Indian Statistical Institute

Mid-Semester Examination First Semester (2006-2007) 3

B.Stat. Third Year

Statistical Inference I

Maximum Marks: 100 Date: 4.9.06 Duration :- 3 hours

Note : Special credit will be given to answers which are complete, precise and to the point. You may answer any part of any question; maximum you can score is 100.

1. (a) Define an MVUE and a zero function. [1+1=2]

(b) State and prove a necessary and sufficient condition for an estimator to be an MVUE in terms of zero functions. [5+5=10]

(c) Let T_1, T_2 be two MVUEs such that the product T_1T_2 is square-integrable. Show that T_1T_2 is an MVUE if either T_2 is bounded or every zero function is square-integrable. Give an example where every zero function is square-integrable. [4+4+2=10]

2. (a) Define completeness and bounded completeness. [1+1=2]

(b) Show that a boundedly complete sufficient statistic is minimal sufficient. [8]

(c) State and prove Basu's theorem. [1+5=6]

(d) Let $X_1, \ldots, X_n$ be i.i.d., each following $U(0,1)$ distribution. Show that

$$E(\frac{\bar{X}}{X_{(n)}}) = \frac{n+1}{2n},$$

where $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ and $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$. [8]

3. (a) Show that a minimal sufficient statistic need not be boundedly complete. [10]

(b) Let $X_1, \ldots, X_n$ be i.i.d., each following $N(\theta, \sigma^2)$, where θ is unknown and σ^2 is known. Find the MVUE of $P_\theta(X_1 \leq c)$ where c is a given real number. [10]

4. (a) Let $X_1, \ldots, X_n$ be i.i.d., each following $N(\mu, \sigma^2)$ where μ and σ^2 are unknown. Find explicitly the MLE of $E_{\mu,\sigma^2}(\Phi(a+bX_1))$ where Φ is the distribution function of $N(0,1)$ and a and b are known constants. [8]

(b) Let $X_1, \ldots, X_n$ be as in part (a), where μ is unknown and σ^2 is known. Find the MVUE of $\exp(t\mu)$ where $t \neq 0$ is a given number. Compare the variance of the MVUE with the Cramer-Rao lower bound. Show that the ratio tends to 1 as $n \to \infty$. [10]

P. T. O

(c) Write a critical essay on the method of maximum likelihood. [10]

(d) Give an example of a family of pdfs $\{f_\theta\}$, where the parameter space is a non-empty open interval, such that there is an unbiased but inconsistent estimator. [7]

5. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d., each following the Bivariate Normal distribution with zero means, unit variances and unknown covariance θ, $-1 < \theta < 1$. Show that $\sum_{i=1}^{n} X_i^2$ as well as $\sum_{i=1}^{n} Y_i^2$ are ancillary but $(\sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} Y_i^2)$ is not. [1+9=10]

INDIAN STATISTICAL INSTITUTE

Mid-semester Examination, B.Stat.-III : (2006 - 2007)

ELECTIVE GEOLOGY

Date : 6. 09. 2006. Maximum Marks :40 Duration : 2 hours

Answer all the following questions

1. A. What is a mineral?
   B. Distinguish between rock and mineral.
   C. Name two rock-forming minerals and the mineral groups in which they belong.
   D. Distinguish between plutonic and volcanic igneous rocks.
   E. How are minerals classified on the basis of their optical properties?

2+2+2+2+2

2. Select the correct word or expression from the three alternatives

A. The streak of a mineral- may vary from it's colour/ always varies from its colour/ never varies from its colour.
B. Quartz/ calcite/ muscovite/ has three sets of cleavages.
C. The majority of the rock -forming minerals are- borates/ molybdates/ silicates.
D. Mica is a phyllosilicate/inosilicate/sorosilicate.
E. Double refraction takes place- in ordinary glass/ opaque minerals/ transparent minerals.
F. In isometric system of crystals the unit lengths of- all the crystallographic axes are equal/ all the crystallographic axes are not equal/ two of the crystallographic axes are equal.
G. Radiometric dating indicates- relative ages of rocks/ absolute ages of rocks/ a range of age of rocks.
H. If an igneous intrusion is seen to have vertically cut through a package of horizontal sedimentary rocks from bottom to top, then it is- younger than the sedimentary rocks/ older the sedimentary rocks/ younger than only the bottom part of the sedimentary rocks.
I. Fossils are generally best preserved in- igneous rocks/ metamorphic rock/ sedimentary rock.
J. According to the Mohs' scale of hardness, Apatite is the- softest mineral/ hardest mineral/ moderately hard mineral.

10 x 1= 10

2. Name and illustrate different types of unconformities. State the "Law of Superposition".

Or

Illustrate polymorphism, solid solution and exsolution with examples. 4+2 = 6

4. What are the implications of "Red shift" and "Microwave background radiation" ? 6
5. Why do we see only one face of the moon from any location on the earth? 4
6. What are the principal physical constituents of the universe? 2
7. Draw perspective diagrams of cube, prism and pyramid. 5

Or

Illustrate the rock cycle with a suitable diagram. 3

Indian Statistical Institute

Mid-Semester Examination: (2006-2007)
Name of the Course: Introduction to Sociology
B.Stat.- III year
Date: 6.9.06 Maximum Marks: 50 Duration: 1½ hour

Group –A

The figures in the margin indicate full marks

Q. 1. Relate the Sociological thinkers with the concept introduced by them and write in your answer script: 1x5 = 5

Thinker	Concept
1) Emile Durkheim	Sympathetic Introspection
2) Auguste Comte	Organic Solidarity
3) Max Weber	Social Darwinism
4) Karl Marx	Social Statics
5) Emile Durkheim	Social Facts

Q. 2. Write the correct answer in your answer script: 1x5 = 5

(a) The 'Positive era' (according to Auguste Comte) will be dominated by 'Military' / 'Scientists & Industrialists'.
(b) Anomic suicide (according to Emile Durkheim) is caused by social depression / over attachment to societal norms.
(c) Max Weber's conception of social change is related to 'Cultural interpretation'/ 'Economic interpretation'.
(d) Feudal society is principally based on agriculture/ industry.
(e) The concept of 'class for itself' (according to Karl Marx) means a class devoid of class conscious / with class consciousness.

3) Write short notes (any two): 5x2 =10

(a). Can the 'Industrial Revolution' be related with the rise of the discipline 'Sociology'?
(b) How did Auguste Comte Describe the 'law of three stages' for the evolution of society?
(c) How did Emile Durkheim describe the reason of 'egoistic suicide'?

P.T.O

Group –B

Q.1. Choose the correct answers: 1x10 = 10

(A) The "structure" of an Institution is described as:
- (i) That of an association
- (ii) That which consists of personal equipment, organization and ritual
- (iii) That of a big group

(B) ..... is called the process of two individuals affecting each Other's behaviour?
- (i) Social stratification
- (ii) Social interaction
- (iii) Socialization

(C) Social class is the factor that characterises the :
- (i) Stability
- (ii) Heredity
- (iii) Vertical mobility

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

(E) Which is the modern trend in the development activity of the village ?
- (i) The villages should become more bigger in size
- (ii) The villages should have all the amenities of the towns
- (iii) The villages should be wiped out and towns set up

(F) What is the basis for the demarcation of the family into matriarchal and patriarchal ?
- (i) Organization
- (ii) Authority
- (iii) Residence

(G) Which is the noblest group among Indian Muslim?
- (i) Pathan
- (ii) Sheikh
- (iii) Moghal

(H) The concept of "Sanscritization and Westernization" is attributed to:
- (i) Ramkrishna Mukherjee
- (ii) Andre Beteille
- (iii) M.N. Srinivas

(I) 'Census sampling' means:
- (i) Covering 5% of the universe
- (ii) Covering 20% of the universe
- (iii) Taking all the respondents available

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

Q.2 . Answer (in short) any five questions: 2x5 =10
- (i) Define Sociology.
- (ii) What does class mean?
- (iii) Define status and role.
- (iv) What is meant by values?
- (v) Are civilization and culture synonymous?
- (vi) What is meant by Integrated Rural Development?
- (vii) Who has introduced the concept of 'dominant caste'?

Q.3. Answer briefly any two questions: 5x2 =10
- (i) Analyse the relation between caste and class in modern Indian society.
- (ii) Discuss Andre Beteille's contribution to Indian Sociology with special reference to caste and politics in Tamilnadu.
- (iii) What is panchayat ? What is its impact on village society?

INDIAN STATISTICAL INSTITUTE

Mid-semester Examination (2006-2007)
Course: B. Stat. (III)
Subject: Introduction to Anthropology and Human Genetics

Date:..6..9..06...Maximum Marks:...60.....Duration:...2hr 15min (135min)...

(Desk calculator and Statistical Tables allowed.)

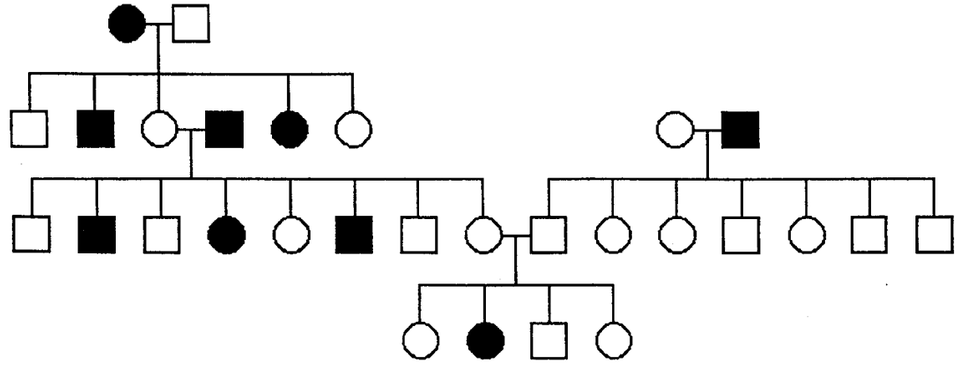
1. Answer any three:

   (i) What does the subject Anthropology deals with? How is it different from other fields of study which deals with Humans? What are the different branches of study in the subject Anthropology?
   (ii) Write a note on Hardy-Weinberg equilibrium and its significance.
   (iii) Write a note on some distinguishing features of Humans
   (iv) Define Adaptation.

   3 X 6 (18)

2. Which best describes the genetics of the afflicting allele in the following pedigree? Give reasons.

   (i) autosomal dominant
   (ii) autosomal recessive
   (iii) X-linked dominant
   (iv) X-linked recessive
   (v) Y-linked dominant
   (vi) Y-linked recessive



   3 + 4 (7)

P.T.O

3. Two populations A and B were tested for MN blood group and the following distribution was observed:

	MN blood group phenotypes		
Population	MM	MN	NN
A	298	489	213
B	348	389	263

Do the two populations differ significantly for this trait? Calculate the allele frequency for M and N alleles for the both populations. Are the two populations in Hardy-Weinberg equilibrium?

3 + 4 + 5 (12)

4. A sample of 1000 individuals in a population was tested for the trait PTC tasting ability. A total of 360 individuals were found to be non taster (tt). Assuming the population to be in Hardy-Weinberg equilibrium, calculate the following:

(i) The frequency of the "tt" genotype.
(ii) The frequency of the "t" allele.
(iii) The frequency of the "T" allele.
(iv) The frequencies of the genotypes "TT" and "Tt."
(v) The frequencies of the two possible phenotypes: Tasters and non tasters.

1 X 5 (5)

5. A certain population is at Hardy-Weinberg equilibrium. In this population, there is an autosomal recessive disease, and the frequency of those affected by the disease is exactly equal to the frequency of heterozygous carriers for that disease. What is q, the frequency of the disease-causing allele, in this population?

(6)

6. If the frequency of the "green" form of red-green color blindness (due to an X-linked locus) is 5 percent among males, what fraction of females would be affected? What fraction of females would be heterozygous? Why the frequency of affected females is much less than frequency of affected males?

3 X 2 (6)

7. Define Race. What are the criteria used for classifying human populations? Why anthropologists are no longer interested in racial classification

(6)

INDIAN STATISTICAL INSTITUTE
Mid – Semester Examination : 2006 –07
B. Stat. (Hons.) III Year
Sample Surveys

Date : 8. 09.2006 **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. A simple random sample of size $n = n_1 + n_2$ with mean $\bar{y}$ is drawn from a finite population , and a simple random sub-sample of size n_1 is drawn from it with mean $\bar{y}_1$. Show that

(a) $\mathrm{Var}(\bar{y}_1 - \bar{y}_2) = S^2\left[\left(\frac{1}{n_1}\right) + \left(\frac{1}{n_2}\right)\right]$, where $\bar{y}_2$ is the mean of the remaining n_2 units in the sample ,

(b) $\mathrm{Var}\,(\bar{y}_1 - \bar{y}) = S^2\left[\left(\frac{1}{n_1}\right) - \left(\frac{1}{n}\right)\right]$,

(c) $\mathrm{Cov}\,(\bar{y}, \bar{y}_1 - \bar{y}) = 0$.

Repeated sampling implies repetition of the drawing of both the sample and the sub-sample .

(9 +8+8) = [25]

2. In a sample of 50 households drawn with SRSWOR from a village consisting of 250 households , only 8 households were found to possess a bicycle . These had 3,5,3,4,7,4,4 and 5 members respectively . Estimate unbiasedly the total number of households in the village possessing a bicycle as well as the total number of persons in such households . Also estimate the RSE's of the estimates by using the unbiased estimates of their variances .

