INDIAN STATISTICAL INSTITUTE

MID-SEMESTRAL EXAMINATION: (2015-2016)

MSQE I and M.Stat II

Microeconomic Theory I

Date: 07.09.2014 Maximum marks: 40 Duration: 2 Hours

Note: Answer **all** questions.

Note: Throughout, $\mathbb{R}^L$ is the L-dimensional Euclidean space. Let

$$\mathbb{R}^L_+ = \{x = (x_1, \cdots, x_L) \in \mathbb{R}^L : x_i \geq 0 \text{ for all } 1 \leq i \leq L\}$$

and

$$\mathbb{R}^L_{++} = \{x = (x_1, \cdots, x_L) \in \mathbb{R}^L : x_i > 0 \text{ for all } 1 \leq i \leq L\}.$$

Q1. Let $\succeq$ be a *preference relation* over a non-empty set of alternatives X. Suppose also that $\succ$ and $\sim$ are the *strict preference relation* and the *indifference relation* respectively associated with $\succeq$.

(i) Show that $\succ$ is transitive. [3]

(ii) Suppose that $\succeq$ is rational and $U : X \to \mathbb{R}$ is a utility function representing $\succeq$, that is, $x \succeq y \Leftrightarrow U(x) \geq U(y)$ for $x, y \in X$. Show that "U is a utility function representing $\succeq$" is equivalent to "$x \succ y \Leftrightarrow U(x) > U(y)$ for any $x, y \in X$". [4]

(iii) Prove or disprove: Suppose that X has only finitely many elements. For any non-empty subset B of X, let $|B|$ denote the number of elements in B. Assume that $\mathscr{B}$ denotes the collection of all non-empty subsets of X. Define choice rules $\mathbb{C}(\cdot)$ on $\mathscr{B}$ by

$$\mathbb{C}(B) := \left\{x \in B : x \succeq y \text{ for all } y \in A \text{ for some } A \subseteq B \text{ with } |A| > \frac{1}{3}|B|\right\}.$$

Then $(\mathscr{B}, \mathbb{C}(\cdot))$ satisfies *WARP*. [3]

(iv) Prove or disprove: If U and V represent $\succeq$ then there is a strictly monotonic function $f : \mathbb{R} \to \mathbb{R}$ such that $V(x) = f(U(x))$. [2]

Q2. Answer **all** questions.

(i) Show that *WARP* is not a sufficient condition to ensure the existence of a rationalizing preference relation. [3]

(ii) Show that for a demand function $x : \mathbb{R}^L_{++} \times \mathbb{R}_{++} \to \mathbb{R}^L_+$ satisfying Walras's law, *WARP* holds for all compensated price changes if and only if for any compensated price change from (p, w) to $(p', w') = (p',\ p' \cdot x(p, w))$, the following inequality holds:

$$(p' - p) \cdot [x(p', w') - x(p, w)] \leq 0,$$

with strict inequality for $x(p', w') \neq x(p, w)$. [5]

(iii) Suppose that X is a non-empty set of alternatives and $(\mathscr{B}, \mathbb{C}_1(\cdot))$ is a choice structure of X that satisfies *WARP*. Consider the choice structure $(\tilde{\mathscr{B}}, \mathbb{C}_2(\cdot))$, where $\tilde{\mathscr{B}} = \{\mathbb{C}_1(B) : B \in \mathscr{B}\}$. Show that if $(\tilde{\mathscr{B}}, \mathbb{C}_2(\cdot))$ satisfies *WARP* then $(\mathscr{B}, \mathbb{C}(\cdot))$ must satisfy *WARP*, where $\mathbb{C}(B) = \mathbb{C}_2(\mathbb{C}_1(B))$ for all $B \in \mathscr{B}$. [4]

Q3. Answer **all** questions.

(i) Prove or disprove: Let $(\mathscr{B}, C(\cdot))$ be a choice structure. The *strict revealed preference relation* $\succ^*$ is defined by $x \succ^* y \Leftrightarrow \exists\ B \in \mathscr{B}$ such that $x, y \in B, x \in C(B)$ and $y \notin C(B)$. If $(\mathscr{B}, C(\cdot))$ satisfies *WARP*, then $\succ^*$ is transitive. [2]

(ii) Suppose that X is a non-empty set of alternatives and $(\mathscr{B}, \mathbb{C}(\cdot))$ is a choice structure of X that satisfies *WARP*. Show that $\succ^*$ and $\succ^{**}$ are identical, that is, for any $x, y \in X$, $x \succ^* y \Leftrightarrow x \succ^{**} y$, where $\succ^*$ is the *strict revealed preference relation*, $x \succ^{**} y \Leftrightarrow [x \succsim^* y$ and $y \not\succsim^* x]$ and $\succsim^*$ is the *revealed preference relation*. [4]

(iii) Let $X = \{x, y, z\}$ and $\mathscr{B} = \{\{x, y\}, \{y, z\}, \{z, x\}, \{x, y, z\}\}$. Suppose that $C(\{x, y\}) = \{x\}$, $C(\{y, z\}) = \{y, z\}$, $C(\{z, x\}) = \{z, x\}$ and $C(\{x, y, z\})$ is a non-empty subset of X. Does $C(\cdot)$ satisfy *WARP*? Justify your answer. [3]

(iv) Let $(\mathscr{B}, C(\cdot))$ be a choice structure. If a rational preference relation $\succsim$ *rationalizes* $C(\cdot)$ relative to $\mathscr{B}$, then show that $C(B_1 \cup B_2) = C(C(B_1) \cup C(B_2))$ for every $B_1, B_2 \in \mathscr{B}$ such that $B_1 \cup B_2 \in \mathscr{B}$ and $C(B_1) \cup C(B_2) \in \mathscr{B}$. [5]

(v) Show that if a demand function $x : \mathbb{R}^L_{++} \times \mathbb{R}_{++} \to \mathbb{R}^L_+$ satisfies *WARP* then it is homogeneous of degree zero. [2]

INDIAN STATISTICAL INSTITUTE

Mid-semester exam. (Semester I: 2015-2016)

Course Name: M. Stat. 2nd year

Subject Name: Analysis of discrete data

Date: 07/09/2015, Maximum Marks: 40. Duration: 2 hrs.

Note: Answer all questions.

1. Geometrically interpret 2×2 contingency tables with given constant margins. [8]

2. Let X and Y be identically distributed categorical random variables. Suppose both X and Y have k categories $C_1, \ldots, C_k$ with probabilities $p_1, \ldots, p_k$, and joint distribution of X and Y is defined by the conditional probabilities

$$P(Y = C_j | X = C_i) = (1 - \alpha)p_j + \alpha I(i = j), \quad i, j = 1, \ldots, k,$$

where $I(i = j)$ is the indicator and $0 \leq \alpha \leq 1$. Find $\tau_{Y|X}$, Goodman and Kruskal's τ, as a function of α. [8]

3. Let X and Y be two nominal categorical random variables with I and J categories. Define $\pi_{ij} = P(X = C_i, Y = D_j)$, $i = 1, \ldots, I$; $j = 1, \ldots, J$. Define the measure of variability $V(\cdot)$ as the probability that two independent guesses are wrong.

   (a) Find $V(Y)$. [4]

   (b) Find the measure of association $A_{Y|X}$ as the proportional reduction of the variability of Y with the knowledge of X. [4]

4. (a) Derive the joint asymptotic distribution of log odds ratios in a 2×4 contingency table. [8]

   (b) Find the P-value of the test for independence for the following table by Fisher's conditional test procedure against one-sided alternative.

	Guess poured first	
Poured first	Milk	Tea
Milk	6	2
Tea	2	6

[8]

INDIAN STATISTICAL INSTITUTE

M.Stat Second Year, First Semester, 2015-16

Mid-Semestral Examination

09·09·2015

Time: $2\frac{1}{2}$ Hours | Statistical Computing | Full Marks: 60

(Answer as many as you can. The maximum you can score is 60.)

1. Use the acceptance-rejection method to generate observations from the following distributions with p.m.f./p.d.f. given by [6 × 2=12]

   (a) $f(x) = \frac{6}{\pi^2 x^2}$; $x = 1, 2, \ldots$.

   (b) $f(x) = \frac{1}{\Gamma(1/3)} e^{-x}\, x^{-2/3}$; $x > 0$.

2. Consider a non-negative bounded continuous function $f : [a, b] \to [0, f_0]$. In order to estimate $\theta = \int_a^b f(t)dt$, a student generated n independent observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ from $U(a, b) \times U(0, f_0)$ and proposed the estimate $\hat{\theta}_1 = \frac{(b-a)f_0}{n} \sum_{i=1}^n I(y_i \leq f(x_i))$. Another student generated n independent observations $z_1, z_2, \ldots, z_n$ from $U(a, b)$ and proposed the estimate $\hat{\theta}_2 = \frac{b-a}{n} \sum_{i=1}^n f(z_i)$. Which of these two estimates will you prefer ? Justify your answer [8]

3. Let F be a continuous distribution with finite second moments. Let $T(F_n)$ be an estimator of $T(F) = \int |x|\, dF(x)$ based on n independent observations $X_1, X_2, \ldots, X_n$.

   (a) Find the jackknife estimator of $Var(T(F_n))$. [4]

   (b) If this jackknife estimator is denoted by V_{jack}, check whether $V_{jack}/Var(T(F_n))$ converges to 1 almost surely. [6]

   (c) Show that the bootstrap version of $T(F_n)$ can be expressed as $\frac{1}{n}\sum_{i=1}^n W_i|X_i|$, where $(W_1, W_2, \ldots, W_n)$ follows a multinomial distribution with parameters $(n, \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$. [4]

4. Consider an auto-regressive model $X_t = \theta X_{t-1} + \epsilon_t$, where $X_0 = 1$, $t = 1, 2, \ldots, n$, and the ϵ_i's are i.i.d. with mean 0 and variance 1.

   (a) Find the least square estimate of θ. [2]

   (b) Describe the residual bootstrap method for constructing a 95% confidence interval for θ. [4]

5. Show that the influence function for the mean of a univariate distribution is unbounded but that for the median is bounded under some appropriate assumptions. [3+5]

6. Let $x_1 = 20.9$, $x_2 = 13.1$, $x_3 = 11.7$, $x_4 = 9.8$, $x_5 = 7.6$, $x_6 = 12.2$, $x_7 = 17.4$ and $x_8 = 10.7$ be $n = 8$ observations from a univariate distribution F.

   (a) Find a value of θ that minimizes $2\sum_{i=1}^8 |x_i - \theta| + \sum_{i=1}^8 (x_i - \theta)$. Is this minimizer unique? Justify your answer. [4+4]

   (b) Find the least median of squares (LMS) and the least trimmed squares (LTS) (based on $\lfloor n/2 \rfloor + 1$ observations) estimates of location of F. [3+5]

INDIAN STATISTICAL INSTITUTE
Mid-Semester Examination: 2015-2016
Course Name: MStat II
Subject Name: Topics in Mathematical Physics

Date: 9 September 2015 **Maximum Marks: 50** **Duration: 2 hrs**

Note: **Answer as many questions as you can. Each question carries 10 marks.**

1. A bob of mass m is hanging from a rigid support by a spring of self-mass M having uniform linear density and spring constant k. Using Lagrangian formalism, show that the time period of small oscillation for such a system, if displaced from its ground state, is given by

$$T = 2\pi\sqrt{\frac{m + M/3}{k}}.$$

[10]

2. (a) Derive Hamilton's equations of motion using variational principle.

   (b) Using the standard Poisson bracket, prove that any physical quantity that commutes with the Hamiltonian of a system leads to a conserved quantity.

[5+5=10]

3. For a central force field the potential energy of the system is a function of the radial coordinate only. Using Hamiltonian formalism find out the equation of motion for a particle moving under such a field. Hence identify centrifugal force of the system.

[(7+3)=10]

4. The Hamiltonian of an oscillating system with mass m and force constant k is given by

$$H = \frac{p^2}{2m} + \frac{1}{2}kq^2,$$

where $\{q,\ p\}$ is a set of generalized coordinate and momentum. Consider a generating function $F = \mu q^2 \cot Q$ (where Q is the new generalized coordinate), show that the motion of the system can be described by simple harmonic oscillation.

[10]

5. (a) Show that for a one-dimensional free particle wave function

$$\psi(x,t) = (2\pi\hbar)^{-1/2}\int_{-\infty}^{\infty} \exp\left[i\left(p_x x - \frac{p_x^2}{2m}t\right)/\hbar\right]\phi(p_x)dp_x$$

the expectation value of the momentum $\langle p_x\rangle$ does not change with time, where the symbols have their usual meanings.

P.T.O

(b) Prove that $\langle xp_x \rangle$ and $\langle p_x x \rangle$ are both non-Hermitian operators. Can you construct a Hermitian operator out of these two variables?

[5+(3+2)=10]

6. For a quantum linear harmonic oscillator, show that the energy is quantized. Hence define ground state energy of the system.
Hint: You can make use of the dimensionless variables $\lambda = \frac{2E}{\hbar\omega}$, $\zeta = x\sqrt{m\omega/\hbar}$ and the wave-function $\psi = \exp[-\zeta^2/2]H(\zeta)$, where $H(\zeta)$ is the Hermite polynomial. The symbols have their usual meanings.

