of demands placed in a unit time interval is a random variable and its distribution is known from past experience. It is further known that the individual demands are independent of each other. The quantity demanded in every demand is a random variable and its distribution is also known from past experience. Suppose the lead time is the time to get replacement after placing an order and assume further that the lead time is fixed and known beforehand.
      i. Write a model for the distribution of the lead time demand in terms of the random variables. [7]
      ii. Explain how the technique of simulation may be used to understand the demand distribution by writing the simulation algorithm. Assume that the empirical distributions (frequency distributions) of the different random variables are available from past data. [10]

INDIAN STATISTICAL INSTITUTE

PGDBA 2015-2016, I Semester

Statistical Structures in Data (BAISI2)

End-semester examination

Maximum marks: 100 | 26 November 2015 | Maximum time: 3 hours

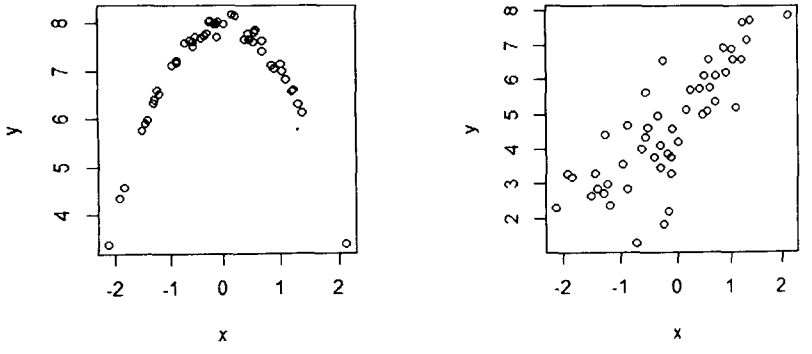
This examination is closed book, closed notes. Non-programmable calculators are allowed. The entire question paper is for 110 marks. The maximum you can score is 100.

1. a. Explain how you would generate 10 independent samples from the distribution function $F(x) = 1 - e^{-\sqrt{x}}$.

   b. Explain how you would generate 10 independent samples from the empirical distribution function of the 10 data values generated in part a.

[2 + 3 = 5]

2. Consider the scatter plots given below.



   a. Which of the data sets would lead to a higher sample correlation between X and Y?

   b. Which data set indicates a stronger dependence between X and Y? Explain.

[2 + 3 = 5]

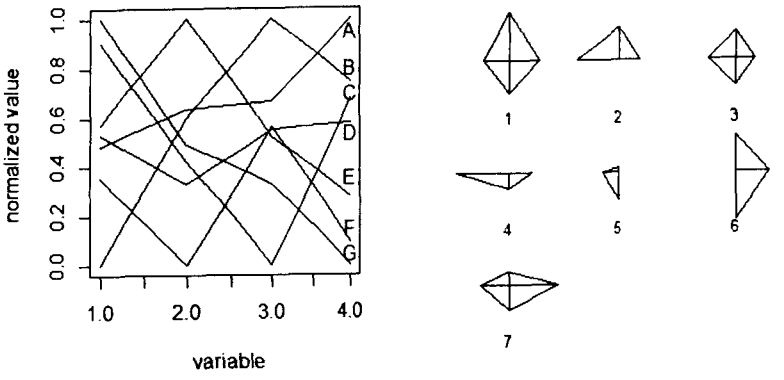
3. a. Determine the copula of the bivariate distribution

$$F(x,y) = \frac{(1-e^{-x})(1-e^{-y^2})}{1-0.5e^{-x-y^2}}, \quad x,y > 0.$$

   b. If the same marginal distributions as above are combined by using the independence copula, what would be the resulting bivariate distribution?

[4 + 1 = 5]

4. The line plots and star plots for seven cases (marked 'A' to 'G') are shown below.



The variable numbers and data labels in the line plot are mixed up in the star plot. After naming the four dimensions in the star plots as North, South, East and West, with obvious interpretations, identify which variable numbers of the line plot correspond to these dimensions.

[5]

5. Let X have the distribution $N_3(\mu, \Sigma)$, where

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}, \quad \mu = \begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{pmatrix}.$$

   a. Find distribution of $X_1 - 2X_2 + X_3$.
   b. Find real numbers a and b such that X_2 and $X_2 - aX_1 - bX_3$ are independent.

[5 + 5 = 10]

6. The paired multivariate data (X_i, Y_i), $i = 1, \ldots, N$, where both X_i and Y_i are p-dimensional random vectors, are assumed to be samples from a $2p$-variate normal distribution.

   a. Suggest a statistical test for checking if X_i and Y_i have the same mean.
   b. Mention a concrete example where this test may be useful.
   c. Give a set of simultaneous confidence intervals for the difference between the mean vectors, with specified coverage probability.

[5 + 2 + 3 = 10]

7. The sample variance covariance matrix of a three-variate data set is found to be

$$S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}.$$

Draw the scree plot for this data set.

