BStat III [2010 - 2011]
Course on Linear Models **Mid-Semester Examination**
Exam Date : Monday, August 30, 2010
Full Marks : 70 **Time : 3 hours** **Instructor : B K Sinha**

NOTE : ALL QUESTIONS ARE COMPULSORY.

Q1. Consider the following expressions for the model expectations of uncorrelated observable random variables [Y_1, Y_2, Y_3, Y_4, Y_5] with the same unknown variance sigma^2 :

E[Y_1]= alpha + gamma; E[Y_2] = alpha + delta;
E[Y_3] = beta + gamma; E[Y_4] = beta + delta = E[Y_5]

All the parameters involved in the model expectations are unknown.

(i) Recast the above in the form of a linear model [Y, XTheta, sigma^2 I] and explain the notations used.
(ii) Write down explicit form of the matrix X and determine its rank.
(iii) Find a pair of uncorrelated error functions arising out of the above model.
(iv) Identify ALL unbiasedly estimable linear parametric functions.
(v) Find an unbiased estimator for sigma^2.
(vi) Explain how you would carry out a test for H_0: [alpha = beta, gamma = delta].

[(2+3) + (2+3) + 5 + 5 + 5 + 5= 30]

Q2. Consider the standard linear model [Y, X Beta, sigma^2 I] where X is a rectangular matrix having full column rank and all model parameters are unknown.

(i) Develop a canonical version of the model.
(ii) Show that every BLUE is necessarily uncorrelated with every error function.
(iii) Show that every BLUE is necessarily unique.

[5 + 5 + 5 = 15]

Q3. (a) Define Sum of Squares [SS] wrt a standard linear model set-up. Explain the notion of Orthogonal Decomposition of the Total SS.
(b) Consider the following model :

E[X]= alpha + gamma; E[Y] = alpha + delta;
E[S] = beta + gamma = E[T]; E[U] = beta + delta = E[V]

where the observable random variables have the same variance and are uncorrelated with each other.
Show that the Total SS i.e., X^2 + Y^2 + S^2 + T^2 + U^2 + V^2 has an Orthogonal Decomposition into FOUR components. Identify these components and give their interpretation.

[(5 + 5) + (10 + 3 + 2) = 25]

INDIAN STATISTICAL INSTITUTE
Mid-Semester Examination : 2010-2011
B.Stat. III Year
Sample Surveys

Date : 01.09.2010 Maximum Marks : 100 Duration : 3 Hours

Use two separate answer sheets for Group A and Group B

Group A (Total Marks : 50)

Notations are as usual.

1. (a) In SRSWR of size n out of N units, derive formulae for π_i and π_{ij}. Let s be an SRSWR of size n out of N units and $\gamma(s)$ denote the number of distinct units in s. Find $E(\gamma(s))$.

$1+3+2=6$

(b) Let $\bar{y}_n$ and $\bar{y}_d$ denote respectively the sample means based on all observations and based on distinct units obtained in an SRSWR of size n out of N units. Show that $\bar{y}_d$ is unbiased for the population mean $\bar{Y}$ and derive $V(\bar{y}_d)$. And also prove that $\bar{y}_d$ is better than $\bar{y}_n$ in estimating $\bar{Y}$.

$2+2+2=6$

2. (a) Explain the linear systematic sampling scheme and write down with proof an unbiased estimator of population total Y from such a sample as surveyed. Discuss why a circular systematic sampling scheme is needed instead.

$2+2+2=6$

(b) Explain why the variance of an estimator cannot be estimated unbiasedly from a single systematic sample.
Explain with proof how you may get over this problem.

$2+5=7$

3. (a) In estimating the population ratio $R=\frac{Y}{X}$ through simple random sampling, show that the bias of the ratio estimator $\hat{R}=\frac{\bar{y}}{\bar{x}}$ is $-\frac{Cov(\hat{R},\bar{x})}{\bar{X}}$.

In SRSWOR, show that an approximate expression for mean square error of $\hat{R}$ is $M_1(\hat{R})=(\frac{N-n}{Nn})\frac{S_y^2+R^2S_x^2-2R\rho S_xS_y}{\bar{X}^2}$, $S_y^2=\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{Y})^2$ and ρ is the correlation coefficient in between y and x.

Show that in SRSWOR, for estimating the population total Y, and if R is +ve, the usual ratio estimation method will be better than the usual simple average estimation method if $\rho > \frac{CV(x)}{2CV(y)}$, where CV(x) denotes the coefficient of variation of x.

2 + 4 + 4 = 10

(b) (i) Derive an exactly unbiased estimator for population ratio R based on SRSWOR. 5

(ii) The following figures of expenditures (Rs.) relate to a group of 10 households.

Serial no.	HH size	Exp. (Rs.) last month
1	7	3470.35
2	6	2716.80
3	5	1873.75
4	4	1693.20
5	3	1393.55
6	6	2198.74
7	2	3178.35
8	5	2708.75
9	6	1873.60
10	4	2175.80

Take an SRSWOR of 4 households and give an exactly unbiased estimate of the per capita last month's expenses of these 10 households.

10

Random number table

25	19	17	50	50	46	26	92	62	41	27	66	85	60	70
54	61	41	41	91	88	83	30	32	75	59	03	58	58	83
97	50	71	35	65	67	15	45	73	09	17	60	68	38	05
96	17	27	35	82	80	77	28	97	11	26	72	02	88	96
21	48	84	49	72	93	48	66	75	82	36	33	77	97	35
85	12	09	36	72	81	06	73	04	02	03	10	81	34	44
49	57	40	54	64	88	97	69	03	12	94	45	86	74	66
07	43	79	37	60	96	75	39	46	33	42	41	29	83	73
80	07	51	15	59	55	24	80	49	12	61	68	00	44	58
40	71	81	93	03	03	60	02	42	53	38	35	05	67	73

Group B (Total Marks : 50)

Answer ANY TWO questions . Marks allotted to each question are given within the parentheses . Standard notations and symbols are used .

1. **The frequency distribution of 232 cities in a country by population size is given in the following table .Calculate the RSE of the estimator of the total population Y (i) when a sample of 50 cities is selected with SRSWOR and (ii) the two largest cities are definitely included in the survey and only 48 cities are drawn from the remaining 230 cities with SRSWOR .**

FREQUENCY DISTRIBUTION OF 232 CITIES BY POPULATION SIZE (000)

population size class	no. of cities	population size class	no. of cities	population size class	no. of cities
50 – 75	81	500 – 550	2	1800 – 1850	1
75 - 100	45	550 – 600	3	1850 – 1950	0
100 – 150	42	600 – 650	1	1950 – 2000	1
150 - 200	14	650 – 700	1	2000 – 2050	0
200 – 250	9	700 – 750	0	2050 – 2100	1
250 – 300	5	750 – 800	1	2100 – 3600	0
300 – 350	6	800 - 850	2	3600 – 3650	1
350 – 400	5	850 - 900	1	3650 – 7850	0
400 – 450	5	900 – 950	2	7850 – 7900	1
450 – 500	2	950 – 1000	0		

(10+15)=[25]

2. **Obtain an estimate of the percentage (P) of unemployed persons in a large city using the data given in the following table and estimate its RSE .**

ESTIMATES OF P BASED ON TWO SIMPLE RANDOM SAMPLES SELECTED WITH REPLACEMENT

Sample Number	Sample size No. of persons	Percent unemployed
1	2345	19.678
2	1789	20.123

(10+15)=[25]

3. **A survey is to be conducted for estimating the total number of literate persons in a town having three communities , some particulars of which are given in the following table based on the results of a pilot study .**

A ROUGH IDEA OF THE TOTAL NUMBER OF PERSONS AND PROPORTIONS OF LITERATE PERSONS

Community	Total number of persons	Percentage of literate persons
1	60,000	40
2	10,000	80
3	30.000	60

(a) Treating the communities as strata and assuming SRSWR in each stratum , allocate a total sample size of 2000 persons to the strata in an optimum manner for estimating the overall proportion of literate persons in the town .

(b) Estimate the efficiency of stratification as compared to unstratified sampling . (10+15)=[25]

4. A sample of 10 villages was drawn from a tehsil with PPSWR , size being the 1951 census population and the relevant data are given in the following table .

1951 CENSUS POPULATION (x) AND CULTIVATED AREA (y) IN ACRES FOR 10 SAMPLE VILLAGES

Village	x	y	Village	x	y
1	5511	4824	6	7357	5506
2	865	924	7	5131	4051
3	2535	1948	8	4654	4060
4	3523	3013	9	1146	809
5	8368	7678	10	1165	1013

Total population of the tehsil in 1951 = 415149

(a) Estimate the total cultivated area Y and its RSE .

(b) Obtain the sample size required to ensure an RSE of 2 % . (15 +10) =[25]

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination: (2009–2010)

B. Stat Third Year

Statistical Inference I

Date: 8/09/2010 Marks: 50 Duration: 2 hours

Attempt all questions

1. Prove or disprove the following statement: "A finite collection of exchangeable random variables must be conditionally *iid*". [5]

2. (a) Express the family of binomial distributions in exponential family form. [4]

   (b) Find the natural parameter and the natural parameter space. [3]

   (c) Find a complete sufficient statistic. [1]

3. Let $X_1, \ldots, X_n$ be conditionally *iid* Unif$(1-\theta, 1+\theta)$ given $\Theta = \theta$.

   (a) Find a one dimensional minimal sufficient statistic for θ. [4]

   (b) Show that $T = (X_{(1)}, X_{(n)})$ is not a complete sufficient statistic. [4]

4. Let $X_1, \ldots, X_n$ be *iid* $N(\mu, \gamma_0^2\mu^2)$ given $\Theta = \mu$, where γ_0 is known and $\mu > 0$.

   (a) Find a minimal sufficient statistic for μ. [5]

   (b) Find
$$E_\mu\left[\frac{n+\gamma_0^2}{1+\gamma_0^2}\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right].$$
[5]

   (c) Is the minimal sufficient statistic found in (a) complete? Justify. [2]

5. Let X_1 and X_2 be conditionally *iid* Poisson(λ) given $\Theta = \lambda$. Is $T = X_1 + 2X_2$ a sufficient statistic? Prove or disprove. [3]

6. Let $X = (X_1, \ldots, X_n)$ be conditionally *iid* Exp(θ) given $\Theta = \theta$, i.e.,
$$f_{X_i|\Theta}(x_i|\theta) = \frac{1}{\theta}\exp(-x_i/\theta), \quad 1 \leq i \leq n.$$

   (a) Find a minimal and complete sufficient statistic T. [3]

(b) Find the expectation of

$$g(X) = \frac{X_n}{\sum_{i=1}^{n} X_i}.$$

[6]

7. Consider a meter that is trying to measure a quantity Θ. Suppose that the meter gives a reading Z, which has $N(\theta, 1)$ distribution given $\Theta = \theta$ if $Z < 2$, but if $Z \geq 2$, the reading is always 2. Let $X = \min\{Z, 2\}$ be the reading. Letting Θ have a conjugate prior of the form $\Theta \sim N(\theta_0, \sigma_0^2)$ for known values of θ_0 and σ_0^2, obtain the posterior distribution of Θ given X.

[5]

INDIAN STATISTICAL INSTITUTE
Mid-Semester Examination : 2010-11(First Semester)
Course Name: **B.Stat III**

Subject Name : **Introduction to Anthropology and Human Genetics**
Date : 9 September 2010 Maximum Marks : 80 Duration : 3 Hours

Answer any **five** (5) questions from the following, all questions carry equal marks

1. How do you define Anthropology? What are the distinctive features of Anthropology?

2. How does Physical Anthropology differ from Biological Anthropology? What are the areas of Biological anthropology?

3. What are the characteristic features of mammals?

4. Why is man unique in the animal kingdom?

5. What are the anatomical changes that have taken place due to the assumption of erect posture?

6. Describe normal karyotypes in man. How Klinefelter syndrome develops?

7. How Huntington's disease occurs? Illustrate the criteria of inheritance of the disease?

8. Short notes on (write any **two**):

   a. Primate
   b. Holism
   c. Turner syndrome
   d. Law of independent assortment

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination (2010-2011)

Subject: Introduction to Sociology
B.Stat. III Year

Date: 9.9.10 Maximum Marks:**25** Duration: Two hours

The figures in the margin indicate full marks

Q 1. Answer any two of the following questions: 5 x 2 = 10

a) Define Sociology. How can Sociology be scientific when the laboratory method can seldom or never be used?

b) Examine Sociology's relation with other social sciences.

c) What is meant by social movements?

Q 2. Write short notes on any two: 4 x 2 =8

a) Globalization.

b) Institution

c) Agrarian society

Q 3. Choose the correct answers: 1 x 7 =7

(a) In which of the following did the class structure develop first?
(i) Tribal Society
(ii) Agricultural Society
(iii) Industrial Society

(b) Who is the father of Sociology?
(i) August Comte
(ii) Ramkrishna Mukherjee
(iii) Darwin

(c) Introduction to Sociology was written by:
(i) Arnold
(ii) A.R.Desai
(iii) Gillin

(d) "The quality of life" was written by:
(i) Ramkrishna Mukherjee
(ii) M.N.Srinivas
(iii) Andre Beteille

(e) The villagers maintain their general behaviouristic patterns because:
(i) They are trained by their families to do so
(ii) They are more cultured
(iii) They generally follow traditions, customs and folkways more resolutely

(f) "Hunting and "Food-gathering" were the primary pursuits of most of the tribesmen in the primitive ages because:
(i) People were ferocious and hunting and food gathering were their past times
(ii) People did not know any other alternative to earn their living
(iii) None of these

(g) "Village is the oldest and permanent community of man", are the words spoken By:
(i) Ogburn and Nimkoff
(ii) Anthony Giddens
(iii) Andre Beteille

Indian Statistical Institute

Mid Semester Examination: (2010–2011)

B.Stat.(Hons.) – III year

Economics III

Date: 11/09/2010 *Maximum Marks –50* *Duration: 2 hours*

Answer ***any two*** questions.

1. (a) State the assumptions of the Classical Linear Regression Model (CLRM) in a multiple regression set up.

(b) Show that the Ordinary Least Squares (OLS) estimator of the parameters is consistent and Best Linear Unbiased Estimator (BLUE).

(c) Consider the model

$$y_i = \beta x_i + \varepsilon_i$$
$$\varepsilon_i = u_i + \delta u_{i-1}, \qquad |\delta| < 1$$
$$E(u_i) = 0$$
$$E(u_i u_j) = \begin{cases} \sigma_u^2, i = j \\ 0, i \neq j \end{cases}$$

(i) Show that the OLS estimator of β is unbiased.

(ii) Assuming δ and σ_u^2 are known, write down the expression for the GLS estimator of β.

[6+9+10=25]

2. (a) What is multicollinearity?

(b) Describe the method of Principal Component Regression in this context.

(c) Describe the Ridge regression technique and show that $Var(\hat{\beta}) - Var(\hat{\beta}_R)$ is positive semidefinite, where $\hat{\beta}$ and $\hat{\beta}_R$ are the OLS and Ridge estimators, respectively.

[3+12+10=25]

3. (a) What do you mean by 'Dummy variables'?

(b) Suppose

$$y = \begin{cases} \alpha_1 + \beta_1 x + \lambda_1 z + \varepsilon & for \quad period \quad 1 \\ \\ \alpha_2 + \beta_2 x + \lambda_2 z + \varepsilon & for \quad period \quad 2 \end{cases}$$

Using dummy variables, how would you test the hypothesis that only the coefficient of x changes between the two periods?

(c) Suppose consumption expenditure y depends on income x, but different levels of income produce different values of marginal propensities to consume (β_i). In particular,

$$\frac{dE(y|x)}{dx} = \begin{cases} \beta_1 \; if \; x < Rs.100 \\ \beta_2 \; if \; Rs.100 \le x < Rs.500 \\ \beta_3 \; if \; x \ge Rs.500 \end{cases}$$

Using Dummy variables estimate the marginal propensity to consume for the three groups (assuming that there is no intercept term) given that $\sum yx$ is 30000; 5,00,000 and 10,00,000 for the three groups, respectively, and $\sum x^2$ is 90000; 25,00,000 and 100,00,000 for the three groups, respectively.

(d) Suppose you want to test r, $r \le k$, independent linear restrictions of the form $R\beta = d_{r\times 1}$,

where $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{pmatrix}$ in $y_{n\times 1} = X\beta + \varepsilon_{n\times 1}$.

Write down R and d to incorporate the following cases: (k=4)

(i) $\beta_1 = \beta_2 = \beta_3 = 0$

(ii) $\beta_1 = \beta_2$ and $\beta_3 = \beta_4$

(iii) $\beta_1 - 3\beta_2 = 5\beta_3$

(iv) $\beta_1 + 3\beta_2 = 0$

[2+5+10+8=25]

INDIAN STATISTICAL INSTITUTE
Mid-Semester Examination: 2010-11(First Semester)

Course Name: OPTIONAL COURSE FOR B III

Subject Name : GEOLOGY

Date : 09.09.2010 **Maximum Marks :40** **Duration : 1hr 30 min**

<u>Answer any three (3) questions from question number 1 to 4</u>

1. **A.** The weight % of Fe in Earth's crust is rather small (nearly 5) and it is the fourth element in the decreasing order of abundance. Why it becomes the most abundant element when we consider the whole earth instead of the crust only. --------- 3

B. What are "Chondrites" and "Achondrites"? ------- 5

C. Name one acid and one basic igneous rock. -- 2

2. **A.** Define "Orthorhombic" crystal system. Name one mineral whose crystals generally belong to "Orthorhombic" system. ------------------2+1 = 3

B. Compare "Triclinic" and "Monoclinic" systems of crystals. ------ 3

C. Draw a trigonal pyramid. -------------------- 2

D. Miller indices of a crystal face are 110; which crystallographic axis is not intercepted by that face? If the face is made to intercept all the three axes with unit lengths, what will be its new Miller indices? ------------ 1+1 = 2

3. What is a mineral? Name one mineral each from the Feldspar and Mica groups. Define "Phyllosilicates" and "Tectosilicates". -------- 3+2+5 = 10

4. Write a note on the factors controlling the texture of a rock. ---------- 10

---

5. Laboratory assignments (done at the GSU Lab) -------------- 10

BStat III [Session : July – November, 2010]
Linear Statistical Models
End-Sem Exam **Full Marks : 100 [Weight : 50%]**
Time : 3 ½ Hours

19/11/

Note : Answer Q1 and any THREE of the rest.

