# INDIAN STATISTICAL INSTITUTE

## Mid-Sem Examination

Post Graduate Diploma in Business Analytics 2017-18 (Semester-I)

*Stochastic Processes and Apllications*

Date: **September 11, 2017** Maximum Marks:100 Duration: **3 Hours**

**Note:** The question paper carries a total of 90 marks. You can answer as much as you can, but the maximum you can score is 80.

1. You are taking a multiple choice test (with no negative mark for a wrong answer) for which you have mastered 70% of the material. Assume this means that you have a 0.7 chance of knowing the answer to a random test question, and that if you don't know the answer to a question then you randomly select among the four answer choices. Finally, assume that this holds for each question, independent of the others. What is your expected score (as a percent) in the test?

   (6)

2. Let $C_1, C_2, \ldots, C_M$ be a partition of the sample space $S$, and $A$ and $B$ be two events. Suppose we know that

   (a) Given $C_i$, $A$ and $B$ are independent, $\forall i \in \{1, 2, \ldots, M\}$

   (b) $B$ is independent of all $C_i$ 's.

   Prove that A and B are independent.

   (6)

3. A box contains three coins: two regular coins and one fake two-headed coin (P(H)=1).

   (a) You pick a coin at random and toss it. What is the probability that it lands heads up?

   (b) You pick a coin at random and toss it, and get heads. What is the probability that it is the two-headed coin?

   (3+5=8)

4. I toss a coin repeatedly. The coin is unfair and $P(H) = p$. The game ends when for the first time two consecutive heads ($HH$) or two consecutive tails ($TT$) are observed. I win if $HH$ is observed and lose if $TT$ is observed. For example if the outcome is $HTHTT$, I lose. On the other hand, if the outcome is $THTHTHH$, I win. Find the probability that I win.

   (10)

5. $M$ leaves her child in daycare twice a week. Being a busy person she is often a few minutes late to pick her up. The daycare has a strict policy that parents need to be on time. They enforce this by charging Rs. 10 per minute for tardiness. Suppose that each day the amount of time in minutes that she is late follows an exponential distribution with mean 6.

   (a) Her child will be in daycare for 100 days this year. Estimate the probability that she will pay more than Rs 6300 in late fees?

   (b) The late fees were not effective in getting $M$ to arrive on time, so the daycare changed the rate to Rs. $10 \times (t^2 + t)$ for $t$ minutes of tardiness. On average, how much will $M$ pay in late fees each day?

   For part (a) we want a numerical answer. For part (b) leave your answer as an integral. Do.not compute it out.

   (12+8=20)

6. Let $X$ have range $[0,3]$ and density $f_X(x) = kx^2$. Let $Y = X^3$.

   (a) Find $k$ and the cumulative distribution function of $X$.

   (b) Compute E(Y ).

   (c) Write down an explicit formula, involving an integral, for Var(Y ).
   (No need to compute the value of the integral.)

   (d) Find the probability density function $f_Y(y)$ for $Y$.

(6+5+4+5=20)

7. Let $X_i$, $i = 1, \ldots, n$ be independent normal random variables that are normally distributed with respective parameters $\mu_i, \sigma_i^2, i = 1, \ldots, n$. Prove that $X = \sum_{i=1}^n X_i$ is normally distributed with parameters $\sum_{i=1}^n \mu_i$ and $\sum_{i=1}^n \sigma_i^2$.

(10)

8. Define a *Random Walk*. Prove that a Random Walk is a Markov chain.

(3+7=10)

# INDIAN STATISTICAL INSTITUTE

## Mid Semester Examination: 2017 – 18

Course Name: PGDBA
Subject Name: Inference (BAISI3)

Date: 12 September 2017 Maximum Marks: 80 Duration: 2.5 Hours

Notes: Answer all questions. The paper carries 85 marks but the maximum you can score is 80.

1. Answer the following:
   a. What do we mean by explanatory and response variables? [2 + 2 = 4]
   b. What is measurement? What are the different scales of measurement? [3 + 3 = 6]
   c. Consider the following problem scenarios and identify the response and explanatory variables in each case. Identify the scale of measurement of these variables. State what distribution may be used to model the response variable and identify the parameters that you may need to estimate.
      i. A commercial bank disbursing loans need to understand how likely are different loan applicant with varying age, gender and yearly income to return the money borrowed on time.
      ii. The Indian Railways find that the number of derailments are unacceptably large. They decide to divide the entire railway network into about 100 geographic regions. They decide to study the number of derailments in different periods of times for each of these regions along with data on the traffic density (total train kilometre per day divided by total length of track), average time gap between trains, minimum time gap between trains, and person-hours required for inspecting the entire track.
      iii. A manufacturing company measures the diameter of a shaft produced in large numbers in one machine. The manufacturer needs to ensure that the actual diameter remains within an interval for all shafts manufactured. Thus the manufacturer needs to assess the risk of producing shafts outside the given range.
      iv. The owner of a shop selling a particular type of chocolates does not want to have a stock out condition (i.e. she wants to avoid conditions when she is unable to satisfy customer demand). However, she is aware that keeping too large a stock would lead to high capital tie-up that would eventually reduce her profitability. She would like to know how much stock needs to be kept so that customer demands can be satisfied in 80%, 90%, 95% and 99% cases. She has observed that the pattern of sales are different across holidays and other days. [4 X 5 = 20]
2. A software service company sells its services to large corporate houses. In order to acquire prestigious clients, it participates in bids invited by large companies where it has not worked before. The sales team of the company observes that the chance of success in any one bid is low and it may often have to participate in several bids before getting the success for the first time. Let X denote the number of trials required to get the first success.
   a. What distribution is X likely to follow? What are its parameter(s)? [2 + 1 = 3]
   b. Suppose the company tries to acquire new clients frequently and they have observed the value of X for 20 new clients – say, $x_1, x_2, \ldots x_{20}$. If you assume that $x_1, x_2, \ldots x_{20}$

are i.i.d., what would that assumption really mean? Would you believe that the assumption is reasonable in practice? [4 + 3 = 7]

c. Suppose the i.i.d. assumptions are valid.
   i. Find the likelihood of the parameter(s) of interest.
   ii. Find the maximum likelihood estimate of the parameter(s)
   iii. Find the standard errors of the estimated parameter(s) [3 + 3 + 4 = 10]

3. Consider the following cases
   a. You are trying to compare the value of sales per day for shops A and B belonging to the same group. In order to present their performance visually, you have decided to draw ogives.
      i. Explain what ogives mean with special reference to this case.
      ii. Suppose you find that the ogive of shop A is everywhere higher than the ogive of shop B. Draw a free hand sketch of these ogives and explain what does this mean. [4 + 2 + 4 = 10]
   b. A shoe manufacturer uses chemical A to reduce the wear of shoe soles. A chemist claims that chemical B would be better. You can measure the actual rate of wear in your laboratory. Suppose you have collected 20 samples of soles manufactured using chemical A and 20 samples manufactured using chemical B. Suggest a suitable way of presenting this data so that differences between chemical A and B, if any, is likely to be brought out clearly. [4]
   c. It is assumed that the height of an adult male is a linear function of the height of both parents. Suppose a model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ has been fitted where $X_1$ is the height of the father, $X_2$ is the height of the mother and Y is the height of the son. Suppose the log likelihood of the entire model was – 110.43. You have noted that when $X_1$ is dropped, the log likelihood becomes – 134.63 and when $X_2$ is dropped, it becomes – 113.23. Assuming these figures are correct, whose height is likely to have more impact on the height of the son – height of mother or father? Explain briefly. [6]
   d. Consider the following data on the frequency of awarding death penalty in murder cases on the basis of the race of the defendant and victim. Out of 151 cases of white defendant and victim, 19 were awarded death penalty. Out of 9 cases of white defendant and black victim, none were awarded death penalty. Out of 63 cases of black defendant and white victim, 11 were awarded death penalty. Finally, out of 103 cases of black defendant and victim, 6 were awarded death penalty. Note that defendant is the person who has committed the crime, in this case murder. Note further that the sentence being awarded to the defendant is a random variable
      i. Present this data suitably in the form of a table
      ii. What distribution is the random variable likely to follow and what parameter(s) are of interest?
      iii. Notice that there are two explanatory variables – the race of the defendant and the race of the victim. Do you think that these variables have an interaction effect on the chance of death penalty being awarded to the defendant? [5 + 2 + 8 = 15]

# Indian Statistical Institute
# Foundations of Database Systems
# Mid Sem Exam

PGDBA 2017-19 1st Semester

Total marks: 60
Date: 13 September 2017
Time: 2 hours

1. Consider the following simplified schema for a telecom provider company (for example, Vodafone) storing their customer and call data. The table *Customer* (with primary key Phone_number) contains the records for all customers registered only with this company. The table *Call_records* contains the records for all outgoing calls made from all telephone numbers registered with this company to any number (within or outside its own network). Sample records for the two tables are given below. Assume that the duration of the calls are in minutes.

| Customer | | |
|---|---|---|
| Phone_number | Name | Billing_address |
| 9051333012 | D. Majumdar | 203 BT Road, Kolkata 700108 |
| 9778347829 | A. Ghosh | 45 Beckbagan Row, Kolkata 700029 |
| 9007283991 | M. Mitra | 205 BT Road, Kolkata 700108 |
| ... | ... | ... |

| Call_records | | | |
|---|---|---|---|
| From_number | To_number | Timestamp | Duration |
| 9778347829 | 9903238448 | 2017-02-23 10:10:35 | 3 |
| 9051333012 | 9007283991 | 2017-02-24 18:34:02 | 5 |
| ... | ... | ... | ... |

   (a) Suppose you want to determine the top phone friendships in a particular month. Define the *talktime* of two people (identified by their phone numbers) to be the total time they have conversed over the phone. Please note that the call could be initiated by either of the two friends. Write an SQL query to output the 10 best pairs of phone friends from the data available with the company in the month of July 2017.

   The output of your query should be a list of 10 such friendships, each with the numbers of the two people and the total duration of their conversation in that month. The list should be ranked by the total duration. [10]

   (b) Now you have to calculate the monthly bill of a particular customer. Suppose, the cost of calls within the same network is Rs. 0.20 and the cost for calls outside the network is Rs. 1.00 per minute. Given any particular number, write an SQL query to calculate the cost of all outgoing calls from that number in a given month. Assume any given number and given month in your query. The query should output just the amount in rupees. [10]

2. Consider a movie database with details about movies with their titles, actors, directors, producers, production houses, date of release, genre, length (in time), country, awards (if any), user ratings and other relevant information which are left to you to incorporate. For the questions below you only need to draw an ER diagram.