(4+6+6+9)=[25]

3. A sampler has two strata with relative sizes W_1 and W_2. He believes that S_1 , S_2 can be taken as equal but thinks that c_2 may be between $2c_1$ and $4c_1$. He would prefer to use proportional allocation but does not wish to incur a substantial increase in variance compared with optimum allocation . For a given cost $C = c_1 n_1 + c_2 n_2$, ignoring the fpc, show that

$$\frac{V_{prop}(\bar{y}_{st})}{V_{opt}(\bar{y}_{st})} = \frac{W_1 c_1 + W_2 c_2}{\left(W_1\sqrt{c_1} + W_2\sqrt{c_2}\right)^2}$$

If $W_1 = W_2$, compute the relative increases in variance from using proportional allocation when $\frac{c_2}{c_1} = 2,4$.

P.T.O.

(20 + 5) =[25]

4. In a directory of 13 houses on a street the persons are listed as follows : M = male adult , F = female adult , m = male child , f = female child .

Household												
1	2	3	4	5	6	7	8	9	10	11	12	13
M	M	M	M	M	M	M	M	M	M	M	M	M
F	F	F	F	F	F	F	F	F	F	F	F	F
f	f	m		m	f	f	m	m	m	f	f	
m	m	f		m	m	f	f		f	m		
f	f			f		m						

Compare the variances given by a systematic sample of one in five persons and a 20% simple random sample for estimating (a) the proportion of males , (b) the proportion of children , (c) the proportion of persons living in professional households (households 1,2,3,12 and 13 are described as professional) . For the systematic sample , number down each column , then go to the top of the next column .

(8 + 8 + 9) = [25]

5. A survey is to be conducted for estimating the total number of literates 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 literates

Community	Total number of persons	Percentage of literates
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 200 persons to the strata in an optimum manner for estimating the overall proportion of literates in the town .
(b) Estimate the efficiency of stratification as compared to unstratified sampling .

(10 + 15) = [25]

6. (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-Semestral Examination

First Semester 2006–2007

B. Stat (Third year)

Differential Equation

Date: 12 September, 2006

Maximum Marks: 40

Duration: 2 hours 30 minutes

Answer all questions.

(1) Let L be a linear differential operator of order n with constant coefficients and let p be the charachteristic polynomial of L. Suppose p has n distinct roots namely $r_1, r_2, \ldots r_n$. Show that for any continuous function b on $\mathbb{R}$, a solution ϕ of the equation $Ly = b$ can be expressed as follows:

$$\phi = \sum_{k=1}^{n} \frac{1}{p'(r_k)} \phi_k$$

where ϕ_k satisfies the equation $(D - r_k)y = b$, $k = 1, 2, \ldots, n$, where $D \equiv \frac{d}{dx}$. 12

(2) Consider the differential equation $y'' + P(x)y' + Q(x)y = 0$, where P and Q are continuous functions on the interval (a, b). Prove the following:

(a) If $\phi_1(x)$ and $\phi_2(x)$ are two solutions of the above equation and have a common zero in (a, b) then one of them is a constant multiple of the other.

(b) If ϕ is a solution of the same equation and $\phi'(x_0) = 0$ for some $x_0 \in (a, b)$, then ϕ is a constant function. 8

(3) Consider the linear differential equation of the following form:

$$x^n y^{(n)} + a_1 x^{n-1} y^{(n-1)} + \cdots + a_{n-1} x y' + a_n y = 0,$$

where a_i's are constants.

Show that the substitution $z = \log x, x > 0$, transforms the equation into a linear differential equation with constant coefficients. Hence show that the indicial polynomial of the given equation is the same as the characteristic polynomial of the transformed equation. 12

(4) Let n be a non-negative integer. Show that the equation

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

has a polynomial solution P_n of degree n.

Prove that $\int_0^\infty x e^{-x} P_n(x) P_m(x) dx = 0$ whenever $m \neq n$. 12

Mid-term Examination (2006-07)

B.Stat. (Hons.) III year

Linear Statistical Models

Maximum time: $2\frac{1}{2}$ hours September 15, 2006 Maximum marks: 60

This test is closed book. The total number of marks of all the questions is 68, and the maximum you can score is 60.]

1. Consider a linear model for independent, normal distributed observations with equal variance σ^2 and mean as under:

$$\begin{aligned} E(y_1) &= \beta_1 + \beta_2 + 2\beta_3, \\ E(y_2) &= \beta_1 + \beta_3, \\ E(y_3) &= \beta_2 + \beta_3, \\ E(y_4) &= -\beta_1 - \beta_2 - 2\beta_3, \\ E(y_5) &= -\beta_1 - \beta_3, \\ E(y_6) &= -\beta_2 - \beta_3. \end{aligned}$$

The parameters β_1, β_2, β_3 and σ^2 are unknown.

(a) Identify a non-estimable function of β_1, β_2 and β_3.

(b) Identify two estimable functions of β_1, β_2 and β_3 which are not multiples of one another.

(c) Identify four linear zero functions (in terms of $y_1, \ldots, y_6$) which are uncorrelated with one another.

(d) Identify the BLUEs (in terms of $y_1, \ldots, y_6$) of the two parametric functions mentioned in part (b), with justification.

(e) Is there any linear zero function which is uncorrelated with the four functions mentioned in part (c)? Explain.

(f) Identify the usual unbiased estimator of σ^2 (in terms of $y_1, \ldots, y_6$), with justification.

(g) Give Bonferroni (simultaneous) confidence intervals for the two estimable parametric functions mentioned in part (b), in terms of $y_1, \ldots, y_6$.

(h) Indicate how the linear hypothesis $\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \mathbf{0}$ can be decomposed into a completely testable hypothesis and a completely non-testable hypothesis.

(i) Give a test statistic for the completely testable hypothesis of part (h) and its null distribution.

(j) Give a tolerance interval with confidence coefficient 0.95, covering 90% of all future observations having mean $\beta_1 + \beta_3$. [1+2+4+5+1+2+2+3+2+3=25]

2. You have samples (x_i, y_i), $i = 1, \ldots, n$ from a bivariate normal distribution with unknown parameters 0, 0, σ_x^2, σ_y^2 and ρ. Express the 'regression of y on x' as a function of these parameters. Derive a 'confidence band' that would contain the entire regression line with probability $1 - \alpha$. How does the shape of this confidence band change with α?

[2+8+5=15]

P.T.O

3. Consider the quadratic regression model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon, \quad E(\epsilon) = 0, \quad Var(\epsilon) = \sigma^2.$$

(a) If $\beta_2 > 0$, determine the value of x which will minimize the expected response.

(b) If you have n uncorrelated observations from the model, how would you test the null hypothesis $\beta = 0$ against the alternative hypothesis $\beta > 0$, so that the probability of type I error is .05?

(c) Assuming that the null hypothesis of part (b) is rejected at the level .05, derive a confidence interval for the optimal value of x obtained in part (a), so that the coverage probability is 0.9. [2+2+6=10]

4. In a linear model $(\boldsymbol{y}, \boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ where $\boldsymbol{X}$ has full column rank, you have to test the hypothesis $\boldsymbol{\beta} = \boldsymbol{Z\theta}$ for a specified matrix $\boldsymbol{Z}$ (having full column rank which is smaller than the rank of $\boldsymbol{X}$) and unspecified vector $\boldsymbol{\theta}$.

(a) Show that the hypothesis can be formulated as the usual linear hypothesis $\boldsymbol{A\beta} = \boldsymbol{\xi}$ for specified matrix $\boldsymbol{A}$ and specified vector $\boldsymbol{\xi}$.

(b) Indicate how you would test this hypothesis. [5+5=10]

5. Suppose that $\boldsymbol{X} = (\boldsymbol{x}_{(1)} : \cdots : \boldsymbol{x}_{(k)})$, $\boldsymbol{D}$ is a diagonal matrix with $\|\boldsymbol{x}_{(j)}\|$ in the jth diagonal position and $\boldsymbol{X}_s = \boldsymbol{XD}^{-1}$, that is, the columns of $\boldsymbol{X}_s$ have unit norm and are proportional to the columns of $\boldsymbol{X}$. Let $\boldsymbol{X}$ have full column rank.

(a) Show that VIF_j is the jth diagonal element of the matrix $(\boldsymbol{X}_s'\boldsymbol{X}_s)^{-1}$.

(b) Show that $VIF_j \geq 1$ for $j = 1, \ldots, k$.

(c) If $\lambda_1 \geq \cdots \geq \lambda_k$ are the eigenvalues of $\boldsymbol{X}_s'\boldsymbol{X}_s$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ are the corresponding eigenvectors, show that

$$VIF_j = \sum_{i=1}^{k} \frac{v_{ij}^2}{\lambda_i}, \quad j = 1, \ldots, k,$$

v_{ij} being the ith element of $\boldsymbol{v}_j$, $i, j = 1, \ldots, k$. [3+3+2=8]

INDIAN STATISTICAL INSTITUTE

Semestral Examination

First Semester 2006–2007

B. Stat (Third year)

Differential Equations

Date: 23 November, 2006

Maximum Marks: 60

Duration: 2 hours 30 minutes

Answer all questions.

(1) Consider the equation

$$y'' + e^x y = 0.$$

(i) Reduce the equation to a system of first order equations and hence prove that the equation has a solution $\phi(x)$ which is defined for all $x \in \mathbb{R}$.

(ii) Prove that every interval of length π in $(0, \infty)$ contains at least one zero of $\phi(x)$ and every interval of length π in $(-\infty, 0)$ contains at most one zero of $\phi(x)$.

State clearly any result that you use to prove the above statements. 8+8+4

(2) (a) Let $M_n(\mathbb{R})$ denote the set of all $n \times n$ matrices with real entries and let $A \in M_n(\mathbb{R})$. Consider the curve $\gamma : \mathbb{R} \longrightarrow M_n(\mathbb{R})$ defined by $\gamma(t) = e^{tA}$. Prove that $\frac{d\gamma}{dt} = A\gamma(t)$. Justify each step in the proof.

(b) Suppose A has distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ with eigenvectors $v_1, v_2, \ldots, v_n$ respectively. Prove that the vector valued functions $e^{\lambda_1 t}v_1, e^{\lambda_2 t}v_2, \ldots, e^{\lambda_n t}v_n$ form a basis for the solution space of the differential equation $X'(t) = AX(t)$, where $X = (x_1, x_2, \ldots, x_n)^T$.

(c) Let A be a 2×2 matrix over the reals. Let A have two distinct eigenvalues λ_1, λ_2 such that $\lambda_1 < \lambda_2 < 0$. Suppose $\gamma(t) = (x_1(t), x_2(t))$ is a solution of the differential equation $X'(t) = AX(t)$, where $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. Prove the following statements:

(i) The point $\gamma(t)$ approaches $(0, 0)$ as $t \longrightarrow \infty$.

(ii) $\lim_{t \longrightarrow \infty} \frac{x_2(t)}{x_1(t)}$ exists (the limit may be finite or infinite).

5+8+8

P.T.O

INDIAN STATISTICAL INSTITUTE

First Semestral Examination (2006-07)

B.Stat. (Hons.) III year

Linear Statistical Models

Maximum time: 3 hours November 28, 2006 Maximum marks: 100

[*This test is closed book. The total number of marks of all the questions is 110, and the maximum you can score is 100.*]

1. Show that a linear parametric function in a linear regression model is estimable if and only if it is identifiable. If a linear parametric function does not have a linear unbiased estimator, show that it does not have any non-linear unbiased estimator. [5+5=10]

2. Consider the following linear model with six observations $y_1, \ldots, y_6$,

$$\begin{aligned} y_1 &= \beta_1 - \beta_2 + \epsilon_1, \\ y_2 &= \beta_1 - \beta_2 + \epsilon_2, \\ y_3 &= \beta_2 - \beta_3 + \epsilon_3, \\ y_4 &= \beta_2 - \beta_3 + \epsilon_4, \\ y_5 &= \beta_3 - \beta_1 + \epsilon_5, \\ y_6 &= \beta_3 - \beta_1 + \epsilon_6, \end{aligned}$$

where β_1, β_2 and β_3 are unknown parameters and $\epsilon_1, \ldots, \epsilon_6$ are uncorrelated errors with mean 0 and variance σ^2. Suppose that you have already computed the following quantities: best linear unbiased estimators (BLUE) $\widehat{\beta}_1 - \widehat{\beta}_2$, $\widehat{\beta}_2 - \widehat{\beta}_3$ and $\widehat{\beta}_3 - \widehat{\beta}_1$ corresponding to the parameters $E(y_1)$, $E(y_3)$ and $E(y_5)$, respectively, the variances of these estimators (each as a multiple of σ^2), and the usual unbiased estimator of σ^2. Subsequently, a fresh observation y_7 with mean $\beta_1 - 2\beta_2 + \beta_3$ becomes available. This observation is uncorrelated with the preceding six observations, and also has variance σ^2.

   (a) Show that $z = y_7 - \widehat{\beta}_1 + 2\widehat{\beta}_2 - \widehat{\beta}_3$ is a linear zero function (LZF) in the augmented linear model with seven observations.

   (b) Show that the above LZF is uncorrelated with all LZFs of the original linear model with six observations.

   (c) Using the above facts, derive an expression for the BLUE of $\beta_1 - \beta_2$ in the augmented model, explicitly in terms of $\widehat{\beta}_1 - \widehat{\beta}_2$ and z.

   (d) Derive an expression for the variance of the estimator of part (c), explicitly in terms of σ^2 and the variance of $\widehat{\beta}_1 - \widehat{\beta}_2$.