Q1. Consider linear regression of Y on a non-stochastic regressor X with homoscedastic errors : $E(Y \mid x) = \alpha + \beta x$.

You are given the following 'regression design' in terms of a choice of x-values and their 'repeat numbers' :

x :	3	6	8	11	13
n_x :	1	2	8	4	5

(a) Work out an explicit expression for the information matrix for the 'natural' parameters [α and β] in the model.

(b) Make a change in the regressor-values by setting $Z = (X - 8) / 5$. Write down the linear regression model in terms of the 'transformed' regressor Z in the form : $E[Y \mid z] = \gamma + \delta z$.
Express α and β in terms of γ and δ.

(c) Work out an explicit expression for the information matrix for the 'derived parameters' γ and δ.

(d) Establish that a modified regression design of the form :

x* :	3	5	6	8	10	11	13
n_{x*} :	3	2	1	8	1	2	3

would maximize the determinant of the information matrix for the natural parameters, given the total number of observations as 20.

(e) Interpret your finding in (d).

[5 + (2 + 3) + 5 + 8 + 2 = 25]

Q2. Consider a standard linear model [$Y, X\beta, \sigma^2 I$] where X is a rectangular matrix, not necessarily possessing full column rank and all model parameters are unknown.

(i) Develop a canonical version of the model. Illustrate with a small example.

(ii) Explicitly establish that any LUE of an estimable linear parametric function, if not already the BLUE, can be converted to the BLUE, by solving a derived system of linear equations. Illustrate with a small example.

(iii) Show that the BLUE of any estimable linear parametric function is necessarily unique in terms of the observable random variables.

[(5 + 5) + (5 + 5) + 5 = 25]

Q3. (a) Write down explicitly the model for two-way classified data with multiple but equal number of observations per 'cell'. Consider both the cases of : 'Without Interaction Terms' and 'With Interaction Terms'.

(b) For both the cases, display the ANOVA Table.

(c) How would you modify the ANOVA Table for a model without any interaction term, in case all the cells possess 'proportional frequencies' ?

(d) For any component source of variation of your choice, work out Expected Value of the Corresponding Sum of Squares for the model in (c) above.

[(2 + 3) + (3 + 5) + 5 + 7 = 25]

Q4. Consider the following model :

E[X]= alpha + gamma; E[Y] = alpha + delta;
E[S_1] = E[S_2] = E[S_3] = beta + gamma;
E[U_1] = E[U_2] = E[U_3] = beta + delta

where the observable random variables X, Y, S_1, S_2, S_3, U_1, U_2 and U_3 have the same variance and are uncorrelated with each other.
Show that the Total SS i.e., $X^2 + Y^2 + \Sigma S^2$_i + ΣU^2_i has an Orthogonal Decomposition into FOUR components. Identify these components and give their interpretation.

[18 + 5 + 2 = 25]

Q5. (a) Write down the ANOCOVA Model wrt One-Way classified data and a single non-stochastic covariate X. Derive conditions for estimability of the 'regression coefficient' and also suggest a test for its significance.

(b) Prepare the ANOCOVA Table for the above set-up in (a). Display expressions for different Sum of Squares and derive an expression for E[Sum of Squares due to Error].

[(5 + 5) + (5 + 5 + 5) = 25]

10/11/2010

INDIAN STATISTICAL INSTITUTE

First-Semester Examination : 2010-2011

B.Stat. III Year

Sample Surveys

Date : 23.11.2010 Maximum Marks : 100 Duration : 3 Hours

Use two separate answer sheets for Group A and Group B

Group A (Total Marks : 50)

Notations are as usual.

Answer any 3 from Q.1 to Q.5 and Q.6 is compulsory.

1. (a) In SRSWOR of size n out of N units, derive the expressions of π_i and π_{ij}. Also prove that $Cov(\bar{x},\bar{y}) = \left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{X})(y_i-\bar{Y})$. 1 + 1 + 6 = 8

   (b) Let $\bar{y}_d$ denote the sample mean based on distinct units obtained in an SRSWR of size n out of N units. Find $V(\bar{y}_d)$. Suggest with proof one unbiased estimator for $V(\bar{y}_d)$ based on distinct units only. 2 + 2 = 4

2. (a) In linear systematic sampling of size n from N = n.k units, where k is an integer, write down with proof an unbiased estimator $\hat{Y}$ of population total Y.

   (b) Explain the concept of Intra-class correlation coefficient ρ with proper derivation, for a variable of interest, say y, for L classes and each class having M units.

   (c) Prove that in above systematic sampling, $V(\hat{Y}) = \frac{N}{n}(N-1)S_y^2[1+(n-1)\rho]$.

   (Use proper definition of ρ in this case).

   2 + 5 + 5 = 12

3. (a) In estimating the population total Y by ratio method of estimation with the auxiliary variable x, through SRSWOR, write down the classical ratio estimator.

   (b) Modify the sampling scheme so that this classical ratio estimator is unbiased.

   (c) Discuss how to estimate the variance of the classical ratio estimator for modified sampling scheme.

   2 + 5 + 5 = 12

4. (a) In cluster sampling of n clusters out of N clusters, i-th cluster having M_i units, derive an unbiased estimator $\hat{\bar{Y}}_C$ for population mean $\bar{Y}$, assuming that the clusters are selected by probability proportional to the number of units in them with replacement.

(b) Derive $V(\hat{\bar{Y}}_C)$.

(c) Show that $\frac{1}{n(n-1)}\sum_{i=1}^{n}(\bar{y}_i-\hat{\bar{Y}}_C)^2$ is an unbiased estimator for $V(\hat{\bar{Y}}_C)$, where $\bar{y}_i$ denotes the mean of y values in the i-th selected cluster.

2 + 5 + 5 = 12

5. (a) Discuss how to obtain an unbiased estimator $\hat{\bar{Y}}$ for population mean $\bar{Y}$ in SRSWOR of size n out of N units, following Hansen and Hurwitz 's technique of dealing with non-response errors.

   (b) Find variance of the estimator $\hat{\bar{Y}}$ in (a).

   (c) Find optimum sample size for a given value V_0 of $V(\hat{\bar{Y}})$ and a given cost function with rates defined as C_0 = cost per unit for 1st attempt of data collection, C_1 = cost of editing and processing data for per unit in response class and C_2 = cost of interviewing and processing data per unit of non-response class.

2 + 5 + 5 = 12

6. The following figures of expenditures (Rs.) relate to a group of 10 households.

Serial no.	Exp. (Rs.) last month
1	3470.35
2	2716.80
3	1873.75
4	1693.20
5	1393.55
6	2198.74
7	3178.35
8	2708.75
9	1873.60
10	2175.80

(a) Estimate the average monthly expenditure of this group of 10 households based on 2 independent circular systematic samples each of size 3.

Use the extract from a random number table given below.

(b) Give an unbiased estimate of the standard error of above estimator.

(c) Also give an estimate of the coefficient of variation in percentage.

5 + 5 + 4 = 14

Random number table

25	19	17	50	50	46	26	92	62	41	27	66	85	60	70
54	61	41	41	91	88	83	30	32	75	59	03	58	58	83
97	50	71	35	65	67	15	45	73	09	17	60	68	38	05
96	17	27	35	82	80	77	28	97	11	26	72	02	88	96
21	48	84	49	72	93	48	66	75	82	36	33	77	97	35
85	12	09	36	72	81	06	73	04	02	03	10	81	34	44
49	57	40	54	64	88	97	69	03	12	94	45	86	74	66
07	43	79	37	60	96	75	39	46	33	42	41	29	83	73
80	07	51	15	59	55	24	80	49	12	61	68	00	44	58

GROUP – B
(Maximum Marks =50)

Answer Question No. 4 and any one question from the rest . Marks allotted to each question are given within the parentheses . Standard notations and symbols are used .

1. (a) A sampler proposes to take a stratified random sample . He expects that his field cost will be of the form $\sum c_h n_h$. His advance estimates of relevant quantities for the two strata are as follows .

Stratum	W_h	S_h	c_h
1	0.4	10	\$4
2	0.6	20	\$9

(i) Find the values of $\frac{n_1}{n}$ and $\frac{n_2}{n}$ that minimize the total field cost for a given value of $\text{Var}(\bar{y}_{st})$.

Find the sample size required under this optimum allocation , to make $\text{Var}(\bar{y}_{st})=1$. Ignore the fpc .

(ii) How much will the total field cost be ?

(b) After the sample in (a) is taken , the sampler finds that his field costs were actually \$2 per unit in stratum 1 and \$12 in stratum 2 .

(i) How much greater is the field cost than anticipated ?

(ii) If he would have known the correct field costs in advance , could he have attained $\text{Var}(\bar{y}_{st})$ =1 for the original estimated field cost in (a) ?

(iii) If your answer to (ii) is 'no' , find the minimum field cost to reduce $\text{Var}(\bar{y}_{st})$ to 1 .

(5+5+5+5+2+3)=[25]

2. Suppose a population consists of N first stage units (f.s.u.'s) and the ith f.s.u. consists of M_i second stage units (s.s.u.'s) . Suppose a sample of f.s.u.'s is selected in n draws according to PPSWR method of sampling using p_i's as the normed size measures for the ith f.s.u. and each time a f.s.u. (say, the ith f.s.u.) is selected , a sample of m_i s.s.u.'s is selected by SRSWOR sampling scheme .
   (a) Obtain an unbiased estimator of the population total .
   (b) Derive an expression for the sampling variance of the proposed unbiased estimator .
   (c) Also obtain an unbiased estimator of the variance of the estimator of the population total .

(5+10+10) =[25]

3. If a sample is selected in 3 draws according to PPSWR sampling scheme , show that the following two estimators are unbiased for the population total Y :

(i) $$t(s,\mathbf{Y})=\begin{cases}\dfrac{y_i}{p_i} \text{ if } \quad v=1 \quad \textit{and} \quad s \quad \textit{contains} \quad 1 \quad \textit{distinct} \quad \textit{unit} \quad i \\ \dfrac{1}{3}\left[\dfrac{y_i}{p_i}+\dfrac{y_j}{p_j}+\dfrac{y_i+y_j}{p_i+p_j}\right] \text{if} \quad v=2 \quad \textit{and} \quad s \quad \textit{contains} \quad 2 \quad \textit{distinct} \quad \textit{units} \quad i,j \\ \dfrac{1}{3}\left[\dfrac{y_i}{p_i}+\dfrac{y_j}{p_j}+\dfrac{y_k}{p_k}\right] \text{if} \quad v=3 \quad \textit{and} \quad s \quad \textit{contains} \quad 3 \quad \textit{distinct} \quad \textit{units} \quad i,j,k\end{cases}$$

where v denotes the number of distinct units in s .

(ii) $$e(s,\mathbf{Y})=\sum_{i\in s}\frac{y_i}{1-(1-p_i)^3}$$

If $y_i \alpha\ p_i$, state giving reasons which one of the above two estimators you would prefer .

(15 +5 +5) =[25]

4 The following table gives the household size and information on whether the household took an agricultural loan from a bank for a random sample of 25 households selected from a population of 515 households in the locality by SRSWOR . Estimate (a) the proportion and the number of households in the locality having agricultural loan and (b) the total number of persons in those households having agricultural loan . Also estimate the relative standard errors of the estimates .

Hh Sl. No.	Whether households have agricultural loan	Hh size	Hh Sl. No.	Whether households have agricultural loan	Hh size
1	Y	5	14	N	6
2	N	7	15	N	10
3	Y	9	16	N	18
4	N	17	17	Y	5
5	Y	3	18	N	9
6	N	7	19	Y	8
7	N	11	20	N	12
8	Y	6	21	N	15
9	Y	8	22	Y	6
10	N	7	23	N	8
11	N	5	24	Y	11
12	Y	14	25	N	17
13	N	12			

Y : Yes , N : No

(10+5+10) = [25]

INDIAN STATISTICAL INSTITUTE

FIRST SEMESTER EXAMINATION, 2010-11

Campus: Kolkata

COURSE NAME : B-STAT III YEAR

SUBJECT NAME : Differential equations

Date : 26-11-2010 **Maximum Marks : 70** **Duration: 3 hours 30 min**

Answer the following questions.

1. Locate and classify all the singular points for finite values of t of the following differential equation :

$$t(1-t)y'' + [c - (a+b+1)t]y' - aby = 0 \ (a, b, c \ \textit{are constants}).$$

(10 marks)

2. By the method of power series, show that the solution ϕ of the differential equation

$$y'' - ty = 0$$

satisfying the initial conditions $\phi(0) = a$, $\phi'(0) = b$ is

$$\phi(t) = a\left[1 + \sum_{m=1}^{\infty} \frac{(1)(4)\ldots(3m-2)}{(3m)!}t^{3m}\right] + b\left[t + \sum_{m=1}^{\infty} \frac{(2)(5)\ldots(3m-1)}{(3m+1)!}t^{3m+1}\right].$$

(8 marks)

3. Obtain Green's function $G(t, s, \lambda)$, defined for $0 \leq t, s \leq \pi$, and $\lambda \neq n^2$ $(n = 1, 2, \ldots)$ for the non-homogeneous boundary value problem

$$y'' + \lambda y = f(t)$$

with boundary conditions $y(0) = 0$, $y(\pi) = 0$. (13 marks)

CONTINUED

4. (a) Find all real valued of solutions $y(t)$ of the differential equation

$$t\frac{dy}{dt} - y = t \quad (-1 < t < 1).$$

(7 marks)

(b) Find all solutions $y : \mathbb{R} \to \mathbb{R}$ to

$$\frac{dy}{dt} = \sqrt{y(y-2)},\ y(0) = 0.$$

(4 marks)

5. Find a function $y(t)$ such that $y^{(4)} + y = 0$ for $t \geq 0$, $y(0) = 0$, $y'(0) = 1$ and

$$\lim_{t\to\infty} y(t) = \lim_{t\to\infty} y'(t) = 0.$$

(7 marks)

6. Determine all real eigenvalues and corresponding eigenfunctions of each of the following boundary value problems.

(a) $y'' + \lambda y = 0$; $(y(0) = 0,\ y'(\pi) = 0)$. (6 marks)

(b) $y'' + y' + (\lambda + 1)y = 0$; $(y(0) = 0,\ y(\pi) = 0)$. (7 marks)

(c) $t^2y'' - \lambda ty' + \lambda y = 0$; $(y(1) = 0,\ y(2) - y'(2) = 0)$. (7 marks)

Note : If the eigenvalues are roots of a transcendental equation which cannot be solved explicitly, give the equation for the eigenvalues and the form of the eigenfunctions.

7. Show that for a Sturm-Liouville problem consisting of the following differential equation

$$L(y) = (p(t)y')' + q(t)y = \lambda r(t)y$$

where p, q *and* r are continuous on $a \leq t \leq b$, and $p(t) > 0$, $r(t) > 0$ on $a \leq t \leq b$, and the boundary conditions

$$\alpha y(a) + \beta y'(a) = 0, \quad \gamma y(b) + \delta y'(b) = 0$$

(where α, β, γ and δ are real constants), all eigenvalues are real and eigenvectors corresponding to different eigenvalues are orthogonal on $a \leq t \leq b$ with respect to the weight function $r(t)$. (13 marks)

INDIAN STATISTICAL INSTITUTE

Semestral Examination: (2009–2010)

B. Stat Third Year

Statistical Inference I

Date: 30/11/2010 Marks: ..100.. Duration: .4 hours.

Attempt all questions

1. Suppose that $\{X_n\}_{n=1}^{\infty}$ are bounded, exchangeable random variables. Let Θ $=\lim_{n\to\infty}\sum_{i=1}^{n} X_i/n$, almost surely. Prove that $Var\,(\Theta) = Cov\,(X_1, X_2)$. [8]

2. Suppose that for every $m = 1, 2, \ldots$,

$$f_{X_1,\ldots,X_m}(x_1,\ldots,x_m) = \frac{2}{(m+1)c_m(x_1,\ldots,x_m)^{m+1}}, \quad \text{if all} \quad x_i \geq 0$$

where $c_m(x_1,\ldots,x_m) = \max\{2, x_1, \ldots, x_m\}$.

(a) Prove that X_i are exchangeable and that these distributions are consistent. [3]

(b) Find the distribution of $Y_n = c_n(X_1,\ldots,X_n)$ and the limit of this distribution as $n \longrightarrow \infty$. [3]

(c) Find the conditional density of X_{n+1} given $X_1 = x_1, \ldots, X_n = x_n$, and assume that $\lim_{n\to\infty} c_n(x_1,\ldots,x_n) = \theta$. Find the limit of the conditional density as $n \longrightarrow \infty$. [3]

(d) Use DeFinetti's representation theorem to show that the prior (the answer to part (b)) and the likelihood (the asnwer to part (c)) combine to give the original joint distribution. [3]

3. Let $X_1, \ldots, X_n$ be *iid* with pmf

$$P(X = x \mid \Theta = \nu) = \frac{1}{x^{\nu}\zeta(\nu)},$$

where $\zeta(\nu) = \sum_{x=1}^{\infty} \frac{1}{x^{\nu}}$.

(a) Find a minimal sufficient statistic for ν. [4]

(b) Is the minimal sufficient statistic complete? Justify. [2]

(c) Find a sufficient statistic which is not minimal. Justify. [2]

4. Suppose that the posterior distribution of Θ is denoted as $F_{\Theta|X}(\theta)$. Define $L(\theta, \delta(x))$ for a nonrandomized rule as

$$L(\theta, \delta(x)) = a(\theta - \delta)I\{\theta \geq \delta\} + b(\delta - \theta)I\{\theta < \delta\},$$

where $a > 0$, $b > 0$ are known constant and $I(\cdot)$ is indicator function.

