4

INDIAN STATISTICAL INSTITUTE

Mid-semester Examination: 2008-2009

B.Stat. (Hons.) 3rd Year. 1st Semester

Linear Statistical Models

Date: September 01, 2008 **Maximum Marks: 70** **Duration: 3 and 1/2 hours**

- This question paper carries 74 points. Answer as much as you can. However, the maximum you can score is 70.
- You should provide as much details as possible while answering a question. You must state clearly any result stated (and proved) in class you may need in order to answer a particular question.

---

1. *Consider a linear model given by* $Y = X\beta + \epsilon$, *where* X *is* $n \times p$, β *is* $p \times 1$, $E(\epsilon) = 0$, *and* $\mathrm{Cov}(\epsilon) = \sigma^2 I_n$.

(a) Suppose $\lambda \notin C(X^T)$. Show that $\lambda^T\beta$ is not identifiable.

(b) Let $\lambda^T\beta$ be estimable. Assume that for any solution $\hat{\beta}$ of the normal equations $X^TX\beta = X^TY$, $\lambda^T\hat{\beta}$ is unique, i.e., independent of the choice of $\hat{\beta}$. Show that $\lambda^T\hat{\beta}$ is a BLUE of $\lambda^T\beta$.

(c) Let $R^2 \stackrel{def}{=} Y^T(I_n - P_X)Y$. Obtain an unbiased estimate of σ^2, which is of the form $c\,R^2$ for a suitable constant c to be chosen by you. Comment on this estimate.

[4+7+(5+3) = 19]

2. *Consider a linear model given by* $Y \sim N_n(X\beta, \sigma^2 I_n)$, *where* X *is* $n \times p$ *and* β *is* $p \times 1$. *Let* Λ *be a known* $p \times q$ *non-null matrix such that* $\Lambda^T = R^TX$ *for some* $n \times q$ *matrix* R. Suppose we wish to test the hypothesis $H_0 : \Lambda^T\beta = 0$ versus $H_1 : \Lambda^T\beta \neq 0$.

(a) Explain how you can recast the testing problem described above as one of testing

$$\begin{aligned} & H_0^* : Y \sim N_n(X_0\gamma, \sigma^2 I_n), \quad X_0 \text{ is } n \times p_0, \ \gamma \text{ is } p_0 \times 1 \\ \text{versus} \quad & H_1^* : H_0 \text{ is false.} \end{aligned} \qquad (*)$$

for some X_0 satisfying $C(X_0) \subsetneq C(X)$, i.e., $C(X_0)$ is a proper subspace of $C(X)$.

(b) Obtain, with adequate justification, a specific choice for X_0 in $(*)$.

(c) Assuming the algebraic expression for a suitable test statistic for testing $(*)$ to be known, obtain a suitable test statistic for testing H_0 which depends on the least squares estimate of $\Lambda^T\beta$ and whose null distribution is F with degrees of freedom depending on $r(\Lambda)$ and $r(P_X)$. [Note. You are required to derive the null distribution.] [6+6+9 = 21]

P.T.O.

3. *Consider a one-way ANOVA model given by* $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, j = 1, \ldots, N_i, i = 1 \ldots, t, \epsilon_{ij}$*'s* $\overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$. *Let* $n \overset{def}{=} \sum_{i=1}^{t} N_i$.

(a) A suitable test statistic for testing $\mathrm{H}_0 : \alpha_1 = \cdots = \alpha_t$ versus $\mathrm{H}_1 : \alpha_i \neq \alpha_j$ for some $i \neq j$ is given by

$$F^* \overset{def}{=} \frac{\sum_{i=1}^{t} N_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{\sum_{i=1}^{t}\sum_{j=1}^{N_i}(Y_{ij} - \bar{Y}_{i\cdot})^2}, \text{ where } \bar{Y}_{i\cdot} \overset{def}{=} \sum_{j=1}^{N_i} \frac{Y_{ij}}{N_i}, \ i = 1, \ldots, t; \ Y_{\cdot\cdot} \overset{def}{=} \sum_{i=1}^{t}\sum_{j=1}^{N_i} \frac{Y_{ij}}{n},$$

with large values of F^* indicating significance. Find the non-null distribution of F^*.

(b) Suppose $t = 3, N_1 = 4, N_2 = 9, N_3 = 4$. Obtain, with adequate justification, two orthogonal contrasts. [8+4 = 12]

4. *Consider a balanced two-way ANOVA model without interaction given by* $Y_{ijk} = \mu + \alpha_i + \eta_j + \epsilon_{ijk}, k = 1, \ldots, N, i = 1, \ldots, s, j = 1, \ldots, t, \epsilon_{ijk}$*'s* $\overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$. Explain how you can develop a suitable test for testing $\mathrm{H}_0 : \alpha_1 = \cdots = \alpha_s$ versus $\mathrm{H}_1 : \mathrm{H}_0$ is false. [Note. You may assume relevant facts about testing of hypotheses, including distribution-theoretic ones, arising in a one-way ANOVA model.] [10]

5. *Consider a balanced two-way ANOVA model with interaction given by* $Y_{ijk} = \mu + \alpha_i + \eta_j + \gamma_{ij} + \epsilon_{ijk}, k = 1, \ldots, N, i = 1, \ldots, s, j = 1, \ldots, t, \epsilon_{ijk}$*'s* $\overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$.

(a) Suppose we wish to test the hypothesis $\mathrm{H}_0 : \gamma_{ij} = \gamma \, \forall \, i, j$ versus $\mathrm{H}_1 : \mathrm{H}_0$ is false. Show that an appropriate sum of squares for testing H_0, whose large values indicate significance, is given by

$$N \sum_{i=1}^{s}\sum_{j=1}^{t} [\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdot\cdot\cdot}]^2,$$

where

$$\bar{Y}_{ij\cdot} \overset{def}{=} \sum_{k=1}^{N} \frac{Y_{ijk}}{N}, \ \bar{Y}_{i\cdot\cdot} \overset{def}{=} \sum_{j=1}^{t}\sum_{k=1}^{N} \frac{Y_{ijk}}{tN}, \ \bar{Y}_{\cdot j\cdot} \overset{def}{=} \sum_{i=1}^{s}\sum_{k=1}^{N} \frac{Y_{ijk}}{sN}, \ i = 1, \ldots, s, \ j = 1, \ldots, t,$$
$$Y_{\cdot\cdot\cdot} \overset{def}{=} \sum_{i=1}^{s}\sum_{j=1}^{t}\sum_{k=1}^{N} \frac{Y_{ijk}}{Nst}.$$

(b) Suppose we wish to test the hypothesis $\mathrm{H}_0 : \alpha_1 = \cdots = \alpha_s$ versus $\mathrm{H}_1 : \mathrm{H}_0$ is false. Explain why this hypothesis cannot be tested. [8+4 = 12]

* * * * * * *Best of Luck!* * * * * *

INDIAN STATISTICAL INSTITUTE

Mid-semestral Examination : 2008-2009

B. Stat. - III Year

Differential Equation

Date : 03. 09. 2008 Maximum Score : 40 Time :3 Hours

Any result that you use should be stated clearly.

(1) Verify that the following differential equation is exact and find the general solution:

$$(\sin\ x \sin\ y - xe^y)dy = (e^y + \cos\ x \cos\ y)dx.$$

[2+5=7]

(2) Find an integrating factor of the differential equation

$$(3x^2y^4 + 2xy)dx + (2x^3y^3 - x^2)dy = 0$$

and find the general solution.

[3+5=8]

(3) Find the particular solution of

$$yy'' = y^2y' + (y')^2,$$

satisfying the conditions $y = -\frac{1}{2}$ and $y' = 1$ when $x = 0$.

[6]

(4) Let $p(x)$ and $q(x)$ be continuous functions on $[a\ , b]$ and $y_1(x)$, $y_2(x)$ be any two solutions of

$$\frac{d^2y}{dx^2} + p(x)\frac{dy}{dx} + q(x)y = 0$$

on $[a\ , b]$. Prove that the Wronskian of $y_1(x)$ and $y_2(x)$ is either identically zero or is never zero on the interval $[a\ , b]$.

[10]

(5) Verify that $y = e^x$ is a particular solution of the differential equation

$$xy'' - (2x + 1)y' + (x_1)y = 0$$

and find the general solution.

[6]

(6) Find a particular soultion of $y'' - y' - 6y = e^{-x}$ by using the method of variation of parameters.

[6]

(7) Find two independent analytic solutions of the equation $y'' + xy' + y = 0$ around 0.

[8]

(8) Consider the initial value problem $\frac{dy}{dx} = x^2|y|$, $y(0) = 0$. Does it admit a unique solution in a neighbourhood of 0? Justify your answer.

[4]

INDIAN STATISTICAL INSTITUTE
Mid – Semester Examination : 2008 -09
B. Stat. (Hons.) III Year
Sample Surveys

Date :5. 09.2008 **Maximum Marks : 100** **Duration : 3 Hours**

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

1. Suppose from a finite population of size 3 , a sample of size 2 is drawn according to SRSWOR . Consider the following estimator

$$\hat{\bar{Y}}_{12} = \frac{1}{2}y_1 + \frac{1}{2}y_2 \ , \ \hat{\bar{Y}}_{13} = \frac{1}{2}y_1 + \frac{2}{3}y_3 \ , \ \hat{\bar{Y}}_{23} = \frac{1}{2}y_2 + \frac{1}{3}y_3$$

where $\hat{\bar{Y}}_{ij}$ is the estimator for the sample that has units (i,j) .