   (e) Derive an expression for the usual unbiased estimator of σ^2 in the augmented model in terms of the corresponding estimator in the original model and z.

   (f) Express in terms of the observations the usual unbiased estimator of σ^2 under the original model.

   (g) Express in terms of the observations the BLUE of $\beta_1 - \beta_2$ under the original model.

   [1+2+4+4+4+2+3=20]

P. T. O

(3) Consider the following system of differential equations:

$$\frac{dx}{dt} = 2xy \qquad \frac{dy}{dt} = y^2 - x^2.$$

Solve the above system and draw a rough sketch of the phase curves. 8

(4) Consider the cylinder C defined as follows:

$$C = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 = 1\}.$$

Let P and Q be two fixed points on the cylinder. For any smooth curve $\gamma : [0, 1] \longrightarrow C$ satisfying $\gamma(0) = P$ and $\gamma(1) = Q$ define the length functional by

$$L(\gamma) = \int_0^1 \left\| \frac{d\gamma}{dt} \right\| dt,$$

where $\|(x, y, z)\| = \sqrt{x^2 + y^2 + z^2}$ for $(x, y, z) \in \mathbb{R}^3$.

Find the extremal(s) of L and describe them on C. 15

3. Consider the model $(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ where $\boldsymbol{\beta} = (\beta_0 : \beta_1 : \beta_2 : \beta_3)'$, and

$$\boldsymbol{X}' = \begin{pmatrix} \boldsymbol{x}_0' \\ \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \boldsymbol{x}_3' \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \end{pmatrix}.$$

The objective of this study is to determine whether the subset model consisting only of $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ will be more suitable than the full model for the purpose of estimating certain LPFs.

(a) Will the 'subset estimator' of $\beta_3 - \beta_0 - \beta_1 - \beta_2$ have smaller MSE than the BLUE from the full model, if the true parameter values are such that $\beta_2 = \beta_3$?

(b) Will the 'subset estimator' of $\beta_1 - \beta_0 - \beta_2 - \beta_3$ have smaller MSE than the BLUE from the full model, if the true parameter values are such that $\beta_2 = \beta_3$?

(c) Can you give an intuitive explanation of the discrepancy in the answers to parts (a) and (b)? [4+3+3=10]

4. Let there be n_i observations of the response (arranged as the $n_i \times 1$ vector $\boldsymbol{y}_i$) for a given combination of the explanatory variables $(\boldsymbol{x}_i)$, $i = 1, \ldots, m$, $n_1 + \cdots + n_m = n$. The plan is to check the adequacy of the model $(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ through a formal test of lack-of-fit, assuming normal errors. Here, $\boldsymbol{y} = (\boldsymbol{y}_1' : \cdots : \boldsymbol{y}_m')'$ and

$$\boldsymbol{X} = (\boldsymbol{x}_1 \otimes \boldsymbol{1}'_{n_1 \times 1} : \cdots : \boldsymbol{x}_m \otimes \boldsymbol{1}'_{n_m \times 1})'.$$

Assume that $m > r = \rho(\boldsymbol{X})$.

(a) Show that the model $(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ is a restricted version of another model, where the response for every given $\boldsymbol{x}_i$ is allowed to have an arbitrary mean.

(b) Obtain the error sum of squares under the unrestricted model (*pure error sum of squares*).

(c) Identify the restriction of part (a) as the hypothesis of adequate fit, and obtain an expression for the sum of squares for deviation from the hypothesis (*lack of fit sum of squares*).

(d) Construct the ANOVA table.

(e) Describe the likelihood ratio test for lack of fit. [4+3+3+3+2=15]

5. Consider the model

$$\begin{aligned} y_{ij} &= \mu + \tau_i + \beta_j + \epsilon_{ij}, \quad i = 1, 2, \ j = 1, 2, 3, \\ \epsilon_{ij} &\overset{i.i.d.}{\sim} N(0, \sigma^2). \end{aligned}$$

where μ, τ_1, τ_2 and $\beta_1, \beta_2, \beta_3$ are unspecified parameters. The following summary statistics are available:

$$\begin{aligned} (\bar{y}_{\cdot 1} - \bar{y}_{\cdot\cdot})^2 &= 1, & (\bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot})^2 &= 4, \\ (\bar{y}_{\cdot 2} - \bar{y}_{\cdot\cdot})^2 &= 2, & (\bar{y}_{2\cdot} - \bar{y}_{\cdot\cdot})^2 &= 5, \\ (\bar{y}_{\cdot 3} - \bar{y}_{\cdot\cdot})^2 &= 3, & \textstyle\sum_{i=1}^{2}\sum_{j=1}^{3}(\bar{y}_{ij} - \bar{y}_{\cdot\cdot})^2 &= 50. \end{aligned}$$

Do the summary statistics support the hypothesis $\tau_1 = \tau_2$? Give a clear answer with justification. [10]

6. Consider observations from t groups following the model

$$\begin{aligned} y_{ij} &= \mu + \tau_i + \epsilon_{ij}, \quad j = 1, \ldots, n_i, \ i = 1, \ldots, t, \\ E(\epsilon_{ij}) &= 0, \qquad\qquad j = 1, \ldots, n_i, \ i = 1, \ldots, t, \\ Cov(\epsilon_{ij}, \epsilon_{i'j'}) &= \begin{cases} \sigma^2 & \text{if } i = i' \text{ and } j = j', \\ \rho\sigma^2 & \text{if } i = i' \text{ and } j \neq j', \\ 0 & \text{if } i \neq i', \end{cases} \end{aligned}$$

where ρ is a known and positive fraction. Derive a test for equality of all group means. Comment on the effect of the value of ρ on the test. [11+4=15]

7. Consider the balanced two-way classification model with three treatments, four parameters and five observations per cell. The last observation for the first treatment-block combination is missing. Using the analysis of covariance formulation for missing observations, describe precisely the likelihood ratio test for 'no interaction'. Clearly identify the correction term(s) for the missing observation, simplifying the expression(s) as much as possible. [15]

8. You have three-way classified count data $\{y_{ijk}\}$ having the Poisson distribution, with

$$\log(E(y_{ijk})) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}, \quad i = 1, \ldots, I, \ j = 1, \ldots, J, \ k = 1, \ldots, K,$$

where the summands on the right hand side are unspecified parameters.

(a) Show that the link function used here is the canonical link function in a generalized linear model.

(b) Interpret the above model and suggest suitable side-conditions for identifiability of the parameters.

(c) Derive maximum likelihood estimators of $E(y_{ijk})$.

(d) Suggest a suitable test for significance of the interaction terms. [2+3+5+5=15]

INDIAN STATISTICAL INSTITUTE

First Semestral Examination, B.Stat.-III : (2006 - 2007)

ELECTIVE GEOLOGY

Date : 1. 12. 2006. Maximum Marks : 60, Duration : 2 and 1/2 hours

Answer all the following questions

1. How can you distinguish oceanic crust from continental crust. (2)

2. Explain the following features of the earth using the principle of isostasy:
   a) the continental crust is thicker than the oceanic crust
   b) the mountains have deep roots
   (4)

3. Derive the radioactive decay equation. (3)

4. Samples of two different rock bodies A and B have yielded the following isotopic compositions

ROCK A	^{87}Rb	^{87}Sr
Sample 1	2	7
Sample 2	4	11

ROCK B	^{87}Rb	^{87}Sr
Sample 1	2	12
Sample 2	3	17

Which of the rocks between A and B is older? (2)

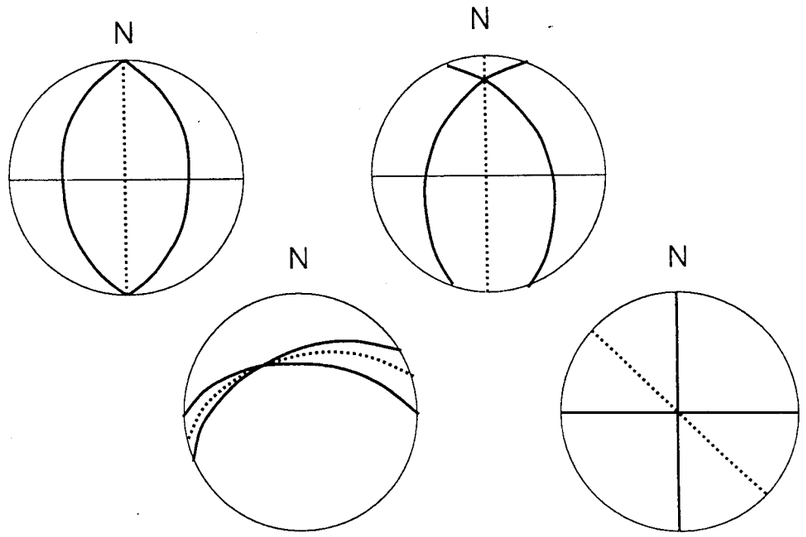
5. Distinguish between P and S seismic waves. (2)

6. Classify folds based on the orientation of fold axis and axial plane. (3)

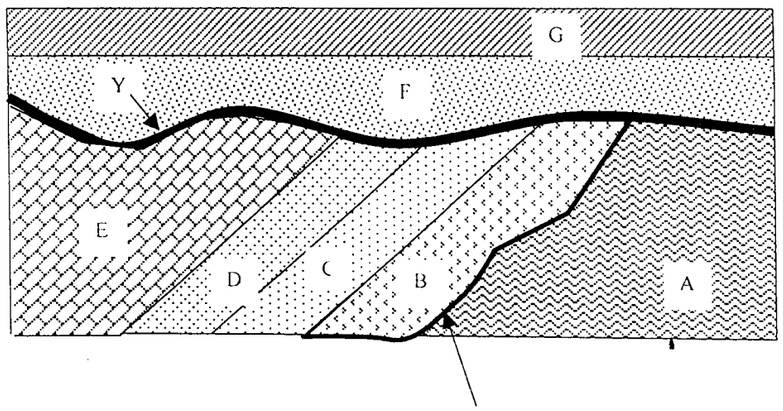
7. Illustrate normal, reverse and strike-slip faults with the help of sketches. (3)

P.T.O

8. Name the folds represented in following stereograms. Bold and dotted lines are limbs and axial planes respectively. (4)



9.



The above figure depicts the organization of strata observed in a vertical sectio in the field. Write down the sequence of geological events and comment on the nature of the surfaces X and Y. A: Granite, B-G: Sedimentary rocks. (4

10. A uniformly dipping coal seam is exposed on the surface at locations A and B. Location B occurs at 200 meter north-east of A. In a vertical bore hole at C (located 300 meter south-east of B) the same coal seam is struck at a depth of 200 meter. At which depth the coal seam will be struck in a vertical bore hole at a location D occurring 200 meter east of A? Assume: (a) the top direction of your answer script as north and (b) 100 meter = 2 centimeter. (5)

11. Define "Monophyletic", "Polyphyletic" and "Paraphyletic" group with simple, schematic, cladograms. (6)

12. Why is Darwin's theory of organic evolution considered as a 'gradualistic theory' ? (4)

13. Why mirror image striping of sea floor magnetism are noted on the opposite sides of the Mid Atlantic Rift? (4)

y do the major volcanoes of the world occur as linear belts? (2)

tify the occurrence of marine fossils within the rocks of the Himalayan ıntain. (2)

ed on the fieldwork in the Asansol area answer the following questions:
a) What are the different types of rocks you have observed in the field?
b) Mention the geological ages of the rocks you have observed.
c) Mention the different geological structures you have observed.
d) What kind of geological measurements you did in the field? (10)

INDIAN STATISTICAL INSTITUTE
Semestral –1 Examination: 2006-2007
B.Stat. III Year
Introduction to Sociology

Date: 1. 12. 06 **Maximum Marks: 100** **Duration: Three Hours**

Q. 1. Answer briefly, **any three** of the following questions:

a) What is meant by 'holistic' theory or 'grand' theory? How do you justify Functionalism as a holistic theory? 5+5 =10

b) What do you mean by Conflict theory? What are the major contrasting features of Conflict theory and Functional theory? 5+5 = 10

c) What is functional theory? What are its major limitations? 5+5 = 10

d) What is meant by 'Relational' approach in sociology? How does it differ from the Conventional, Attributional approach? 5+5 = 10

Q. 2. Answer **any three** of the following questions:

a) Define social research. What is survey research? Describe the merits and demerits of face-to-face interview and telephone surveys. 5+5+5 = 15

b)Discuss the role and limitations of statistical method in social research. Illustrate with an example. 10+5 =15

c) What is measurement? Discuss the relationship between theory and measurement. Write briefly on the interrelationship between measurements and statistics in terms of sociological research. 5+5+5 =15

d) What is objectivity? How can objectivity be maintained in sociological research? Point out the limits of objectivity in sociological research. 5+5+5 =15

Q. 3. Write short notes on **any three** of the following: 3×5 = 15

a) Questionnaire
b) Different stages of analysis of data
c) Observational method of data collection
d) Interview method of data collection

P.T.O.