(a) Find δ that minimizes the posterior risk $r(\delta|x)$. [4]

(b) "If $a= b$ then δ is the posterior mean"-true or false. Justify. [2]

5. Suppose that $\{X_n\}_{n=1}^{\infty}$ are *iid* Ber(θ), $\aleph = \{a_0, a_1\}$, and

$$L(\theta, a) = \begin{cases} 0 & \text{if } (\theta \leq 1/2 \text{ and } a = a_0) \text{ or if } (\theta > 1/2 \text{ and } a = a_1) \\ 1 & \text{otherwise} \end{cases}$$

Let $X = (X_1, \ldots, X_n)$ and let n be even. Define $Y = \sum_{i=1}^{n} X_i$ and let the prior for θ be $U(0,1)$.

(a) If $Y = y$ successes are observed in n trials, then find the posterior of θ. [4]

(b) Find a randomized and a nonrandomized *Formal Bayes* rules and show that their posterior risks are equal. [6]

6. Let δ_0 be a randomized rule and T be a sufficient statistic. Define $\delta_1(t)(A)$ $=E\left(\delta_0(X)(A)|T = t\right)$.

(a) "$\delta_1(t)(A)$ will not depend on θ"- Is the statement correct? Justify. [2]

(b) Let $h : \aleph \to \mathbb{R}^+$ be a non negative simple function, prove that

$$E\left\{\int h(a)d\delta_0(X)(a)\middle|T = t\right\} = \int h(a)d\delta_1(t)(a).$$

[Hint: Any simple function $\phi(x)$ is written as $\phi(x) = \sum_{i=1}^{n} a_i I_{A_i}(x)$, where $A_i = \{x : \phi(x) = a_i\}$.] [4]

(c) Find the risk functions $R(\theta, \delta_1)$ and $R(\theta, \delta_0)$ and show that they are equal. [6]

7. (a) Suppose that Θ is k-dimensional and that $f_{X|\Theta}(x|\theta)$ is the density of X given $\Theta = \theta$. Also suppose that all the regularity conditions related to the Fisher Information hold. Denote $I_X(\theta)$ be the Fisher Information matrix and let $I_{X,i,j}(\theta)$ be the (i,j)th element of $I_X(\theta)$. Further, assume that two derivatives can be passed

under the integral sign. Given that θ, θ_0 are two elements of Ω (parameter space), let $\mathcal{I}_X(\theta_0;\theta)$ denote the Kullback-Leibler Information matrix. Show that

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\mathcal{I}_X(\theta_0;\theta)\bigg|_{\theta=\theta_0} = I_{X,i,j}(\theta_0).$$

[4]

(b) Let X have an exponential family of distributions, then show that

$$I_X(\theta) = -\left(\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log(c(\theta))\right)\right),$$

where $c(\theta)$ is the normalizing constant for the exponential family. [4]

8. Let Ω = (0,1) and let $\{X_n\}_{n=1}^{\infty}$ are *iid* Ber(θ). Suppose $X = (X_1,\ldots,X_n)$. Let $\xi > \theta$ and Θ be discrete with

$$f_\Theta(y) = \begin{cases} \pi_0 & \text{if } y = \theta \\ 1-\pi_0 & \text{if } y = \xi. \end{cases}$$

Then show that

$$P\left(\Theta = \xi \middle| X = x\right) = \left(1 + \frac{1-\pi_0}{\pi_0}\exp\{(\mathcal{I}_X(p_n;\theta) - \mathcal{I}_X(p_n;\xi))\,n\}\right)^{-1},$$

where $x = \sum_{i=1}^n x_i$, $p_n = x/n$, and $\mathcal{I}_X(p_n;\theta)$, $\mathcal{I}_X(p_n;\xi)$ are Kullback-Leibler information of p_n & θ and p_n & ξ, respectively. [8]

9. Let $\Omega = (0,\infty)\times(0,\infty)$, $\chi = \mathbb{R}^3$, and $\aleph = \mathbb{R}^+$. Suppose that X_1, X_2, X_3 are *iid* U(α,β), where $\theta = (\alpha,\beta)$. Let

$$L(\theta,a) = \left(\frac{\alpha+\beta}{2} - a\right)^2,$$

and $\delta_0 = \bar{X}$.

(a) Let T be a 2-dimensional sufficient statistic for (α,β). Find a rule $\delta_1(T)$ whose risk function is as good as δ_0. [4]

(b) Find the risk functions $R(\theta,\delta_0)$ and $R(\theta,\delta_1)$. [4]

10. Let $X = (X_1,\ldots,X_n)$, where $X_i \overset{iid}{\sim} N(\theta,1)$. Consider testing $H : \Theta = \theta_0$ versus $A : \Theta \neq \theta_0$. Let $0 < \pi_0 = Pr\,(\Theta = \theta_0 \mid H) < 1$ and let g be a uniform prior density over $(\theta_0 - c, \theta_0 + c)$ for some $c > 0$, and let x be such that $\bar{x}_n = \sum_{i=1}^n x_i/n = \theta_0 + \frac{2\tau_{\alpha/2}}{n}$, where $\tau_{\alpha/2}$ is the upper $\alpha/2$ point of $N(0,1)$. Then, using the 0-1 loss function compare classical two-sided size α and Bayes tests for all such $\bar{x}_n$. [10]

11. Let the parameter space be the open interval $\Omega = (0, 100)$. Let X_1 and X_2 be conditionally independent given $\Theta = \theta$ with $X_1 \sim Poisson(\theta)$ and $X_2 \sim Poisson(100 - \theta)$. We are interested in the hypothesis $H : \Theta \leq c$ versus $A : \Theta > c$.

   (a) Show that there is no UMP level α test of H versus A. [2]

   (b) Is there any UMP level α test conditional on the ancillary $T = X_1 + X_2$? [3]

   (c) Find a prior distribution for Θ such that the conditional UMP level α test given T is to reject H if $Pr\,(H \text{ is true} \mid X_1 = x_1, X_2 = x_2) < \alpha$. [5]

BStat III [2010 - 2011]
Course on Linear Models **Mid-Semester Examination**
Exam Date : Monday, August 30, 2010
Full Marks : 70 **Time : 3 hours** **Instructor : B K Sinha**

NOTE : ALL QUESTIONS ARE COMPULSORY.

Q1. Consider the following expressions for the model expectations of uncorrelated observable random variables [Y_1, Y_2, Y_3, Y_4, Y_5] with the same unknown variance sigma^2 :

E[Y_1]= alpha + gamma; E[Y_2] = alpha + delta;
E[Y_3] = beta + gamma; E[Y_4] = beta + delta = E[Y_5]

All the parameters involved in the model expectations are unknown.

(i) Recast the above in the form of a linear model [Y, XTheta, sigma^2 I] and explain the notations used.
(ii) Write down explicit form of the matrix X and determine its rank.
(iii) Find a pair of uncorrelated error functions arising out of the above model.
(iv) Identify ALL unbiasedly estimable linear parametric functions.
(v) Find an unbiased estimator for sigma^2.
(vi) Explain how you would carry out a test for H_0: [alpha = beta, gamma = delta].

[(2+3) + (2+3) + 5 + 5 + 5 + 5= 30]

Q2. Consider the standard linear model [Y, X Beta, sigma^2 I] where X is a rectangular matrix having full column rank and all model parameters are unknown.

(i) Develop a canonical version of the model.
(ii) Show that every BLUE is necessarily uncorrelated with every error function.
(iii) Show that every BLUE is necessarily unique.

[5 + 5 + 5 = 15]

Q3. (a) Define Sum of Squares [SS] wrt a standard linear model set-up. Explain the notion of Orthogonal Decomposition of the Total SS.
(b) Consider the following model :

E[X]= alpha + gamma; E[Y] = alpha + delta;
E[S] = beta + gamma = E[T]; E[U] = beta + delta = E[V]

where the observable random variables have the same variance and are uncorrelated with each other.
Show that the Total SS i.e., X^2 + Y^2 + S^2 + T^2 + U^2 + V^2 has an Orthogonal Decomposition into FOUR components. Identify these components and give their interpretation.

[(5 + 5) + (10 + 3 + 2) = 25]

INDIAN STATISTICAL INSTITUTE
Mid-Semester Examination : 2010-2011
B.Stat. III Year
Sample Surveys

Date : 01.09.2010 Maximum Marks : 100 Duration : 3 Hours

Use two separate answer sheets for Group A and Group B

Group A (Total Marks : 50)

Notations are as usual.

1. (a) In SRSWR of size n out of N units, derive formulae for π_i and π_{ij}. Let s be an SRSWR of size n out of N units and $\gamma(s)$ denote the number of distinct units in s. Find $E(\gamma(s))$.

$1+3+2=6$

(b) Let $\bar{y}_n$ and $\bar{y}_d$ denote respectively the sample means based on all observations and based on distinct units obtained in an SRSWR of size n out of N units. Show that $\bar{y}_d$ is unbiased for the population mean $\bar{Y}$ and derive $V(\bar{y}_d)$. And also prove that $\bar{y}_d$ is better than $\bar{y}_n$ in estimating $\bar{Y}$.

. $2+2+2=6$

2. (a) Explain the linear systematic sampling scheme and write down with proof an unbiased estimator of population total Y from such a sample as surveyed. Discuss why a circular systematic sampling scheme is needed instead.

$2+2+2=6$

(b) Explain why the variance of an estimator cannot be estimated unbiasedly from a single systematic sample.
Explain with proof how you may get over this problem.

$2+5\ =7$

3. (a) In estimating the population ratio $R=\frac{Y}{X}$ through simple random sampling, show that the bias of the ratio estimator $\hat{R}=\frac{\bar{y}}{\bar{x}}$ is $-\frac{Cov(\hat{R},\bar{x})}{\bar{X}}$.

In SRSWOR, show that an approximate expression for mean square error of $\hat{R}$ is $M_1(\hat{R})=(\frac{N-n}{Nn})\frac{S_y^2+R^2S_x^2-2R\rho S_xS_y}{\bar{X}^2}$, $S_y^2=\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{Y})^2$ and ρ is the correlation coefficient in between y and x.

Show that in SRSWOR, for estimating the population total Y, and if R is +ve, the usual ratio estimation method will be better than the usual simple average estimation method if $\rho > \dfrac{CV(x)}{2CV(y)}$, where CV(x) denotes the coefficient of variation of x.

2 + 4 + 4 = 10

(b) (i) Derive an exactly unbiased estimator for population ratio R based on SRSWOR. 5

(ii) The following figures of expenditures (Rs.) relate to a group of 10 households.

Serial no.	HH size	Exp. (Rs.) last month
1	7	3470.35
2	6	2716.80
3	5	1873.75
4	4	1693.20
5	3	1393.55
6	6	2198.74
7	2	3178.35
8	5	2708.75
9	6	1873.60
10	4	2175.80

Take an SRSWOR of 4 households and give an exactly unbiased estimate of the per capita last month's expenses of these 10 households.

10

Random number table

25	19	17	50	50	46	26	92	62	41	27	66	85	60	70
54	61	41	41	91	88	83	30	32	75	59	03	58	58	83
97	50	71	35	65	67	15	45	73	09	17	60	68	38	05
96	17	27	35	82	80	77	28	97	11	26	72	02	88	96
21	48	84	49	72	93	48	66	75	82	36	33	77	97	35
85	12	09	36	72	81	06	73	04	02	03	10	81	34	44
49	57	40	54	64	88	97	69	03	12	94	45	86	74	66
07	43	79	37	60	96	75	39	46	33	42	41	29	83	73
80	07	51	15	59	55	24	80	49	12	61	68	00	44	58
40	71	81	93	03	03	60	02	42	53	38	35	05	67	73

Group B (Total Marks : 50)

Answer ANY TWO questions . Marks allotted to each question are given within the parentheses . Standard notations and symbols are used .

1. The frequency distribution of 232 cities in a country by population size is given in the following table .Calculate the RSE of the estimator of the total population Y (i) when a sample of 50 cities is selected with SRSWOR and (ii) the two largest cities are definitely included in the survey and only 48 cities are drawn from the remaining 230 cities with SRSWOR .

FREQUENCY DISTRIBUTION OF 232 CITIES BY POPULATION SIZE (000)

population size class	no. of cities	population size class	no. of cities	population size class	no. of cities
50 – 75	81	500 – 550	2	1800 – 1850	1
75 - 100	45	550 – 600	3	1850 – 1950	0
100 – 150	42	600 – 650	1	1950 – 2000	1
150 - 200	14	650 – 700	1	2000 – 2050	0
200 – 250	9	700 – 750	0	2050 – 2100	1
250 – 300	5	750 – 800	1	2100 – 3600	0
300 – 350	6	800 - 850	2	3600 – 3650	1
350 – 400	5	850 - 900	1	3650 – 7850	0
400 – 450	5	900 – 950	2	7850 – 7900	1
450 – 500	2	950 – 1000	0		

(10+15)=[25]

2. Obtain an estimate of the percentage (P) of unemployed persons in a large city using the data given in the following table and estimate its RSE .

ESTIMATES OF P BASED ON TWO SIMPLE RANDOM SAMPLES SELECTED WITH REPLACEMENT

Sample Number	Sample size No. of persons	Percent unemployed
1	2345	19.678
2	1789	20.123

(10+15)=[25]

3. A survey is to be conducted for estimating the total number of literate persons in a town having three communities , some particulars of which are given in the following table based on the results of a pilot study .

A ROUGH IDEA OF THE TOTAL NUMBER OF PERSONS AND PROPORTIONS OF LITERATE PERSONS

Community	Total number of persons	Percentage of literate persons
1	60,000	40
2	10,000	80
3	30.000	60

(a) Treating the communities as strata and assuming SRSWR in each stratum , allocate a total sample size of 2000 persons to the strata in an optimum manner for estimating the overall proportion of literate persons in the town .

(b) Estimate the efficiency of stratification as compared to unstratified sampling . (10+15)=[25]

4. A sample of 10 villages was drawn from a tehsil with PPSWR , size being the 1951 census population and the relevant data are given in the following table .

1951 CENSUS POPULATION (x) AND CULTIVATED AREA (y) IN ACRES FOR 10 SAMPLE VILLAGES

Village	x	y	Village	x	y
1	5511	4824	6	7357	5506
2	865	924	7	5131	4051
3	2535	1948	8	4654	4060
4	3523	3013	9	1146	809
5	8368	7678	10	1165	1013

Total population of the tehsil in 1951 = 415149

(a) Estimate the total cultivated area Y and its RSE .

(b) Obtain the sample size required to ensure an RSE of 2 % . (15 +10) =[25]

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination: (2009–2010)

B. Stat Third Year

Statistical Inference I

Date: 8/09/2010 Marks: 50 Duration: 2 hours

Attempt all questions

1. Prove or disprove the following statement: "A finite collection of exchangeable random variables must be conditionally *iid*". [5]

2. (a) Express the family of binomial distributions in exponential family form. [4]

   (b) Find the natural parameter and the natural parameter space. [3]

   (c) Find a complete sufficient statistic. [1]

3. Let $X_1, \dots, X_n$ be conditionally *iid* Unif$(1-\theta, 1+\theta)$ given $\Theta = \theta$.

   (a) Find a one dimensional minimal sufficient statistic for θ. [4]

   (b) Show that $T = (X_{(1)}, X_{(n)})$ is not a complete sufficient statistic. [4]

4. Let $X_1, \dots, X_n$ be *iid* $N(\mu, \gamma_0^2\mu^2)$ given $\Theta = \mu$, where γ_0 is known and $\mu > 0$.

   (a) Find a minimal sufficient statistic for μ. [5]

   (b) Find

$$E_\mu \left[\frac{n+\gamma_0^2}{1+\gamma_0^2} \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2 \right].$$

   [5]

   (c) Is the minimal sufficient statistic found in (a) complete? Justify. [2]

5. Let X_1 and X_2 be conditionally *iid* Poisson(λ) given $\Theta = \lambda$. Is $T = X_1 + 2X_2$ a sufficient statistic? Prove or disprove. [3]

6. Let $X = (X_1, \dots, X_n)$ be conditionally *iid* Exp(θ) given $\Theta = \theta$, i.e.,

$$f_{X_i|\Theta}(x_i|\theta) = \frac{1}{\theta}\exp(-x_i/\theta), \quad 1 \le i \le n.$$

   (a) Find a minimal and complete sufficient statistic T. [3]

(b) Find the expectation of

$$g(X) = \frac{X_n}{\sum_{i=1}^{n} X_i}.$$

[6]

7. Consider a meter that is trying to measure a quantity Θ. Suppose that the meter gives a reading Z, which has $N(\theta, 1)$ distribution given $\Theta = \theta$ if $Z < 2$, but if $Z \geq 2$, the reading is always 2. Let $X = \min\{Z, 2\}$ be the reading. Letting Θ have a conjugate prior of the form $\Theta \sim N(\theta_0, \sigma_0^2)$ for known values of θ_0 and σ_0^2, obtain the posterior distribution of Θ given X.

[5]

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination : 2010-11(First Semester)

Course Name: **B.Stat III**

Subject Name : **Introduction to Anthropology and Human Genetics**

Date : 9 September 2010 Maximum Marks : 80 Duration : 3 Hours

Answer any **five** (5) questions from the following, all questions carry equal marks

1. How do you define Anthropology? What are the distinctive features of Anthropology?

2. How does Physical Anthropology differ from Biological Anthropology? What are the areas of Biological anthropology?

3. What are the characteristic features of mammals?

4. Why is man unique in the animal kingdom?

5. What are the anatomical changes that have taken place due to the assumption of erect posture?

6. Describe normal karyotypes in man. How Klinefelter syndrome develops?

7. How Huntington's disease occurs? Illustrate the criteria of inheritance of the disease?

8. Short notes on (write any **two**):

   a. Primate
   b. Holism
   c. Turner syndrome
   d. Law of independent assortment

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination (2010-2011)

Subject: Introduction to Sociology
B.Stat. III Year

Date: 9.9.10 Maximum Marks:**25** Duration: Two hours

The figures in the margin indicate full marks

Q 1. Answer any two of the following questions: 5 x 2 = 10

a) Define Sociology. How can Sociology be scientific when the laboratory method can seldom or never be used?
b) Examine Sociology's relation with other social sciences.
c) What is meant by social movements?