[5]

8. The variance covariance matrix of a random vector $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$ is

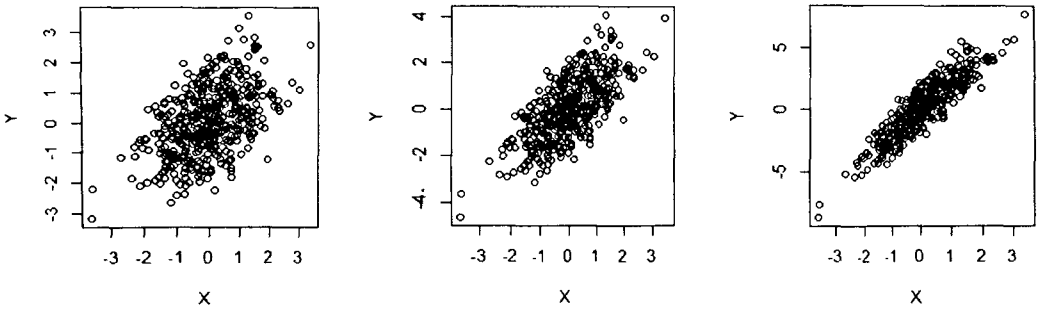
$$\Sigma = \begin{pmatrix} 8 & 2 & 4 \\ 2 & 2 & 2 \\ 4 & 2 & 4 \end{pmatrix}.$$

It is known that the random vector has a normal distribution.

a. Calculate the conditional variance of X, given Y and Z.
b. Calculate the multiple correlation coefficient of X with Y and Z.
c. Calculate the partial correlation of Y and Z, given X.
d. Which of the above calculations would be valid if the joint distribution of the random vector is not normal?

$[3 + 2 + 3 + 2 = 10]$

9. The three scatter-plots given below are obtained from samples (X,Y) generated from the bivariate normal distribution with correlation coefficients 0.5, 0.7 and 0.9, respectively. In each of the three cases, the two marginal distributions are standard normal. All the scatters are seen to be elongated along a line that is at about 45^0 with the X-axis.



a. Consider the line through origin inclined at an angle θ with the X-axis. Show that the distance of the origin from the foot of the perpendicular drawn from the point (X,Y) on this line is $X \cos\theta + Y \sin\theta$.
b. If each of X and Y have zero mean and unit variance, and they have correlation ρ, what is the variance of $X \cos\theta + Y \sin\theta$?
c. Consider the projection of all the points of a scatter on the line described in part a. How would the spread of these projections depend on θ?
d. By considering the principal components analysis of the theoretical variance covariance matrix of the bivariate normal distribution described above, explain why the elongation of the three scatters happens at about 45^0, irrespective of the value of the correlation coefficient.
e. Explain why the elongation is more prominent for a larger value of the correlation coefficient.
f. What would happen when the correlation is negative?

$[3 + 1 + 1 + 6 + 3 + 1 = 15]$

10. In an investigation on the consistency, determinants and uses of *accounting* and *market-value* measures of profitability, a factor analysis of *accounting profit measures* and *market estimates of economic profits* was conducted. The correlation matrix of *accounting historical, accounting replacement* and *market-value* measures of profitability for a sample of firms is as follows.

Variable	HRA	HRE	HRS	RRA	RRE	RRS	Q	REV
Historical return on assets (HRA)	1							
Historical return on equity (HRE)	0.738	1						
Historical return on sales (HRS)	0.731	0.520	1					
Replacement return on assets (RRA)	0.828	0.688	0.652	1				
Replacement return on equity (RRE)	0.681	0.831	0.513	0.887	1			
Replacement return on sales (RRS)	0.712	0.543	0.826	0.867	0.692	1		
Market Q ration (Q)	0.625	0.322	0.579	0.639	0.419	0.608	1	
Market relative excess value (REV)	0.604	0.303	0.617	0.563	0.352	0.610	0.937	1

The following rotated principal component estimates of factor loadings for a three factor model were obtained.

Variable	Estimated factor loadings		
	F_1	F_2	F_3
Historical return on assets (HRA)	0.433	0.612	0.489
Historical return on equity (HRE)	0.125	0.892	0.234
Historical return on sales (HRS)	0.296	0.238	0.887
Replacement return on assets (RRA)	0.406	0.708	0.483
Replacement return on equity (RRE)	0.198	0.895	0.283
Replacement return on sales (RRS)	0.331	0.414	0.789
Market Q ration (Q)	0.928	0.160	0.294
Market relative excess value (REV)	0.910	0.079	0.355
Cumulative proportion of total variance explained	0.287	0.628	0.908

a. Using the estimated factor loadings, determine the specific variances and communalities.

b. Assuming that estimated loadings less than 0.4 are small, interpret the three factors. Does it appear, for example, that *market-value measures* provide evidence of profitability distinct from that provided by *accounting measures*? Can you separate *accounting historical measures* of profitability from *accounting replacement measures*?

[8 + 7 = 15]

11. The variance-covariance matrix of a large data set, with nine variables has 2000 cases and nine variables, has been approximated by using principal components analysis with *five* principal components. An analyst claims that an alternative approximation obtained by using factor analysis with only *three* principal factors produces comparable approximation error (i.e., sum of squares of the approximation error matrix corresponding to the two approximations are comparable). The analyst claims that the latter model is more parsimonious (i.e., it has fewer number of free parameters). Is this claim correct? Explain.

