PGDBA

(POST GRADUATE DIPLOMA IN BUSINESS ANALYTICS)

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination

Post Graduate Diploma in Business Analytics 2016-17 (Semester-I)

Stochastic Processes and Apllications

Date: **September 19, 2016** Maximum Marks: **80** 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. Ignoring leap days, the days of the year can be numbered 1 to 365. Assume that birthdays are equally likely to fall on any day of the year. Consider a group of n people, of which you are not a member. An element of the sample space Ω will be a sequence of n birthdays (one for each person).

   (a) Define the probability function P for Ω.

   (b) Consider the following events:
   A: someone in the group shares your birthday
   B: some two people in the group share a birthday
   C: some three people in the group share a birthday
   Carefully describe the subset of Ω that corresponds to each event.

   (c) Find an exact formula for $P(A)$. What is the smallest n such that $P(A) > 0.5$?

   (3+6+(4+3)=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. **(15)**

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. **(15)**

4. Recall that an exponential random variable $X \sim exp(\lambda)$ has mean $\frac{1}{\lambda}$ and pdf given by $f(x) = \lambda e^{-\lambda x}$ on $x \geq 0$.

   (a) Compute $P(X \geq x)$.

   (b) Suppose that $X1$ and $X2$ are independent exponential random variables with mean $\frac{1}{\lambda}$. Let $T = \min(X1, X2)$ Find the cdf of T.

   (c) Suppose we are testing 3 different brands of light bulbs B1, B2, and B3 whose lifetimes are exponential random variables with mean 1/2, 1/3, and 1/5 years, respectively. Assuming that all of the bulbs are independent, what is the expected time before one of the bulb fails.

   (3+5+8=16)

5. Let $N_1(t)$ and $N_2(t)$ be two independent Poisson processes with rates λ_1 and λ_2 respectively. Define $N(t) = N_1(t) + N_2(t)$. Show that $N(t)$ is a Poisson process with rate $\lambda = \lambda_1 + \lambda_2$. **(12)**

6. Define a nonhomogeneous Poisson Process. Show that for a nonhomogeneous Poisson process with rate $\lambda(t)$, the number of arrivals in any interval is a Poisson random variable with appropriate parameter. **(4+12=16)**

INDIAN STATISTICAL INSTITUTE

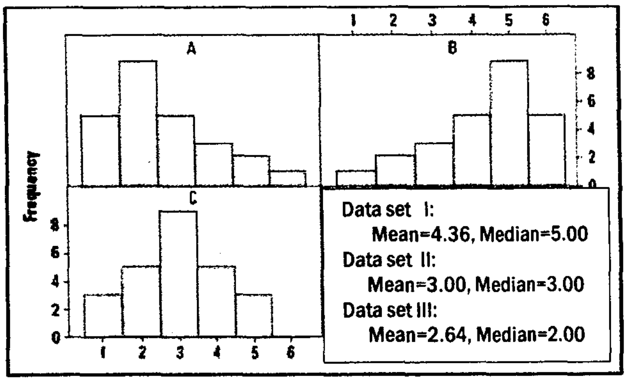
POST-GRADUATE DIPLOMA IN BUSINESS ANALYTICS
Mid-Semester Examination: 2016–17 (Semester I)

Course: Stat2: Statistical Structures in Data

Date: September 20, 2016 Maximum Marks: 50 Duration: 2 hr

Note: The maximum you can score is 50.

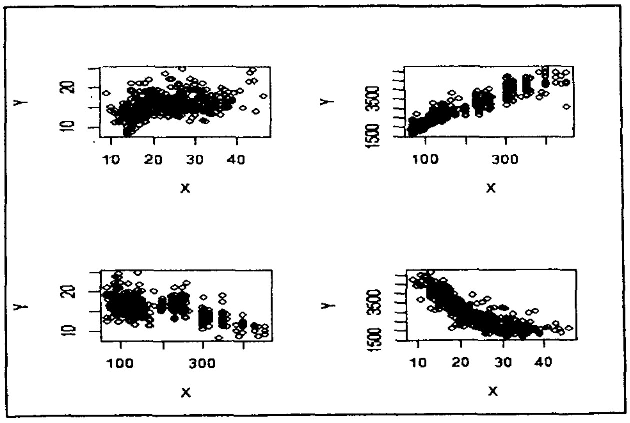
1. Given below are three histograms, labelled A, B and C, together with some descriptive measures corresponding to three discrete univariate data sets I, II and III.



Data set I:
Mean=4.36, Median=5.00
Data set II:
Mean=3.00, Median=3.00
Data set III:
Mean=2.64, Median=2.00

Match each histogram with the data set it represents best, giving reasons. [5]

2. Consider the following scatterplots:



Match each with exactly one of the following correlation coefficients:
0.4, -0.8, -0.5, 0.9.

[5]

3. State, with explanation, whether the following statements, in respect of a dataset containing 50 observations, are *true* or *false*:

   a. Increasing each observation by 9 increases the mean by 450.
   b. Reducing each observation by half reduces the mean by half.
   c. Multiplying each observation by 5 increases the variance 5 times.
   d. Changing the sign of each observation does not change the standard deviation.
   e. Adding 100 to each observation increases the standard deviation by 100.

[5]

4. An investigator, who has collected data on monthly incomes from 500 families in Kolkata, suddenly notices that the highest observation in the dataset has been wrongly noted as Rs. 50,00,000 instead of Rs. 50,0000The remaining observations range from Rs. 5,000 to 1,00,000. In what way will this mistake affect the mean, the mode and the median, and by how much? [5]

5. A company manufactures rubber balls having certain specifications. From past experience, it has been found that about 5% of the balls are defective, that is, do not meet the specifications. If 20 balls are selected at random, what is the probability that
   a. none of the balls is defective?
   b. not more than three balls are defective?

[1+4=5]

6. A continuous random variable with probability density function

$$f(x) = \begin{cases} \dfrac{\alpha x_0^{\alpha}}{x^{\alpha+1}}, & \text{if } x \geqslant x_0, \\ 0 & \text{otherwise,} \end{cases}$$

is said to have a Pareto distribution with parameters $\alpha\ (> 0)$ and x_0. If $\alpha > 2$, find the cumulative distribution function, the mean and the mode of the distribution.

[2+2+1=5]

7. The joint probability density function of two random variables, X and Y, is given by

$$f(x,y) = \frac{1}{\pi\sqrt{3}} e^{-2\left(x^2 - xy + \frac{y^3}{3}\right)}, \quad -\infty < x, y < \infty.$$

Deduce the regression of Y on X. [5]

8. For a pair of random variables x and y, the regression lines are

$$3x + 2y = 25 \text{ and } 6x + y = 30.$$

   a) Which of these is the regression line of x on y? Justify your answer.
   b) What will be the equation of the SD line?

[3+2=5]

9. The scores of 10,000 applicants taking a nation-wide selection test for admission to a certain university are observed to have mean 54 and standard deviation 10. If the cut-off score for selection is 75, estimate the number of applicants selected, making appropriate assumptions.

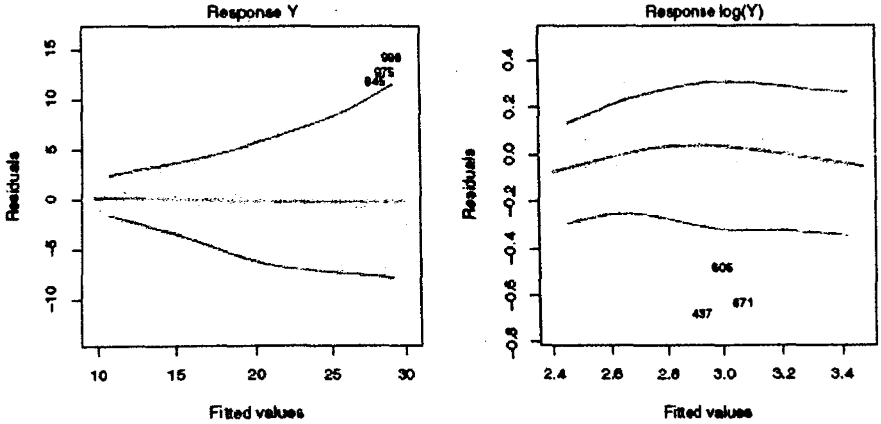
[5]

10. The least-squares regression line of y on x and of x on y are respectively, $x + 3y = 0$ and $3x + 2y = 0$. The sample standard deviation, s_x, of x is 1. Define new random variables $u = x + y$ and $v = x - y$.

   a. Show that $\text{cov}(u, v) = s_x^2 - s_y^2$, where s_y is the sample standard deviation of y.
   b. Determine the least-squares regression line of v on u.

[2+3=5]

11. What do each of the following residual plots, each obtained after fitting a least-squares regression line to a bivariate dataset, indicate?



Indian Statistical Institute
Foundations of Database Systems
Mid Sem Exam

PGDBA 2016-18 1st Semester

Total marks: 40
Date: 21 September 2016
Time: 2 hours 30 minutes

Instruction: Each question carries 5 marks. For every question asking you to write an SQL query, you must explain the logic behind it, that is, why your query should work.

1. Refer to the *institute* schema in Appendix B. For each department, which are the classrooms present in the same building as the department? Write an SQL query to list all the classrooms numbers, with building names, in the same building as the department, for each department.

2. From the schema *institute* in Appendix B, identify at least three entities and two relations. Explain how they are connected either in words or by an entity-relationship (ER) diagram.

3. Write an SQL query to determine the highest number of times any employee has switched his or her department in the *employees* database. Refer to Appendix A for details of the schema.

4. Consider the following SQL statement to be executed on a database based on the *institute* schema. Explain what would be the output, or effect of the statement on the database.

```
delete from dept_name where budget < 10000;
```

5. Given a student's id (student.ID) x, write an SQL query to output all the instructors who have taught the student. Refer to Appendix B for details of the schema.

6. Write an SQL query on the *employees* database to output the employee id and full name of all the employees who have never been a manager.

7. Let R and S are two relations with attributes $R(A, B, C, D)$ and $S(A, E, F, G)$. Suppose the following are given:

   (a) (A) is the primary key in R.

   (b) (A, E) is the primary key in S.

   (c) In S, (A) is a foreign key with reference to $(R.A)$.

   Then prove that the number of records in S is greater than or equal to the number of records in R.

8. The schema *employees* represent an organization where every department has a manager and consequently, all the employees in that department have a single manager. Suppose the organizational structure was different such that:

   - Every department has one manager.
   - Under the manager of the department, there can be other non-manager employees and managers of non-manager employees (in other words, a department manager could have a tree of depth at most two under him or her).

   How would you change the schema (by dropping, altering or adding some tables) to represent the above mentioned scenario?