Q. 4. Choose the correct answer and rewrite the answer: 10×1 = 10

a) Who is the author of "Grammar of Politics"?
i) Laski ii) Durkheim
iii) Leacock iv) Locke

b) What is meant by 'Sex ratio'?
i) The number of females per 1000 males in a population
ii) The number of males per 1000 females in a population.
iii) The ratio between the number of males and females in a population.
iv) The ratio between the number of adult males and females in a population.

c) Which among the following is a tentative proposition, the authenticity of which is yet to be ascertained?
i) Index
ii) Introduction
iii) Conclusion
iv) Hypothesis

d) A set of techniques used to measure attractions and repulsions during interpersonal relations in quantitative and diagrammatic terms is called:
i) Sociometry
ii) Social Statistics
iii) Social theory
iv) Social survey

e) The ratio of recorded live births in one year to the total mid-year population multiplied by 1000 is known as:
i) Inverse transition
ii) Crude birth rate
iii) Fertility rate
iv) Fecundity

f) Which one of the following methods consists of observation, recording, classification, hypothesis, verification and prediction?
i) Scientific method
ii) Deductive method
iii) Ideal-type method
iv) Historical method

g) India is a/an:
i) Complex society
ii) Agricultural society
iii) Socialistic society
v) Traditional society

h) What are the two types of 'observations'?
i) Formal and informal
ii) Open and closed
iii) General and specific
v) None of the above

i)Who wrote the book "The Study of Man"?
i) Ralph Linton
ii) MacIver
iii) Ramkrishna Mukherjee
iv) Garner

j) Who wrote the book "Six Villages of Bengal"?
i) Garner
ii) Ramkrishna Mukherjee
iii) MacIver and Page
iv) A. R. Desai

INDIAN STATISTICAL INSTITUTE

First Semester Examination (2006-2007)
B.Stat. III year
Introduction to Anthropology and Human Genetics

Date: 1.12.06 Maximum Marks:.....30.....Duration...2hr.....

Attempt any four Questions
Each question carries 8 marks. The maximum you can score is 30

1)

A) Match the following

i) Darwin	a.	Laws of inheritance
ii) Lamarck	b.	Inbreeding coefficient
iii) Wright	c.	Organic evolution
iv) Malaria	d.	ABO blood group
v) Landsteiner	e.	Inheritance of acquired characteristics
vi) Mendel	f.	Sickle cell anemia

(3)

B) Two men Mr. X and Mr. Y having blood group AB and B respectively claimed parentage of a lost infant whose blood group is O. Who can be excluded as father of the infant? If the second person is the father of the infant, explain if it is possible to determine the blood group of the mother?

(3)

C) Choose the correct response (2)

i) *If natural selection causes both homozygotes (AA and aa) for a trait to die in early childhood, the result for the population will be:*

a) extinction
b) elimination of the recessive allele in one generation
c) only hèterozygous individuals will survive to reproduce

ii) *If natural selection is against individuals who are heterozygous (Aa) for a particular trait and it always causes death in early childhood, the result for the population will be:*

a) only homozygous individuals (AA and aa) will survive to reproduce
b) elimination of the recessive allele (a) in one generation
c) only heterozygous births will occur, which will ultimately result in extinction.

P.T.O

iii) If nature selects only against people who are homozygous recessive (aa) for a particular trait and it always causes death in early childhood, the result for the population will be:

a) a progressive decrease in the recessive allele (a)
b) an elimination of the recessive allele (a) in one generation
c) a gradual increase in the number of people who are heterozygous (Aa)

iv) In order for a new recessive allele (created by a mutation) to be selected for or against by nature, it must:

a) be expressed in the phenotype of an individual
b) appear in the genotype of an individual
c) be inherited by at least 10% of a population
d) be inherited by at least 50% of a population

2) Define (any 4) (1 to 2 sentences, appox 30 words) 4X2 = 8

i) Fertility
ii) Fecundity
iii) Cross sectional and longitudinal growth studies
iv) Inbreeding coefficient
v) Genetic drift
vi) Mutation

3) Write short note on the following (any 2) (appox 100 words) 2X4 = 8

i) Discuss salient features of Darwin's theory of evolution
ii) Anagenesis and Cladogenesis
iii) Erythroblastosis fetalis and Rh incompatibility
iv) Microevolution and Macroevolution

4) What is polymorphism? How can balanced polymorphism be maintained for traits which are disadvantaged. Discuss citing example of sickle cell trait.

8

5) What are the major stresses on man at high altitude? Why is high altitude unique as a habitat of man?

5+3 = 8

INDIAN STATISTICAL INSTITUTE
First Semester Examination: 2006-07
B. Stat. III Year
Statistical Inference I

Date:05.12.06 Maximum Marks: 100 Duration: 3 Hours

Answer all questions. Special credit will be given to answers which are complete, to the point and precise.

1. (a) Define a minimal sufficient statistic. Is it unique ? Explain by means of an example.

(b) If $X_1,....,X_n$ are i.i.d $U(\theta_1, \theta_2)$ random variables where $-\infty < \theta_1 < \theta_2 < \infty$, then show that $(X_{(1)}, X_{(n)})$ is minimal sufficient for (θ_1, θ_2).

(c) Show that a minimal sufficient statistic need not be boundedly complete.

[(1+1) + 6 + 6 = 14]

2. Let

$$P_\theta(X=-1)=\theta,\quad P_\theta(X=k)=(1-\theta)^2\theta^k,\quad k=0,1,2,....$$

Find all parametric functions which admit MVUEs.

[10]

3. (a) State and prove the Bhattacharya lower bound to the variance of an unbiased estimator.

(b) Give an application of the result stated in (a).

[(2 + 7) + 7 = 16]

4. (a) State and prove the Neymon – Pearson lemma.

(b) Let $X_1,....,X_n$ be iid $U(\theta, \theta+1)$ random variables where $-\infty < \theta < \infty$. Does there exist an UMP test for testing $H_0: \theta \le 0$ vs. $H_1: \theta > 0$?

[(2 +8) + 10 = 20]

5. (a) Consider the following distriution

$$f_\theta(x) = c(\theta)\exp(\theta T(x))h(x).$$

Show that for the testing problem $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$, there exists an UMPU test of size α.

(b) Find an UMPU test of size α for the testing problem

$$H_0: \sigma = \sigma_0 \quad \text{vs.} \quad H_1: \sigma \neq \sigma_0$$

based on the iid observations $X_1, ..., X_n$ from a $N(0; \sigma^2)$ population.

[8 + 7 = 15]

P.T.O

6. Give an example where there exists no consistent estimator of any non-constant parametric function.

[10]

7. (a) Let X follow Bin$(n;\theta)$, $0<\theta<1$. Find a minimax estimator of θ.

(b) Let $X_1,\ldots,X_n$ be iid $N(\theta;1)$, $a\le\theta<\infty$ where a is a known real number. Show that the sample mean $\overline{X}$ is inadmissible but $\overline{X}$ is a minimax estimator of θ.

[7 + (1 + 7) = 15]

INDIAN STATISTICAL INSTITUTE
First Semestral Examination : 2006-2007
B.Stat.(Hons.) III Year
Sample Surveys

Date : 08.12.06 **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.(a) If the sample size required to estimate the proportion of workers in a population with an RSE of α % is n in SRSWR , determine the sample size required to estimate the proportion of non-workers with the same precision .

(b) From an SRSWOR sample of n units a random sub-sample of m units are duplicated and added to the original sample . Show that the mean based on (n+m) units is an unbiased estimator of the population mean and its variance is greater than the variance of the mean based on n units .

(10 + 15) = [25]

2. (a) Find the bias in $\hat{Y}_R=\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{x_i}.X$ as an estimator of the population total Y of the study variable y under SRSWOR and an unbiased estimator of the bias .Hence find an unbiased estimator of Y utilizing the information on the auxiliary variable where X is the population total of the auxiliary variable x .

(b) Explain why it is not generally possible to estimate unbiasedly the sampling variance of the estimated mean based on a single systematic sample . What do you mean by interpenetrating network of sub-samples ? Explain how this technique can be utilized in estimating unbiasedly the sampling variance of the estimated mean in case of circular systematic sampling .

(13 +12)=[25]

3.(a) Derive an approximate expression for the mean square error of the regression estimator of the population mean based on SRSWOR sampling .

(b) Compare the precisions of the ratio and the regression estimators both based on SRSWOR sampling scheme .

(15 + 10) = [25]

4. Suppose a population consists of N first stage units and the ith first stage unit consists of M_i second stage units , i = 1,2,..., N . Suppose a sample of n first stage units is drawn from the population by SRSWOR sampling scheme and if the ith first stage unit is selected , a sample of m_i second stage units is selected again by SRSWOR sampling scheme . Obtain an unbiased estimator of the population total on the basis of the sample drawn and derive an expression for its sampling variance . Also obtain an unbiased estimator of the sampling variance .

(5+10+10)=[25]

5.(a) Describe how you would unbiasedly estimate the population proportion of an attribute based on a stratified simple random sample . Derive an expression for the sampling variance of the estimator .

P.T.O.

(b) Derive Neyman's optimum allocation formula under the set-up in (a) and also an expression for the variance of the estimated proportion under Neyman's optimum allocation formula .

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

6. In a demographic survey , it is proposed to use stratified sampling taking the districts in a region as strata . The relevant data are given in the following table .

District Sl. No. (h)	No. of villages (N_h)	Average population per village ($\bar{Y}_h$)	Standard deviation (S_h)
1.	1953	487	564
2.	1664	829	931
3.	1381	822	996
4.	1174	1083	1167
5.	531	1956	1940
6.	1391	664	625
7.	1996	456	779
8.	1951	372	556
9.	3369	339	591

(a) Assuming that the cost of enumeration and tabulation per person is $\frac{1}{4}$ th of a rupee and the overhead cost is Rs.10,000 , determine the optimum values of n_h's that would minimize the sampling variance of the estimator of the overall population mean for a given expected total cost of Rs.80,000 when villages are selected using SRSWR from each stratum .

(b) For the same value of the total sample size n obtained in (a) find the values of n_h's when the allocation is made in proportion to $N_h S_h$ and obtain the cost -efficiency of the procedure as compared to that of (a) .

(10 + 15) =[25]

INDIAN STATISTICAL INSTITUTE
First Semester Backpaper Examination: 2006-07

B. Stat. III Year
Statistical Inference I

Date:07.02.07 **Maximum Marks: 100** **Duration: 3 Hours**

Answer all questions.

1. a) State Fisher – Neyman factorization theorem. Prove it for discrete probability distributions.

   b) Let $X_1,...,X_n$ be iid random variables following $N(\mu,\sigma^2)$ where $-\infty<\mu<\infty$, $0<\sigma^2<\infty$. Show that $\left(\sum_1^n X_i, \sum_1^n X_i^2\right)$ is sufficient for (μ,σ^2). Is it minimal sufficient ? Justify your assertion.

   [(1+9) + (5+5) = 20]

2. Define an MLR family of distributions. Give an example of such a family whose supports depend on the underlying parameter. Justify your answer.

   [2+5=7]

3. Let $X_1,...,X_n$ be as in Q.1 (b). Find the MVUE of $P(X_1>c)$

   where c is a known constant. Find also the MVUE of the density of X_1 at x in case σ^2 is known; here x is a known real number.

   [10+6=16]

4. a) Let $X_1,...,X_n$ be iid Bin $(1;\theta)$ where $0<a\le\theta\le b<1$ for some known constants a, b with a < b. Find the MLE of θ.

   b) Let $X_1,...,X_n$ be as in Q.1 (b). Describe the likelihood ratio test for testing $H_0:\mu=\mu_0$ versus $H_1:\mu\neq\mu_0$.

   [6+9=15]

5. a) State and prove Cramer – Rao Inequality.

   b) Give an example where this inequality can be used to get an MVUE. State clearly the results you are using.

   [(2+8) + 7 = 17]

P.T.O

6. a) If the loss is squared error, show that a Bayes estimator, if unbiased, has Bayes risk zero.

a) Suppose the risk function $R(\theta,\delta)$ is continuous in θ for each δ where - $\infty < \theta < \infty$. Let π be a prior distribution of θ such that the Bayes risk $r(\pi,\delta) < \infty$ for any δ and that $P_\pi(a<\theta<b)>0$ for any a < b.

Show that a Bayes rule with respect to π is admissible.

c) Let $X_1,\ldots,X_n$ be iid $N(\theta;1), -\infty<\theta<\infty$.

If the loss is squared error, show that the sample mean $\overline{X}$ is both admissible and minimax estimator of θ.

[5+8+ (4+3) = 20]

7. Let the distribution of X be given by

x	0	1	2	3
$P_\theta(X=x)$	θ	2θ	$\cdot 9-2\theta$	$\cdot 1-\theta$

Where $0 < \theta < 0.1$. For testing $H_0:\theta=0.05$ vs. $H_1:\theta>0.05$ at level $\alpha = 0.05$, determine which of the following tests, if any, is UMP :

(i) $\phi(0)=1,\ \phi(1)=\phi(2)=\phi(3)=0;$
(ii) $\phi(1)=0.5,\ \phi(0)=\phi(2)=\phi(3)=0;$
(iii) $\phi(3=1,)\quad \phi(0)=\phi(1)=\phi(2)=0.$

[5]

*****

INDIAN STATISTICAL INSTITUTE

Semestral Examination (Back-paper)

First Semester 2006–2007

B. Stat (Third year)

Differential Equations

Date: 9. 2. 2007

Maximum Marks: 100

Duration: 3 hours

Answer all questions.

(1) Consider the differential equation $y^{(n)}+a_1(x)y^{(n-1)}+\cdots+a_n(x)y=0$, where $a_i(x)$, $i=1,2,\ldots a_n$ are real valued functions defined on some open interval I of the real line.

(a) Show that every complex valued solution of this equation must be C^∞ if a_i's are C^∞ functions.

(b) Let $x_0 \in I$ and let $(\alpha_0,\alpha_1,\ldots,\alpha_{n-1}) \in \mathbb{R}^n$. If ϕ is a solution of the above equation which further satisfies the initial conditions

$$\phi(x_0)=\alpha_0, \phi'(x_0)=\alpha_1,\ldots,\phi^{(n-1)}(x_0)=\alpha_{n-1}$$

then show that ϕ must be a real valued function.