   (a) Determine at least five entities (none of which are weak entities) from the above scenario and describe them in the form of an ER diagram, with relevant attributes for these entities. [10]

   (b) Determine at least one weak entity with the relevant attributes and the discriminators in your diagram. Identify the identifying entity for the same in the diagram. [4]

   (c) Also specify at least six relationships between these entities in the diagram. [6]

3. Consider two tables $X(\underline{A}, B, C)$, with primary key $A$ and $Y(\underline{A, D}, E)$ with primary key $(A, D)$. If $X$ and $Y$ both have 50 records each, then can you determine the number of records in $X$ *right outer join* $Y$? [10]

4. Find a positive integer $n$ for which the Elias $\gamma$-encoding uses lesser number of bits than the variable byte encoding. Find another integer $m$ for which the variable byte encoding takes lesser number of bits than the Elias $\gamma$-encoding. For all cases, write the encoding with brief explanation. [10]

INDIAN STATISTICAL INSTITUTE

PGDBA 2017-2018, I Semester

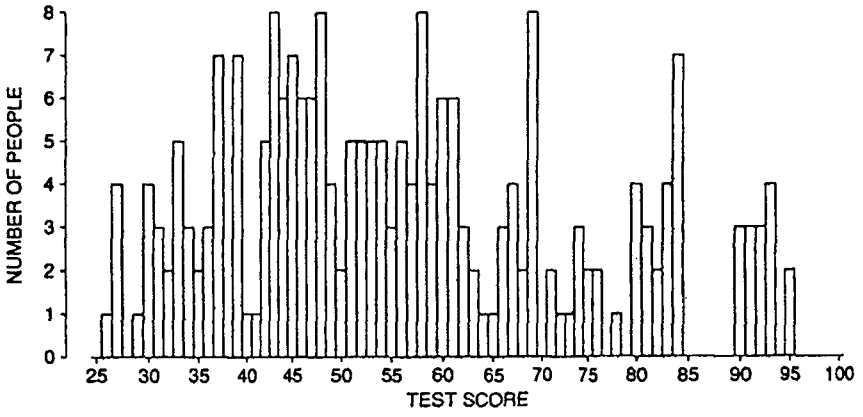
Statistical Structures in Data

Mid-semester examination

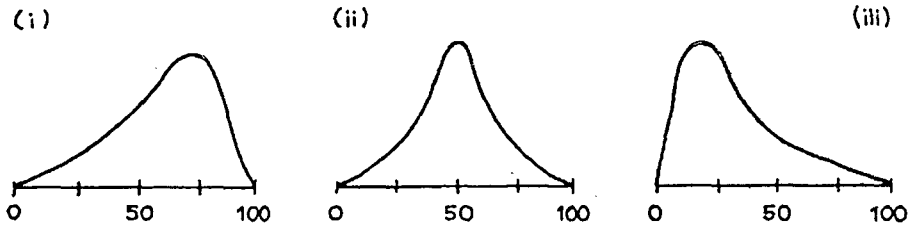
Maximum marks: 60 | 14 September 2017 | Maximum time: 2 hours

*This examination is closed book, closed notes. Non-programmable calculators are allowed. The questions carry a total of 65 marks. The maximum you can score is 60.*

1. Fifteen persons were selected from a pool of 223 candidates through a competitive examination. By examining the bar-chart of the frequency distribution of their scores (given below), explain whether there is any reason to suspect unfairness in the examination. [6]




2. Approximate shapes of histograms of three data sets are given below.




   a. In scrambled order, the averages of the three data sets are 30, 50 and 60. Match the histograms with the averages.

   b. For each histogram, indicate whether the median is less than, about the same as or greater than the average.

   c. For one of the data sets, the standard deviation is larger than the average. Which one is it? Explain. [3+3+3=9]

3. A certain data set has first and third quartile as 42 and 74. Assuming that the data set is a sample from a normally distributed population, and using the fact that $\Phi(0.6745) = 0.75$, calculate the probability that a single random number drawn from the same population would be greater than 91. [5]

4. Draw a freehand sketch of the plot generated from the following lines of R code (remember to label the axes and put appropriate tick marks on them).

```
y <- -log(runif(10))
qqplot(qnorm(ppoints(10), mean = 1, sd = 1), y)
```

[15]

P.T.O

5. According to a study conducted at Kaiser Permanente in Walnut Creek, California (findings published in the *American Journal Epidemiology*, 1977), users of oral contraceptives have a higher rate of cervical cancer than non-users, even after adjusting for age, education and marital status. Investigators concluded that the pill causes cervical cancer.

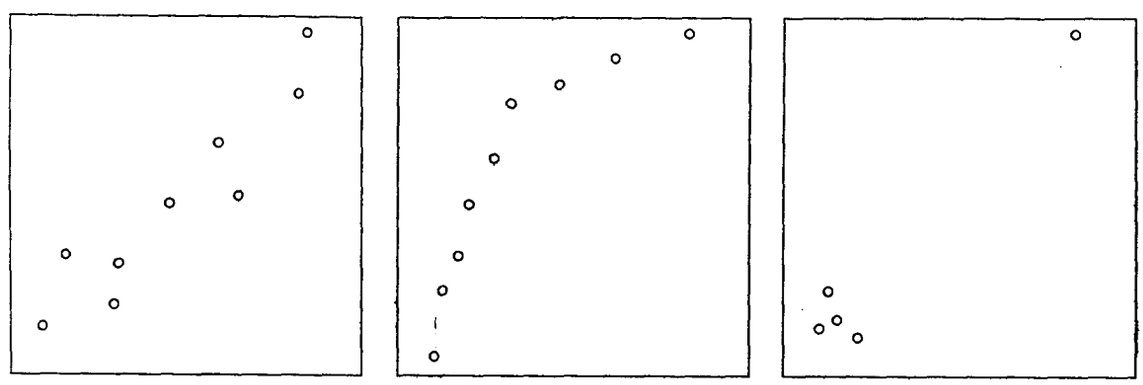
   a. Is this a controlled experiment or an observational study?

   b. Why did the investigators adjust for age, education and marital status?

   c. Women using the pill were likely to differ from non-users on another factor that might affect the risk of cervical cancer. What factor is that?

   d. Were the conclusions of the study justified by the data? Explain. [1+3+3+3=10]

6. For the three plots shown below, indicate (with reasons) the Pearson correlation coefficient will be (a) much larger than, (b) about the same as, or (c) much smaller than Spearman's rank correlation coefficient. [5]



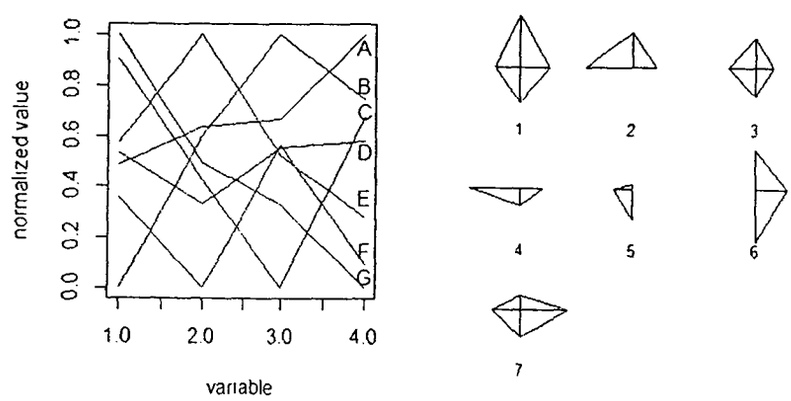
(i) (ii) (iii)

7. Identify the copula function implied by the bivariate distribution function

$$F(x,y) = (1-\theta)(1-e^{-x})\left(1-\frac{1}{1+y^2}\right) + \theta\min\left\{1-e^{-x}, 1-\frac{1}{1+y^2}\right\},\ x,y>0,$$

where $\theta$ is a parameter taking values in the interval (0,1). Identify the nature of dependence for the extreme values of $\theta$. [5+2=7]

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




The variable numbers 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. [8]

# INDIAN STATISTICAL INSTITUTE

## Mid-Sem Examination

Post Graduate Diploma in Business Analytics 2017-18 (Semester-I)

*Stochastic Processes and Apllications*

Date: **November 13, 2017** Maximum Marks:40 Duration: **1 hour 15 minutes**

1. $R$, $B$ and $k$ are positive integers. An urn initially contains $R$ red and $B$ black balls. A ball is drawn at random and its colour is noted. The ball is put back in the urn alongwith $k$ balls of the same colour. The process is repeated many times.

   (a) Calculate the probability that the second ball drawn is red?

   (b) Calculate the probabilility that the $i$th draw returns a red ball?

   (c) Given that the second ball drawn is red, calculate the probability the first ball was also red?

   (d) Calculate the probability that the first three balls are of the same colour.

   (3+5+4+4=16)

2. In a city with one hundred taxis, 1 is blue and 99 are green. A witness observes a hit-and-run by a taxi at night and recalls that the taxi was blue, so the police arrest the blue taxi driver who was on duty that night. The driver proclaims his innocence. A scientist tests the witness' ability to distinguish blue and green taxis under conditions similar to the night of accident. The data suggests that the witness sees blue cars as blue 99% of the time and green cars as blue 2% of the time. Use Probability theory to find out if there is a case for reasonable doubt.

   **(12)**

3. Let $X$ and $Y$ denote the horizontal and vertical miss distances when a bullet is fired at a target. Assume that

   (a) $X$ and $Y$ are independent continuous random variables having differntiable density functions.

   (b) The joint density $f(x,y) = f_X(x)f_Y(y)$ of $X$ and $Y$ depends on $(x,y)$ only through $x^2+y^2$.

   Derive the distribution of $X$.

   **(12)**

**INDIAN STATISTICAL INSTITUTE**

**FIRST SEMESTER EXAMINATION: 2017 – 18**

**Course Name: PGDBA**

**Subject Name: Inference (BAISI3)**

**Date: 21.11.2017** **Maximum Marks: 100** **Duration: 3 hours**

**Notes, if any: Answer any four questions.**

1. A box contains 7 tickets. Five tickets belong to students and the other two belong to faculty. Two tickets are drawn from the box at random without replacement (i.e. a ticket drawn is not returned to the box) to determine the two winners. The null hypothesis is that the draw is random. The alternative hypothesis is that the box is rigged such that the first ticket drawn belongs to a faculty member and the second ticket is then randomly selected from the remaining six tickets.
   a. Are the hypotheses simple or composite? Explain briefly. [5]
   b. Suppose the decision rule is to reject the null hypothesis if both tickets drawn belong to faculty members. Find $\alpha$. Find power. [10]
   c. Suppose, instead, the decision rule is to reject the null hypothesis if the first ticket drawn belongs to a faculty member. Find $\alpha$. Find power. [10]
2. A chemical manufacturing company believes that the yield of a particular chemical is adversely impacted by its moisture content, i.e. the yield would be lower in case the moisture is high. In order to verify the belief, the organization decided to collect some data from the production log books where the level of moisture as well as the yield are recorded for every batch being produced by the company.
   a. In this specific context
      i. Identify the explanatory (X) and response (Y) variables. [1 + 1 = 2]
      ii. What distributions are X and Y likely to follow? [1 + 1 = 2]
   b. Suppose the organization has identified 100 batches with high yield and 100 batches with low yield from their past production records. It was noted that among the batches with high yield, only 37 batches had high moisture. It was also observed that in the batches with lower yield, 66 batches had high moisture. [In this exercise both yield and moisture were marked as high or low, i.e. they were converted into binary variables].
      i. Present the data suitably in a 2 X 2 table. [5]
      ii. Note that the analyst has first identified batches with high and low yield and then found whether the moisture was high or low. Keeping this in view, write down the conditional probabilities that could be estimated from the 2 X 2 table. Can you use these conditional probabilities directly (without using any other information) to verify the claim that moisture impacts yield? [4 + 4 = 8]
      iii. Suppose you want to test the claim that moisture impacts yield.
         a. Write the problem in terms of tests of hypotheses, i.e. write the null and the alternative hypotheses. Identify the parameters of the distributions on which you are attempting to carry out tests. [6]
         b. What do type I and type II errors mean in this case? [2]
3. Answer the following
   a. Let $Y_1, Y_2, \ldots Y_n$ be a random sample from a Poisson distribution with mean $\lambda$.
      i. What distribution does $U = Y_1 + Y_2 + \ldots + Y_n$ follow? [4]
      ii. Find the likelihood $L(\lambda / \mathbf{y})$ of $\lambda$ given that $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$. [We have used bold face y to dentoe the vector of observed Y values]. [5]
      iii. Show that $L(\lambda / \mathbf{y}) = h(\mathbf{y}). L^*(\lambda / \mathbf{y})$ – where $h(\mathbf{y})$ does not depend on $\lambda$ and $L^*(\mathbf{y})$ can be computed even when the individual values of $y_1, y_2, \ldots, y_n$ are not known but only n and the sample average is known. [4]