[5]

12. Consider the following data on one predictor variable Z_1 and two responses Y_1 and Y_2.

Y_1	Y_2	Z_1
5	-3	-2
3	-1	-1
4	-1	0
2	2	1
1	3	2

a. Determine the least squares estimates of the parameters in the bivariate linear regression model

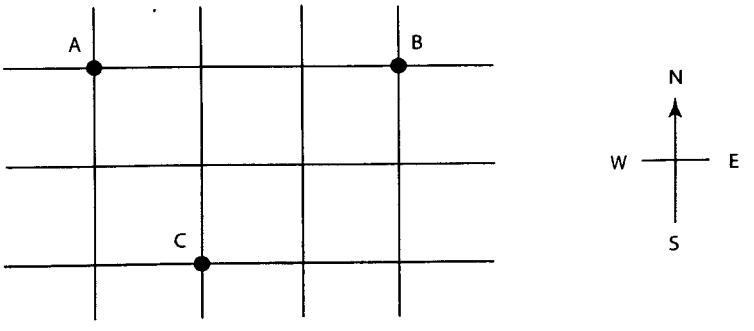
$$\begin{aligned} Y_{i1} &= \beta_{01} + \beta_{11} Z_{i1} + \varepsilon_{i1}, \\ Y_{i2} &= \beta_{02} + \beta_{12} Z_{i1} + \varepsilon_{i2}, \end{aligned} \qquad i = 1,2,3,4,5.$$

b. For the 5×2 response data matrix, calculate the matrix of fitted values and the residual matrix.

c. Calculate the usual unbiased estimator of the 2×2 variance-covariance matrix of the model error vector $\begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix}$, for the given data.

$[4 + 3 + 3 = 10]$

13. In a planned town, all the streets run from East to West or from North to South, and the successive streets are 200 meters apart. Three stores are located in street corners as indicated in the figure.



a. Compute the distance matrix for the three stores, where the distance is measured as the shortest walking distance via streets (i.e., no diagonal path is allowed).

b. Give a one-line argument, justifying that there is a multidimensional scaling with two dimensions, which makes Kruskal's measure of stress equal to zero.

c. Express the three data points in terms of the two components of a multidimensional scaling that makes the stress measure equal to zero.

d. Is the scaling unique (i.e., can there be another answer to part c)?

$[2 + 3 + 4 + 1 = 10]$

INDIAN STATISTICAL INSTITUTE

End-Semester Examination : 2015–16

Course : Post Graduate Diploma in Business Analytics (First Year)

Subject : Computing for Data Sciences : BAISI–4 for PGDBA–I

Date : 27 November 2015 | Maximum Marks : 100 | Duration : 3 Hours

You are allowed to use the Scribes and/or the Lecture Notes during the examination. Answer ALL questions.

Problem 1 [25]

Suppose that a single-feature dataset contains 100 observations (x, y), where the feature values x are independent samples drawn from $X \sim \mathcal{N}(0, 4)$, and y are the corresponding values of the target/output variable. The following R code was executed to fit an *ordinary least squares* linear model, and the corresponding output was obtained in terms of the coefficients and the residuals. Figure 1a represents the dataset (x, y) and the linear model $\hat{y} = h_\theta(x) = \theta_0 + \theta_1 x$ obtained.

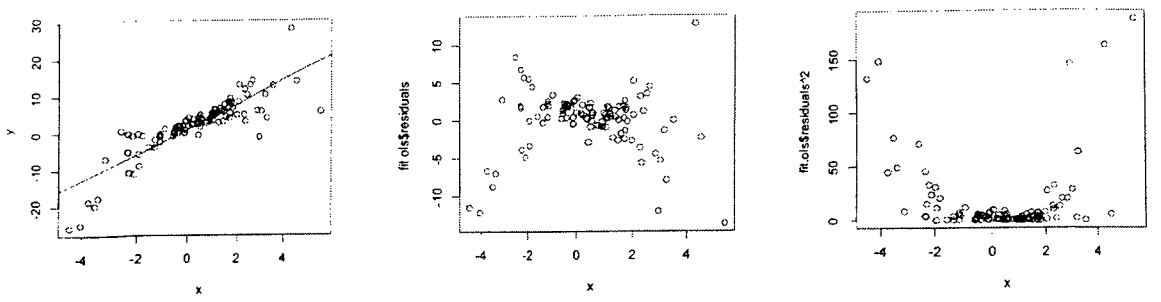
```
fit.ols <- lm(y~x)
fit.ols$coefficients
summary(fit.ols$residuals)
plot(x, y)
abline(fit.ols$coefficients)
plot(x, fit.ols$residuals)
plot(x, fit.ols$residuals^2)
```

```
(Intercept)           x
   0.859571    3.383583

    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-13.6900  -0.8885   0.6999   0.0000   1.9110  12.7600
```

Interpret the output. Suppose that the actual model for (x, y) is approximately linear, of the form $y = a_0 + a_1 x + \epsilon$, with a Gaussian error component $\epsilon \sim \mathcal{N}(\mu_\epsilon, \sigma_\epsilon^2)$. From the distribution of the *residuals* $(y - \hat{y})$, as in Figure 1b, and the distribution of the residuals squared, as in Figure 1c, infer if the distribution of ϵ, and in particular, if μ_ϵ and σ_ϵ, depend on x – explain your answer.

In such a scenario, will a *weighted least squares* model fit the dataset better? If not, explain why. If yes, suggest a suitable weight vector $\mathbf{w}(x)$ that can be used for the weighted linear regression.



(a) Linear model (b) Residuals (c) Residuals squared

Figure 1: Ordinary least squares regression on the dataset (x, y)

Problem 2 [25]

The following R code was executed on the stock iris dataset, to obtain the *decision tree* as shown in Figure 2. Given that there are exactly four features in the iris dataset – Sepal.Length, Sepal.Width, Petal.Length and Petal.Width – what can you say about the relative importance of the features in classifying the dataset into the three species - setosa, versicolor and virginica?