(c) Prove that the solution space of the differential equation is generated by n linearly independent real valued functions. 6+7+9

(2) Find the complete set of solutions of the equation

$$y''+4y=\cos x. \qquad 12$$

(3) Consider the differential equation $y''+a_1(x)y'+a_2(x)y=0$, where a_1 and a_2 are real valued functions defined on $\mathbb{R}$. Let ϕ_1 and ϕ_2 be two linearly independent real valued solutions of this differential equation.

(a) Show that the function $\phi_1\phi_2'-\phi_2\phi_1'$ is everywhere non-vanishing.

(b) Hence prove that ϕ_1 vanishes exactly once between any two successive zeros of ϕ_2.

(c) Consider specific functions for $a_1(x)$ and $a_2(x)$ (both of them can not be identically zero). Solve the corresponding differential equation and show that the solutions ϕ_1 and ϕ_2 support the results (a) and (b) obtained above. 8+12+6

P.T.O

(4) Consider the equation $\frac{dy}{dx} = f(x)(\cos y)^n + g(x)(\sin y)^n$, where f and g are continuous functions on $\mathbb{R}$. Show that every initial value problem for this equation has a solution which is defined on all of $\mathbb{R}$. 12

(5) Write down the general solution of the system of equations

$$\frac{dx}{dt} = 2x - y; \qquad \frac{dy}{dt} = y.$$

Find the equations of the phase curves and draw a rough sketch of them near the critical point.

6+6+4

(6) Let $f : \mathbb{R}^3 \longrightarrow \mathbb{R}$ be a function given by $f(x, y, z) = a(x)z^2 + 2b(x)yz + c(x)y^2$, where a, b and c are continuous functions on $\mathbb{R}$. Let (x_1, y_1) and (x_2, y_2) be two fixed points in $\mathbb{R}^2$. Define a functional I by

$$I(\mathbf{y}) = \int_{x_1}^{x_2} f(x, \mathbf{y}(x), \mathbf{y}'(x))dx,$$

where $\mathbf{y}$ is a real valued function on $\mathbb{R}$ with continuous second derivative satisfying the boundary conditions $\mathbf{y}(x_1) = y_1$ and $\mathbf{y}(x_2) = y_2$. Show that the extremals of I are solutions to some second order linear differential equation. 12

INDIAN STATISTICAL INSTITUTE

First Semester Backpaper Examination (2006-07)

B.Stat. (Hons.) III year

Linear Statistical Models

Maximum time: 3 hours **Date: 15.2.07** Maximum marks: 100

[This test is closed book. The total number of marks of all the questions is 110, and the maximum you can score is 45. Except for question 2, you can use any result that was proved in class.]

1. Prove that a linear unbiased estimator of an estimable parametric function in a linear model is the corresponding BLUE if and only if it is uncorrelated with every linear zero function. [8]

2. Suppose that $\boldsymbol{X} = (\boldsymbol{x}_{(1)} : \cdots : \boldsymbol{x}_{(k)})$, $\boldsymbol{D}$ is a diagonal matrix with $\|\boldsymbol{x}_{(j)}\|$ in the jth diagonal position and $\boldsymbol{X}_s = \boldsymbol{X}\boldsymbol{D}^{-1}$, that is, the columns of $\boldsymbol{X}_s$ have unit norm and are proportional to the columns of $\boldsymbol{X}$. Let $\boldsymbol{X}$ have full column rank.

   (a) Show that VIF_j is the jth diagonal element of the matrix $(\boldsymbol{X}_s'\boldsymbol{X}_s)^{-1}$. [Thus, it is insensitive to a change in scale of the corresponding variable.]

   (b) Show that $VIF_j \geq 1$ for $j = 1, \ldots, k$.

   (c) If $\lambda_1 \geq \cdots \geq \lambda_k$ are the eigenvalues of $\boldsymbol{X}_s'\boldsymbol{X}_s$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ are the corresponding eigenvectors, show that

$$VIF_j = \sum_{i=1}^{k} \frac{v_{ij}^2}{\lambda_i}, \quad j = 1, \ldots, k,$$

   v_{ij} being the ith element of $\boldsymbol{v}_j$, $i, j = 1, \ldots, k$. [4+4+4=12]

3. Let $\boldsymbol{p}_1'\boldsymbol{\beta}$ and $\boldsymbol{p}_2'\boldsymbol{\beta}$ be two estimable LPFs in the linear model $(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ with normally distributed errors, and suppose that $\lambda = \boldsymbol{p}_1'\boldsymbol{\beta}/\boldsymbol{p}_2'\boldsymbol{\beta}$. Find a 95% confidence interval of λ in the following manner.

   (a) Find the mean and variance of $a = \boldsymbol{p}_1'\widehat{\boldsymbol{\beta}} - \lambda\boldsymbol{p}_2'\widehat{\boldsymbol{\beta}}$, where λ is the true value of the ratio of the parameters and $\boldsymbol{p}_1'\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{p}_2'\widehat{\boldsymbol{\beta}}$ are BLUEs. Can a be called an LZF?

   (b) Determine the distribution of $(a^2/Var(a)) \times (\sigma^2/\widehat{\sigma}^2)$, where $\widehat{\sigma}^2$ is the usual estimator of σ^2.

   (c) Show that the ratio of part (b) is less than a given constant if and only if a quadratic function in λ is negative. Using this fact, obtain a two-sided confidence interval for λ. [(2+1)+2+5=10]

4. You have observations (x_i, y_i), $i = 1, \ldots, n$. Given the x_is, the y_is are independent and have normal distribution with mean $\alpha + \beta x_i$ and variance σ^2. The parameters α, β and σ^2 are unknown. Derive an expression (in as simplified form as possible) for a confidence band for the straight line $y = \alpha + \beta x$ with exact coverage probability 0.95. [10]

Derive an exact multiple comparisons test for all pairs of group means in a one way classified data model with normally distributed observations. Is is possible to extend this procedure to two-way classified data with one observation per cell? Explain. [7+3=10]

6. Consider the two-way classification model with m observations per cell. We wish to test the hypothesis $\gamma_{ij} = 0$ for all i, j (that is, no interaction). Find the testable part of this hypothesis and interpret the result. What is the untestable part of the hypothesis?

[4+4=8]

7. Describe a linear model with three-way classified data with a single observation per cell and no interaction, the classification being according to two types of block factors (with b and h levels, respectively) and a treatment effect (t levels). Give the ANOVA table.

[3+5=8]

8. set of t drugs, each having d dose levels, are administered to subjects divided into b blocks. Each dose level of every drug is applied to m subjects of every block, while the allocation is completely random. The response is a measure of degree of relief caused by the drug.

   (a) Write down a suitable nested model for this set-up and derive the ANOVA table.

   (b) How will you test the hypotheses that the various dose levels of Drug 1 do not have different effects? [4+4=8]

9. Describe the Analysis of covariance table for two-way classified data with single observation per cell when there is a single covariate. Adapt Tukey's one-degree of freedom test for interaction to the present situation, and indicate explicitly the test statistic, along with its null distribution. [4+11=15]

10. Explain what is meant by a generalized linear model and derive expressions for the mean and variance of the response on terms of the explanatory variables. Simplify these expressions for the case when the canonical link function is used. [2+2+2=6]

11. You have three-way classified count data $\{y_{ijk}\}$ having the Poisson distribution, with

$$\log(E(y_{ijk})) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}, \quad i = 1, \ldots, I, \ j = 1, \ldots, J, \ k = 1, \ldots, K,$$

where the summands on the right hand side are unspecified parameters.

   (a) Interpret the above model and suggest suitable side-conditions for identifiability of the parameters.

   (b) Derive maximum likelihood estimators of the parameters and of $E(y_{ijk})$.

   (c) Suggest a suitable test for significance of the interaction term. [2+5+5=15]

Indian Statistical Institute
Mid-semester Examination : Second Semester(2006-2007)
B.Stat(hons.) III
Statistical Inference II
Date :- 19.02.2007 Maximum Marks: 50 Duration :- 3 Hours

This paper contains questions worth 55 points. Answer as many as you can. The maximum you can get is 50 points.

1. (a) Let $X_1, \ldots, X_n$ be i.i.d. observations from some continuous distribution F. We want to test $H_0 : F(x) = F_0(x)$ for all $x \in \mathrm{R}$, where $F_0(.)$ is a completely specified continuous distribution. Show that under H_0, D_n^+ and D_n^- have the same distribution, where these are the one-sided one-sample Kolmogorov-Smirnov test criteria.

(b) Suppose we want to estimate the distribution function of a continuous population using the empirical distribution function of a sample of size n from the population, such that the probability is approximately 0.90 that the absolute value of the error of the estimate does not exceed 0.25 anywhere on the real line. How large should n be to ensure this ? [4+2=6]

2. In a random arrangement of 8 A's and 8 B's in a line, find the probability of observing an A-run of length at least 5. Prove your answer. [4]

3. Let $X_i, i = 1, \ldots, n$ be i.i.d. F where $F(x) = G(x - \delta)$ where $G(.)$ is a continuous distribution symmetric around zero and δ is unknown.

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

(b) Write down the Hodges-Lehmann estimtor for δ.

(c) Can you heuristically motivate this estimator of part (b) using the symmetry assumption ? [5+1+2=8]

4. Let $X_1, \ldots, X_n$ be i.i.d. F and $Y_1, \ldots, Y_n$ be i.i.d. G where F and G are continuous distribution functions. In the following $D_{n,n}$ and $D_{n,n}^+$ denote respectively two-sided and one-sided Kolmogorov-Smirnov two-sample test criteria. Let $H_0 : F(x) = G(x)\ \forall x \in \mathrm{R}$.

(a) Find $P_{H_0}(D_{n,n}^+ \geq \frac{k}{n})$, where $k = 0, 1, \ldots, n$. Prove your answer.

(b) Using part (a), find the limiting value of $P_{H_0}(D_{n,n}^+ > \lambda\sqrt{\frac{2\log n}{n}})$ as $n \to \infty$ where $\lambda > 0$. Prove your answer.

(c) Show that the test based on $D_{n,n}$ is consistent for testing $H_0 : F(x) = G(x)$ $\forall x \in \mathrm{R}$ against the alternative that F and G are not identical. You can directly use the fact that the upper $\alpha\%$ point of $D_{n,n}$ tends to zero as $n \to \infty$.

(d) Find the value of $D_{4,4}$ for the sample arrangement $XYYXXYYX$. Find the p-value of this sample arrangement for testing against the alternative that F and G are not identical. [3+5+4+4=16]

P.T.O

5. Let $X_i, i = 1, \ldots, m$ be i.i.d. F and $Y_i, i = 1, \ldots, n$ be i.i.d. G where $G(x) = F(x - \Delta)$ for all $x \in \mathrm{R}$ and some $\Delta \in \mathrm{R}$, where $F(.)$ is a continuous distribution function.

(a) Without using any approximation, show that $\pi(\Delta)$ defined as $P_{F,G}(W_{XY} \geq c)$ is a non-decreasing function of $\Delta \in \mathrm{R}$. Here c is the upper 5% point of the null distribution (i.e when $F(x) = G(x)\ \forall x \in \mathrm{R}$) of W_{XY} and W_{XY} equals the number of pairs (i, j) such that $X_i < Y_j$, $1 \leq i \leq m$, $1 \leq j \leq n$.

(b) Let us consider the case $m = n = 30$, $F(x) = \Phi(\frac{x-\zeta}{2})$ and $G(x) = \Phi(\frac{x-\zeta-1}{2})$, where $\Phi(.)$ is the c.d.f. of the standard normal distribution and ζ is unknown. Find the approximate value of $\pi(1)$, where $\pi(.)$ is defined in part (a).

(c) Consider $H_0 : F(x) = G(x), \forall x \in \mathrm{R}$. Let the generic notation $(W_s = r|k, l)$ 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 greater than $N = m + n$,

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

[5+3+4=12]

6. Let $X_i, i = 1, \ldots, m$ be i.i.d. F and $Y_i, i = 1, \ldots, n$ be i.i.d. G where $G(x) = F(x - \Delta)$ for all $x \in \mathrm{R}$ and some $\Delta \in \mathrm{R}$, where $F(.)$ is a continuous distribution function. Let $\hat{\Delta}$ be the Hodges-Lehmann estimator of Δ. Then show that

(a) Distribution of $\hat{\Delta} - \Delta$ in this shift-model setup is the same for all $\Delta \in \mathrm{R}$.

(b) $\hat{\Delta}$ is symmetrically distributed about Δ if either (i) $F(.)$ is symmetric about some point $\mu \in \mathrm{R}$ or (ii) $m = n$.

(c) If the product mn is odd then $P(\hat{\Delta} < \Delta) = P(\hat{\Delta} > \Delta)$. [1+5+3=9]

Table A Normal distribution

Each table entry is the cumulative probability P, right tail from the value of z to plus infinity, and also left tail from minus infinity to $-z$, for all $P \leq .50$. Read down the first column to the first decimal value of z, and over to the correct column for the second decimal value; the number at the intersection is P.

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002

Source: Adapted from Table 1 of Pearson, E. S., and H. O. Hartley, eds. (1954), *Biometrika Tables for Statisticians*, Volume 1, Cambridge University Press, Cambridge, England, with permission of the Biometrika Trustees.