[10]

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination : 2015-16(First Semester)

M. Stat. II year

Topology& Set Theory

Date: 11.09.15 Maximum Marks : 80 Duration : 2 1/2 hrs

Note: Answer as many as you can. The maximum you can score is 80. Notation is as used in the class.

1. (a) When is a set said to be countably infinite?
   Show that $\mathbb{Z}^+ \times \mathbb{Z}^+$ is countably infinite, where $\mathbb{Z}^+$ is the set of positive integers.
   (Show all the steps)

   (b) Give an example of an uncoutable set. Prove that the set is uncountable

   (c) Show that any two closed intervals $[a, b], [c, d]$ are bijectively equivalent, where $a < b, c < d$ are reals. Are they homeomorphic? Justify.

   [8+5+7]

2. Show that the standard topology on $\mathbb{R}$ has a countable basis. Hence, or otherwise, show that the standard topology on $\mathbb{R}^n$ has a countable basis [7]

3. Let A, B be subsets of a topological space X

   (a) Show that $\overline{A \bigcup B} = \overline{A} \bigcup \overline{B}$

   (b) Prove or disprove $\overline{A \bigcap B} = \overline{A} \bigcap \overline{B}$.

   (c) Show that if $A \subseteq X$ and $B \subseteq Y$, then $\overline{A \times B} = \overline{A} \times \overline{B}$

   [5+4+5]

4. Let $f : X \longrightarrow Y$, where X and Y are topological spaces.

   (a) Show that f is continuous iff for every subset $A \subseteq X$, $f(\overline{A}) \subseteq \overline{f(A)}$.

(b) Show that if X is metrizable, then f is continuous iff for every sequence $x_n \to x$ in X, $f(x_n) \to f(x)$ in Y.

[8+7]

5. Let (X, d) be a metric space. Show that $d : X \times X \longrightarrow I\!R$ is continuous. [7]

6. Consider the product space $I\!R^\omega$, where $I\!R$ is equipped with the standard topology. *Define* a metric on $I\!R^\omega$, that realizes the product topology [5]

7. Show that if X and Y are connected then so is $X \times Y$. [8]

8. (a) Let $\mathbf{R}^n \subseteq I\!R^\omega$ be the set of all sequences $\mathbf{x} = (x_1, x_2, ...)$ such that $x_i = 0$ for $i > n$. Is $\mathbf{R}^n$ connected ? Justify.

   (b) Let $I\!R^\infty$ be the union of these spaces. Is $I\!R^\infty$ connected? Justify.

   (c) Find the closure of $I\!R^\infty$ in $I\!R^\omega$.

   (d) Conclude that $I\!R^\omega$ is connected.

[3+3+4+3]

INDIAN STATISTICAL INSTITUTE
Mid-Term Examination: 2015-2016
MS (Q.E.) II and MStat II Year
Econometric Methods II/ Econometric Methods

Date: 11th September 2015 Maximum Marks 30 Duration 2 hours

All notations are self-explanatory. This question paper carries a total of 35 marks. You can answer any part of any question. But the maximum that you can score is 30. Marks allotted to each question are given within parentheses.

1. Consider the multiple linear regression model as $Y = X\beta + \epsilon$, where X is stochastic. Assume that data are independent across observations. Suppose $E(\epsilon_i|X_i) \neq 0$ but there are available instruments Z with $E(\epsilon_i|Z_i) = 0$ and $V(\epsilon_i|Z_i) = \sigma_i^2$, where dim(Z)> $dim(X)$. We consider the GMM estimator $\hat{\beta}$ that minimizes
$G_N(\beta) = [\frac{1}{N}\sum Z_i(Y_i - X_i'\beta)]'W_N[\frac{1}{N}\sum Z_i(Y_i - X_i'\beta)]$.
   a. Derive the limit distribution of $\sqrt{N}(\hat{\beta} - \beta)$.
   b. If errors are homoscedastic what optimal choice W_N would you use?
   c. If errors are heteroscedastic what optimal choice W_N would you use?

[8 +6+6=20]

2. Consider the following panel data model:

$$y_{it} = \alpha_t + x_{it}'\beta + \varepsilon_{it},$$

Where the x_{it} (k× 1) are time-individual varying regressors. Let $x_t = (x_{1t}, x_{2t}, \ldots, x_{Nt})'$. Assume that $E[\varepsilon_{it}|x_t, \alpha_t] = 0$, and $E[\alpha_t|x_t] \neq 0$. $\sigma_\varepsilon^2 = Var(\varepsilon_{it})$. ε_{it} is independent across individuals and also over time.

Provide a consistent estimator of β. Prove the consistency.

[5+10=15]

INDIAN STATISTICAL INSTITUTE
MID-SEMESTRAL EXAMINATION 2015

M.STAT 2nd year. Advanced Design of Experiments

September 14, 2015, Total marks 30 Duration: $1\frac{1}{2}$ hours
Answer all questions.

Keep your answers brief and to the point. Do not give an elaborate proof when you are asked to give a brief justification for your answer.

1. a) Can there exist a Hadamard matrix of order 15? Justify your answer with a proof.

   b) Obtain the elements of GF(11). [Hint: primitive root: 2]

   c) Obtain **only the first 3 columns** of the Hadamard matrix of order 12 which can be constructed using the elements obtained in (b) above.

   d) Will an orthogonal array $OA(12, 11, 2, 2)$ exist? Give a **brief justification** for your answer.(Actual array need not be constructed)

   e) Indicate the construction of a balanced incomplete block design with $v = b = 11, r = k = 5$. (No proof needed, show any two blocks of the design only). [4+2+2+2+2 = 12]

2. Answer True or False to the following statements. Give **brief justifications** for your answers in each case.

   a) An orthogonal array $OA(24, 12, 2, 3)$ will not exist.

   b) Any $N \times k'$ subarray of an orthogonal array $OA(N, k, s, t)$ $(k > k')$, will also be an orthogonal array.

   c) An orthogonal array $OA(25, 4, 5, 2)$ may not exist.

   e) A universal optimal design for treatment effects, with 6 treatments and 6 blocks, each block being of size 6, will not exist. [$5 \times 2 = 10$]

3. a) Explain the statistical significance of the E-optimality criterion.

   b) Prove the E-optimality property of balanced incomplete block designs for estimating full sets of orthonormal treatment contrasts under the usual model and in the appropriate class. [4+4=8]

Date: 14.9.2015
Time: 2 hours

INDIAN STATISTICAL INSTITUTE

Statistical Methods in Genetics – I
M-Stat (2nd Year) 2015-2016
Mid-Semester Examination

This paper carries 30 marks.

1. Suppose, in every generation of a certain population, a fraction α practises self mating, while the remaining $(1-\alpha)$ fraction of the population practises random mating. The initial genotype frequencies at a biallelic locus in this population are D_0, H_0 and R_0. Examine whether the genotype frequencies reach equilibria. If so, what are the equilibrium values? If not, provide suitable justification. [15]

2. The genotype distribution of a random set of 100 individuals at an autosomal biallelic locus in an inbred population is as follows:

Genotype	Frequency
AA	52
Aa	33
aa	15

Obtain a 95% confidence interval for the allele frequency of *A* based on its maximum likelihood estimate. [Hint: Model the genotype counts in the general trinomial framework] [15]

Indian Statistical Institute
Midterm Examination
First Semester, 2015-2016 Academic Year
M.Stat. 2nd Year
Topics in Bayesian Inference

Date: 17 September, 2015 Maximum Marks : 35 Duration: $2\frac{1}{2}$ Hours

Answer as many questions as you can. The maximum you can score is 35.

1. Describe a Bayesian's way of evaluation of performance of a statistical procedure emphasizing how it differs from that of a frequentist. [3]

2. Let $X_1, \ldots, X_n$ be i.i.d. observations having a $N(\mu, \sigma^2)$ distribution where both μ and σ^2 are unknown. Suppose the prior on (μ, σ^2) is given by $\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$. Derive $100(1-\alpha)\%$ credible regions for μ and σ^2 based on this. Explain all your steps. [8]

3. Let $X_1, \ldots, X_n$ be i.i.d. observations with a common density $f(x|\theta)$, $\theta \in \Theta = \{\theta_1, \theta_2\}$. Consider a prior distribution (π_1, π_2) for θ.
   (a) Find the posterior distribution of θ.
   (b) Find the Bayes estimate of θ for a 0-1 loss function (loss is 0 for a correct estimate and 1 for a wrong estimate).
   (c) Suppose
   $$E_{\theta_i} \log \frac{f(X_1|\theta_i)}{f(X_1|\theta_j)} > 0 \quad \text{for } i \neq j,$$
   where $i, j \in \{1, 2\}$. Show that the posterior distribution is consistent at each θ_i. [2+4+4=10]

4. State the extra assumption used in proving the Bernstein-von Mises Theorem about posterior normality in addition to the usual assumptions for proving asymptotic normality of consistent roots of the likelihood equation. Explain mathematically the role of this assumption in the proof. [2+4=6]

5. Give an example where the Bayes estimate (with respect to squared error loss) of a parameter is not strongly consistent. Prove your answer. [4]

6. Describe the technique of Laplace approximation of an integral. Show that using this technique one can asymptotically approximate the posterior probability of the parameter (suitably centred and scaled) lying in an interval by the probability of an appropriate normal random variable lying in that interval. [3+4=7]

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination: (2015-2016)

M.STAT. – II Yr.

Pattern Recognition and Image Processing

Date: 18.09.2015 Maximum marks: 50 Duration: 2 hrs

Answer as many questions as you can. Maximum you can score is 50.

1. Let C_1 and C_2 be two classes having the pdf's p_1 and p_2 respectively where

$$p_1(x) = 1/(2+\varepsilon), \text{ if } x \in (-2, \varepsilon)$$
$$= 0, \text{ otherwise}$$

and

$$p_2(x) = 1/(2+\varepsilon), \text{ if } x \in (-\varepsilon, 2)$$
$$= 0, \text{ otherwise}$$

where $0 < \varepsilon < 1$.

Let P_i be the a priori probability of C_i, $i = 1, 2$.

a) If $P_1 = P_2 = 0.5$, find the Bayes error. Is the Bayes boundary unique? Justify your answer.

b) If $P_1 = 0.6$ and $P_2 = 0.4$, find the Bayes error. Is the Bayes boundary unique? Justify your answer. (3+3)+(3+3) = 12

2. Let f be a positive valued function defined on the interval [0,31]. Suppose the maximum value of f occurs only at 7. Consider a Genetic Algorithm where the string length of the chromosome is 5. Suppose you have only the crossover operation where the probability of crossover is $0<p_c<1$. If the initial population is {01011,00101},

(a) what are the possible populations in the next generation ?

(b) what is the probability that the maximum value of f will be attained in this generation ? 3 + 4 = 7

3. Let C_1 and C_2 be two circles with the same radius, say, 4, but with different centers that are unknown. Suppose 10 distinct points which approximately lie on the circumference of each C_i are given to you. Describe, with justifications, an algorithm to estimate the centers of the circles.

[Hints : First find the smallest rectangle containing the candidate centers.] 10

4. (a) Suppose a set S of hue values in degrees is given as S = {10,90,110,200,250, 320,350}. Draw the dendrogram for single linkage clustering for the dataset S, keeping the circular nature of the data in mind. If one wants 3 clusters, what will they be ?

(b) Suppose in a colour image there are some red or near-red patches. Describe an algorithm to change these patches to blue or near-blue, keeping the colour of the other patches in the input image unchanged. Give your own definition of "near-red" or "near-blue" providing justifications. 7 +7 = 14

5. a) What is the full form of CCD ? How is a digital image formed by it ?
   b) Name two different types of cameras each of which can form images by sensing electromagnetic radiation outside the visible range.
   c) What is the full form of TIFF? What are its advantages over JPEG image format?
   d) Define spatial and gray level resolutions of a digital image.

(1+2)+2+(1+2)+3 = 11

6. a) An image I(x, y) of size 100×100 is given. Obtain necessary expressions to compute the pixel values of the zoomed image J(p, q) of size 150×150 by bilinear interpolation method.
   b) What is the distance of a pixel (x, y) from its farthest 24-neighbour ?
   c) Write down the name and expression of a basic image transformation which can be used for image enhancement.
   d) Write an algorithm to remove noise from a given image.
   e) Write down spatial filters for extraction of horizontal, vertical and diagonal edges from an input image.
   f) What do you mean by the statement "gradient of an image is linear"?

4+1+2+2+3+1 =13

INDIAN STATISTICAL INSTITUTE

Semester exam. (Semester I: 2015-2016)

Course Name: M. Stat. 2nd year

Subject Name: Analysis of discrete data

Date: 16.11.2015, Maximum Marks: 60. Duration: 3 hrs.

Note: Answer all questions.