Q 2. Write short notes on any two: 4 x 2 =8

a) Globalization.
b) Institution
c) Agrarian society

Q 3. Choose the correct answers: 1 x 7 =7

(a) In which of the following did the class structure develop first?
(i) Tribal Society
(ii) Agricultural Society
(iii) Industrial Society

(b) Who is the father of Sociology?
(i) August Comte
(ii) Ramkrishna Mukherjee
(iii) Darwin

(c) Introduction to Sociology was written by:
(i) Arnold
(ii) A.R.Desai
(iii) Gillin

(d) "The quality of life" was written by:
(i) Ramkrishna Mukherjee
(ii) M.N.Srinivas
(iii) Andre Beteille

(e) The villagers maintain their general behaviouristic patterns because:
(i) They are trained by their families to do so
(ii) They are more cultured
(iii) They generally follow traditions, customs and folkways more resolutely

(f) "Hunting and "Food-gathering" were the primary pursuits of most of the tribesmen in the primitive ages because:
(i) People were ferocious and hunting and food gathering were their past times
(ii) People did not know any other alternative to earn their living
(iii) None of these

(g) "Village is the oldest and permanent community of man", are the words spoken By:
(i) Ogburn and Nimkoff
(ii) Anthony Giddens
(iii) Andre Beteille

Indian Statistical Institute

Mid Semester Examination: (2010–2011)

B.Stat.(Hons.) – III year

Economics III

Date: 11/09/2010 *Maximum Marks –50* *Duration: 2 hours*

Answer ***any two*** questions.

1. (a) State the assumptions of the Classical Linear Regression Model (CLRM) in a multiple regression set up.

   (b) Show that the Ordinary Least Squares (OLS) estimator of the parameters is consistent and Best Linear Unbiased Estimator (BLUE).

   (c) Consider the model
$$\begin{aligned} &y_i = \beta x_i + \varepsilon_i \\ &\varepsilon_i = u_i + \delta u_{i-1}, \qquad |\delta| < 1 \\ &E(u_i) = 0 \\ &E(u_i u_j) = \begin{cases} \sigma_u^2, i = j \\ 0, i \neq j \end{cases} \end{aligned}$$

   (i) Show that the OLS estimator of β is unbiased.

   (ii) Assuming δ and σ_u^2 are known, write down the expression for the GLS estimator of β.

[6+9+10=25]

2. (a) What is multicollinearity?

   (b) Describe the method of Principal Component Regression in this context.

   (c) Describe the Ridge regression technique and show that $Var(\hat{\beta}) - Var(\hat{\beta}_R)$ is positive semidefinite, where $\hat{\beta}$ and $\hat{\beta}_R$ are the OLS and Ridge estimators, respectively.

[3+12+10=25]

3. (a) What do you mean by 'Dummy variables'?

(b) Suppose

$$y = \begin{cases} \alpha_1 + \beta_1 x + \lambda_1 z + \varepsilon \quad for \quad period \ 1 \\ \\ \alpha_2 + \beta_2 x + \lambda_2 z + \varepsilon \quad for \quad period \ 2 \end{cases}$$

Using dummy variables, how would you test the hypothesis that only the coefficient of x changes between the two periods?

(c) Suppose consumption expenditure y depends on income x, but different levels of income produce different values of marginal propensities to consume (β_i). In particular,

$$\frac{dE(y|x)}{dx} = \begin{cases} \beta_1 \ if \ x < Rs.100 \\ \beta_2 \ if \ Rs.100 \leq x < Rs.500 \\ \beta_3 \ if \ x \geq Rs.500 \end{cases}$$

Using Dummy variables estimate the marginal propensity to consume for the three groups (assuming that there is no intercept term) given that $\sum yx$ is 30000; 5,00,000 and 10,00,000 for the three groups, respectively, and $\sum x^2$ is 90000; 25,00,000 and 100,00,000 for the three groups, respectively.

(d) Suppose you want to test r, $r \leq k$, independent linear restrictions of the form $R\beta = d_{r\times 1}$,

where $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{pmatrix}$ in $y_{n\times 1} = X\beta + \varepsilon_{n\times 1}$.

Write down R and d to incorporate the following cases: (k=4)

(i) $\beta_1 = \beta_2 = \beta_3 = 0$

(ii) $\beta_1 = \beta_2$ and $\beta_3 = \beta_4$

(iii) $\beta_1 - 3\beta_2 = 5\beta_3$

(iv) $\beta_1 + 3\beta_2 = 0$

[2+5+10+8=25]

INDIAN STATISTICAL INSTITUTE
Mid-Semester Examination: 2010-11(First Semester)

Course Name: OPTIONAL COURSE FOR B III

Subject Name : GEOLOGY

Date : 09.09.2010 **Maximum Marks :40** **Duration : 1hr 30 min**

Answer any three (3) questions from question number 1 to 4

1. **A.** The weight % of Fe in Earth's crust is rather small (nearly 5) and it is the fourth element in the decreasing order of abundance. Why it becomes the most abundant element when we consider the whole earth instead of the crust only. --------- 3

   B. What are "Chondrites" and "Achondrites"? ------- 5

   C. Name one acid and one basic igneous rock. -- 2

2. **A.** Define "Orthorhombic" crystal system. Name one mineral whose crystals generally belong to "Orthorhombic" system. -----------------2+1 = 3

   B. Compare "Triclinic" and "Monoclinic" systems of crystals. ------ 3

   C. Draw a trigonal pyramid. -------------------- 2

   D. Miller indices of a crystal face are 110; which crystallographic axis is not intercepted by that face? If the face is made to intercept all the three axes with unit lengths, what will be its new Miller indices? ------------ 1+1 = 2

3. What is a mineral? Name one mineral each from the Feldspar and Mica groups. Define "Phyllosilicates" and "Tectosilicates". -------- 3+2+5 = 10

4. Write a note on the factors controlling the texture of a rock. ---------- 10

---

5. Laboratory assignments (done at the GSU Lab) -------------- 10

BStat III [Session : July – November, 2010]
Linear Statistical Models
End-Sem Exam Full Marks : 100 [Weight : 50%]
Time : 3 ½ Hours

Note : Answer Q1 and any THREE of the rest. 19/11

Q1. Consider linear regression of Y on a non-stochastic regressor X with homoscedastic errors : $E(Y \mid x) = \alpha + \beta x$.

You are given the following 'regression design' in terms of a choice of x-values and their 'repeat numbers' :

x :	3	6	8	11	13
n_x :	1	2	8	4	5

(a) Work out an explicit expression for the information matrix for the 'natural' parameters [α and β] in the model.

(b) Make a change in the regressor-values by setting $Z = (X - 8) / 5$. Write down the linear regression model in terms of the 'transformed' regressor Z in the form : $E[Y \mid z] = \gamma + \delta z$.
Express α and β in terms of γ and δ.

(c) Work out an explicit expression for the information matrix for the 'derived parameters' γ and δ.

(d) Establish that a modified regression design of the form :

x^*	:	3	5	6	8	10	11	13
n_{x^*}	:	3	2	1	8	1	2	3

would maximize the determinant of the information matrix for the natural parameters, given the total number of observations as 20.

(e) Interpret your finding in (d).

[5 + (2 + 3) + 5 + 8 + 2 = 25]

Q2. Consider a standard linear model $[Y, X\beta, \sigma^2 I]$ where X is a rectangular matrix, not necessarily possessing full column rank and all model parameters are unknown.

(i) Develop a canonical version of the model. Illustrate with a small example.

(ii) Explicitly establish that any LUE of an estimable linear parametric function, if not already the BLUE, can be converted to the BLUE, by solving a derived system of linear equations. Illustrate with a small example.

(iii) Show that the BLUE of any estimable linear parametric function is necessarily unique in terms of the observable random variables.

[(5 + 5) + (5 + 5) + 5 = 25]

Q3. (a) Write down explicitly the model for two-way classified data with multiple but equal number of observations per 'cell'. Consider both the cases of : 'Without Interaction Terms' and 'With Interaction Terms'.

(b) For both the cases, display the ANOVA Table.

(c) How would you modify the ANOVA Table for a model without any interaction term, in case all the cells possess 'proportional frequencies' ?

(d) For any component source of variation of your choice, work out Expected Value of the Corresponding Sum of Squares for the model in (c) above.

[(2 + 3) + (3 + 5) + 5 + 7 = 25]

Q4. Consider the following model :

E[X]= alpha + gamma; E[Y] = alpha + delta;
E[S_1] = E[S_2] = E[S_3] = beta + gamma;
E[U_1] = E[U_2] = E[U_3] = beta + delta

where the observable random variables X, Y, S_1, S_2, S_3, U_1, U_2 and U_3 have the same variance and are uncorrelated with each other.
Show that the Total SS i.e., $X^2 + Y^2 + \Sigma S^2$_i + ΣU^2_i has an Orthogonal Decomposition into FOUR components. Identify these components and give their interpretation.

[18 + 5 + 2 = 25]

Q5. (a) Write down the ANOCOVA Model wrt One-Way classified data and a single non-stochastic covariate X. Derive conditions for estimability of the 'regression coefficient' and also suggest a test for its significance.

(b) Prepare the ANOCOVA Table for the above set-up in (a). Display expressions for different Sum of Squares and derive an expression for E[Sum of Squares due to Error].

[(5 + 5) + (5 + 5 + 5) = 25]

BK Sinha

10/11/2010

INDIAN STATISTICAL INSTITUTE
First-Semester Examination : 2010-2011
B.Stat. III Year
Sample Surveys

Date : 23.11.2010 Maximum Marks : 100 Duration : 3 Hours

Use two separate answer sheets for Group A and Group B

Group A (Total Marks : 50)

Notations are as usual.
Answer any 3 from Q.1 to Q.5 and Q.6 is compulsory.

1. (a) In SRSWOR of size n out of N units, derive the expressions of π_i and π_{ij}. Also prove that $Cov(\bar{x},\bar{y}) = \left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{X})(y_i-\bar{Y})$. 1 + 1 + 6 = 8

(b) Let $\bar{y}_d$ denote the sample mean based on distinct units obtained in an SRSWR of size n out of N units. Find $V(\bar{y}_d)$. Suggest with proof one unbiased estimator for $V(\bar{y}_d)$ based on distinct units only. 2 + 2 = 4

2. (a) In linear systematic sampling of size n from N = n.k units, where k is an integer, write down with proof an unbiased estimator $\hat{Y}$ of population total Y.
(b) Explain the concept of Intra-class correlation coefficient ρ with proper derivation, for a variable of interest, say y, for L classes and each class having M units.
(c) Prove that in above systematic sampling, $V(\hat{Y}) = \frac{N}{n}(N-1)S_y^2[1+(n-1)\rho]$.
(Use proper definition of ρ in this case).
2 + 5 + 5 = 12

3. (a) In estimating the population total Y by ratio method of estimation with the auxiliary variable x, through SRSWOR, write down the classical ratio estimator.
(b) Modify the sampling scheme so that this classical ratio estimator is unbiased.
(c) Discuss how to estimate the variance of the classical ratio estimator for modified sampling scheme.
2 + 5 + 5 = 12

4. (a) In cluster sampling of n clusters out of N clusters, i-th cluster having M_i units, derive an unbiased estimator $\hat{\bar{Y}}_C$ for population mean $\bar{Y}$, assuming that the clusters are selected by probability proportional to the number of units in them with replacement.

(b) Derive $V(\hat{\bar{Y}}_C)$.

(c) Show that $\frac{1}{n(n-1)}\sum_{i=1}^{n}(\bar{y}_i-\hat{\bar{Y}}_C)^2$ is an unbiased estimator for $V(\hat{\bar{Y}}_C)$, where $\bar{y}_i$ denotes the mean of y values in the i-th selected cluster.

2 + 5 + 5 = 12

5. (a) Discuss how to obtain an unbiased estimator $\hat{\bar{Y}}$ for population mean $\bar{Y}$ in SRSWOR of size n out of N units, following Hansen and Hurwitz 's technique of dealing with non-response errors.

   (b) Find variance of the estimator $\hat{\bar{Y}}$ in (a).

   (c) Find optimum sample size for a given value V_0 of $V(\hat{\bar{Y}})$ and a given cost function with rates defined as C_0 = cost per unit for 1st attempt of data collection, C_1 = cost of editing and processing data for per unit in response class and C_2 = cost of interviewing and processing data per unit of non-response class.

2 + 5 + 5 = 12

6. The following figures of expenditures (Rs.) relate to a group of 10 households.

Serial no.	Exp. (Rs.) last month
1	3470.35
2	2716.80
3	1873.75
4	1693.20
5	1393.55
6	2198.74
7	3178.35
8	2708.75
9	1873.60
10	2175.80

(a) Estimate the average monthly expenditure of this group of 10 households based on 2 independent circular systematic samples each of size 3.

Use the extract from a random number table given below.

(b) Give an unbiased estimate of the standard error of above estimator.

(c) Also give an estimate of the coefficient of variation in percentage.

5 + 5 + 4 = 14

Random number table

25	19	17	50	50	46	26	92	62	41	27	66	85	60	70
54	61	41	41	91	88	83	30	32	75	59	03	58	58	83
97	50	71	35	65	67	15	45	73	09	17	60	68	38	05
96	17	27	35	82	80	77	28	97	11	26	72	02	88	96
21	48	84	49	72	93	48	66	75	82	36	33	77	97	35
85	12	09	36	72	81	06	73	04	02	03	10	81	34	44
49	57	40	54	64	88	97	69	03	12	94	45	86	74	66
07	43	79	37	60	96	75	39	46	33	42	41	29	83	73
80	07	51	15	59	55	24	80	49	12	61	68	00	44	58

GROUP – B
(Maximum Marks =50)

Answer Question No. 4 and any one question from the rest . Marks allotted to each question are given within the parentheses . Standard notations and symbols are used

.

1. (a) A sampler proposes to take a stratified random sample . He expects that his field cost will be of the form $\sum c_h n_h$. His advance estimates of relevant quantities for the two strata are as follows .

Stratum	W_h	S_h	c_h
1	0.4	10	\$4
2	0.6	20	\$9

(i) Find the values of $\frac{n_1}{n}$ and $\frac{n_2}{n}$ that minimize the total field cost for a given value of $\text{Var}(\bar{y}_{st})$.

Find the sample size required under this optimum allocation , to make $\text{Var}(\bar{y}_{st})=1$. Ignore the fpc .

(ii) How much will the total field cost be ?

(b) After the sample in (a) is taken , the sampler finds that his field costs were actually \$2 per unit in stratum 1 and \$12 in stratum 2 .

(i) How much greater is the field cost than anticipated ?

(ii) If he would have known the correct field costs in advance , could he have attained $\text{Var}(\bar{y}_{st}) = 1$ for the original estimated field cost in (a) ?

(iii) If your answer to (ii) is 'no' , find the minimum field cost to reduce $\text{Var}(\bar{y}_{st})$ to 1 .

(5+5+5+5+2+3)=[25]

2. Suppose a population consists of N first stage units (f.s.u.'s) and the ith f.s.u. consists of M_i second stage units (s.s.u.'s) . Suppose a sample of f.s.u.'s is selected in n draws according to PPSWR method of sampling using p_i's as the normed size measures for the ith f.s.u. and each time a f.s.u. (say, the ith f.s.u.) is selected , a sample of m_i s.s.u.'s is selected by SRSWOR sampling scheme .
   (a) Obtain an unbiased estimator of the population total .
   (b) Derive an expression for the sampling variance of the proposed unbiased estimator .
   (c) Also obtain an unbiased estimator of the variance of the estimator of the population total .

(5+10+10) =[25]

3. If a sample is selected in 3 draws according to PPSWR sampling scheme , show that the following two estimators are unbiased for the population total Y :

(i) $$t(s,\mathbf{Y})=\begin{cases}\dfrac{y_i}{p_i} if \quad v=1 \quad and \quad s \quad contains \quad 1 \quad distinct \quad unit \quad i\\ \dfrac{1}{3}\left[\dfrac{y_i}{p_i}+\dfrac{y_j}{p_j}+\dfrac{y_i+y_j}{p_i+p_j}\right] if \quad v=2 \quad and \quad s \quad contains \quad 2 \quad distinct \quad units \quad i,j\\ \dfrac{1}{3}\left[\dfrac{y_i}{p_i}+\dfrac{y_j}{p_j}+\dfrac{y_k}{p_k}\right] if \quad v=3 \quad and \quad s \quad contains \quad 3 \quad distinct \quad units \quad i,j,k\end{cases}$$

where v denotes the number of distinct units in s .

(ii) $$e(s,\mathbf{Y})=\sum_{i\in s}\frac{y_i}{1-(1-p_i)^3}$$

If $y_i\alpha\; p_i$, state giving reasons which one of the above two estimators you would prefer .

(15 +5 +5) =[25]

4 The following table gives the household size and information on whether the household took an agricultural loan from a bank for a random sample of 25 households selected from a population of 515 households in the locality by SRSWOR . Estimate (a) the proportion and the number of households in the locality having agricultural loan and (b) the total number of persons in those households having agricultural loan . Also estimate the relative standard errors of the estimates .