P.T.O

Table F Kolmogorov-Smirnov one-sample statistic

Each table entry is the value of a Kolmogorov-Smirnov one-sample statistic D_n for sample size n such that its right-tail probability is the value given on the top row.

n	.200	.100	.050	.020	.010
1	.900	.950	.975	.990	.995
2	.684	.776	.842	.900	.929
3	.565	.636	.780	.785	.829
4	.493	.565	.624	.689	.734
5	.447	.509	.563	.627	.669
6	.410	.468	.519	.577	.617
7	.381	.436	.483	.538	.576
8	.358	.410	.454	.507	.542
9	.339	.387	.430	.480	.513
10	.323	.369	.409	.457	.489
11	.308	.352	.391	.437	.468
12	.296	.338	.375	.419	.449
13	.285	.325	.361	.404	.432
14	.275	.314	.349	.390	.418
15	.266	.304	.338	.377	.404
16	.258	.295	.327	.366	.392
17	.250	.286	.318	.355	.381
18	.244	.279	.309	.346	.371
19	.237	.271	.301	.337	.361
20	.232	.265	.294	.329	.352
21	.226	.259	.287	.321	.344
22	.221	.253	.281	.314	.337
23	.216	.247	.275	.307	.330
24	.212	.242	.269	.301	.323
25	.208	.238	.264	.295	.317
26	.204	.233	.259	.290	.311
27	.200	.229	.254	.284	.305
28	.197	.225	.250	.279	.300
29	.193	.221	.246	.275	.295
30	.190	.218	.242	.270	.290
31	.187	.214	.238	.266	.285
32	.184	.211	.234	.262	.281
33	.182	.208	.231	.258	.277
34	.179	.205	.227	.254	.273
35	.177	.202	.224	.251	.269
36	.174	.199	.221	.247	.265
37	.172	.196	.218	.244	.262
38	.170	.194	.215	.241	.258
39	.168	.191	.213	.238	.255
40	.165	.189	.210	.235	.252

For $n > 40$, right-tail critical values based on the asymptotic distribution can be calculated as follows:

.200	.100	.050	.020	.010
$1.07/\sqrt{n}$	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.52/\sqrt{n}$	$1.63/\sqrt{n}$

Source: Adapted from L. H. Miller (1956), Table of percentage points of Kolmogorov statistics, *Journal of the American Statistical Association*, 51, 111–121, with permission.

INDIAN STATISTICAL INSTITUTE
MID- SEMESTERAL EXAMINATION: 2006 -2007
Subject: Design of Experiments
B. Stat. III Year

Date of Examination: 21.02.07 Maximum Marks: 75 Duration: 2½ hours

1. Answer all questions
2. The Paper carries 85 Marks But the maximum you can score is 75

1) Two batches of steel components have arrived. Physical properties (three in number) are measured taking samples from each batch. The manager wants to know if the physical properties are the same for both the batches. Is it an experimental study? Justify your answer.
[2]

2) What are the basic principles of experimental design ? What purposes do the serve ?
[2 x 3 = 6]

3) Consider the problem of optimization of paper helicopter design parameters discussed in the class. Can you suggest a possible blocking factor and at least two noise factors? Justify your answer.
[3 + 3 = 6]

4) An agricultural engineer obtained the following data on two methods of drying corn and asked you for statistical analysis :

Drying Rate

With Preheater	**With Preheater**
16	20
12	10
22	21
14	10
19	12

Under what circumstances would you be justified using

(a) a paired t test or

(b) an unpaired t test

[3 + 3 = 6]

5) A fibre optic cable company wants to compare the performances of its fibre ribbon production lines. Average meters of ribbon produced before a defect is observed are treated as the response and the variance of the response as estimated from past data is - $10^5 m^2$. The engineers wish to detect a difference in average defect rate of .5 km. between the production lines. See table on page 4.

$\sigma^2 = 10^5 \cdot 10^6$

(a) How many samples should be produced by each line in order to detect the average difference of 0.5 km with 90% probability? ($\alpha = .05$) [Hint: $\Phi^2 = n\Delta^2/(2a\sigma^2)$]

(b) Suppose that the variance of the response after the experiment was estimated to be 1.5 x $10^5 m^2$. What power did the test actually have for detecting the difference of 0.5 km in the response?

[7 + 2 = 9]

6) 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 are some of the shortcomings of this design?
(b) What difficulties will they cause in the analysis?
(c) Give a proper design with exactly 12 classes. You are free to choose the number of teachers. What is this design?

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

7) A researcher wants to study the effect of propeller types on gas consumption for powerboats; 4 propeller types to be compared. Four boats and four drivers are available for the experiment. It was decided to complete all the trials of a single replicate in a single day.

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

(b) What type of experiment is this? Write down the model.

(c) Is there a factor not considered for experimentation that might have an effect on the outcome? Can you suggest an improved design considering this factor ? Write down the model

(d) What are some of the possible noise factors ?

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

8) 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

Batch	Day 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: Sum of squares of all 25 observations = 1071
Sum of squares of Day totals = 4383
Sum of squares of Day 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 = $t_{\alpha/2,\upsilon}\sqrt{(2MS_E/n)}$, υ is the error d.f

[3+1+10+8+3 = 25)]

9) 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 Expt. of	Yield	Order Expt. of	Yield	Order Expt. of	Yield	Order Expt. of	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 of they were run. See the graph and comment on the relevant model assumptions. Can you find out what went wrong ?

[4+4 =8]

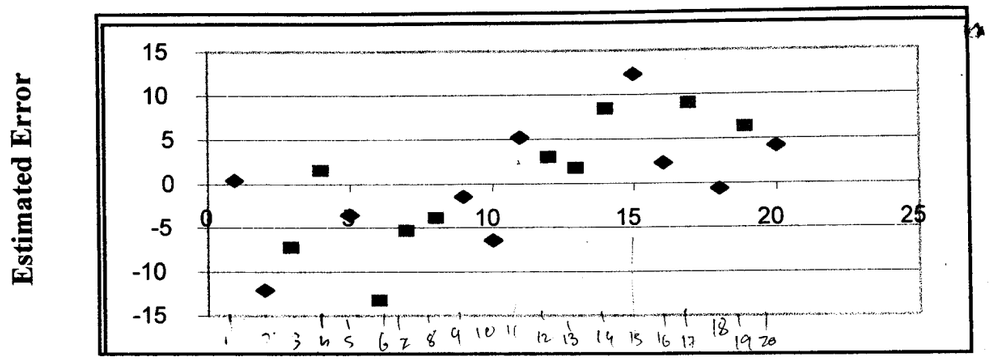


Table A.7 (continued) Power of the F-test: $\pi(\phi) = P(F_{\nu_1,\nu_2,\phi} > F_{\nu_1,\nu_2,\alpha})$

$\nu_1 = 3$, $\alpha = 0.05$

ν_2	1.00	1.33	1.67	2.00	2.33	2.67	3.00	3.33	3.67	4.00	4.33
5	0.19	0.31	0.46	0.61	0.75	0.86	0.93	0.97	0.99	0.99	1.00
6	0.21	0.35	0.51	0.67	0.81	0.90	0.96	0.98	0.99	1.00	1.00
7	0.22	0.37	0.55	0.72	0.85	0.93	0.97	0.99	1.00	1.00	1.00
8	0.24	0.40	0.58	0.75	0.87	0.95	0.98	0.99	1.00	1.00	1.00
9	0.25	0.41	0.60	0.77	0.89	0.96	0.99	1.00	1.00	1.00	1.00
10	0.25	0.43	0.63	0.79	0.91	0.97	0.99	1.00	1.00	1.00	1.00
12	0.27	0.45	0.66	0.82	0.93	0.98	0.99	1.00	1.00	1.00	1.00
15	0.28	0.48	0.69	0.85	0.94	0.98	1.00	1.00	1.00	1.00	1.00
20	0.30	0.51	0.72	0.87	0.96	0.99	1.00	1.00	1.00	1.00	1.00
30	0.32	0.54	0.75	0.90	0.97	0.99	1.00	1.00	1.00	1.00	1.00
60	0.34	0.57	0.78	0.92	0.98	1.00	1.00	1.00	1.00	1.00	1.00
1000	0.36	0.60	0.81	0.93	0.98	1.00	1.00	1.00	1.00	1.00	1.00

INDIAN STATISTICAL INSTITUTE

Periodical Examination

B. Stat. - III Year (Semester - II)

Data Base Management System

Date : 23.02.2007 Maximum Marks : 90 Duration : 3 Hours

1.(a) Suppose we have a database consisting of the following four relation schemes
Branch-scheme=(branch-name, assets, branch-city)
customer-scheme=(customer-name, street, customer-city)
Deposit-scheme=(branch-name, account-number, customer-name, balance)
Borrow-scheme=(branch-name, loan-number, customer-name, amount)
Express the following queries in relational algebra.

(i) delete all accounts at branches located in city 'Kolkata',

(ii) Print the customers those have loan only on the branches where they have accounts,

(iii) Find the customer having largest loan amount in the bank.

1.(b) Give an expression in tuple relational calculus for any one of the above mentioned queries.

1.(c) In relational algebra, express the division operator in terms of fundamental operations.

[(6+10+6)+5+5=32]

2. Consider the following proposed rule for functional dependencies:
If $\alpha -> \beta$ and $\gamma -> \beta$ then $\alpha -> \gamma$

Prove that this rule is not sound (by showing a relation r which satisfies $\alpha -> \beta$ and $\gamma -> \beta$ but does not satisfy $\alpha -> \gamma$) [5]

3. State Armstrong's Axioms and prove that Armstrong's Axioms are complete. [13]

P.T.O

INDIAN STATISTICAL INSTITUTE

Mid Semester Examination, 2^{nd} Semester, 2006-07

Statistics Comprehensive, B.Stat 3^{rd} Year

Date: 27.02.07

Time: 3 hours

Answer any three questions. The maximum you can score is 40.

1. Consider a linear regression set up:

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon \tag{1}$$

where $\epsilon \sim (0, \sigma^2 I)$.

(a) Show that the least squares estimates of β_1 and β_2 can be expressed as

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'(Y - X_2\hat{\beta}_2),$$

and

$$\hat{\beta}_2 = (X_2'QX_2)^{-1}X_2'QY,$$

where $Q = I - X_1(X_1'X_1)^{-1}X_1'$.

(b) Suppose that β_1^*, is the least squares estimate of β_1 obtained from the equation

$$Y = X_1\beta_1 + \epsilon \tag{2}$$

and β_2^* is the least squares estimate of β_2 obtained on regressing the residuals Y^R in (2) on X_2 [under the assumptions that the residuals are uncorrelated]

i. Obtain expressions for β_1^* and β_2^*.

ii. Show that β_1^* and β_2^* are biased in general. When are they unbiased?

iii. If β_2^* is a scalar, show that $\beta_2^* = (1 - R^2)\hat{\beta}_2$, where R is the multiple correlation coefficient between X_2 and X_1.

(c) Suppose that β_2^0 is the least squares estimate of β_2 obtained on regressing Y^R on the residuals X_2^R, after regressing X_2 on X_1. Show that $\beta_2^0 = \hat{\beta}_2$. [15]

P.T.O

4. Find the minimal cover of the following dependency set F
Attribute set $R = \{A, B, C, D, E, G, H\}$

F: $AB \rightarrow CG$ $\quad CG \rightarrow AE$
$BH \rightarrow C$ $\quad A \rightarrow CH$
$AEG \rightarrow C$ $\quad AD \rightarrow GEH$
$ABH \rightarrow ED$ $\quad AC \rightarrow B$

Is your minimal cover unique? [7+3=10]

5. Given the relation scheme $R\{A, B, C, D\}$ and functional dependency $A \rightarrow C$
$D \rightarrow C$ and $BD \rightarrow A$, prove that the decomposition of R onto $R_1\{A, B\}$, $R_2\{A, C, D\}$ and $R_3\{B, C, D\}$ is not lossless. [10]

6. Conclude whether the following decomposition is lossless or lossy. State also under that decomposition whether they preserve dependency.

$R = \{A, B, C, D, E, F, G, H, I\}$

$A \rightarrow BDE$ $\quad B \rightarrow ACDE$
$DE \rightarrow A$ $\quad C \rightarrow DF$
$D \rightarrow IG$ $\quad I \rightarrow G$
$F \rightarrow G$ $\quad FD \rightarrow H$

$R_1\{A, B, C, D, E\}$, $R_2\{C, D, F\}$, $R_3\{D, I, G\}$ and $R_4\{F, D, H, G\}$, [10]

7. Let a prime attribute be one that appears in atleast one candidate key. Let α and β be sets of attributes such that $\alpha \rightarrow \beta$ holds but $\beta \rightarrow \alpha$ does not hold. Let A be an attribute that is not in α and not in β for which $\beta \rightarrow A$ holds. We say that A is *transitively* dependent on α.

Prove that a relation scheme R is in 3NF with respect to a set F of functional dependencies if and only if there are no nonprime attribute A in R for which A is transitively dependent on a candidate key for R. [10]

2. (a) A study on the benefits of Vitamin C showed that 90% of the people suffering from a cold who took Vitamin C got over their cold in a week. Do you think that the study design was appropriate to make conclusions on the effectiveness of Vitamin C? Explain.

(b) Consider two sets of bivariate data on (x, y). The means of x, the means of y, the variances of x and the variances of y are the same for the two data sets, while x and y are positively correlated in both the sets. If the two datasets are pooled, can the correlation coefficient between x and y be zero in the pooled set?