1. Find the sample value of the measures of association Goodman-Kruskal's τ for the following three tables.

1	3	10	6	7
2	3	10	7	6
1	6	14	12	5
0	1	9	11	3

1	6	14	12	5
0	1	9	11	3
1	3	10	6	7
2	3	10	7	6

12	1	5	6	14
11	0	3	1	9
6	1	7	3	10
7	2	6	3	10

[6]

2. Let X, Y and Z be identically distributed categorical random variables. Suppose each of X, Y and Z have n categories $C_1, \ldots, C_n$ with probabilities $p_1, \ldots, p_n$, and

$$P(X = C_i | Y = C_j, Z = C_k) = (1-2\alpha)p_i + \alpha I(i = j) + \alpha I(i = k), \quad i, j, k = 1, \ldots, n,$$

where $I(i = j)$ is the indicator and $0 \le \alpha \le 0.5$. Find $\tau_{X|Y,Z}$, the multiple association measure corresponding to Goodman and Kruskal's τ, as a function of α. [10]

3. Define kappa and generalized kappa as measures of agreement. Calculate kappa for a 4×4 table having $n_{ii} = 5$ for all i, $n_{i,i+1} = 15$, $i = 1, 2, 3$, $n_{41} = 15$, and $n_{ij} = 0$ otherwise. Explain why strong association does not imply strong agreement. Also find the value of generalized kappa by considering suitable weights and interpret the results. [4+2+4]

4. Describe Cochran-Mantel-Haenszel test for conditional independence. Carry out the test procedure for the following data of a multicenter clinical trial.

Center	Treatment	Success	Failure
1	Drug	11	25
	Control	10	27
2	Drug	16	4
	Control	22	10
3	Drug	14	5
	Control	7	12
4	Drug	2	14
	Control	1	16
5	Drug	6	11
	Control	0	12
6	Drug	1	10
	Control	0	10
7	Drug	1	4
	Control	1	8
8	Drug	4	2
	Control	6	1

[4+8]

5. Discuss the latent variable approach to model ordinal categorical variables with possible covariates. Write down the likelihood assuming normality of the latent variable. How can you model a bivariate response vector using latent variable approach where one variable is ordinal categorical and the other is continuous? Write down the likelihood in this context. [3+2+4+4]

6. Five groups of animals were exposed to a dangerous substance in varying concentration. Let n_i be the number of animals and y_i be the number that died in the ith concentration.

Concentration	n_i	y_i
1×10^{-5}	6	0
1×10^{-4}	6	1
1×10^{-3}	6	4
1×10^{-2}	6	6
1×10^{-1}	6	6

Describe the fit of an appropriate logistic model for π_i (probability of death) as a function of $\log_{10}$(concentration). [Discuss the applicability of Newton-Raphson procedure in this context.] How can you test for LD_{50} in this context? [5+4]

Indian Statistical Institute
Semestral Examination
First Semester, 2015-2016 Academic Year
M.Stat. 2nd Year
Topics in Bayesian Inference

Date: 18.11.2015 | Maximum Marks : 65 | Duration: 3 Hours

Answer as many questions as you can. The maximum you can score is 65.

1. Suppose, given $\theta \in (0,1)$, $X_1, \ldots, X_n$ are iid Bernoulli(θ) random variables. Assume that θ has a uniform prior distribution on the open interval $(0,1)$. Show that the probability of the event $\{X_{n+1} = 1\}$ given that $X_i = 1$ for all $1 \leq i \leq n$, increases to 1 as n increases to infinity. [6]

2. (a) Suppose $X_1, \ldots, X_n$ are iid with common density $f(x|\theta)$, where $\theta \in \mathcal{R}$. Suppose θ is assigned a prior distribution $\pi(\theta)$. Suppose $\theta = \theta_0$ is the true value of θ. Assuming appropriate conditions and using an appropriate version of the asymptotic normality of posteriors under such conditions, prove that $\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n) \to 0$ with probability one (under P_{θ_0}), where $\tilde{\theta}_n$ and $\hat{\theta}_n$ denote the posterior mean and the MLE, respectively.

   (b) Suppose now that $X_1, \ldots, X_n$ are iid $N(\theta, 1)$ and $\theta \sim \pi(\theta)$. Suppose $\theta = \theta_0$ is the true value of θ. Prove that under appropriate conditions to be stated by you, the difference between the posterior mean and the MLE, upon proper rescaling, converges in distribution (under P_{θ_0}) to a non-degenerate limit. [6+12=18]

3. Define the Jeffreys prior. Describe in a few sentences the justification of thinking of Jeffreys prior as a noninformative/low information prior. [6]

4. (a) Argue that use of improper non-informative priors in model selection problems may lead to Bayes factors which are defined only upto arbitrary multiplicative constants. Can the use of diffuse proper prior be a solution to this problem ? Explain your answer.

(b) Consider the general model selection problem and derive the intrinsic prior determining equations starting with Arithmetic Intrinsic Bayes Factors. Argue that in the nested model selection problem, the solution suggested by Berger and Pericchi to these equations indeed satisfies the equations. [(2+5)+(6+4)=17]

5. Consider p independent random samples, each of size n from p normal populations $N(\theta_j, \sigma^2)$, $j = 1, \ldots, p$ where σ^2 is known. Our problem is to estimate $\theta_1, \ldots \theta_p$. We assume that $\theta_1, \ldots, \theta_p$ are iid $N(\eta_1, \eta_2)$, where $\eta_1 \in R$ and $\eta_2 > 0$ are unknown constants. Find the Parametric Empirical Bayes (PEB) estimate of the vector of θ's. Explain in what sense this estimator might be more appealing than the simple vector of sample means as the estimator of the vector of θ's in this setup. [6+2=8]

6. (a) Suppose our task is to evaluate $E_f(h(X)) = \int h(x)f(x)dx$ where f is a probability density such that it is very difficult to sample directly from f. Describe how the Importance Sampling technique can be used to approximate $E_f(h(X))$. Suppose now we are given the class of importance functions $u(x)$ such that $E_u\left[h^2(X)\frac{f^2(X)}{u^2(X)}\right] < \infty$. Show that among this class, the importance function u which minimizes the variance of the importance sampling estimator (for any given value of m, the number of samples drawn for the approximation) is given by

$$u(x) = u^*(x) = \frac{|h(x)|f(x)}{\int |h(x)|f(x)dx}.$$

(b) Suppose we have a random observation X from a $N(\theta, 1)$ distribution and θ is assumed to have a Cauchy(μ, τ) distribution where both μ and τ are known constants. The problem is to approximate the posterior mean and variance of θ given X. Describe how one can approximate these quantities using the teachnique of Gibbs sampling. [(2+6)+7=15]

Indian Statistical Institute

MStat II Year | Semester Examination | 20 Nov, 2015

Advanced Design of Experiments

Answer all questions. Total score: 60 **Time: Three hours**

1. State whether each of the following statements is True or False and justify your answer clearly in each case: [$3 \times 6 = 18$]

   (a) A Hadamard matrix of order $2(s+1)$, where s is an odd prime power, always exists.

   (b) There does not exist a symmetric balanced incomplete block (BIB) design with 15 treatments and 15 blocks, each block of size 7.

   (c) An orthogonal array $OA(32, 16, 2, 3)$ exists.

   (d) If a design d is D-optimal for inferring on a parameter θ in a class $\mathcal{D}$, then it need not remain D-optimal in $\mathcal{D}$ for inferring on a non-singular transformation of θ.

   (e) A $3 \times 4 \times 2$ factorial experiment will lead to 24 independent contrasts belonging to the 3-factor interaction effect.

   (f) In a factorial setup, treatment contrasts belonging to any two distinct interactions are mutually orthogonal.

2. (a) Give an example of an experimental situation where you would recommend a row-column design.

   (b) Write down the usual model for analyzing row-column designs.

   (c) Suppose a row-column design is to be planned with t treatments, $2t$ rows and q columns, $q < t$. Then, under your model, identify a possible design which, if it exists, will be universally optimal for treatment effects in the class of all row-column designs with these parameters. Prove your claim. (You can assume the form of the C-matrix, construction of the design is not needed)

   [2+2+8=12]

3. (a) What are the necessary conditions on the design parameters for a strongly balanced crossover design to exist?

   (b) Are the above necessary conditions also sufficient? Justify your answer.

   (c) Construct a strongly balanced uniform crossover design with 3 treatments and the minimum possible number of subjects and time periods.

   (d) Construct a strongly balanced crossover design with 3 treatments, 4 periods and the minimum possible number of subjects.

   (e) Let d^* be a strongly balanced uniform design. Show that (in usual notations) $C_{d^*12} = 0$.(Assume the form of C_{d^*12} without proof.) [3 + 4 + 2 + 2+ 4=15]

P.T.O

4. (a) Consider a $s_1 \times s_2 \times \ldots \times s_m$ factorial experiment run as a connected design d. Prove that if BLUE's of every two mutually orthogonal contrast belonging to a particular main effect are uncorrelated, then that main effect is balanced.

(b) Show that if d has balance and if also the BLUE's of contrasts belonging to distinct interactions are uncorrelated, then the C-matrix must satisfy a certain algebraic form to be specified by you.

(c) What is meant by a N-run resolution (2,3) plan in the set up of a $2 \times 3 \times 4$ factorial experiment?

(d) An optimal main effects plan is required for a factorial experiment in 5 factors with each factor at 2 levels. Indicate the construction of such a plan with the smallest possible number of runs. [4+4+3+4=15]

INDIAN STATISTICAL INSTITUTE

Semester Examination : 2015-16(First Semester)

M. Stat. II year

Topology& Set Theory

Date: 24/11/15 Maximum Marks :90 Duration : 3 hrs

Note: Answer as many as you can. The maximum you can score is 90. Notation is as used in the class.

1. (a) Let A be a *proper* subset of X and B be a proper subset of Y. If X and Y are connected spaces, then show that $X \times Y - A \times B$ is connected.

   (b) Show that $\mathbb{R}$ and $\mathbb{R}^2$ are not homeomorphic. [7+4]

2. Let X be a compact space and let $\mathcal{F}$ be a family of closed subsets of X with the finite intersecting property. Show that $\bigcap_{f \in \mathcal{F}} F \neq \phi$. [5]

3. (a) Show that a path-connected space is connected.

   (b) Let $A \subseteq \mathbb{R}^2$ be a countable set. Show that $\mathbb{R}^2 - A$ is path-connected.

   [5+7]

4. (a) Show that a non-empty compact Hausdorff space with no isolated point is uncountable.

   (b) Show that a compact subspace of a Hausdorff space is closed.

   [7+7]

5. Let (X, d) be a compact metric space and $f : X \to X$ be an isometry. Show that f is a homeomorphism. [8]

6. When is a topological space said to be normal?
   Show that a regular, second countable space is normal [9]

7. Show that if X is metrizable and A, B are two disjoint closed sets in X, then there is a continuos function $f : X \to [0, 1]$ such that $f(A) = \{0\}$ and $f(B) = \{1\}$. [5]

8. Assume that the uniform metric $\overline{\rho}$ on $\mathbb{R}^{[0,1]}$ is complete. Show that $C[0,1]$, the space of real-valued continuos functions with the sup metric, is complete. [7]

9. Consider $\mathbb{R}$ with the usual topology. Show that a countable intersection of dense open sets in $\mathbb{R}$ is dense.
Hence, or otherwise, prove that $\mathbf{Q}$, the set of rationals, is not a G_δ set.
Is there a real-valued function which is continuos precisely at the rationls? If not, why not? [10+6]

10. Let R be an equilalence relation on a topological space X. Define the quotient topology on X/R
Let P denote the projection of X onto the quotient space X/R. Show that if P is open and R is closed in $X \times X$ then X/R is Hausdorff. [3+5]

INDIAN STATISTICAL INSTITUTE
First Semester Examination: 2015-2016
Course Name: MStat II
Subject Name: Topics in Mathematical Physics

Date: 23 November 2015 **Maximum Marks: 90** **Duration: 3 hrs**

Note: Answer as many questions as you can. Each question carries 15 marks.

1. (a) Derive Lagrange's equation of motion using variational principle.
   (b) Show that if Lagrangian of a system is independent of time, corresponding total energy of the system is conserved.

[9+6=15]

2. A small solid homogeneous cylinder of radius r rolls without slipping inside a stationary large cylinder of radius R.
   (a) Using Lagrangian formalism, show that for small angular displacement, the period of oscillation of the rolling cylinder is given by
   $$T = \left[\frac{6\pi^2(R-r)}{g}\right]^{1/2}.$$
   (b) Verify your results using Hamiltonian formalism as well.

[8+7=15]

3. (a) Check if the following transformation is canonical :
   $$Q = \frac{1}{p}\ ;\ P = qp^2$$
   where $\{q,\ p\}$ and $\{Q,\ P\}$ are two sets of generalized coordinates and momenta.
   (b) The Hamiltonian of a system with mass m and force constant k is given by
   $$H = \frac{p^2}{2m} + \frac{1}{2}kq^2,$$
   where $\{q,\ p\}$ are generalized coordinate and momentum. Consider a generating function $F = \mu q^2 \cot Q$ (Q is the new generalized coordinate), show that the motion of the system can be described by simple harmonic oscillation.

[7+8=15]

4. (a) Two objects A and B travel with velocities $\frac{4}{5}c$ and $\frac{3}{5}c$ respectively (with respect to a stationary observer sitting on the earth) along the same straight line in the same direction. How fast (with respect to the stationary observer) should another object C travel between them, so that it feels that both A and B are approaching C at the same speed?