Note that there are 150 observations in total, with 50 observations each for the three species - setosa, versicolor and virginica. From the decision tree shown in Figure 2, calculate the *information gain* at the first level of the tree (i.e., at node 1 or the root node), in terms of Shannon entropy.

```
irisFit <- ctree(Species ~ Sepal.Length + Sepal.Width + Petal.Length
                                          + Petal.Width, data=iris)
plot(irisFit)
```

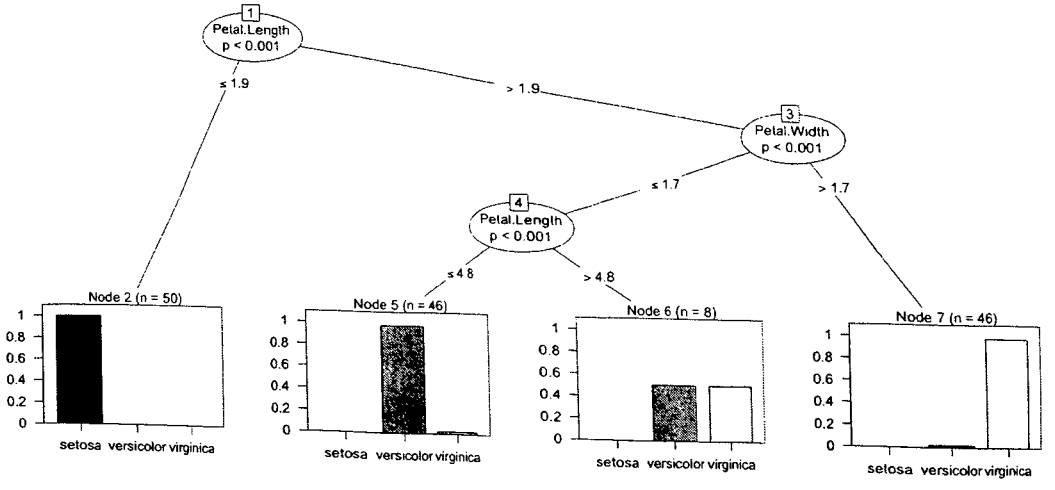


Figure 2: Decision Tree for the iris dataset

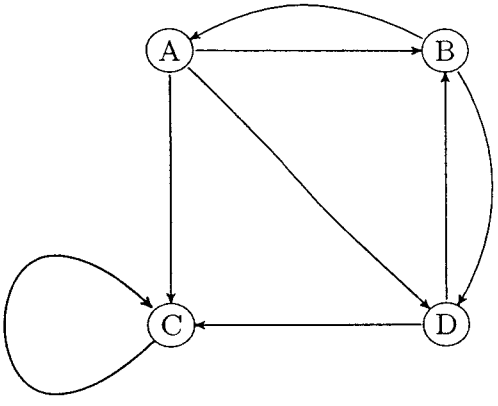
To test the classification accuracy of the decision tree on the training data itself (i.e., on the iris dataset), the following R code was executed, and the corresponding output was obtained. Comment on the accuracy of the classifier in terms of Type-I and Type-II errors in identifying each species.

```
trainPred <- predict(irisFit, iris)
table(trainPred, iris$Species)
```

```
trainPred     setosa versicolor virginica
  setosa          50          0         0
  versicolor       0         49         5
  virginica        0          1        45
```

Problem 3 [25]

Assuming the *random surfer* model of transition in a graph, where the surfer is equally likely to visit any connected node from the current position, create the transition matrix for the graph:



What do you expect to be the output of the following R code, where `transMat` denotes the transition matrix you created? You do not need to compute the eigenvectors. Justify your answer explaining the connection between the eigenvector `transVec` with the notion of PageRank in the network.

```
transEig <- eigen(transMat)
transVec <- transEig$vec[,1]
transVec
```

Does the eigenvector `transVec` provide a rational depiction of PageRank in this network? If so, justify your answer. If not, provide a solution to fix the computation for PageRank in the network.

Problem 4 [25]

Suppose that G is an undirected graph consisting of 6 vertices $\{A, B, C, D, E, F\}$, and an unknown number of edges. You are provided with neither the adjacency matrix of the graph, nor with the list of edges in it. However, you are provided with the following piece of R code that was executed on the graph G, and the corresponding output, as follows.

```
degMat <- diag(degree(G))
adjMat <- get.adjacency(G)
lapMat <- degMat - adjMat
lapEig <- eigen(lapMat)
lapEig
```

```
$values
[1] 4.561553e+00 3.000000e+00 3.000000e+00 3.000000e+00 4.384472e-01 5.313210e-16

$vectors
           [,1]         [,2]          [,3]       [,4]       [,5]      [,6]
[1,]  0.6571923  0.066054535  0.000000e+00  0.5735592 -0.2609565 0.4082483
[2,]  0.1845241 -0.734936838  2.807975e-02 -0.2059434  0.4647051 0.4082483
[3,]  0.1845241  0.668882304 -2.807975e-02 -0.3676157  0.4647051 0.4082483
[4,] -0.1845241 -0.060922638 -7.065490e-01 -0.2835670 -0.4647051 0.4082483
[5,] -0.6571923  0.066054535  7.979728e-17  0.5735592  0.2609565 0.4082483
[6,] -0.1845241 -0.005131897  7.065490e-01 -0.2899922 -0.4647051 0.4082483
```

Interpret the output to take an educated guess about the structure of the graph G, and draw a regular vertex-edge layout of the graph to depict your guess. From your reconstruction of the graph G, find the adjacency matrix (denoted by `adjMat` in the R code), the degree matrix (denoted by `degMat` in the R code), and the corresponding Laplacian matrix (denoted by `lapMat` in the R code).