Hh Sl. No.	Whether households have agricultural loan	Hh size	Hh Sl. No.	Whether households have agricultural loan	Hh size
1	Y	5	14	N	6
2	N	7	15	N	10
3	Y	9	16	N	18
4	N	17	17	Y	5
5	Y	3	18	N	9
6	N	7	19	Y	8
7	N	11	20	N	12
8	Y	6	21	N	15
9	Y	8	22	Y	6
10	N	7	23	N	8
11	N	5	24	Y	11
12	Y	14	25	N	17
13	N	12			

Y : Yes , N : No

(10+5+10) = [25]

INDIAN STATISTICAL INSTITUTE

FIRST SEMESTER EXAMINATION, 2010-11

Campus: Kolkata

COURSE NAME : B-STAT III YEAR

SUBJECT NAME : **Differential equations**

Date : 26-11-2010 **Maximum Marks : 70** **Duration: 3 hours 30 min**

Answer the following questions.

1. Locate and classify all the singular points for finite values of t of the following differential equation :

$$t(1-t)y'' + [c - (a+b+1)t]y' - aby = 0 \ (a, b, c \ are \ constants).$$

(10 marks)

2. By the method of power series, show that the solution ϕ of the differential equation

$$y'' - ty = 0$$

satisfying the initial conditions $\phi(0) = a$, $\phi'(0) = b$ is

$$\phi(t) = a\left[1 + \sum_{m=1}^{\infty} \frac{(1)(4)\ldots(3m-2)}{(3m)!} t^{3m}\right] + b\left[t + \sum_{m=1}^{\infty} \frac{(2)(5)\ldots(3m-1)}{(3m+1)!} t^{3m+1}\right].$$

(8 marks)

3. Obtain Green's function $G(t, s, \lambda)$, defined for $0 \leq t, s \leq \pi$, and $\lambda \neq n^2$ $(n = 1, 2, \ldots)$ for the non-homogeneous boundary value problem

$$y'' + \lambda y = f(t)$$

with boundary conditions $y(0) = 0$. $y(\pi) = 0$. (13 marks)

CONTINUED

4. (a) Find all real valued [illegible] solutions $y(t)$ of the differential equation

$$\frac{dy}{dt} - y = t \quad -1 < t < 1.$$

(7 marks)

(b) Find all solutions $y : \mathbb{R} \to \mathbb{R}$ to

$$\frac{dy}{dt} = \sqrt{y(y-2)},\ y(0) = 0.$$

(4 marks)

5. Find a function $y(t)$ such that $y^{(4)} + y = 0$ for $t \geq 0$, $y(0) = 0$, $y'(0) = 1$ and

$$\lim_{t \to \infty} y(t) = \lim_{t \to \infty} y'(t) = 0.$$

(7 marks)

6. Determine all real eigenvalues and corresponding eigenfunctions of each of the following boundary value problems.

(a) $y'' + \lambda y = 0$; $(y(0) = 0,\ y'(\pi) = 0)$. (6 marks)

(b) $y'' + y' + (\lambda + 1)y = 0$; $(y(0) = 0,\ y(\pi) = 0)$. (7 marks)

(c) $t^2 y'' - \lambda t y' + \lambda y = 0$; $(y(1) = 0,\ y(2) - y'(2) = 0)$. (7 marks)

Note : If the eigenvalues are roots of a transcendental equation which cannot be solved explicitly, give the equation for the eigenvalues and the form of the eigenfunctions.

7. Show that for a Sturm-Liouville problem consisting of the following differential equation

$$L(y) = (p(t)y')' + q(t)y = \lambda r(t)y$$

where p, q *and* r are continuous on $a \leq t \leq b$, and $p(t) > 0$, $r(t) > 0$ on $a \leq t \leq b$, and the boundary conditions

$$\alpha y(a) + \beta y'(a) = 0, \quad \gamma y(b) + \delta y'(b) = 0$$

(where α, β, γ and δ are real constants), all eigenvalues are real and eigenvectors corresponding to different eigenvalues are orthogonal on $a \leq t \leq b$ with respect to the weight function $r(t)$. (13 marks)

INDIAN STATISTICAL INSTITUTE

Semestral Examination: (2009–2010)

B. Stat Third Year

Statistical Inference I

Date: 30/11/2010 Marks: 100 Duration: 4 hours

Attempt all questions

1. Suppose that $\{X_n\}_{n=1}^{\infty}$ are bounded, exchangeable random variables. Let $\Theta = \lim_{n\to\infty} \sum_{i=1}^{n} X_i/n$, almost surely. Prove that $Var\,(\Theta) = Cov\,(X_1, X_2)$. [8]

2. Suppose that for every $m = 1, 2, \ldots,$

$$f_{X_1,\ldots,X_m}(x_1,\ldots,x_m) = \frac{2}{(m+1)c_m(x_1,\ldots,x_m)^{m+1}}, \quad \text{if all} \quad x_i \geq 0$$

where $c_m(x_1,\ldots,x_m) = \max\{2, x_1,\ldots,x_m\}$.

(a) Prove that X_i are exchangeable and that these distributions are consistent. [3]

(b) Find the distribution of $Y_n = c_n(X_1,\ldots,X_n)$ and the limit of this distribution as $n \longrightarrow \infty$. [3]

(c) Find the conditional density of X_{n+1} given $X_1 = x_1,\ldots,X_n = x_n$, and assume that $\lim_{n\to\infty} c_n(x_1,\ldots,x_n) = \theta$. Find the limit of the conditional density as $n \longrightarrow \infty$. [3]

(d) Use DeFinetti's representation theorem to show that the prior (the answer to part (b)) and the likelihood (the asnwer to part (c)) combine to give the original joint distribution. [3]

3. Let $X_1,\ldots,X_n$ be *iid* with pmf

$$P(X = x \mid \Theta = \nu) = \frac{1}{x^{\nu}\zeta(\nu)},$$

where $\zeta(\nu) = \sum_{x=1}^{\infty} \frac{1}{x^{\nu}}$.

(a) Find a minimal sufficient statistic for ν. [4]

(b) Is the minimal sufficient statistic complete? Justify. [2]

(c) Find a sufficient statistic which is not minimal. Justify. [2]

4. Suppose that the posterior distribution of Θ is denoted as $F_{\Theta|X}(\theta)$. Define $L(\theta, \delta(x))$ for a nonrandomized rule as

$$L(\theta, \delta(x)) = a(\theta - \delta)I\{\theta \geq \delta\} + b(\delta - \theta)I\{\theta < \delta\},$$

where $a > 0$, $b > 0$ are known constant and $I(\cdot)$ is indicator function.

(a) Find δ that minimizes the posterior risk $r(\delta|x)$. [4]

(b) "If $a= b$ then δ is the posterior mean"-true or false. Justify. [2]

5. Suppose that $\{X_n\}_{n=1}^{\infty}$ are *iid* Ber(θ), $\aleph = \{a_0, a_1\}$, and

$$L(\theta, a) = \begin{cases} 0 & \text{if } (\theta \leq 1/2 \text{ and } a = a_0) \text{ or if } (\theta > 1/2 \text{ and } a = a_1) \\ 1 & \text{otherwise} \end{cases}$$

Let $X = (X_1, \ldots, X_n)$ and let n be even. Define $Y = \sum_{i=1}^{n} X_i$ and let the prior for θ be $U(0, 1)$.

(a) If $Y = y$ successes are observed in n trials, then find the posterior of θ. [4]

(b) Find a randomized and a nonrandomized *Formal Bayes* rules and show that their posterior risks are equal. [6]

6. Let δ_0 be a randomized rule and T be a sufficient statistic. Define $\delta_1(t)(A)$ $=E\left(\delta_0(X)(A)\middle|T = t\right)$.

(a) "$\delta_1(t)(A)$ will not depend on θ"- Is the statement correct? Justify. [2]

(b) Let $h : \aleph \to \mathbb{R}^+$ be a non negative simple function, prove that

$$E\left\{\int h(a)d\delta_0(X)(a)\middle|T = t\right\} = \int h(a)d\delta_1(t)(a).$$

[Hint: Any simple function $\phi(x)$ is written as $\phi(x) = \sum_{i=1}^{n} a_i I_{A_i}(x)$, where $A_i = \{x : \phi(x) = a_i\}$.] [4]

(c) Find the risk functions $R(\theta, \delta_1)$ and $R(\theta, \delta_0)$ and show that they are equal. [6]

7. (a) Suppose that Θ is k-dimensional and that $f_{X|\Theta}(x|\theta)$ is the density of X given $\Theta = \theta$. Also suppose that all the regularity conditions related to the Fisher Information hold. Denote $I_X(\theta)$ be the Fisher Information matrix and let $I_{X,i,j}(\theta)$ be the (i, j)th element of $I_X(\theta)$. Further, assume that two derivatives can be passed

under the integral sign. Given that θ, θ_0 are two elements of Ω (parameter space), let $\mathcal{I}_X(\theta_0; \theta)$ denote the Kullback-Leibler Information matrix. Show that

$$\frac{\partial^2}{\partial\theta_i \partial\theta_j} \mathcal{I}_X(\theta_0; \theta)\Big|_{\theta=\theta_0} = I_{X,i,j}(\theta_0).$$

[4]

(b) Let X have an exponential family of distributions, then show that

$$I_X(\theta) = -\left(\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log(c(\theta))\right)\right),$$

where $c(\theta)$ is the normalizing constant for the exponential family. [4]

8. Let $\Omega = (0,1)$ and let $\{X_n\}_{n=1}^{\infty}$ are *iid* Ber(θ). Suppose $X = (X_1, \ldots, X_n)$. Let $\xi > \theta$ and Θ be discrete with

$$f_\Theta(y) = \begin{cases} \pi_0 & \text{if } y = \theta \\ 1 - \pi_0 & \text{if } y = \xi. \end{cases}$$

Then show that

$$P\left(\Theta = \xi \middle| X = x\right) = \left(1 + \frac{1-\pi_0}{\pi_0} \exp\{(\mathcal{I}_X(p_n; \theta) - \mathcal{I}_X(p_n; \xi))\, n\}\right)^{-1,}$$

where $x = \sum_{i=1}^{n} x_i$, $p_n = x/n$, and $\mathcal{I}_X(p_n; \theta)$, $\mathcal{I}_X(p_n; \xi)$ are Kullback-Leibler information of p_n & θ and p_n & ξ, respectively. [8]

9. Let $\Omega = (0, \infty) \times (0, \infty)$, $\chi = \mathbb{R}^3$, and $\aleph = \mathbb{R}^+$. Suppose that X_1, X_2, X_3 are *iid* U(α, β), where $\theta = (\alpha, \beta)$. Let

$$L(\theta, a) = \left(\frac{\alpha + \beta}{2} - a\right)^2,$$

and $\delta_0 = \bar{X}$.

(a) Let T be a 2-dimensional sufficient statistic for (α, β). Find a rule $\delta_1(T)$ whose risk function is as good as δ_0. [4]

(b) Find the risk functions $R(\theta, \delta_0)$ and $R(\theta, \delta_1)$. [4]

10. Let $X = (X_1, \ldots, X_n)$, where $X_i \overset{iid}{\sim} N(\theta, 1)$. Consider testing $H : \Theta = \theta_0$ versus $A : \Theta \neq \theta_0$. Let $0 < \pi_0 = Pr(\Theta = \theta_0 \mid H) < 1$ and let g be a uniform prior density over $(\theta_0 - c, \theta_0 + c)$ for some $c > 0$, and let x be such that $\bar{x}_n = \sum_{i=1}^{n} x_i/n = \theta_0 + \frac{2\tau_{\alpha/2}}{n}$, where $\tau_{\alpha/2}$ is the upper $\alpha/2$ point of $N(0,1)$. Then, using the 0-1 loss function compare classical two-sided size α and Bayes tests for all such $\bar{x}_n$. [10]

11. Let the parameter space be the open interval $\Omega = (0, 100)$. Let X_1 and X_2 be conditionally independent given $\Theta = \theta$ with $X_1 \sim Poisson(\theta)$ and $X_2 \sim Poisson(100 - \theta)$. We are interested in the hypothesis $H : \Theta \leq c$ versus $A : \Theta > c$.

(a) Show that there is no UMP level α test of H versus A. [2]

(b) Is there any UMP level α test conditional on the ancillary $T = X_1 + X_2$? [3]

(c) Find a prior distribution for Θ such that the conditional UMP level α test given T is to reject H if $Pr\,(H \text{ is true} \mid X_1 = x_1, X_2 = x_2) < \alpha$. [5]

INDIAN STATISTICAL INSTITUTE

First-Semester Examination: 2010-11

Subject: Introduction to Sociology
B.Stat. III Year

Date: 2.12.10 Maximum Marks: **50** Duration: Three hours

The figures in the margin indicate full marks

Answer any five questions of the following:

1. What is objectivity? How can objectivity be maintained in sociological research? Point out the limits of objectivity in sociological research. 2.5+2.5+5 =10

2. What are sources of hypothesis? Discuss the qualities of a workable hypothesis. Cite examples. 5+5=10

3. Relate the Sociological Thinkers with the Concepts introduced by them: 1.5 X 6+1(if all correct) = 10

Serial Number	THINKER	CONCEPT
1	Auguste Comte	Theory of Evolution
2	Herbert Spencer	Positivism
3	Emile Durkheim	Alienation
4	Karl Marx	Suicide
5	Max Weber	Dialectics
6	Karl Marx	Concept of Social Actions

P.T.O.

4. Write Short Notes (any two) 5 X 2 =10

(a) Community Development
(b) Family
(c) Panchayati Raj
(d) Bureaucracy

5. What do you mean by Empowerment?
What is the relation of equality of women with empowerment of women? 5+5= 10

6. Answer any five of the following 2 X 5 = 10
(a) Who is the Father of Sociology?
(b) Which year was the 'Hindu Succession Act' passed?
(c) Who are the major proponents of Conflict Theory?
(d) What are the three types of leadership according to Max Weber?
(e) Who has written 'Origin of Species'?
(f) Which year was the dowry prohibition (Amendment) Bill passed?

Indian Statistical Institute
Semester Examination
B.Stat. III, year 2010
Subject: Anthropology and Human Genetics
Full Marks 80 Date: 2. 12 .10 Time : 3 hours

(Answer any **five** questions from the following, all questions carry **equal** marks)

1. Define organic evolution? Discuss various theories of organic evolution?

2. State Lamarck's theory of evolution? How does it differ from Darwin's theory?

3. What are the stresses on man at high altitude? How do human beings cope with high altitude stress?

4. Define 'adaptation' and 'acclimatization'? State the differences between adaptation and acclimatization with suitable examples.

5. What type of genetic disorder prevailed in the family of Queen Victoria? Describe the mode of inheritance of the disease with the help of pedigree (hypothetical).

6. Discuss nature – nurture controversy.

7. In a randomly mating population of 499 individuals, the distribution of following blood groups are A=193, B=95, AB=54 and O=157. Calculate the phenotype and genotype frequencies of the blood groups in the population.

8. Describe briefly Hardy-Weinberg equilibrium. Does this model applicable in any existing population? Discuss.

9. Write short notes on any **two**-

(a) Multifactorial inheritence
(b) Sex-limited traits
(c) Aneuploidy
(d) Fertility

Indian Statistical Institute

First Semestral Examination: (2010–2011)

B.Stat.(Hons.) – III year

Economics III

Date: 3 · 12 · 10 *Maximum Marks 100* *Duration: 3 hours*

This question paper carries 105 marks. The maximum you can score is 100.

Answer any *three* questions.

1. (a) A researcher has data on enrolment, *N*, and annual recurrent expenditure, *EXP*, measured in Indian rupees, for 50 nursery schools in an Indian city for 2006 and hypothesizes that the cost function is of the quadratic form

$$EXP = \beta_1 + \beta_2 N + \beta_3 NSQ + u, \text{ where } NSQ \text{ is the square of } N.$$

He fits the following equation:

$$\underset{}{\hat{EXP}} = \underset{(12{,}908)}{17{,}999} + \underset{(133)}{1{,}060}\, N - \underset{(0.30)}{1.29}\, NSQ, \qquad R^2 = 0.74 \qquad (1.1)$$

Suspecting that the regression was subject to heteroscedasticity, the researcher runs the regression twice more, first with the 19 schools with lowest enrolments, then with the 19 schools with the highest enrolments. The residual sums of squares (*RSS*) in the two regressions are 8.0 million and 64.0 million, respectively.

The researcher defines a new variable, *EXPN*, expenditure per student, as *EXPN* = *EXP/N*, and fits the equation

$$\hat{EXPN} = \underset{(90)}{1{,}080} - \underset{(0.25)}{1.25}\, N + \underset{(6{,}000)}{16{,}114}\, NREC \qquad R^2 = 0.65 \qquad (1.2)$$

where *NREC* = 1/*N*. He again runs regressions with the 19 smallest schools and the 19 largest schools and the residual sums of squares are 900,000 and 600,000.

(i) Explain what is meant by heteroscedasticity and describe the consequences of its presence in a regression model.

(ii) Describe the Goldfeld–Quandt test for heteroscedasticity and explain why only under certain conditions it may detect heteroscedasticity.

(iii) Perform a Goldfeld–Quandt test for heteroscedasticity on both of the regression specifications. [The critical value at 5% level of significance is 2.33]

(iv) Explain why the researcher ran the second regression.

(v) R^2 is lower in regression (2) than in regression (1). Does this mean that regression (1) is preferable?

(b) Show that, for a first order autoregressive model with positive coefficient, the autocorrelation function (ACF) declines geometrically.

(c) In the multiple regression model $y = X\beta + \varepsilon$, if ε's are correlated with the regressors, what is the appropriate method of estimation? Briefly describe the method.

[(5+6+6+5+3) + 5 +5 = 35]

2. (a) A researcher has data on the average annual rate of growth of employment, *e*, and the average annual rate of growth of GDP, *x*, both measured as percentages, for a sample of 27 developing countries and 23 developed ones for the period 1985–1995. He runs simple regressions of *e* on *x* for the whole sample, for the developed countries only, and for the developing countries only, with the following results:

whole sample	$\hat{e} = -0.56 + 0.24x$ (0.53) (0.16)	$R^2 = 0.04$ $RSS = 121.61$
developed countries	$\hat{e} = -2.74 + 0.50x$ (0.58) (0.15)	$R^2 = 0.35$ $RSS = 18.63$
developing countries	$\hat{e} = -0.85 + 0.78x$ (0.42) (0.15)	$R^2 = 0.51$ $RSS = 25.23$

Now he defines a dummy variable *D* that is equal to 1 for the developing countries and 0 for the others.