(a) Prove that $\hat{\bar{Y}}_{ij}$ is an unbiased estimator of the population mean $\bar{Y}$.

(b) Obtain the sampling variance of $\hat{\bar{Y}}_{ij}$.

(c) Hence or otherwise show that $\hat{\bar{Y}}_{ij}$ is more efficient than the sample mean $\bar{y}$ if

$$y_3(3y_2 - 3y_1 - y_3) > 0 \ .$$

(5+10+10)=[25]

2. Two dentists A and B make a survey of the state of the teeth of 200 children in a village . Dr. A selects a simple random sample of 20 children drawn without replacement and counts the number of decayed teeth for each child , with the following results .

Number of decayed teeth/child	0	1	2	3	4	5	6	7	8	9	10
Number of children	8	4	2	2	1	1	0	0	0	1	1

Dr. B , using the same dental techniques , examines all 200 children , recording merely those who have no decayed teeth . He finds 60 children with no decayed teeth .Estimate the total number of decayed teeth in the village children , (a) using A's results only , (b) using both A's and B's results . (c) Are the estimates unbiased ? (d) Which estimate do you expect to be more precise and why ?

(7+8+5+5)=[25]

3. With two strata , a sampler would like to have $n_1 = n_2$, for administrative convenience , instead of using the values given by the Neyman optimum allocation . If $V(\bar{y}_{st})$ and $V_{opt}(\bar{y}_{st})$ denote the variances given by the $n_1 = n_2$ and the Neyman optimum allocations , respectively , show that the fractional increase in variance is given by

P.T.O.

$$\frac{V(\bar{y}_{st}) - V_{opt}(\bar{y}_{st})}{V_{opt}(\bar{y}_{st})} = \left(\frac{r-1}{r+1}\right)^2$$

where $r = \frac{n_1}{n_2}$ as given by Neyman optimum allocation .

For the following example , what would be the fractional increase in variance by using $n_1 = n_2$ instead of the optimum ?

Stratum	W_h	S_h
1	0.8	2
2	0.2	4

(15+10)=[25]

4. A population of 360 households (numbered 1 to 360) in Baltimore is arranged alphabetically in a file by the surname of the head of the household . Households in which the head is nonwhite occur at the following numbers : 28 , 31-33 , 36-41, 44,45 ,47 ,55 , 56 ,58 ,68 ,69 ,82 ,83 , 85 ,86 , 89-94 ,98 ,99 ,101 ,107-110 ,114,154 ,156, 178 ,223 ,224 ,296 , 298-300 , 302-304 ,306-323 , 325-331,333,335-339 ,341 ,342 .(The nonwhite households show some "clumping" because of an association between surname and colour .)
Compare the precision of a 1-in-8 systematic sample with a simple random sample of the same size for estimating the proportion of households in which the head is nonwhite .

[25]

5. (a) Describe how you would select a PPSWR sample in n draws by Lahiri's method . Show that a sample selected according to this method is really a PPS sample .
(b) If N is not a multiple of n , what are the shortcomings of linear systematic sampling ? Describe how you can either modify the sampling procedure or suggest a suitable method of estimation so as to get rid of the shortcomings in case N is not a multiple of n .

(13 +12) =[25]

INDIAN STATISTICAL INSTITUTE

Mid- Semester Examination: 2008-2009

B. Stat. III Year

Statistical Inference I

Date: 09.9.08 **Maximum Marks: 30** **Duration: 2 hours**

All answer should be complete and rigorous and to the point. If you use a result proved in the class, state it clearly. Answer all questions.

1. Let $\{X_n\}_{n\geq 1}$ be iid $N(\theta;\sigma^2)$ where $0<\sigma^2<\infty$ is known and the mean θ is unknown, $-\infty<\theta<\infty$.

(a) Find the MVUE T of $P_\theta(X_1\leq u)$ where u is known and T is based on $X_1,.....,X_n$. Find the MVUE of $Var_\theta(T)$ which is a function of $X_1,.....,X_n$; write down the MVUE as explicitly as possible.

[3+5=8]

(b) Is T consistent?

[1]

2. Let $X_1,.....,X_n$ be iid $U(\theta_1-\theta_2,\theta_1+\theta_2)$. Does there exist an MVUE of θ_1/θ_2 ?

[6]

3.(a) Let $X_n\rightarrow^{P} c$, and $a_n\rightarrow a$, $b_n\rightarrow b$. Show that

$$a_nX_n+b_n\rightarrow^{P}(ac+b)$$

[2]

(b) Show that under some assumptions (to be stated by you), the MVUE of $g(\theta)$ is consistent for $g(\theta)$.

[2+3=5]

4. Let $X_1,.....,X_n$ be iid $N(0;1)$. Show that S and $(X_{(n)}-X_{(1)})/(X_{(n)}-X_{(n-1)})$ are independent where $S^2=\sum_{i=1}^{n}(X_i-\bar{X}_n^2)$.

[3]

5. Let $X_1,.....,X_{2n+1}$ be iid $N(\mu;\sigma^2)$. Find

$$E(X_n^*|\bar{X}_n)$$

where X_n^* is the median of $X_1,.....,X_{2n+1}$.

[5]

***

INDIAN STATISTICAL INSTITUTE

Mid Term Examination: (2008-2009)

ELECTIVE GEOLOGY

Date : 12.09.2008 Maximum Marks :40 Duration : 2 hours

B. Stat. - III yr.

1. Write short notes on any three of the following topics --------- 5x3= 15

   a. Bowen's Reaction series
   b. Mineral Stability series
   c. Geological Time Scale
   d. Silicate minerals
   e. Moh's scale of hardness

2. Distinguish between the following (any eight) ----------2.5x8=20

   a. Current ripple and wave ripple
   b. Plutonic and volcanic rocks
   c. Granite and Sandstone
   d. Mohorovicic discontinuity and Gutenberg discontinuity
   e. Feldic and Mafic minerals
   f. Laminar flow and turbulent flow
   g. True dip and apparent dip
   h. Stress and strain
   i. Anticlines and antiforms
   j. Class I and Class III folds
   k. Dip slip faults and strike slip faults
   l. Triclinic system and Monoclinic system of crystals

3. What is a Mineral? What is a rock? What do you understand by the term "Rock cycle"?—1.5+1.5+2=5

OR

4. Illustrate using block (3D) diagrams, how folds and faults can produce repetition of strata. ---------- 5

Indian Statistical Institute

First Semestral Examination: (2008–2009)

B.Stat.(Hons.) – III year

Economics III

Date: 12/9/2008 *Maximum Marks –50* *Duration: 2 hours*

Answer ***any two*** questions.

1. (a) State and explain the assumptions underlying the Classical Linear regression Model (CLRM).

(b) Consider a simple regression model $y_i = \alpha + \beta x_i + \varepsilon_i$, $i=1,2,...,n$, for which all classical assumptions of CLRM are satisfied except that $E(\varepsilon_i) = \lambda z_i$, where z_i is non-stochastic and such that $\sum (x_i - \bar{x})(z_i - \bar{z}) = 0$. Determine the bias (if any) in estimating the parameters by OLS.

(c) Show that $E(e'e) = \sigma^2 (n-k-1)$, where e is the vector of OLS residuals, k is the number of regressors and $\text{var}(\varepsilon_i) = \sigma^2$, $i=1,2,...,n$.

[10+8+7=25]

2. (a) What do you mean by a "dummy variable"?

(b) A regression equation explaining household expenditure on recreation (y) as a function of income (x) was estimated as

$$y_t = \underset{(10)}{-25} + \underset{(16)}{35D_{t2}} + \underset{(15)}{40D_{t3}} - \underset{(8)}{15D_{t4}} + \underset{(0.02)}{0.05x_t} + e_t$$

where figures in parentheses are the standard errors, D's are the quarterly dummies with the 1st quarter (January-March) left out.

Determine the values of the estimated coefficients given that

(i) the equation contains four seasonal dummies and no constant term

(ii) the equation contains four seasonal dummies and a constant term, but the coefficients of the seasonal dummies represent deviations from the annual average and their sum id equal to zero.

(c) What is dummy variable trap?

P.T.O

(d) 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,

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

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.

[2+8+5+10 = 25]

3. (a) What is multicollinearity?

(b) Describe the procedure of detecting multicollinearity using 'condition number' and 'variance proportions'.

(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]

4. (a) What are the different types of specification errors that may arise in formulating a regression model?

(b) Suppose the model $y = X_{n\times(k+1)}\beta + \varepsilon$ has been misspecified as $y = Z_{n\times(r+1)}\delta + u$, where $r<k$. Then show that the OLS estimator $\hat{\delta}$ is a biased estimator of β, and that the estimate of the residual variance from the misspecified model is biased upwards.

(c) Suppose in the regression model $y = X_{n\times(k+1)}\beta + \varepsilon$, $E(\varepsilon\varepsilon') = \Omega$. Which assumption of the CLRM does it violate? How do you obtain an unbiased and efficient estimator of β?