Good luck! ☺

Booklet

.....
Supervisor's Signature

INDIAN STATISTICAL INSTITUTE

Semestral Examination

PGDBA

Subject: Fundamentals of Database Systems

Maximum Marks: 100 **Duration: 3 Hours**

Roll No: _____ Date: 30/11/2015

1. Mobile phones and all electronics communication devices are strictly prohibited inside the examination hall. Anybody Found in possession of such devices, even in switched off mode, will be expelled and his/her candidature will be cancelled.
2. Check for correctness the entries you have made above.
3. Your name should not be written anywhere inside this booklet.
4. All answers should be written in ink; pencil may be used for drawing and for numerical work.
5. Use the left side blank paper for rough work. Sheets should not be torn out.
6. Do not leave the examination hall until you have handed over this booklet to the invigilator.
7. Candidate will be subject to disciplinary action for violation of examination rule or for improper behavior in the examination hall.

Question No	1	2	3	4	5	Total
Marks Allotted	20	19	20	20	21	100
Marks Obtained						

INDIAN STATISTICAL INSTITUTE

End-Semester Examination: 2015-16

Course Name: PGDBA

Subject Name: Fundamentals of Data Base Systems

Date: 30/11/2015 **Maximum Marks: 100** **Duration: 3 Hours**

Note: Answer all questions at the assigned place. Mention your assumptions (if any).

1. Consider the following schema:

   Suppliers (supplier_id, name, city, state)
   Parts (part_id, name, size, colour)
   Catalog (supplier_id, part_id, quantity, price)

   a) Draw the Hierarchical Organizations and show graphically some example data. **[10]**

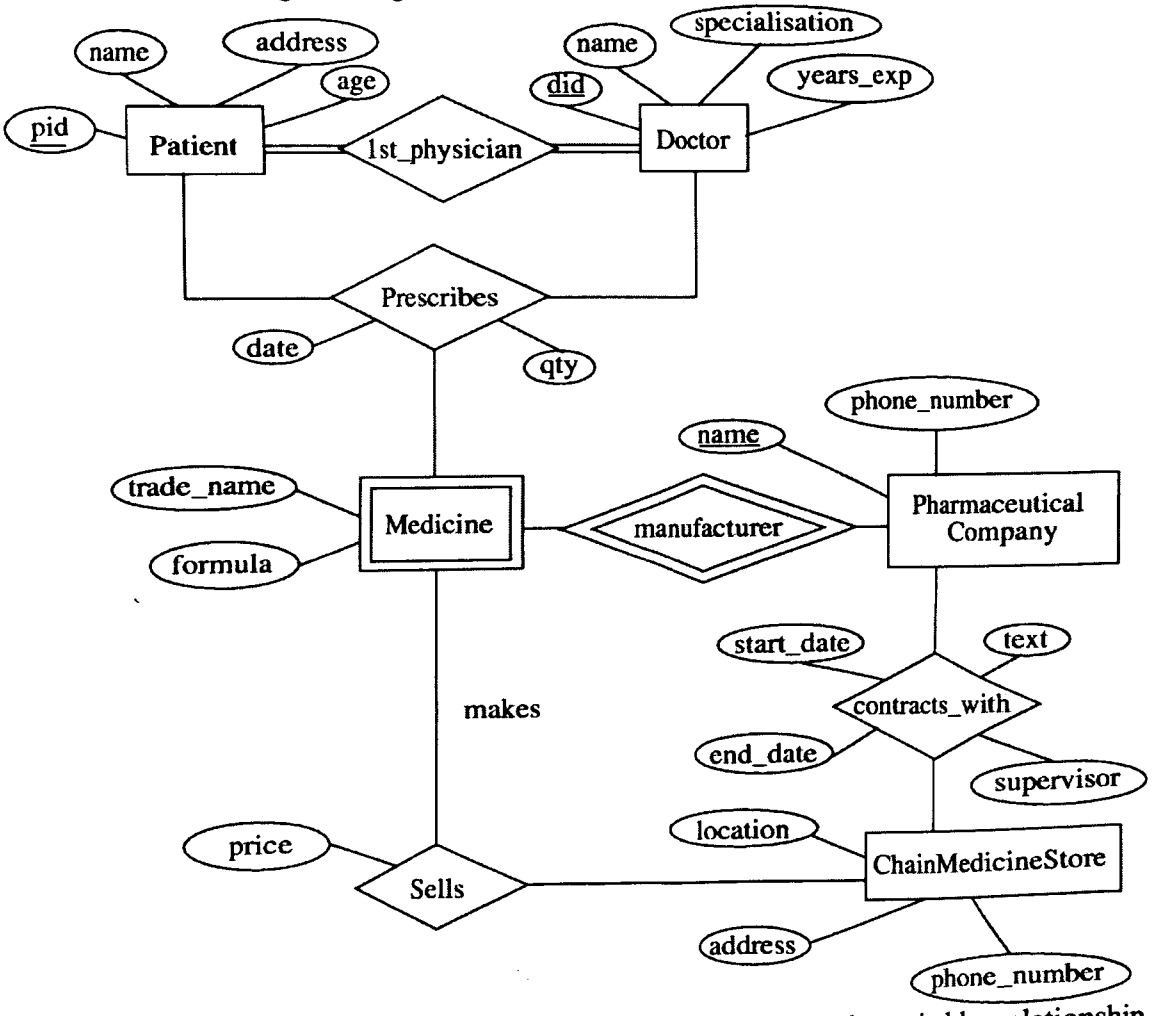
Answer 1a)

b) Draw the Directed Graph and show graphically some example data. **[10]**

Answer 1b)