(c) The Hardy-Weinberg Equilibrium law implies that at any locus with alleles $A_1, A_2, \ldots$, the probability distribution of the genotypes is as follows: $P(A_iA_i) = p_i^2\ \forall i$, and $P(A_iA_j) = 2p_ip_j$, $\forall i, j$, $i \neq j$, where $\sum_i p_i = 1$. Data are collected on 157 patients suffering from Alcoholic Cirrhosis. The genotype frequencies at a marker acid phosphatase (ACP) are as follows:

Genotype	Frequency
A_1A_2	5
A_1A_2	55
A_1A_3	15
A_2A_2	65
A_2A_3	5
A_3A_3	1

Do the above data provide evidence that the marker ACP is in Hardy-Weinberg Equilibrium? [3+4+8]

3. (a) Given a random sample $X_1, \ldots, X_n$ from the $U(0, \theta)$ model, find the UMVUE of θ. Show that it is a consistent estimator of θ.

(b) An ecologist, who is interested in studying the effect of ground area on the number of moss plants growing in that area, collected data on the number of plants $(y_1, y_2, \ldots, y_n)$ and the corresponding ground areas $(x_1, x_2, \ldots, x_n)$. Suppose the conditional distribution of y_i given x_i is modeled as Poisson(λx_i). Examine whether the maximum likelihood estimator of λ based on these data is the UMVUE for λ.

(c) Consider the family of distributions $\mathcal{P} = \{P_N : N \geq 1\}$ where

$$P_N\{X = k\} = \begin{cases} \frac{1}{N} & \text{for } k = 1, 2, \ldots, N \\ 0 & \text{otherwise} \end{cases}$$

i. Show that $\mathcal{P}$ is complete.

ii. Fix a positive integer n. Suppose

$$\phi_0(k) = \begin{cases} 0 & \text{for } k = 1, 2, \ldots, n-1, n-2 \\ a & \text{for } k = n \\ -a & \text{for } k = n+1 \end{cases}$$

Obtain $E_N\{\phi_0(X)\}$, and show that $\mathcal{P} - \{P_n\}$ is not complete. [4+5+6]

4. (a) Consider the linear model $(Y, X\beta, \sigma^2 I)$ with the restriction that $A\beta = 0$. Show that $p'\beta$ is estimable under this model if and only if there exists a vector l such that $X'l - p \in C(A')$.

(b) An experiment is conducted to test if there is any difference between blood pressures in the left and the right brachial arteries, and if so to determine the difference. For this experiment, 100 persons are available. Consider the following two study designs.

Design 1: The 100 persons are divided into two groups of 50 at random. In the first group, left blood pressure is observed and in the second group, right blood pressure is observed.

Design 2: On each of the 100 persons, both left and right blood pressures are observed.

Formulating appropriate criteria for comparing the two designs, investigate which design is better.

(c) A survey is conducted among television set owners in two towns with 10000 and 1000000 television set owners respectively to estimate the proportion of owners who watch KBC with Shahrukh Khan. Simple random sampling without replacement is to be used for each town. A total of 1000 owners is to be surveyed. If it is of interest to estimate the overall proportion with the least variance, how would you allocate the 1000 owners to be surveyed between the two towns? [5+5+5]

INDIAN STATISTICAL INSTITUTE

Mid-semester Examination – Semester II : 2006-2007

B.Stat. (Hons.) III Year

Introduction to Stochastic Processes

Date : 02.03.07 **Maximum Score : 40 pts** Time : 3 Hours

Note : This paper carries questions worth a **total** of **50 MARKS**. Answer as much as you can. The **MAXIMUM** you can score is **40 MARKS**.

In the following, **MC** is an abbreviation for a time-homogeneous Markov Chain.

1. Let $\{X_n, n \geq 0\}$ be a MC on some state space S. Starting from the definition, prove the following. Also, in each case, find a formula for the conditional probability in terms of the transition probabilites of the chain.
   (a) $P(X_{n+3} \neq X_{n+2} \,|\, X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i) = P(X_3 \neq X_2 \,|\, X_0 = i)$, for $n \geq 0$ and $i_0, i_1, \ldots, i_{n-1}, i \in S$.
   (b) $P(X_{n+1} = j \,|\, X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i, X_{n+2} = k, \ldots, X_{n+m} = k_m)$
   $= P(X_1 = j \,|\, X_0 = i, X_2 = k)$, for $n \geq 0, m \geq 2$ and $i_0, \ldots, i_{n-1}, i, k, k_3, \ldots, k_m \in S$. (4+4)=[8]

2. Consider a MC $\{X_n, n \geq 0\}$ on the state space $S = \{1, 2, 3, 4\}$ and having transition probability matrix $P = \begin{pmatrix} \frac{1}{4} & 0 & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \\ 0 & \frac{3}{4} & 0 & \frac{1}{4} \end{pmatrix}$.
   (a) Find $P(X_2 = 2, X_1 \neq 2 \,|\, X_0 = 2)$.
   (b) If the chain has initial distribution given by the vector $\mu = (\frac{1}{3}, \frac{1}{3}, 0, \frac{1}{3})'$, find $P(X_2 = 3)$.
   (c) Classify the states as recurrent and transient giving justification.
   (d) With usual notations, find $E_1(N_1)$. (3+3+3+3)=[12]

3. Consider a MC $\{X_n, n \geq 0\}$ on the state space $S = \{1, 2, 3, 4, 5, 6\}$ and having transition probability matrix $P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{5} & \frac{2}{5} & \frac{1}{5} & 0 & \frac{1}{5} \\ 0 & 0 & 0 & \frac{1}{6} & \frac{1}{3} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{3}{4} \end{pmatrix}$.
   (a) Find ρ_{25}.
   (b) If chain has equal probability of starting from each of the 6 states, then what is the probabily that the state 5 will be visited infinitely many times? (4+4)=[8]

4. Consider a system in which a machine in use is replaced instantly, whenever it breaks down, by a new machine. Assume that the lifetimes of the successive machines installed are independent positive integer valued random variables with a common distribution given by $P(L = i) = \frac{1}{i(i+1)}$. Denote X_n to be the age of the machine in operation on the nth day.
   (a) Show that $\{X_n, n \geq 0\}$ is a MC on the set of positive integers and find the transition probabilites.
   (b) With usual notations, find $P_1(T_1 > n)$. Hence classify the states as recurrent and transient.
   (c) Find $E_1(T_1)$ and $E_4(T_1)$. (4+4+5)=[13]

5. (a) Let $\{X_n\}$ be a sequence of real random variables. Show that if τ_1 and τ_2 are two stopping times for the sequence $\{X_n\}$, then so also is $\tau = \tau_1 + \tau_2$.
   (b) Let $\{X_n, n \geq 0\}$ be a simple (not necessarily symmetric) random walk on integers and let τ be a finite stopping time. Define $Y_n = X_{\tau+n} - X_\tau$, $n \geq 0$. Show that $\{Y_n, n \geq 0\}$ is a random walk starting at 0. (4+5)=[9]

INDIAN STATISTICAL INSTITUTE
Second Semestral Examination: 2006 -2007
Subject: Design of Experiments
B. Stat. III Year

Date of Examination: 11.05.07 Maximum Marks: 100 Duration: 3½ hours

Note:
1. Answer all questions
2. The Paper carries 115 Marks But the maximum you can score is 100

1) Explain the meaning of *"Block what you can randomize what you can't"* [2+2=4]

2) 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 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?
(c) Modify the plan so that all the above properties are satisfied.
(d) How will you randomize while conducting the experiment?

[3+6+2+2 = 13]

3) (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.

[7+5 = 12]

4) 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.

(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.

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

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

[3+2+3+7+5+7=27]

5) How will you design a 2^{5-1} design in eight blocks of size 2 so that main effects are clear of block effects?

[8]

6) What are *Hierarchical Ordering*, *Effect Sparsity*, and *Effect Heredity Principles*? Explain their role in construction and analysis of confounded factorial and fractional factorial designs.

[2 x 3 = 6]

7) 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 experiment do you propose?

(b) Identify the generators and write down the defining relations for the fraction you use? Write down the alias 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 which constitute the design

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

[1+5+2 +2+ 6+4 = 20]

8) 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 yield values are given in the table on page 3 of 3.

(i) How will you estimate the experimental error variance in this case? State two alternative approaches. State all the assumptions you need to make.
(ii) Analyze the data using normal probability plots. Prepare the ANOVA table.
(iii) What would be the best possible operating conditions?
(iv) 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.

[5+17+2+3=25]

Variables and their levels

	Variable	Unit	-	+
1	Concentration of catalyst	(%)	5	7
2	Concentration of NaOH	(%)	40	45
3	Agitation speed	(rpm)	10	20
4	Temperature	(^{o}F)	150	180

Design Layout and yield

Trt No	Var-1	Var-2	Var-3	Var-4	Impurity (Coded)	Contrasts without the deviser
1.	-	-	-	-	3.8	79.6
2.	+	-	-	-	4.0	0.2
3.	-	+	-	-	2.7	-13.8
4.	+	+	-	-	3.0	0.4
5.	-	-	+	-	5.8	7.8
6.	+	-	+	-	5.6	-0.8
7.	-	+	+	-	3.0	-6.0
8.	+	+	+	-	3.2	1.4
9.	-	-	-	+	5.9	17.4
10.	+	-	-	+	6.2	-0.8
11.	-	+	-	+	5.3	0.8
12.	+	+	-	+	5.0	-0.6
13.	-	-	+	+	7.9	-0.4
14.	+	-	+	+	7.5	0.2
15.	-	+	+	+	5.3	0.2
16.	+	+	+	+	5.4	0.8

INDIAN STATISTICAL INSTITUTE

Semestral Examination

B. Stat. - III Year (Semester - II)

Database Management System

Date : 14.5.07 Maximum Marks : 100 Duration : 3.30 Hours

Note : You may answer any part of any question, but maximum you can score is 100.

1. Consider the following set of requirements for a university database that is used to keep track of students' transcripts.

   The university keeps track of each student's name, student number, current address, permanent address and phone, birth-date, gender, year of study (1st, 2nd, 3rd), and degree program (B.Stat., B.Math. M.Tech, ..., Ph.D.). Some user applications need to refer to the city, country, post code of the student's permanent address and to the student's last name. Each department is described by name, department code, office number and office phones. Both name and code have unique values for each department. Each course has a course name, description, code number, number of semester-hours and offering department. The value of the code number is unique for each course.

   Draw an entity relationship (ER) diagram for the above database. Your ER diagram should reflect all the information listed above as closely as possible, without adding extra unnecessary constraints. Clearly show the primary keys, existence dependencies, and cardinality constraints. State and justify any additional assumptions that you may make. [14]

2. Suppose that a relation R has six attributes $ABCDEF$ and that the following functional dependencies hold over these attributes:

   $A \rightarrow B$
   $BC \rightarrow E$
   $ED \rightarrow A$
   $F \rightarrow B$

   Decompose R into a BCNF relational schema. Show a decomposition tree describing your decomposition. (Relation R should appear at the root of the tree. If one decomposition step decomposes relation R into relations S and T using a functional dependency $X \rightarrow Y$, then S and T should appear as children of R in the tree, and R should be labeled with $X \rightarrow Y$ to indicate the dependency that was used to decompose it.) [7+3]

P.T.O.

3. Consider the following relations:

Deposit(branch-name, account-number, customer-name, customer-address, balance)
Borrow(branch-name, loan-number, customer-name, customer-address, amount)
Branch(branch-name, assets, branch-city)
Customer(customer-name, street, customer-city)

Write the relational algebra expressions for the following queries and optimize them:

a) Find the names of all customers of "Dunlop" branch who have an account there but not a loan in that branch.

b) Find the assets and names of all branches, which have depositors living in "Kolkata" city. [5+5]

4. Consider the join $R \bowtie_{R.a=S.b} S$, given the following information about the relations to be joined. The cost metric is the number of page I/Os unless otherwise noted, and the cost of writing out the result should be uniformly ignored.

- Relation R contains 200,000 tuples and has 20 tuples per page.
- Relation S contains 4,000,000 tuples and also has 20 tuples per page.
- Attribute a of relation R is the primary key for R. Each tuple of R joins with exactly 20 tuples of S.
- 1,002 buffer pages are available.

Write an efficient algorithm for the above join. What is the cost of joining R and S using your algorithm? [6+6]

5. Consider a relation R with four attributes $ABCD$. You are given the following dependencies: $A \rightarrow B$, $B \rightarrow D$, $CD \rightarrow AB$.

a) List all keys for R.

b) Is R in 3NF?

c) Is R in BCNF? [4+3+3]

6. Let R be a relation scheme and $P = (R_1, R_2)$ a decomposition of R. Let D be a set of functional and multivalued dependencies on the attributes of R. Then P has a lossless join if and only if $(R_1 \bigcup R_2) \twoheadrightarrow (R_1 - R_2)$. [10]

7. From the definition of multivalued dependency, prove that $A \twoheadrightarrow B$, $B \twoheadrightarrow C$ implies $A \twoheadrightarrow (C - B)$ [10]

P.T.O.

8. Schedule S_1 is conflict-equivalent to schedule S_2 if S_2 can be derived from S_1 by a sequence of swaps of non-conflicting actions.