P.T.O

(b) Using 4-vector notation, prove that the mass-energy relation of a relativistic point particle is

$$E^2 = p^2c^2 + m_0^2c^4,$$

where the symbols have their usual meanings.

[7+8=15]

5. (a) Prove that $\hat{x}\hat{p}_x$ and $\hat{p}_x\hat{x}$ are both non-Hermitian operators.

(b) For a one-dimensional quantum linear harmonic oscillator, the Hamiltonian operator is given by

$$\hat{H} = \frac{\hat{p}_x^2}{2m} + \frac{1}{2}m\omega^2\hat{x}^2,$$

where $\hat{x}$ and $\hat{p}_x$ are two Hermitian operators satisfying the usual commutation algebra. Using raising and lowering operators

$$a_\pm = \mp\frac{i}{\sqrt{2}}\left[\frac{\hat{p}_x}{(m\hbar w)^{1/2}} \pm i\left(\frac{mw}{\hbar}\right)^{1/2}\hat{x}\right]$$

find out the energy eigenvalue of the system.

[(3+3)+9=15]

6. (a) In Dirac equation, the Hamiltonian for Fermions (say, electrons) is given by

$$\hat{H} = c\alpha_i\hat{p}_i + \hat{\beta}m_0c^2.$$

(i) Prove that α_i and $\hat{\beta}$ are Hermitian, traceless matrices with eigenvalue $= \pm 1$.
(ii) Show that the probability density of such particles is positive definite.

(b) What are the major differences between Bose-Einstein and Fermi-Dirac statistics?

[(5+6)+4=15]

7. (a) Derive (i) Gauss law and (ii) Ampere's law from the electromagnetic field tensor

$$F_{\mu\nu} \equiv \begin{bmatrix} 0 & E_x/c & E_y/c & E_z/c \\ -E_x/c & 0 & B_z & -B_y \\ -E_y/c & -B_z & 0 & B_x \\ -E_z/c & B_y & -B_x & 0 \end{bmatrix}$$

where E_i and B_i are the components of electric and magnetic fields respectively.

(b) Show that covariant derivative of a Rank 1 tensor is non-commutative and it gives rise to a Rank 4 tensor, called Riemann curvature tensor.

[(4+6)+5=15]

Date: 23.11.2015

INDIAN STATISTICAL INSTITUTE

M.Stat Second Year, First Semester, 2015-16

Semestral Examination

Time: $3\frac{1}{2}$ Hours | Statistical Computing | Full Marks: 100

(Answer as many as you can. The maximum you can score is 100).

1. (a) Let 0.3762, 0.7198 and 0.1505 be three random numbers independently generated from the $U(0,1)$ distribution. Using these random numbers, generate a random vector (X_1, X_2) from the bivariate distribution with cumulative distribution function given by

$$F(x_1, x_2) = 1 - \exp\{-x_1\} - \exp\{-x_2\} + \exp\{-x_1 - x_2 - 0.5x_1x_2\}, \quad \text{where } x_1, x_2 \geq 0.$$

[6]

   (b) When a coin is tossed, 'Head' and 'Tail' appear with probability p and $1-p$, respectively $(0 < p < 1)$. Using this coin, is it possible to generate an unbiased estimate of $1/p$? If you think it is possible, write down the algorithm. If you think it is not possible, give reasons. [4]

2. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed as $U(0,\theta)$, where $\theta > 0$. Define $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$. Let $\{X_1^*, X_2^*, \ldots, X_n^*\}$ be a random sample of size n drawn from $\{X_1, X_2, \ldots, X_n\}$ with replacement and $\{X_1^+, X_2^+, \ldots, X_n^+\}$ be a random sample of size n drawn from $U(0, X_{(n)})$. Define $X_{(n)}^* = \max\{X_1^*, X_2^*, \ldots, X_n^*\}$ and $X_{(n)}^+ = \max\{X_1^+, X_2^+, \ldots, X_n^+\}$. Check which of the following quantities converge(s) to zero as the sample size n diverges to infinity.

   (a) $\sup_x \mid P(X_{(n)}^+ \leq x) - P(X_{(n)} \leq x) \mid$

   (b) $\sup_x \left| P\left(\frac{n(X_{(n)} - X_{(n)}^+)}{X_{(n)}} \leq x\right) - P\left(\frac{n(\theta - X_{(n)})}{\theta} \leq x\right) \right|$

   (c) $\sup_x \left| P\left(\frac{n(X_{(n)} - X_{(n)}^*)}{X_{(n)}} \leq x\right) - P\left(\frac{n(\theta - X_{(n)})}{\theta} \leq x\right) \right|$ [4+4+4]

3. Consider a data cloud of the form $\{(x_i, y_i),\ i = 1, 2, \ldots, 21\}$, where $x_i = i - 11$ and $y_i = 21x_i^2 + 11x_i + 8$ for $i = 1, 2, \ldots, 21$.

   (a) Show that the regression depth of any linear fit cannot exceed 8. Find a linear fit $y = \alpha + \beta x$, which has regression depth 8. Is this choice unique ? Justify your answer. [4+2+2]

   (b) Derive the limiting value of the Nadaraya-Watson estimate of the regression function f at $x = 4.5$, when the kernel K is Gaussian and the bandwidth h shrinks to zero. [4]

   (c) Find the limiting value of the local linear estimate of the regression function f at $x = 4.5$ when the kernel K is Gaussian and the bandwidth h diverges to infinity. [4]

P.T.O

4. Prove or disprove the following statements.

(a) LMS (least median of squares) and LTS (least trimmed squares) estimators of location are one-dimensional analogs of MVE (minimum volume ellipsoid) (based on 50% observations) and MCD (minimum covariance determinant) estimators of location, respectively. [4]

(b) Let $t_0 < t_1 < \ldots < t_k$ be a set of points in R. Consider a function f defined on $[t_0, t_k]$, which is continuous and linear in $[t_{i-1}, t_i]$ for all $i = 1, 2, \ldots, k$. Let $\mathcal{C}$ be the class of all such continuous functions. If $f_0(t) = 1$, $f_1(t) = t$ and $f_i(t) = max\{0, t - t_{i-1}\}$ for $i = 2, 3, \ldots, k$ and all $t \in [t_0, t_k]$, then $f_0, f_1, \ldots, f_k$ form a basis for $\mathcal{C}$. [5]

(c) Consider a logistic linear regression problem involving a binary valued response variable Y and a p dimensional covariate $\mathbf{X}$. Let S_i be the convex hull formed by the $\mathbf{X}$-observations with $Y = i$ $(i = 0, 1)$. If S_0 and S_1 are disjoint, the maximum likelihood estimate of the parameters of the logistic regression model will not exist. [5]

5. Give examples to show that

(a) Small reduction in impurity value is not always a good choice as a stopping rule when a regression tree is constructed. [3]

(b) For fitting a linear spline with power spline basis, placement of equi-spaced knots may not be a good option. [3]

(c) Spatial median of a continuous distribution can lie outside the support of the distribution. [3]

(d) Half-space median of a two-dimensional continuous distribution can have half-space depth smaller than 0.5. [3]

6. (a) Let $x_1, x_2, \ldots, x_n$ be n observations generated from a mixture of three univariate normal distributions $N(0,1)$, $N(1,1)$ and $N(-1,1)$ with mixing proportions p^2, $2pq$ and q^2 $(p+q = 1)$, respectively. Use the expectation-maximization (EM) algorithm (clearly state the choice for the initial value of p and the convergence criterion) to estimate p. [5]

(b) Show that the EM algorithm can be viewed as a special case of minorization-maximization algorithm. [4]

(c) Show that the iteratively re-weighted least squares method used for LAD (least absolute deviations) regression can be viewed as a majorization-minimization algorithm. [5]

7. (a) Describe the Metropolis-Hastings algorithm for generating observations from a target distribution π. Show that if $X_t \sim \pi$, all subsequent observations generated by this algorithm will follow the same distribution. [2+4]

(b) Use the Gibbs sampling algorithm to generate observations from the following distribution.

$$f(x, y, z) = C_0 \, y^{x+z}/(1+y)^{10+z}, \; x = 0, 1, 2, \ldots, 10, \; 0 < y < 1, \; z = 0, 1, 2, \ldots,$$

where C_0 is a normalizing constant. [4]

8. Computer Assignments [20]

Date: November 24, 2015
Time: 3 hours

Statistical Methods in Genetics – I
M-Stat (2nd Year)
First Semester Examination 2015-16

The paper carries 60 marks. This is an open notes examination. Answer all questions.

Question 1

1. Consider the following genotype data at a biallelic locus on 200 randomly chosen individuals in each of three populations:

Genotype	Population 1	Population 2	Population 3
AA	25	30	30
AB	95	95	100
BB	80	70	70

Do the above data provide evidence that the allele frequencies at this locus differ across the three populations? [10]

Question 2

(a) Consider a recessive disorder controlled by an autosomal biallelic locus. Suppose 100 nuclear families each comprising two unaffected parents and one offspring and 60 nuclear families each comprising exactly one parent affected and one offspring are randomly selected from the population. It was found that in the first group of families, there were 10 affected offspring while in the second group of families, there were 18 affected offspring. Obtain the maximum likelihood estimate of the prevalence of the disease based on the above data.

(b) Suppose the initial genotype frequencies at an autosomal biallelic locus are according to Hardy-Weinberg equilibrium proportions. If the fitness coefficients corresponding to the genotypes are proportional to the initial genotype frequencies, explain whether the allele frequencies at the locus reach non-trivial equilibrium values. [7 + 5]

P.T.O.

Question 3

What is the probability that an uncle and a niece have different genotypes at an autosomal biallelic locus with minor allele frequency 0.3? [8]

Question 4

Consider two autosomal biallelic loci with alleles (*D*,*d*) and (*M*,*m*) respectively. The following are the genotype data on two nuclear families:

Family 1: father is *DM*/*dm*, mother is *Ddmm*, offspring are *DDMm*, *DdMm* and *DDmm*

Family 2: both parents are *Dm*/*dM*, offspring are *DdMM*, *DDmm* and *ddMm*

Test using the LOD score approach whether the two loci are linked. [10]

Question 5

(a) Compute the expected identical-by-descent score of a pair of sibs, both heterozygous at an autosomal biallelic locus with equally frequent alleles.

(b) Consider L ordered loci such that the recombination fraction between locus i and locus j is $\theta_{i,j}$; $i \leq i,j \leq L$. Show that the recombination fraction between locus 1 and L is given by:

$$\tfrac{1}{2}\left\{1 - \prod_{i=1}^{L-1} (1-2\,\theta_{i,i+1})\right\}$$ [5+5]

Question 6

Given genotype data on a random set of individuals at two autosomal biallelic loci, explain how you would test whether there exists linkage disequilibrium between the two loci. [10]

INDIAN STATISTICAL INSTITUTE

Semestral Examination: 2015-2016

M.S. (Q.E.) II and M.Stat. II Year

Econometric Methods II/ Econometric Methods

Date: 26.11.15 Maximum Marks 100 Duration 3 hours

All notations are self-explanatory. This question paper carries a total of 110 marks. You can answer any part of any question. But the maximum that you can score is 100. Marks allotted to each question are given within parentheses.

1. Let $y_t = \mu + \varepsilon_t$, $\varepsilon_t = u_t \times (\alpha_0 + \alpha_1 \varepsilon_{t-1}^2)^{0.5}$, $u_t \sim i.i.d. N(0,1)$, $\alpha_0 > 0, \alpha_1 \geq 0$.
   a. Compute the first and second moments for y_t (i) conditional on y_{t-1}; and (ii) unconditional.
   b. Write down the likelihood function based on T number of observations on y_t.
   c. Do you feel that the ML estimator of μ will be more efficient than that of the OLS estimator? Give the intuitive reason.
   d. Describe how would you test for $H_0: \alpha_1 = 0$.
   e. How will you estimate all the model parameters by GMM? Write down the moment conditions appropriately.

   [(2+6)+7+5+7+8=35]

2. Let $y_t = \mu + \lambda\, h_t + \varepsilon_t$, $\varepsilon_t = u_t \times (\alpha_0 + \alpha_1 \varepsilon_{t-1}^2)^{0.5}$, $u_t \sim i.i.d. N(0,1)$, $\alpha_0 > 0, \alpha_1 \geq 0$.
   a. Derive the unconditional mean of y_t.
   b. Show that y_t is auto correlated. Derive the autocorrelation structure.

   [6+9=15]

3. Define temporal aggregation for an ARCH process. Show that as the sampling interval gets larger, the ARCH effect dies out. [3+12=15].

4. Consider the dynamic panel data model: $y_{it} = \alpha_i + \rho\, y_{i\,t-1} + \beta x_{it} + \varepsilon_{it\,,t=1,2,\ldots T;\, i=1,2,\ldots N}$. ε_{it} is i.i.d. with all ideal conditions. x_{it} is purely exogenous. $\alpha_i's$ are *i.i.d.* random variables with mean α and variance σ_α^2.
   a. Show that the OLS estimator of ρ is inconsistent for finite number of time series observations.
   b. Propose a GMM estimator of ρ . Write the moment conditions appropriately. Show that the proposed GMM estimator is consistent even when T is fixed.