A SQL Statements to create schema *employees*

```
CREATE TABLE employees (
    emp_no      INT            NOT NULL,
    birth_date  DATE           NOT NULL,
    first_name  VARCHAR(14)    NOT NULL,
    last_name   VARCHAR(16)    NOT NULL,
    gender      ENUM ('M','F') NOT NULL,
    hire_date   DATE           NOT NULL,
    PRIMARY KEY (emp_no)
);

CREATE TABLE departments (
    dept_no     CHAR(4)        NOT NULL,
    dept_name   VARCHAR(40)    NOT NULL,
    PRIMARY KEY (dept_no),
    UNIQUE  KEY (dept_name)
);

CREATE TABLE dept_manager (
   dept_no      CHAR(4)        NOT NULL,
   emp_no       INT            NOT NULL,
   from_date    DATE           NOT NULL,
   to_date      DATE           NOT NULL,
   KEY         (emp_no),
   KEY         (dept_no),
   FOREIGN KEY (emp_no)  REFERENCES employees (emp_no)    ON DELETE CASCADE,
   FOREIGN KEY (dept_no) REFERENCES departments (dept_no) ON DELETE CASCADE,
   PRIMARY KEY (emp_no,dept_no)
);

CREATE TABLE dept_emp (
    emp_no      INT            NOT NULL,
    dept_no     CHAR(4)        NOT NULL,
    from_date   DATE           NOT NULL,
    to_date     DATE           NOT NULL,
    KEY         (emp_no),
    KEY         (dept_no),
    FOREIGN KEY (emp_no)  REFERENCES employees   (emp_no)  ON DELETE CASCADE,
    FOREIGN KEY (dept_no) REFERENCES departments (dept_no) ON DELETE CASCADE,
    PRIMARY KEY (emp_no,dept_no)
);
```

```
CREATE TABLE titles (
    emp_no      INT             NOT NULL,
    title       VARCHAR(50)     NOT NULL,
    from_date   DATE            NOT NULL,
    to_date     DATE,
    KEY         (emp_no),
    FOREIGN KEY (emp_no) REFERENCES employees (emp_no) ON DELETE CASCADE,
    PRIMARY KEY (emp_no,title, from_date)
);

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

B SQL Statements to create schema *institute*

```
create table classroom (
    building        varchar(15),
    room_number     varchar(7),
    capacity        numeric(4,0),
    primary key (building, room_number)
);

create table department (
    dept_name       varchar(20),
    building        varchar(15),
    budget          numeric(12,2) check (budget > 0),
    primary key (dept_name)
);

create table course (
    course_id       varchar(8),
    title           varchar(50),
    dept_name       varchar(20),
    credits         numeric(2,0) check (credits > 0),
    primary key (course_id),
    foreign key (dept_name) references department on delete set null
);

create table instructor (
    ID              varchar(5),
    name            varchar(20) not null,
    dept_name       varchar(20),
    salary          numeric(8,2) check (salary > 29000),
    primary key (ID),
    foreign key (dept_name) references department on delete set null
);

create table student (
    ID              varchar(5),
    name            varchar(20) not null,
    dept_name       varchar(20),
    tot_cred        numeric(3,0) check (tot_cred >= 0),
    primary key (ID),
    foreign key (dept_name) references department on delete set null
);
```

```
create table section (
    course_id        varchar(8),
    sec_id           varchar(8),
    semester         varchar(6)
            check (semester in ('Fall', 'Winter', 'Spring', 'Summer')),
    year             numeric(4,0)
            check (year > 1701 and year < 2100),
    building         varchar(15),
    room_number      varchar(7),
    time_slot_id     varchar(4),
    primary key (course_id, sec_id, semester, year),
    foreign key (course_id) references course on delete cascade,
    foreign key (building, room_number) references classroom
        on delete set null
);

create table teaches (
    ID               varchar(5),
    course_id        varchar(8),
    sec_id           varchar(8),
    semester         varchar(6),
    year             numeric(4,0),
    primary key (ID, course_id, sec_id, semester, year),
    foreign key (course_id,sec_id, semester, year) references section
                on delete cascade,
    foreign key (ID) references instructor
                on delete cascade
);

create table takes (
    ID               varchar(5),
    course_id        varchar(8),
    sec_id           varchar(8),
    semester         varchar(6),
    year             numeric(4,0),
    grade            varchar(2),
    primary key (ID, course_id, sec_id, semester, year),
    foreign key (course_id,sec_id, semester, year) references section
                on delete cascade,
    foreign key (ID) references student
                on delete cascade
);
```

```
create table advisor (
    s_ID            varchar(5),
    i_ID            varchar(5),
    primary key (s_ID),
    foreign key (i_ID) references instructor (ID)
                on delete set null,
    foreign key (s_ID) references student (ID)
                on delete cascade
);

create table time_slot (
    time_slot_id    varchar(4),
    day             varchar(1),
    start_hr        numeric(2) check (start_hr >= 0 and start_hr < 24),
    start_min       numeric(2) check (start_min >= 0 and start_min < 60),
    end_hr          numeric(2) check (end_hr >= 0 and end_hr < 24),
    end_min         numeric(2) check (end_min >= 0 and end_min < 60),
    primary key (time_slot_id, day, start_hr, start_min)
);

