PGDBA 2018-20 Semester 1: Foundations of Database Systems
Mid Sem Exam
Time: 2 hours and 30 minutes
All questions carry equal marks.

Part A

Please refer to the SQL construct statements below for the questions in this part. These tables represent a simplified version of human resource data for an organization at a given point of time.

```
CREATE TABLE employees (
    emp_no      INT             NOT NULL,
    birth_date  DATE            NOT NULL,
    first_name  VARCHAR(32)     NOT NULL,
    last_name   VARCHAR(32)     NOT NULL,
    email       VARCHAR(64)     NOT NULL,
    profile     VARCHAR(2048),
    gender      ENUM ('M','F')  NOT NULL,
    PRIMARY KEY (emp_no)
);

CREATE TABLE managers (
    emp_no          INT             NOT NULL,
    manager_emp_no  INT             NOT NULL,
    PRIMARY KEY (emp_no),
);

CREATE TABLE salaries (
    emp_no      INT             NOT NULL,
    salary      INT             NOT NULL,
    FOREIGN KEY (emp_no) REFERENCES employees (emp_no) ON DELETE CASCADE,
    PRIMARY KEY (emp_no)
);
```

1. Suppose an employee's writes a profile description of 39 characters (1 character = 1 byte). How many bytes would be consumed to store his/her profile in the column `profile` of the table `employees`? Explain.

2. Write an SQL query to output the percentage of female employees in the organization.

3. Write an SQL query (with brief explanation) to output the names of all the employees who are not managers.

4. Write an SQL query (with brief explanation) to output the average salary of the managers.

5. Suppose every employee, except the general manager, must have exactly one manager, who is also another employee. The table `managers` stores this information, by storing the employee numbers of the employees and their manager's. Then, write an SQL query (with brief explanation) to output the full name and the profile of the general manager of the organization, in the following format

   ```
   full_name       profile
   ```

6. Is there any employee whose salary is greater than the salary of his/her manager? Write an SQL query (with brief explanation) to find out the answer.

   ```
   full_name       manager_email
   ```

Part B

A distributed Hadoop file system contains customer transaction data for a retail store. Every time a customer makes a purchase, the tuples (`customer_id, product_id, timestamp`) are stored in the filesystem for all products purchased by the customer. Consider this data as input to design MapReduce algorithms (map and reduce functions) for the following problems. Write appropriate explanation for each of the solutions.

1. Describe a MapReduce algorithm to determine which customer has bought the most number of distinct products from the store.

2. Describe a MapReduce algorithm to compute all pairs of customers (A,B) such that A and B have bought the same product on the same day during the same hour of day from this store at least once. In your solution, you can assume there are obvious ways (functions) to convert timestamp into day and hour of day.

INDIAN STATISTICAL INSTITUTE

Midsemester Examination : (2018-2019)

PGDBA 1st Year

Statistical Structures in data

Date: 12. 09. 2018 Maximum marks: 60 Duration: 2 hours.

Note: This paper carries 65 points. Maximum you can score is 60.

1. Suppose that male and female students are applying to a particular college for admission in three different streams. Illustrate Simpson's paradox with hypothetical data for this scenario. [10]

2. Suppose the mean, median and the standard deviation of a list of 10 numbers was calculated to be 15, 14 and 3 respectively. It was later found that one of the number in the list was mistakenly recorded as 1 which should have been 11. Calculate the correct values of the mean, median and standard deviation. [10]

3. Using the Cauchy-Schwarz inequality or otherwise, prove that the measure of Kurtosis $\beta_2 = \frac{m_4}{m_2^2}$ is always greater than or equal to the measure of skewness $\beta_1 = \frac{m_3^2}{m_2^3}$ [8]

4. The average height and weight of a group of students turned out to be 5 ft 6 inches and 65 kilograms with standard deviations 3 inches and 10 kilograms respectively. Using the regression equation for predicting weight from height, the estimated weight of a 6 ft tall student was calculated to be 80 kilograms. Predict the height of a student whose weight is 60 kilograms. [10]

5. Suppose Y is regressed on X_1, X_2 and X_3 with an intercept term, and the following matrices are computed.

$$\text{Y'Y= } 5000 \quad \text{Y'X= } (20, 30, 50, -40), \quad \text{X'X= } \begin{bmatrix} 20 & 0 & 0 & 0 \\ 0 & 19 & 3 & 0 \\ 0 & 3 & 1 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

[P.T.O.]

(a) Compute the mean and variance of Y. [5]

(b) Compute the regression equation. [6]

(c) Compute the analysis of variance of Y by deriving total, regression and residual sum of squares. Produce the ANOVA table. [8]

(d) Compute the estimate of error variance. Find the estimated variance for each regression coefficient. Also find the estimated covariances between them. [8]

INDIAN STATISTICAL INSTITUTE

Midsemester Examination : (2018-2019)

PGDBA 1st Year

Statistical Structures in data

Date: 12. 09. 2018 Maximum marks: 60 Duration: 2 hours.

Note: This paper carries 65 points. Maximum you can score is 60.

1. Suppose that male and female students are applying to a particular college for admission in three different streams. Illustrate Simpson's paradox with hypothetical data for this scenario. [10]

2. Suppose the mean, median and the standard deviation of a list of 10 numbers was calculated to be 15, 14 and 3 respectively. It was later found that one of the number in the list was mistakenly recorded as 1 which should have been 11. Calculate the correct values of the mean, median and standard deviation. [10]

3. Using the Cauchy-Schwarz inequality or otherwise, prove that the measure of Kurtosis $\beta_2 = \frac{m_4}{m_2^2}$ is always greater than or equal to the measure of skewness $\beta_1 = \frac{m_3^2}{m_2^3}$ [8]

4. The average height and weight of a group of students turned out to be 5 ft 6 inches and 65 kilograms with standard deviations 3 inches and 10 kilograms respectively. Using the regression equation for predicting weight from height, the estimated weight of a 6 ft tall student was calculated to be 80 kilograms. Predict the height of a student whose weight is 60 kilograms. [10]

5. Suppose Y is regressed on X_1, X_2 and X_3 with an intercept term, and the following matrices are computed.

$$\text{Y'Y= } 5000 \quad \text{Y'X= } (20, 30, 50, -40), \quad \text{X'X= } \begin{bmatrix} 20 & 0 & 0 & 0 \\ 0 & 19 & 3 & 0 \\ 0 & 3 & 1 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

[P.T.O.]

(a) Compute the mean and variance of Y. [5]

(b) Compute the regression equation. [6]

(c) Compute the analysis of variance of Y by deriving total, regression and residual sum of squares. Produce the ANOVA table. [8]

(d) Compute the estimate of error variance. Find the estimated variance for each regression coefficient. Also find the estimated covariances between them. [8]

INDIAN STATISTICAL INSTITUTE

Mid – Semester Examination: 2018 - 19

Course Name: PGDBA

Subject Name: Inference

Date: 13 September 2018 **Maximum Marks: 100** **Duration: 3 hours**

Notes, if any: The paper carries 105 marks. Answer as much as you can. However, the maximum you can score is 100.

1. In 1970s, some legal experts believed that the verdict of death penalty in cases of homicide (murder) in the US depended on the race (white / black) of the victim and the defendant (the person accused of the murder). A detailed study covering 326 cases where death penalty was awarded to the defendant was carried out and the following were observed. Out of 151 cases where both defendant and victim were white, 19 were awarded death penalty. Out of 9 cases where defendant was white and victim black, none were awarded death penalty. Out of 63 cases where defendant was black and victim white, 11 were awarded death penalty. Finally, out of 103 cases where both defendant and victim were black, 6 were awarded death penalty.
    a. Construct a two way table to present the number of black and white defendants being awarded death penalty or a lighter sentence irrespective of the race of the victim. [5]
    b. Use the above table to estimate the relative risk of a white defendant being awarded death penalty vis-à-vis black defendant irrespective of the race of the victim. Do this table provide any evidence of racial discrimination against black defendants? [3]
    c. Present the entire data in tabular format showing race of both the defendant as well as the victim. [5]
    d. Do you think that the data support the belief that the race of the victim (white or black) has an impact on the chance of awarding death penalty? Explain, preferably with a visual presentation. [7]
2. A telecom service provider carried out a survey to understand the importance the customers attach to the different aspects of telecom service. A total of 357 customers were randomly selected and they were asked to provide importance rating for different aspects of service in a 7 point (1 to 7) scale, where 1 implies almost no importance to the particular characteristic and 7 indicates the highest level of importance. The data collected for two different characteristics are given below.