20.11.2017

b. Explain briefly the concept of likelihood ratio test (i.e. how the concept of likelihood ratio is used to construct the test statistic) for testing a null hypothesis $H_0$ against the alternative hypothesis $H_1$. [6]

c. What is the scale of measurement for the following random variables? Give a very brief (not exceeding 3 or 4 lines) explanation for each
   i. Your PGDBA roll number
   ii. The weight lost (or gained) expresesed in percentage when on a particular diet [both weight loss / gain and percentage change are computed with respect to weight at the start of the programme]
   iii. The ratings provided by different viewers of a particular TV programme on a ten point [1 – 10] scale [3 X 2 = 6]

4. A manufacturing company produces certain products using 5 different machines. The company suspects that the propotion of defective items produced by the machines are different. The company wants to check the suspicion and collects $n_1, n_2, \ldots, n_5$ products randomly from each machine. Suppose the number of defective items were found to be $d_1, d_2, \ldots, d_5$ respectively.
   a. Provide a tabular structure to present this data. [3]
   b. Suppose it is believed that machine 1 is the best.
      i. Show how the relative risk of getting defective items from any machine j, j = 2, 3, 4 or 5 with respect to machine 1 may be computed. [2]
      ii. Suppose the standard errors of the estimated relative risks are 0.27, 0.38, 0.22 and 0.29 for machines 2, 3, 4 and 5 with respect to machine 1 respectively. Suppose from the estimated relative risks it was found that machines 2, 3, 4 and 5 are 32%, 39%, 45% and 34% more likely to produce defective items. Do you support the claim that machine 1 is really better than all the others with respect to production of defective items? [Notice that Relative Risk for machine j, j = 2,3,4,5 vs.machine 1 is P(a randomly selected item from machine j is defective) / P(a randomly selected item from machine 1 is defective)]. [6]
   c. Suppose you want to check whether the machines are different with respect to production of defective items
      i. Write this problem in terms of testing of hypotheses [3]
      ii. Explain how you will carry out the test (i.e. how you will construct the test statistic, what distribution will it follow and what rule will be used to raject the null hypothesis). [6]
   d. Suppose 100 items were collected randomly from machine 3 and they were checked. Ten (10) items were found to be defective. Construct an approximate 95% confidence interval for the true population proportion defective. Do you think that the chance that the interval constructed and reported by you contains the population proportion defective is about 95%? Provide a brief explanation. [3 + 2 = 5]

5. A particular type of shell (a shell is an ammunition fired from large guns) used by the army is not supposed to explode even if it hits a target in a distance of 20 meters or less from the muzzle of the gun [i.e. the point where the shell leaves the barrel of the gun]. If the shell hits a target after 20 meters, it should fire and it must definitely fire if it hits a target at 200 meters or beyond.
   a. Draw the target distance vs. probability of firing curve. Show the ideal curve and the curve that we are actually expected to get. [3 + 3 = 6]
   b. When a consignment of shells consisting of a few hundred shells are offered to the army, they fire 20 rounds [i.e. they fire 20 shells chosen at random] where a target is placed at a distance of 20 meters. The army rejects the entire lot in case even a single round explodes. Does this procedure provide adequate confidence that shells will not explode when it accidentally hits something placed at a distance of less than 20 meters during actual warfare? Explain briefly. [Assume that a chance of failure > 1% should be considered large as shells are fired in quick succession during actual war]. [7]

20.11.2017

c. These shells are supposed to have an avearge range of 5400 meters and the variation should not be more than plus or minus 200 meters. The army fires 100 rounds and measures the actual distance traversed by each shell. [The range is nothing but the actual distance].

   i. Note that the distance traversed by individual shells are random variables. What distribution is it likely to follow? [2]

   ii. Express the problem of accepting or rejecting a consignment of shells on the basis of the 100 rounds fired in terms of confidence intervals of the parameters to be estimated. Suggest how the confidence intervals may be constructed in this particular case. Explain what statistic will be used and what distribution are the statistics likely to follow. [4 + 6 = 10]

**INDIAN STATISTICAL INSTITUTE**

**FIRST SEMESTER EXAMINATION: 2017 – 18**

**Course Name: PGDBA**

**Subject Name: Inference (BAISI3)**

**Date: 21.11.2017** **Maximum Marks: 100** **Duration: 3 hours**

**Notes, if any: Answer any four questions.**

1. A box contains 7 tickets. Five tickets belong to students and the other two belong to faculty. Two tickets are drawn from the box at random without replacement (i.e. a ticket drawn is not returned to the box) to determine the two winners. The null hypothesis is that the draw is random. The alternative hypothesis is that the box is rigged such that the first ticket drawn belongs to a faculty member and the second ticket is then randomly selected from the remaining six tickets.
   a. Are the hypotheses simple or composite? Explain briefly. [5]
   b. Suppose the decision rule is to reject the null hypothesis if both tickets drawn belong to faculty members. Find $\alpha$. Find power. [10]
   c. Suppose, instead, the decision rule is to reject the null hypothesis if the first ticket drawn belongs to a faculty member. Find $\alpha$. Find power. [10]
2. A chemical manufacturing company believes that the yield of a particular chemical is adversely impacted by its moisture content, i.e. the yield would be lower in case the moisture is high. In order to verify the belief, the organization decided to collect some data from the production log books where the level of moisture as well as the yield are recorded for every batch being produced by the company.
   a. In this specific context
      i. Identify the explanatory (X) and response (Y) variables. [1 + 1 = 2]
      ii. What distributions are X and Y likely to follow? [1 + 1 = 2]
   b. Suppose the organization has identified 100 batches with high yield and 100 batches with low yield from their past production records. It was noted that among the batches with high yield, only 37 batches had high moisture. It was also observed that in the batches with lower yield, 66 batches had high moisture. [In this exercise both yield and moisture were marked as high or low, i.e. they were converted into binary variables].
      i. Present the data suitably in a 2 X 2 table. [5]
      ii. Note that the analyst has first identified batches with high and low yield and then found whether the moisture was high or low. Keeping this in view, write down the conditional probabilities that could be estimated from the 2 X 2 table. Can you use these conditional probabilities directly (without using any other information) to verify the claim that moisture impacts yield? [4 + 4 = 8]
      iii. Suppose you want to test the claim that moisture impacts yield.
         a. Write the problem in terms of tests of hypotheses, i.e. write the null and the alternative hypotheses. Identify the parameters of the distributions on which you are attempting to carry out tests. [6]
         b. What do type I and type II errors mean in this case? [2]
3. Answer the following
   a. Let $Y_1, Y_2, \ldots Y_n$ be a random sample from a Poisson distribution with mean $\lambda$.
      i. What distribution does $U = Y_1 + Y_2 + \ldots + Y_n$ follow? [4]
      ii. Find the likelihood $L(\lambda / \mathbf{y})$ of $\lambda$ given that $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$. [We have used bold face y to dentoe the vector of observed Y values]. [5]
      iii. Show that $L(\lambda / \mathbf{y}) = h(\mathbf{y}). L^*(\lambda / \mathbf{y})$ – where $h(\mathbf{y})$ does not depend on $\lambda$ and $L^*(\mathbf{y})$ can be computed even when the individual values of $y_1, y_2, \ldots, y_n$ are not known but only n and the sample average is known. [4]

20.11.2017

b. Explain briefly the concept of likelihood ratio test (i.e. how the concept of likelihood ratio is used to construct the test statistic) for testing a null hypothesis $H_0$ against the alternative hypothesis $H_1$. [6]

c. What is the scale of measurement for the following random variables? Give a very brief (not exceeding 3 or 4 lines) explanation for each
   i. Your PGDBA roll number
   ii. The weight lost (or gained) expresesed in percentage when on a particular diet [both weight loss / gain and percentage change are computed with respect to weight at the start of the programme]
   iii. The ratings provided by different viewers of a particular TV programme on a ten point [1 – 10] scale [3 X 2 = 6]

4. A manufacturing company produces certain products using 5 different machines. The company suspects that the propotion of defective items produced by the machines are different. The company wants to check the suspicion and collects $n_1, n_2, \ldots, n_5$ products randomly from each machine. Suppose the number of defective items were found to be $d_1, d_2, \ldots, d_5$ respectively.
   a. Provide a tabular structure to present this data. [3]
   b. Suppose it is believed that machine 1 is the best.
      i. Show how the relative risk of getting defective items from any machine j, j = 2, 3, 4 or 5 with respect to machine 1 may be computed. [2]
      ii. Suppose the standard errors of the estimated relative risks are 0.27, 0.38, 0.22 and 0.29 for machines 2, 3, 4 and 5 with respect to machine 1 respectively. Suppose from the estimated relative risks it was found that machines 2, 3, 4 and 5 are 32%, 39%, 45% and 34% more likely to produce defective items. Do you support the claim that machine 1 is really better than all the others with respect to production of defective items? [Notice that Relative Risk for machine j, j = 2,3,4,5 vs.machine 1 is P(a randomly selected item from machine j is defective) / P(a randomly selected item from machine 1 is defective)]. [6]
   c. Suppose you want to check whether the machines are different with respect to production of defective items
      i. Write this problem in terms of testing of hypotheses [3]
      ii. Explain how you will carry out the test (i.e. how you will construct the test statistic, what distribution will it follow and what rule will be used to raject the null hypothesis). [6]
   d. Suppose 100 items were collected randomly from machine 3 and they were checked. Ten (10) items were found to be defective. Construct an approximate 95% confidence interval for the true population proportion defective. Do you think that the chance that the interval constructed and reported by you contains the population proportion defective is about 95%? Provide a brief explanation. [3 + 2 = 5]

5. A particular type of shell (a shell is an ammunition fired from large guns) used by the army is not supposed to explode even if it hits a target in a distance of 20 meters or less from the muzzle of the gun [i.e. the point where the shell leaves the barrel of the gun]. If the shell hits a target after 20 meters, it should fire and it must definitely fire if it hits a target at 200 meters or beyond.
   a. Draw the target distance vs. probability of firing curve. Show the ideal curve and the curve that we are actually expected to get. [3 + 3 = 6]
   b. When a consignment of shells consisting of a few hundred shells are offered to the army, they fire 20 rounds [i.e. they fire 20 shells chosen at random] where a target is placed at a distance of 20 meters. The army rejects the entire lot in case even a single round explodes. Does this procedure provide adequate confidence that shells will not explode when it accidentally hits something placed at a distance of less than 20 meters during actual warfare? Explain briefly. [Assume that a chance of failure > 1% should be considered large as shells are fired in quick succession during actual war]. [7]

20.11.2017

c. These shells are supposed to have an avearge range of 5400 meters and the variation should not be more than plus or minus 200 meters. The army fires 100 rounds and measures the actual distance traversed by each shell. [The range is nothing but the actual distance].

   i. Note that the distance traversed by individual shells are random variables. What distribution is it likely to follow? [2]

   ii. Express the problem of accepting or rejecting a consignment of shells on the basis of the 100 rounds fired in terms of confidence intervals of the parameters to be estimated. Suggest how the confidence intervals may be constructed in this particular case. Explain what statistic will be used and what distribution are the statistics likely to follow. [4 + 6 = 10]

# INDIAN STATISTICAL INSTITUTE
# First Semester Examination: 2017-18

PGDBA 2017-19 1st Year
Foundations of Database Systems

Date: 22 November 2017
Maximum marks: 100
Time: 3 Hours

**For every answer you must explain the logic behind it.**

1. Consider two vectors $\mathbf{u}$ and $\mathbf{v}$ of the same dimension $n$, stored as tables `U(ind,val)` and `V(ind,val)` with the same schema. A row $(i, u_i)$ of table `U` specifies that the $i$-th element of vector $\mathbf{u}$ has value $u_i$ (similarly for $\mathbf{v}$, respectively). Only the non-zero entries of the vectors are stored in the corresponding tables. For example, the if the vector $\mathbf{u}$ equals $(0, 1, 3, 0, 2, 0)$, then it is represented in table `U` as:

| ind | val |
|---|---|
| 2 | 1 |
| 3 | 3 |
| 5 | 2 |

Write relational algebra expressions or SQL queries to compute each of the following. Explain why your solution would work.