create table prereq (
    course_id       varchar(8),
    prereq_id       varchar(8),
    primary key (course_id, prereq_id),
    foreign key (course_id) references course
                on delete cascade,
    foreign key (prereq_id) references course
);
```

INDIAN STATISTICAL INSTITUTE

Mid – Semester Examination: 2016 - 17

Course Name: PGDBA

Subject Name: Inference

Date: 23 September 2016 Maximum Marks: 100 Duration: 3 hours

Notes, if any: The paper carries 110 marks. Answer as much as you can. However, the maximum you can get is 100.

1. What is the random variable and the parameter to be compared / estimated in the following business problems (i.e. in each case specify the random variable to be studied and the parameter to be estimated for the purpose of decision making / comparison). State your assumptions, if any, clearly.
    a. An aviation expert wants to compare the performance of two different airlines with respect to on time arrival.
    b. An automobile engineer wants to find whether the fuel efficiency (measured in kilometres travelled per litre of fuel) of a particular brand of car has increased after making a modification.
    c. The head of traffic police of a particular city wants to assess whether safety measures adopted by the traffic police have led to reduction of number of major accidents. Assume that "major" accidents are well defined.
    d. In a particular company the time required to process invoices is seen to have a lot of variation. A consultant has suggested some changes in the process and claims that the change would lead to reduction of variation. You have past data on time to process invoices. You need to verify the claim of the consultant. [5 X 4 = 20]
2. Answer the following
    a. 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). Do you think the samples qualify as random samples? If not, why not? [5]
    b. What are retrospective and prospective cohort studies? Give one example of each of the studies. [6 + 6 = 12]
    c. What is the definition of measurement? What are the different scales of measurement? Give definition only, examples are not required. [2 + 4 = 6]
    d. Identify the scales of measurement of the following
        i. Calendar date
        ii. Temperature measured in Fahrenheit scale
        iii. Weight of an individual
        iv. City of birth of a person
        v. Level of satisfaction of a customer of a fast food restaurant [5 X 1 = 5]
    e. Explain the concepts of explanatory and response variables with one example each. [5]
    f. In the following situations, identify the response and the response variables
        i. Economic status and attitude towards political parties
        ii. Age and occurrence heart disease
        iii. Fuel efficiency (measured in terms of average distance travelled per litre of fuel) and weight of a car
        iv. Hours of exercise and average loss of body weight per month. [2 X 4 = 8]

P.T.O

g. In business analytics we often come across different types of "missing data". Explain briefly the three different types of missingness of data with one example each. [3 X 3 = 9]

3. We are aware that four major themes of EDA are resistance, residuals, re-expression, and revelation.
   a. Explain briefly the meaning of "resistance"? [4]
   b. Suppose we want to summarize the dispersion observed in a sample. We may use standard deviation, inter quartile range, and range. Which of three statistics has the maximum resistance and which one has the least. Explain briefly. [3]
   c. Suppose we want to re-express the collected data so that it has reasonably constant variance. For example, we might have collected data on crime rates in large cities across countries. We want to see whether the dispersion of crime rate across cities in a country is dependent on the median crime rate of that country. How would you use the spread versus level plot in this situation? [5]
   d. Suppose an analyst wants to compare the performance of two retail stores with respect to daily sales. How can box plot be used for this purpose? You must state the data that needs to be collected and the statistics need to be computed for this purpose. [3]
   e. What is standard error (SE)? How is the standard error of a maximum likelihood estimator estimated? [2 + 3 = 5]

4. Answer the following
   a. What is an unbiased estimator? Give an example. [3]
   b. Suppose we have used $s^2 = \Sigma(x_i - \text{x-bar})^2/n$ – where x-bar is the usual arithmetic mean and the summation is over all values of i (i.e. i= 1, 2...n) as an estimator of variance. Do you think it is an unbiased estimator? Explain. [5]
   c. What is a consistent estimator? Give an example. [3]
   d. Suppose you are trying to estimate the parameter p for a geometric distribution. In this context
      i. Write the likelihood function [4]
      ii. Suppose you have tossed a coin and counted the number of tails before the first head appears. Suppose you have carried out the experiment twice and the observations are 3, and 0. Plot the likelihood function. [5]

INDIAN STATISTICAL INSTITUTE

Post-Graduate Diploma in Business Analytics

Semestral Examination: 2016–17(First Semester)

STATISTICAL STRUCTURES IN DATA (BAISI2)

Date: November 28, 2016 Maximum Marks: 100 Duration: 3 hr

1. The Dirichlet distribution of order $k \geq 2$ and parameters $a_1, a_2, \ldots, a_k > 0$ has the probability density function
$$f(\boldsymbol{x}|\boldsymbol{a}) = \frac{1}{B(\boldsymbol{a})} \prod_{i=1}^{k} x_i^{a_i},$$
where $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)'$, $\boldsymbol{a} = (a_1, a_2, \ldots, a_k)'$, $0 < x_i < 1$, $x_1 + x_2 + \ldots + x_k = 1$ and
$$B(\boldsymbol{a}) = \frac{\prod_{i=1}^{k} \Gamma(a_i)}{\Gamma(\sum_{i=1}^{k} a_i)}.$$
   (a) Set $k = 3$ and show that the marginal distribution of (X_2, X_3) has a Dirichlet distribution with appropriate parameters.
   (b) Hence determine the regression of X_1 on (X_2, X_3).
   (c) Write down the general definition of the multiple correlation coefficient of X_1 on the predictor variables $X_2, X_3, \ldots, X_k$, and hence deduce the multiple correlation coefficient of X_1 on (X_2, X_3) if X_1, X_2 and X_3 have a joint Dirichlet distribution.

[6+8+6=20]

2. Consider a data set consisting of daily readings of the following air quality values for a number of days:
   (a) `Ozone`: Mean ozone in parts per billion from 1300 to 1500 hours
   (b) `Solar.R`: Solar radiation in Langleys in the frequency band 40007700 Angstroms
   (c) `Wind`: Average wind speed in miles per hour at 0700 and 1000 hours
   (d) `Temp`: Maximum daily temperature in degrees Fahrenheit

The following summary is obtained after using the R function `lm` to regress `Solar.R` on all the other variables and interactions:

	Estimate	Std. Error	t value	Pr(> $\lvert t \rvert$)
(Intercept)	332.526452	380.603247	0.874	0.3843
`Ozone`	-7.145736	6.956791	-1.027	0.3068
`Wind`	-38.633851	30.432875	-1.269	0.2071
`Temp`	-2.702812	5.058902	-0.534	0.5943
`Ozone:Wind`	1.649706	0.778109	2.120	0.0364
`Ozone:Temp`	0.091476	0.083880	1.091	0.2780
`Wind:Temp`	0.497654	0.412634	1.206	0.2306
`Ozone:Wind:Temp`	-0.018470	0.009246	-1.998	0.0484

Residual standard error: 19.2 on 103 degrees of freedom
Multiple R^2: 0.6883
F-statistic: 32.49 on 7 and 103 DF, p-value: $< 2.2 \times 10^{-16}$

P.T O

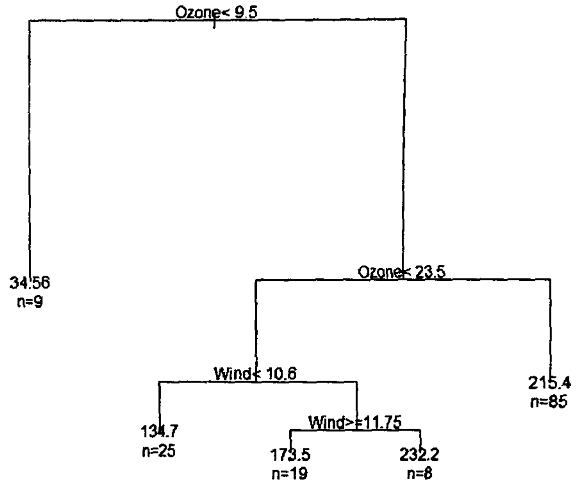
(a) Write down the complete model that has been assumed, specifying the assumptions made.

(b) Specify clearly the hypotheses that have been tested through the t-tests to which the last two columns correspond. Write down clearly the conclusion that you can make from these tests.

(c) What can you say about the validity of the model that you started with, on the basis of the information given above? Provide adequate justification.

[6+(4+4)+6=20]

3. (a) For the dataset in Problem 2, consider the unpruned regression tree, given below, with `Solar.R` as the response variable and the remaining three variables as predictors.



i. List the ways in which the regression tree model is different from the least-squares regression model in Problem 2, highlighting the issue of variable selection.

ii. Given that $1 \le$ `Ozone` ≤ 200, $0 \le$ `Wind` ≤ 25 and $50 \le$ `Temp` ≤ 100, sketch the partition of the predictor space generated by the tree, labeling the sub-regions by the value of the response variable within it.

(b) Describe any adaptive regression technique known to you, stating clearly the rationale behind it and the main steps involved in implementing it.

[(6+6)+8=20]

4. Let $\boldsymbol{X} = (X_1, X_2, X_3)'$ be a trivariate random vector with mean vector 0 and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & 0 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ 0 & \rho\sigma^2 & \sigma^2 \end{bmatrix},$$

where $\rho \in (-1/\sqrt{2}, 1/\sqrt{2})$.

(a) Find the principal components of $\boldsymbol{X}$.

(b) If $\rho = 0.5$ and $\sigma^2 = 2$, determine the proportion of the total population variance that is explained by

i. the first principal component,

P.T.O

ii. the first two principal components.

(c) Will the principal components remain the same if they are determined from the correlation matrix

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{bmatrix}$$

rather than from $\boldsymbol{\Sigma}$? Justify your answer.

[12+4+4=20]

5. Let X_1, X_2, X_3, X_4 and X_5 denote respectively the weekly rates of return of stocks (over a fixed period) of five companies A, B, C, D and E, listed on the New York Stock Exchange. Data on $n = 100$ weekly rates of return was used to obtain the sample correlation matrix

$$\boldsymbol{R} = \begin{bmatrix} 1 & .577 & .509 & .387 & .462 \\ & 1 & .599 & .389 & .322 \\ & & 1 & .436 & .426 \\ & & & 1 & .523 \\ & & & & 1 \end{bmatrix}$$

Based on this, the estimates of factor loadings for a 2-factor model were computed by the maximum likelihood method. These are given in the following table, together with the estimated varimax-rotated factor loadings:

Variable	MLEs of factor loadings		MLEs of rotated factor loadings	
	F_1	F_2	F_1^*	F_2^*
X_1	0.684	0.189	0.601	0.377
X_2	0.694	0.517	0.850	0.164
X_3	0.681	0.248	0.643	0.335
X_4	0.621	-0.073	0.365	0.507
X_5	0.792	-0.442	0.208	0.883

(a) Using the estimated unrotated factor loadings, obtain estimates of the specific variances and communalities, explaining briefly the underlying theory.

(b) In the unrotated factor model, suggest an interpretation of the two factors, given that the companies A, B and C are chemical-manufacturing companies, while D and E are petroleum companies.

(c) In what manner does the interpretation of the factors change after rotation? Explain.

[12+4+4+=20]

INDIAN STATISTICAL INSTITUTE

FIRST SEMESTER EXAMINATION: 2016 – 17

Course Name: PGDBA

Subject Name: Inference (BAISI3)

Date: 29.11.2016 **Maximum Marks: 100** **Duration: 3 hours**

Notes, if any: Answer questions 7 and 8 and any 3 from the rest

1. Answer the following
   a. What are the four different scales of measurement? Give definition only, examples are not required. [4]
   b. Identify the type of variable in the given situations
      - Anxiety rating – none, mild, moderate, severe, very severe
      - Patient survival in number of months
      - Fuel efficiency of an automobile measured as the number of kilometres travelled per litre of fuel
      - Favourite beverage – water, juice, milk, soft drink, wine [4 X 1 = 4]
   c. Explain briefly the meaning of prospective and retrospective cohort studies with examples. [4 + 4 = 8]
   d. Explain briefly the concepts of odds ratio and relative risks using a 2 X 2 table. [2 + 2 = 4]
2. Each of 100 multiple choice questions in an examination has four possible answers, one of which is correct. Suppose, a student guesses the answers by selecting one of the alternatives randomly. Suppose further that the student has guessed all the answers.
   a. Specify the distribution of (n_1, n_2, n_3, n_4) where n_j is the number of times the student picked the the answer j, j = 1, 2, 3, 4. [2]
   b. What will be the mean and variance of n_j, j = 1, 2, 3, 4? [6]
   c. Let X be the number of correct answers given by the student.
      - What distribution will X follow? Specify the parameters of the distribution. [2]
      - Find the mean and standard deviation of X. [2]
      - Would it be surprising if the student scores 50 or more marks, assuming that each question has one mark without any negative marking? Explain. [8]
3. Answer the following
   a. What are type I and type II errors? [3 + 3 = 6]
   b. What are simple and composite hypotheses? [2 + 2 = 4]
   c. Explain briefly the concept of p-value. [4]
   d. What are the null and alternative hypotheses in the following cases
      - The average efficiency of a process is known to be 95%. A technologist claims to have discovered a method that would increase the efficiency. [3]
      - A business analyst claims that the conditional expectation of daily sale in a retail shop is a linear function of three explanatory variables X_1, X_2, and X_3. [3]
4. You want to compare the performance of three different gels with respect to their capability to provide relief from pain. The outcome may be "completely cured", "partially cured", or "not cured", i.e. at three different levels. Suppose you have looked at N persons randomly who had the disease and had used one of the gels. You have noted the gel consumed and the outcome.
   a. Explain briefly how you will present the data in the form of a table. What is this table called? [4]

P.T.O

b. Suppose you want to check whether the gels are equally effective or not. Explain how this may be checcked. You must use the structure you used to present the data. You need to explain clearly the steps to be carried out. [6]

c. Suppose the three gels are A, B and C respectively. Suppose you want to compare the performance of these three gels with respect to their ability to completely cure pain. Explain briefly how the performance may be presented graphically. [5]

d. Suppose the gels may be used to provide relief in two different types of sprains. Suppose the data collected shows that gel A is more effective than gel B in curing both types of sprains. Does it mean that, when the data are pooled, gel A would definitely turn out to be more effective than gel B? Explain briefly. [5]

5. Answer the following

a. Suppose a software product has k defects, where k is unknown. A tester tests the product and finds y defects ($y \le k$). Notice that the number of defects detected is a random variable.

- What could be the distribution of this random variable and what would be its parameters? [2]
- Use the method of moments to estimate the parameters. [6]
- Do you foresee any difficulty in this method of estimation? [2]

b. The company XYZ Ltd. manufactures paints for automobiles and sells the paints to car manufacturer ABC Ltd. XYZ Ltd. needs to give a guarantee about the average drying time of its paints. Thus XYZ Ltd. needs to specify that the average drying time would not exceed some u minutes. Unfortunately, neither the average drying time nor its variability are known to XYZ Ltd. However, it is known that the drying time follows a normal distribution. XYZ Ltd. had collected n observations and had computed sample mean and standard deviation as t and s respectively. Explain how they would specify the drying time. [3]

c. A textile company manufactures fabric. They have observed that the number of blemishes on a square meter of cloth being manufactured has a large variability. Let X be the random variable giving the number of defects on a cloth of area 1 square meter.

- What distribution is X likely to follow? What parameter will it have? [2]
- Suppose you have collected data on n square meters of cloth randomly and have observed x_1, x_2, ...x_n defects respectively. Write the likelihood function for the parameter of X. [3]
- Obtain the maximum likelihood estimator and its variance. [2]

6. Answer the following

a. Explain the concepts of Wald, Score and Likelihood Ratio tests with example. [3 X 5 = 15]

b. What assumptions are required to estimate relative risks? Explain briefly. [3]

c. When can odds ratio be considered to be approximately equal to relative risk? [2]

7. Indicate whether the following are true or false

a. A 100(1 − α)% confidence interval for the parameter β is the set of β_0 for which the test H_0: $\beta = \beta_0$ has a p value exceeding 100α%

b. Relative risk takes any value between 0 and 1.

c. The standard error of an unbiased estimator is smaller than the standard error of a biased estimator

d. The odds ratio can be estimated from the data collected through a case control study

e. The mean function plot is used to understand the form of relationship between two variables

P.T.O

f. The box plot gives a three point summary of a random variable

g. The power function is used to minimize the chance of rejecting a null hypothesis when it is true

h. The higher the curvature of the likelihood function, the higher the standard error of the corresponding maximum likelihood estimator

i. We use the F test to test the equality of two variances

j. Non-sampling errors generally pose a bigger problem than sampling errors [10 X 2 = 20]

8. Answer the following

a. Suppose you have collected data on two variables x and y as given below. Draw a mean function plot of x vs. y and comment about the relationship between x and y. Do you think that x and y are linearly related? [6]

X	Y
10.0	9.14
8.0	8.14
13.0	8.74
9.0	8.77
11.0	9.26
14.0	8.10
6.0	6.13
4.0	3.10
12.0	9.13
7.0	7.26
5.0	4.74

b. A group of 305 people went for an outing. After eating food some people fell ill. It was noted that people ate salad and / or crabmeat and this could be the reason behind their falling ill. The data on consumption of different food are as follows

- 120 people who ate both crabmeat and salad fell ill
- 4 people who ate crabmeat but did not eat salad fell ill
- 80 people who ate both crabmeat and salad did not fall ill
- 31 people who ate crabmeat but did not eat salad did not fall ill
- 22 people who did not eat crabmeat but ate salad fell ill
- 1 person who ate neither salad nor crabmeat fell ill
- 24 people who ate salad but did not eat crab meat did not fall ill
- 23 people who ate neither salad nor crab meat did not fall ill

i. Present the data in a suitable tabular manner.

ii. How many people fell ill and how many did not fall ill?

iii. Find the relative risk of falling ill after consuming crabmeat.

iv. Do you think salad is an effect modifier? [6 + 1 + 3 + 4 = 14]

Indian Statistical Institute
Foundations of Database Systems
End Sem Exam

PGDBA 2016-18 1st Semester

Total marks: 50
Date: 30 November 2016
Time: 3 Hours

For every answer you must explain the logic behind it.

1. Refer to Appendix A which shows the creation of the schema *institute*. For each of the SQL statements below, explain what would be the output or effect on the tables of the schema *institute*.

   (a) `insert into dept_emp values (1020,'HIST','2015-01-01','2016-01-01');`

   (b)
   ```
   select emp_sal.emp_no, first_name, last_name, max_sal from employees,
   ((select emp_no, max(salary) as max_sal from salaries group by emp_no) as emp_sal)
   where emp_sal.emp_no = employees.emp_no and emp_sal.max_sal < 50000
   ```

   4 × 2 = 8 marks

2. Refer to Appendix A which shows the creation of the schema *institute*. Write SQL queries which would answer the following questions.

   (a) What is the distribution (as `(gender, number of employees)`) of male and female employees who have ever received a salary greater than 80000?

   (b) Find the maximum salary received by all employees who have never become a manager.

   4 × 2 = 8 marks

3. Suppose there is a file D, containing a very large number of integers with very few duplicates, in a Hadoop Distributed File System (HDFS). In other words, all the integers cannot be stored in one node, and the number of distinct integers is close to the total number of integers N. Describe how would you compute a histogram of the data using MapReduce. The histogram should show the distribution of the data for k equal sized intervals and there is no prior information about the smallest and largest integer.

   6 marks

4. Examine the relation R shown in Table 1. Explain if the relation is susceptible to *insertion*, *deletion* and *update* anomalies. For each of these anomalies, either explain why the relation is not susceptible to it, or give an example of such an anomaly.

   2 × 3 = 6 marks

Staff ID	Staff Name	Client ID	Client Name	Vehicle Number	Appointment Date
15	V. Kumar	495	A. Banerjee	WB 24K 3382	5 Oct 2016
15	V. Kumar	1342	R. Jain	WB 06F 2878	10 Oct 2016
15	V. Kumar	1201	P. Ghosh	WB 02AH 2239	9 Nov 2016
8	A. Yadav	781	D. Srivastava	WB 02AA 3892	11 Nov 2016
12	S. Banerjee	1891	D. Sinha	WB 06K 1093	24 Nov 2016
15	V. Kumar	495	A. Banerjee	WB 24K 3382	1 Dec 2016
12	S. Banerjee	573	K. Saha	WB 06C 7728	2 Dec 2016
8	A. Yadav	573	K. Saha	WB 06J 1002	12 Dec 2016

Table 1: A relation showing the vehicle service appointment data for a service center.

5. Find a *canonical cover* F_c for the set F of all functional dependencies that hold for the relation R as shown in Table 1.

6 marks

6. Decompose the relation R as shown in Table 1 to achieve a normal form (2-NF, 3-NF or BCNF), the best as you can. State which normal form you achieved and prove your claim.

6 marks

7. Suppose an inverted index with positions is created from a collection C of text documents. The index contains, for each word w present in the collection, the list of documents in which the w is present, a score for each of those documents and the list of positions of the word w in each of those documents. Show that the size of the inverted index is of the same order as size of the original text collection.

5 marks

8. Suppose in a market basket scenario, the following association rules are detected with a minimum support 25% and a minimum confidence 80%:

$$\{A,B\} \rightarrow \{C\}$$

$$\{B,C\} \rightarrow \{D\}$$

From this information, determine the minimum support of the itemset $\{A, D\}$ as accurately as possible.

5 marks

A SQL Statements to create schema *employees*