Rating	Variable					
	Store Experience			Consistency of Service Delivery		
	Frequency	Proportion	Cumulative Proportion	Frequency	Proportion	Cumulative Proportion
1	4	0.011	0.011	1	0.003	0.003
2	6	0.017	0.028	1	0.003	0.006
3	7	0.020	0.048	5	0.014	0.020
4	35	0.100	0.148	34	0.095	0.115
5	122	0.341	0.489	97	0.272	0.387
6	134	0.375	0.864	155	0.434	0.821
7	49	0.136	1.000	64	0.179	1.000
Total	357	1.000		357	1.000	

Note that the rating given by any randomly selected person for any particular characteristic (store experience or consistency of service delivery) is the random variable. Thus we are looking at two random variables (one for each characteristic) with similar distribution.

a. What distribution are the random variables expected to follow? What are their parameters? [2 + 2 = 4]
b. Draw ogives for both these random variables. [8]
c. Which of the two characteristics seem to be more important to the customers? Explain using the ogives. [6]
d. A box plot gives a 5 point summary of a random variable. What are these 5 points? Just give the names, no explanation is necessary. [2]

3. A company bought 1,00,000 electric batteries from a manufacturer with an understanding that the consignment will have no more than 3% defective items. As it is not possible to test each battery, the company selected a random sample of 400 batteries and tested them for defects. Ten batteries turned out to be defective. Assume that the number of defective batteries in a set of batteries tested follows a Binomial distribution.
   a. Do you think that the assumption of Binomial distribution is reasonable? Explain. [4]
   b. What is the sample proportion in this case? What is its standard error? [2 + 3 = 5]
   c. Can the company safely assume that the consignment does not have more than 3% defective items? Explain. [8]
   d. Notice that a defective battery may have more than one defect while a non-defective battery will have no defect. Suppose 6 of the 10 defective batteries have 1 defect each, 3 has 2 defects and 1 has 3 defects.
      i. What distribution is the number of defects on a battery likely to follow?
      ii. Estimate the parameter(s) of the distribution in this case. [1 + 2 = 3]
4. A box contains very large number of tickets. A number is written on each ticket and it is known that the average of these numbers is 100 with an SD of 20. Suppose 400 draws were made randomly from this box
   a. Estimate the chance that the average of the draws will be in the range 99 to 101. What distributional assumption did you make and why? [3 + 3 = 6]
   b. Suppose the average is unknown and the average of the 400 draws turns out to be 101. Construct a 95% two sided confidence interval for the population mean. What is the population in this case? [4 + 2 = 6]
   c. How will the confidence interval change in case the standard deviation is unknown and has been estimated to be 20 from the sample of size 400? [4]
   d. What is a consistent estimator? Is the sample average a consistent estimator of the population average? [4]
5. We often come across surveys conducted by newspapers or TV channels where people are asked to comment whether they support some issue or not (typically a yes / no type of an answer).
   a. Identify the sample and the population in this case. [2 + 2 = 4]
   b. Suppose you are looking at a set of respondents and counting the number of people supporting an issue (people who said yes in response to a question). What distribution is the random variable likely to follow? What assumptions are you making? [2 + 3= 5]
   c. In a particular case an issue was supported by 80% of the 400 respondents. Estimate the relevant parameters and construct their confidence intervals. In case you think that the parameter(s) cannot be estimated, give reasons. [6]
6. In a particular year 64 students wrote an examination in ISI. Their scores averaged 72.5 out of 100. The sample standard deviation was 16. Can you compute the standard error of the sample average and construct a confidence interval for the population average using an appropriate distribution? Explain. Specify the sample and the population and the parameters you are estimating. [10]

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination: 2018

Course Name: Post Graduate Diploma in Business Analytics (PGDBA)

Subject: Computing for Data Sciences

Date: 14.09.2018

Full Marks: 60 Duration: 2.5 hrs.

Answer any four questions.

1. a) Compute an eigen-decomposition of the following matrix A:

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$$

b) With the help of the eigen-decomposition thus computed, calculate the 5-th power of matrix A i.e. A^5.

10+5 = 15

2. a) Find the orthogonal projection of the vector $v = [1, 1, 1]^T$ onto the plane **W** spanned by the orthonormal vectors $[0, 1, 0]^T$ and $\left[-\frac{4}{5}, 0, \frac{3}{5}\right]^T$.

b) Give an example of a 3×3 matrix A and a vector b ($\in \mathbb{R}^3$) such that the solutions of $Ax = b$ form a line in $\mathbb{R}^3$, $b \neq 0$, and all the entries of the matrix A are nonzero. Find all solutions of x.

c) Find a basis for the null space of the following matrix A:

$$A = \begin{pmatrix} 1 & 0 & 0 & 3 & 2 \\ 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 1 & 2 & 3 \end{pmatrix}$$

5+5+5 = 15

3. a) Perform the Singular Value Decomposition (SVD) on the matrix and show the necessary steps:

$$A = \begin{pmatrix} 2 & 1 \\ 2 & -1 \\ 1 & 0 \end{pmatrix}$$

b) Compute the Moore-Penrose pseudo-inverse of A by using the SVD obtained for Part (a).

10 + 5 = 15

4. You are given a quadratic polynomial $f(x_1, x_2, x_3)$:

$$f(x_1, x_2, x_3) = 2x_1^2 - 2x_1x_2 - 4x_1x_3 + x_2^2 + 2x_2x_3 + 3x_3^2 - 2x_1 + 2x_3$$

a) Write the polynomial $f(x_1, x_2, x_3)$ in the form $f(x) = x^T Ax - b^T x$, where $x = [x_1, x_2, x_3]^T$, A is a real symmetric matrix, and b is some constant vector.

b) Find the point $x^* = [x_1^*, x_2^*, x_3^*]^T$ where $f(x_1, x_2, x_3)$ attains an extremum or stationary value.

c) Is this point $x^* = [x_1^*, x_2^*, x_3^*]^T$ a minimum, maximum, or saddle point of some kind? Justify your answer with suitable optimality tests.

4 + 6 + 5 = 15

5. Consider minimization of the function:

$$f(x_1, x_2) = -12x_2 + 4x_1^2 + 4x_2^2 + 4x_1x_2$$

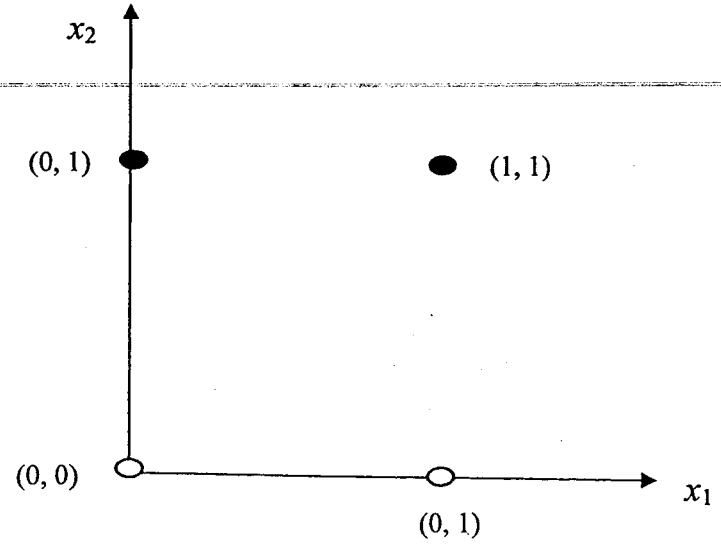
by the method of steepest descent. Let us start with the initial solution $x^0 = [0, 1]^T$. Hand trace the steepest descent algorithm with exact line search to determine the optimal learning rate parameter at every iteration. Continue up to the first 4 iterations and report the norm of the final solution gradient. Show all your work.

15

6. a) Is the following function $f: \mathbb{R}^3 \rightarrow \mathbb{R}$ convex? Justify.

$$f(x_1, x_2, x_3) = 3x_1^2 - 4x_1x_2 + x_2^2 + x_3^2$$

b) Consider a simple classification problem where the training set contains 4 points in $\mathbb{R}^2$ as shown in the following figure.



The black circles in the figure denote two data points (with feature vectors (0, 1) and (1, 1)) belonging to one class with class label y = +1 and the white circles denote the remaining points (with feature vectors (0, 0) and (0, 1)) belonging to the other class with class label y = -1.

Design a linear support vector machine classifier for this dataset and derive the equation of the separating line by using the Lagrangian multiplier method. Show all your work.

5 + 10 = 15

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination: 2017-18

PGDBA 2019-20

Statistics Comprehensive

Date: ~~10~~ 17 September 2018 | Maximum marks: 30 | Duration: 2 hours

Answer any five out of the six questions.