(a) The sum $\mathbf{u} + \mathbf{v}$ of the two vectors $\mathbf{u}$ and $\mathbf{v}$.

(b) The dot product $\mathbf{u}^T\mathbf{v}$ of the two vectors $\mathbf{u}$ and $\mathbf{v}$.

(10 + 10 = 20 marks)

2. Suppose an asymmetric social network similar to Twitter (that is, users follow other users, but mutual following is not essential) stores its data in the table `following(src_id VARCHAR(32), dest_id VARCHAR(32))`, where every row (*src_id*, *dest_id*) indicates that the user with ID *src_id* follows the user with ID *dest_id*. The primary key of the table is `(src_id, dst_id)`.

Write a relational algebra expression or SQL query to output all the user ids of the users who are followers of followers (but not direct followers) of the user with id `iamsachin`. Provide a brief explanation for your solution.

(10 marks)

3. Let $R = (A, B, C, D, E, G)$ be a relation with functional dependencies $A \rightarrow BC$, $C \rightarrow D$ and $AE \rightarrow G$.

   (a) Define prime and non-prime attributes.

   (b) Show that $R$ is not in 2-NF.

   (c) Provide a decomposition of $R$ such that the resulting relations are in 2-NF.

   $(4 + 8 + 8 = 20$ marks$)$

4. Let $R = (A, B, C, D, E, G)$ be a relation with functional dependencies $AB \rightarrow CD$ and $D \rightarrow EG$.

   (a) Show that $R$ is in 2-NF.

   (b) Show that $R$ is not in BCNF.

   (c) Provide a decomposition of $R$ such that the resulting relations are in BCNF.

   $(8 + 4 + 8 = 20$ marks$)$

5. Let $R$ be a relation and let $X, Y \subseteq R$ be attribute sets in $R$. Prove or disprove the following: $(X^+Y^+)^+ = (XY)^+$. Note that, by $XY$, we denote the union of two attribute sets $X$ and $Y$.

   (10 marks)

6. Suppose two large text files are residing in HDFS such that no single node contains the entire content of any of the files. Write map and reduce functions (with explanation) so that the output consists of all distinct words which are present in both of the files. For example, if the first file reads as `"there was a brown crow and there was a black panther"`, and the second file reads `"once upon a time there was a king"`, then the output should be `"there was a brown crow and black panther once upon time king"`. The order of the words in the sequence does not matter.

   (10 marks)

7. Let `R(A, B, C)` and `S(B, D, E)` be two tables residing in HDFS, so that none of the tables have all of its records in a single node of the cluster. Explain how to perform

   ```
   R natural left outer join S
   ```

   using MapReduce. Write your map and reduce functions.

   (10 marks)

# INDIAN STATISTICAL INSTITUTE

## Semestral Examination

Post Graduate Diploma in Business Analytics 2017-18 (Semester-I)

***Stochastic Processes and Apllications***

Date: **November 24, 2017** Maximum Marks:**100** Duration: **3 Hours**

**Note:** The question paper carries a total of 118 marks. You can answer as much as you can, but the maximum you can score is 100.

1. Suppose $A$ and $B$ are events with $0 < P(A) < 1$ and $0 < P(B) < 1$.
   (a) If $A$ and $B$ are disjoint can they be independent?
   (b) If $A$ and $B$ are independent can they be disjoint?
   (c) If $A \subset B$ can they be independent?

   (4+4+2=10)

2. Aruna plays with Naren the following game. First Aruna randomly chooses 4 cards out of a 52-card deck, memorizes them, and places them back into the deck. Then Naren randomly chooses 8 cards out of the same deck. Aruna wins if Narens cards include all cards selected by her. What is the probability of this happening?

   (8)

3. In a TV game show there are three closed doors. Behind one of which there is a car, the other doors have nothing behind them. Let us label the door with the car behind it $a$ and the other two doors $b$ and $c$. In the game the contestant chooses a door and then the host chooses a door, so we can label each outcome as contestant followed by host, e.g., $ab$ means the contestant chose $a$ and host chose $b$. The door chosen by the host is opened and it reveals that it has nothing behind it. The contestant is now given the option of staying with her choice or switch.
   (a) Suppose the contestants strategy is to switch. List all the outcomes in the event 'the contestant wins a car. What is the probability the contestant wins?
   (b) Redo part (a) with the strategy of not switching.

   (10+10=20)

4. In a Markov chain, let $x$ be a recurrent state. Suppose $x$ leads to $y$. Prove that $y$ is recurrent and $\rho_{xy} = \rho_{yx} = 1$.

   (20)

5. Consider a Poisson process with rate $\lambda$. Let random variable $N$ be the number of arrivals in $(0, t]$ and $M$ be the number of arrivals in $(0, t + s]$, where $t, s \geq 0$.
   (a) Find the conditional PMF of $M$ given $N$.
   (b) Find the joint PMF of $N$ and $M$.
   (c) Find the conditional PMF of $N$ given $M$.
   (d) Find $E(N|M)$.

   (5+5+5+5=20)

6. Prove that if the interarrival times of a counting process are independent and identically distributed random variables following $exponential(\lambda)$ distribution, then the process is a Poisson process

   (12+8=20)

7. Derive Markov's Inequality, Chebyshev's inequality and Chernoff's inequality. Let $X$ follow Binomial$(n, \frac{1}{3})$. Find bounds for $P(X \geq \frac{3}{4n})$ using the above inequalities.

   (2+5+5)+(2+3+3)=20)

INDIAN STATISTICAL INSTITUTE

Semestral Examination

Post Graduate Diploma in Business Analytics 2017-18 (Semester-I)

*Stochastic Processes and Apllications*

Date: November 24, 2017 Maximum Marks:100 Duration: 3 Hours

**Note:** The question paper carries a total of 118 marks. You can answer as much as you can, but the maximum you can score is 100.

1. Suppose $A$ and $B$ are events with $0 < P(A) < 1$ and $0 < P(B) < 1$.
   (a) If $A$ and $B$ are disjoint can they be independent?
   (b) If $A$ and $B$ are independent can they be disjoint?
   (c) If $A \subset B$ can they be independent?

   (4+4+2=10)

2. Aruna plays with Naren the following game. First Aruna randomly chooses 4 cards out of a 52-card deck, memorizes them, and places them back into the deck. Then Naren randomly chooses 8 cards out of the same deck. Aruna wins if Narens cards include all cards selected by her. What is the probability of this happening?

   (8)

3. In a TV game show there are three closed doors. Behind one of which there is a car, the other doors have nothing behind them. Let us label the door with the car behind it $a$ and the other two doors $b$ and $c$. In the game the contestant chooses a door and then the host chooses a door, so we can label each outcome as contestant followed by host, e.g., $ab$ means the contestant chose $a$ and host chose $b$. The door chosen by the host is opened and it reveals that it has nothing behind it. The contestant is now given the option of staying with her choice or switch.
   (a) Suppose the contestants strategy is to switch. List all the outcomes in the event 'the contestant wins a car. What is the probability the contestant wins?
   (b) Redo part (a) with the strategy of not switching.

   (10+10=20)

4. In a Markov chain, let $x$ be a recurrent state. Suppose $x$ leads to $y$. Prove that $y$ is recurrent and $\rho_{xy} = \rho_{yx} = 1$.

   (20)

5. Consider a Poisson process with rate $\lambda$. Let random variable $N$ be the number of arrivals in $(0, t]$ and $M$ be the number of arrivals in $(0, t + s]$, where $t, s \geq 0$.
   (a) Find the conditional PMF of $M$ given $N$.
   (b) Find the joint PMF of $N$ and $M$.
   (c) Find the conditional PMF of $N$ given $M$.
   (d) Find $E(N|M)$.

   (5+5+5+5=20)

6. Prove that if the interarrival times of a counting process are independent and identically distributed random variables following $exponential(\lambda)$ distribution, then the process is a Poisson process

   (12+8=20)

7. Derive Markov's Inequality, Chebyshev's inequality and Chernoff's inequality. Let $X$ follow Binomial$(n, \frac{1}{3})$. Find bounds for $P(X \geq \frac{3}{4n})$ using the above inequalities.

   (2+5+5)+(2+3+3)=20)

# INDIAN STATISTICAL INSTITUTE

First Semester Examination: 2017-18

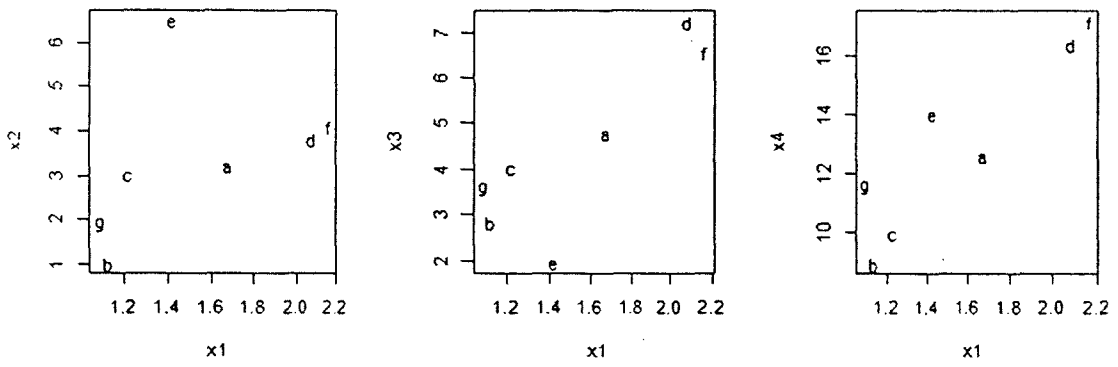
POST GRADUATE DIPLOMA IN BUSINESS ANALYTICS

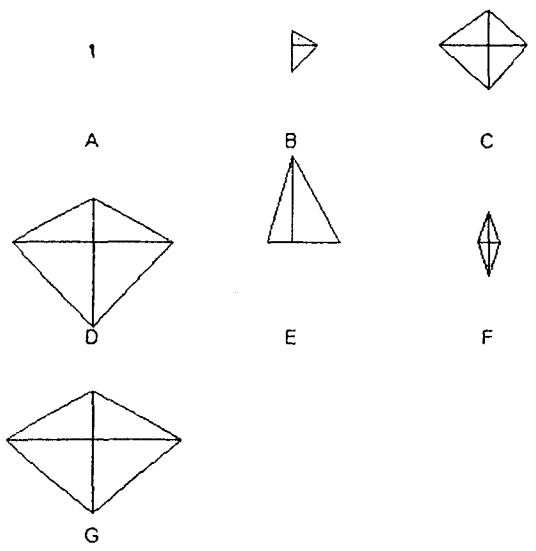
Statistical Structures in Data (BAISI2)

Date: 27 November 2017 Maximum marks: 100 Duration: 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. Pairwise scatterplots of four variables x1, x2, x3 and x4 with seven cases (marked 'a' to 'g') are shown below.




The star plots for the data set are shown, with the cases marked as 'A' to 'G', in random order. The variables are also mixed up. Match the case labels of the star plot with the case labels in the scatter plots.