[4+12+9=25]

INDIAN STATISTICAL INSTITUTE

MID-SEMESTER EXAMINATION

Introduction to Anthropology & Human Genetics
B III

Date: 12.9.08 **Marks : 40**

Answer any four questions [Duration: 2 hours]

1. Choose the correct answer from the options given in (a) through (e) [2 X 5 = 10]

(a) Anthropology is defined as the :
- (i) study of man in its totality
- (ii) study of animals of different kinds
- (iii) study of primates

(b) Biological anthropology is defined as :
- (i) study of human evolution, biological variation and adaptation
- (ii) study of physiological variation in man
- (iii) study of species differentiation

(c) social anthropology has many overlaps with :
- (i) psychology
- (ii) sociology
- (iii) geography

(d) Paleoanthropology deals with:
- (i) human biology and culture of the past societies
- (ii) future prediction of past populations
- (iii) spatial distribution of human populations

(e) Medical anthropology incorporates:
- (i) studies related to people's perspectives of health and illness
- (ii) study of physical measurements
- (iii) study of languages

2. List the characteristic human features that distinguish humans from all animals. [10]

3. (a) Plio-pleistocene hominid sites are found in which countries ? [3]

(b) How many genera of hominids are found in these countries? [3]

(c) Provide the time range (in million years ago) of their first appearance. [2]

(d) How many species in each genus of hominid have been found in these countries? [2]

P.T.O

4. Fill-in the blanks with the correct option in case of (a) through (e) [2 X 5= 10]

(a) Stone tools are not usually directly associated with .....
   (i) Australopithecus
   (ii) Homo habilis
   (iii) Homo erectus
   (iv) Homo neanderthalensis

(b) The earliest stone tool industry is called the .....
   (i) Cambrian
   (ii) Chopper-chopping
   (iii) Oldowan
   (iv) Assemblage

(c) The long bone in man that forms the upper arm is called .....
   (i) Ulna
   (ii) Radius
   (iii) Humerus
   (iv) Fibula

(d) Smaller and less robustly built australopithecines are known as ..... ones
   (i) Docile
   (ii) Gracile
   (iii) Muscular
   (iv) Dwarf

(e) The earliest australopithecine species probably is .....
   (i) Ausralopithecus aferensis
   (ii) Australopithecus robustus
   (iii) Australopithecus africanus
   (iv) Australopithecus aethiopicus

5. (a) List 5 characteristic features of mammals [5]

   (b) List 5 characteristic features of primates [5]

6. Define the following terms: (a) Epidemiology; (b) Disease; (c) Adaptation; (d) Environment and (e) Demography [2 X 5]

INDIAN STATISTICAL INSTITUTE

First Semestral Examination: 2008–2009

B.Stat. (Hons.) 3rd Year. 1st Semester

Linear Statistical Models

Date: November 21, 2008 **Maximum Marks: 70** **Duration: 4 hours**

- This question paper carries 86 marks. Answer question no. 6 and as much as you can, from the test. However, the maximum you can score is 70.
- You should provide as much details as possible while answering a question. You have to state clearly any result stated (and proved) that may be needed to answer a particular question.

---

1. *Suppose* $\boldsymbol{Y} \sim \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, *where* $\mathbf{X}$ *is an* $n \times (p+1)$ *matrix with fixed (non-random) entries and having rank* $p+1$, $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$, $\sigma > 0$. Denote by $(\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\sigma}^2)^{\mathrm{T}}$, the MLE of $(\boldsymbol{\beta}^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$. It is known that $\hat{\boldsymbol{\beta}}$ is an unbiased estimate of $\boldsymbol{\beta}$. State and prove an appropriate result that demonstrates optimality of $\hat{\boldsymbol{\beta}}$ within a suitable class of unbiased estimates of $\boldsymbol{\beta}$. [9]

2. *Let* $Y_{ij} \sim \mathrm{N}(\mu_i, \sigma^2)$, $j = 1, \ldots, n_i$, $i = 1, \ldots, k$, *be independent.*

(a) Let $\mu_i = \beta_0 + \beta_1 x_i$, $i = 1, \ldots, k$, for unknown parameters $\beta_0, \beta_1 \in \mathbb{R}$, and known constants x_i's, not all equal. Obtain the least squares estimate of $(\beta_0, \beta_1)^{\mathrm{T}}$.

(b) Under a suitable condition on the n_i's to be stated by you, develop a suitable test for testing

$$\mathrm{H}_0 : \mu_i = \beta_0 + \beta_1 x_i,\ i = 1, \ldots, k, \text{ for unknown parameters } \beta_0, \beta_1 \in \mathbb{R}$$
$$\text{versus} \quad \mathrm{H}_1 : \mathrm{H}_0 \text{ is false,}$$

where x_i's are known constants, not all equal. [7+10 = 17]

3. *Consider a balanced two-way ANOVA model, with single observation in each cell.* Develop, with adequate reasons, a suitable test for interaction. [**Note.** You are required to derive the null distribution of the test you develop.] [11]

4. *Consider a two-way nested ANOVA model given by* $Y_{ijk} = \mu + \alpha_i + \eta_{ij} + \epsilon_{ijk}$, $k = 1, \ldots, N_{ij}$, $j = 1, \ldots, t_i$, $i = 1, \ldots, s$, ϵ_{ijk}'s $\overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$. Develop the ANOVA table for this set-up. [**Note.** You are required to provide in the table **(a)** the sources of variation, **(b)** the degrees of freedom of the sum of squares, **(c)** matrix-theoretic and algebraic expressions of the sum of squares, and **(d)** expectations of the corresponding mean squares – both matrix-theoretic and algebraic expressions.] [13]

P.T.O.

5. *Consider a balanced one-way ANOVA model given by* $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $j = 1, \ldots, N$, $i = 1, \ldots, 4$, ϵ_{ij}'s $\overset{i.i.d.}{\sim}$ $N(0, \sigma^2)$. Consider the problem of simultaneous testing of the hypotheses

$$H_0^{(1)} : \alpha_1 - \alpha_2 = 0 \text{ versus } H_1^{(1)} : H_0^{(1)} \text{ is false,}$$

and

$$H_0^{(2)} : \alpha_3 - \alpha_4 = 0 \text{ versus } H_1^{(2)} : H_0^{(2)} \text{ is false,}$$

with experimentwise error rate not exceeding a given number $\alpha \in (0, 1)$.

(a) Describe a solution to this problem based on Scheffe's method.

(b) Describe a solution to this problem based on least significance method.

(c) Explain how you can compare the solutions in (a) and (b) above. [5+5+5 = 15]

6. *Consider a balanced two-way ANOVA model without interaction given by* $Y_{ijk} = \mu + \alpha_i + \eta_j + \epsilon_{ijk}$, $k = 1, \ldots, N$, $i = 1, \ldots, s$, $j = 1, \ldots, t$, ϵ_{ijk}'s $\overset{i.i.d.}{\sim}$ $N(0, \sigma^2)$. Consider the problem of testing simultaneously the hypotheses

$$H_0^{(\alpha)} : \alpha_1 = \cdots = \alpha_s \text{ versus } H_1^{(\alpha)} : H_0^{(\alpha)} \text{ is false,}$$

and

$$H_0^{(\eta)} : \eta_1 = \cdots = \eta_t \text{ versus } H_1^{(\eta)} : H_0^{(\eta)} \text{ is false.}$$

Suppose that the two level-α F-tests are performed and the simultaneous hypothesis is rejected if at least one of the level-α F-test rejects the corresponding hypothesis. Show that the experimentwise error rate is at most $1 - (1 - \alpha)^2$. [8]

7. *Consider a one-way ANOVA model with a single covariate given by* $Y_{ij} = \mu + \alpha_i + \xi x_{ij} + \epsilon_{ij}$, $j = 1, \ldots, N_i$, $i = 1, \ldots, t$, ϵ_{ij}'s $\overset{i.i.d.}{\sim}$ $N(0, \sigma^2)$. Assume for every $i = 1, \ldots, t$, not all of the x_{ij}'s are equal.

(a) Show that any contrast in the α's is estimable.

(b) Find the BLUE of a contrast $\sum_{i=1}^{t} \lambda_i \alpha_i$, where $\sum_{i=1}^{t} \lambda_i = 0$.

(c) Find the variance of the BLUE obtained in (b) above. [4+4+5 = 13]

* * * * * * *Best of Luck!* * * * * * *

INDIAN STATISTICAL INSTITUTE
First Semester Examination: 2008-09
B. Stat. III Year
Statistical Inference I

Date: 25.11.08 **Maximum Marks: 70** **Duration: $3\frac{1}{2}$ Hours**

(a) All answers should be complete, rigorous and to the point. If you are using a result proved in the class, state it clearly.

(b) Answer any five questions. Maximum you can score is 70.

1 (a) Let T be an unbiased estimator of $g(\theta)$ with $E_\theta(T^2) < \infty$ for each θ. Show that T is an MVUE of $g(\theta)$ iff $E_\theta(T\,Z) = 0$ for each θ whenever Z is a zero function with $E_\theta(Z^2) < \infty$ for each θ.

[6]

(b) Give two applications of the above result.

[6+6=12]

2. Let $X_1,\ldots,X_n$ be iid from $N(\theta, \sigma^2)$ where θ and σ^2 are both unknown. Let u be a known constant.

(a) Find the MVUE of $P_\theta(X_1 \le u)$; give details.

[12]

(b) Find the MVUE of the density function of X_1 at u.

[6]

3. Let $X_1,\ldots,X_n$ be iid from the exponential distribution

$$f_\theta(x) = \frac{1}{\theta}\exp(-x/\theta)\,I_{(0,\infty)}(x).$$

Let $g(\theta) = P_\theta(X_1 > u)$ where u is a known constant.