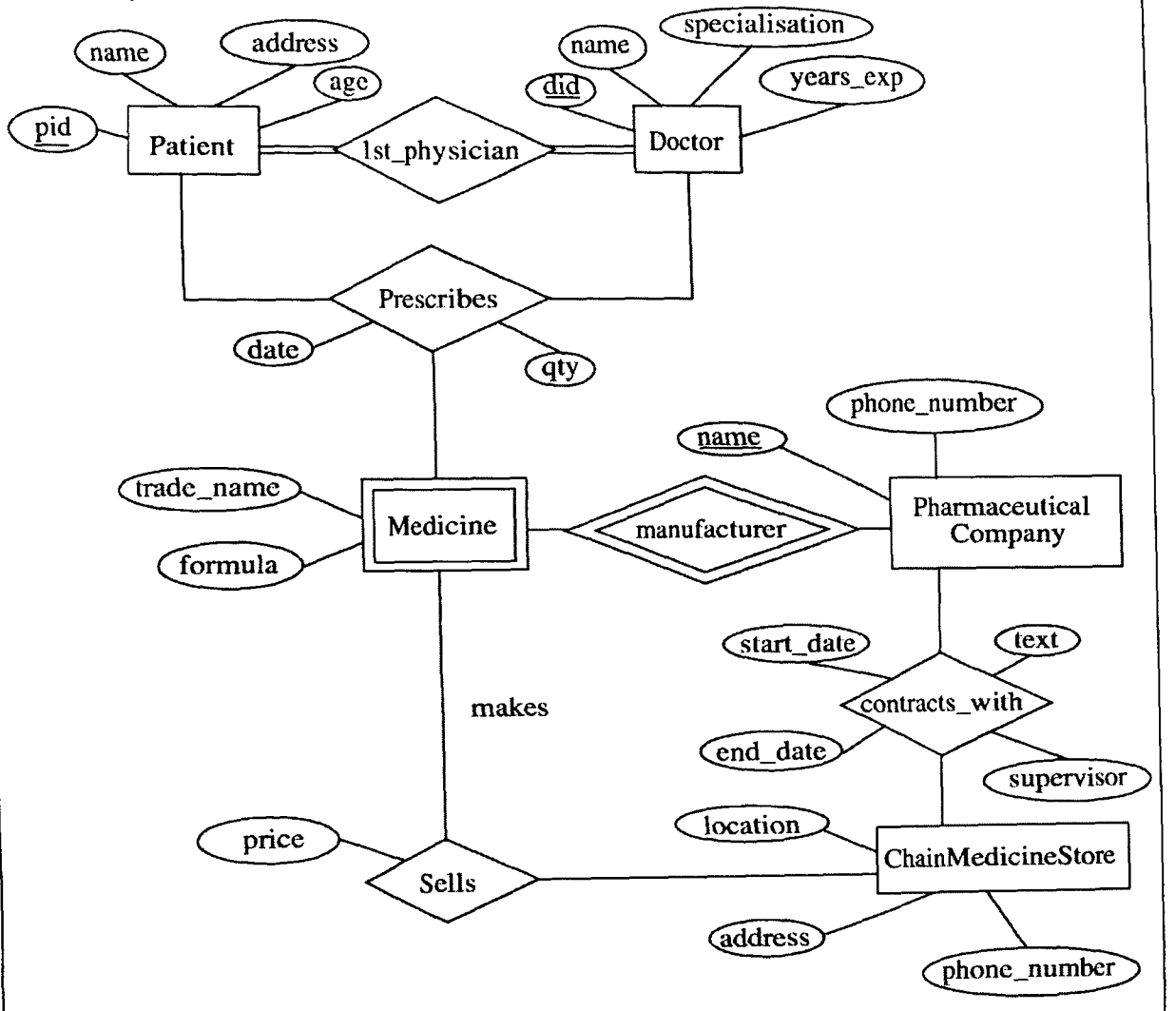
[10+10=20]

2. Consider the following ER Diagram:



a) Correct, if you can identify any mistakes. What will be the suitable relationship mappings (1:1, 1:n,n:1 and n:m). [07]

Answer 2a)



[12]

b) Formulate the problem statement.

Answer 2b)

[07+12=19]

3. a) In a range selection a range-partitioned attribute, it is possible that only one disk may need to be accessed. Describe the benefits and drawbacks of this property. **[06]**

Answer 3a)

b) List the possible type of failure in a distributed system. [06]

Answer 3b)

(c) Consider the following relation with schema {A,B,C,X,Y} with the following Functional Dependencies:

$A \rightarrow B$

$B \rightarrow C$

$X,Y \rightarrow A$

Decompose in BCNF. **[08]**

Answer 3c)

[06+06+08=20]

4. Write whether the following statements are **TRUE** or **FALSE**.

i. *Normalisation is a step by step reversible process of replacing a give collection of relations by successive collections in which the relations have a progressively simpler and more regular structure.*

ii. *Any two-attribute (atomic) relation may or may not be in BCNF.*

iii. *In Oracle, all views are not updatable.*

iv. *Distinct clause in select command eliminates rows that have exactly same contents in each column.*

v. *The USER_SYNONYMS view can provide information about private synonyms.*

vi. *The user SYSTEM owns all the base tables and user-accessible views of the data dictionary.*

vii. *All the dynamic performance views prefixed with V$ are accessible to all the database users.*

viii. *The USER_OBJECTS view can provide information about the tables and views created by the user only.*

ix. *DICTIONARY is a view that contains the names of all the data dictionary views that the user can access.*

x. *The data dictionary is created and maintained by the database administrator.*

xi. *The data dictionary views can consist of joins of dictionary base tables and user-defined tables.*

xii. *The usernames of all the users including the database administrators are stored in the data dictionary.*

xiii. *The USER_CONS_COLUMNS view should be queried to find the names of the columns to which a constraint applies.*

xiv. *Views with the same name but different prefixes, such as DBA, ALL and USER, use the same base tables from the data dictionary*

xv. *A simple view in which column aliases have been used cannot be updated.*

xvi. *A subquery used in a complex view definition cannot contain group functions or joins.*

xvii. *Rows cannot be deleted through a view if the view definition contains the DISTINCT keyword.*

xviii. *Rows added through a view are deleted from the table automatically when the view is dropped.*

xix. *The OR REPLACE option is used to change the definition of an existing view without dropping and re-creating it.*

xx. *The WITH CHECK OPTION constraint can be used in a view definition to restrict the columns displayed through the view.*