```
CREATE TABLE employees (
    emp_no      INT             NOT NULL,
    birth_date  DATE            NOT NULL,
    first_name  VARCHAR(14)     NOT NULL,
    last_name   VARCHAR(16)     NOT NULL,
    gender      ENUM ('M','F')  NOT NULL,
    hire_date   DATE            NOT NULL,
    PRIMARY KEY (emp_no)
);
CREATE TABLE departments (
    dept_no     CHAR(4)         NOT NULL,
    dept_name   VARCHAR(40)     NOT NULL,
    PRIMARY KEY (dept_no),
    UNIQUE  KEY (dept_name) # i.e. there can be an index with unique key dept_name
);
CREATE TABLE dept_manager (
   dept_no      CHAR(4)         NOT NULL,
   emp_no       INT             NOT NULL,
   from_date    DATE            NOT NULL,
   to_date      DATE            NOT NULL,
   KEY         (emp_no),
   KEY         (dept_no),
   FOREIGN KEY (emp_no)  REFERENCES employees (emp_no)    ON DELETE CASCADE,
   FOREIGN KEY (dept_no) REFERENCES departments (dept_no) ON DELETE CASCADE,
   PRIMARY KEY (emp_no,dept_no)
);
CREATE TABLE dept_emp (
    emp_no      INT             NOT NULL,
    dept_no     CHAR(4)         NOT NULL,
    from_date   DATE            NOT NULL,
    to_date     DATE            NOT NULL,
    FOREIGN KEY (emp_no)  REFERENCES employees   (emp_no)  ON DELETE CASCADE,
    FOREIGN KEY (dept_no) REFERENCES departments (dept_no) ON DELETE CASCADE,
    PRIMARY KEY (emp_no,dept_no)
);
CREATE TABLE titles (
    emp_no      INT             NOT NULL,
    title       VARCHAR(50)     NOT NULL,
    from_date   DATE            NOT NULL,
    to_date     DATE,
    FOREIGN KEY (emp_no) REFERENCES employees (emp_no) ON DELETE CASCADE,
    PRIMARY KEY (emp_no,title, from_date)
);
CREATE TABLE salaries (
    emp_no      INT             NOT NULL,
    salary      INT             NOT NULL,
    from_date   DATE            NOT NULL,
    to_date     DATE            NOT NULL,
    FOREIGN KEY (emp_no) REFERENCES employees (emp_no) ON DELETE CASCADE,
    PRIMARY KEY (emp_no, from_date)
);
```

INDIAN STATISTICAL INSTITUTE

Semestral Examination

Post Graduate Diploma in Business Analytics - I Year, 2016-2017 (Semester - I)

Stochastic Processes and Applications

Date : 01.12.2016 Maximum Marks : 100 Duration : 3.0 Hours

Note: The question paper is of 112 marks. Answer as much as you can, but the maximum you can score is 100.

(Q1) A box contains two coins: an unbiased coin and a biased coin with $P(H) = \frac{2}{3}$. A coin is chosen at random and tossed once. We define the random variable X as a Bernoulli random variable associated with this coin toss, i.e., $X = 1$ if the result of the coin toss is heads and $X = 0$ otherwise. Then the remaining coin is taken and tossed once. The random variable Y is defined as a Bernoulli random variable associated with the second coin toss. Find the joint distribution of X and Y. Are X and Y independent? [10+5=15]

(Q2) Let $N_1(t)$ and $N_2(t)$ be two independent Poisson processes with rates $\lambda_1 = 1$ and $\lambda_2 = 2$, respectively. Let $N(t)$ be the merged process $N(t) = N_1(t) + N_2(t)$.

(a) Find the probability that $N(1) = 2$ and $N(2) = 5$.

(b) Given that $N(1) = 2$, find the probability that $N_1(1) = 1$.

[5+7 = 12]

(Q3) A dice is rolled repeatedly. Which of the following are Markov chains? For those that are, supply the transition matrix.

(a) The largest number M_n shown up to the nth roll.

(b) The number N_n of sixes in n rolls.

(c) At time r, the time C_r since the most recent six.

(d) At time r, the time B_r until the next six.

[5 × 4 = 20]

(Q4) We consider the Markov chain $\{X_n\}_{n=0,1,2,\cdots}$ with state space S = 1,2,3, initial state $X_0 = 2$, and transition matrix

$$P = \begin{bmatrix} 1 & 0 & 0 \\ p & 1-p-q & q \\ 0 & 0 & 1 \end{bmatrix}, \qquad p, q > 0, p+q < 1.$$

(a) Show that the time $T \geq 1$ when $\{X_n\}$ first changes its value follows a Geometric distribution.

(b) Show also that X_T is independent of T.

(c) Find the distribution of T.

[10+10+10=30]

(Q5) Taxis are waiting in a queue for passengers to come. Passengers for those taxis arrive according to a Poisson process with an average of 60 passengers per hour. A taxi departs as soon as two passengers have been collected or 3 minutes have expired since the first passenger has got in the taxi. Suppose you get in the taxi as first passenger. What is your average waiting time for the departure? [15]

(Q6) Describe the classical decomposition of the Time series data. Define Autoregressive and Moving Average Processes. Derive the condition that enables us to invert an AR(1) process to an MA(∞) process. [6+4+10=20]

INDIAN STATISTICAL INSTITUTE

POST-GRADUATE DIPLOMA IN BUSINESS ANALYTICS
Semestral (Back-paper/Supplementary) Examination: 2016–17
(Semester I)

Course: Statistical Structures in Data

Date: December 22, 2016 Maximum Marks: 100 Time 3 hr.

1. Let X be a continuous random variable having a probability density function

$$f(x) = \begin{cases} 0 & \text{if } x < -a, \\ \dfrac{2(x+a)}{a(a+b)} & \text{if} -a \le x \le 0, \\ \dfrac{2(b-x)}{b(a+b)} & \text{if } \ 0 < x \le b, \\ 0 & \text{if } x > b, \end{cases}$$

where a and b are positive constants.

a) Express the mean and the variance of this distribution in terms of a and b.
b) What is the mode of the distribution?
c) Fit the aforementioned distribution to the following grouped frequency distribution constructed from 120 observations on a continuous random variable X and test the goodness of fit by a Pearson χ^2-statistic at the 5% level.

Mid-point of Class interval	-15	-5	5	15	25	35
Frequency	5	35	40	20	15	5

[(2+4)+2+(8+2)=20]

2.

a) The joint probability density function of two random variables, X and Y, is given by

$$f(x,y) = \frac{1}{\pi\sqrt{3}} e^{-2\left(x^2 - xy + \frac{y^2}{3}\right)}, \quad -\infty < x, y < \infty.$$

Deduce the regression of Y on X.

(Please turn over.)

b) For a pair of random variables x and y, the regression lines are

$$3x + 2y = 25 \text{ and } 6x + y = 30.$$

i. Which of these is the regression line of x on y? Justify your answer.
ii. What will be the equation of the SD line?

c) The least-squares regression line of y on x and of x on y are respectively, $x + 3y = 0$ and $3x + 2y = 0$. The sample standard deviation, s_x, of x is 1. Defining new random variables $u = x + y$ and $v = x - y$, determine the least-squares regression line of v on u.

[8+(4+2)+6=20]

3. In a customer preference survey, customers were asked to rate several attributes of a new ready-to-eat product. The attributes were:

- Taste (X_1)
- Good buy for money (X_2)
- Flavour (X_3)
- Suitability as a snack (X_4)
- Source of energy (X_5)

The responses were tabulated and the correlation matrix **R** was found to be

$$\mathbf{R} = \begin{bmatrix} 1 & 0.02 & \mathbf{0.96} & 0.42 & 0.01 \\ & 1 & 0.13 & 0.71 & \mathbf{0.85} \\ & & 1 & 0.50 & 0.11 \\ & & & 1 & 0.79 \\ & & & & 1 \end{bmatrix}$$

Based on this, the estimates of factor loadings for a 2-factor model were computed by the maximum likelihood method. These are given in the following table, together with the estimated varimax-rotated factor loadings:

Attribute	MLEs of Factor Loadings		MLEs of Rotated Factor Loadings	
	F_1	F_2	F_1^*	F_2^*
X_1	0.56	0.82	0.02	0.99
X_2	0.78	-0.53	0.94	0.01
X_3	0.65	0.75	0.13	0.98
X_4	0.94	-0.11	0.84	0.43
X_5	0.80	-0.54	0.97	-0.02

a) Using the estimated unrotated factor loadings, obtain estimates of the specific variances and communalities, explaining briefly the underlying theory.
b) In the unrotated factor model, suggest an interpretation of the two factors.
c) In what manner does the interpretation of the factors change after rotation? Explain.

[12+4+4=20]

4. Describe in detail how a regression tree is fit to a data set containing n observations on p predictor variables, $X_1, X_p, \cdots, X_p$, and one response variable Y. Also discuss how an optimal tree is obtained using the principle of cost-complexity pruning.

[12+8=20]

5. Let **X** be a trivariate random variable, presumed to have mean vector equal to the null vector **0**. Based on 120 observations on **X**, the following dispersion matrix is obtained:

$$\mathbf{S} = \begin{pmatrix} 2.00 & 3.33 & 1.33 \\ & 8.00 & 4.67 \\ & & 7.00 \end{pmatrix}.$$

a) Determine the principal components of **X**.
b) How many principal components would you need to explain 90% of the total variation in the data? Justify your answer.
c) Will the principal components remain the same if you compute them from the sample correlation matrix **R** rather than from **S**? Explain.

[10+5+5=20]

INDIAN STATISTICAL INSTITUTE

SUPPLEMENTARY EXAMINATION: 2016 – 17

Course Name: PGDBA 2016 - 2018

Subject Name: Inference (BAISI3)

Date: 23.12.2016 **Maximum Marks: 100** **Duration: 3 hours**

Notes, if any: The paper carries 110 marks. Answer all questions. However, the maximum you can score is 100.

1. Answer the following
   a. Consider the following random variables. Identify the distributions they are likely to follow and their parameters.
      i. The time taken to provide service in a garage
      ii. The number of customers arriving in a bank in a given time interval
      iii. The height of adult males belonging to a particular race
      iv. The amount of error observed in repeat measurement of weights of a particular box
      v. Number of 1, 2, 3, 4, 5, and 6s obtained in n throws of a 6 faced dice [5 X 2 = 10]

   b. Consider the time taken to provide service in a garage. Suppose you have observed the actual time to provide service in n cases. Suppose the observed times are $x_1, x_2, \ldots, x_n$ respectively. Obtain the maximum likelihood estimate for the average service time. Do you think, this is a consistent estimate? Explain. [4 + 3 = 7]

   c. Suppose the six faced dice was thrown n times and 1s, 2s ...,6s appeared $n_1, n_2, \ldots n_6$ times. Thus $\Sigma x_j = n, j = 1..6$. Obtain the maximum likelihood estimate of the chance of getting a 3 on a single throw. What is the variance of this estimate? [4 + 4 = 8]

2. Suppose a particular dimension of a product is critical for its accurate functioning. The dimension is measured by the producer as well as by the customer. Suppose n products numbered 1, 2, ... n have been collected and their dimensions as measured by the manufacturer were $x_1, x_2, \ldots x_n$ respectively (i.e. the j^th product had the dimension x_j, j = 1, 2, ...n). The same products were measured by the customer and the dimensions were $y_1, y_2, \ldots y_n$. Thus x_j and y_j are measurement of the same product. It is known that the measurement has a large variability and hence x_j and y_j are likely to be different even though they measure the same quantity. A product is acceptable in case its dimension is less than or equal to μ_0. It may be assumed that the measured dimension follows a normal distribution with unknown mean and variance.
   a. There is a belief that the measurement by the manufacturer and the customer do not match on an average.
      i. Formulate this problem as a problem of test of hypothesis. Write the null and the alternative hypotheses in terms of the parameters. [5]
      ii. Write down the test statistic for testing the proposed hypotheses. [4]

   b. Suppose the manufacturer measured the dimension of a particular product four times and got the values as x_1, x_2, x_3, and x_4 respectively. Suppose the true value of the dimension of the product is μ (unknown). Suppose we assume that the measured values may be described in terms of the model $x_j = \mu + \varepsilon_j$.
      i. What would be the distribution of ε_j? [2]

ii. How would you use the measured values to decide whether the particular product is acceptable or not? Explain using the theories of hypotheses testing / confidence interval. Formulate the hypotheses clearly. [8]

c. It is claimed by the manufacturers that their measurements have lower variability compared to the variability of the measurement undertaken by the customers.

i. What data will you collect to verify this claim? [2]

ii. Formulate this problem as the problem of test of hypotheses. Write the null and alternative hypotheses clearly. [4]

3. 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. Identify the explanatory and response variables in this context. [3]

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.

i. Present the data suitably. [5]

ii. Someone claims that it is a prospective cohort study. Do you agree? If not, what kind of study is this? Explain. [5]

iii. Do you think that the data support the claim that higher moisture leads to lower yield? Explain. [7]

iv. Suppose you want to test the claim that moisture impacts yield.

a) Write this in terms of tests of hypotheses, i.e. write the null and the alternative hypotheses. [4]

b) How would you interpret type I and type II errors in this case? [6]

4. Indicate whether the following are true or false. Do not provide any explanation – only state whether the statement is true or false.

a. A box has a large number of red and white balls. It is believed that the number of white and black balls are the same. However, someone claims that 60% of the balls are white. You have set up null and alternative hypotheses to test the belief against the proposed alternative. Both the null and the alternative hypotheses are simple hypotheses.

b. Range is a more resistant estimator of dispersion compared to standard deviation.

c. The chance that a 95% confidence interval constructed for a parameter θ contains the true value is at least 95%.

d. Relative risk close to 1 indicates dependence of the variables being studied.

e. TV channels conduct surveys by requesting their viewers to provide opinion on specific topics. These samples can be considered to be random samples.

f. Log of odds ratio can take any numeric value, including negative values.

g. The standard error of an unbiased estimator is smaller than the standard error of a biased estimator

h. Retrospective cohort studies use data collected for other studies

i. The mean function plot is used to understand the form of relationship between two variables

j. Box plot can be used to understand the pattern of variation of data collected in nominal scale

k. The power function is used to minimize the chance of rejecting a null hypothesis when it is true

l. The higher the curvature of the likelihood function, the higher the standard error of the corresponding maximum likelihood estimator

m. Z test is used to test equality of means when variability is unknown

n. Non-sampling errors generally pose a bigger problem than sampling errors

o. Type I error gives the probability of accepting the alternative hypothesis when it is not true.

[15 X 2 = 30]

INDIAN STATISTICAL INSTITUTE

Supplementary Semestral Examination

Post Graduate Diploma in Business Analytics 2016-17 (Semester-I)

Stochastic Processes and Apllications

Date: **December 26, 2016** Maximum Marks: **100** Duration: **3 Hours**

Note: All Notations have their usual meanings. The question paper carries a total of 110 marks. You can answer as much as you can, but the maximum you can score is 100.

1. A total of n people randomly take their seats around a circular table with n chairs. No two people have the same height. What is the expected number of people who are shorter than both of their immediate neighbors?

   (12)

2. Define an absorbing state. Show that if a is an absorbing state, then $P^n(x,a) = P_x(T_a \leq n)$, $n \geq 1$.

   (2+10=12)

3. Define a transient and a recurrent state. Show that if y is a transient state. Prove that $P_x(N(y) < \infty) = 1$ and $G(x,y) = \frac{\rho(x,y)}{1-\rho(x,y)}$ for $x \in S$.

   (4+(6+6)=16)

4. Let C be a finite irreducible closed set of states. Prove that every state in C is recurrent.

   (20)

5. An urn contains 3 red balls and 2 blue balls. A ball is drawn. If the ball is red, it is kept out of the urn and a second ball is drawn from the urn. If the ball is blue, then it is put back in the urn and a red ball is added to the urn. Then a second ball is drawn from the urn.

   a) What is the probability that both balls drawn are red?

   b) If the second drawn ball is red, what is the probability that the first drawn ball was blue?

   c) Suppose the process i continued. Let X_n denote the number of red balls in the urn before the $n-th$ draw. Is X_n a Markov chain. Justify your answer.

   (7+8+15=30)

6. Describe the classical decomposition of the Time Series Data. Define Autoregressive and Moving Average Processes. Derive the condition that enables us to invert an $AR(1)$ process to an $MA(\infty)$ process.

   (6+4+10=20)

PGDCA

(POST GRADUATE DIPLOMA IN COMPUTER APPLICATIONS)

Indian Statistical Institute

Semester_1 Examination, PGDCA: 2016-2017
Subject: Introduction to Computer Architecture and systems software (101)
Total marks:110, Full marks:100, Duration:180 mins

1.(a) What do you mean by system software?
(b) With 'n' variables, what would be the maximum number of possible logical expressions?
(c) What GATE(S) is/are called universal GATE and why?
(d) Implement full adder using minimum number of NAND GATEs.(2+1+2+5)

2.(a) $f(A,B,C,D)=\sum(0,1,4,5,8,9,13,15)$, minimize f.
(b) Implement 16:1 MUX using 4:1 and 2:1 MUX.
(c) Implement $f(A,B,C,D)=\sum(1,4,5,7,9,12,13)$ using 4:1 MUX.(2+3+5)

3.(a) Why is there Race around condition in JK flip flop and show how master-slave JK flip-flop removes it?
(b) Convert D flip-flop into JK flip-flop. (1+4+5)

4.(a) Show that Counter can be used as frequency divider and draw the timing diagram.
(b) Design 3 bit UP/DOWN counter and draw the timing diagrams. (4+6)

5.(a) Show the IEEE754 binary representation of the number (-1.75)10 in single precision.
(b) Write Booth's algorithm of multiplication and multiply 14 and -5, use 5 bits to store multiplicand and 5 bits for multiplier and 10 bits for product register. (4+3+3)

6. What do you understand by addressing mode and effective address? An instruction is stored at location 300 with its address field at location 301. The address field has value 400. A processor register R1 contains the number 200. Calculate the effective address if the addressing mode of the instruction is (i) Direct (ii) Immediate (iii) Relative (iv) Register indirect.(1+1+2+2+2+2)

7.(a)A CPU has a 32 KB direct mapped cache with 128 byte block size. Suppose 'A' is a two dimensional array of size 512 X 512 with elements that occupy 8 byte each. Consider the following two C code segments, p1 and p2.