   [9+(4+12)=25]

5. Write down the advantages of having panel data. [5]

P.T.O.

6. Consider the panel data model: $y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}$ $_{,t=1,2,....T;\ i=1,2,.....N}$. ε_{it} is i.i.d. with all ideal conditions. $\alpha_i's$ are iid random variables with mean α and variance σ_α^2. Suppose that you are not sure whether $E[\varepsilon_{it}\ |x_{it}] = 0\ or\ not.$

   a. In such situation, propose an consistent estimator of β. Prove your results analytically.

   b. How will you test for H_0: $E[\varepsilon_{it}\ |x_{it}] = 0$? Discuss the test procedure clearly.

[7+8=15]

INDIAN STATISTICAL INSTITUTE, KOLKATA
FINAL EXAMINATION: FIRST SEMESTER 2015 - '16
M.STAT II YEAR

26/11/15

Subject	:	**Functional Analysis**
Time	:	3 hours
Maximum score	:	50

Attempt all the problems. Please use a new page to answer each problem and make sure that the problem number in the margin can be read, even after stapling. If you attempt the same problem several times, please strike out all the attempts except the correct one before submitting your answer script. Justify every step in order to get full credit of your answers. All arguments should be clearly mentioned on the answer script. Points will be deducted for missing or incomplete arguments.

(1) Let $\mathcal{H}$ be a Hilbert space. Find the closure of the space of compact operators $\mathcal{K}(\mathcal{H})$ in $\mathcal{L}(\mathcal{H})$ with respect to the strong operator topology. Prove it.

[6 marks]

(2) The Volterra operator $V : L^2[0,1] \to L^2[0,1]$ is defined by

$$Vf\ (x) = \int_0^x f(y)\ dy \text{ for all } f \in L^2[0,1].$$

a) Show that $V^n f\ (x) = \int_0^x f(y) \frac{(x-y)^{n-1}}{(n-1)!} dy$ for all $n \in \mathbb{N}$.

b) Show that $\sigma(V) = \{0\}$.

c) Show that V is a compact operator.

d) Find V^*.

[5+4+2+2 = 13 marks]

(3) Consider bounded linear maps S and $T : L^2[0,\infty] \to L^2[0,\infty]$ be defined by

$$(Sf)(x) = f(x+1) \quad \text{and} \quad (Tf)(x) = f(x) + f(x+2)$$

a) Show that $\sigma(S) = \{z \in \mathbb{C} : |z| \leq 1\}$

b) Find $\sigma(T)$ (You may assume that $\sigma(p(S)) = p(\sigma(S))$ for any polynomial p).

[(5+3)=8 marks]

(4) a) Show that an idempotent operator on a Hilbert space is compact if and only if it has finite rank.

b) Show that if $T : \mathcal{H} \to \mathcal{K}$ is a compact operator and $x_n \to x$ weakly in $\mathcal{H}$, then $Tx_n \to Tx$ in $\|\cdot\|$ of $\mathcal{K}$.
Hint: If X is a Banach space and $\{x_n\}_n$ is a sequence in X which converges weakly to $x \in X$ such that every subsequence $\{x_{n_k}\}_k$ has a norm-convergent subsequence, then show that $\{x_n\}_n$ converges to x in norm of X.

[(5+5)=10 marks]

(5) For all $n \in \mathbb{N}$, define $p_n : l^1(\mathbb{N}) \to [0, \infty)$ by $p_n(f) = |f(n)|$ for all $f \in l^1(\mathbb{N})$. Show that:

a) $\{p_n : n \in \mathbb{N}\}$ is a separating family of seminorms,

b) the unit ball of $l^1(\mathbb{N})$ is compact with respect to the topology induced the family $\{p_n : n \in \mathbb{N}\}$.
Hint: A net $\{f_\alpha\}_\alpha$ in $l^1(\mathbb{N})$ converges to $f \in l^1(\mathbb{N})$ with respect to the topology induced by the above seminorms if and only if $f_\alpha(n) \underset{\alpha}{\longrightarrow} f(n)$ for all $n \in \mathbb{N}$

[(3+6)=9 marks]

(6) Let (X, Ω, μ) be a σ-finite measure space. Take $\varphi \in L^\infty(X, \Omega, \mu)$. Define $M_\varphi : L^2(X, \Omega, \mu) \to L^2(X, \Omega, \mu)$ by $(M_\varphi f)(x) = \varphi(x)f(x)\ \forall x \in X$.

a) Prove that the spectrum of M_φ is the essential range of φ, namely,
$$\{\lambda \in \mathbb{C} : \mu(\varphi^{-1}(U)) > 0 \text{ for each open set } U \text{ containing } \lambda\}$$

b) When is M_φ a projection?

c) When is M_φ a unitary operator?

[(5+2+2)=9 marks]

Indian Statistical Institute

Semester I Examination 2015-2016

M. Stat. - II Year

28 - 11 - 15

Subject: Pattern Recognition & Image Processing

Maximum Marks: 100 Duration: 3 hrs.

(Answer as many questions as you can)

1. Let the components of the vector $X = (x_1, \ldots, x_5)^T$ be binary valued (0 or 1). Let there be two classes with equal a priori probabilities, and let

$$p_{ik} = \text{Prob}(x_i = 1 \mid \text{class } k), \quad i = 1, \ldots, 5 \text{ and } k = 1, 2$$

Suppose $p_{i1} = p$ $(0.5 < p < 1)$ and $p_{i2} = 1 - p$, for $i = 1, \ldots, 5$. All the components x_i are statistically independent for both the classes.

a) Derive the Bayes classification rule for the above problem.
b) Find the Bayes error. [8 + 6 = 14]

2. (a) Define the Fisher criterion for a 2-class problem with equal a priori probabilities to obtain the Fisher discriminant function.

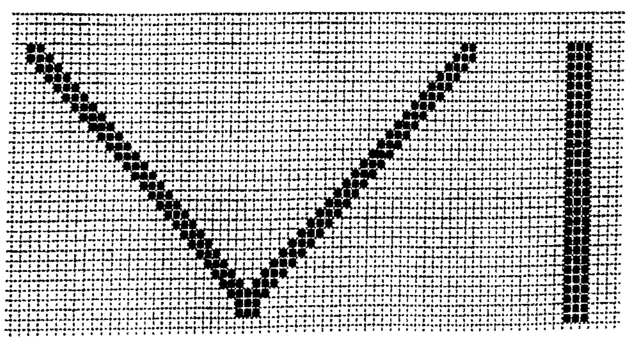
(b) Consider two classes (with equal a priori probabilities) in which the features (X,Y) have bivariate normal distributions with respective parameters as

$$\mu_1 = (0, 5)^T, \quad \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix} \quad \text{and} \quad \mu_2 = (5, 0)^T, \quad \Sigma_2 = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$$

Find the linear discriminant function that maximizes the Fisher criterion. Also find the threshold that minimizes the error. [3 + (8 + 3) = 14]

3. Provide the set of parameters that define a hidden Markov model. Describe an algorithm to find the most likely state sequence for a given sequence (of length T) of observation vectors, with computational complexity not higher than $O(Tn^2)$, where n is the number of states. [4 + 8 = 12]

4. Define a connected component in a binary image. Consider the binary image below in which the object pixels are darker. Describe a linear time (in terms of the number of pixels in the image) algorithm to assign unique labels to the connected components present in the image. [3 + 7]



P.T.O.

5. (a) Consider a two-class classification problem where the classes are known to be linearly separable. Describe a gradient descent algorithm based on the perceptron criterion function to determine an optimal linear discriminant. Prove that the algorithm converges after a finite number of iterations.

(b) Show the architecture, with appropriate weights, of a multilayer perceptron for the XOR problem. Are the weights unique? Justify your answer.

[(6 + 7) + 7 = 20]

6. (a) Define the two most basic operations of mathematical morphology.
(b) Are these two basic operations inverses of each other?
(c) State two properties each of these two morphological operations.
(d) Give examples of a binary image and a structuring element such that one of these two operations is required to be applied on the image exactly two times to remove all object pixels from the image. [3 + 1 + 2 + 4 = 10]

7. (a) Write down two non-linear transformations for gray-level image, one each for (i) stretching darker regions while suppressing brighter regions and (ii) suppressing darker regions while stretching brighter regions. Can the power law be used for both these transformations? Justify your answer.
(b) Write down the 8-neighbours for the pixel (0,0) situated at the top-left corner.
(c) Design a 3×3 filter or mask for image smoothing such that its center pixel has the maximum weight while the weights at other pixels vary inversely with their distance from the center pixel. Does this filter have any advantage over the simple averaging filter? Justify your answer. [(2+2+2) + 1 + 3 = 10]

8. (a) Describe Otsu's thresholding method in brief.
(b) Does it require modeling of probability density functions of object and background pixels?
(c) How does Otsu's thresholding method differ from the K-means algorithm?
(d) Which discriminant criterion is maximized or minimized by the Otsu's method?
(e) What are the drawbacks of Otsu's method? [2 + 1 + 2 + 2 + 3 = 10]

9. (a) What is skeletonization of a binary image?
(b) Describe a method of skeletonization.
(c) What are the possible defects of the skeleton obtained by the algorithm you described?
(d) How may these defects be corrected? [2 + 7 + 2 + 3 = 14]

INDIAN STATISTICAL INSTITUTE

M.Stat Second Year, First Semester, 2015-16

Semestral (Backpaper) Examination

Time: 3 Hours | Statistical Computing | Full Marks: 100

1. (a) Let $\mathbf{x} = (0.102,\ 3.456)'$ be an observation generated from a bivariate normal distribution with the location parameter $(1,2)'$ and the scatter matrix $\begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$. Without using any additional random number, generate an observation from the uniform distribution on the ellipse $E = \{(x,y) :\ 5x^2 - 8xy + 5y^2 \leq 9\}$. [8]

   (b) Let $X_1, X_2, \ldots, X_n$ be independent observations from a normal distribution with the location parameter μ and the scale parameter σ^2. If the prior distribution of (μ, σ^2) is given by $\pi(\mu, \sigma^2) \propto 1/\sigma^2$, give an algorithm for generating observations from the posterior distribution of (μ, σ^2). Using these observations, how will you construct a 95% credible region (you may consider a rectangular region) for (μ, σ^2). [5+3]

2. Consider a moving average process $X_t = \mu + \varepsilon_t + \frac{1}{2}\varepsilon_{t-1}$ for $t = 1, 2, \ldots$, where the ε_ts are i.i.d. with the mean 0 and the variance 1.

   (a) Check whether $\bar{X}_n = \sum_{i=1}^n X_i/n$ is a consistent estimator of μ. [4]

   (b) Find the asymptotic distribution of $Y_n = \sqrt{n}(\bar{X}_n - \mu)$. [4]

   (c) Consider a bootstrap sample $X_1^*, \ldots, X_n^*$ drawn from $\{X_1, \ldots, X_n\}$ with replacement. Define $\bar{X}_n^* = \sum_{i=1}^n X_i^*/n$ and $Y_n^* = \sqrt{n}(\bar{X}_n^* - \bar{X}_n)$. Check whether the asymptotic distribution of Y_n^* matches with that of Y_n. [6]

3. (a) Show that in the case of a univariate continuous distribution, half-space median, simplicial median and spatial median all coincide with the usual univariate median. [6]

   (b) If F is a spherically symmetric multivariate distribution with the centre at the origin, show that $D(x, F)$, spatial depth of $\mathbf{x}$ with respect to F, is a decreasing function of $\|\mathbf{x}\|$. [8]

4. Consider the following data set.

x	1	2	3	4	5	6	7	8	9	10
y	18.6	15.7	10.8	6.9	3.5	-0.9	-5.6	-8.7	-13.8	-16.5

   (a) Find the equation of the line $y = \alpha + \beta x$ that passes through the point $(x_3, y_3)'$ and minimizes $\sum_{i=1}^{10} |y_i - \alpha - \beta x_i|$, where $(x_i, y_i)'$ denotes the i-th $(i = 1, 2, \ldots, 10)$ observation. Compute the regression depth of this line [4+2]

   (b) What will be the equation of the line in (a) if the observation $(2, 15.7)'$ is replaced by (i) $(32, 15.7)'$ (ii) $(2, -15.7)'$. Comment on the robustness of LAD (least absolute deviations) regression based on your findings. [4]

   (c) Find the LMS (least median of squares) and LTS (least trimmed squares) (based on 50% observations) regression line of Y on X. [2+2]

P.T.O

5. Let $x_1, x_2, \ldots, x_n$ be n observations generated from a mixture of three univariate normal distributions $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ and $N(\mu_3, \sigma_3^2)$ ($\mu_1 < \mu_2 < \mu_3$) with mixing proportions p^2, $2pq$ and q^2 ($p + q = 1$), respectively.