Prove or disprove each of the following statements.

a) If two schedules are conflict equivalent, then their precedence graphs are identical.

b) If two schedules involve the same set of transactions, and have identical precedence graphs, then they are conflict equivalent. [5+5]

9. One can determine whether a particular point (a, b) is in the set by performing a range query with range $[a : a] \times [b : b]$. Show that performing such a range query on a kd-tree takes time $O(\log n)$. [10]

10. You are given the cryptarithmetic puzzle:

```
  S E N D
+ M O R E
---------
M O N E Y
```

The goal of the puzzle is to substitute numbers (from zero to nine) for letters, so that the addition works out. There are some constraints your solution should respect:

a) The same number should be used for a given letter, throughout. For example, if you guess "5" for the letter E, then E should get the value "5" at all the places it occurs.

b) Different letters should get different numbers, e. g., you cannot assign "4" to both E and to M.

c) None of the numbers SEND, MORE, or MONEY has any leading zeroes, i.e., they do not begin with zeroes.

Explain how you will solve this puzzle by creating database tables and writing a query.

[25]

INDIAN STATISTICAL INSTITUTE

Semestral Examination – Semester II : 2006-2007

B.Stat. (Hons.) III Year

Introduction to Stochastic Processes

Date : 18.05.07 **Maximum Score : 60** Time : 3 Hours

Note : This paper carries questions worth a **total** of **68 MARKS**. Answer as much as you can. The **MAXIMUM** you can score is **60 MARKS**.

1. A markov chain on $S = \{1, 2, 3, 4, 5\}$ has transition matrix $P = \begin{pmatrix} \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ 0 & \frac{2}{5} & \frac{2}{5} & \frac{1}{5} & 0 \\ 0 & \frac{3}{10} & \frac{2}{5} & \frac{3}{10} & 0 \\ 0 & \frac{1}{5} & \frac{2}{5} & \frac{2}{5} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix}$.
(a) Does this chain have a stationary distribution? How many? Find the stationary distributions, if there are any.
(b) If the chain has initial distribution $(0, 0, \frac{7}{10}, \frac{3}{10}, 0)$, then find the expected value of the time taken by the chain to return to its initial state. ((3+5)+4)=**[12]**

2. For a markov chain, Let i be state such that $i \rightsquigarrow i$. Denoting $A = \{n \geq 1 : p_{ii}^{(n)} > 0\}$ and $B = \{n \geq 1 : \rho_{ii}^{(n)} > 0\}$, show that the set A consists precisely of all possible finite sums of elements of the set B and hence prove that the period of the state i equals the g.c.d. of the set B. (7+5)=**[12]**

3. Let $\{X_n,\, n \geq 0\}$ be a markov chain on $S = \{1, 2, \ldots, d\}$, with d an absorbing state and all others transient. Fix a transient state i. Define a markov chain $\{Y_n,\, n \geq 0\}$ whose transtion probabilites are the same as those for $\{X_n\}$ **except that** $P(Y_1 = i \,|\, Y_0 = d) = 1$.
(a) Argue carefully that the $\{Y_n\}$-chain has a unique stationary distribution and that this stationary distribution π satisfies $0 < \pi_d < 1$.
(b) Show that for the $\{X_n\}$-chain, the expected time till absorption, given that the chain started at i, equals $(1 - \pi_d)/\pi_d$, where π is as in (a). (7+5)=**[12]**

4. (a) State what is meant by the **Strong Markov Property** of a markov chain.
(b) Let $\{X_n,\, n \geq 0\}$ be an irreducible positive recurrent markov chain on S and let π denote its (unique) stationary distribution. Show that if f is a real-valued function on S such that $\sum_{i \in S} |f(i)|\pi_i < \infty$, then $\frac{1}{n} \sum_{l=0}^{n} f(X_l) \longrightarrow \sum_{i \in S} f(i)\pi_i$ with probability 1, whatever be the initial distribution of the chain. [You may use SLLN for i.i.d. random variables and any related result in probability, but state them clearly!] (3+9)=**[12]**

5. Let $\{N_t,\, t \geq 0\}$ denote a Poisson Process with intensity λ.
(a) Denote $X_t = (N_t - \lambda t)^2$, $t \geq 0$. For $0 \leq s_1 < \cdots < s_k < s < s + t$, find the conditional expectation of $X_{s+t} - X_s$ given $N_{s_1} = n_1, \ldots, N_{s_k} = n_k, N_s = n$. [You may use the fact that for a random variable Z with Possion (μ) distribution, $E(Z) = \mu$, $E(Z^2) = \mu^2 + \mu$.]
(b) Given that $N_t = 6$, find the probability that at least one of these six arrivals happened after time $2t/3$ and at least one before time $t/3$. (6+6)=**[12]**

6. Consider a system consisting of a collection of particles evolving in the following manner. Each particle, independently of others, lives for a random amount of time having $Exp(\lambda)$ distribution, at the end of which it either splits into two new particles or completely disappears with probabilities p and $q(= 1 - p)$ respectively. Also new particles immigrate into the system according to a Poisson Process with rate μ. Identify the state space and the Q-matrix of the continuous time markov chain representing the number of particles in the system. Also find the transition probabilities of the **embedded** chain. (5+3)=**[8]**

Indian Statistical Institute
Second Semestral Examination : (2006-2007) B.Stat(Hons.) III
Statistical Inference II
Date:- 21.05.07 Maximum Marks:- 50 Duration :- 3 hours

This paper contains questions worth a total of 55 points. Answer as much as you can. The maximum you can get is 50 points.

1. (a) Consider i.i.d. observations $X_1, X_2, \ldots$ and SPRT with boundaries $0 < B < 1 < A < \infty$ for testing a simple null hypothesis H_0 versus a simple alternative H_1 concerning $X_1, X_2, \ldots$. Show that the SPRT terminates with probability 1 under H_1.
(b) Let X_i, $i = 1, 2, \ldots$ be i.i.d. Bernoulli (θ), where $0 < \theta < 1$. Consider SPRT for testing $H_0 : \theta = \frac{1}{3}$ against $H_1 : \theta = \frac{1}{2}$ where the boundaries satisfy $0 < B < 1 < A < \infty$. **Without using Stein's lemma**, show directly that the SPRT terminates with probability one if the true θ equals 0.42.
[6+4=10]

2. Let $X_1, \ldots$ be i.i.d. Poisson(θ). We want to test $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$ $(\theta_1 > \theta_0)$. Derive SPRT (with boundaries replaced by Wald's approximation) of target strength (α, β), where $\alpha > 0$, $\beta > 0$ are small positive numbers and $\alpha + \beta < 1$. Find approximate expressions for the OC and ASN functions of this test procedure. Indicate how one can plot the OC function in a graph. [1+5+2=8]

3. Let $X_1, \ldots, X_n$ be i.i.d. observations drawn from some continuous distribution. The null hypothesis is that the sample is from a $N(\mu, \sigma^2)$ distribution where both μ and $\sigma > 0$ are unknown. Consider the one-sided Kolmogorov-Smirnov test statistic with μ and σ estimated by the sample mean $\bar{X}$ and the sample s.d. S respectively, i.e consider

$$\hat{D}_n^+ = \sup_t \{F_n(t) - \Phi((t - \bar{X})/S)\},$$

where $F_n(.)$ is the empirical distribution function and $\Phi(.)$ is the c.d.f. of the standard normal distribution. Show that the null distribution of $\hat{D}_n^+$ does not depend on μ and σ. [5]

4. Let $X_1, \ldots, X_m$ be i.i.d. with distribution function F and $Y_1, \ldots, Y_n$ be i.i.d. with distribution function G, where both F and G are continuous and strictly increasing.
(a) Show that in the case Y is stochastically smaller than X, one can write $G(t) = F(t - \Delta(t))$ for some function $\Delta(.)$, such that $\Delta(t) \leq 0$ for all t. Find an explicit expression for $\Delta(t)$.
(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 using part (a) show that if Y is stochastically smaller than X, then $P(U \leq c_\alpha) \leq \alpha$. [3+4=7]

P.T.O

Year	Divorce Rate
1945	3.5
1955	2.3
1965	2.5
1975	4.8
1985	5.0

5. The table above is based on figures which appeared in the *World Almanac and Book of Facts.* It shows divorce rates per 1000 population in the United States in different years. Is there any evidence in the data of a positive trend in divorce rates with the passage of time ? Justify your answer using some appropriate measure of association.

[3]

6. Consider the two sample problem.
(a) Describe the Wald-Wolfowitz Runs Test for testing $H_0 : F(x) = G(x)$ for all x vs. $H_1 : F(x) \neq G(x)$ for at least some x, where $X_1, \ldots, X_m$ are i.i.d. with distribution F and $Y_1, \ldots, Y_n$ are i.i.d. with distribution G, both F and G being continuous.
(b) Will this test be appropriate for testing against one sided alternatives (say, Y's are stochastically larger than the X's) ? Explain. [2+2=4]

7. State Hoeffding's theorem on the asymptotic distribution of one-sample U-statistic of order m ($m \geq 1$). Briefly indicate the main steps in proving this result. Using this result, find the asymptotic distribution of ${s_n}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ (suitably normalized) where $X_1, \ldots, X_n$ are i.i.d. F with $E_F(X_i^4) < \infty$.

[1+5+3=9]

8. Let $X_1, \ldots, X_n$ be i.i.d. $F(x-\theta)$ where F is a continuous distribution symmetric around 0 and its derivative at 0, denoted $f(0)$, is positive and $\mathrm{Var}_F(X_i) = {\sigma_F}^2 < \infty$. We want to test $H_0 : \theta = 0$ versus $H_1 : \theta > 0$. Derive a general expression for the asymptotic relative efficiency of the sign test relative to the t test in this problem and evaluate it when (i) $F = N(0,1)$ and (ii) $F =$ Double exponential with density $f(x) = \frac{1}{2}e^{-|x|}$, $-\infty < x < \infty$.

[7+2=9]

INDIAN STATISTICAL INSTITUTE

Second Semestral Examination(Back Paper)

B. Stat. - III Year (Semester - II)

Database Management System

Date : 19.7.07 Maximum Marks : 100 Duration : 3.00 Hours

1. Consider the following relation:

$Shipping(ShipName, ShipType, VoyageID, Cargo, Port, Date)$

With the functional dependencies:

$ShipName \rightarrow ShipType$

$VoyageID \rightarrow ShipName, Cargo$

$ShipName, Date \rightarrow VoyageID, Port$

a) Identify the candidate keys.
b) Normalize to 2NF
c) Normalize to 3NF
d) Normalize to BCNF [15]

2. Consider tables $R(A,B,C)$, $S(C,D)$, and $T(D,E)$ where the notations have their usual meanings.

a) Transform the query,
$\pi_{R.B,S.D,T.E}(\sigma_{(R.A=10)\wedge(R.C=S.C)\wedge(S.D=T.D)\wedge(R.A>T.E)}(R \times S \times T))$ into an equivalent query that:
- Contains no cross products
- Performs projections and selections as early as possible.

b) Suppose we have the following statistics:
$| R | = 1,000$; $| \pi_A R | = 1,000$; $| \pi_B R | = 100$; $| \pi_C R | = 500$;
$| S | = 5,000$; $| \pi_C S | = 300$; $| \pi_D S | = 10$;
$| T | = 4,000$; $| \pi_D T | = 4,000$; $| \pi_E T | = 1,500$.
Estimate the number of the rows returned by the following queries:
i. $\sigma_{A=10} R$
ii. $\sigma_{A=10 \vee B=\text{"}Bart\text{"}} R$
iii. $R \bowtie S$
iv. $R \bowtie S \bowtie T$ [9+(4+4+4+4)]

P.T.O.

3. Consider the following relations:

Deposit(branch-name, account-number, customer-name, balance)
Borrow(branch-name, loan-number, customer-name, amount)
Branch(branch-name, assets, branch-city)
Customer(customer-name, street, customer-city)

Express the following queries in relational algebra and SQL :

a) Delete all accounts at branches located in city 'Kolkata'.

b) Print the customers who have loan only on the branches where they have accounts.

c) Find the customer having largest loan amount in the bank. [8+8+9]

4. This database system will be used by the airport to keep track of its airplanes, the airplanes' owners, the airport employees, and the pilots. Following is a description of the information to be contained in the database:

- Each airplane has a unique registration number, is of a particular type, and is stored in a particular hangar.
- Each airplane type has a unique model number, a capacity, and a weight.
- Each hangar has a unique hangar number, a capacity, and a location.
- Airplanes may be owned by individuals and/or corporations. A given airplane may have multiple owners. The date on which each owner purchased an airplane must be stored. (Note that different owners may have purchased their portion of an airplane on different dates.)
- Each individual airplane owner has a unique owner ID number, a name, an address, a date of birth, and a phone number.
- Each corporate airplane owner has a unique owner ID number, a name, an address, and a phone number.
- Each employee has a unique Social Security number, a name, an address, a job title, and a salary.
- Information must be maintained about the type(s) of airplane that each employee is qualified to work on.
- Each airplane undergoes service many times. A service record indicates the date that the airplane was serviced, a work code for the type of work done, and the number of hours of work of that type done on the airplane on that date. More than one employee may be involved in a certain type of work on an airplane at a given time, and an employee may be involved in working on more than one airplane on a given date. (Note that the date of service and the work code will uniquely identify a service record for a particular airplane; thus to uniquely identify a service record, it will be

P.T.O.

necessary to include the airplane registration number, the date of service, and the work code.)

Draw an entity relationship (ER) diagram for the above database. Clearly show the primary keys, existence dependencies, and cardinality constraints. [25]

5. For each of the following decompositions of $R = ABCDYZ$, with the set of functional dependencies $F = \{AB \rightarrow Z, AC, Y \rightarrow C, ZB \rightarrow D, BD \rightarrow Z\}$, say whether the decompositions are (i) dependency preserving, and (ii) lossless join.

   a) ABYD, ABCYZ

   b) ACD, YC, YZ, ABDZ [10]