```
P1: for(i=0;i<512;i++){
for(j=0;j<512;j++){
x+=A[i][j];
}
}
P2: for(i=0;i<512;i++){
for(j=0;j<512;j++){
x+=A[j][i];
}
}
```

p1 and p2 are executed independently with the same initial state, namely, the array A is not in cache and i,j,x are in registers. Let the number of cache misses experienced by p1 be M1 and that for p2 be M2. Find M1/M2.

(b) 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. Find the percentage of block miss in cache if LRU replacement policy is used. (5+5)

8.Design and describe each logical part <u>precisely</u> of a 32 bit ALU that can perform Addition, Substraction, OR and AND operations. (10)

9.(a) What do you mean by ROM? Explain different types of ROM.
(b) What do you mean by the locality of reference?
(c) Express X=(A+B)*(C+D) instruction into (i) One address instruction and (ii) Zero address instruction. (1+2+2+2*5/2)

10. (a) Define different types of Interrupts? Explain Interrupt driven I/O mode.
(b) What is the disadvantage of Interrupt driven I/O mode?
(c) If 'n' bits are used to represent an integer then what would be the minimum number in 2's complement representation and maximum number in 1's complement representation?((2+5)+1+(1+1))

11.(a) What do you mean by clock? What are the differences between Latch and flip-flop?
(b)Assume a program that runs in 10 sec on a computer 'A', which has a 400MHz clock. Suppose you are trying to build a computer 'B' that will run the same program in 6 sec and requires 1.2 times as many clock cycles as computer 'A' for this program. What clock rate should your computer 'B' have ?
(c) Suppose you have two implementations of the same instruction set architecture. Machine A has clock cycle time 1 ns and CPI of 2.0 for some program and machine B has clock cycle time 2 ns and CPI of 1.2 for the same program. Which machine is faster? ((1+2)+3+4)

Indian Statistical Institute

Course Name : PGDCA

First Semester Examination (2016-17)

Paper Name: Introduction to Programming (102)

Time: 3 Hours **Marks: 100** **Date: 14.12.2016**

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

For all the programs, assume input should be taken from keyboard. Also take note that ANCI-C standard should be followed.

1. (a) Write a program in C that will implement merge sort that sort and merge two different size arrays of integer numbers.

(b) Write a program in C using array to implement binary search of searching an integer number from a set of integer numbers.

(c) Write program in C that will multiply two matrices (of integer numbers) of order 3 by 2 and 2 by 3.

(d) Differentiate between array and structure.

[7+6+7+2 = 22]

2. (a) Differentiate between recursion and iteration.

(b) Write a program in C using recursion to find out the value of the n th term in a Fibonacci series (the value of n should be taken from user).

(c) "Is switch-case a replacement of if"-critically comment.

(d) Write a program in C to find out whether a year (taken as user input) is leap year or not using conditional/ternary operator [Take into account all possible cases of leap year].

(e) Differentiate between function and macro.

[2+6+2+4+2 = 16]

3. (a) Write a program in C that will implement circular queue of integer numbers using array.

(b) Evaluate the following comma separated expression given in postfix notation (using stack)

5, 6, 2, +, *, 12, 4, /, -

(c) What are the advantages of linked list over array?

(d) Write a program in C that will perform the following operations on single linked list(where each node will contain integer values as data part) considering all possible situations:

(i) add a node at the end of the list. (ii) delete a node from any position from the list.

[7+4+2+(4+4) = 21]

4. (a) Write a program in C that will calculate the number of vowels present in a sentence which should be taken as input from keyboard.

(b) Write a user defined function in C which is equivalent to strcat().

(c) Differentiate between malloc() and calloc().

(d) Write the difference between algorithm and flowchart.

(e) Write algorithm and flowchart to calculate the maximum of three integer numbers.

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

5. (a) What do you mean by ADT (Abstract Data Type)? Differentiate between linear and non-linear data structure.

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

Inorder: DBHEAIFJCG Preorder: ABDEHCFIJG

Draw the binary tree and also make the binary tree a threaded binary tree using inorder threading.

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

(d) Draw an AVL tree with the following set of nodes (show all the rotations):

64, 1, 44, 26, 13, 110, 98, 85

[(2+2)+(3+4)+(2+2)+6 = 21]

6. What will be the output of the following?

(a)
```
int x=5;
int main()
 {
   int x=10;
   printf("%d ",x);
   fun(x);
   return(0);
 }
void fun(int  y)
 {
   printf("%d %d",x, y);
 }
```

(b)
```
#define x   5+3
int main()
 {
    int y;
    y=x*x*x;
    printf("%d",y);
    return(0);
 }
```

(c)
```
int main()
 {
    int x=10;
    printf("%d %d %d", x!=10, x=15, x>10);
    return(0);
 }
```

(d)
```
int main()
 {
    int i=5;
    int *p;
    p=&i;
    magic(&p);
    printf("%d",*p);
    return(0);
    }
  void magic(int **p)
    {
      int x=10;
      *p=&x;
      printf("%d",**p);
    }
```