1. If 4 married couples are arranged in a row, find the probability that no husband sits next to his wife. [6]

2. Consider an unending sequence of independent trials, where each trial is equally likely to result in any of the outcomes A, B or C. Given that outcome C is the last of the three outcomes to occur, find the conditional probability that the first two outcomes are identical. [6]

3. Suppose that it takes at least 9 votes from a 12-member jury to convict a defendant. Suppose also that the probability that a juror votes a guilty person innocent is 0.2, whereas the probability that the juror votes an innocent person guilty is 0.1. If each juror acts independently and if 65 percent of the defendants are guilty, find the probability that the jury renders a correct decision. What percentage of defendants is convicted? [6]

4. Suppose that the life distribution of an item has the hazard rate function $\lambda(t) = t^3, t > 0$.

   (a) What is the probability that a new item survives to age 2?
   (b) What is the probability that a 1-year-old item will survive to age 2? [3+3=6]

5. Consider the bivariate distribution function of the random variables X and Y:

$$F_{XY}(x,y) = \begin{cases} (1-e^{-x})\Phi^{1/2}(y), & 0 < x \leq -\ln[1-\Phi(y)], \quad -\infty < y < \infty, \\ (1-e^{-x})^{1/2}\Phi(y), & x > -\ln[1-\Phi(y)], \quad -\infty < y < \infty, \end{cases}$$

   where Φ is the standard normal distribution function.

   (a) Determine the marginal distributions of X and Y.
   (b) Determine the copula for this bivariate distribution. [3+3=6]

6. A health insurer tracks the monthly status of a retired person as *Healthy* (state 1), *Sick* (state 2) or *Dead* (state 3). The (one-step) transition probability matrix is

$$P = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.5 & 0.3 & 0.2 \\ 0 & 0 & 1 \end{pmatrix}.$$

   (a) What is the probability that a healthy person under this insurance will stay healthy after two months?
   (b) Show that the *Sick* state is transient.
   (c) What is the expected number of months that a healthy person will be sick? [2+2+2=6]

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination: 2017-18

PGDBA 2019-20

Stochastic Processes and Applications

Date: 19 September 2018 Maximum marks: 30 Duration: 2 hours

Answer any five out of the six questions.

1. If 4 married couples are arranged in a row, find the probability that no husband sits next to his wife. [6]

2. Consider an unending sequence of independent trials, where each trial is equally likely to result in any of the outcomes A, B or C. Given that outcome C is the last of the three outcomes to occur, find the conditional probability that the first two outcomes are identical. [6]

3. Suppose that it takes at least 9 votes from a 12-member jury to convict a defendant. Suppose also that the probability that a juror votes a guilty person innocent is 0.2, whereas the probability that the juror votes an innocent person guilty is 0.1. If each juror acts independently and if 65 percent of the defendants are guilty, find the probability that the jury renders a correct decision. What percentage of defendants is convicted? [6]

4. Suppose that the life distribution of an item has the hazard rate function $\lambda(t) = t^3, t > 0$.

   (a) What is the probability that a new item survives to age 2?
   (b) What is the probability that a 1-year-old item will survive to age 2? [3+3=6]

5. Consider the bivariate distribution function of the random variables X and Y:

$$F_{XY}(x,y) = \begin{cases} (1-e^{-x})\Phi^{1/2}(y), & 0 < x \leq -\ln[1-\Phi(y)], \quad -\infty < y < \infty, \\ (1-e^{-x})^{1/2}\Phi(y), & x > -\ln[1-\Phi(y)], \quad -\infty < y < \infty, \end{cases}$$

   where Φ is the standard normal distribution function.

   (a) Determine the marginal distributions of X and Y.
   (b) Determine the copula for this bivariate distribution. [3+3=6]

6. A health insurer tracks the monthly status of a retired person as *Healthy* (state 1), *Sick* (state 2) or *Dead* (state 3). The (one-step) transition probability matrix is

$$P = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.5 & 0.3 & 0.2 \\ 0 & 0 & 1 \end{pmatrix}.$$

   (a) What is the probability that a healthy person under this insurance will stay healthy after two months?
   (b) Show that the *Sick* state is transient.
   (c) What is the expected number of months that a healthy person will be sick? [2+2+2=6]

INDIAN STATISTICAL INSTITUTE

Supplementary Mid-Semester Examination: 2017-18

PGDBA 2019-20

Stochastic Processes and Applications

Date: 26 October 2018 | Maximum marks: 30 | Duration: 2 hours

Answer any five out of the six questions.

1. If there are 12 strangers in a room, what is the probability that no two of them celebrate their birthdays in the same month? [6]

2. A and B alternate rolling a pair of dice, stopping either when A rolls the sum 9 or when B rolls the sum 6. Assuming that A rolls first, find the probability that the final roll is made by A. [6]

3. A communications channel transmits the digits 0 and 1. However, due to static, the digit transmitted is incorrectly received with probability 0.2. Suppose that we want to transmit a message consisting of one binary digit. To reduce the chance of error, we transmit 00000 instead of 0 and 11111 instead of 1. If the receiver of the message uses "majority" decoding, what is the probability that the message will be wrong when decoded? What independence assumptions are you making? [6]

4. If X is uniformly distributed over $(-1, 1)$, derive

   (a) $P\left(|X| > \frac{1}{2}\right)$;

   (b) the density function of the random variable $|X|$. [3+3=6]

5. Suppose the random variables X and Y have marginal distributions $F_X(x)$ and $F_Y(y)$ and the copula of their joint distribution is $C(u, v)$. Express the cumulative distribution function of $Z = \min\{X, Y\}$ in terms of these three functions. [6]

6. Consider a Markov chain having state space $\{0,1,\ldots,6\}$ and transition matrix

$$P = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{8} & \frac{1}{4} & \frac{1}{8} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

   (a) Determine which states are transient and which states are recurrent.

   (b) Find ρ_{0y}, $y = 0,1,\ldots,6$. [3+3=6]

INDIAN STATISTICAL INSTITUTE

First Semester Examination: 2018-19

POST GRADUATE DIPLOMA IN BUSINESS ANALYTICS

Stochastic Processes and Applications (BAISI1)

Date: 26 November 2018 Maximum marks: 100 Duration: 3.5 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. Five balls are randomly chosen, without replacement, from an urn containing 5 red, 6 white, and 7 blue balls. Find the probability that at least one ball of each colour is chosen. [10]

2. Two local factories, A and B, produce radios. Each radio produced at factory A is defective with probability 0.05, whereas each one produced at factory B is defective with probability 0.01. Suppose you purchase two radios that were produced at the same factory, which is equally likely to have been either factory A or factory B. If the first radio that you check is defective, what is the conditional probability that the other one is also defective? [7]

3. (a) The number of eggs X laid on a tree by an insect of a certain type is a Poisson random variable with parameter λ. However, X is observed only when it is not zero, as we do not know whether the insect had been on the leaf if there is no egg. Let Y be the observed number of eggs, which means

$$P(Y = i) = P(X = i|X > 0), \quad i = 1,2,3, \ldots.$$

Calculate $E(Y)$.

(b) The amount of time a customer spends at a railway reservation counter, while being served by a booking clerk, is an exponential random variable with mean 3 minutes. If there is a single customer being served when you stand in the queue, what is the probability that your turn will not arrive for another 5 minutes? How does your answer change if there is another person in the queue before you? [4+(4+3)=11]

4. Consider the bivariate distribution of the random variables X and Y

$$F_{XY}(x,y) = \frac{(1-e^{-x})(1-e^{-y^2})}{1-\theta e^{-x-y^2}}, \quad 0 < x,y < \infty.$$

(a) Determine the marginal distribution functions of X and Y.
(b) Determine the copula of the joint distribution F_{XY}.
(c) For which value of θ does the answer in (b) reduce to the independence copula?
(d) Suppose Z is another random variable with marginal distribution

$$F_Z(z) = 1-(1+z)e^{-z}, \quad 0 < z < \infty.$$

If the bivariate distribution of X and Z has the same copula as in part (c), determine that bivariate distribution. [2+3+1+3=9]

P T O

5. There are N balls in a box, each coloured either green or red. At the end of every minute, a ball is taken out of the bag at random and is replaced by a ball of the other colour. Let X_n be the number of green balls in the box after n replacements.