[10]

2. Suppose $P(X > x) = e^{-2x}$ and $Y = -\sqrt{X}$.

   (a) What is the copula of the joint distribution of $X$ and $Y$?
   (b) What is the marginal (cumulative) distribution of $Y$?
   (b) If the same marginal distributions as above are combined by using the independence copula, what would be the resulting bivariate distribution?

[2 + 2 + 1 = 5]

3. Suppose $X$ and $Y$ are positive valued random variables, $U = 2X - 3$ and $V = Y^2 + 2Y + 1$. Explain whether the following statements are true or false, with reasons.

   (a) Mean of $V$ is smaller than the variance of $Y$.
   (b) Correlation between $U$ and $Y$ is generally the same as that between $X$ and $Y$.
   (c) Correlation between $U$ and $V$ is generally the same as that between $X$ and $Y$.
   (d) Rank correlation between $n$ realized values of $U$ and $V$ is the same as that between the corresponding set of realized values of $X$ and $Y$.
   (e) Copula of the distribution of $U$ and $V$ is generally the same as that of $X$ and $Y$.

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

4. The mean of a random vector $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$ is $\begin{pmatrix} 10 \\ 0 \\ 5 \end{pmatrix}$, and its variance covariance matrix is

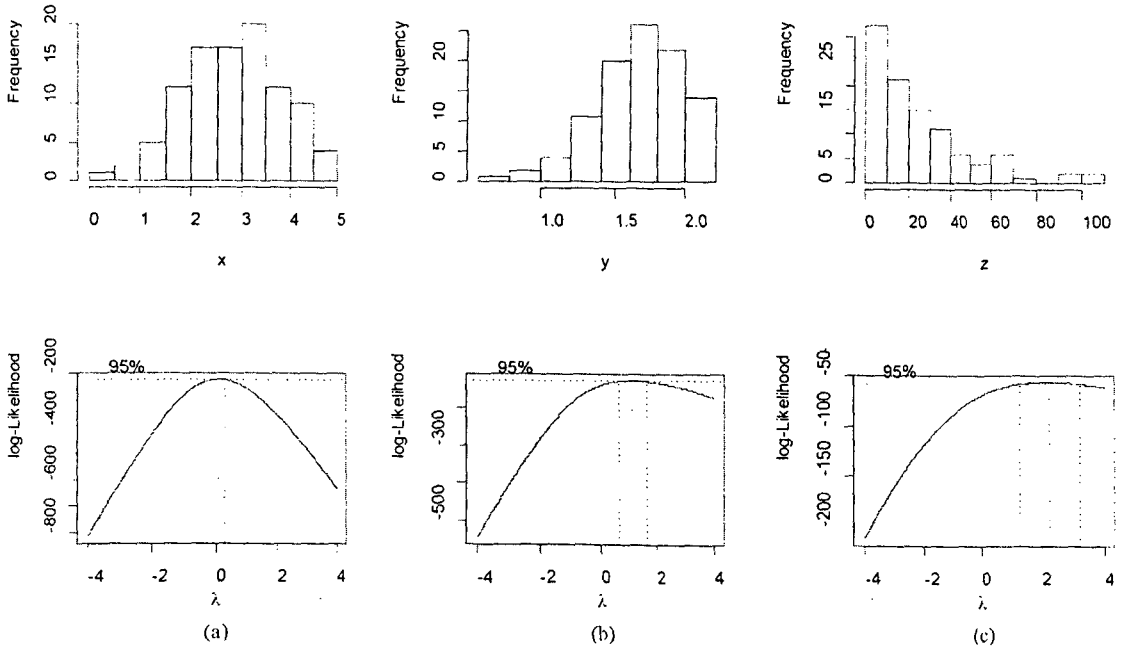
$$\Sigma = \begin{pmatrix} 8 & 2 & 4 \\ 2 & 2 & 2 \\ 4 & 2 & 4 \end{pmatrix}.$$

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

   (a) What is the conditional distribution of $Z$, given $X$ and $Y$?
   (b) Calculate the partial correlation of $Y$ and $Z$, given $X$.
   (c) Give examples of linear combination of $X$ and $Y$ that has (i) largest positive correlation, (ii) zero correlation and (iii) largest negative correlation with $Z$.

   $[5 + 5 + (2 + 2 + 1) = 15]$

5. Associate the three histograms with the correct plot of likelihood against the Box-Cox transformation parameter.




[5]

6. The R output of a multiple regression and related ANOVA is given below.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.3087     2.8294   8.591 1.84e-09
ihp         719.9060   137.3845   5.240 1.30e-05
wt           -3.2037     0.5771  -5.552 5.48e-06
---
Residual standard error: 2.22 on 29 degrees of freedom
Multiple R-squared:  0.873,     Adjusted R-squared:  0.8643
F-statistic: 99.71 on 2 and 29 DF,  p-value: 1.007e-13