```
(e)  int main()
       {
         int a=5;
         int c;
        c=++a + ++a + a++;
        printf("%d %d",a,c);
        return(0);
       }
```

[2×5 = 10]

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

Indian Statistical Institute

Course Name : PGDCA

First Semester Examination (2016-17)

Paper Name: Introduction to Programming (102)

Time: 3 Hours **Marks: 100** **Date: 14.12.2016**

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

For all the programs, assume input should be taken from keyboard. Also take note that ANCI-C standard should be followed.

1. (a) Write a program in C that will implement merge sort that sort and merge two different size arrays of integer numbers.

   (b) Write a program in C using array to implement binary search of searching an integer number from a set of integer numbers.

   (c) Write program in C that will multiply two matrices (of integer numbers) of order 3 by 2 and 2 by 3.

   (d) Differentiate between array and structure.

   [7+6+7+2 = 22]

2. (a) Differentiate between recursion and iteration.

   (b) Write a program in C using recursion to find out the value of the n th term in a Fibonacci series (the value of n should be taken from user).

   (c) "Is switch-case a replacement of if"-critically comment.

   (d) Write a program in C to find out whether a year (taken as user input) is leap year or not using conditional/ternary operator [Take into account all possible cases of leap year].

   (e) Differentiate between function and macro.

   [2+6+2+4+2 = 16]

3. (a) Write a program in C that will implement circular queue of integer numbers using array.

(b) Evaluate the following comma separated expression given in postfix notation (using stack)

5, 6, 2, +, *, 12, 4, /, -

(c) What are the advantages of linked list over array?

(d) Write a program in C that will perform the following operations on single linked list(where each node will contain integer values as data part) considering all possible situations:

(i) add a node at the end of the list. (ii) delete a node from any position from the list.

[7+4+2+(4+4) = 21]

4. (a) Write a program in C that will calculate the number of vowels present in a sentence which should be taken as input from keyboard.

(b) Write a user defined function in C which is equivalent to strcat().

(c) Differentiate between malloc() and calloc().

(d) Write the difference between algorithm and flowchart.

(e) Write algorithm and flowchart to calculate the maximum of three integer numbers.

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

5. (a) What do you mean by ADT (Abstract Data Type)? Differentiate between linear and non-linear data structure.

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

Inorder: DBHEAIFJCG Preorder: ABDEHCFIJG

Draw the binary tree and also make the binary tree a threaded binary tree using inorder threading.

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

(d) Draw an AVL tree with the following set of nodes (show all the rotations):

64, 1, 44, 26, 13, 110, 98, 85

[(2+2)+(3+4)+(2+2)+6 = 21]

6. What will be the output of the following?

(a)

```
int x=5;
int main()
 {
   int x=10;
   printf("%d ",x);
   fun(x);
   return(0);
 }
void fun(int  y)
 {
   printf("%d %d",x, y);
 }
```

10 5 10

(b)

```
#define x   5+3
int main()
  {
    int y;
    y=x*x*x;
    printf("%d",y);
    return(0);
  }
```

38

(c)

```
int main()
  {
    int x=10;
    printf("%d %d %d", x!=10, x=15, x>10);
    return(0);
  }
```

1 15 0

(d)

```
int main()
  {
    int i=5;
    int *p;
    p=&i;
    magic(&p);
    printf("%d",*p);
    return(0);
```

10 10

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

8 21

```
(e) int main()
      {
        int a=5;
        int c;
       c=++a + ++a + a++;
       printf("%d %d",a,c);
       return(0);
      }
```

[2×5 = 10]

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

Indian Statistical Institute

Semester_1 Examination, PGDCA: 2016-2017
Subject: Operating System(104)
Total marks:110, Full marks:100, Duration:180 mins

1.(a) What are the functions of Operating System? (b) What is multiprocessor operating system? Mention the main advantages and disadvantages of multiprocessor system (c) What is a Real Time Operating Systems? Mention the types of real time operating system with examples. . (3+1+3+1+2)=10

2. Explain the contents of process control block of a process. Draw the process state diagram of a process and explain each components. (5+5)=10

3.(a) Consider a system with 'N' Processors What can be the minimum and maximum number of processes which may be present in the ready, running, and block states? (b) Find the average waiting time of the processes if they use shortest remaining time first scheduling algorithm by using following table.

Process number	Arrival time	Burst time
1	3	4
2	4	2
3	5	1
4	2	6
5	1	8
6	2	4

(5+5)=10

4. (a) What are the neccessary and sufficient conditions? (b) Consider processes p1,p2,p3 and p4 arriving in the ready queue in the same order at time zero. If burst time requirement of these processes are 4,1,8,1 respectively then what is the completion time of process p1 by using Round Robin scheduling if the time quantum is of 1 unit? (c) Using the priority scheduling and the table given below, find the finishing time of processes. Processes can perform their I/O operation independently.

process	Arrival time	priority	CPU Time	I/O Time	CPU Time	I/O Time
P1	0	2	1	5	3	10
P2	2	3(Low)	3	3	1	15
P3	3	1(High)	2	3	1	9

- (2+4+4)=10
-

5. (a) What do you mean by Inter-Process Communication(IPC)? (b) On the basis of IPC what are the types of processes? Give the differences between them. (c) Write peterson's algorithm for two process synchronization. (1+4+5)=10

-

6. (a) Describe the different applications of semaphores usind examples with code. (b) Write a Pseudocode solution for Reader-Writer problem. (5+5)=10

7. (a) There are five processes(p0,p1,p2,p3,p4) and three resource type(A,B,C) . Maximum need of each process of each type of resources are {(7,5,3), (3,2,2), (9,0,2), (2,2,2), (4,3,3)} and current allocations are {(0,1,0), (2,0,0), (3,0,2), (2,1,1), (0,0,2)} respectively. Find a safe sequence. (b) Describe the techniques for a system to recover from deadlock. (5+5)=10

8. (a) Each process p_i, i=1 to 9 executes the following code. Initially semaphore mutex=1.

```
    Repeat
    p(mutex)
    critical section
    v(mutex)
    forever
```

And process p10 executes following code