(i) Explain the role of the dummy variable in estimating the coefficients of the equations for the two types of countries using a single equation.

(ii) What are the coefficients of this regression equation?

(iii) Compute an appropriate statistic for testing the researcher's hypothesis that the slope and intercept of the equations for the two types of countries are equal. Specify the degrees of freedom and the distribution it follows.

(b) In a linear probability model with binary dependent variable and a single explanatory variable explain why the OLS method of estimation of the parameters produce biased and inconsistent estimates. Name some appropriate methods of estimation in such a case.

(c) Consider the model with two explanatory variables:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon \ .$$

Suppose X_1, X_2 and Y are measured with error and we measure

$$x_1 = X_1 + u_1, x_2 = X_2 + u_2 \text{ and } y = Y + v.$$

Assume that u_1, u_2 and v are mutually uncorrelated and uncorrelated with X_1, X_2 and Y. Are the OLS estimates consistent? Justify your answer.

[(4+5+6) +10 + 10 =35]

3. (a) What are distributed lag models?

(b) Describe the geometric lag model and rationalize the model in terms of (i) Adaptive Expectation Model and (ii) Partial Adjustment Model.

(c) What is "Koyck transformation"? Explain its relevance in the context of estimation of Partial Adjustment Models.

(d) Describe the Almon polynomial lag model. In an empirical exercise how does one determine the lag length and the degree of the polynomial?

[4+ (8+8) + 5 + 10 = 25]

4. (a) A researcher has the following data for a sample of 1,000 manufacturing enterprises on the following variables, each measured as an annual average for the period 2001–2005: *G*, average annual percentage rate of growth of sales; *R*, expenditure on research and development; and *A*, expenditure on advertising. *R* and *A* are measured as a proportion of sales revenue. He hypothesizes the following model:

$$G = \beta_1 + \beta_2 R + \beta_3 A + \varepsilon_G \qquad (4.1)$$

$$R = \alpha_1 + \alpha_2 G + \varepsilon_R \qquad (4.2)$$

where ε_G and ε_R are disturbance terms distributed independently of each other. A second researcher believes that expenditure on quality control, *Q*, measured as a proportion of sales revenue, also influences the growth of sales, and hence that the first equation should be written

$$G = \beta_1 + \beta_2 R + \beta_3 A + \beta_4 Q + \varepsilon_G \qquad (4.1^*))$$

A and *Q* may be assumed to be exogenous variables.

(i) Explain what is meant by 'endogenous' and 'exogenous'.

(ii) Explain what is meant by a reduced form equation, and derive the reduced form equation for G for the first researcher.

(iii) Explain why ordinary least squares (OLS) would be an inconsistent estimator of the parameters of equation (4.2). [*Note*: You are NOT asked to derive an expression for the large-sample bias and no credit will be given for doing so.]

(iv) Comment on the identification status of the models proposed by the two researchers.

(v) Both researchers use two stage least squares (2SLS) to estimate α_2 in (4.2). Do you think that the 2SLS estimator obtained by the second researcher will produce 'better' results than the 2SLS estimator used by the first researcher if the growth equation is given by *(4.1*)*? Give reasons for your answer. (*Note:* Be specific about what you mean by 'better'.]

(b) Describe the Instrumental Variables (IV) approach to estimating a single equation of a simultaneous equations model. Specify the instrument that yields the 2SLS estimator as an IV estimator.

[(3 +5 + 4 +10 + 5)+ 8 = 35]

Introduction to Stochastic Processes
B. Stat. IIIrd Year Semester 2
Indian Statistical Institute

Mid-semestral Examination

Time: 2 Hours Full Marks: 25
Date: February 21, 2011

This is an OPEN NOTE examination. You are allowed to us **only own handwritten classnotes and solutions to homework problems.**

1. Conclude, with justification, whether the following statements are true or false: [$4 \times 3 = 12$]
   (a) For an irreducible Markov chain with d states, for any pair of states i and j, there exists $n \leq d$, such that $p_{ij}^{(n)} > 0$.
   (b) For an irreducible finite state Markov chain, there exists n, such that for any pair of states i and j, we have $p_{ij}^{(n)} > 0$.
   (c) For two states i and j in a Markov chain with $i \rightsquigarrow j$, if i is positive recurrent, then so is j.
   (d) Consider an irreducible Markov chain on the state space S with transition matrix P. Then a Markov chain on the state space $\tilde{S} = S \times S$ with transition matrix $\tilde{P}$ given by
   $$\tilde{p}_{(i_1,i_2),(j_1,j_2)} = p_{i_1j_1}p_{i_2j_2}$$
   is also irreducible.
2. Show that a branching process with the progeny count probability generating function ϕ is a Markov chain, such that the (i,j)-th element of the transition matrix is the coefficient of t^j in $\phi(t)^i$. [3]
3. Consider a queueing system, where, at every time epoch, one customer leaves the queue, if there is one, and a random number of new customers arrive, where the number of arrivals has probability generating function $A(s) = a_0 + a_1 s + a_2 s^2 + \cdots \infty$. Let X_n denote the queue length after time n. Obtain the transition matrix P.

   Show that P is irreducible if and only if $0 < a_0 \leq a_0 + a_1 < 1$. Assume, for the rest of the problem, that the chain is irreducible.

   Let π satisfy $\pi = \pi P$ and let $\Pi(s) = \pi_0 + \pi_1 s + \pi_2 s^2 + \cdots \infty$ be the corresponding generating function. Show that
   $$\Pi(s) = \frac{\pi_0 A(s)}{1 - \frac{1-A(s)}{1-s}}.$$
   Hence or otherwise show that the chain is positive recurrent if and only if the mean arrival count $\rho = \sum_k k a_k$ satisfies $\rho < 1$ and in this case $\pi_0 = 1 - \rho$. [2+3+3+2=10]

INDIAN STATISTICAL INSTITUTE
MID- SEMESTERAL EXAMINATION: 2010 -2011
Subject: Design of Experiments
B. Stat. III Year

Date of Examination: 22.02.11 Maximum Marks: 60 Duration: 2 hours

1. Answer all questions
2. The Paper carries 70 Marks But the maximum you can score is 60

1) What are the fundamental principles of experimental design and what purposes do the serve ?

[10]

2) An engineer approaches a statistician with the following data.

Nine samples were taken from two streams, four from one and five from the other, and the following data obtained:

Pollution level in stream-1 (ppm)	Pollution level in stream-2 (ppm)
16	9
12	10
14	8
11	6
	5

Based on the data it is claimed that the stream-2 is cleaner than stream-1.

The statistician asked the following questions:

1) When the data taken? *2) All in one day?* *3) On different days?*

4) Were data taken during the same time period for the two streams?

5) Were temperatures of the two streams the same? *6) Why were these points chosen?*

7) Are they representative? *8) Are they comparable?*

a) Why did he ask these questions?

b) Are there any other question(s) that he should have asked?

c) Is there any set of answers to these questions (and the question(s) you suggest in (b) above) that would justify a *t* test to draw conclusions?

[3 + 2 + 5 = 10]

3) Experimentation was conducted to compare three teaching methods A B and C. Five teachers were trained in all methods and taught a total of twelve classes. Assignments were made in the following manner:

Teacher 1:	Method A, Method A, Method B
Teacher 2:	Method A, Method C
Teacher 3:	Method B, Method B
Teacher 4:	Method A, Method B, Method C
Teacher 5:	Method A, Method B

The response was a test score for the students in the classes.

(a) What kind of design is this

(b) What are some of the shortcomings of this design?

(c) What difficulties will they cause in the analysis?

(d) Give a proper design with exactly 12 classes. You are free to choose the number of teachers. What is this design?

[2+4+4+(4+1) = 15]

4) In each of the following situations

a) Identify the factors and that are being considered and their types. Give appropriate reasons in each case.

b) Suggest an appropriate design.

c) Suggest at least one noise factor.

i. In a steel melting shop there are two furnaces. The sources of raw material are the same for both the furnaces. It was suspected that the batches of raw materials may yield different results. Only one heat (batch) of steel is produced per day per furnace. From each heat of steel a single sample is collected and chemical analysis is carried out and carbon percentages for each heat of steel are recorded. A team of engineers wants to know if the two furnaces produce steel of same chemical compositions.

[5]

ii. An industrial engineer is conducting an experiment on eye focus time. He is interested in the effect of the distance of the object from the eye on the focus time. Four different distances are of interest. He has five persons available for the experiment.

[5]

5) The effect of five different ingredients (A, B, C, D and E) on the reaction time of a chemical process is being studied. Each batch of new material is only large enough to permit five runs to be made. Furthermore, each run requires approximately 90 minutes, so only five runs can be made in a day. The design used and the data are given in the following table-

Day

Batch	1	2	3	4	5	Row Totals
1	A = 8	B = 7	D = 1	C = 7	E = 3	26
2	C = 11	E = 2	A = 7	D = 3	B = 8	31
3	B = 4	A = 9	C = 10	E = 1	D = 5	29
4	D = 6	C = 8	E = 6	B = 6	A = 10	36
5	E = 4	D = 2	B = 3	A = 8	C = 8	25
Col Totals	33	28	27	25	34	147

Note:		
Note:	Sum of squares of all 25 observations =	1071
	Sum of squares of Day totals =	4383
	Sum of squares of Batch totals =	4399

(a) What design was used? Write down the model explaining the terms

(b) Write down the hypothesis the experimenter wants to test.

(c) Carry out analysis of variance

(d) Carry out tests for the equality of all pairs of treatment means

(e) draw conclusions

Note: LSD = $t_{\alpha/2,\upsilon}$ √(2MS$_E$/n), υ is the error d.f

[4+1+8+8+4 = 25)]

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination : (2010-2011)

B.Stat. III Year

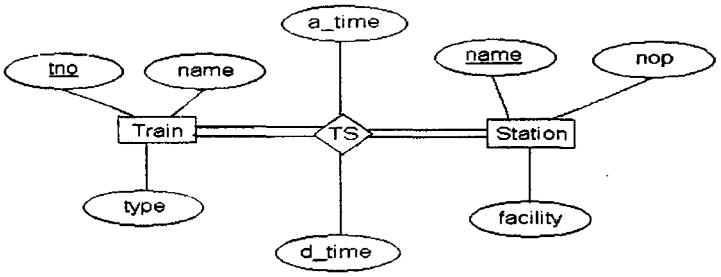
Database Management Systems

Date: 25.02.2011 **Maximum Marks: 30** **Duration: 2 Hrs.**

1. Consider the following schema and form the required queries:
Suppliers (sid, sname, address)
Parts (pid, pname, color)
Catalog (sid, pid, price)
The key fields are underlined. Attributes sid and pid indicate the unique ids for a supplier and a part respectively.
Form the following queries using SQL/ Relational Algebra
a) List the name and address of such suppliers who can supply red carpets.
b) List the name of such parts supplied by "XYZ Co" where the price is greater than Rs.500 but less than Rs.1000.

(5x2=10)

2. Find the errors in the ER diagram shown below and rectify them.



The key attributes are underlined. 'tno' signifies 'train number', ' a_time' and 'd_time' are the arrival and departure time of a train, 'nop' is the 'number of platforms' in a station. Each station may have different facilities like, refreshment room, retiring room, cloak room etc, shown by the attribute 'facility'. All stations may not have all the facilities. So, the attribute 'facility' is multivalued. Station name is unique, but the train name may not be unique.
Arrival time of the originating station of a train and the departure time of the final destination of a train are assigned with the value 9999. In all other cases, value of time is represented in terms of 24 hours.

a) Draw the corrected ER diagram.
b) To design a relational schema, derive a set of relations using the standard mapping rules from the corrected ER diagram.
c) Form the following queries using relational algebra:
   i) Find the name of the originating station and the final destination of the train number 1054.
   ii) Find the list of stations (provide the station name only) that do not have any retiring room facility.

(4+5+5+6=20)

-x-

Midterm Examination
Statistical Inference II
B. Stat. Third Year
Second Semester
2010-2011 Academic Year

Date : 28.02.11 Maximum Marks: 40 Duration :- 2 hours and 15 minutes

Answer as many questions as you can. The maximum you can score is 40.

1. Let $X_1, \ldots, X_n$ be iid observations from some unknown continuous distribution F. We want to test $H_0 : F(x) = F_0(x)$ for all $x \in R$ versus $H_1 : F(x) \neq F_0(x)$ for at least one $x \in R$. Here $F_0(.)$ is a completely specified continuous distribution function.

   (a) Show that the power $P_F(D_n \geq k_\alpha)$ of the one-sample two-sided Kolmogorov-Smirnov test is bounded below by $\sup_x P_F(|F_n(x) - F_0(x)| \geq k_\alpha)$, where k_α is the upper α-point of the distribution of D_n under H_0. [2]

   (b) Show that under H_0, D_n^+ and D_n^- have the same distribution, where symbols have their usual meanings. [3]

2. What is the probability of observing an A-run of length at least 5 in a random arrangement of 8 A's and 8 B's in a line ? Prove your assertion. [4]

3. Let $X_i, i = 1, \ldots, n$ be iid F where $F(x) = G(x - \delta)$ for all $x \in R$, where $G(.)$ is the distribution function of a continuous distribution symmetric around zero and δ is unknown.

   (a) Find a $100(1 - \alpha)\%$ confidence interval for the unknown δ. [4]

   (b) Write down the Hodges-Lèhmann estimator for δ. [1]

   (c) Can you intuitively motivate the estimator in part (b) ? [2]

4. Consider the two-sample problem where one has a random sample $X_1, \ldots, X_m$ from a distribution F and an independent random sample $Y_1, \ldots, Y_n$ from a distribution G, where F and G are both assumed to be continuous.

   (a) Describe the Wald-Wolfowitz runs test for testing $H_0 : F(x) = G(x)$ for all $x \in R$ versus $H_1 : F(x) \neq G(x)$ for at least one $x \in R$. [1]

   (b) Will this test be reasonable for testing that the Y's are stochastically larger than the X's ? Explain your answer. [2]

5. Consider a two-sample testing problem as in question 4 and you use the one-sided or two-sided two-sample Kolmogorov-Smirnov test procedure.

   (a) Consider the case when $m = n$. Suppose H_1 is true. Show that the Kolmogorov-Smirnov two-sided test based on $D_{n,n}$ rejects H_0 with probability tending to 1 when n tends to infinity. You may assume the fact that the upper α point of $D_{n,n}$ under H_0 tends to 0 as $n \to \infty$. [3]

P.T.O.

(b) Consider the case when $m = n$ and H_0 is true. Find the limiting value of $P(D^+_{n,n} > \lambda\sqrt{\frac{2\log n}{n}})$ as $n \to \infty$, where $\lambda > 0$. Prove your answer. [5]

(c) Show in this context that under H_0, for all m, n and $c \in R$

$$P(D^+_{m,n} < c) = P(D^+_{n,m} < c) = P(D^-_{m,n} < c) = P(D^-_{n,m} < c),$$

where $D^+_{n,m}$ and $D^-_{n,m}$ denote the one-sided two-sample Kolmogorov-Smirnov test statistics when one has an X sample of size n and a Y sample of size m. [5]

6. Let X_i, $i = 1, \ldots, m$ be iid F and Y_i, $i = 1, \ldots, n$ be iid G where $G(x) = F(x - \Delta)$ for all $x \in R$ and some $\Delta \in R$. Assume that $F(.)$ is a continuous distribution function.

(a) Without using any approximation, show that $\pi(\Delta)$ defined as $P_{F,G}(W_{XY} > c)$ is a non-decreasing function of $\Delta \in R$, where $c > 0$ and W_{XY} denotes the number of pairs (i, j) such that $X_i < Y_j$, $1 \leq i \leq m$, $1 \leq j \leq n$. [4]

(b) Let $\hat{\Delta}$ be the Hodges-Lehmann estimator of Δ. Then show that

i. Distribution of $\hat{\Delta} - \Delta$ is the same for all $\Delta \in R$. [2]

ii. If $m = 9$ and $n = 7$, then $P(\hat{\Delta} < \Delta) = P(\hat{\Delta} > \Delta)$. [5]

DATABASE MANAGEMENT SYSTEMS
B. STAT. IIIRD YEAR SEMESTER 2
INDIAN STATISTICAL INSTITUTE

Mid-semestral Examination (Supplementary)

Time: 2 Hours Full Marks: 30
Date: March 18, 2011

1. Consider the following schema and form the required queries

```
    Sailors (sid, sname, rating, age)
    Boats (bid, bname, color)
    Reserves (sid, bid, day)
```

The key fields are underlined. Attributes `sid` and `bid` indicate the unique ids for `sailor` and `boat` respectively.

Form the following queries using SQL/Relational Algebra:

(a) Find the names of `sailors` who have reserved all `boats`.

(b) Find the names of `sailors` who have not reserved any green colored `boat`.

[5× 2=10]

2. Explain the following terms briefly: [3×5=15]

(a) attribute;

(b) participation constraint;

(c) domain;

(d) entity;

(e) many to many relationship set.

3. A company database needs to store information about employees (identified by ssn, with salary and phone as attributes); departments (identified by dno, with dname and budget as attributes); and children of employees (with name and age as attributes). Employees work in a department; each department is managed by an employee; the company is not interested in information about a child once the parent leaves the company.

Draw the ER diagram that captures this information. [5]

Supplementary Midterm Examination
Statistical Inference II
B. Stat. Third Year
Second Semester
2010-2011 Academic Year

Date : 21.03.11 Maximum Marks: 40 Duration :- 2 hours and 15 minutes

Answer as many questions as you can. The maximum you can score is 40.