Analysis of Variance Table

Model 1: mpg ~ 1
Model 2: mpg ~ ihp + wt
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     31 1126.05
2     29  142.96  2    983.09 99.711 1.007e-13
```

(a) What is the fitted linear regression equation in terms of the response and explanatory variables mentioned in the output?

(b) What are the residual sum of squares and the associated degrees of freedom?

(c) What are the interpretations of the last four numbers reported in the last line of the output?

(d) What is the estimated variance of the model error?

(e) Why is `Pr(>F)` much smaller than the values of `Pr(>|t|)` for the two regressors?

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

7. Consider the bivariate linear regression model with independent and normally distributed errors

$$Y_{i1} = \beta_{01} + \beta_{11} Z_i + \varepsilon_{i1},$$
$$Y_{i2} = \beta_{02} + \beta_{12} Z_i + \varepsilon_{i2},$$
$$E\begin{pmatrix}\varepsilon_{i2}\\ \varepsilon_{i2}\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix}, D\begin{pmatrix}\varepsilon_{i2}\\ \varepsilon_{i2}\end{pmatrix} = \Sigma, \qquad i = 1, \ldots, n.$$

(a) Give expressions for the maximum likelihood estimators of the regression coefficients and $\Sigma$.

(b) Are the above estimators unbiased? If not, give expression(s) for unbiased estimator(s).

(c) What are the variances of the estimated regression coefficients?

(d) Give expressions for Bonferroni simultaneous confidence intervals of $\beta_{11}$ and $\beta_{12}$ with joint coverage probability $1 - \alpha$.

(e) If $Z_i$ is a binary indicator of a group membership (say, men and women), express the hypothesis of *homogeneity of group means* in terms of the beta-coefficients.

(f) If there are four groups, explain (in plain terms, without formal proof) why the above model with a single explanatory variable cannot be used to test for *homogeneity of group means*, describe an alternative linear model that can serve the purpose, and express the hypothesis in terms of the parameters of that model.

$[4 + 2 + 2 + 4 + 2 + (2 + 3 + 1) = 20]$

8. Suppose $X$ and $Y$ are continuous random variables with bivariate density $f(x,y)$.

   (a) What is meant by the regression of $Y$ on $X$? How does it relate to prediction?
   (b) If $f(x,y) = xe^{-x(y+1)}$, $x \geq 0$, $y \geq 0$, find the equation of the regression of $Y$ on $X$.
   (c) Give an example of $f(x,y)$, which is not bivariate normal, such that the regression of $Y$ on $X$ is a linear function of $X$.

$$[(1+1)+4+3=8]$$

9. The variance covariance matrix of a standardized random vector $\boldsymbol{X} = (X_1 : X_2 : \cdots : X_p)'$ is

$$\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix},$$

   for some positive fraction $\rho$.

   (a) Show that the vector $\boldsymbol{u} = (1:1:\cdots:1)'$ is an eigenvector of $\Sigma$, and identify the corresponding eigenvalue.
   (b) Show that any vector $\boldsymbol{v}$ that is orthogonal to $\boldsymbol{u}$ is also an eigenvector of $\Sigma$, and identify the corresponding eigenvalue.
   (c) List all the eigenvalues of $\Sigma$, with reason.
   (d) Draw the scree plot for $p = 10$ and $\rho = 0.5$. How many principal components are needed to explain 90% of the total variance of $\boldsymbol{X}$?
   (e) Write an orthogonal factor model of $\boldsymbol{X}$ for its factor analysis with the smallest number of factors, and describe all parts of the model. Explain why there cannot be a linear model with fewer factors. Can this model be written in another way after factor rotation?

$$[2+4+2+(3+2)+(4+2+1)=20]$$

10. Four multivariate data points $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$ and $\boldsymbol{X}_4$ have the Euclidean distance matrix

$$\begin{pmatrix} 0 & 7.44 & 18.8 & 21.4 \\ 7.44 & 0 & 19.2 & 21.9 \\ 18.8 & 19.2 & 0 & 2.73 \\ 21.4 & 21.9 & 2.73 & 0 \end{pmatrix}.$$

    An analyst has carried out multidimensional scaling of the data, and arrived at the following bivariate representation of the four data points:

$$\boldsymbol{X}_1 : (1.5,5.9);\ \boldsymbol{X}_2 : (11.9,12.6);\ \boldsymbol{X}_3 : (1.6,0.2);\ \boldsymbol{X}_4 : (12.9,14.4).$$

   (a) Examine the findings and report if there is anything wrong.
   (b) Can you suggest any alternative bivariate representation of the data that would have a smaller value of Kruskal's stress, without doing any computation?
   (c) Suggest another bivariate representation, distinct from your answer in part (b), which would have exactly the same value of Kruskal's stress.

$$[3+2+2=7]$$

# INDIAN STATISTICAL INSTITUTE

## First Semester Examination : 2017–18

Post Graduate Diploma in Business Analytics (First Year)

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

Date : 29 November 2017 | Maximum Marks : 60 | Duration : 3 Hours

Answer any THREE questions out of the following. All questions carry equal marks.

### Problem A : Numerical Algorithms [10 + 8 + 2 = 20]

1. Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and a vector $\mathbf{b} \in \mathbb{R}^m$, propose a linear algebraic numerical algorithm to minimize $||\mathbf{Ax} - \mathbf{b}||_2$, where $\mathbf{x} \in \mathbb{R}^n$. What is this process more popularly known as, especially if $\mathbf{A}$ is a dataset with $m$ observations and $n$ predictors, and $\mathbf{b}$ is the set of $m$ corresponding values of the response variable? [10]
2. Is it possible to modify this method to minimize $||\mathbf{Ax} - \mathbf{b}||_2^2 + \lambda||\mathbf{x}||_2^2$, where $\mathbf{x} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ is a given constant? Justify. What is this process more popularly known as, especially if $\mathbf{A}$ is a dataset with $m$ observations and $n$ predictors, and $\mathbf{b}$ is the set of $m$ corresponding values of the response variable? [8]
3. What is this process more popularly known as, especially if $\mathbf{A}$ is a dataset with $m$ observations and $n$ predictors, and $\mathbf{b}$ is the set of $m$ corresponding values of the response variable, if we minimize $||\mathbf{Ax} - \mathbf{b}||_2^2 + \lambda||\mathbf{x}||_1^2$ instead (note the norm carefully)? [2]

### Problem B : Linear Regression [4 + 4 + 6 + 6 = 20]

Suppose you are given the `summary()` and `plot()` of the following linear model in R.

`lm(response variable ~ all predictors)`

1. In a model where there are more than one predictors, how do you judge if there is *any* linear relationship between the predictors and the response at all? Justify. [4]
2. In a model where there are more than one predictors, how do you judge which predictors appear to have a statistically significant relationship to response? Justify. [4]
3. Would you build a smaller model by discarding *insignificant* predictors at this stage, or would you require any further information to make such a decision? Justify. [6]
4. What aspects of the *residuals* would you explore in the `plot()` of `lm()`, as stated above, and what measures would you take based on these observations? Briefly comment. [6]

P.T.O

## Problem C : Logistic Regression [4 + 4 + 6 + 6 = 20]

Suppose you are given the `summary()` of the following logistic regression model in R.

`glm(response variable ~ all predictors, family = binomial(link='logit'))`

1. Assuming that the predictors in this model are $\{X_1, X_2, \ldots, X_p\}$ and the response is $Y$, write the logistic model in its equation form, involving the `logit` link function. [4]
2. How do you judge the accuracy of the logistic model with respect to the *training data*? Explain in terms of Null Deviance, Residual Deviance and AIC on training data. [4]
3. How would you set the threshold for probability while predicting with the logistic model? Is setting a threshold of $p > 0.5$, the common default value, a good choice? Justify. [6]
4. Given two logistic models, one smaller than the other (with a subset of predictors), how would you compare the two in terms of accuracy on the training data? Justify. [6]

## Problem D : Model Selection [6 + 8 + 6 = 20]

1. What do you mean by *overfitting* a model? Explain in terms of *bias* and *variance* of the model. Why is it a good practice not to *overfit* your model to the training data? [6]
2. How would you ensure that your model does not overfit training data? Compare validation set approach, leave-one-out cross validation and $k$-fold cross validation in this regard. [8]
3. Suppose you have multiple predictors in a dataset, and fit a complete model to start with. How would you select the best possible subset of predictors? Analyse the complexity. [6]

## Problem E : Tree-based Models [5 + 3 + 6 + 6 = 20]

Suppose you are given a training dataset for classification, with a binomial response variable.

1. If there are $p$ predictors $\{X_1, X_2, \ldots, X_p\}$, some numeric and some categorical, how would you choose the *optimal split* at the root (or any) level of a decision tree? Justify. [5]
2. Would this choice for *optimal split*, at the root (or any) level, be different if you choose different notions of information – Shannon Entropy vs. Gini Index? Justify. [3]
3. Is it possible to improve the model by using more than one tree? What happens if the trees are identical? How would you ensure independence between the trees in such a model? [6]
4. What is the notion of *overfitting* in case of a decision tree, and how would you control it? Is it the same in case of a bagging model or a random forest? Justify. [6]

P.T.O

## Problem F : Clustering Methods [6 + 4 + 4 + 6 = 20]

Suppose you are given a multivariate dataset with $m$ observations and $p$ predictors.

1. Describe the method of $k$-means to segregate these data-points in $k$ different clusters. [6]
2. Is there a guarantee for convergence in $k$-means? Is it possible to improve the convergence and stability of the method by selecting the initial centroids more effectively? [4]
3. What is the primary limitation of the clusters formed by the $k$-means algorithm? Is it practical to use $k$-means for Gaussian mixtures? How would you solve this issue? [4]
4. How would you choose the *optimal* value of $k$ in case of $k$-means algorithm? Explain in terms of *within* and *between* variances of the clusters. Would this optimal value of $k$ change if the notion of distance is changed in case of $k$-means – $\ell_2$ vs. cosine, say? Justify. [6]

## Problem G : Singular Value Decomposition [4 + 6 + 6 + 4 = 20]

1. Suppose that the *full* Singular Value Decomposition of an $m \times n$ matrix $\mathbf{X} = U\Sigma V^T$ is:

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

   What are the dimensions of each matrix $(U, \Sigma, V)$ in this representation, where for example, $\mathbf{X}$ is $m \times n$? How can you determine the Rank of $\mathbf{X}$ given this SVD representation? [4]
2. As per the above representation of the SVD of $\mathbf{X}$, determine the dimension and rank of each of the slices (matrices) $\mathbf{Z}_i = \sigma_i \mathbf{u}_i \mathbf{v}_i^T$, where $1 \leq i \leq r$. Is is possible to reconstruct the original matrix $\mathbf{X}$ given the slices (matrices) $\mathbf{Z}_i$ for $1 \leq i \leq r$? [6]
3. Is is possible to reconstruct the original matrix $\mathbf{X}$ given the slices (matrices) $\mathbf{Z}_i$ for $1 \leq i \leq k$, where $k$ is strictly less than $r$? If so, provide such a construction. If not, provide a low-rank *approximate* reconstruction of $\mathbf{X}$ using the available slices (matrices) $\mathbf{Z}_i$ for $1 \leq i \leq k$, and comment on the quality of such an approximation. [6]
4. How would you choose the *optimal* number of slices (matrices) $\mathbf{Z}_i$, for $1 \leq i \leq k$, that are required to reconstruct the original matrix $\mathbf{X}$ with *sufficient* accuracy? [4]

*Best of luck!* ☺

# INDIAN STATISTICAL INSTITUTE

## First Semester Examination, PGDCA: 2017-2018
## Subject: Introduction to Computer Architecture and systems software (101)
## Total marks: 110, Full marks: 100, Duration: 180 min
## *Examination Date: 11th December, 2017*

1. (a) Write an assembly program for MIPS processor to calculate the average of integers present in an array of integers in RAM.

   (b) Write an assembly program for MIPS processor that finds the largest integer from three user given integers. (5+5)

2. (a)What are the differences between level triggered and edge triggered sequential circuits?

   (B) "00 ->01 ->11 ->10 ->00" is the states of a counter. Implement this counter using D Flip-Flops.

   (c) Implement SR flip-flop and SR latch using NAND Gates. (2+5+3)

3. (a) What are the different types of instruction, Explain each in brief.

   (b)Explain different types of shift instructions with examples.

   (c) Which data structure is suitable to implement 'call and return' instruction? (4+4+2)

4. (a) What do you understand by BUS in the computer system? What are the types of BUSes? Write their purposes.

   (b) Implement DATA BUS using multiplexers such that 8 registers each of 4 bits can transfer data among themselves. (4+6)

5. (a) A Cache has the following specifications:
   Number of sets = 128
   2 ways set Associative
   Cache size = 4Kbytes
   Main memory has 21 bit address
   Calculate the size of the cache blocks, number of cache blocks and memory required for tag directory.

   (b)What are the different types of cache misses. Explain each in brief. (6+4)

6. (a) f(a,b,c,d) is represented using the Karnaugh map shown below. 'X' is don't care term. Minimize 'f' and implement using basic gates.

| cd \ ab | 00 | 01 | 11 | 10 |
|---|---|---|---|---|
| 00 | 1 | X | X | 1 |
| 01 | X | | | 1 |
| 11 | | | | |
| 10 | 1 | | | X |

(b) If P, Q, R are Boolean variables, then minimize (P + Q')(PQ' + PR)(P'R' + Q') .

(c) If ' n ' bits are used to represent a number in 2's complement form, then what are maximum and minimum number? (5+3+2)

7. (a) Consider a 4-way set associative cache(initially empty) with total 16 cache blocks. The main memory consists of 256 blocks and the request for memory blocks in the following order: 0, 255, 1, 4, 3, 8, 133, 159,216,129,63,8,48,32,73,92,155.Calculate the percentage appear of block hit in cache if LRU block replacement policy is used.

(b) A system consists of two level of cache L1 with hit ratio 0.6 and L2 with hit ratio 0.5. Suppose the access time of L1, L2 and main memory are 1 ns, 10 ns and 100 ns respectively. Calculate the average memory access time of the system. (6+4)

8. (a)Match List-I with List-II.

| List-I | List-II |
|---|---|
| A. Pointer | 1. Indirect addressing mode |
| B. Position independent code | 2. Immediate addressing mode |
| C. Constant operand | 3. Relative addressing mode |

(b) If the last operation performed on a computer with an 8-bit word has an addition in which the two operands were 00000010 and 00000011, what would be the value of the Overflow, Sign, Zero and Carry flags ?

(c) A 4 byte long PC-relative branch instruction is fetched from memory address $(512)_{10}$ and while its execution, the branch is made to location $(885)_{10}$. What is the unsigned displacement present in the instruction? (relative value) (3+4+3)

9. (a)What are the advantage of pipelined system over non-pipelined system?

   (b) What is the relation between efficiency '$\eta$' and speed-up 'S' of a pipelined system having 'm' stages?

   (c) Consider a pipelined system with four stages: IF, ID, EX, WB. Following chart shows the clock cycles required by each instruction to compete each stage

| Instructions | Instruction Fetch(IF) | Instruction Decode(ID) | Instruction Execute(EX) | Write back(WB) |
|---|---|---|---|---|
| $I_0$ | 1 | 1 | 2 | 1 |
| $I_1$ | 2 | 2 | 3 | 1 |
| $I_2$ | 2 | 2 | 2 | 2 |
| $I_3$ | 2 | 1 | 1 | 1 |
| $I_4$ | 3 | 2 | 1 | 2 |

   Draw the phase-time diagram. Calculate the efficiency and speed-up of the system for the above 5 instructions. (2+2+6)

10. (a)Show the IEEE754 binary representation of the number $(-63.75)_{10}$ in single precision format.

    (b)Write Booth's algorithm of multiplication.

    (c) Multiply -2 and 3 using 4 bits to store multiplicand and 4 bits for multiplier and 8 bits for product registers respectively. (6+3+3)

11. (a) What do you mean by interrupt?

    (b)What are differences between vectored and non-vectored interrupt?

    (c) Describe the interrupt driven I/O transfer mode. What is disadvantage of this I/O transfer mode? (2+3+5)

# Indian Statistical Institute

## Course Name : PGDCA

## First Semester Examination (2017-18)

## Paper Name: Introduction to Programming (102)

**Time: 3 Hours and 30 minutes** **Marks: 100** **Date: 13.12.2017**

**The question paper is for 110 marks, answer as many as you can, you can get at most 100 marks.**

**NOTE: If element types are not mentioned, assumed them as integers. Write answers in your own language as much as possible. If no input/output device is specified, assume Keyboard as your standard input device and Monitor as your standard output device.**

1. (a) Write a program in C to implement a STACK of 5 elements.

   (b) Evaluate the following expression (given in postfix notation) using a STACK.
   5, 6, 2, +, *, 12, 4, /, -

[ 6+4 = 10 ]

2. (a) Differentiate between array of pointers and pointer to an array in C.

   (b) Write a program in C to search an element from an array of 10 elements using binary search.

   (c) "Array in C does not have boundary checking"-justify the statement.

[ 3+5+2 = 10 ]

3. (a) What is the difference between call by value and call by reference? Explain with a segment of code in C.

   (b) Define Recursion. Write a program in C to calculate the GCD (Greatest Common Divisor) of two numbers (taken as user input) using recursion.

[ 4+(2+4) = 10 ]

4. (a) Write a program in C to find out whether a year (taken as user input) is leap year or not using conditional/ternary operator.

(b) Differentiate between the keywords 'break' and 'continue' in C.

(c) Write a program in C to display the following pattern for n rows (n should be user input)