Find the MVUE of $g(\theta)$; give details.

[18]

4. (a) Define a minimal sufficient statistic.

[1]

(b) Show that a boundedly complete, sufficient statistic is minimal sufficient.

[9]

(c) Is the converse of (b) true ? Justify.

[8]

5. (a) Write down the Bhattacharya lower bound to the variance of an unbiased estimator of $g(\theta)$, stating clearly the assumptions.

[5]

(b) Using (a). find the MVUE of θ^2 based on $X_1,\ldots,X_n$ where $X_1,\ldots,X_n$ are iid $N(\theta, \sigma^2)$, σ^2 being known.

[7]

P.T.O.

(c) Let $X_1,...,X_n$ be independent, and X_n follow $N(\theta, \sigma_n^2)$ where the σ_n^2 are known and satisfy $\sum_{n=1}^{\infty} \sigma_n^{-2} < \infty$. Let $g(\theta)$ be a non-constant function of θ. Show that there does not exist any consistent estimator of $g(\theta)$.

[6]

6. (a) Let $X_1,..,X_n$ be iid following U $(0,\theta)$, $0 < \theta < \infty$. Find the UMP level α test for testing $H_0 : \theta = \theta_0$ VS. $H_1 : \theta \neq \theta_0$ $(0 < \alpha < 1)$.

[6]

(b) Let $\{ f_\theta : \theta \in \Theta \}$ be MLR in T. Consider the testing problem $H_0 : \theta \leq \theta_0$ Vs. $H_1 : \theta > \theta_0$. Let $0 < \alpha < 1$. Consider the following statements :

(i) There exists a test function $\Phi(X)$ such that $\Phi(x) = 1.\ \gamma, 0$ according as $T(x) >, =, < c$, and $E_{\theta_0}(\Phi(X)) = \alpha$.

(ii) The test function defined in (i) is UMP level α for the above testing problem.

(iii) The power function $\beta(\theta)$ of the test function defined in (i) is a strictly increasing function of θ, provided $0 < \beta(\theta) < 1$.

(iv) For any $\theta < \theta_0$, the test function defined in (i) minimizes the probability of Type I error among all tests satisfying the condition $E_{\theta_0}(\Phi) = \alpha$.

Assuming the validity of (i), prove (ii), (iii) and (iv).

[12]

7 (a) Show that under suitable assumptions (to be stated by you), the MP tests are consistent.

[1+5=6]

(b) Let $X_1,....,X_n$ be iid following Poisson (θ), $0 < \theta < \infty$. Let the loss function be $L(\theta, a) = (\theta - a)^2/\theta$.

Using the Cramer – Rao Inequality, show that $\overline{X}$, the sample mean, is admissible and minimax.

[6+6=12]

8. (a) Show that if a minimal complete class of decision rules exists, then it consists of all admissible rules.

[10]

(b) Find the MLE of $\theta = (\alpha, \beta)$ where $X_1,...,X_n$ are iid having the p.d.f.

$$f_\theta(x) = \frac{1}{\beta} \exp\left(-(x - \alpha)/\beta\right) I_{[\alpha,\infty]}(x).$$

[5]

(c) Give an example where the MLE is inconsistent.

[3]

*****

INDIAN STATISTICAL INSTITUTE

Semestral Examination : 2008-2009

B. Stat. - III Year

Differential Equation

Date : 28.11. 2008 | Maximum Score : 60 | Time :3 Hours

This paper carries a total of 70 marks. Answer as many questions or parts thereof. But the maximum you can score is 60. Any result that you use should be stated clearly.

(1) Let $y_1(x)$ and $y_2(x)$ be two linearly independent solutions of

$$y'' + P(x)y' + Q(x)y = 0,$$

where $P(x)$ and $Q(x)$ are continuous functions.

- Prove that $y_1(x)$ and $y_2(x)$ cannot have a common zero.
- Suppose x_1 and x_2 are two successive zeros of $y_2(x)$. Prove that $y_1(x)$ must have a zero in between x_1 and x_2.

[2+5=7]

(2)
- Find the Frobenius series solutions of Gauss's hypergeometric equation

$$x(x-1)y'' + [c - (a+b+1)x]y' - aby = 0,$$

where $a,\ b,\ c \in \mathbb{R}$; c being non-integral.
- Discuss, in necessary details, the convergence aspects of these series solutions.

[14+2=16]

(3) For the Bessel's equation of order n (real)

$$x^2y'' + xy' + (x^2 - n^2)y = 0,$$

find out the Frobenius series solution $J_n(x)$ about the regular singular point which is bounded at the origin ($x = 0$).

[9]

(4) Prove that
- $\frac{d}{dx}[x^n J_n(x)] = x^n J_{n-1}(x)$,
- $\frac{d}{dx}[x^{-n} J_n(x)] = -x^{-n} J_{n+1}(x)$,
- $J_{n-1}(x) + J_{n+1}(x) = \frac{2n}{x} J_n(x)$.

[3+3+2=8]

(5) Using the integral representations

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt;\ x > 0,\ B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt;\ x, y > 0,$$

prove that $B(x+1, y) + B(x, y+1) = B(x, y)$.

[8]

P. T. O

(6) Using the method of variational calculus, deduce Euler's differential equation for extremization (stationarity) of the integral

$$I = \int_0^4 \{xy' - (y')^2\} dx,$$

for fixed values of $y(0)$, $y(4)$ and find out the 'stationary' function $y(x)$ with boundary conditions $y(0) = 0$, $y(4) = 3$.

[6+2 =8]

(7)
- Find out the Laplace transform of the function $\sin at$.
- If $L[f(t)]$ denotes the Laplace transform of a function $f(t)$ then prove that

$$L[f''(t)](s) = sL[f'(t)](s) - sf(0) - f'(0).$$

- Solve the initial-value problem

$$y'' + y = \sin t;\ y(0) = 1,\ y'(0) = 2,$$

using the method of Laplace transform.

[4+4+6= 14]

INDIAN STATISTICAL INSTITUTE
First Semester Examination : 2008-09
B.Stat.(Hons.) III Year
Sample Surveys

Date : 02.12.2008 **Maximum Marks :100** **Duration : 3 Hours**

Answer Question No. 6 and ANY THREE questions from the rest . Marks allotted to each question are given within the parentheses .Standard notations and symbols are used.

1. After the decision to take a simple random sample had been made , it was realized that y_1 would be unusually low and y_N would be unusually high . For this situation , consider the following estimator of the population mean $\bar{Y}$

$$\begin{aligned}\hat{\bar{Y}}_S &= \bar{y}+c \quad \text{if the sample contains } y_1 \text{ but not } y_N \\ &= \bar{y}-c \quad \text{if the sample contains } y_N \text{ but not } y_1 \\ &= \bar{y} \quad \text{for all other samples}\end{aligned}$$

where c is a constant .

(a) Prove that $\hat{\bar{Y}}_S$ is an unbiased estimator of $\bar{Y}$.

(b) Prove that $\mathrm{Var}\left(\hat{\bar{Y}}_S\right)=(1-f)\left[\frac{S^2}{n}-\frac{2c}{(N-1)}(y_N-y_1-nc)\right]$ where S^2 is the population variance with divisor $(N-1)$ and $f=\frac{n}{N}$.

(c) Show that $\mathrm{Var}\left(\hat{\bar{Y}}_S\right)<\mathrm{Var}\,(\bar{y})$ if $0<c<\frac{(y_N-y_1)}{n}$. (10+10+5)=[25]

2. (a) A sampler proposes to take a stratified random sample . He expects that his field costs 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 $\mathrm{Var}(\bar{y}_{st})$.

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

P.T.O.

(iii) 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]

3. (a) Show that if N is a multiple of n , the variance of the estimated mean $\bar{y}_{sy}$ based on a linear systematic sample of size n can be written as

$$\text{Var}(\bar{y}_{sy}) = \frac{\sigma^2}{n}[1+(n-1)\rho_c]$$

where ρ_c is the intra-class correlation coefficient and σ^2 is the population variance with divisor N .

Show that $-\frac{1}{n-1} \le \rho_c \le 1$.

(b) Explain why it is not generally possible to unbiasedly estimate the sampling variance of the estimated mean based on a single systematic sample .

(8+7+10)=[25]

4. 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]

5.(a) Obtain an approximate expression for the bias and mean square error of the ratio estimator of the population mean based on SRSWOR sampling scheme .

(b) Derive an approximate expression for the mean square error of the linear regression estimator of the population mean based on SRSWOR sampling scheme .Compare it with that of the ratio estimator of the population mean based on SRSWOR sampling scheme .

(10+10+5)=[25]

6. 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 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			

(10+5+10)=[25]

Indian Statistical Institute

First Semestral Examination: (2008–2009)

B.Stat.(Hons.) – III year

Economics III

Date: 9/1/2009 *Maximum Marks 100* *Duration:* $2\frac{1}{2}$ *hours*

Part I

This part of the Question Paper carries 88 marks. The maximum you can score is 80.

Answer any ***four*** questions.

1. (a) Consider a linear regression model $y = X\beta + \varepsilon$, where all assumptions of a CLRM are satisfied except that $E(\varepsilon\varepsilon') = \Omega$, a positive definite matrix. What are the consequences of applying OLS to this model? Describe a method that will yield a more efficient estimator than the OLS estimator.