Answer 4

i.	ii.	iii.	iv.	v.
vi.	vii.	viii.	ix.	x.
xi.	xii.	xiii.	xiv.	xv.
xvi.	xvii.	xviii.	xix.	xx.

[10x2=20]

5. Find out the correct option(s):

i. A *Functional dependency between two or more non-key attributes is called*

A. Transitive dependency

B. Partial transitive dependency

C. Functional dependency

D. Partial functional dependency

E. None of the above

ii. *Who proposed the relational model?*

A. Bill Gates

B. Edgar Frank Codd

C. Herman Hollerith

D. Charles Babbage

E. None of them

iii. *A Relation is a*

A. Subset of a Cartesian product of a list of attributes

B. Subset of a Cartesian product of a list of domains

C. Subset of a Cartesian product of a list of tuple

D. Subset of a Cartesian product of a list of relations

E. None of the given option

iv. *Which of the following is true regarding Referential Integrity?*

A. Every primary-key value must match a primary-key value in an associated table

B. Every primary-key value must match a foreign-key value in an associated table

C. Every foreign-key value must match a primary-key value in an associated table

D. Every foreign-key value must match a foreign-key value in an associated table

E. None of them

v. *Identify the option that correctly matches the data types with the values.*

1 INTERVAL YEAR TO MONTH	**a** '2003-04-15 8:00:00 -8:00'
2 TIMESTAMP WITH LOCAL TIME ZONE	**b** '+06 03:30:16.000000'
3 TIMESTAMP WITH TIME ZONE	**c** '17-JUN-03 12.00.00.000000 AM'
4 INTERVAL DAY TO SECOND	**d** '+02-00'

A. 1-d, 2-c, 3-a, 4-b

B. 1-b, 2-a, 3-c, 4-d

C. 1-b, 2-a, 3-d, 4-c

D. 1-d, 2-c, 3-b, 4-a

E. None of the above

vi. Identify the option that contains the steps in the correct sequence in which the Oracle server evaluates a correlated subquery.

1) *The WHERE clause of the outer query is evaluated.*

2) *The candidate row is fetched from the table specified in the outer query.*

3) *The procedure is repeated for the subsequent rows of the table, till all the rows are processed.*

4) *Rows are returned by the inner query, after being evaluated with the value from the candidate row in the outer query.*

A. 4, 2, 1, 3
B. 4, 1, 2, 3
C. 2, 4, 1, 3
D. 2, 1, 4, 3
E. None of the above

vii. You executed the following SQL statements in the given order:

```
CREATE TABLE orders(
order_id NUMBER(3) PRIMARY KEY,
order_date DATE,
customer_id number(3)
);
INSERT INTO orders VALUES (100, '10-mar-2007',222);
ALTER TABLE orders MODIFY order_date NOT NULL;
UPDATE orders SET customer_id=333;
DELETE FROM orders;
```

The DELETE statement results in the following error:

ERROR at line 1: ORA-00942: table or view does not exist

What would be the outcome?

A. All the statements before the DELETE statement would be rolled back.

B. All the statements before the DELETE statement would be implicitly committed within the session.

C. All the statements up to the ALTER TABLE statement would be committed and the outcome of UPDATE statement would be rolled back.

D. All the statements up to the ALTER TABLE statement would be committed and the outcome of the UPDATE statement is retained uncommitted within the session.

E. None of these

Answer 5

i.	ii.	iii.	iv.	v.	vi.	vii.

[7x3=21]

INDIAN STATISTICAL INSTITUTE

PGDBA 2015-2016, I Semester

Statistical Structures in Data (BAISI2)

Back-paper examination

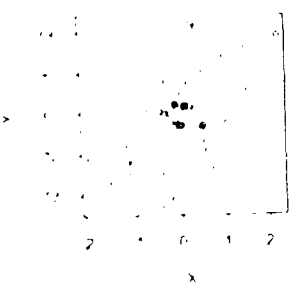
Maximum marks: 100 December 2015 Maximum time: 3 hours

This examination is closed book, closed notes. Non-programmable calculators are allowed. The entire question paper is for 100 marks. The maximum you can score is 45.

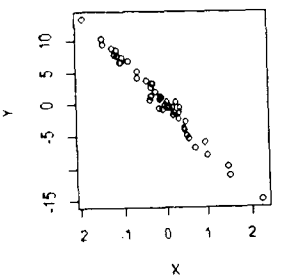
1. Six scatter diagrams for hypothetical data are shown below. The correlation coefficients, in scrambled order, are:

$$-0.99, \quad -0.8, \quad -0.4, \quad 0.05, \quad 0.5, \quad 0.9.$$

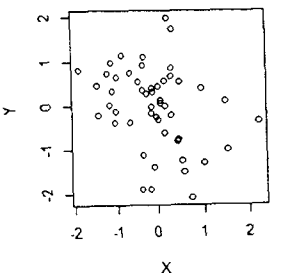
Match the scatter diagrams with the correlation coefficients.