   (a) Explain why the stochastic process $X_n, n \geq 0$ with state space $S = \{0,1,\dots,N\}$ is a Markov chain.
   (b) Determine the transition probabilities of the Markov chain.
   (c) Explain whether the chain is irreducible.
   (d) Show that the chain is periodic and determine the periods of all the states.
   (e) Show that for this process, the stationary distribution is given by

$$\pi(j) = \binom{N}{j}\frac{1}{2^N}, \qquad \text{for } j = 0,1,2,\dots,N.$$

   [Hint: Do not deduce; simply verify whether this is adequate.]
   (f) Explain whether $\lim_{n\to\infty} P(X_n = j | X_0 = 0) = \pi(j)$. [1+3+1+2+3+1=11]

6. (a) Consider the Poisson process $X(t),\ t \geq 0$ with parameter λ. The process is observed at a random time T, which is independent of $X(t)$ and has the exponential distribution rate parameters ν. Find the distribution of $X(T)$, the count of events till time T.
   (b) Find the transition function of the two-state birth and death process by solving the forward equation. [4+8=12]

7. Let w_t, for $t = 0, \pm1, \pm2, \dots$ be a set of independent and identically distributed (i.i.d.) random variables having the normal distribution $N(0,\sigma^2)$, and consider the time series

$$y_t = w_t w_{t-1}.$$

   (a) Determine the mean and autocovariance function of y_t.
   (b) Explain why the process y_t is weakly stationary.
   (c) Is y_t strictly stationary?
   (d) Determine the partial autocorrelation function (PACF) of y_t. [4+1+4+1=10]

8. Consider the process y_t defined by the equation

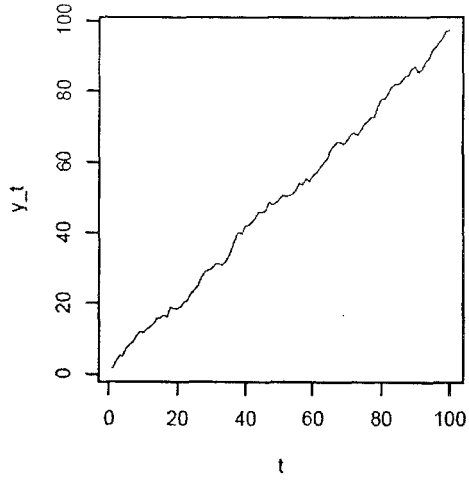
$$y_t = \varphi y_{t-1} + w_t, \qquad t = 1,2,\dots,n,$$

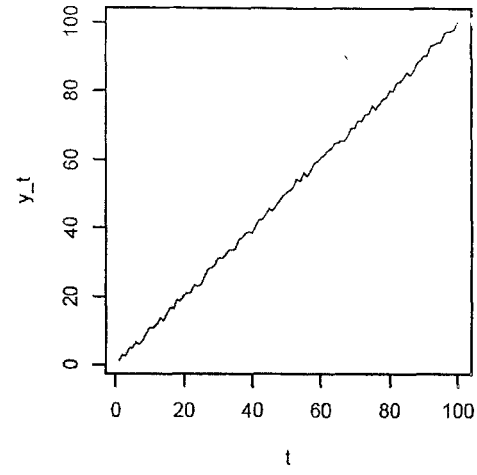
where φ is an unspecified parameter with $|\varphi| < 1$, w_t is white noise with mean 0 and variance σ^2, and y_0 is uncorrelated with this white noise sequence and has mean 0 and variance σ_0^2.

   (a) For which value of σ_0^2 is the process y_t stationary? Explain.
   (b) If y_t is indeed stationary and $y_1, y_2, \dots, y_n$ are observed, derive the estimator of φ by the method of moments (i.e., by matching the first two sample autocovariances with their expressions derived from the model). *You can assume that the process has zero mean.* [7+3=10]

PTO

9. Examine carefully the plots of two simulated time series given below.





One of the plots represents a linear trend added to simulated white noise, while the other plot represents a simulated random walk with constant drift. The simulation models are

$$y_t = \beta_0 + \beta_1 t + w_t, \quad w_t \sim N(0, \sigma^2), \qquad i.i.d., \qquad t = 1,2,\ldots,100;$$

$$y_t = \delta + y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \tau^2), \qquad i.i.d., \qquad t = 1,2,\ldots,100.$$

(a) By examining the two plots carefully, use your best guess to identify them, with reasons.

(b) Guess the values of the parameters of the two models, with reasons, from the given plots. [5+5=10]

10. Consider the process

$$x_t = \beta_0 + \beta_1 t + z_t, \quad t = 0,1,2,\ldots,$$

where β_0 and β_1 are fixed constants and z_t is a weakly stationary process with zero mean.

(a) Determine whether the process x_t is weakly stationary.

(b) Determine whether the process $\nabla x_t = x_t - x_{t-1}$ is weakly stationary.

(c) If z_t is a stationary ARMA(1,2) process, what sort of process is ∇x_t? What is its mean? [1+4+(3+2)=10]

11. Suppose x_1, x_2, x_3, x_4 are part of a weakly stationary time series with mean $\mu = 3$ and autocovariance function

$$\gamma(h) = \begin{cases} 1 & \text{if } h = 0, \\ 0.4 & \text{if } |h| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Calculate the best linear predictor (BLP) of x_4 in terms of x_1, x_2, x_3 that minimizes the mean squared prediction error. [You can use any result that was proved in class.]

(b) Calculate the mean squared prediction error of the BLP of part (a).

(c) Compute the mean squared prediction error of the BLP of x_4 in terms of x_3 alone. Is it in expected order with the answer to part (b)? Explain. [6+2+(1+1)=10]

Indian Statistical Institute
Foundations of Database Systems
End Semester Examination

PGDBA 2018-20 1st Semester

Total marks: 50
Date: 28 November 2018
Time: 2:30 Hours

Each question carries 5 marks.

1. Write an SQL query to get the second highest salary from the `Employee` table, one instance of which is shown below.

```
Employee
+----+--------+
| Id | Salary |
+----+--------+
| 1  | 100    |
| 2  | 200    |
| 3  | 300    |
+----+--------+
```

For example, given the above Employee table, the query should return 200 as the second highest salary. If there is no second highest salary, then the query should return `null`.

```
+---------------------+
| SecondHighestSalary |
+---------------------+
| 200                 |
+---------------------+
```

2. Define *candidate key* and *superkey* for a relation. Is any candidate key also a superkey? Is any superkey also a candidate key? Explain.

3. If R itself (the set of all attributes of the relation R) is a candidate key for R, then prove or disprove: for any two different attributes A and B of R, the functional dependency $A \rightarrow B$ cannot hold.

4. Give an example of a relation R and a set of functional dependencies $\mathcal{F}$ between it's attributes so that R is in 2NF, but not in 3NF or BCNF. Justify your answer.

5. Let $R = (A, B, C, D, E, G)$ with functional dependencies $\mathcal{F} = \{A \to BCD, B \to G, D \to E\}$. Show that R is not in BCNF. Demonstrate a dependency preserving and lossless decomposition of R into multiple relations.

6. Let R be a relation with functional dependencies $\mathcal{F}$ and $A, B \in R$ be attributes in R. Prove or disprove: If the canonical cover $\mathcal{F}_c$ of $\mathcal{F}$ does not contain any functional dependency of the form $AB \to Z$ for some $Z \subseteq R$, then $A^+B^+ = (AB)^+$.

7. What is a *sparse index*? Can both the primary and secondary indexes be sparse? Explain your answers.

8. Recall that in a database D of transactions, the support of an itemset A is defined as the fraction of transactions in which all items in A were bought. Prove or disprove: for itemsets X and Y, $|X| \leq |Y| \implies \text{support}(X) \geq \text{support}(Y)$.

9. Suppose you have a 3TB file containing one integer per line in a 30-node Hadoop cluster, each of which has 1TB hard disk space and 8GB RAM. Write *map* and *reduce* functions to compute the top-10 integers from the given file.

10. Suppose M is an $m \times n$ matrix each of whose columns are 0-1 vectors in n dimensions (m and n are very large). Define a version of min-hash function for hashing the columns of M to hash buckets (one dimension). If two columns v and w of M has Jaccard similarity s, then what can be said about the min-hash values of v and w? State the property, no proof is required.

INDIAN STATISTICAL INSTITUTE

Semester Examination : (2018-2019)

PGDBA 1st Year

Statistical Structures in Data

Date: 30. 11. 2018 Maximum marks: 100 Duration: 3 hours.

Note: This paper carries 110 points. Maximum you can score is 100.

1. The average height and weight of a group of students turned out to be 5 ft 6 inches and 65 kilograms respectively. Using the regression equation for predicting weight from height, the estimated weight of a 6 ft tall student was calculated to be 80 kilograms.