```
1 2 3 4 5
1 2 3 4
1 2 3
1 2
1
```

(output for n=5)

[ 3+3+4 = 10 ]

5. (a) Differentiate between an algorithm and a flowchart.

(b) Write a segment of code in C to allocate memory dynamically for a two-dimensional array of size m by n (m and n should be user input).

(c) What are the benefits of using switch-case over if- else if- else.

[ 3+4+3 = 10 ]

6. (a) Write a user defined function in C which is equivalent to strcpy ().

(b) Write a program in C that will calculate the number of vowels present in a sentence.

(c) Differentiate between structure and union in C.

[ 3+4+3 = 10 ]

7. (a) What are the advantages of linked list over array?

(b) Write functions in C that will perform the following operations on single linked list:

(i) Insert a node at any specified position. (ii) Delete a node from any specified position.

[ 4+ (3+3) = 10 ]

8. (a) Suppose inorder and preorder traversals of a binary tree are as follows:

Inorder: DBHEAIFJCG Preorder: ABDEHCFIJG

Construct the binary tree and also make the binary tree a threaded binary tree using in-order threading.

(b) What is BST (Binary Search Tree)? Why do we need height balancing for BST?

[ 6+(2+2) = 10 ]

9. (a) Suppose we have an array of elements 15, 7, 11, 18, 22, 14, 1, 20, 9 respectively. Show all the steps for first two passes in both selection and bubble sort.

(b) Write a program in C to find transpose of a matrix of order 3 by 4.

[ 6+4 = 10 ]

10. (a) Write a program in C that merges lines alternately from two files and writes the results to a new file. If one file has less number of lines than the other, the remaining lines from the larger file should be simply copied into the target file.

(b) What are the differences between opening a file in "w" mode and in "a" mode?

[ 7+3 = 10 ]

11. What will be the output of the following segment of code? (Assume 32 bit Turbo C/C++ Compiler)

(a)
```
int x=20;
void main()
 {
   int x=5;
   printf("%d ",x);
   fun(x);
 }
void fun(int  y)
 {
   printf("%d %d",x, y);
 }
```

(b)
```
#define SQUARE(r) r*r