   (b) Describe a test for testing the presence of first order autocorrelation in a given time series.

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

   [10 + 6 + 6 = 22]

2. (a) Consider the following model:

$$C_t = \nu_0 + \nu_1 Y_t + \varepsilon_{1t} \quad \text{(Consumption function)}$$
$$I_t = \delta_0 + \delta_1 Y_{t-1} + \varepsilon_{2t} \quad \text{(Investment function)}$$
$$Y_t = C_t + I_t \quad \text{(Income identity)}$$

   Reduce the three-equation model to a single equation and examine if any assumption of the CLRM is violated. Give reasons for your answer.

   (b) Show that the OLS estimator of the parameters is biased but consistent.

   (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.

   [7 + 7 + 8 = 22]

P.T.O

3. (a) What are distributed lag models? Define "impact multiplier" and "equilibrium multiplier".

   (b) Describe the geometric lag model and rationalize the model in terms of (i) Adaptive expectation model and (ii) Partial adjustment model.

[6 + 16 =22]

4. (a) Describe the general structural form of a simultaneous equations model (SEM) explaining all the terms you use. Obtain the reduced form of the equation system.

   (b) Write down the general rules (order and rank conditions) of determining the identification status of a structural equation.

   (c) Discuss the identifiability status of each of the equations in the following SEM.

$$y_t = \beta\, x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + \varepsilon_{1t}$$

$$x_{3t} = \alpha_1 y_t + \alpha_2 z_{1t} + \alpha_2 z_{2t} + \varepsilon_{2t}$$

[8 + 8 + 6=22]

5. (a) Describe the Two Stage Least Squares (2SLS) procedure for estimating a SEM.

   (b) Explain the Instrumental Variables (IV) approach to estimating a single equation of a SEM. Derive the 2SLS estimator as an IV estimator.

   (c) Describe the K-class estimator. Derive the condition under which the K-class estimator is consistent.

[8 + 9 + 5 =22]

Part II

Practical assignment [20]

INDIAN STATISTICAL INSTITUTE
FIRST Semester Examination: 2008-09
B.Stat. (Hons) III year
Introduction to Anthropology and Human Genetics

Date: 9.1.2009 Maximum Marks: 40 Duration: 2.30 Hours

1. **What is Hardy-Weinberg Equilibrium? Why is that it is so important?** (5 x 2)

or

1a. **What is population structure and population composition? Write briefly about the interrelationship between the two?** (5 x 2)

2. **Colour blindness in man is due to a sex-linked recessive gene. A survey of 500 men from a local population revealed that 10 were colour blind.** (5 x 2)
   (a) What is the gene frequency of the normal allele in the population?
   (b) What percentage of the females in this population would be expected to be normal?

3. **The allele frequencies at the ABO locus in a population are 0.27, 0.08 and 0.65 for the alleles corresponding to blood types A, B and O. The alleles for blood types A and B are co-dominant and the allele for blood type O is recessive. Calculate the genotype frequencies at this locus assuming Hardy-Weinberg equilibrium?** (5 x 2)

4. **Define Environment? What are the different aspects of the environment? Illustrate how different aspects of environment affect biological functions in human groups?** 10

or

4a. **Write short notes on any two of the following?** (5 x 2)
   a. Human Fertility **b. Culture** **c. High altitude**

INDIAN STATISTICAL INSTITUTE

First Semester Examination: 2008 - 2009

B.Stat. - III Year

Geology

Date:9.1..09 **Maximum marks 50** **Duration 2 hours**

1. Write short notes on the following topics (any three) (5 X 3 = 15)
   (a) Island arc
   (b) Continental drift theory
   (c) Major postulates of Darwin's theory of organic evolution
   (d) Different types of unconformities observed within the lithological successions of the earth.

2. Distinguish between the following pairs (any three) (5X 3 = 15)
   (a) Constructive and destructive plate boundaries
   (b) Continental rift and mid oceanic rift
   (c) Monophyletic group and polyphyletic group
   (d) Orthorhombic and isometric crystal systems.

3. Answer **any four** of the following questions: (5X 4 = 20)

   (a) "Present is the key to the past"- Justify the validity of the statement. (5)

   (b) Name a few primary sedimentary structures that are helpful in finding the direction of younging in a sedimentary succession. Describe one of those structures. (2+3=5)

   (c) Why is quartz considered as the most abundant mineral in clastic sedimentary rocks? (5)

   (d) Why is ^{14}C not used to date very old rocks like the Precambrian rocks? (5)

   (e) "Minerals do not necessarily have a fixed and absolute chemical formulae" – Explain with at least one example. (5)

   (f) Why do fossils generally provide the relative ages of their host rocks? (5)

***

INDIAN STATISTICAL INSTITUTE

Mid-semestral Examination 2008 - 2009

B.Stat Third year : Stochastic processes

Date 27- 02- 2009 Maximum marks 40 Duration : 2 hrs

Justify your answers.

1. We have a Markov chain with state space $\{0, 1, 2 \cdots, 20\}$ and transition matrix (p_{ij}). Suppose that the state 2 leads to the state 5. Show that for some $n \leq 20$; $p_{2,5}^{(n)} > 0$.

[3]

2. Let $S = \{-1, +1\}^{100}$ be the set of all sequences of $+1$ and -1 of length 100. Here is the transition mechanism for a Markov chain with state space S. If you are at $\sigma = (\sigma_1, \cdots, \sigma_{100}) \in S$ then you pick two integers $1 \leq i \neq j \leq 100$ at random and interchange σ_i and σ_j to get the new state. Describe all the communicating classes, explain which of these classes are recurrent and for each recurrent class calculate its period.

[5+3+3]

3. Let S be the set of all non-negative integers. I have a number $\lambda > 0$ and a coin whose chance of heads in a single toss is p, where $0 < p < 1$. I have a box containing some balls. If there are i balls this morning, I do the following in the evening. I toss my coin (independently) for each ball; keep the ball if heads up and remove the ball if tails up. Also independently I add certain number of balls and the number of balls added is Poisson with parameter λ. Every day I do this independently. Let X_n be the number of balls on the n-th morning.

   Calculate $P(X_{n+1} = j | X_n = i) = p_{ij}$, say. Calculate $p_{0j}^{(n)}$. You must simplify your answer. Calculate $\lim_n p_{0j}^{(n)}$. Explain why this limit is the same as $\lim_n p_{ij}^{(n)}$ for each i.

[5+10+2]

4. Let P be an irreducible idempotent (that is, $P^2 = P$) stochastic matrix. Let i and j be two states. Show that $p_{ij} = p_{jj}$. (Hint: One idea is Mean Ergodic theorem).

[4]

P.T.O

5. Define the terms 'essential state' and 'transient state'.

In a finite state Markov chain show that a transient state must be inessential.

State with reasons if the above statement remains true in an infinite state chain.

[4+3+2]

INDIAN STATISTICAL INSTITUTE
MID- SEMESTERAL EXAMINATION: 2008 -2009
Subject: Design of Experiments
B. Stat. III Year

Date of Examination: 02.03.09 Maximum Marks: 80 Duration: 2½ hours

1. Answer all questions
2. The Paper carries 90 Marks But the maximum you can score is 80
3. Give to the point answers. Marks will be deducted for lengthy answers

1) a) What are the different components of experimental error? Explain with examples.

b) Explain the meaning of "*Block what you can randomize what you can't*" with an example.

[2.5 x 3 + 2.5 = 10]

2) To study the effect of different diets on water consumption by animals' an experiment was conducted. Water consumptions were measured over a standard period of time for animals on three diets: Standard, New-1, and New-2. The data are as follows:

Animals ->	1	2	3	4	5	6	7	8	9	10	Total
Standard	14.2	10.3	10.7	9.7	13.1	20.0	10.7	14.7	12.2	11.4	127.0
New-1	15.7	9.3	12.4	10.0	14.7	21.8	12.3	16.0	13.0	12.7	137.9
New-2	16.3	10.4	11.9	13.2	13.7	24.9	13.3	16.2	14.3	13.0	147.2
Total	46.2	30.0	35.0	32.9	41.5	66.7	36.3	46.9	39.5	37.1	412.1

$14.2^2 + ... + 13.0^2 = 6023.17$

$127^2 + ... + 147.2^2 = 56813.3$

$46.2^2 + ... + 37.1^2 = 17966.95$

(a) What design was used? Give justification

(b) Write down the model explaining the terms

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

(d) Carry out analysis of variance

(e) Draw conclusions

[3+1+1+5+2 = 12)]

3) 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.

d) Also answer specific questions, if any, in each case

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.

[6]

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.

[4]

iii. Grain size of a particular type of aluminium product is of concern. The company produces the product in four furnaces. Each furnace is known to have its own unique operating characteristics. The process engineers suspect that stirring rate affects the grain size and decided to experiment with four different stirring rates. It was also suspected that the line voltage may also have an effect but is uncontrollable, however it could be measured. Furnace temperature was decided to be fixed at a suitable level.

[6]

iv. A market research team designs an experiment to determine which factors will most influence customers to purchase a product. They run a survey which asks questions about several levels of price, packaging and delivery. Suggest two factors not considered in this study that might affect the outcome? Justify.

[6 + 2 = 8]

v. An industrial engineer wants to investigate the effect of four assembly methods (A, B, C, D) 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. The engineer also suspected that the workplaces used by the four operators may represent another source of variation.