   (a) What will be the predicted weight of a 5 ft tall student? [5]

   (b) Given that the standard deviations of heights and weights were 3 inches and 12 kilograms respectively, find the correlation coefficient between heights and weights. [5]

2. Suppose Y is regressed on X_1, X_2 and X_3 with an intercept term which yielded

$$\text{Y'Y}= 5000 \quad \text{Y'X}= (20, 30, 50, -40), \quad \text{X'X}= \begin{bmatrix} 20 & 0 & 0 & 0 \\ 0 & 19 & 3 & 0 \\ 0 & 3 & 1 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

   (a) Find the multiple correlation coefficient between Y and X_1, X_2 and X_3. [10]

   (b) Test whether the regression is effective at the significance level .05. [10]

3. Let $\mathbf{X}$ be distributed as $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}' = [1, -1, 2]$ and

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

   (a) Find the conditional distibution of X_1, given that $X_3 = 3$. [8]

   (b) Find the conditional distibution of X_1, given that $X_2 = 0$ and $X_3 = 3$. Compare this with your answer in part(a) and comment. [8+4]

[P. T. O.]

4. Suppose $\mathbf{X}$ is a random vector having $N_3(0, \mathbf{\Sigma})$ distribution where

$$\mathbf{\Sigma} = \begin{bmatrix} 4 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 4 \end{bmatrix}$$

(a) Find the principal components, and the proportion of total variance explained by the first two of them. [12 + 3]

(b) If instead the principal components were computed from the correlation matrix, would they be the same as in part (a)? Give a complete explanation with computation if necessary. [5]

5. (a) Describe the Principal Factor Analysis method. [10]

(b) Suppose in this method, all the initial estimates of the specific variances are given by the same constant c. Derive the estimated factor loadings in this case. [10]

6. In an ecological study of the feeding behaviour of birds, the number of hops between flights was counted for several birds.

Number of hops	Frequency
1	48
2	31
3	20
4	9
5	6
6	5
7	4
8	2
9	1
10	1
11	2
12	1

(a) Fit an appropriate distribution to the above data and estimate the relevant parameter(s).

(b) Calculate the chi-square statistic for goodness of fit. Comment on the results. [10+10]

INDIAN STATISTICAL INSTITUTE

End-Semestral Examination: 2018

Course Name: Post Graduate Diploma in Business Analytics (PGDBA)

Subject: Computing for Data Sciences

Date: 03.12.2018

Full Marks: 100

Duration: 3.5 hrs.

Answer as much as you can.

1. a) Consider the problem of minimizing the following function:

$$f(x_1, x_2) = x_1^2 + 2x_2^2 - 2x_1x_2 - 2x_2 + 2x_1.$$

i) Compute the optimal solution vector $x^* = [x_1^*, x_2^*]^T$ to this problem. Is the optimal solution unique? Give reasons for your answer. 2+1+3 = 6

ii) Suppose the Newton's algorithm is used to minimize the same function $f(x_1, x_2)$ in Q. 1(a) (i). Starting from the point $x^0 = [2,2]^T$, hand trace the iterations of the Newton's algorithm either till convergence or up to 3 iterations, whichever occurs earlier. (6)

b) Find the orthogonal projection of the vector $v = [2, 1, -1]^T$ onto the subspace spanned by the vectors $v_1 = [1,0,1]^T$ and $v_2 = [1,1,0]^T$. (6)

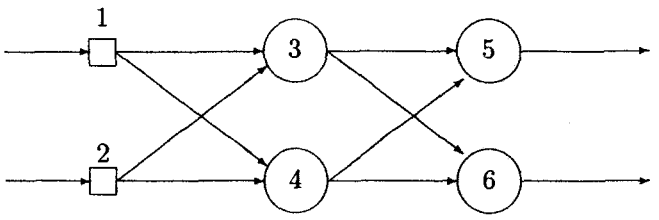
c) You are given the following matrix:

$$A = \begin{pmatrix} 0 & -1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & -1 \end{pmatrix}.$$

Calculate the sum and product of the eigenvalues of the matrix AA^T without explicitly computing the same eigenvalues and show your steps. (7)

2. a) The following diagram represents a feed-forward neural network with one hidden layer:

P TO



A weight on connection between nodes i and j is denoted by w_{ij} , for example, w_{13} is the weight on the connection between nodes 1 and 3. The following table lists all the weights in the network:

$w_{13} = -2$	$w_{35} = 1$
$w_{23} = 3$	$w_{45} = -1$
$w_{14} = 4$	$w_{36} = -1$
$w_{24} = -1$	$w_{46} = 1$

Suppose each of the nodes 3, 4, 5, and 6 uses the following activation function:

$$\varphi(v) = \begin{cases} 1 & \text{if } v \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where v denotes the weighted sum of a node. Each of the input nodes 1 and 2 can only receive binary values (0 or 1). Calculate the output of the network (y_5 and y_6) for the following input patterns:

Pattern:	P_1	P_2	P_3	P_4
Node 1:	0	1	0	1
Node 2:	0	0	1	1

5+5 = 10

b) The Performance of two classifiers are listed in the following table on ten test instances for a two-class classification problem along with the actual class labels.

Data #	Actual Class	Class predicted by Classifier 1	Class predicted by Classifier 2
1	+	+	+
2	+	-	+
3	-	-	-
4	-	+	+
5	+	+	-
6	-	+	-
7	+	+	+
8	+	+	-
9	-	-	-
10	-	+	-

Prepare the confusion matrix and compare the classifiers 1 and 2 on the basis of precision, recall, and F-score. Show all your calculations.

3+3+3 = 9

PTO

c) Briefly discuss the vanishing gradient problem in connection to the backpropagation learning in multi-layer neural networks. How does ReLU type activation functions provide some resistance to the vanishing gradient problem? 3+3 = 6

3. a) In the following table, a dataset is provided with four categorical attributes: Sky, Temperature, Humidity, and Wind associated with a target class Play Football. Derive a simple decision tree classifier by using the information gain criterion for deciding root and other nodes of the tree. Clearly indicate the calculations in each layer. Draw the final trained decision tree where at each node the splitting criterion is indicated.

(12)

Day	Sky	Temperature	Humidity	Wind	Play_Football
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weal	Yes
14	Rain	Mild	High	Strong	No

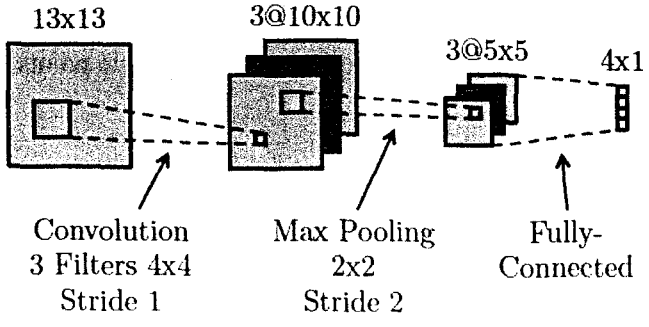
b) Suppose you are given a data matrix consisting of four data points in $\mathbb{R}^2$ in the following form:

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

i) Compute the covariance matrix for the sample points. (Warning: Observe that X is not centered.) Then compute the unit eigenvectors and the corresponding eigenvalues of the covariance matrix. (8)

ii) Suppose we use PCA to project the points onto a one dimensional space. For each of the four sample points in X, write the coordinate (in the principal coordinate space, not in $\mathbb{R}^2$) that the point is projected to. (5)

4. a) Below is a diagram of a small convolutional neural network that converts a 13×13 image into 4 output values. The network has the following layers/operations from input to output: convolution with 3 filters, max pooling, ReLU, and finally a fully-connected layer. For this network we will not be using any bias/offset parameters. Please answer the following questions about this network.



PTO

i) How many weights in the convolutional layer do we need to learn?

ii) How many ReLU operations are performed on the forward pass?

iii) How many weights do we need to learn for the entire network? 3+3+3 = 9

b) Consider the following toy image matrix and the two simple 3×3 filters.

1	0	1	0	0	1
0	1	0	0	1	0
1	0	0	1	0	0
1	0	0	0	0	0
1	0	0	1	1	1
0	0	1	0	0	0

Image Matrix

-1	-1	1
-1	1	-1
1	-1	-1

Filter 1

-1	-1	-1
1	1	1
-1	-1	-1

Filter 2

Derive the feature map that will be obtained after one round of convolution (with stride = 1) of the entire image matrix without zero pooling and maxpooling with filters 1 and 2 on the image matrix.

4+4 = 8

c) Table 3 contains artificial data from songs that can either be classified as Pop or Rock. Using some predefined measures, three real-valued features Vocals, Dynamics, Timbre are generated for six songs. Using the *k*-Nearest Neighbour classifier, predict the Music class for the following two data points: (i) (5.5, 3.3, 2.4), and (ii) (3.1, 5.4, 6.5). Show your work. You are free to choose a suitable value for *k*.