```
    repeat
    v(mutex)
    critical section
    p(mutex)
    forever
```

What is the maximum number of processes that may be present in the critical section at any point of time?

(b) Explain the criteria that should meet for solution of a critical section problem?

(c)Consider a system with processes p1,p2,p3 and p4. Each process demand 5,6,9,1 tape drive respectively. What would be the minimum number of tape drive require to make system deadlock free? (5+3+2)=10

9.(a) What are the differences between static and dynamic loading? (b) Consider a system with logical address space of 1280M words and physical address of 24 bits. The physical address space is divided into 8K frames. Calculate the page size and number of pages in the logical address space. (c) Consider a memory system of two level with TLB access time 60 ns. What hit ratio is required to reduce the effective memory access time from 300 ns without TLB, to 250 ns with TLB? (3+4+3)=10

10.(a)Draw and explain the paging hardware with TLB precisely. (b) What is the difference between secrecy and privacy? Mention some mechanism of authentication((5+5)=10

11. Consider page reference string 7,0,1,2,0,3,0,4,2,3,0,3,2,1,2,0,1,7,0,1.
(a) If first in first out page replacement algorithm is used. Find the difference of the number of page faults when (i) three frames available to the process and (ii) four frames are available to the process. (b) When the process uses least recently used(LRU) page replacement algorithm then find the percentage of page hit. (6+4)=10

Indian Statistical Institute

Course Name : PGDCA

First Semester Examination (2016-17)

Paper Name: DBMS (103)

Time: 3 Hours **Marks: 100** **Date: 19.12.2016**

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

1. (a) What are the advantages of DBMS over a flat file system?

   (b) Draw and briefly explain the 3 level ANSI-SPARC architecture for DBMS.

   (c) What do you mean by logical and physical data independence?

[4+4+4 = 12]

2. (a) Define closure of a set of functional dependencyies.

   (b) Consider two sets of FD's

   F = {B->CD, AD->E, B->A} G = {B->CDE, B->ABC, AD->E}

   Prove that both the FD's are equivalent.

   (c) Assume a relation schema R (A,B,C,D,E,F,G,H,I,J) with the following sets of FD's {AB->C, BD->EF, AD->GH, A->I, H->AJ}.

   Find the key/keys.

[2+4+5 = 11]

3. (a) Define foreign key.

   (b) Differentiate between natural join and equijoin.

   (c) Assume the following schema and write SQL, operator tree (using relational algebra operators) and finally write the optimized operator tree of the following query.

   EMP (fname, lname, ssn, bdate, address, sex, salary, super_ssn, dno)

DEPT (dname, dnumber, mgr_ssn, mgr_start_date)

PROJ (pname, pnumber, plocation, dnum)

Query: For every project located in 'Houston', list the project number, the controlling dept. number, dept. manager's last name, salary and address.

[2+3+(2+2+2) = 11]

4. (a) Differentiate between superkey and candidate key.

(b) Write the relational algebra query to find out all the employees first name and their supervisors' last name (Assume the schema of Question no 3(c)).

(c) Differentiate between inner join and outer join.

(d) What are the roles of a database administrator?

[2+3+2+3 = 10]

5. (a) Explain with examples (i) deletion anomaly (ii) modification anomaly.

(b) Define BCNF. Give one example of a relation which is in 3NF but not in BCNF.

[(4+4)+(2+3) = 13]

6. (a) Consider a relation schema EMP_DEPT(Ename,Ssn,Bdate,Address,Dnumber,Dname,Dmgrssn) where **Ssn** is the primary key. Consider the following Functional dependency hold in relation schema EMP_DEPT : Dnumber->{Dname,Dmgrssn}.

Find out whether this table is in 3NF or not. If not, then make decomposition to make it in 3NF.

(b) Define denormalization. Why do we need denormalization?

(c) What is the use of data dictionary?

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

7. (a) What is week entity set in ERD? State with an example.

(b) Differentiate between generalization and specialization. Give one example of specialization lattice.

(c) What do you mean by ternary relationship in ERD? State with an example.

[3+(3+2)+3 = 11]

8. (a) State and briefly explain the ACID properties that every transaction must follow.

(b) Briefly explain the dirty read problem occurred because of concurrent transaction execution.

(c) Discuss the immediate update recovery technique for multiuser environment.

[4+3+4= 11]

9. (a) Consider three transactions T1, T2 and T3 and the following non-serial schedule S1. Draw the precedence graph for S1 and state whether the schedule is serializable or not. If the schedule is serializable, then write down the equivalent serial schedule(s).

S1: r3(y);r3(z);r1(x);w1(x);w3(y);w3(z);r2(z);r1(y);w1(y);r2(y);w2(y);r2(x);w2(x).

(b) Write the read_lock(), write_lock() and unlock() algorithms for shared-exclusive locking.

[5+(2+2+2) = 11]

10. Assume the following schemas and write SQL of the following queries.

CUST(c_id,c_name,area,ph_no),MOVIE(mv_no,title,type,star,price),

INVOICE(inv_no,mv_no,c_id)

(i) Find the name of the customers who stay in an area whose second letter is 'b' or 'd'.

(ii) Change the price of the movie titled 'Lagaan' to Rs. 250.00.

(iii) Find the name and phone no of customers who have been issued movie of type 'thriller'.

(iv) Add a column named director with type varchar2 and size 20 in the movie table.

(v) Count separately the number of movies in the 'action' and 'thriller' types.

[2×5= 10]

INDIAN STATISTICAL INSTITUTE
GIRIDIH

Group – A (Discrete Mathematics)
Use separate answer sheets

Full Marks: 50 **Max. Time: 2 hrs.**

21/12/2016

Answer any 5:

1. a) A code is being written using four symbols a, b, c and d.

   i) How many 12-digit code-words are there that use exactly 3 of each symbol? (2)

   ii) If a 12-digit code-word is chosen at random, what is the probability that it will use exactly 5 a's, 4 b's, 2 c's and 1 d? (3)

   b) Suppose we have 27 distinct odd positive integers all less than 100. ('Distinct' means that no two numbers are equal). Show that there is a pair of numbers whose sum is 102. (5)

2. a) Suppose that there are p different kinds of objects, each in infinite supply. Let a_k be the number of permutations of k objects chosen from these objects. Find a_k explicitly by using exponential generating functions. (4)

   b) Consider RNA chains having four bases A, G, C, and U. Find the number of 7-link RNA chains having 2 A's, 1 G, 3 C's, and 1 U. (3)

   c) Find out the number of RNA chains of length k if we assume an arbitrarily large supply of each base (bases are same as in 2. b)). (3)

3. a) Define derangements. Let D_n be the derangements of n objects. Find the recurrence for D_n and derive the values of D_n for n=5, 7. (1.5 + 1.5 = 3)

   b) A straight line separates a plane into two disjoint regions. Suppose we have n straight lines such that no two of them are parallel and no three of them intersect at the same point. Into how many disjoint regions do these lines divide the plane? Derive the corresponding recurrence. Use the recurrence to find the number of disjoint regions when n=5. (1+1+1 = 3)

   c) Solve the following recurrence under the given initial conditions

$$a_n = a_{n-1} + 2a_{n-2}, \quad a_0 = 2, \ a_1 = 7 \qquad (4)$$

4. a) Define a generating function. Solve the following recurrence using a generating function

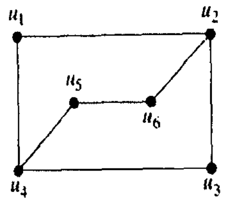
$$a_{k+1} = 2a_k + 4^k, a_1 = 3 \qquad (5)$$

   b) Suppose that $f(n+1) = f(n)f(n-1)$ for all $n \geq 1$ and $f(0) = f(1) = 2$. Find $f(n)$. (5)

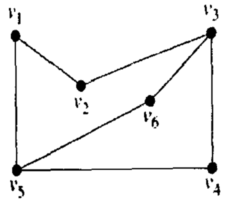
P.T.O

5. a) Define isomorphism between graphs. (2)

b) Check whether the following two graphs are isomorphic or not. If isomorphic, then find the correspondences. (3)

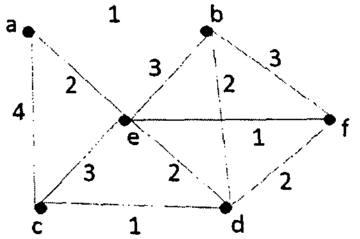


G



H

c) Using Prim's algorithm find the minimum spanning tree of the following connected weighted graph. (5)



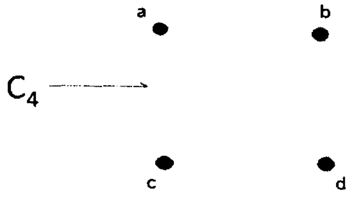
6. a) Define Hamiltonian path and Eulerian path of a graph each with an example. (2+2=4)

b) Define chromatic number of a graph. What is the chromatic number of a bipartite graph? What is the chromatic number of a complete graph?

(1+1+1 =3)

c) Let P(G, x) be the number of ways to color a graph G with at most x colors. Consider K_2 and if x colors are available, then $P(K_2, x)=x(x-1)$. Find the expression for $P(C_4, x)$.

(3)



7. a) Evaluate the following prefix and postfix expressions. (2)

$+-*235/\uparrow 234$ $\qquad$ $723*-4\uparrow 93/+$

b) Show that $\neg(p \vee (\neg p \wedge q))$ and $\neg p \wedge \neg q$ are logically equivalent using both truth tables and laws.

(4)

c) Check whether the following compound proposition is a tautology or not. (Verify through both truth tables and laws.)

$$(\neg q \wedge (p \rightarrow q)) \rightarrow \neg p$$

(4)

INDIAN STATISTICAL INSTITUTE
GIRIDIH

Group – B (Statistics)

Use separate answer sheets

Full Marks: 50

21/12/2016

1. Write true or false: 1 x 10 = 10

   i) A discrete variable assumes only integer values.

   ii) The difference between upper and lower limits is the same as that between upper and lower boundaries.

   iii) Step diagrams are used for showing cumulative frequencies of a discrete variable.

   iv) Mean of first n natural numbers is (n + 1)/2.

   v) Standard deviation of a variable is affected by change of both origin and scale.

   vi) All odd order central moments are zero for any symmetrical frequency distribution.

   vii) For a negatively skew distribution, mean < median < mode.

   viii) For any distribution, first order central moment is zero.

   ix) If A and B are disjoint events, then probability of occurring both A and B is zero.

   x) If one of the linear regression coefficients is greater than one, then the other is less than one.

2. Select the correct one: 1 x 10 = 10

   i) Grade obtained in an examination is
   (a) attribute, (b) discrete variable, (c) continuous variable, (d) none of these.

   ii) For a random variable X, the value of its distribution function $F(x) = P(X \leq x)$ at $x = \infty$ is
   (a) < 1, (b) > 1, (c) 1, (d) none of these.

   iii) If sample space is finite and an event A is defined on the sample space such that $P(A) = 1$, then A is always
   (a) a sure event, (b) an impossible event, (c) an elementary event, (d) none of these.

   iv) For any two events A and B, $P(A$ or $B)$ is
   (a) $\leq P(A) + P(B)$, (b) $\geq P(A) + P(B)$, (c) $= P(A) + P(B)$, (d) none of these.

   v) If $y = 5 + x$, then
   (a) var(x) > var(y), (b) var(x)=var(y), (c) var(x)<var(y), (d) none of these.

   vi) Sum of squares of the deviations of the variate values from A is minimum, when A is
   (a) mean, (b) median, (c) mode, (d) none of these.

   vii) When a variable takes two distinct values with equal probabilities, the third order central moment is
   (a) < 0, (b) > 0, (c) 0 (d) none of these.

   viii) The event 'occurrence of A but not B' can be expressed as
   (a) $A \cap B$, (b) $(A \cup B)^c$, (c) $A \cap B^c$, (d) none of these.

P.T.O.

ix) If $P(A \cap B) = P(A).P(B)$, then the events A and B are
(a) independent, (b) may not be independent, (c) mutually disjoint, (d) none of these.

x) If $u = -2x$, $v = -7y$, then cov(u,v) is
(a) $-14\, cov(x, y)$ (b) $14\, cov(x, y)$ (c) $cov(x, y)$ (d) none of these.

3. Answer the following questions: 5 x 6 = 30

i) If a variable x takes the values 1, 2, ..., r with $F_1, F_2, ..., F_r\ (= n)$ as the corresponding less-than type cumulative frequencies, then prove that $\bar{x} = (r + 1) - \frac{1}{n}\sum_{i=1}^{r} F_i$, $\bar{x}$ = mean of x,
OR, If two variables x and y are linearly related in the form $y = a + bx$, then prove that median of y = a + b × median of x.

ii) Prove that the geometric mean of the ratio of two variables is the ratio of their geometric means.
OR, The harmonic mean and geometric mean of two positive observations are 12 and 18 respectively. Find their arithmetic mean.

iii) Show that the standard deviation cannot be smaller than mean deviation about mean.
OR, Show that if two variables x and y are linearly related in the form $y = a + bx$, then prove that $Q(y) = |b|\, Q(x)$, where Q(x) and Q(y) are quartile deviations of x and y respectively .

iv) Show that the standard deviation of the first n odd positive integers is equal to that of the first n even positive integers.
OR, For two values (say a and b, $a < b$) of a variable x, the mean and the standard deviation are respectively 25 and 4. Find a and b.

v) If A_1 and A_2 are two events, which are not necessarily mutually exclusive, then prove that
$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2).$$
OR, Five different letters are put at random inside 5 addressed envelopes. Find the probability of putting exactly 2 letters in the correct envelopes.

vi) Out of the two lines of regression given by $x + 2y - 5 = 0$ and $2x + 3y - 8 = 0$, which one is the regressión line of x on y and y on x ? Find means of x and y.
OR, If X follows N(μ, σ^2), find its median.

Indian Statistical Institute

Semester-2, PGDCA

End-semester Examination (4th May, 2017)

Subject: Computer Network (205)

Total Marks: 110   Maximum marks: 100   Duration: 240 mins

1. (a) What are the components of Data Communication in Computer Network? Explain each.(b) What are the advantages of Star topology over Mesh topology? (c) What is DNS?
(5+3+2)

2. (a) What are the differences between Limited Broadcasting and Directed Broadcasting? (b) Suppose you are given a Class 'C' network with id 200.1.2.0. Divide the given network into four equal sized sub network and calculate Subset id and Directed broad cast address of each sub networks & number of hosts in each sub network. (2+2+3+3)

3. (a) What do you mean by Queuing delay in computer network? (b) Suppose 'N' packets are to be sent from sender to receiver using Stop And Wait ARQ protocol and 'P' is the probability of losing packets. Then calculate the total number of packets required to be transmit in order to send 'N' packets successfully. (c) Suppose Transmission time, Propagation time and Band width of a network are 1 ms, 49.5 ms and 40Mbps respectively. Calculate the efficiency and through put of GB10(Go Back 10) flow control. (2+4+4)

4. (a) If Propagation time and Band width of a channel are 1 ms and 4Mbps then, in order to detect collision in CSMA/CD access control method, what would be minimum packet size to send? (b) What is the advantage of Token Passing Access Control method over CSMA/CD? (c) Derive the formula to calculate the efficiency for Token Passing Access Control method. (4+2+4)

5. (a) Given the data-word 1010011010 and the divisor 10111, show the generation of the code word at the sender side and also show the checking of the CRC code word at receiver side (assume no error in code word during transmission). (b) If the Data word is 10011010 then using hamming coding scheme derive Code word. Flip any one bit of received message and show how the scheme corrects it. (4+3+3)

6. (a) Describe in brief the main responsibilities of ISO-OSI Data Link Layer. (b) Given bits pattern is 101010. Encode the pattern using Differential Manchester Encoding. (c) What is the function of IANA? (5+3+2)

7. (a) What is the difference between Physical address and Logical address? (b) Can a MAC address be unique globally? (c) Which layer is responsible for Multiplexing and Demultiplexing? (d) If station A and station B are transmitting data over a medium and data collides first time. What is the probability that A transmit before B next time if they follow Exponential Back-off algorithm ? (3+1+1+5)

8. (a) In Distance Vector Routing algorithm F{6,3,4,1,2,0}, B{2,0,1,3,1,5}, A{0,1,2,3,2,5} and delay of F, B, A to D are 2,1,3. Calculate new vector table for D . (b) What is the disadvantage of Distance Vector Routing algorithm? Explain

with example. (c) Mention the differences between Routing and Flooding. (5+3+2)

9. (a) What could be the maximum and minimum TCP Header size? (b)What do you mean by Socket in Computer Network? Write its uses. (c) Write the Port address of Telnet. (d) In TCP if band width is 1 MBps then Calculate the Wrap around time (Number of bits in sequence field is 32). (3+3+1+3)

10. (a) Let X and Y are the total number of keys required for a set of 'n' individuals to communicate with each other using the secret key and the public key crypto system respectively. Find X and Y . (b) Encrypt the message "PGDCA EXAM" using shift cipher with a key -6.(c) Consider a plain text message '6' . Use RSA algorithm and two prime numbers 11 and 3. Now compute private key, public key and cipher text message.
(2+2+6)

11. (a) What do you mean by Switching? (b) What are the differences between virtual circuit switching and Data gram switching? (c) Define the capacity of a link. Bandwidth and Propagation time of a fulduplex link are 5MBps and 6ms respectively. Calculate the capacity of the link. (2+3+2+3)

Indian Statistical Institute

Course Name : PGDCA

Second Semester Examination (2016-17)

Paper Name: Software Engineering (203)

Time: 3 Hours 30 minutes Marks: 100 Date: 02.05.2017

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

1. (a) Differentiate between white-box testing and black-box testing.

(b) Differentiate between alpha testing, beta testing and acceptance testing.

(c) Why stubs and drivers are needed in unit testing?

(d) Design the equivalence class test cases for a program that reads two pairs (m1,c1) and (m2,c2) defining two straight lines of the form y=mx+c. The program computes the intersection point of the two straight lines and displays the point of intersection.

(e) Draw the CFG(Control Flow Graph) of the following program and subsequently find the cyclomatic complexity.