[8]

4) An experimenter wanted to use a Latin Square Design (LSD) but instead, used the following design

Row	Column 1	2	3
1	A	B	C
2	B	A	C
3	C	A	B

Are all treatment contrasts estimable? Give reasons for your answer.

[4]

5) The yield of a chemical process was measured using five batches of raw material, five acid concentrations and five standing times (A, B, C, D, E)

Batch	Acid Concentration 1	2	3	4	5
1	A = 26	B = 16	C = 19	D = 16	E = 13
2	B = 18	C = 21	D = 18	E = 11	A = 21
3	C = 20	D = 12	E = 16	A = 25	B = 13
4	D = 15	E = 15	A = 22	B = 14	C = 17
5	E = 10	A = 24	B = 17	C = 17	D = 14

Note: Sum of all 25 observations = 430
Sum of squares of all 25 observations = 7832
Sum of squares of Batch totals = 37030
Sum of squares of Acid Concentration totals = 37102

(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 ($\alpha = 0.05$)

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

(e) draw conclusions

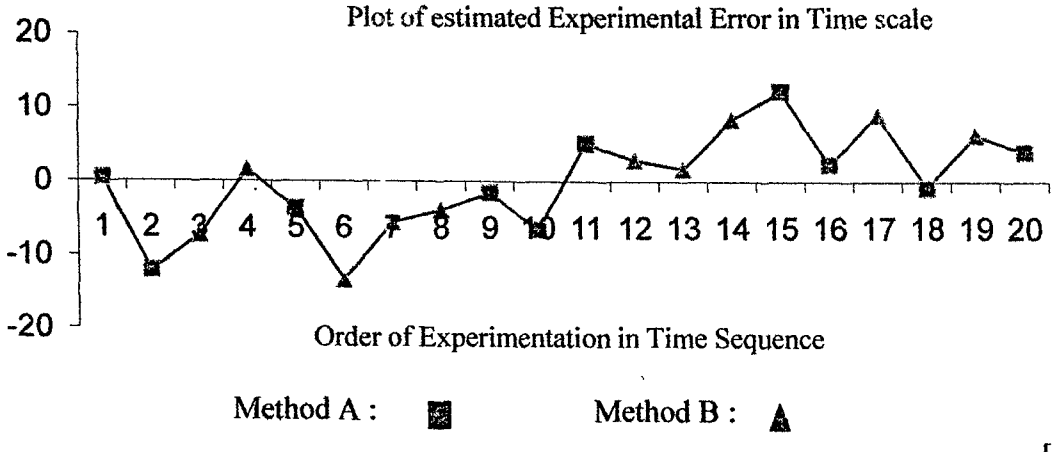
Note: LSD = $t_{\alpha/2,\upsilon}\sqrt{(2MS_E/n)}$, υ is the error d.f

[3+1+10+8+2 = 24)]

6) 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 and comment on the relevant model assumptions. Can you find out what possibly went wrong?



[4+4 = 8

Indian Statistical Institute

Mid Semestral Examination: (2008–2009)

B.Stat.(Hons.) – III year

Statistics Comprehensive

USE SEPERATE ANSWERSCRIPTS FOR GROUP A & GROUP B.

Group A

(Answer all questions)

Date: 4.03.09 *Maximum Marks –50* *Duration: 1:30* ***hours***

1. Let $X_1, X_2, \ldots, X_n$ be iid Bernoulli trials with success probability p. We would like to test

$$H_0: p = \frac{1}{2} \quad vs \quad H_1: p = \frac{3}{4}$$
$$\text{with } \alpha = .05 \quad and \quad p \le .05$$

(a) Is it possible to carry out the test for

(i) $n = 10$?
(ii) $n = 1000$?
(iii) Can you make a comment on the basis of the above results?

(b) Can you present 'some' data in the above set up so that while testing

$$H_0: p = \frac{1}{2} \quad vs \quad H_1: p = \frac{3}{4}$$
$$and$$
$$H_0: p = \frac{3}{4} \quad vs \quad H_0: p = \frac{1}{2}$$

H_0 gets rejected in both the cases.

[5+5+5+10=25]

2. Consider a district with 5 villages having population 50, 100, 100, 100 & 50 households respectively. First two villages are chosen using ppswr and from each of the chosen villages, 5 households are chosen using SRSWR. In each household, the number of children of age less than or equal to 5 is recorded. We would like to estimate the total number of such children in the district. The observations are

0, 3, 1, 1, 1 from village 2
1, 2, 2, 1, 1 from village 4

(a) What would be your estimate and estimate of the variance of your estimate?
(b) Explain π_i, π_{ij}'s in this context and compute them.

$[\overline{5+5} + \overline{5+10} = 25]$

P.T.O

(Group B)

Date : March 4, 2009 Full Marks : 50 Time : 1½ hours

Answer both the questions. Marks allotted to a question are indicated in brackets at the end.

Q 1. Consider a two-way classified data in p rows and q columns with y_{ij} as observation in cell (i,j). Assume the usual linear model with general mean effect μ, row effects α_i 's and column effects β_j's for analysis of the data. Suppose now that the observations in cells (r,c) and (r*, c*) are missing, where $r \neq r^*$ and $c \neq c^*$. Applying missing plot technique, obtain expressions for w_1^* and w_2^* in terms of the available observations. Also derive expressions for the BLUE's of $(\alpha_i - \alpha_{i^*})$, and their variances as well.

[12+7+6=25]

Q 2. An experiment on grain yield (y) for five varieties of a crop in three replicates (blocks) was conducted in order to see if the varieties differed in respect of the grain yields or not. To increase sensitivity of the experiment, data on a covariate (x), the number of plants per plot was also collected. The data is reproduced below. Using analysis of variance and covariance, analyse the data and draw conclusions. Also give the best estimates of the differences of the mean yields of the varieties, alongwith their standard errors.

[15+10=25]

Variety	Block 1		Block 2		Block 3	
	y	(x)	y	(x)	y	(x)
1	48.25	(227)	56.25	(226)	48.34	(259)
2	99.50	(248)	85.50	(218)	54.50	(234)
3	43.50	(249)	58.50	(256)	48.50	(270)
4	52.50	(264)	43.50	(252)	40.20	(248)
5	83.31	(271)	65.25	(263)	61.13	(259)

(a) [redacted]

(b) [redacted]

[redacted]

INDIAN STATISTICAL INSTITUTE
Mid-Semestral Examination
B. Stat. III year : 2008–2009
Database Management Systems

Date: 06-03-2009 Marks: 60 Time: 3 Hours

The questions are for 72 marks. The maximum marks you can get is 60.
Answer any part of any question. Answer all the parts of a question at the same place.

1. Zuji Records has decided to store information about musicians who perform on its albums (as well as other company data) in a database.

   Each musician that records at Zuji has an Id Number which is unique, a name, an address, and a phone number. Poorly paid musicians often share the same address, and no address has more than one phone. Each instrument used in songs recorded at Zuji has a unique identification number, a name (e.g., guitar, synthesizer, flute) and a musical key (e.g., C, B-flat, E-flat). Each album recorded on the Zuji label has a unique identification number, a title, a copyright date, a format (e.g., CD or MC), and an album identifier. Each song recorded at Zuji has a title and an author. Each musician may play several instruments, and a given instrument may be played by several musicians. Each album has a number of songs on it, but no song may appear on more than one album. Each song is performed by one or more musicians, and a musician may perform a number of songs. Each album has exactly one musician who acts as its producer. A musician may produce several albums, of course.

   Design a conceptual schema for Zuji and draw an ER diagram for your schema. The preceding information describes the situation that the Zuji database must model. Be sure to indicate all key and cardinality constraints and any assumptions you make. Identify any constraints you are unable to capture in the ER diagram and briefly explain why you could not express them.

   (16)

2. Consider the following schema:

   Suppliers(*sid*: integer, *sname*: string, *sadd*: string)
   Parts(*pid*: integer, *pname*: string, *color*: string)
   Catalog(*sid*: integer, *pid*: integer, *cost*: real)

   The key fields are underlined, and the domain of each field is listed after the field name. Therefore *sid* is the key for *Suppliers*, *pid* is the key for *Parts*, and *sid* and *pid* together form the key for *Catalog*. The *Catalog* relation lists the prices charged for parts by *Suppliers*. Write the following queries in relational algebra, tuple relational calculus, and domain relational calculus:

   (P.T.O)

(a) Find the *names* of suppliers who supply some red part.

(b) Find the *sids* of suppliers who supply every part.

(c) Find the *pids* of parts supplied by at least two different suppliers.

(6+6+6 = 18)

3. Consider the $Supplier - Parts - Catalog$ schema from the previous question. State what the following queries compute:

   (a) $\pi_{sname}((\pi_{sid,sname}((\sigma_{color=\text{'red'}}Parts) \bowtie (\sigma_{cost<200}Catalog) \bowtie Suppliers)) \cap (\pi_{sid,sname}((\sigma_{color=\text{'green'}}Parts) \bowtie (\sigma_{cost<200}Catalog) \bowtie Suppliers)))$

   (b) $\pi_{sname}(\pi_{sid}((\sigma_{color=\text{'red'}}Parts) \bowtie (\sigma_{cost<100}Catalog) \bowtie Suppliers))$

   (c) $\{T \mid \exists T1 \in Catalog(\exists X \in Parts((X.color = \text{'red'} \vee X.color = \text{'green'}) \wedge X.pid = T1.pid) \wedge T.sid = T1.sid)\}$

   (d) $\{\langle X \rangle \mid \langle X, Y, Z \rangle \in Catalog \wedge \forall \langle A, B, C \rangle \in Parts (C \neq \text{'red'} \vee \exists \langle P, Q, R \rangle \in Catalog(Q = A \wedge P = X))\}$

(2+2+2+2 = 8)

4. (a) Define an m-way search tree. What are the additional constraints you require to define a B-Tree? What is a 2-3 tree?