(8)

Example No.	Vocals	Dynamics	Timbre	Music
1	5.1	3.6	1.4	Pop
2	4.9	3.2	1.4	Pop
3	4.6	3.4	1.3	Pop
4	6.5	3.1	5.2	Rock
5	6.1	3.3	5.4	Rock
6	5.9	3.2	5.1	Rock

5. a) Recall the loss function for *k*-means clustering with *k* clusters, *n* sample points $x_1, x_2, \ldots x_n$ and the *k* cluster centers:

$$L = \sum_{j=1}^{k} \sum_{x_i \in S_j} \left\| x_i - m_j \right\|^2,$$

where S_j refers to the set of data points that are closer to center m_j than to any other cluster center.

PTO

i) Instead of updating the centers m_j by computing the mean, suppose L is minimized with the batch gradient descent while holding the sets S_j fixed. Derive the update formula for the center m_1 with learning rate (step size) η. (5)

ii) Derive the update formula for m_1 with stochastic gradient descent on a single sample point x_i. Use learning rate η. (3)

iii) Recall that in the update step of the standard algorithm, we assign each cluster center to be the mean (centroid) of the data points closest to that center. It turns out that a particular choice of the learning rate η (which may be different for each cluster) makes the two algorithms (batch gradient descent and the standard k-means algorithm) to have identical update steps. Let's focus on the update for the first cluster, with center m_1. Calculate the value of η so that both algorithms perform the same update for m_1. (7)

b) Write a short note on any one of the following: (10)

i) Dropout in deep neural networks
ii) Random forest classifiers
iii) The Adagrad algorithm

INDIAN STATISTICAL INSTITUTE

FIRST SEMESTER EXAMINATION: 2018 – 19

Course Name: PGDBA

Subject Name: Inference (BAISI3)

Date: 5.12.2018 **Maximum Marks: 100** **Duration: 3 hours**

Notes, if any: Answer question 5 and any three from the rest

1. Customers of a bank makes calls to a call centre for information or service requests. The bank has outsourced the job to a third party and has stipulated that 90% of the calls need to be attended within 30 seconds. It may be assumed that the time between a call getting connected and attended is known.
   a. Identify the random variable being studied and the parameter to be estimated. [2 + 2 = 4]
   b. Suppose the bank wants to check whether the call centre is able to meet the stipulated requirement or not. Formulate the problem as a problem of hypothesis testing.
      i. Write the null and the alternative hypotheses clearly. [2 + 2 = 4]
      ii. Are these hypotheses simple or composite? Explain briefly. [3 + 3 = 6]
      iii. Suppose you have observed 10 random calls and found that 3 of them had to wait for more than 30 seconds. You are asked to state whether there is sufficient evidence that the call centre is unable to meet the stipulation? How would you address the problem? [6]
   c. Suppose a parametric model was developed to estimate the log odds of a person getting a heart problem given age measured in completed years. The findings are given below:

Variable	Coefficient
Age	0.111
Constant	–5.309

      i. It was found that the log likelihood of the fitted model is -53.68 and log likelihood of the constant only (null) model is –57.45. How will you test whether the coefficient of age is significantly different from zero or not? [3]
      ii. Suppose you know that 95% confidence interval of age does not include zero. In that case is it possible that the estimated SE of the coefficient of age is 0.08? [2]

2. The measurement of a particular characteristic of a product is known to have a large variation but no bias. The measured value is, therefore, a random variable.
   a. Write a model to express the random variable. State your assumptions clearly. [5]
   b. Suppose a sample of n products have been selected and each product was measured twice using the same instrument. How you will estimate the measurement variance? [6]
   c. Suppose the variance of the measurement is known to be 2. Suppose the two different measured values for a particular product are 3 and 5 respectively. Can you say with at least 99% confidence that the true value for the product being measured is not over 6.5? Explain specifying distributional and other assumptions. [8]
   d. Suppose the instrument has a positive but unknown bias, say b. Is it possible to estimate the true value of the product characteristic in this case? Is it possible to estimate the variance of measurement? Explain briefly. [3 + 3 = 6]

3. Suppose a finite population has four elements having values 4, 6, 7 and 10. A random sample of size 2 is drawn from the population without replacement and the sample median is computed as $(X_1 + X_2) / 2$ when X_1, X_2 is the random sample.
   a. What distribution does the sample median follow and what are its parameter(s)? [2]
   b. Plot the distribution function for the sample median. [3]
   c. Is the sample median an unbiased estimate of the population median? Explain breifly. [5]

PTO

d. What is the standard error of the sample median? [5]

e. Suppose the average yield of a chemical process being followed by a manufacturing company is 95%. A new method is suggested and it was claimed that the new method improves the yield. In order to test the claim, 16 batches of chemicals were produced and the average yield was found to be 97%. Assuming that the yields of batches follow a normal distribution with standard deviation 4%, would you agree that the new method is indeed more effective? Provide your argument by stating the hypotheses and computing the approximate p value. [5]

f. A company manufacturing detergent powder wishes to get an idea about the quality of their product as perceived by the users. They conduct a campaign on television and collect data on the responses given by users watching some programmes. Identify the study and the target populations and comment whether the estimated proportions are likely to be reliable or not. [5]

4. A manufacturing company has three machines A, B and C. Each of the machines produce the same product continuously in a manner such that the products are independent of each other within and across machine. It is known that the probability of a randomly chosen product being defective are 0.1, 0.2 and 0.3 for machines A, B and C respectively. Suppose a random sample of n products have been selected from some machine, i.e. the entire sample was drawn from the same machine. Let the sample be $X_1, X_2, \ldots X_n$, where $X_j = 0$ if the j^{th} product is defective and 1 otherwise. Thus each X_j, $j = 1, 2, ..n$ is a Bernoulli trial and the probability of defective item depend on the chosen machine. Let $Y = X_1 + X_2 + \ldots + X_n$.

   a. Suppose $n = 3$ and none of the products are defective. Plot the likelihood function and obtain the maximum likelihood estimate of the parameter. [5]

   b. Do you need to know the individual values of $X_1, X_2, \ldots X_n$ to obtain the maximum likelihood estimate of the of the parameters? If yes, why. If not, why not. [4]

   c. Suppose 50% of the products are produced by machine A, 30% by B and 20% by C. Suppose one product was collected at random from some machine and it was found to be non defective. Incorporate the information on production frequency and obtain an estimate of the parameter. [6]

   d. A coin is tossed 5 times. At each toss the experimenter observes whether it is a head or a tail. Simultaneously, a blindfolded person who claims to have extrasensory perception states whether the coin turns up heads or tails. The null hypothesis is that the person predicts with probability 0.5 while the alternative is that $p > 0.5$. It was decided to accept that the person has extrasensory power in case all predictions are correct.

      i. What is the critical region? [2]

      ii. What is α? [2]

      iii. Draw the graph of the power function. [6]

5. Answer the following

   a. Amateur mountain climbers in a region continuously debate about the level of difficulty of climbing two steep hills A and B. Some believe that climbing A is easier while some others believe that climbing B is easier. In order to test the claim data were collected on 200 climbers who tried to climb both hills. It was noted that 70 of them succeeded in climbing both, 40 failed in both, 60 succeeded in climbing A but failed in B.

      i. Put the data in the form of a 2 X 2 contingency table.

      ii. Do you think there is enough evidence that some hill is indeed more difficult to climb. You may assume that the skill of a climber remains same irrespective of the hill being climbed. [4 + 8 = 12]

   b. The shells fired during war are not supposed to explode even if they hit a target within 20 meters since an explosion so close to the gun may lead to injury of the soldier firing the gun. However, if the shell hits a target at a distance of over 20 meters, it should definitely explode. Draw the ideal curve depicting the probability of explosion of the shell against the distance of the target being hit. [3]

c. An IT company having 4 service departments hires over one thousand freshmen every year. The HR department conducts training for the hired persons and allocate them to the 4 service departments. Before allocation, an examination is conducted to test the ability of the students to write code. Students in lower 25^{th} percentile are considered to be weak, the ones in $25^{th} - 75^{th}$ percentile are considered average and students in the highest quartile (i.e. 75^{th} percentile onward) are considered strong.

   i. Suggest a possible format to present the data showing distribution of weak, average and strong candidates in the four service departments. [4]

   ii. Every service department claims that they get larger share of weak students while the HR department claims that the allocation is made randomly and is fair. How will you test the claim of the HR? Explain the method briefly and state the hypotheses clearly. [6]

INDIAN STATISTICAL INSTITUTE

First Semester Backpaper Examination: 2018-19

POST GRADUATE DIPLOMA IN BUSINESS ANALYTICS

Stochastic Processes and Applications (BAISI1)

Date: 26 December 2018 Maximum marks: 45 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. How many people have to be in a room in order that the probability that at least two of them celebrate their birthday in the same month is at least 0.5? Assume that all possible monthly outcomes are equally likely. [10]

2. A parallel system functions whenever at least one of its components works. Consider a parallel system of n components, and suppose that each component works independently with probability 0.5. Find the conditional probability that component 1 works given that the system is functioning. [7]

3. Each of 50 students in a class independently has a certain disease with probability 0.001. This disease will show up in a blood test, and to facilitate matters, blood samples from all 50 students are pooled and tested.