void main()
 {
    int y;
    y=64/SQUARE(4);
    printf("%d",y);
 }
```

(c)
```
void main()
 {
    int x=15;
    printf("%d %d %d", x==15, x=20, x>10);
 }
```

(d)
```
void main()
 {
    int i=5;
    int *p;
    p=&i;
    magic(&p);
    printf("%d",*p);
 }
void magic(int **p)
 {
    int x=10;
    *p=&x;
    printf("%d",**p);
 }
```

(e)
```
void main()
 {
    int a=5,b=10;
    b=a++ + ++b;
    a=++a + b++;
    printf("%d %d",a,b);
 }
```

[ 2×5 = 10 ]

**********END**********

# INDIAN STATISTICAL INSTITUTE

## First Semester Examination, PGDCA: 2017-2018
## Subject: DBMS (103)
## Attain all questions; you will get maximum 100 marks
## Duration: 180 min
## *Examination Date: 15th December, 2017*

1. (a) What do you mean by Data? Give examples.
   (b)What are database administrator's responsibilities?
   (c) Write note on Meta Data.

   (2+5+3)

2. Construct an E-R diagram for a car insurance company whose customers own one or more cars each. Each car has associated with it zero to any number of recorded accidents. Each insurance policy covers one or more cars, and has one premium payments associated with it. Each payment is for a particular period of time, and has an associated due date, and the date when the payment was received.

   10

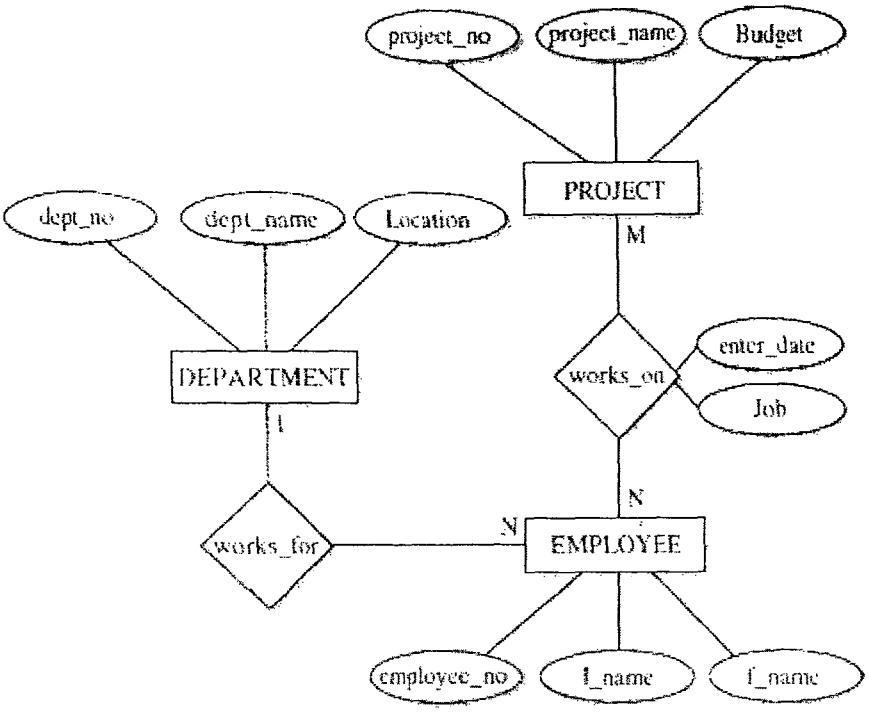
3. (a) Define foreign key.
   (b). Give an example that makes difference between Primary key, Candidate key, Alternative key and Super key.
   (c) Suppose R(A,B,C,D,E) is a relation Which has two candidate key {A} and {B}. How many Super key does 'R' have?
   (d)Why is a table called Relation in Relation database?

   (2+3+3+2)

4. (a) Define functional dependency. Give example.
   (b) With example, make the differences between trivial, non-trivial and semi- trivial functional dependency.

(c) Suppose R(E,F,G,H,I,J,K,L,M,N) is a relation and functional dependency { EF→ G, F→I J, EH→KL, K→M, L→N} is applicable on R. Derive Candidate key of R. How many super key does R have? (2+3+5)

5. Convert the following ER diagram into relations. 10




6. (a) Consider a relation R= {M, N, O,P, Q, R, S, T} with the following set of dependencies:

MN→Q

M→RQ

N →R

R→ST

Next consider the following set of decompositions for the relation schema R: D1= {R1,R2,R3,R4}: R1={M,N,O,P}, R2={M,P,Q}, R3= {N,R} and R4={R,S,T}

D2= {R1,R2,R3,R4}: R1={M,N,O}, R2= {P,Q}, R3= {N,R} and R4= {R,S,T}

Which of the above decomposition has/ have lossless join property?

(b) Consider the relation R= (A, B, C, D, E). Functional dependencies FD1 and FD2 are applicable on R.

FD1. A->B, AB->C, D->AC, D->E

FD2. A->BC, D->AE

Are FD1 and FD2 equivalent? (5+5)

7. (a) Why do we do normalization?

(b)When do we call a functional dependency partial dependency? Give example.

(c) Define transitive functional dependency. Give example.

(d) Given R (A, B, C, D, E, F) and FDs={AB->C, D->E, E->F, D->A}

What is the highest normal form of R? If it is not in BCNF then convert it into BCNF.

(2+2+2+4)

8. (a) What are the different types of joins. Explain with example.

(b) Consider following Relations A, B, and C given below:

A

| Id | **Name** | Age |
|---|---|---|
| 12 | Arun | 60 |
| 15 | Shreya | 24 |
| 99 | Rohit | 11 |

B

| Id | Name | Age |
|---|---|---|
| 15 | Shreya | 24 |
| 25 | Hari | 4 |
| 98 | Rohit | 20 |
| 99 | Rohit | 11 |

C

| Id | Name | Age |
|---|---|---|
| 10 | Ram | 2 |
| 99 | Sita | 1 |

How many tuples does the result of the following relation algebra expression contain? Assume that the schema of A∪B is the same as that of A and C.

$(A \cup B) \bowtie (\pi_{A.Id>40}(A) \cup \pi_{C.Id<15}(C))$ (5+5)

9. (a) What are the differences between Primary and Secondary indexing?

(b) Suppose that we have an ordered file of 30000 records on a disk with a block of size 1024 byte. Records are of fixed size and are un-spanned. Size of a record is of 100 byte. Suppose we have created secondary index on the key field of size 9 byte and block pointer of size 6 byte. Calculate the average number of blocks access to search a record using with and without index.

(c) What do you mean by clustering index? What are its disadvantages?

(3+5+2)

10. Consider a database with followings relations:

Employee(Name, SSN, Address, Salary, Sex, Dno)

Department( Dname, Dnumber, Mgrssn)

Dept_location(Dnumber, Dlocation)

Note: A particular department (e.g. IT) may present ii more than one place.

Answer the following queries:

(a) List the Employee's names who has/have highest salary in his department.

(b) Get the name and Address of the employees who work for the department located in 'Howrah'

(c) Retrieve the Name and Salary of all the managers except department located at Howrah .

(d) Get the employees name who has second highest salary. (2+2+4+2)

11. Write Short note on the followings:

(a) An algorithm to implement relational operator PROJECTION ($\pi$) for single relation R.

(b) DDL (Data Definition Language)

(c) NULL value of Oracle 10g

(d) DQL (Data Query Language) (2.5 *4)

# Indian Statistical Institute

## Course Name : PGDCA

## First Semester Examination (2017-18)

## Paper Name: Operating System (104)

**Time: 3 Hours** **Marks: 100** **Date: 18.12.2017**

**The question paper is for 110 marks, answer as many as you can, you can get at most 100 marks.**

1. (a) Differentiate between paging and segmentation.

   (b) Consider a logical address space of 512 pages with a 16 kb page size, mapped onto a physical memory of 32 frames.

   i)How many bits are required in the logical address?

   ii)How many bits are required in the physical address?

   (c) What is virtual memory?

   [ 4+4+2 = 10 ]

2. (a) Differentiate between external and internal fragmentation.

   (b) Assuming we store some page table entries into TLB and also assuming 100 milisec required for each main memory reference and 20 milisec required for each TLB reference, find the percentage of profit for 90 percent hit ratio (TLB hit) in comparison with no TLB present.

   (c) What do you mean by hierarchical paging?

   [ 4+4+2 = 10 ]

3. (a) What do you mean by page fault? What are the sequences of events that occur when page fault occurs?

(b) Consider the following page reference string:

7,2,3,1,2,5,3,4,6,7,1,0,5,4,6,2,3,0,1

Assuming pure demand paging with three frames, how many page faults would occur for optimal page replacement algorithm?

(c) What do you mean by thrashing?

[ (1+3)+4+2 = 10 ]

4. (a) What do you mean by degree of multiprogramming? Differentiate between multiprogramming and multitasking.

(b) When do RR scheduling works like FCFS scheduling?

(c) Consider the following set of processes:

| Process | Burst time | Arrival time | Priority |
|---|---|---|---|
| P1 | 7 | 1 | 2 |
| P2 | 9 | 0 | 3 |
| P3 | 3 | 4 | 4 |
| P4 | 6 | 6 | 1 |

Draw the Gantt Chart and find out average waiting time and average turnaround time for the SRTF (Shortest Remaining Time First) scheduling algorithm.

[ (1+2)+2+5 = 10 ]

5. (a) What is context switching?

(b) What do you mean by dual mode CPU operation?

(c) Discuss the following with respect to multiple processor scheduling

i) Processor Affinity ii) Load Balancing.

[ 2+2+(3+3) = 10 ]

6. (a) Differentiate between seek time and latency time with respect to disk management.

(b) What is swapping?

(c) Consider a disk queue with requests for IO to blocks on tracks 98, 183, 37, 122, 31, 124, 55, 67. Find out the total head movement for serving those requests in SSTF algorithm (Assume initially head is at track 50).

[ 3+2+5 = 10 ]

7. (a) What is semaphore?

(b) What do you mean by busy waiting/spinlock? Write the modified wait() and signal() algorithm for semaphore that does not suffer from spinlock.

[ 2+(2+6) = 10 ]

8. (a) What are the necessary conditions for deadlock?

(b) Consider the following snapshot of a system:

| Process | Allocation | Max | Available |
|---|---|---|---|
| | A B C | A B C | A B C |
| P0 | 0 1 0 | 7 5 3 | 3 3 2 |
| P1 | 2 0 0 | 3 2 2 | |
| P2 | 3 0 2 | 9 0 2 | |
| P3 | 2 1 1 | 2 2 2 | |
| P4 | 0 0 2 | 4 3 3 | |

i) Find a safe sequence to prove that the current allocation state is deadlock free.

ii) If a request from process P1 arrives for (1, 0, 2), can the request be granted immediately? Justify your answer using banker's algorithm.

[ 4+(3+3) = 10 ]

9. (a) What are the benefits of multithreading?

(b) Solve the producer-consumer synchronization problem (for bounded buffer) with semaphore (write only code segment, don't need explanation).

(c) Differentiate between virus and worm.

[ 3+4+3 = 10 ]

10. (a) Write Unix command/commands to perform the following tasks:

(i) Find the user-ids of all the users who are currently logged on to your system.

(ii) Record your login session in a file named "LOGFILE".

(iii) Perform deletion of all files in your current directory.

(iv) Display the contents of a file named "PGDCA" with one page at a time.

(v) Count the number of characters in a file called "EMPLOYEE".

(vi) Suppose we have two sorted files "FILE1" and "FILE2". Select lines not common to both files.

(vii) Create a directory named "DIR1" under your parent directory.

(viii) Suppose the owner (user) does not have executable permission for the file named "ARRAY". Assign execute permission to the owner (user).

(ix) List all files and sub-directories(including files in sub-directories) of your current directory.

(b) What is the difference between the commands "pwd" and "echo $HOME"?

[ 9×2+2 = 20 ]

**********END**********

# Indian Statistical Institute

## Course Name : PGDCA

## First Semester Examination (2017-18)

## Paper Name: Discrete Mathematics (105)

**Time: 1 hour and 30 minutes** **Marks: 50** **Date: 21.12.2017**

**The question paper is for 55 marks, answer as many as you can, you can get at most 50 marks.**

1. (a) Differentiate between complete binary tree and full binary tree.

   (b) Prove that the number of pendant vertices in a binary tree (strict) of n vertices is n-1.

   [ 2+3 = 5 ]

2. (a) Define Euler graph.

   (b) Prove that the number of vertices of odd degree in a graph is always even.

   [ 2+3 = 5 ]

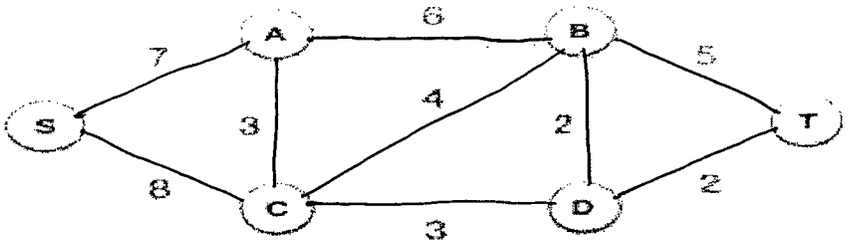
3. (a) What is the difference between a spanning tree and a fundamental circuit of a graph?

   (b) A coke machine is to identify, by a sequence of tests, the coin that is put into the machine. Only penis, nickels, dimes and quarters can go through the slot. Let us assume that the probabilities of a coin being a peny, a nickel, a dime and a quarter are 0.05, 0.15, 0.5, 0.3 respectively. Each test has the effect of partitioning the four types of coins into two complementary sets and asserting the unknown coin to be in one of the two sets. If the time taken for each test is the same, what sequence of tests will minimize the average expected time taken by the coke machine to identify the coin? Give answer by constructing a binary tree (strict) with minimum weighted path length for pendant vertices.

   [ 2+3 = 5 ]

4. (a) What do you mean by minimum spanning tree?

(b) Find out a minimum spanning tree of the following graph using Prim's algorithm.




[ 1+4 = 5 ]

5. (a) Write down the Dijkstra algorithm for finding out shortest path between two specified vertices.

   (b) What is the number of edges present in a complete graph of 6 vertices?

   (c) What is the rank and nullity of a disconnected graph of 3 components with 9 vertices and 16 edges?

[ 5+1+2 = 8 ]

6. Write recurrence relation that counts number of disc movements in the problem of Tower of Hanoi and solve the recurrence relation you wrote.

[5]

7. Suppose you have five letters {A,C,D,G,P}. Calculate the dictionary Rank of word 'PGDCA' among all five lettered words with the following conditions:

   (i) Repetitions are not allowed and length of each word is five.

   (ii) Repetitions are allowed and length of each word is five.

[3.5*2]

8. There are 10 persons (p1,p2,p3,.....p10) in a family. In how many ways you can make them sit around a circular table such that person p3 always sits in front of p10. Note: Seats are numbered.

[5]

9. In the given series 1! + 4! +7! + 10! + 13! + ..... 999!. What is the second last digit (second least significant digit) in sum of the above series?

[5]

10. Suppose there are 10 married couples i.e. 10 husbands numbered H1,H2,.....H10 and 10 wives numbered W1,W2,...W10. You have to find a group of 4 members such that " If H2 is selected then W3 will be selected". How many ways you can form the group.

[5]

**********END**********

# INDIAN STATISTICAL INSTITUTE

First Semester Backpaper Examination: 2017-18

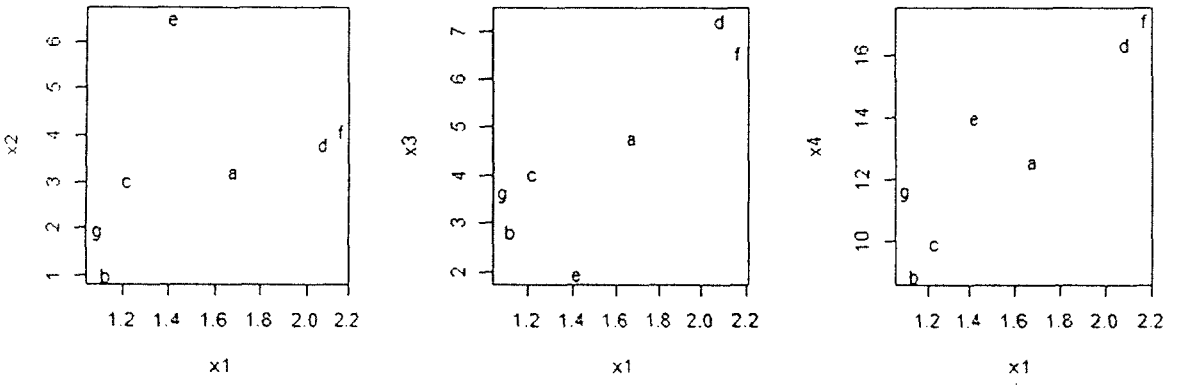
POST GRADUATE DIPLOMA IN BUSINESS ANALYTICS

Statistical Structures in Data (BAISI2)

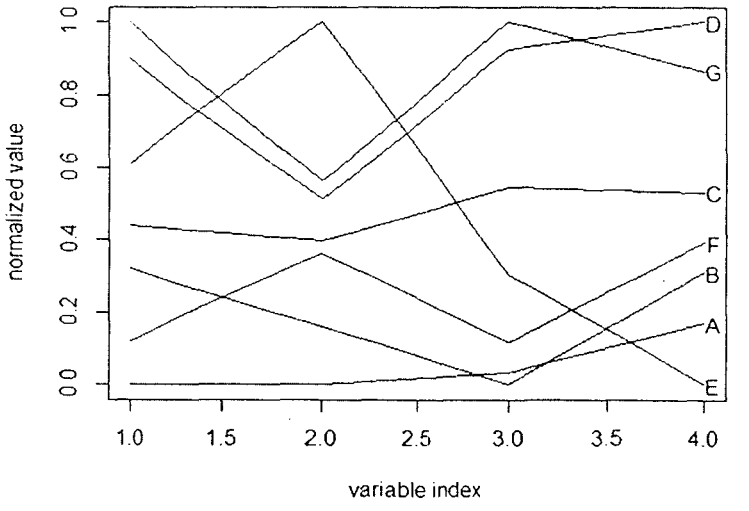
Date: 21 December 2017 Maximum marks: 100 Duration: 3 hours

*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. Pairwise scatterplots of four variable x1, x2, x3 and x4 with seven cases (marked 'a' to 'g') are shown below.




The line plots for the data set are shown below, with the cases marked as 'A' to 'G', in random order. The variables are also mixed up. Match the case labels of the line plot with the case labels in the scatter plots.




[10]

2. Derive an expression for the bivariate normal copula, with correlation coefficient $\rho$.

[10]

3. Five scores are shown below, in original units and in standard units. Fill in the blanks.

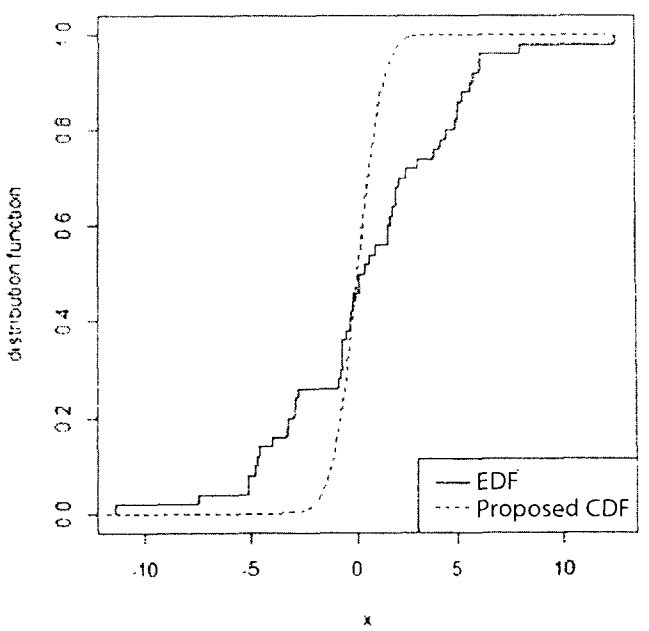
| | | | | |
|---|---|---|---|---|
| 54 | 69 | 62 | 42 | ___ |
| 0.8 | 1.8 | ___ | ___ | −1.4 |

[5]

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

[5]

5. 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 QQ plot for this data set. Label the axes clearly.




[10]

6. How does one check the normality of univariate data graphically and through a statistical test? If the univariate data are found to be normal, how can one check multivariate normality?

[2 + 3 = 5]

7. Let $\boldsymbol{X}$ have the distribution $N_3(\boldsymbol{\mu}, \Sigma)$, where

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

(a) Find distribution of $2X_1 - X_2 + X_3$.
(b) Find a linear function of $X_1$ and $X_3$ that is independent of $X_2$.
(c) Specify the conditional distribution of $X_1$, given $X_2 = x_2$ and $X_3 = x_3$.
(d) Calculate the partial correlation of $X_2$ and $X_3$, given $X_1 = x_1$.
(e) Calculate the multiple correlation coefficient of $X_1$ with $X_2$ and $X_3$.

[5 + 5 + 5 + 5 + 5 = 25]

8. 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]

9. Consider the bivariate linear regression model with independent and normally distributed errors

$$\begin{aligned} Y_{i1} &= \beta_{01} + \beta_{11} X_i + \varepsilon_{i1}, \\ Y_{i2} &= \beta_{02} + \beta_{12} X_i + \varepsilon_{i2}, \end{aligned} \qquad E\begin{pmatrix} \varepsilon_{i2} \\ \varepsilon_{i2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, D\begin{pmatrix} \varepsilon_{i2} \\ \varepsilon_{i2} \end{pmatrix} = \Sigma, \qquad i = 1, \ldots, n.$$

(a) Give expressions for the least squares estimators of the regression coefficients.
(b) Obtain the variance-covariance matrix of the estimated regression coefficients.
(c) Show that the $i^{\text{th}}$ leverage is $h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2}$.

[3 + 7 + 5 = 15]

10. The variance-covariance matrix of a data set, with 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]