   (b) What is the maximum number of key values in a B-Tree of order m having height l?

   (c) Write down the insertion algorithm for a B-Tree.

   (d) Explain the insertion algorithm for a B-Tree of order 3 with the key values 78, 51, 36, 45, 88, 79, a, 99, 61, 53, 35 (arriving in this order), where a is the last two digit of your roll number. Explain each insertion with proper details.

   (e) Explain briefly how the B-Tree data structure can be efficiently used to implement indexing in a file.

((2+2+2)+6+6+6+6 = 30)

INDIAN STATISTICAL INSTITUTE

203 B. T. ROAD, KOLKATA - 700 108

SEMESTRAL EXAMINATION

06-05-2009 B III - Stochastic Processes Time: Three hours

Maximum marks you can score is 60.

Provide proper justifications.

1. Consider the discrete torus, $S = \{0, 1, \cdots, a-1\} \times \{0, 1, \cdots, b-1\}$, where addition is coordinate-wise, modulo a and modulo b respectively. From (x, y), you move to $(x-1, y)$ with probability $3/4$ and to $(x, y-1)$ with probability $1/4$.

   (a) When is the chain irreducible?

   (b) Show that it is aperiodic iff g.c.d.$(a, b) = 1$.

   (c) In the irreducible case, calculate the invariant distribution.

   [2+6+4 =12]

2. Consider the Markov chain with state space $\{0, 1, 2 \cdots\}$ and transition probabilities given by $p_{i,0} = \alpha_i$ and $p_{i,i+1} = 1 - \alpha_i$. Here $0 < \alpha_i < 1$ for each $i \geq 0$.

   (a) Show that the chain is irreducible and aperiodic.

   (b) Show that the chain is recurrent iff $\sum \alpha_i$ diverges.

   [2+10 =12]

3. Accidents in Kolkata follow a Poisson process with parameter λ, while accidents in Mumbai follow a Poisson proces with parameter μ. he two processes are independent.

   (a) Let X be the number of accidents in Mumbai that occurred between the second and third accidents of Kolkata. Calculate the distribution of X.

   (b) Show that the total number of accidents, Kolkata and Mumbai put together, is a Poisson process.

   (c) Given that the total number of accidents so far is four, what are the chances that at least three of them occurred in Kolkata.

   [5+5+2 =12]

P. T. O

4. A small barber shop operated by a single barber has room for at most two customers (including the one being served). Potential customers arrive at a Poisson rate of three per hour. Successive service times are independent exponential random variables with mean 1/4-th hour. Referring to the steady state, answer the following questions.

   (a) What is the average number of customers in the shop?

   (b) What proportion of potential customers are lost?

   (c) If the barber could work twice as fast, what proportion of potential customers would be lost?

   [5+2+5 =12]

5. Assume that X_t, the number of laptops in the hostel at time t, is a pure death process with death rates $\mu_n = n\mu$.

   (a) Explain what are Kolgomorov forward equations for a continuous parameter Markov chain (do NOT derive) and identifying clearly the parameters involved write down the equations for the process (X_t).

   b) Assume that $X_0 = N$. Show that

$$P(X_t = i) = \binom{N}{i} e^{-i\mu t}(1 - e^{-\mu t})^{(N-i)t} ; \qquad i = 0, 1, \cdots, N.$$

   [4+8=12]

6. (a) Define 'taboo probabilities'.

   (b) State and prove the first entrance decomposition for taboo probabilities.

   (c) State (do Not prove) the ratio limit theorem.

   [2+4+2 =8]

♣*Good Luck* ♠

INDIAN STATISTICAL INSTITUTE
Second Semestral Examination: 2008 -2009
Subject: Design of Experiments
B. Stat. III Year

Date of Examination: 15.05.09 Maximum Marks: 100 Duration: 3 hours

Note:
1. Answer all questions
2. The Paper carries 100 Marks but the maximum you can score is 80.
3. Assignments carry 20 marks.

What are the basic principles of experimental design? Explain their role in design and analysis of experiments.

[3 x 3 = 9]

An engineer wants to compare five treatments A, B, C, D and E. The experimental units after receiving the treatments are to be heat treated in a furnace. There are 5 racks in the furnace. The engineer approaches you with the following plan for the experimentation.

Rack 1:	A, A, B B, D
Rack 2:	A, C, C, C, D
Rack 3:	B, B, B, C, D
Rack 4:	A, B, B, C, E
Rack 5:	A, A, B, D, E

(a) What type of experimentation is this? Write down the underlying model. What are the necessary assumptions?

(b) What purposes do connectedness, orthogonality, and balance serve in a block design ? Indicate which of them are satisfied and which are not by the given experiment ?

(c) Modify the plan so that all the above properties are satisfied.

(d) How will you randomize while conducting the modified experiment?

[3+6+3+2 = 14)]

(a) State two alternative definitions of connectedness and show that they are equivalent.

(b) Derive a necessary and sufficient condition for a connected block design to be orthogonal.

[(8+6)) = 14)]

P.T.O

4) (a) Construct a pair of mutually orthogonal 8 x 8 Latin Squares.

(b) Explain an experimental situation where the pair of Latin squares in (a) could provide a suitable design taking care of all sources of variation. Also explain how to construct the design from such Latin Squares, and explain how the design should be randomized.

[(8+7)) = 15)

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

(a) What type of experiment will be appropriate ?

(b) Identify the generators and write down the defining relations. Write down the alia structures of the effects to be estimated.

(c) Write down the underlying model and assumptions.

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

(e) Give the treatment combinations.

(f) Prepare the table of signs to estimate the effects.

[1+5+2 +2+ 6+4 = 2

6) An experiment was run in a chemical industry in which following four variables were considered each at two levels.

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

The process engineer has run a single replicate of a 2^4 design for minimizing the impurity level of the product.

The industry has two similar facilities for producing the product. Only eight runs could be made in a single day. Accordingly the following two sets of treatment combinations were ru in the two different facilities and all the sixteen runs were completed in a day.

Facility-1:	(1)	ab	ac	bc	ad	bd	cd	abcd
Facility -2:	a	b	c	d	abc	abd	acd	bcd

The following table gives the design, the response and the effect estimates.

Trt No	A	B	C	D	Impurity (Coded)	Estimated Effects
1.	-	-	-	-	3.8	4.975
2.	+	-	-	-	5.5	0.025
3.	-	+	-	-	4.2	-1.725
4.	+	+	-	-	3	0.05
5.	-	-	+	-	7.3	0.975
6.	+	-	+	-	5.6	-0.1
7.	-	+	+	-	3	-0.75
8.	+	+	+	-	4.7	0.175
9.	-	-	-	+	7.4	2.175
10.	+	-	-	+	6.2	-0.1
11.	-	+	-	+	5.3	0.1
12.	+	+	-	+	6.5	-0.075
13.	-	-	+	+	7.9	-0.05
14.	+	-	+	+	9	0.025
15.	-	+	+	+	6.8	0.025
16.	+	+	+	+	5.4	-1.4

(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 normal probability plots. A normal probability plot of the factorial effects is given. Prepare the ANOVA table.