   (a) Use an EM algorithm (clearly state the choice for the initial values of the model parameters and the convergence criterion) to estimate $p, \mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2$ and σ_3^3. [8]

   (b) How will you modify your algorithm if it is known that $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$? [4]

   (c) Will the model parameters be identifiable if the condition $\mu_1 < \mu_2 < \mu_3$ is not given? Justify your answer. [2]

6. (a) Consider a data set $\{(x_i, y_i);\ i = 1, \ldots, n\}$, where $y_i = \psi(x_i)$ for $= 1, \ldots, n$ and ψ is monotonically increasing. For any fixed choice of the bandwidth h, check whether the Nadaraya-Watson estimate $\hat{f}_h$ based on Gaussian kernel is also monotonically increasing. [5]

   (b) Consider a classification problem with J competing classes. If a node t of a classification tree contains p_j proportion observations from the j-th class ($j = 1, 2, \ldots, J$), its impurity function is given by $i(t) = \psi(p_1, p_2, \ldots, p_J)$, where ψ is concave and symmetric in its arguments. Show that (i) ψ is maximized when $p_j = 1/J$ for $j = 1, 2, \ldots, J$, (ii) it is minimized when $\max p_j = 1$ and (iii) for any split of a node t, the reduction in the impurity function, $\Delta i(t)$, is non-negative. [3+3+3]

7. (a) Show that the iteratively re-weighted least squares method used for quantile regression can be viewed as a majorization-minimization algorithm. [6]

   (b) The results of one-day international series played in 2014 are given below.

Home Team	AS	SR	IN	WI	NZ	AS	EN	SR	SA	EN	PK	WI	NZ	SA
Visiting Team	NZ	SA	EN	PK	SA	PK	WI	IN	EN	AS	SR	IN	PK	AS
Result	4-1	2-1	3-0	2-2	2-3	3-0	2-1	3-2	4-1	3-3	2-1	2-5	4-1	4-3

   Describe how you will rank these 8 teams based on their performance in 2014 using an appropriate Bradley-Terry model. [8]

Indian Statistical Institute

Semester I Backpaper Examination 2015-2016

M. Stat. - II year 11/02/2016

Subject: Pattern Recognition & Image Processing

Maximum Marks: 100 Duration: 3 hrs.

(Answer as many questions as you can)

1. (a) Derive the update rules to iteratively estimate the parameters (P, μ_1, σ_1^2, μ_2, σ_2^2) of the following mixture distribution.

$$p(x) = P\, p_1(x) + (1-P)\, p_2(x),\ 0 < P < 1$$

where $p_k(x)$ denotes the univariate normal density with mean μ_k and variance σ_k^2, k=1,2.

(b) Describe how the above estimation procedure can be used in segmenting a grey level image into object and background. [10 + 4 = 14]

2. Consider a two-class classification problem in which the univariate feature X has class-conditional probability density function $p_k(x)$ in the k-th class, k=1,2. The density functions $p_k(x)$ are given below.

$$\begin{aligned} p_1(x) &= 1 \ \text{ if } 0 < x < 0.5 \\ &= 1.5 - x \ \text{ if } 0.5 \le x < 1.5 \\ &= 0 \ \text{ otherwise} \end{aligned}$$

$$p_2(x) = p_1(0.5 - x) \text{ for all } x.$$

(a) Find the Bayes error.

(b) Is the Bayes classifier unique? Justify your answer. [8 + 6 = 14]

3. In genetic algorithms, describe with illustrations the operations that are needed to obtain the population at generation (t + 1) from the population at generation t. [10]

4. (a) Describe an algorithm to rotate a binary image of size $n \times n$ around the center of the image by an arbitrary angle θ in the clockwise direction.

(b) Describe how the Hough transform is used to detect straight line segments in a binary image. [7 + 7 = 14]

5. (a) Write down the expressions for conversion of color image pixels from RGB space to HSV space.

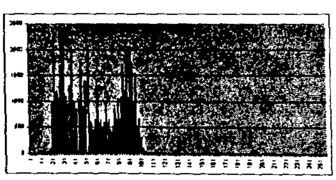
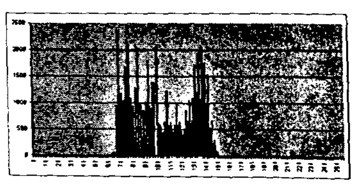
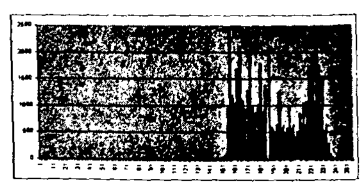
(b) Describe the K-means algorithm for colour image segmentation in HSV space. [7 + 8 = 15]

P.T.O

6. (a) Consider a binary image **B** of size 128 × 128 with object pixels denoted by 1 and background pixels denoted by 0. Consider a grid element **S** of size 3 × 3 each position of which contains 1. Let the top-left corner of **S** be its origin or reference point. Suppose **S** is translated over **B** by placing its origin (reference point) over each pixel of the image **B**. If **S** is wholly contained in the object part of **B**, then the corresponding pixel of an output image **B'** (initialized by 0 at each pixel) of the same size as **B**, is converted to 1. What is this operation called in the context of mathematical morphology? What is its dual operation? Define it.
(b) Explain why opening and closing operations can be considered as complementary to each other.
(c) Are the two operations in (b) idempotent?
(d) Explain one use of any one of the two operations mentioned in (b).
[(2+1+2)+2+1+2 = 10]

7. (a) What is the relationship between skeletonization and thinning of binary image?
(b) Describe an algorithm for thinning an image.
(c) Show a structuring element and describe an algorithm for obtaining the boundary of an object using the structuring element. [2 + 7 + (1 + 5) = 15]

8. (a) Describe the basic limitation of histogram-based methods of image processing.
(b) Gray level histograms of an image and its two transforms are shown below. What characteristic of the image is actually modified?

(c) Write down two masks, one each for horizontal and vertical step detection respectively in an image.
(d) What is the border effect of convolution of an image by a mask of size 3 × 3 ? How does one solve this problem ? [3 + 1 + 3 + 3 = 10]

9. State three discriminant criteria to determine the goodness of image thresholding. Are they equivalent to each other? Explain. Which one among these three is the simplest and why?
[6 + 2 + 2 = 10]

INDIAN STATISTICAL INSTITUTE

Semester exam. (Semester I: 2015-2016) [Back paper]

Course Name: M. Stat. 2nd year

Subject Name: Analysis of discrete data

Date: 11/02/2016 Maximum Marks: 100. Duration: 3 hrs.

Note: Answer all questions.

1. Define Theil's entropy measure of association for nominal responses. Show that it reduces to the form
$$-\frac{\sum_i \sum_j \pi_{ij} \log(\pi_{ij}/\pi_{i+}\pi_{+j})}{\sum_j \pi_{+j} \log \pi_{+j}},$$
where π_{ij} is the cell probability of the (i,j)th cell and $\pi_{i+} = \sum_j \pi_{ij}$, $\pi_{+j} = \sum_i \pi_{ij}$. [4+5]

2. Geometrically interpret *Risk Ratio* $= 1$ for 2×2 contingency tables. [10]

3. Discuss Fisher scoring for a regression using general link function. Discuss the special case of logistic regression. Show that both Newton-Raphson and Fisher's scoring methods are same for a logistic model for estimating the parameters. [5+3+3]

4. Data on pre- and post-operative conditions (classified as bad, moderate, good) of 100 patients are given along with their age, sex and another important covariate related to the initial condition of the disease. Give a latent variable based model and illustrate an approach to test whether there is significant improvement in the operation or not. Discuss any computational problem that might be encountered in the analysis. [6+7+4]

5. Suppose a fixed number of 20 patients are each treated by the three treatments A, B and C, and the number of successes by the treatments are respectively 8, 6 and 5. Let θ_{AB} be the odds ratio of treatment A relative to treatment B. Find the asymptotic joint distribution of $(\log \theta_{AB}, \log \theta_{AC})'$. Suggest an intuitive test for $H_0 : \theta_{AB} = 1, \theta_{AC} = 1$ against the alternative $H_1 : \theta_{AB} \geq 1, \theta_{AC} \geq 1$ with at least one strict inequality. [12+3]

6. Derive Fisher's exact test statistic for testing H_0 : independence, for the 2 × 2 table. Carry out Fisher's exact test against both one-sided and two-sided alternatives.

3	1
1	3

[4+4+4

7. (a) Describe an iterative proportional fitting for loglinear model. Compare the procedure with the Newton-Raphson method. [7+

(b) Consider the 2 × 2 × 2 three-way table with cell frequencies $\{n_{ijk}\}$, where $n_{111} = 12$, $n_{112} = 15$, $n_{121} = 25$, $n_{122} = 22$, $n_{211} = 12$, $n_{212} = 10$, $n_{221} = 1$[illegible] $n_{222} = 20$. Illustrate the iterative proportional fitting for the models (XY, YZ) and (XY, XZ, YZ). [6+

Indian Statistical Institute
Backpaper Examination
First Semester, 2015-2016 Academic Year
M.Stat. 2nd Year
Topics in Bayesian Inference

11/02/2016

Total Marks : 100

Duration: 3 Hours

Answer all questions

1. Give an example of an inference problem where classical inference gives a paradoxical answer while Bayesian inference gives reasonable answer. [5]

2. Let X be a random variable taking values $0, 1, 2, \cdots$ with p.m.f. involving an unknown real parameter $\theta > 0$. Find a model for X and a prior for θ such that the marginal density of X is negative binomial. [5]

3. (a) Define a highest posterior density (HPD) credible region for an unknown parameter. [2]

   (b) Suppose $Y_1, \ldots, Y_{n_1}$ is a random sample of size n_1 from a normal population $N(\theta_1, \sigma^2)$ whereas $Y_1, \cdots, Y_{n_2}$ is an independent random sample of size n_2 from another normal population $N(\theta_2, \sigma^2)$. Here $(\theta_1, \theta_2, \sigma^2)$ are unknown parameters. Assuming the prior $\pi(\theta_1, \theta_2, \sigma^2) \propto \frac{1}{\sigma^2}$, derive the posterior distribution of $\eta = \theta_1 - \theta_2$ and hence find a $100(1-\alpha)\%$ HPD credible set for η. [15]

4. Stating appropriate assumptions, prove the Bernstein-von Mises Theorem about posterior normality. [25]

5. Stating appropriate assumptions, prove that posterior consistency holds for all parameter values when the parameter space is finite. [15]

6. Describe the technique of Laplace approximation of an integral. Can this simplify Bayesian hypothesis testing in any way ? Explain your answer. [3+1+2=6]

7. Suppose we observe $\mathbf{X} = (X_1, \ldots, X_n)$. Under the model M_0, X_i's are iid $N(0,1)$ and under the model M_1, X_i's are iid $N(\theta, 1)$ $\theta \in \mathcal{R}$. Starting with the noninformative prior $\pi(\theta) = 1$ under M_1, argue using direct asymptotic approximations of Arithmetic Intrinsic Bayes Factor (AIBF) that a $N(0,2)$ density can be taken as an intrinsic prior for this problem. [10]

8. Consider p independent random samples, each of size n from p normal populations $N(\theta_j, \sigma^2)$, $j = 1, \ldots, p$ where σ^2 is known. Our problem is to estimate $\theta_1, \ldots \theta_p$. We assume that $\theta_1, \ldots, \theta_p$ are iid $N(\eta_1, \eta_2)$, where $\eta_1, \eta_2)$ has a prior distribution $\pi(\eta_1, \eta_2)$.

   (a) Explain the Hierarchical Bayes (HB) approach in this problem and describe how the HB estimates "borrow strength" from the whole data at hand. [13]

   (b) Explain why a James-Stein type shrinkage estimator might be preferable to the usual vector of sample means as an estimator of $(\theta_1, \ldots, \theta_p)$ where p is large. [4]

INDIAN STATISTICAL INSTITUTE

First Semestral Back Paper Examination: 2015-16

M.STAT 2ndyear. Advanced Design of Experiments. 15/02/16

Total marks 100. Duration: Three hours

Answer all questions.

1. (a) Define the A-optimality criterion and explain its statistical significance.
   (b) Prove that a Randomized block design with parameters v, b, k will be A-optimal for treatment effects in the class of all block designs with parameters v, b, k.
   (c) Prove that a Latin square design with t treatments is universally optimal in the class of all row-column designs with t treatments and t rows and t columns. [(2+4)+6+8=20]

2. (a) Show that a Hadamard matrix of order 18 cannot exist.
   (b) Construct a Hadamard matrix of order 12 after clearly stating the result you use to construct this. (Proof of the result is not required).
   (c) Prove that the existence of a Hadamard matrix of order $4t$ implies the existence of a symmetric BIB design (the parameters of the BIB design are to be determined by you).
   (d) Construct the blocks of the symmetric BIB design from the Hadamard matrix constructed by you in (a) above. [5+5+5+5=20]

3. a) Define a balanced uniform crossover design. What are the necessary conditions that its parameters must satisfy in order that such a design may exist?

   b) Construct a balanced uniform design with 5 treatments, 10 subjects and 5 periods.

   c) From the design in b) above construct a strongly balanced design in 5 treatments and 10 subjects.

   d) Give a model for analyzing data from an experiment with a crossover design.

   e) Under the model in (d) above, show that the design in (b) has a completely symmetric information matrix for direct effects. (You do not have to derive the form of the information matrix, you can assume its form). [(2+3)+5+5+5=20]

4. (a) Give an example of an experimental situation where you would use a factorial experiment.
   (b) What are the advantages of using a factorial experiment over varietal experiments?