   (a) What is the (approximate) probability that the blood test will be positive (that is, at least one person has the disease)?

   (b) If the blood test yields a positive result, what is the probability that more than one person has the disease? [2+5=7]

4. Suppose that the travel time from your home to your office is normally distributed with mean 40 minutes and standard deviation 7 minutes. If you want to be 95 percent certain that you will not be late for an office appointment at 1 P.M., what is the latest time that you should leave home? If you take this strategy for three 1 P.M. appointments on three different days, what is the probability that you would be late for at least one of those appointments? [5+3=8]

5. Consider a Markov chain on the nonnegative integers such that, starting from x, the chain goes to state $x + 1$ with probability p, $0 < p < 1$, and goes to state 0 with probability $1 - p$.

   (a) Show that this chain is irreducible.

   (b) If T_0 is the waiting time in state 0, and P_0 indicates probability conditional on the current state being state 0, find $P_0(T_0 = n)$, $n > 1$.

   (c) Show that the chain is recurrent.

   (d) Calculate the stationary distribution of the chain. [2+4+4+3=13]

6. The number of claims received by an insurer is a Poisson process $X(t)$, $t \geq 0$ with parameter λ. The sizes of the successive claims $Y_1, Y_2, Y_3, \ldots$ are independent and have the exponential distribution with mean μ. Calculate the mean and the variance of the aggregate claims till time t, $S(t) = Y_1 + Y_2 + \cdots + Y_{X(t)}$. [5+4=9]

7. Consider the time series

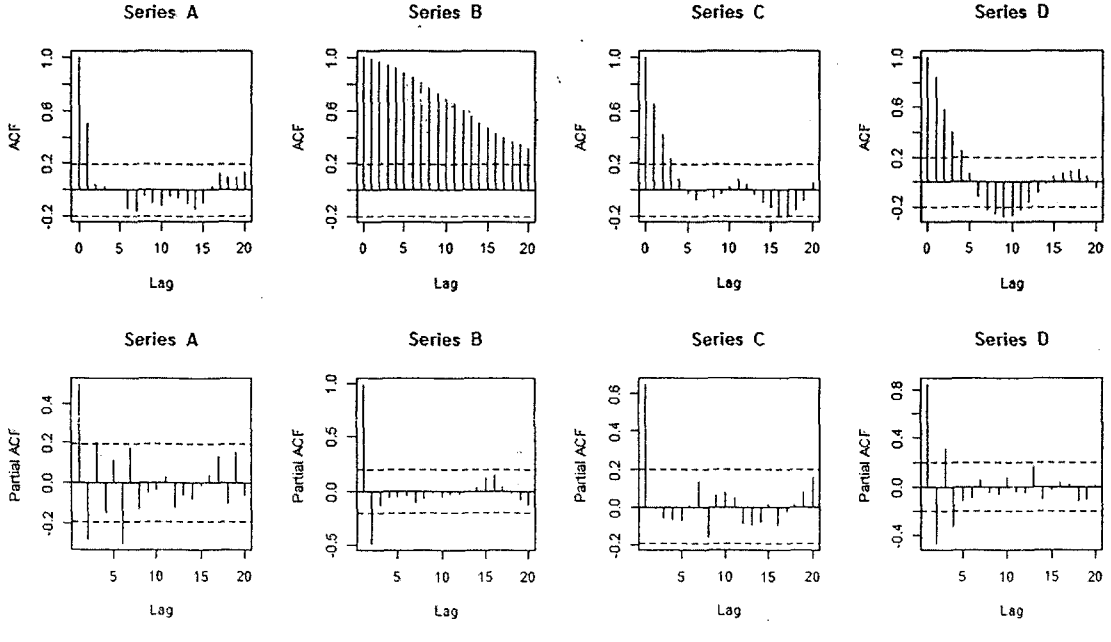
$$x_t = U\cos(2\pi\omega t) + V\sin(2\pi\omega t), \quad t = 1,2,3,\ldots,$$

where ω is a known constant and U and V are independent random variables with zero means and $E(U^2) = E(V^2) = \sigma^2$.

(a) Show that this series is weakly stationary and find its autocorrelation function.
(b) Find the best linear predictor of x_4 in terms of x_1 and x_2.
(c) Find mean squared prediction error of the predictor of part (b).
(d) Find the best linear predictor of x_4 in terms of x_1, x_2 and x_3 and compare its mean squared prediction error with that of part (c). Explain your findings.

[3+3+1+(4+1+1)=13]

8. The plots given below show sample ACF and partial ACF for four series (in scrambled order), which consist of 100 consecutive observations from an AR(1), an MA(1), an ARMA(1,1) and an ARIMA(1,1,1) process. The AR parameter is $\varphi = 0.6$, and the MA parameter is $\theta = 0.9$.



Identify which series (A, B, C and D) is likely to have been generated from which model, with detailed commentary on the anticipated and observed nature of each plot. [10]

9. (a) Consider the stationary process x_t defined by the equation

$$x_t = 0.8x_{t-1} - 0.15x_{t-2} + w_t - 0.3w_{t-1},$$

where w_t is 0 mean white noise with variance σ^2. Identify the order of this ARMA process. Is it causal? Is it invertible?

(b) If the process y_t is defined by the equation

$$y_t = y_{t-1} - 0.5y_{t-2} + x_t - x_{t-1},$$

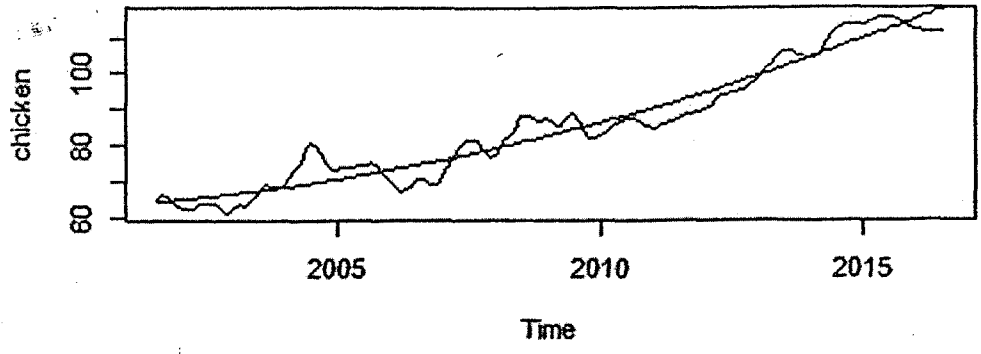
where x_t is as in part (a), identify the order of this ARMA process. Is it causal? Is it invertible? [(3+1+1)+(4+1+1)=11]

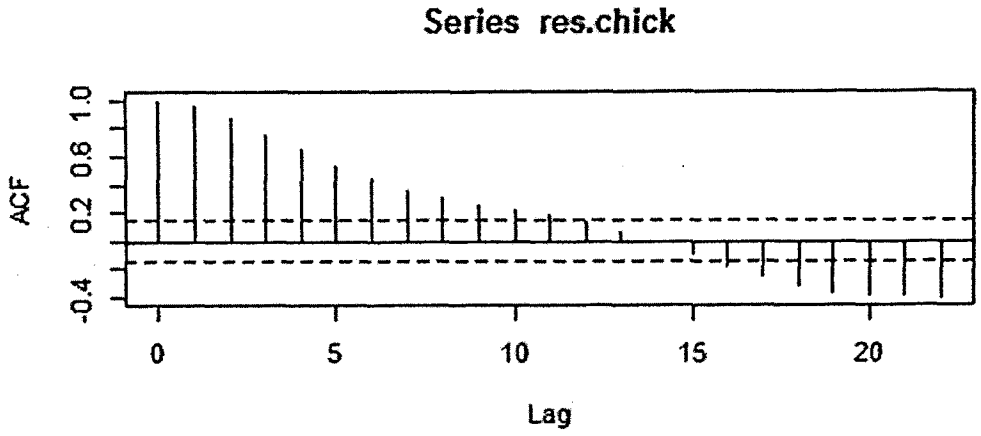
10. A data set on the average monthly price (per pound) of chicken in the US from mid-2001 to mid-2016 (180 months) is analysed through the following code.