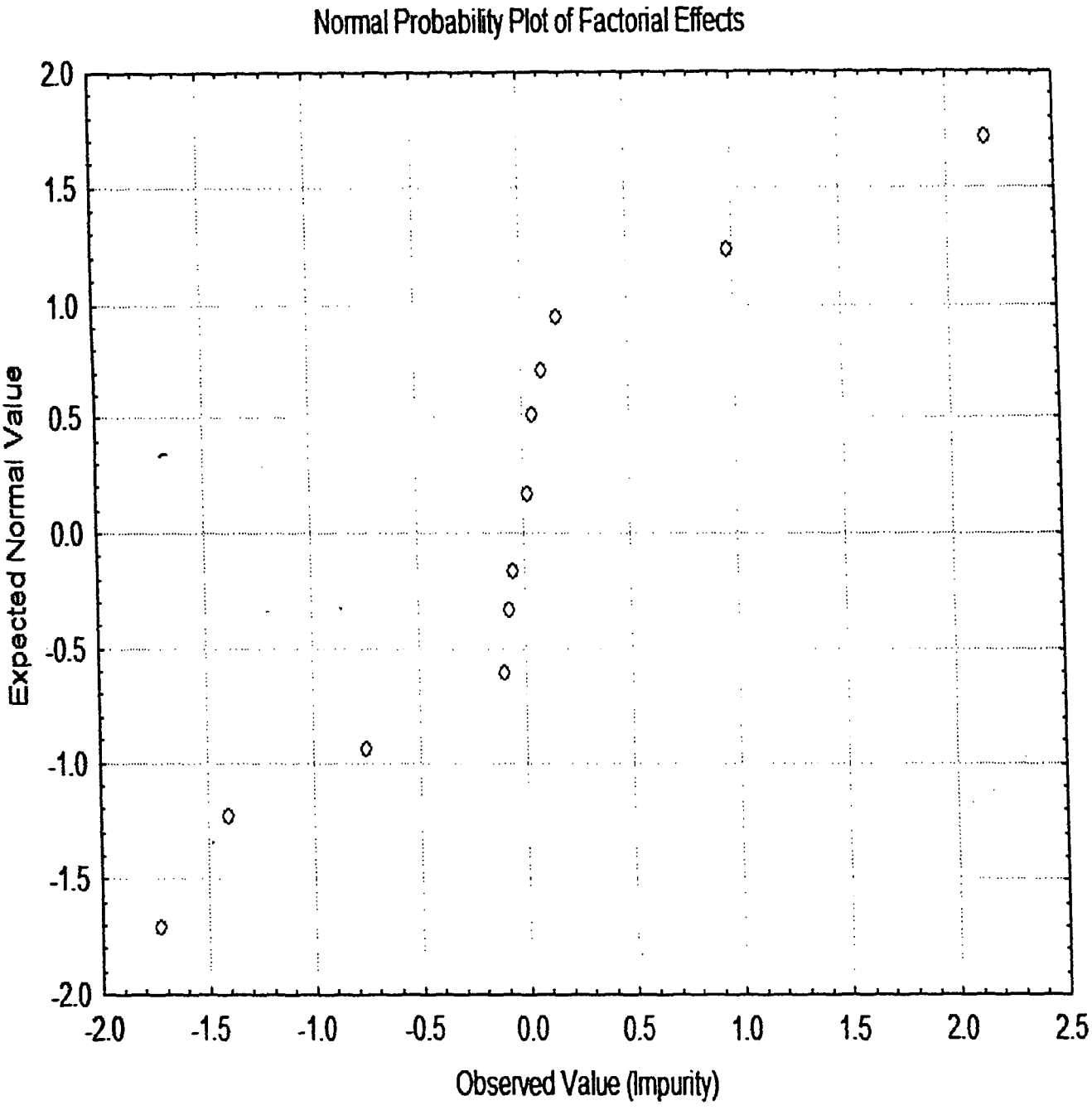
(c) Do the two facilities give different levels of impurities? Justify your answer.

(d) What would be the best possible operating conditions?

(e) Assuming three factor and higher order interactions to be zero, compute an estimate of the error variance. Is there any discrepancy with the earlier result? Comment on your finding.

(f) Is there an indication of hidden replication? Can you reanalyse the data considering the hidden replication, if any ?

[5 +10+2+5+3+3=28]



7) Assignments. [20]

INDIAN STATISTICAL INSTITUTE

Second Semester Examination : 2008–2009

Course Name: B. Stat. III year

Subject Name: Database Management Systems

Date: 18-05-2009 Marks: 100 Time: 3 Hours

The questions are for 110 marks. The maximum marks you can get is 100.
Answer any part of any question. Answer all the parts of a question at the same place.

1. Consider the following relations containing airline flight information:
   Flights(*flno*: integer, *from*: string, *to*: string, *distance*: integer, *departs*: time, *arrives*: time)
   Aircraft(*aid*: integer, *aname*: string, *cruisingrange*: integer)
   Certified(*eid*: integer, *aid*: integer)
   Employees(*eid*: integer, *ename*: string, *salary*: integer)

   Note that the Employees relation describes pilots and other kinds of employees as well; every pilot is certified for some aircraft (otherwise, he or she would not qualify as a pilot), and only pilots are certified to fly. The primary key fields are underlined, and the domain of each field is listed after the field name. Write the following queries in relational algebra, and SQL:

   (a) Find the *eids* of pilots certified for some Boeing aircraft.

   (b) Find the *aids* of all aircraft that can be used on non-stop flights from London to Kolkata.

   (c) Find the names of pilots who can operate planes with a range greater than 3,000 miles but are not certified on any Boeing aircraft.

   (d) Find the *eids* of employees who make the highest salary.

   (e) Find the *eids* of employees who are certified for exactly three aircraft.

   $(5 \times 4 = 20)$

2. Define the terms BCNF, 3NF, and 4NF. Compare BCNF and 3NF. (10+5=15)

3. Consider a relation schema over the attributes $ABCDEFGH$ and the following functional dependencies:
   $ABH \rightarrow C$ $A \rightarrow DE$ $BGH \rightarrow F$
   $F \rightarrow ADH$ $BH \rightarrow DE$
   Find a lossless decomposition into BCNF. Explain the steps. (10)

4. Consider a relation schema over the attributes $ABCDEFGH$ and the following functional dependencies:
   $A \rightarrow E$ $BE \rightarrow D$ $AD \rightarrow BE$ $BDH \rightarrow E$
   $AC \rightarrow E$ $F \rightarrow A$ $E \rightarrow B$ $D \rightarrow H$
   $BG \rightarrow F$ $CD \rightarrow A$
   Find a canonical cover, then decompose into lossless 3NF. (8+7=15)

(P.T.O)

5. Consider a relation schema over the attributes $ABCDE$ and the following multivalued dependencies:
   $A \twoheadrightarrow BC \qquad B \twoheadrightarrow CD \qquad E \twoheadrightarrow AD$
   Find a lossless decomposition into 4NF. Explain the steps. (10)

6. (a) Write down the deletion algorithm for a B-Tree.
   (b) Construct a 2-3 tree with the key values 20 to 29 (total 10 integers, coming in ascending order) and then show how the root can be deleted.

   $(10 + 10 = 20)$

7. (a) Explain the Log-Based recovery scheme with examples.
   (b) Explain the ideas of testing Conflict and View Serializability.

   $(10 + 10 = 20)$

INDIAN STATISTICAL INSTITUTE

Semestral Examination, 2nd Semester, 2008-09

B. Stat III

Statistical Inference II

Date: November 03, 2009 **Maximum Marks: 100** **Duration: 3 and 1/2 hours**

- This question paper carries 125 points. Answer as much as you can. However, the maximum you can score is 100.
- You should provide as much details as possible while answering a question.

---

1. Consider two small positive numbers α and β with $0 < \alpha + \beta < 1$. Consider Wald's SPRT for simple hypotheses with target strength (α, β), in the case of i.i.d. observations. Find approximate expressions for average sample number (ASN) under H_0 and H_1 using Wald's approximation. [14 points]

2. Suppose that $X_1, \ldots, X_n$ is a random sample from an unknown continuous distribution F. Consider the Kolmogorov-Smirnov one sample statistic based on the above data for testing $H_0 : F = F_0$ against $H_1 : F \neq F_0$. Show that the distribution under H_0 does not depend on F_0. [13 points]

3. Let X be a random variable with a continuous distribution function F. Let K_p be the p-th quantile of F.

   (a) Find the expression for $P(X_{(r)} < K_p < X_{(s)})$ in terms of r, s and p using binomial probabilities.

   (b) Describe the test procedure to test

   $$H_0 : K_p = K \text{ versus } H_1 : K_p > K,$$

   with a critical region based on order statistics and approximate level of significance α. [12+12=24 points]

4. X is a random variable with density $f_\theta(x)$. Let $X_1, X_2, \ldots$ be a sequence of sample observations from f_θ.

   (a) What are the sequential test procedures to test hypothesis on θ?

   (b) Discuss the sequential probability ratio test for the simple hypotheses. In which class of procedures is it optimal?

   (c) Let $Z(x) = \ln \dfrac{f_{\theta_0}(x)}{f_{\theta_1}(x)}$, where ln represents the natural logarithm. N is a stopping time in SPRT. Let

   $$S_N = Z(X_1) + Z(X_2) + \ldots + Z(X_N).$$

P.T.O

Given that $E(|Z(X)|) < \infty$, and $E(N) < \infty$, prove that

$$E(S_N) = E(N)E(Z(X)).$$

[8+8+14 = 30 points]

5. Let X_1, X_2, X_3 be i.i.d. F and Y_1, Y_2, Y_3 be i.i.d. G. Derive the sampling distribution of the number of runs R when $F = G$. [8 points]

6. Let X be continuous and $X_1, X_2, \ldots, X_n$ are i.i.d. observations on X. Let $X_{(1)}$ and $X_{(n)}$ be the minimum and the maximum of the X_is.

   For $n = 5$, compute

   $$P\{P[X_{(1)} < X < X_{(n)}] \geq 0.8\}.$$

   [12 points]

7. Consider Aesop's classic experiment where one tortoise raced against one hare, and the tortoise eventually won the race. Suppose that the variable of interest is the amount of time taken to complete the race. Suppose that Aesop is dissatisfied with his experiment as the sample size is too small, and conducts a second race with $m = 4$ tortoises and $n = 5$ hares. The order in which they finished the second race is:

   T T T H H H T H H.

   So the first three positions were claimed by tortoises, the next three by hares, and so on. Thus if X represents the time taken by a tortoise to complete the race, and Y represents the time taken by a hare to complete the race, the above represents a permutation of 4 observations from the X population and 5 observations from the Y population.

   Based on the above data, perform a Wilcoxon Rank Sum test at an approximate level of $\alpha = 0.05$ to determine whether the populations are identical against the alternative that the X population is stochastically larger. [12 points]

8. The following data are lifetimes (in hours) of batteries for two different brands.

   Brand A: 45, 55, 30, 40.

   Brand B: 45, 55, 60, 40.

   Perform an appropriate test of hypothesis to determine whether the brands have the same median life. Use an approximate level of $\alpha = 0.05$. [12 points]