P.T.O

(c) For a $3 \times 4 \times 2$ experiment with factors F_1, F_2 and F_3, obtain a full set of orthonormal contrasts for the interaction F_2F_3. Justify your answer.

(d) When is the above factorial design said to be balanced?

(e) Show that the design is balanced for F_2F_3 only if every two mutually orthogonal contrasts belonging to F_2F_3 are uncorrelated. [$5 \times 4 = 20$]

5. Consider a fraction d of a 2^7 factorial consisting of the eight treatment combinations

$$0000000, 1000000, 0100000, 0010000, 0001000, 0000100, 0000010, 0000001.$$

Let $Y(i_1 \ldots i_7)$ be the observation arising from a typical treatment combination $i_1 \ldots i_7$ in the fraction. Suppose that all interactions are absent and consider the linear model

$$E\{Y(i_1 \ldots i_7)\} = \beta_0 + \sum_{i=1}^{7}(2i_j - 1)\beta_j,$$

where β_0 is the general mean and β_j represents the main effect of the jth factor. As usual, assume that the errors are uncorrelated and homoscedastic.

Write

$$\beta = (\beta_0, \beta_1, \ldots, \beta_7)' \text{ and } Y = (Y(0000000), Y(1000000), \ldots, Y(0000001))'$$

for the parametric vector, and the 8×1 vector of observations arising from d, respectively.

(a) Write down the 8×8 design matrix X, where $E(Y) = X\beta$.

(b) Show that X is nonsingular.

(c) Hence conclude that the BLUE of β, i.e. $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_7)'$ is given by

$$\hat{\beta} = X^{-1}Y.$$

(d) Use the result in (c) to obtain $\hat{\beta}_1$ explicitly as a linear function of the elements of Y. [Hint: Solving a system of linear equations may be easier than finding X^{-1} explicitly.]

(e) What is the variance of $\hat{\beta}_1$? What can you say about the variances of $\hat{\beta}_2, \ldots, \hat{\beta}_7$ without finding these in detail?

(f) On the basis of your findings in (e), will you recommend the use of the fraction d ? Or, can you suggest some other design which will perform even better? Justify your answer.

[3+3+3+3+3+5=20]

Back Paper

INDIAN STATISTICAL INSTITUTE

Semestral Examination: 2015-2016

M.S. (Q.E.) II and M.Stat. II Year

Econometric Methods II/ Econometric Methods

Date: 17/02/16 Maximum Marks 100 Duration 3 hours

All notations are self-explanatory. Marks allotted to each question are given within parentheses.

1. Consider the multiple linear regression model $Y = X\beta + \epsilon$, where X is stochastic. Assume that data are independent across observations. Suppose $E(\epsilon_i|X_i) \neq 0$ but there are available instruments Z with $E(\epsilon_i|Z_i) = 0$ and $V(\epsilon_i|Z_i) = \sigma_i^2$, where $dim(Z) > dim(X)$. We consider the GMM estimator $\hat{\beta}$ that minimizes $G_N(\beta) = [\frac{1}{N}\sum Z_i(Y_i - X_i'\beta)]' W_N [\frac{1}{N}\sum Z_i(Y_i - X_i'\beta)]$, where is an appropriate weight matrix.
   a. Suggest an Instrumental Variable estimator for β using the entire vector of instruments.
   b. Show that the estimator suggested in (a) can be viewed as a GMM estimator as defined above.

   [10+10=20]

2. Consider the dynamic panel data model: $y_{it} = \alpha_i + \rho\, y_{i\,t-1} + \beta x_{it} + \varepsilon_{it\,,t=1,2,\dots,T;\, i=1,2,\dots,N}$. ε_{it} is i.i.d. with all ideal conditions. x_{it} is purely exogenous. $\alpha_i's$ are iid random variables with mean α and variance σ_α^2.
   a. Show that the OLS estimator of ρ is inconsistent for finite number of time series observations.
   b. Propose a GMM estimator of ρ. Write the moment conditions appropriately. Show that the proposed GMM estimator is consistent even when T is fixed.

   [9+(4+12)=25]

3. Consider the following panel data model:

$$y_{it} = \alpha_t + x_{it}'\beta + \varepsilon_{it},$$

where the x_{it} (k× 1) are time-individual varying regressors. Let $x_t = (x_{1t}, x_{2t}, \dots, x_{Nt})'$. Assume that $E[\varepsilon_{it}|x_t,\ \alpha_t] = 0$, and $E[\alpha_t|x_t] \neq 0$. $\sigma_\varepsilon^2 = Var(\varepsilon_{it})$.

   a. Provide a consistent estimator of β. Prove the consistency.
   b. How will you estimate α_t consistently? Prove the consistency.
   c. Discuss how will you test for the assumption $E[\alpha_t|x_t] = 0$.

   [10+10+10=30]

P.T.O

4. Let $y_t = \mu + \beta x_t + \varepsilon_t$, $\varepsilon_t = u_t \times (\alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2)^{0.5}$, $u_t \sim i.i.d\ N(0,1)$, x_t is a single regressor, $\alpha_0 > 0, \alpha_1 \geq 0, \alpha_2 \geq 0$.
   a. Find the autocorrelation function of ε_t and ε_t^2.
   b. Show that the OLS estimator of β is consistent.
   c. Propose a better estimator of β than that of the OLS estimator. Give your logic.
   d. How will you test for $H_0: \alpha_i = 0$? $\forall i = 1, 2$.

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

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination : Semester II (2015-2016)

Course Name : BSDA (M. Stat. 2nd year)

Subject Name : Statistical Methods in Biomedical Research

Date : 22 February, 2016, Maximum Marks : 40. Duration : 2 hrs.

1. Describe Efron's biased coin design (BCD) for allocation with two treatments, say A and B. [4]

   What happens if we take $p = 1$ in BCD? [3]

   Derive the expected proportion of allocation by A using BCD with $p = 2/3$. [4]

2. Discuss the concept of type I error spending functions in group sequential analysis. Discuss how type I error spending functions can be constructed by accumulating boundary crossing probabilities. If there are only two groups, give one form of type I error spending function which will spend the total type I error in 9:16 way in the two groups (assuming the first group is observed after 60% time elapses). [4+8+4]

3. (a) Coeliac disease is a condition that impairs the ability of the gut to absorb nutrients. A useful measure of nutritional status is the bicep's skinfold thickness, which has standard deviation 2.3 mm in this population. A new nutritional programme is proposed and is to be compared with the present programme. If two groups of equal size are compared at the 5% significance level, how large should each group be if there is to be 90% power to detect a change in mean skinfold of 0.5 mm? How many would I need if the power were 80%? [2]

   (b) Suppose I can recruit 300 patients, what difference can I detect with 80% power? [2]

   (c) Suppose I decide that a change of 1 mm in mean skinfold is of interest after all. How many patients do I need for a power of 80%? [2]

   (d) What would be the effect on this value if 2.3 mm underestimates σ by 20% and if it overestimates σ by 20%? [4]

   (e) Assuming that 2.3 mm is a satisfactory estimate of σ, what sample sizes would we need to achieve 80% power to detect a mean difference of 1 mm if we opted to allocate patients to the new and the control treatments in the ratio 2:1? [3]

Indian Statistical Institute
Second Midsemestral 1 Examination 2015-16
M. Stat. II yr

Statistical Inference II

Date: February 22 , 2016 Maximum marks: 60 Duration: 2 hrs.

Answer all Questions. *Paper carries 65 points.*

1 (a) Cosider a finite population consisting of N units and let $\mathcal{E}$ and $\mathcal{F}$ denote SRSWR and SRSWOR sampling experiments of size k respectively. Compute defficiencies of $[\mathcal{E}||\mathcal{F}]$ and $[\mathcal{F}||\mathcal{E}]$ (by explicitly computing necessary Markov kernels) respectively when $k = 3$.

(b) Find out the right kernel to transfer computation of expectations between the following experiments : (a) $X_1, X_2, \ldots, X_n$ are iid Ber(p) ($n > 1$) and, (b) $\tau = \min\{k \geq r : \sum_1^k Y_i \geq r\}$ where $r > 1$ is given integer and, $Y_1, Y_2, \ldots$ is an infinite Bernoulli sequnce with success probability p.

[(8+8) + 10 = 26]

2. Consider the standard Borel space $(\mathbb{R}, \mathcal{B})$.
Let $\mathcal{S} = \{A \subseteq \mathbb{R} : A$ countable or A^c countable $\}$ denote the standard *countable-cocountable* σ-algebra on $\mathbb{R}$ (include the null set in countable as convention).
(a) Describe the set of measurable functions on $(\mathbb{R}, \mathcal{S})$.
(b) Suppose X is a random variable with N$(\theta, 1)$ distribution on $(\mathbb{R}, \mathcal{B})$ ($\theta \in \mathbb{R}$ is the location). Is $\mathcal{S}$ sufficient for θ ? Justify.
(c) Suppose X is Poi(θ) defined on $(\mathbb{R}, \mathcal{B})$. Is $\mathcal{S}$ sufficient for θ ? Justify.

[5+5+5 = 15]

3. A statistic T is pairwise sufficient for a model $\{P_\theta : \theta \in \Theta\}$ if it is sufficient for every two point submodel $\{P_{\theta_0}, P_{\theta_1}\}$, $\theta_0, \theta_1 \in \Theta$. Show that
(a) If Θ is countable and T is pairwise sufficient then T is sufficient for $\{P_\theta : \theta \in \Theta\}$.
(b) If $\{P_\theta : \theta \in \Theta\}$ is a dominated family by some σ-finite measure and T is pairwise sufficient then T is sufficient for $\{P_\theta : \theta \in \Theta\}$.

[12+12] = 24

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination : Semester II (2015-16)

M. Stat. II Year

Survival Analysis

Date: 24.2.2016   Maximum marks: 60   Time: 135 minutes

(Total mark is 65. Calculator can be used.)

1. (a) Derive the hazard rate for log-normal life distribution.

   (b) A study that began in 1990 included only disease-free subjects. A subject, who was disease-free and 18 years old in 1990, was included in the study and found to have contracted the disease 12 years later. As time is measured in years, write down the likelihood contribution from this subject with $f(\cdot)$ denoting the density for disease onset time.

   (c) In a censoring scheme, any surviving subject at a prefixed time t_1 is censored with probability 0.5 and all the surviving subjects at a prefixed time t_2 ($> t_1$) are censored. Prove that it is a special case of random censoring.

   (d) Consider random right censored data from $Weibull(\lambda, p)$ life distribution with known $p > 0$. Obtain the maximum likelihood estimate of λ.

   (e) In the modification of Gehan's test, suggested by Efron, find the u_{ij} score for the case $x_{1i} < x_{2j}$, $\delta_{1i} = 0$, $\delta_{2j} = 0$.

   [3+3+3+3+3=15]

2. (a) Suppose the hazard rate $\lambda(t)$ is given by λ_i, if $t \in I_i$, where $I_1, \cdots, I_k$ form a partition of $[0, \infty)$. Derive $S(t)$ and $f(t)$.

   (b) Consider T following $Geometric(p)$ distribution with p.m.f. $p_t = p(1-p)^t$, $t = 0, 1, 2, \cdots$. Prove its memoryless property. Based on random right censored data from this life distribution, obtain the maximum likelihood estimate of p.

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

3. Consider right censored life time data from the model which has constant hazard α up to a given time t_0 and then an increased hazard $\alpha + \delta$ after time t_0 with $\delta \geq 0$. Describe the score test for $\mathcal{H}_0 : \delta = 0$ against the one-sided alternative. Give details.

   [12]

4. (a) Describe the nonparametric maximum likelihood estimate (NPMLE) of the arbitrary survival function $S(t)$ based on random right censored data giving all details (Variance calculation is not necessary). Mention one of its uses.

   (b) Consider the following data on the times of remission (in weeks) of leukemia patients (+ indicating right censored observation).
   6+, 6, 6, 6, 7, 9+, 10+, 10, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+.
   Obtain the Kaplan-Meier survival curve giving details of the calculations. Also obtain the Nelson-Aalen estimate of the cumulative hazard curve. Plot both the

Kaplan-Meier survival curve and the survival estimate obtained from the Nelson-Aalen estimate in the same graph (rough plot will do) with some comparative comments.

[(7+1)+(7+3+3)=21]

5. Describe the log-rank test for comparing two survival distributions based on random right censored data from each population. Mention one of its limitations. [6+1=7]

INDIAN STATISTICAL INSTITUTE

Mid-Semester of 2nd Semester Examination : 2015-16

Course Name : M.Stat. 2nd Year

Subject Name : Advanced Sample Surveys

Date : Feb 26, 2016

Duration : 3 hrs

Note: Use separate answer sheets for two groups.

Group B (Total Marks = 25)

Answer all questions.

1. Describe how an optional randomized response technique can be used to estimate a sensitive population proportion by a general sampling design with $p(s)$ being the selection probability of a sample s of respondents. Also give a variance estimator.