1. (a) Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ under H_0 where both μ and σ are unknown. Consider the one-sided Kolmogorov-Smirnov test criterion with μ and σ estimated by $\bar{X}$ and S respectively where $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$, i.e consider

$$D_n^+ = \sup_t \left\{ F_n(t) - \Phi(\frac{t - \bar{X}}{S}) \right\}.$$

Here $F_n(.)$ is the empirical distribution function and $\Phi(.)$ is the distribution function of $N(0,1)$. Show that under H_0, the distribution of D_n^+ does not depend on μ and σ. [6]

(b) Suppose you have iid observations X_1, X_2 from some unknown continuous distribution F. We want to test $H_0 : F(x) = F_0(x)$ for all $x \in R$ versus $H_1 : F(x) \neq F_0(x)$ for at least one $x \in R$, using the two-sided Kolmogorov-Smirnov test. Here $F_0(.)$ is a completely specified continuous distribution function. Find $P_{H_0}(D_2 < \frac{1}{4} + v)$ where $v \in R$. [6]

2. Let X_i, $i = 1, \ldots, N$ be iid F where F is the distribution function of a continuous distribution symmetric around zero. Let S be the number of i's such that $X_i > 0$ in the sample. Show that under F,

$$P\left[T_1 = t_1, , \ldots, T_s = t_s, S = s\right] = \frac{1}{2^N},$$

where $1 \leq s \leq N$ and $t_1 < \ldots < t_s$ is any set of possible values of $T_1 < \ldots < T_s$, where $T_1 < \ldots < T_s$ are the ordered ranks of the ***absolute values*** of the X observations that are positive. Here the rankings are with respect to the whole set of N absolute values of the X's. [7]

3. Consider the setup in Problem 2 where F is the distribution function of a continuous distribution symmetric around δ where δ is unknown. We want to test $H_0 : \delta = 0$ vs. $H_1 : \delta \neq 0$ using the Wilcoxon Signed Rank Test. Show that this test rejects H_0 with probability tending to 1 if the true F is such that $P_F(X_1 + X_2 > 0) \neq \frac{1}{2}$. [7]

4. Assuming no ties, find the value of $D_{3,4}$ for the ordered sample arrangement $X\ Y\ Y\ X\ X\ Y\ Y$, where the X's are an iid sample from F while the Y's are an iid sample from G where F and G are distribution functions of unknown continuous distributions. Find the P-value of this sample arrangement for testing against the alternative that F and G are not identical. [2+4=6]

5. Let $X_1, \ldots, X_m$ be iid with distribiution F and $Y_1, \ldots, Y_n$ be iid with disribution G where both F and G are strictly increasing and continuous.

   (a) Show that in the case Y is stochastically smaller than X, one can write $G(t) = F(t - \Delta(t))$ for some function $\Delta(t) \leq 0$ for all $t \in R$. Find an explicit expression for $\Delta(t)$. [3]

   (b) Let c_α be such that $P_{F=G}(U \leq c_\alpha) = \alpha$ where

   $$U = \{ \text{ number of pairs } (i,j) \text{ such that } Y_j < X_i \}.$$

   Then show that in case Y is stochastically smaller than X, then $P(U \leq c_\alpha) \leq \alpha$. [4]

6. Consider the Wilcoxon Rank Sum test in the two sample problem, where under H_0, the two populations are assumed to have the same unknown continuous distribution. Let $(W_s = r|k,l)$ denote the event of observing the sum W_s of the Y ranks to be r in an ordered arrangement of a sample of size k from the X-population and a sample of size l from the Y-population. Show that for all permissible values of w larger than $N = m + n$, $P_{H_0}(W_s = w|m,n)$ equals

$$\frac{m}{N} P_{H_0}(W_s = w|m-1,n) + \frac{n}{N} P_{H_0}(W_s = w - N|m,n-1). \qquad [4]$$

INTRODUCTION TO STOCHASTIC PROCESSES
B. STAT. IIIRD YEAR SEMESTER 2
INDIAN STATISTICAL INSTITUTE

Semestral Examination
Time: 3 Hours 30 minutes Full Marks: 50
Date: May 2, 2011

This is an OPEN NOTE examination. You are allowed to use **only own handwritten classnotes and solutions to homework and assignment problems.**

1. Consider a branching process with offspring distribution Bin(2, p). The number of individuals present at the 0-th generation is distributed as Poisson(λ). Find out the extinction probability. [5]

2. A person has r umbrellas, which he keeps distributed between his home and office. At each time point, if the person is in office, he goes to his home and vice versa. At start of each commute, it rains independently with probability p. If it rains at start of a commute and an umbrella is available, the person carries it to the other location. If it is not raining, he does not carry any umbrella. Let X_n denote the number of umbrellas available at his location at n-th time point.

   Obtain the transition matrix and the stationary distribution.

   Fix $0 < \alpha < 1$. If the person wants to guarantee that the probability of getting wet under the equilibrium distribution is at most α, when he knows the probability of rain is p, how many umbrellas must he have? If he does not know p and yet wants to guarantee this probability, irrespective of the probability of rain, how many umbrellas must he have? [3+8+4+3=18]

3. Consider a transition matrix P on the state space $S = \{0, 1, 2, \ldots\}$ with $p_{i,i+1} = 1/(i+1)$ and $p_{i,0} = i/(i+1)$ for $i \geq 0$. Show that P is irreducible. Is it transient or recurrent? [5+7=12]

4. Let $\{N_t : t \geq 0\}$ be PP(λ) and $0 < s < t$. Show that conditioned on the event $[N_t = n]$, the conditional distribution of N_s is Bin(n, s/t). [5]

5. Find out the covariance between B_s and B_t, where $\{B_t : 0 \leq t \leq 1\}$ is a Brownian bridge. [5]

6. The local chowmein shop at Bonhoogly crossing has place for 4 customers to sit. During the busy evening hours, the hungry students of ISI arrive at the shop like a Poisson process with parameter 20. If a student finds the shop to be full on arrival, the student moves to another eatery. However, if there is at least one place available, the student grabs it. The shop can serve each customer independently of each other and of their arrivals on a first come first served basis, where the service time is exponentially distributed with parameter 12. The hungry students eat up the chowmein instantly on being served and leave the shop to allow their friends to enjoy their meals. What is the limiting distribution of number of students at the shop? [5]

INDIAN STATISTICAL INSTITUTE
Second Semestral Examination: 2010 -2011
Subject: Design of Experiments
B. Stat. III Year

Date of Examination: 04.05.11 Maximum Marks: 100 Duration: 3½ hours

Note:
1. Answer all questions
2. The Paper carries 115 Marks but maximum you can score is 100
3. Answer Group A and Group B in separate Answer sheets.

Group A: 50 Marks

1) Describe the missing plot technique, and show that it leads to the valid error as well as a valid least square estimator of the parameter vector. Also show how covariance analysis with dummy covariates is an alternative to the missing plot technique. [(5+10) + 10 = 25]

2) Give the two definitions of connectedness of a block design, and show that they are equivalent. Give an example of a disconnected but orthogonal block design. Develop the ANOVA of a general block design. [(5+10) + 10 = 25]

Group B: 65 Marks

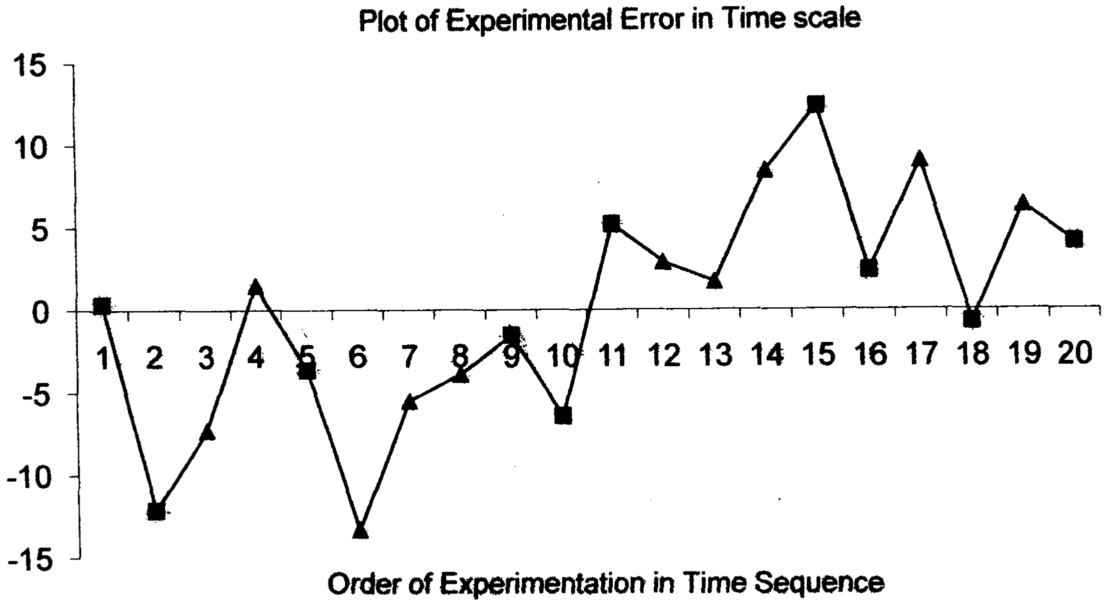
1) A chemical reaction was studied by making 10 runs with a new supposedly improved method (B) and 10 runs with the standard method (A). Following yield results were obtained.

Method A				Method B			
Order of Expt.	Yield	Order of Expt.	Yield	Order of Expt.	Yield	Order of Expt.	Yield
1	52.6	11	57.4	3	64.7	12	74.9
2	40.1	15	64.5	4	73.5	13	73.7
5	48.6	16	54.6	6	58.7	14	80.4
9	50.7	18	51.5	7	66.5	17	81
10	45.8	20	56.3	8	68.1	19	78.3

Error component in each trial was estimated and plotted in the order they were run. See the graph on the following page and comment on the relevant model assumptions. Do you think something went wrong? If your answer is 'NO', justify the answer. If your answer is 'YES', can you suggest what possibly went wrong?

//

3)



Method A : ■ Method B : ▲

[10]

2) An engineer wants to conduct a factorial experiment with the following five factors (each at two levels): temperature, concentration, pH, agitation rate and catalyst type. The experimenter from his domain knowledge expects that only the main effects and two-factor interactions (temperature x concentration) and (temperature x catalyst type) are likely to be present. Resource constraints restrict the size of the experimentation to eight runs.

(a) What type of design should be used? Identify the generators and write down the defining relations for the fraction you propose to use? Write down the alias structures of the effects to be estimated.

(b) Write down the underlying model and assumptions.

(c) What is the resolution of this design? Justify your answer

(d) Give the treatment combinations which constitute the design along with the table of signs to estimate the effects.

[9+3 +3+ 8 = 23]

An experiment is run on an operating chemical process in which four variables are changed in accordance with a randomized factorial plan. The layout of the design and some intermediate computations are given in the following page.

(a) Explain two alternative approaches of estimating experimental error variance in such a situation. Indicate which of Hierarchical Ordering, Effect Sparsity, and Effect Heredity Principles you have used and where in the two alternative approaches.

(b) Analyze the data using the half normal probability plot (given later) of the factorial effects. Prepare the ANOVA table.

(c) What would be the best possible operating conditions in light of the findings in (b) (select the numerically better levels of the factors in respect of impurity, need not do any formal tests) ?

(d) Fit a model that could be used to predict the impurity level in terms of the factors that you have identified as important.

(e) Is there an indication of hidden replication? Can you re-analyse the data considering the hidden replication, if any? How?

[7+10+7+3 +5 =32]

Table 1: Variables and their levels for the chemical process experiment

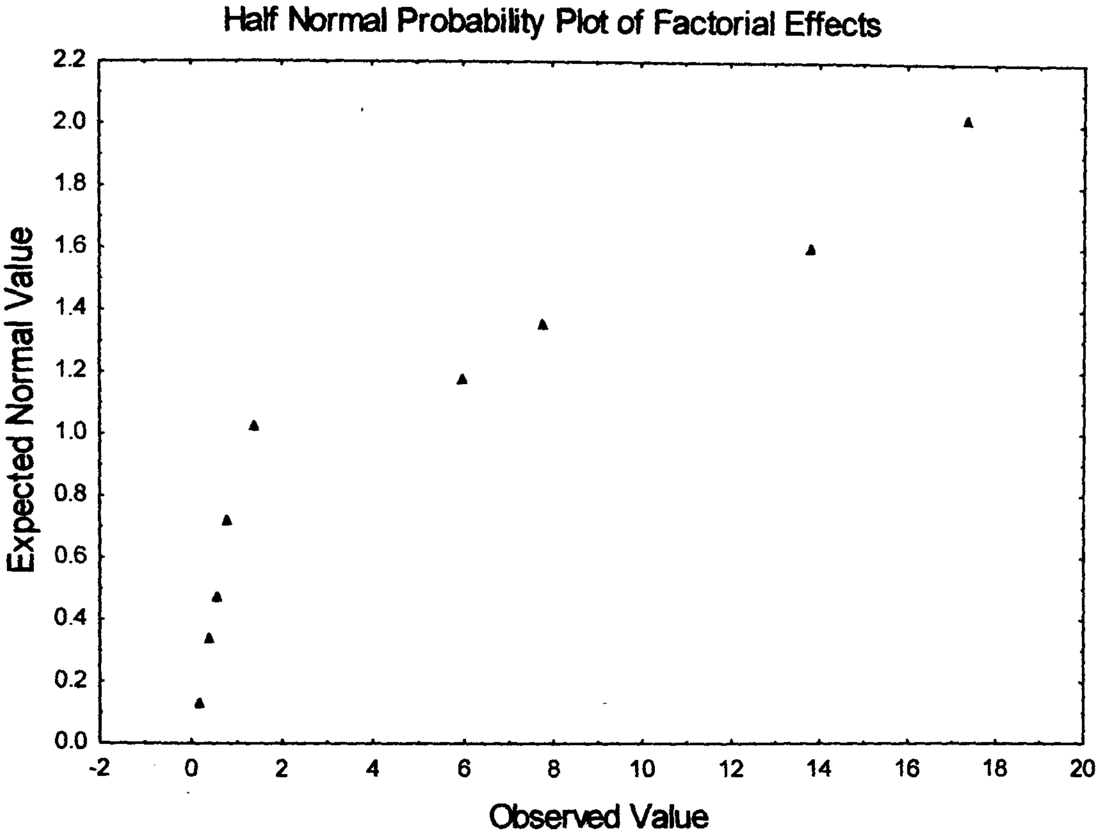
	Variable	Code	Unit	-	+
1	Concentration of catalyst	A	(%)	5	7
2	Concentration of NaOH	B	(%)	40	45
3	Agitation speed	C	(rpm)	10	20
4	Temperature	D	(°F).	150	180

Table 2: Design Layout for the experiment on a chemical process

Trt No	A	B	C	D	Contrasts in the last column of Yates' Table
1.	-	-	-	-	79.6
2.	+	-	-	-	0.2
3.	-	+	-	-	-13.8
4.	+	+	-	-	0.4
5.	-	-	+	-	7.8
6.	+	-	+	-	-0.8
7.	-	+	+	-	-6.0
8.	+	+	+	-	1.4
9.	-	-	-	+	17.4
10.	+	-	-	+	-0.8
11.	-	+	-	+	0.8
12.	+	+	-	+	-0.6
13.	-	-	+	+	-0.4
14.	+	-	+	+	0.2
15.	-	+	+	+	0.2
16.	+	+	+	+	0.8

Note: You may use the following tabulated F values:-

		$\alpha = 0.05$	$\alpha = 0.01$
$F_{1,5}$	:	6.61	16.26
$F_{1,8}$	:	5.32	11.26
$F_{1,9}$	:	5.12	10.56
$F_{1,10}$	:	4.96	10.04
$F_{1,11}$	:	4.84	9.65
$F_{1,12}$	:	4.75	9.33


Half Normal Probability Plot of Factorial Effects
Expected Normal Value
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-2
0
2
4
6
8
10
12
14
16
18
20
Observed Value

INDIAN STATISTICAL INSTITUTE

Second Semester Examination: 2010-2011

B.Stat. 3rd Year

Subject Name: Database Management Systems

Date: 06.05.2011 **Maximum Marks: 70** **Duration: 3 Hrs.**

Time: 3 hr **Subject: DBMS** **Marks: 70**

Note: Answer all questions.

1. Consider the following schema:

   Suppliers(sid, ssname, address)

   Parts(pid, pname, colour)

   Catalog(sid, pid, cost)

   Write the following queries in relational algebra or SQL.

   a. Find the parts (pid) that are supplied by at least two different suppliers.
   b. Find the parts (pid) supplied by every suppliers at a price less than Rs.250/-.

   State what the following queries compute:

   c. $\pi_{sname}\,(\pi_{sid}(\sigma_{colour=\text{`blue'}}\text{Parts})\bowtie(\,\sigma_{cost<100}\text{Catalog})\bowtie \text{Suppliers})$

   d. $\pi_{sname}\,(\pi_{sid}((\sigma_{colour=\text{`blue'}}\text{Parts})\bowtie(\,\sigma_{cost<100}\text{Catalog})\bowtie \text{Suppliers}))$

   e. $(\pi_{sname}\,((\sigma_{colour=\text{`blue'}}\text{Parts})\bowtie(\,\sigma_{cost<100}\text{Catalog})\bowtie \text{Suppliers}))\cap$

   $(\pi_{sname}\,((\sigma_{colour=\text{'yellow'}}\text{Parts})\bowtie(\,\sigma_{cost<100}\text{Catalog})\bowtie \text{Suppliers}))$

   2x6 + 3x4 = 24

2. A data centre, involved in the design of software products, has two types of employees - Technical and Administrative. Technical staffs are again of three types - Data-entry-operator, Programmer and System-analyst. A data-entry-operator is judged by his speed-of-entry in key depressions per second. A programmer has an assigned language in which he normally writes his programs (like C, COBOL etc). A system-analyst has a field of

specialization (like, API-designer, System-software-specialist, Database-consultant etc). In general, an employee is identified by his Employee-no. Besides this, for each employee, the company maintains the following information - name, address, date-of-birth, date-of-joining and monthly-salary. The company also maintains information on the number of projects where each technical staff is involved. If a technical staff is not involved in any project, this field should contain zero. Each project is identified by a unique Project-no. Each project also has a unique name. The other information maintained for each project are - project-name, budget, starting-date, expected-date-of-completion, organization-name, organization-address. The organization-name and address identify the organization that has given the project to the company. Each technical staff may be associated with one or more projects or with no project at all. One of the technical staff associated with a project is its project-leader. All technical staff may not be the project-leader of any project.

a. From the above description, draw an appropriate ER/EER diagram.
b. Map the entity sets and their relationships to a suitable set of relations using the standard mapping rules.