```
int compute_gcd(int x, int y) {
1. while(x!=y) {
2.          if(x>y) then
3.              x=x-y;
4.          else    y=y-x;
5.              }
6.          return x;
                         }
```

(f) "Branch coverage is stronger testing strategy then Statement coverage" justify the statement using an example.

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

2. (a) What are the shortcomings of classical waterfall model? What do you mean by phase containment of error?

   (b) Why spiral model is called a meta model?

   (c) Why we need software configuration management? Explain briefly the Reserve and Restore operation in configuration control.

   (d) Explain briefly the risk assessment phase of risk management. What is risk leverage?

   (e) Let us consider a satellite based mobile communication product. Identify the type of risk in each of the following cases:

   (i) What if the project cost escalates to a large extent than what was estimated?

   (ii) What if the mobile phones become too large for people to conveniently carry?

   (iii) What if hand off between satellites become too difficult to implement?

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

3. (a) Assume that the size of an organic type software product has been estimated to be 32000 LOC. Assume that the average salary of software developers is Rs. 15,000/- per month. Determine the effort required to develop the product, development time and cost to develop the product (Follow Basic COCOMO).

   (b) What are the shortcomings of LOC as project size estimation technique?

   (c) Let us consider the following C program:

```
main()
 {
   int a,b,c,avg;
   scanf(“%d %d %d”,&a,&b,&c);
   avg=(a+b+c)/3;
   printf(“avg=%d”,avg);
 }
```

Find out the estimated length and volume of the above program using Halstead Software Science.

(d) The following table indicates the various tasks involved in completing a software project, the corresponding activities and the estimated effort for each task in person-months(PM):

Notation	Activity	Effort in PM
T1	Requirement specification	1
T2	Design	2
T3	Code actuator interface module	2
T4	Code sensor interface module	5
T5	Code user interface part	3
T6	Code control processing part	1
T7	Integrate and Test	6
T8	Write user manual	3

The precedence relation Ti<{Tj,Tk} implies that the task Ti must complete before either task Tj or task Tk can start.The following precedence relation is known to hold among different tasks:

T1< T2<{T3,T4,T5,T6}<T7

T1<T8

Draw the

(i) Activity Network Diagram showing all the Critical paths.

(ii) Gantt Chart representation for the project.

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

4. (a) What are the characteristics of a good software design?

(b) What do you mean by functionally independent modules? Differentiate between Fan-in and Fan-out in layered arrangement of modules.

(c) What is coupling? Enumerate the different types of coupling that might exist between two modules(Give examples of each).

(d) "Today's robustness is tomorrow's requirement"- justify the statement.

(e) What are the different types of maintenance performed in the life cycle of a software product?

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

5. (a) A software needs to be developed for a Supermarket. First of all,customers first need to register themselves by giving their name, address and telephone number and PAN number. After registration, each customer will get a unique id and the system will generate a message that the customer is successfully registered. If some information has not been entered by the customer, then the system displays a prompt to enter the missing value. If an already registered user want to register, then the system displays a message that the customer is already registered (by matching the PAN number of customers, which is mandatory field at the time of registration). With all things purchased, customers come to the cash and delivery section, where the sales clerks of the Supermarket enter the purchase details and id of the customers and the system will display a message of having successfully registered the sale. The manager of the Supermarket select a winner of a prize scheme by a lottery (by randomly selecting one unique id) and give gold coins to him.

Draw a Use-Case diagram using UML syntax with mainline sequence and alternate sequences (if any) for the above system.

(b) What are the ways we can factor out the commonalities among Use Cases? Explain each with example.

(c) How does activity diagram differ from a flow chart?

(d) A software system needs to be developed for student admission procedure in an undergraduate institution. The academic section of the institution will first check the student records by verifying their marks in 10 and 10+2 and also by checking their date of birth. Then academic section will generate a unique student id and as per that student id, the accounts section will receive the fees. Also, students has to come to hostel office of the institution where they have to submit the hostel fees and hostel office will allot them hostel room. The hospital associated with the institution will conduct medical examination and store hospital record (that includes blood pressure, sugar, eyesight, physical disability and any other past disease.). The department section will allow students to register in their specific courses. The activities of hostel office, hospital and department can be done in parallel. After completion of all these activities, the academic section will issue identity card to the students.

Draw an Activity diagram using UML syntax for the above system.

(e) What are the different views supported by the diagrams in UML?

[9+6+3+10+2=30]

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

Indian Statistical Institute

Course Name : PGDCA

Second Semester Examination (2016-17)

Paper Name: Object Oriented Programming (201)

Time: 3 Hours and 30 minutes **Marks: 100** **Date: 06.05.2017**

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

1. (a) How encapsulation supports data hiding? Explain with a program.

   (c) What do you mean by singleton class in C++?

   (d) Write a class to represents a vector. Each vector has three components i, j, k and each component should have an integer value associated with it. Include the member functions to perform the following task:

   i) To create the vector.

   ii) To multiply the vector with a scalar value.

   iii) To print the vector in the form (i, j, k).

   Write a C++ program to test your class.

   [5+2+7=14]

2. (a) When and why we need to make a function inline?

   (b) Create two classes DM and DB to store the value of distances. DM stores distances in meter and centimeter and DB in feet and inch.

   Write a program that can read values for the class objects and add one object of DM with another object of DB. Use a friend function in C++ to carry out the addition operation. The object that stores the result may be a DM object or DB object, depending on the units in which the results are required. The display should be in the format of feet and inch or meter and centimeter depending on the object on display.

(c) Why static data member is called a class variable?

[(2+2)+9+2=15]

3. (a) Differentiate between function overloading and function overriding.

(b) How run time polymorphism is achieved using virtual function in C++? Explain with a program.

(c) We cannot have virtual constructors but we can have virtual destructors in C++. Why?

(d) When is a friend function compulsory to overload an operator? Give an example.

(e) Create a class **complex** with two integer data members real and imag. Write a program in C++ to overload the subtraction(-) operator to subtract two complex numbers (your operator overloading function should be a member function of the complex class).

[4+5+3+2+6=20]

4. (a) What is copy constructor? Why is a reference to an existing object passed as an argument to the copy constructor in C++ instead of passing it by value?

(b) Differentiate between private, public and protected access specifier/visibility modes of members of a class in C++.

(c) When and why do we need to make a base class virtual?

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

5. (a) A template can be considered as a kind of macro. Then, what is the difference between template and macro?

(b) Write a program in C++ using function template to perform linear search of a set of integer numbers and float numbers.

(c) In C++,both ios::app and ios::ate place the file pointer at the end of the file(when it is opened).What then, is the difference between them?

[2+6+2=10]

6. (a) What is the advantage of bytecode in Java?

(b) How command line arguments are passed in Java? Explain with a program.

(c) What is garbage collection?

(d) Why do we need to call super inside a derived class constructor in Java?

[2+5+3+2=12]

7. (a) Differentiate between a class and an interface in Java.

(b) "Java does not support multiple inheritance but via interface we can get the flavor of multiple inheritance"-justify the statement using a program.

(c) What are the differences between String class and StringBuffer class in Java? Write a program in Java that will replace all occurrences of a word in a sentence with another word. [3+5+(2+4)=14]

8. (a) Differentiate between throw, throws and finally.

(b) What will happen if all the user defined catch blocks in your program in Java failed to catch the exception occurred?

(c) Which are the two ways by which we can create child threads in Java?

(d) Write a program in Java that will correctly implement the producer-consumer problem using inter-thread communication.

[3+2+2+8=15]

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

Indian Statistical Institute

Semester-2 PGDCA 2016-2017

End-semester Examination (8th May, 2017)

Subject: Web Technology (202)

Total Marks: 110 Maximum marks: 100 Duration:180 mins

1. (a) What do you mean by Web Technology? (b) What are the types of web pages in term of its contents? Explain dynamic web pages.(c) Write a short note on Universal Resource Locator(url).

   (2+3+5)

2. (a) Write an HTML code to use image as a link to an another HTML page. (b) What is the use of 'div' tag in HTML?(c) What is the advantage of External Style Sheet over embedded styles in terms of web page development?

   (3+2+5)

3. (a) What are the differences between HTML and dynamic HTML? (b) How XML is different from HTML?(c) Describe 3-tier Client/Server architecture with diagram.

   (2+3+5)

4. (a) Write the differences between HTTP GET and HTTP POST methods. (b)What is the significance of "for...in" statement in JavaScript? Give example.(c) Write a program in JavaScript to validate an email address. (2+3+5)

5. (a) What types of pop-up boxes can JavaScript create? Explain with example. (b) What do you mean by an event in JavaScript? (c) What is class selector? Give proper example.

   (5+2+3)

6. (a) Write a short note on JavaBean. (b) Write a Java Bean class, which will calculate the area and perimeter of a rectangle given its length and breadth as input. Then write a JSP code which will access the JavaBean Class to calculate the area and perimeter of a rectangle given its length and breadth. (4+6)

7. (a) Describe the Servlet Application Architecture with diagram (b) What is the importance of 'Action' tag in JSP? Give proper example. (c) What do you mean by an Object in JSP? Explain with proper example. (4+3+3)

8. (a) Write a brief note on server-side scripting.
(b) Explain the Architecture of JSP with diagram. (5+5)

9. (a) Write a JSP program to get all parameters from client input and show them in a table. (b) Write a short note on web container. (6+4)

10. Write a JSP program to store the roll_number, name and age of the students in an institute as a part of a Students database and retrieve any record by a particular roll_number.

10

11. (a) Explain the JDBC steps to establish a connection with the database . (b) Write a short note on Apache Tomcat server. (c) What do you mean by JDBC driver?

(5+3+2)