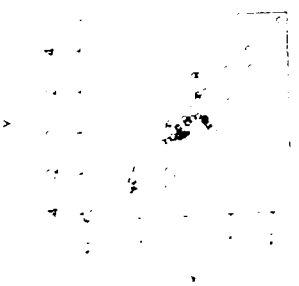
A



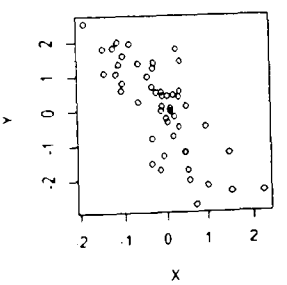
B



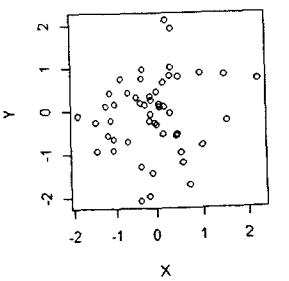
C



D



E



F

[6]

2. A personality test is applied to a large group of students. Five scores are shown below, in original units and in standard units. Fill in the blanks.

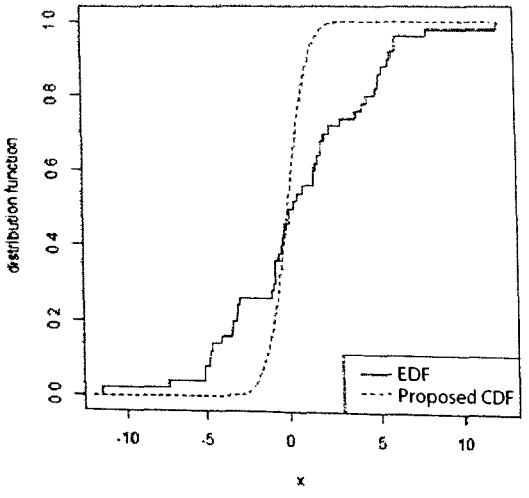
79	64	52	72	—
1.8	0.8	—	—	−1.4

[7]

3. You are looking at a list of 100 test scores, which have been converted to standard units. The first 10 entries are $-1.4, 3.5, 1.2, -0.13, 4.3, -5.1, -7.2, -11.3, 1.8, 6.3$. Do the numbers look reasonable, or is something wrong with the data? Explain.

[5]

4. The overlaid plots of the empirical distribution function (EDF) of a data set (sample size 50) and a proposed CDF are shown below. Draw a free-hand sketch of the PP plot for this data set. Label the axes clearly.



[10]

5. The data (X_i, Y_i, Z_i), $i = 1, \ldots, N$, are samples from the multinomial distribution with parameters n, p_1, p_2 and p_3 with $p_1 + p_2 + p_3 = 1$. Assuming that the parameter n is known, determine the maximum likelihood estimators of p_1, p_2 and p_3.

[7]

6. Consider a bivariate normal distribution with $\mu_1 = 0, \mu_2 = 2, \sigma_{11} = 2, \sigma_{22} = 1, \rho_{12} = 0.5$.

   a. Write the expression for the probability density function for this particular distribution.

   b. Sketch the constant-density contour that contains 95% of the probability.

   c. Specify the conditional distribution of X_1, given that $X_2 = x_2$, where the two random variables have the distribution described above.

[2 + 6 + 2 = 10]

7. Describe the Box-Cox family of transformations. How can the parameter of this transformation be chosen, so that normality of the transformed random variable is enhanced? Assuming that each element of a random vector is transformed separately, how can the multivariate normality of the transformed random vector be enhanced?

[2 + 4 + 4 = 10]

8. Consider four observations from the bivariate data, written as the data matrix

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

   a. Evaluate Hotelling's T^2 for testing $\mu = \begin{pmatrix} 7 \\ 11 \end{pmatrix}$, using the above data.

   b. Specify the distribution of an appropriately scaled version of T^2.

   c. Write an expression for the p-value of the test, simplified as much as possible.

[6 + 2 + 2 = 10]

9. A researcher considered three indices measuring the severity of heart attacks. The values of these indices for 40 heart attack patients arriving at a hospital emergency room produced the summary statistics

$$\bar{X} = \begin{pmatrix} 46.1 \\ 57.3 \\ 50.4 \end{pmatrix} \quad \text{and} \quad S = \begin{pmatrix} 101.3 & 63.0 & 71.0 \\ 63.0 & 80.2 & 55.6 \\ 71.0 & 55.6 & 97.4 \end{pmatrix}.$$

   All three indices are evaluated for each patient. Judge the differences in pairs of mean indices using 95% simultaneous confidence intervals.

[10]

10. Give an expression for the sample multiple correlation coefficient in a linear regression model, and explain its relation with the conditional and unconditional variances of the response.

[5]

11. Show that the value of a leverage in a linear model is always between 0 and 1.

[5]

12. Consider the variance-covariance matrix

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}. \quad -1 < \rho < 1.$$

a. Find the principal components and the proportion of the total population variance explained by each.

b. Draw the scree plot.

c. Suppose Σ is estimated in the usual manner from samples of the underlying distribution, and a principal components analysis is carried out. Draw a free hand sketch of the scatter plot of the first two principal components, which you expect to see for $\rho = 0.6$. Show the scales of the axes clearly.

d. Show that a factor analysis model with a single factor is appropriate for the above Σ.

[5 + 1 + 4 + 5 = 15]