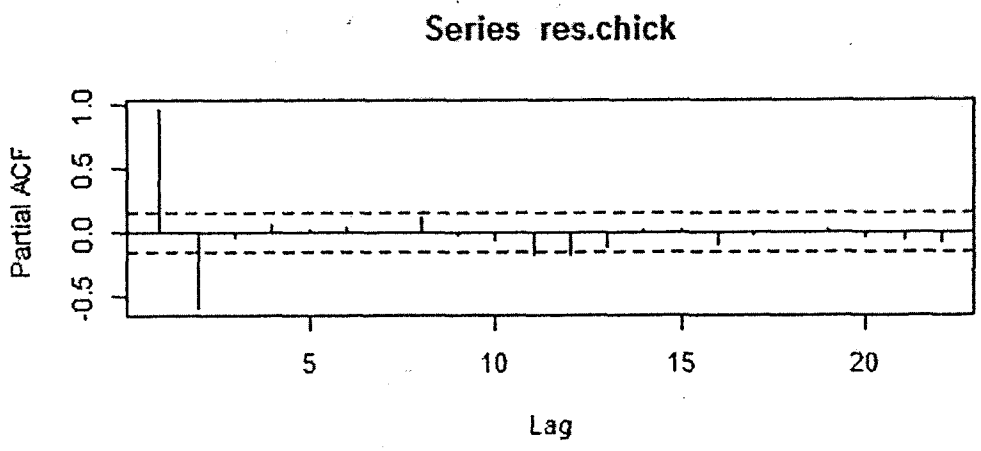
```
par(mfrow=c(3,1))
library(astsa)
# Chicken price data
ts.plot(chicken,col=4)
time1 <- time(chicken)
time2 <- time1^2
fit2 <- lm(chicken~time1+time2)
lines(as.vector(time1),predict(fit2,newdata = cbind(time1,time2)))
summary(fit2)
res.chick = resid(fit2)
acf(res.chick)
pacf(res.chick)
```

The resulting output and the plots are shown next.

```
Call:
lm(formula = chicken ~ time1 + time2)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6538 -3.1601 -0.2308  3.0222 11.2929

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.111e+05  7.056e+04   8.661 2.85e-15 ***
time1       -6.119e+02  7.025e+01  -8.711 2.09e-15 ***
time2        1.532e-01  1.748e-02   8.762 1.53e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.933 on 177 degrees of freedom
Multiple R-squared:  0.9423, Adjusted R-squared:  0.9417
F-statistic:  1446 on 2 and 177 DF,  p-value: < 2.2e-16
```







After studying the code, the output and the plots, write a summary of the apparent purpose of the analysis, the findings and the possible directions of improved analysis. [12]

INDIAN STATISTICAL INSTITUTE

First Semester Backpaper Examination: 2018-19

POST GRADUATE DIPLOMA IN BUSINESS ANALYTICS

Stochastic Processes and Applications (BAISI1)

Date: 26 December 2018 Maximum marks: 45 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. How many people have to be in a room in order that the probability that at least two of them celebrate their birthday in the same month is at least 0.5? Assume that all possible monthly outcomes are equally likely. [10]

2. A parallel system functions whenever at least one of its components works. Consider a parallel system of n components, and suppose that each component works independently with probability 0.5. Find the conditional probability that component 1 works given that the system is functioning. [7]

3. Each of 50 students in a class independently has a certain disease with probability 0.001. This disease will show up in a blood test, and to facilitate matters, blood samples from all 50 students are pooled and tested.

   (a) What is the (approximate) probability that the blood test will be positive (that is, at least one person has the disease)?

   (b) If the blood test yields a positive result, what is the probability that more than one person has the disease? [2+5=7]

4. Suppose that the travel time from your home to your office is normally distributed with mean 40 minutes and standard deviation 7 minutes. If you want to be 95 percent certain that you will not be late for an office appointment at 1 P.M., what is the latest time that you should leave home? If you take this strategy for three 1 P.M. appointments on three different days, what is the probability that you would be late for at least one of those appointments? [5+3=8]

5. Consider a Markov chain on the nonnegative integers such that, starting from x, the chain goes to state $x + 1$ with probability p, $0 < p < 1$, and goes to state 0 with probability $1 - p$.

   (a) Show that this chain is irreducible.

   (b) If T_0 is the waiting time in state 0, and P_0 indicates probability conditional on the current state being state 0, find $P_0(T_0 = n)$, $n > 1$.

   (c) Show that the chain is recurrent.

   (d) Calculate the stationary distribution of the chain. [2+4+4+3=13]

6. The number of claims received by an insurer is a Poisson process $X(t)$, $t \geq 0$ with parameter λ. The sizes of the successive claims $Y_1, Y_2, Y_3, \ldots$ are independent and have the exponential distribution with mean μ. Calculate the mean and the variance of the aggregate claims till time t, $S(t) = Y_1 + Y_2 + \cdots + Y_{X(t)}$. [5+4=9]

7. Consider the time series

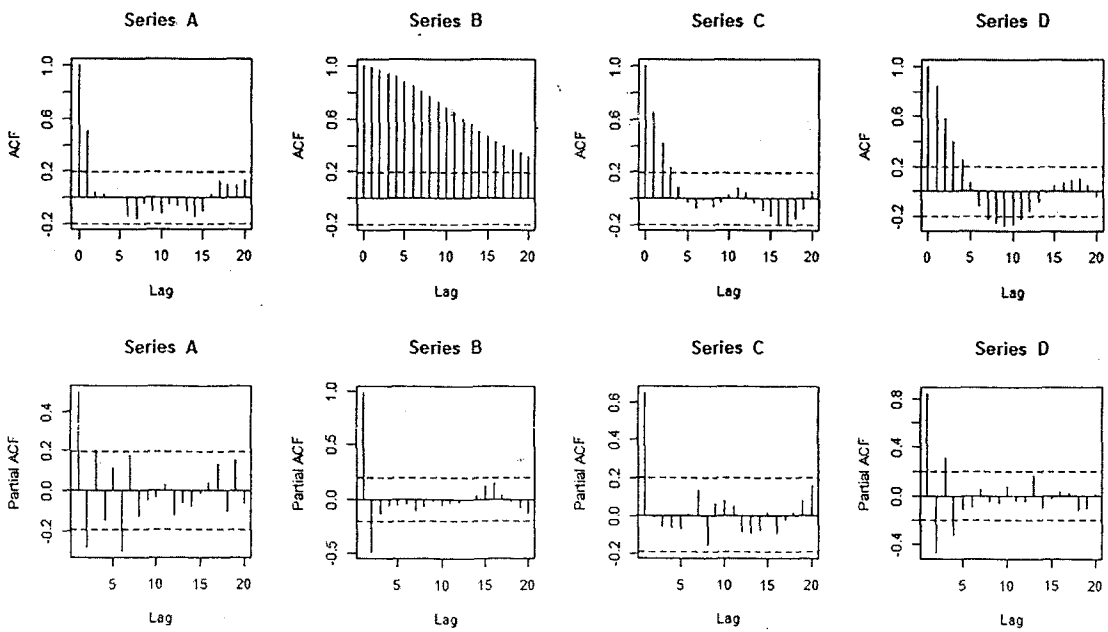
$$x_t = U\cos(2\pi\omega t) + V\sin(2\pi\omega t), \quad t = 1,2,3,\ldots,$$

where ω is a known constant and U and V are independent random variables with zero means and $E(U^2) = E(V^2) = \sigma^2$.

(a) Show that this series is weakly stationary and find its autocorrelation function.
(b) Find the best linear predictor of x_4 in terms of x_1 and x_2.
(c) Find mean squared prediction error of the predictor of part (b).
(d) Find the best linear predictor of x_4 in terms of x_1, x_2 and x_3 and compare its mean squared prediction error with that of part (c). Explain your findings.

[3+3+1+(4+1+1)=13]

8. The plots given below show sample ACF and partial ACF for four series (in scrambled order), which consist of 100 consecutive observations from an AR(1), an MA(1), an ARMA(1,1) and an ARIMA(1,1,1) process. The AR parameter is $\varphi = 0.6$, and the MA parameter is $\theta = 0.9$.



Identify which series (A, B, C and D) is likely to have been generated from which model, with detailed commentary on the anticipated and observed nature of each plot. [10]