15+15=30

3. Consider the relation for published books:

Book(book_title, author_name, book_type, price, author_affiliation, publisher)

Suppose the following dependencies exist:

book_title → publisher, book_type

book_type → price

author_name → author_affiliation

a. What normal form is the relation in? Explain your answer.
b. Apply 3NF. Check if it is in BCNF.

6+10=16

Semestral Examination
Statistical Inference II
B. Stat. Third Year
Second Semester
2010-2011 Academic Year

Date : 09.05.11 Maximum Marks: 60 Duration :- $3\frac{1}{2}$ hours

Answer as many questions as you can. The maximum you can score is 60.

1. Let X_i, $i = 1, 2, \ldots$ be iid observations from Bernoulli (θ), where $0 < \theta < 1$ is unknown. Consider the SPRT for testing $H_0 : \theta = \frac{1}{3}$ against $H_1 : \theta = \frac{1}{2}$ where boundaries satisfy $0 < B < 1 < A < \infty$. Show, **without using Stein's lemma**, that the SPRT terminates with probability one under $\theta = \frac{2}{3}$. [7]

2. Suppose X_i, $i = 1, 2, \ldots$ are iid observations having exponential distribution with density
$$f_\sigma(x) = \frac{1}{\sigma} e^{-\frac{x}{\sigma}} I_{x>0}.$$
Consider the SPRT with target strength $\alpha = 0.05 = \beta$ for testing $H_0 : \sigma = 1$ against $H_1 : \sigma = 2$, where Wald's approximations for the boundaries are used. Compare the approximate average sample numbers under H_0 and H_1 with the minimum sample size required by the MP test for testing H_0 against H_1 with error probabilities at most α and β. You can (i) approximate the distribution of a standardized chi-square random variable by the standard normal and use the facts that (ii) $P(Z \leq 1.645) \approx 0.95$ where $Z \sim N(0,1)$, (iii) $\ln(19) \approx 2.94$ and (iv) $\ln(2) \approx 0.70$. [10]

3. Let $\{X_i\}_{i \geq 1}$ be a sequence of random variables and the joint density of $(X_1, \ldots, X_m)$ be $p_{jm}(x_1, \ldots, x_m)$ under hypothesis H_j, $j = 0, 1$. Let $0 < B < 1 < A < \infty$ and consider the SPRT of strength (α, β) that stops first time when $\lambda_m = \frac{p_{1m}}{p_{0m}} \geq A$ or $\leq B$. H_0 is rejected if $\lambda_n \geq A$, accepted if $\lambda_n \leq B$ and no decision is made if $n = \infty$, where n is the stopping time of the test. Assume $\alpha > 0$ and $\beta > 0$. Will the Wald inequalities connecting A, B, α and β remain true ? Justify your answer. Note that we are not assuming that the X_i's are iid and it is not guaranteed that the SPRT terminates with probability 1 under H_0 or H_1. [7]

4. Let $X_1, X_2, \ldots$ be iid $N(\mu, \sigma^2)$, where both μ and σ are unknown. Let $\alpha \in (0,1)$. Derive a confidence interval of confidence coefficient at least $1 - \alpha$ for μ, of width at most $2l$ where $l > 0$. Jusify your answer. Can you propose an unbiased estimate of μ of variance at most 1 ? Prove your answer. [10]

5. Two judges A and B rank five competitors as $\{5, 3, 1, 2, 4\}$ and $\{5, 2, 3, 1, 4\}$ respectively, where the i-th element in each vector refers to the ranking of the i-th competitor by the respective judge. Compute two nonparametric measures of association between the two judges' rankings. [6]

P.T.O.

6. Let $X_1, \ldots X_n$ be iid $F(x - \theta)$ where F is a continuous distribution having a continuous density f which is symmetric around 0 and $f(0) > 0$ and $Var_F(X_i) = \sigma_F^2 < \infty$. We want to test $H_0 : \theta = 0$ versus $H_1 : \theta > 0$. Derive a general expression for the ARE of the sign test relative to the t-test in this problem and evaluate it when $F = N(0,1)$ and F is the double exponential distribution with density $f(x) = \frac{1}{2}e^{-|x|}$, $-\infty < x < \infty$. [10]

7. Consider $X_1, \ldots, X_n$ iid with distribution F where $F \in \mathcal{F}$, $\mathcal{F}$ being some class of distributions.

   (a) Define the U-statistic U_n for unbiased estimation of $\theta = \theta(F)$ based on a symmetric kernel $h(x_1, \ldots, x_m)$.

   (b) Prove that $nVar(U_n) \rightarrow m^2\sigma_1^2$ as $n \rightarrow \infty$, where symbols have usual meanings.

   (c) Write down the formula for $\hat{U_n}$, the projection of U_n. Sketch the main argument in showing that the asymptotic distribution of appropriately centred and scaled one-sample U-statistic is normal under suitable assumptions.

   (d) Let $\mathcal{F}$ be the calss of $Bernoulli(\theta)$ distributions as $\theta \in (0,1)$. Find those values of $\theta \in (0,1)$ for which the asymptotic distribution of a suitably centered and scaled sample variance s^2 is standard normal. If for some θ this result is not valid, find suitable centering and scaling such that under that θ, the asymptotic distribution is non-degenerate and find the asymptotic distribution. [1+5+4+6=16]

INTRODUCTION TO STOCHASTIC PROCESSES
B. STAT. IIIRD YEAR SEMESTER 2
INDIAN STATISTICAL INSTITUTE

Backpaper Examination
Time: 3 Hours Full Marks: 100
Date: June 27, 2011

This is an OPEN NOTE examination. You are allowed to use **only own handwritten classnotes and solutions to homework and assignment problems.**

1. Consider a branching process with offspring distribution NegBin(2, p). Suppose there is only one individual at the 0-th generation. Find out the probability that the process becomes extinct for the first time at the third generation. [9]

2. Let $f(s) = as^2+bs+c$, where a, b, c are positive and $f(1) = 1$. If the probability of extinction π satisfies $\pi \in (0,1)$, show that $\pi = c/a$. [12]

3. Let $\{r_i\}_{i=0}^{\infty}$ be a probability vector, that is, $0 < r_i < 1$ and $\sum_{i=0}^{\infty} r_i = 1$. Consider a transition matrix P on the state space $S = \{0,1,2,\ldots\}$ with $p_{0,i} = r_i$ and $p_{i,i-1} = 1$ for $i \geq 1$. Show that P is irreducible. Show that the chain is recurrent. Classify the chain as null or positive recurrent based on the probability vector $\{r_i\}$. [6+6+10=22]

4. Show that

$$\sup_{n\geq 1} p_{ij}^{(n)} \leq f_{ij} \leq \sum_{n=1}^{\infty} p_{ij}^{(n)}.$$

Hence show that
(i) $i \rightsquigarrow j$ if and only if $f_{ij} > 0$, (ii) $i \leftrightsquigarrow j$ if and only if $f_{ij}f_{ji} > 0$. [5+5+2=12]

5. An irreducible Markov chain with invariant distribution π is called *time reversible* if $\pi_i p_{ij} = \pi_j p_{ji}$ for all i, j.
(a) For an irreducible doubly stochastic transition matrix, show that uniform distribution is invariant and the chain is time reversible if and only if the matrix is symmetric. [3+7=10]
(b) Show that an irreducible chain with transition matrix P is time reversible if and only if

$$p_{ii_1}p_{i_1i_2}\cdots p_{i_ki} = p_{ii_k}p_{i_ki_{k-1}}\cdots p_{i_1i},$$

for all natural numbers k and states $i, i_1, \ldots, i_k$. [10]

6. Due to the stress of coping with the studies, a student begins to suffer from headaches of random severities. The headaches occur according to a Poisson process with parameter λ. The severities of the headaches are independent of the time of occurrence and the severities themselves are independently and identically distributed as unit exponential random variables. (Assume headaches are instantaneous and have zero duration.)
The student will need to seek the help of the medical centre as soon as a headache of severity greater than $c > 0$ attacks. Compute the probability that the student will need to seek help in the time interval $[0,t]$. [15]

7. Let $\{B_t : 0 \leq t \leq 1\}$ be standard Brownian motion. If c is a positive real number, show that $\{Z_t = c^{-1}B_{c^2t} : t \geq 0\}$ is also standard Brownian motion. [10]

Backpaper Examination
Statistical Inference II
B. Stat. Third Year
Second Semester
2010-2011 Academic Year

30.06.11

Maximum Marks: 100 Duration :- 3 hours

Answer all questions

1. Show, stating appropriate assumptions, that the null distribution of Kolmogorov-Smirnov one-sample goodness of fit test criterion is indepndent of the assumed null distribution from which the data are generated. [10]

2. Consider the one sample problem: $X_1, \ldots, X_n$ are iid F where F is continuous with unknown unique median M. We want to test $H_0 : M = 0$ versus $H_1 : M > 0$. Show that the sign test is consistent for this testing prolem. [10]

3. Consider the two-sample testing problem of testing equality of two continuous distributions against the alternative that one distribution is stochastically larger than the other. Suppose two samples of size n each are drawn from the two distributions and the one sided Kolmogorov-Smirnov test statistic $D^+_{n,n}$ is used. Find $P_{H_0}(D^+_{n,n} \geq \frac{k}{n})$, where $k = 0, 1, \ldots, n$, where under H_0 the two distributions are identical. Prove your answer. [10]

4. Stating appropriate assumptions, prove asymptotic normality of the null distribution of the Wilcoxon signed rank test statistic for the problem of testing that the median of a symmetric continuous distribution is zero. [15]

5. State and prove Stein's lemma about the termination property of SPRT. [15]

6. State and prove the fundamental identity of sequential analysis. [10]

7. Stating appropriate assumptions, prove the optimality of SPRT in terms of its average sample number under both H_0 and H_1 among tests whose type I and type II errors are bounded above by α and β respectively where $0 < \alpha$, $0 < \beta$ and $\alpha + \beta < 1$. [20]

8. Write a few paragraphs explaining the concept of Asymptotic Relative Efficiency of tests due to Pitman. [10]

BACK PAPER

INDIAN STATISTICAL INSTITUTE
Second Semestral Examination: 2010 -2011
Subject: Design of Experiments
B. Stat. III Year

Date of Examination: 01/07/11 Maximum Marks: 100 Duration: 3 hours

Note: Answer all questions. The Paper carries 100 marks

1) Answer true or false (give a brief justification in each case)

(a) A randomised block design is an orthogonal design.

(b) If all pairwise treatment contrasts are estimable from a block design then the full set of orthonormal treatment contrasts are also estimable.

(c) Any incomplete block design may or may not be an orthogonal design

(d) In a connected block design with v treatments, the rank of the treatment estimation space is v – 1.

(e) From a single replicate of a 2^k factorial experiment, one can test significance of all main effects and all interactions.

(f) The objectives met by a designed experiment can also be achieved by regression analysis methodology using happenstance data.

[3 x 6 = 18]

2) An industrial engineer wants to investigate the effect of four assembly methods on the assembly time for a colour television component. Four operators were chosen for the study. Moreover, it is known that the fatigue factor affects the assembly time. The time required for the last assembly may be greater than the time required for the first assembly, regardless of the method. To account for this source of variation, order of assembly was chosen as a factor. Besides these it was felt that the workplace might also affect the assembly time and was considered as a factor.

(a) Identify all factors that are being considered and their types.

(b) What type of an experimental design should be used? Write down the model.

(c) Give a layout of the design and describe how you will randomize while experimenting with the design.

(d) Give the ANOVA table in the case the laid out experiment is repeated with a new set of four operators.

[3+2+3+7=15]

3) (a) Show that there are at most $v - 1$ mutually orthogonal Latin squares of order v,

(b) Construct the complete set of mutually orthogonal Latin squares of order 4

[5+7=12

4) (a) Construct a 2^{5-1} fractional factorial design of highest possible resolution. Justify your answer.

(b) Show how the design in (a) may be run in two blocks of eight treatments each. In the design constructed by you are there any main effects or two factor interactions confounded with blocks?

(c) Indicate where you have used (which of) Hierarchical Ordering, Effect Sparsity, and Effect Heredity principles in (a) and (b).

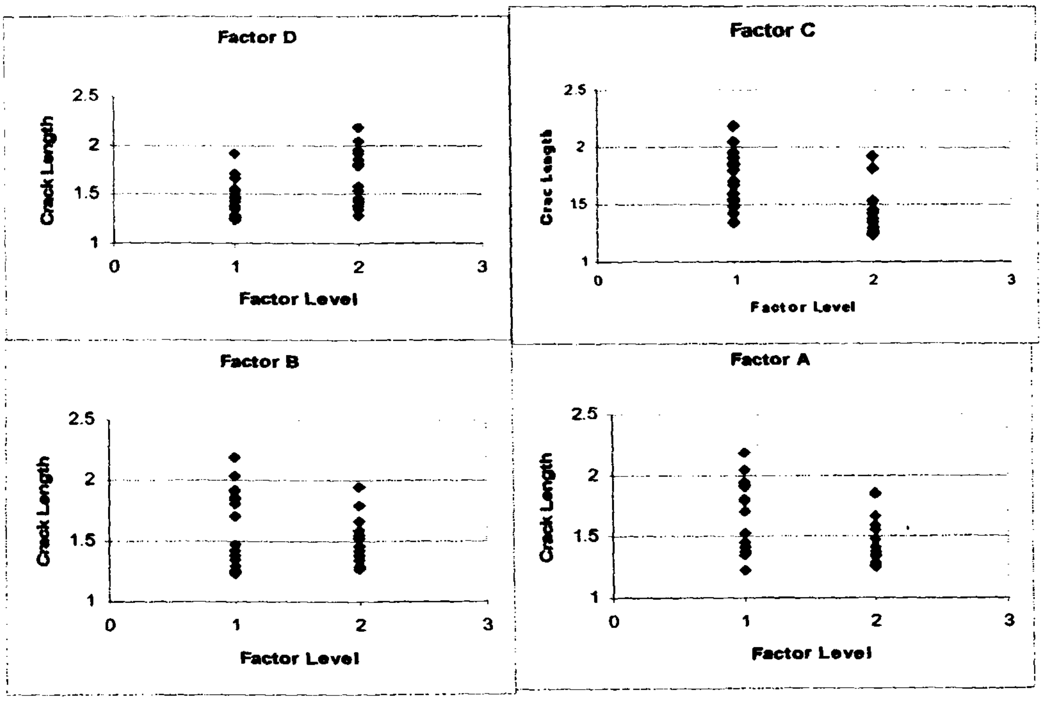
(d) Outline the method of analysis that you would adopt, if only a single replicate were run. State all the necessary assumptions you need to make.

[(5+1) + 7 + 4 + (6+2) = 25]

5) A nickel-titanium alloy is used to make components for jet turbine aircraft engines. Cracking is a potentially serious problem in the final part, as it can lead to non-recoverable failure. A test is run at the parts producer plant to determine the effect of four factors on cracks. The four factors are pouring temperature (A), titanium content (B), heat treatment method (C) and amount of grain refiner used (D). Two replicates of a 2^4 experiment are run and the length of crack (in mm) induced in a sample coupon subjected to a standard test is measured. The data are shown in the next page.

i) When do we use 2 level factors as in the present experiment?
ii) Carry out the ANOVA. Which factors affect cracking?
iii) Prepare average response curves for the significant effects
iv) Obtain the combination that minimizes crack length.
v) The graph below shows level wise plots of the responses for all the factors. What does it indicate? Which of the assumptions may be violated?

[3+13+7+3+4 =30



Design layout and the Response (Problem 5)

Expt No.	Factors				Replicates			Contrasts without the deviser
	A	**B**	**C**	**D**	**I**	**II**	**I + II**	
1.	-	-	-	-	1.71	1.91	3.62	49.95
2.	+	-	-	-	1.42	1.48	2.9	-3.41
3.	-	+	-	-	1.35	1.53	2.88	-1.93
4.	+	+	-	-	1.67	1.55	3.22	1.47
5.	-	-	+	-	1.23	1.38	2.61	-4.67
6.	+	-	+	-	1.25	1.26	2.51	-0.15
7.	-	+	+	-	1.46	1.42	2.88	1.29
8.	+	+	+	-	1.29	1.27	2.56	-0.23
9.	-	-	-	+	2.04	2.19	4.23	3.59
10.	+	-	-	+	1.86	1.85	3.71	-1.81
11.	-	+	-	+	1.79	1.95	3.74	-1.73
12.	+	+	-	+	1.42	1.59	3.01	-0.21
13.	-	-	+	+	1.81	1.92	3.73	-0.55
14.	+	-	+	+	1.34	1.29	2.63	-0.07
15.	-	+	+	+	1.46	1.53	2.99	-0.19
16.	+	+	+	+	1.38	1.35	2.73	2.33

Sum of observations of Replicate I = 24.48 Sum of observations of Replicate II = 25.47

Sum of Squares of 16 treatment totals = 160.0649 Sum of squares of all 32 observations = 80.138

Note: You may use the following tabulated F values:-

	$\alpha = 0.05$	$\alpha = 0.01$
$F_{1,16}$:	4.49	8.53
$F_{2,16}$:	3.63	6.23