(5)

2. Describe the notion of protection of privacy in randomized response literature. For Warner's model and with SRSWR of respondents, find the condition ensuring maximum protection of privacy and conclude regarding its behaviour with efficiency in estimation.

(5)

3. Describe how the population correlation coefficient can be estimated by a sample obtained by a general sampling scheme allowing positive first and second order inclusion probabilities. Describe how you can estimate the measure of error by Linearization Technique.

$(5 + 5 = 10)$

4. Describe the method of Jack-knifing to obtain an improved estimator along with a measure of its error starting with an approximately unbiased estimator for a population parameter whatever it is linear or non-linear.

(5)

INDIAN STATISTICAL INSTITUTE
Mid Semestral Examination: 2015-16

M. Stat. II Year
Advanced Sample Survey

Date: 26/02/2016 **Maximum Marks: 25** **Duration: 3 Hours**

Answer any 2 questions each carrying 10 marks.

Assignment records to be submitted on the date of exam. Carry 5 marks

Time allowed : 3 hours for Group A together with Group B.

Answers to Group A and Group B must be given in separate answer books.

Group – A

1 Given raw survey data, derive with full proof the minimal sufficient statistics. Illustrate giving necessary details usefulness if any of such statistics.

2 Give an account of how the Rao, Hartley and Cochran's strategy may be employed showing how it fares visa-a-vis the strategy of Hansen and Hurwitz.

3 Prove that Horvitz and Thompson's estimator is admissible among all unbiased estimators for a finite population total.

*****

Statistical Methods in Public Health
Mid-Semestral Examination
M.Stat. II Year, 2015-2016
Total Marks - 70
Time - 2 hrs. 30 min.

Indian Statistical Institute
Kolkata 700 108, INDIA

Attempt all questions:

Group A

1. A prey population follows the law of logistic growth, the interaction among the predator population follows Holling type-II functional response and the growth of the predator population declines due to natural mortality.

(i) Based on the above assumptions, write down the simple predator-prey model.

(ii) Find out the condition(s) for which the system is uniformly bounded.

(iii) Under what condition(s) the system is said to be permanent? Interpret the results from ecological point of view.

(iv) In which condition(s), the above model can be represented as a simple epidemic model?

[2 + 3 + 8 + 2 = 15]

2. (i) Stating clearly the basic assumptions and write down Kermack-McKendrick model. In which condition(s), the model can be represented as a simple logistic growth model and hence find out the solution of the model with proper initial conditions.

(ii) From the solution, find out the conditions for which the disease will be epidemic or endemic.

(iii) A new strain of influenza is introduced into a town with 1200 inhabitants by two visitors. Assume that the average infective is in contact with 0.4 inhabitants per day and that the average duration of the infective period is 6 days. Will the infection die out or will the flu persist?

[8 + 3 + 4 =15]

Group B

3. Let us consider the following Blumberg's growth equation

$$\frac{dx(t)}{dt} = rx(t)^{\gamma}\left[1 - \left(\frac{x(t)}{K}\right)\right]^{\theta}, \qquad (*)$$

INDIAN STATISTICAL INSTITUTE

MID-SEMESTRAL EXAMINATION: (2015-2016)

MSQE I and M.Stat II

Microeconomic Theory II

Date: 29.02.2016 Maximum marks: 40 Duration: 2 Hours

Note: Answer **all** questions.

Note: Throughout, $\mathbb{R}^\ell$ is the ℓ-dimensional Euclidean space. Let

$$\mathbb{R}^\ell_+ = \{x = (x_1, \cdots, x_\ell) \in \mathbb{R}^\ell : x_i \geq 0 \text{ for all } 1 \leq i \leq \ell\}$$

and

$$\mathbb{R}^\ell_{++} = \{x = (x_1, \cdots, x_\ell) \in \mathbb{R}^\ell : x_i > 0 \text{ for all } 1 \leq i \leq \ell\}.$$

Note: Throughout, $\mathscr{E} = \{I; \mathbb{R}^\ell_+; (\succeq_i, \omega_i)_{i\in I}\}$ is an economy, where I is the set of agents containing m many elements; $\mathbb{R}^\ell_+$ is the consumption set of each agent; and $\succeq_i$ and ω_i are the preference and initial endowment of agent i, respectively. Suppose further that $\succ_i$ and $\sim_i$ are the strict preference and indifference relations associated with a rational preference relation $\succeq_i$ for all $i \in I$. A price is an element of $\mathbb{R}^\ell \setminus \{0\}$. Assume

$\mathscr{W}(\mathscr{E})$: the set of Walrasian equilibrium allocations of $\mathscr{E}$;

$\mathscr{C}(\mathscr{E})$: the core of $\mathscr{E}$;

$\mathscr{P}(\mathscr{E})$: the set of Pareto optimal allocations of $\mathscr{E}$.

Q1. Answer any **five** questions.

(i) Show that the demand set $D_i(p, \omega_i, \succeq_i) \neq \emptyset$ for a price $p \in \mathbb{R}^\ell_{++}$ and an upper semi-continuous rational preference $\succeq_i$.

(ii) Assume that $N = \{1, 2\}$ and $\ell = 2$. Suppose that $\succeq_i$ is represented by a utility function for $i = 1, 2$. Let

$$\begin{cases} \omega_1 = (0, 4), & U_1(x, y) = \sqrt{x} + \sqrt{y}; \\ \omega_2 = (2, 2), & U_2(x, y) = x. \end{cases}$$

Show that $((0, 6), (2, 0))$ is not an Walrasian equilibrium allocation of $\mathscr{E}$.

(iii) If $\succeq_i$ is strictly convex for all $i \in I$, then show that $\mathscr{W}(\mathscr{E}) \subseteq \mathscr{P}(\mathscr{E})$.

(iv) Suppose that $\succeq_i$ is strictly monotone on $\mathbb{R}^\ell_+$ such that everything in $\mathbb{R}^\ell_{++}$ is strictly preferred to anything on the boundary of $\mathbb{R}^\ell_+$ and $p \cdot \omega_i = 0$. Verify whether $D_i(p, \omega_i, \succeq_i) = \emptyset$.

(v) Let $I = \{1, 2\}$ and $\ell = 2$. Suppose that the preference relation $\succeq_i$ is represented by a utility function U_i for $i = 1, 2$. Given that

$$\begin{cases} \omega_1 = (1,6), & U_1(x,y) = \min\{x,y\}; \\ \omega_2 = (5,0), & U_2(x,y) = \min\{x,2y\}. \end{cases}$$

Find the set of Walrasian equilibrium of $\mathscr{E}$.

(vi) If an allocation is both a Walrasian equilibrium allocation and a quasi-equilibrium allocation, then show that it is a Pareto optimal allocation. $[5 \times 5 = 25]$

Q2. Answer any **two** questions.

(i) If $\{p_k : k \geq 1\} \subseteq \mathbb{R}^\ell_{++}$ satisfies $p_k \to p \in \mathbb{R}^\ell_{++}$, then show that there exists a bounded subset M_i of $\mathbb{R}^\ell_+$ such that the demand set $D_i(p_k, \omega_i, \succeq_i) \subseteq M_i$ holds for each $i \in I$ and $k \geq 1$.

(ii) Let $B_i(p)$ denote the budget set of agent $i \in I$ for a given price p. Suppose that $x \in \mathbb{R}^\ell_{++} \cap B_i(p)$ and $p \cdot y \geq p \cdot \omega_i$ for all $y \in \mathbb{R}^\ell_+$ satisfying $y \succ_i x$. If $\succeq_i$ is continuous and strictly monotone, then show that $p \in \mathbb{R}^\ell_{++}$.

(iii) Show that $\mathscr{C}(\mathscr{E})$ is a compact set. $[2 \times 7.5 = 15]$

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination: 2015–16(Second Semester)

M. Stat. II Year

Applied Multivariate Analysis

Date: March 1, 2016 | Maximum Marks: 50 | Duration: 2 hr

Note: The maximum you can score is 50.

1. Let $\boldsymbol{X}$ be a d-dimensional feature vector for a c-class supervised pattern recognition problem. For each class, the components of $\boldsymbol{X}$ are assumed to be mutually independent. In addition, each component X_i of $\boldsymbol{X}$ is binary-valued, assuming the values 1 or 0 with probabilities p_{ij} and $1 - p_{ij}$ respectively, in the jth class, $i = 1,2,\dots,d$, and $j = 1,2,\dots,c$.

   a. Deduce the Bayes classification rule for this problem explicitly assuming 0-1 loss, if the prior probability for class j is π_j.

   b. If $c = 2$, d is an odd integer, and $\pi_1 = \pi_2 = \frac{1}{2}$, show that the Bayes error is

   $$\sum_{k=0}^{(d-1)/2} \binom{d}{k} p^k (1-p)^{d-k},$$

   if $p_{i1} = p \ (> \frac{1}{2})$ and $p_{i2} = 1 - p$ for $i = 1,2,\dots,d$.

   c. How will the classifier in part (a) change if the loss λ_{ij} incurred in classifying an observation from class j into class i is of the general type, that is, not necessarily equal to 0 or 1? State clearly all the assumptions that you need to make.

   [7+7+(5+1)=20]

2. Consider two bivariate random variables $\boldsymbol{X} = (X_1, X_2)'$ and $\boldsymbol{Y} = (Y_1, Y_2)'$. The correlation matrix of $(X_1, X_2, Y_1, Y_2)'$ is

   $$\begin{pmatrix} 1 & 0.5 & 0.7 & 0.7 \\ 0.5 & 1 & 0.7 & 0.7 \\ 0.7 & 0.7 & 1 & 0.4 \\ 0.7 & 0.7 & 0.4 & 1 \end{pmatrix}.$$

   a. Show that the first canonical correlation coefficient ρ_1 of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is $\rho_1 = 0.97$, clearly explaining how you arrive at this result.

   b. Hence deduce the first canonical correlation vectors for $\boldsymbol{X}$ and $\boldsymbol{Y}$.

   [10+(5+5)=20]

3. In correspondence analysis, explain how the contingency chi-square statistic is decomposed into a number of components, and comment on the significance of these components. [10]

4. Assignment. [5]

INDIAN STATISTICAL INSTITUTE

M.Stat 2nd Year, 2015-16

Mid-semestral Examination

Subject: Theory of Games and Statistical Decisions

Full Marks: 40 Duration: 2 hrs Date: 02.03.2016

ATTEMPT ALL QUESTIONS

1. Give examples of matrices A and B such that

(i) $v(A+B) > v(A) + v(B)$

(ii) $v(A+B) < v(A) + v(B)$

(iii) $v(A+B) = v(A) + v(B)$

where $v(X)$ denotes value of mixed extension of the matrix game X. [3+3+2]

2. Show that $v(A)$ is increasing and continuous function of the matrix $A \varepsilon R^{mn}$, where $v(A)$ denotes value of the mixed extension of the matrix game for an $m \times n$ matrix A. [8]

3. $f: A \times B \rightarrow R$ is a function. When do we call (a_0, b_0) to be a saddle point of $f(a,b)$? Deduce necessary and sufficient condition for the existence of saddle points using max inf and min sup. [1+11]

4. Let $A: n \times n$ be a diagonal matrix with diagonal entries $a_1, a_2, \ldots\ldots\ldots, a_n$ respectively.

(i) If $a_i > 0$, for all i, find $v(A)$.

(ii) If $a_i < 0$, for all i, find $v(A)$.

(iii) If $a_i < 0$ and $a_j > 0$ for some $i \neq j$, find $v(A)$. [7+3+2]

....

Statistical Methods in Public Health
Semestral Examination
M.Stat. II Year, 2015-2016
Total Marks - 100
Time - 3 hrs. 30 mins

Indian Statistical Institute
Kolkata 700 108, INDIA

26.04.16

Attempt all questions:

Group A

1. Let us define X be an $(n \times q)$ longitudinal data matrix in which the ith row corresponds to a q - variate size measurements available at q equispaced time points on the ith of the n individuals. Let us also assume that the ith row $X_i = (X_i(1), ..., X_i(q))' \sim N_q(\theta, \Sigma)$, where $X_i(t)$ = size at time point t for the ith individual, and $E(X_i(t)) = \theta(t) = f(\phi, t)$, is a suitable growth curve, $t = 1, \ldots, q$ and $\theta = (\theta(1), \theta(2), \ldots, \theta(q))'$. We define two commonly used estimates of "relative growth rate" (RGR) for any time interval $(t, t+1]$,

$$(i)\ \ \bar{R}(t) = \ln\left(\frac{\bar{X}(t+1)}{\bar{X}(t)}\right) \quad and \quad (ii)\ \ \bar{R}(t) = \frac{1}{n}\sum \ln\left[\frac{X_i(t+1)}{X_i(t)}\right]$$

Assuming that the conditional mean $E(\bar{R}(t)/\bar{X}(t))$ has a "theta - logistic" structure and the mean size is bounded by twice the carrying capacity show that $(\bar{R}(t), \bar{X}(t))$ asymptotically follows the bivariate normal distribution under both the estimates of RGR.

[10]

2. Suppose we are interested in testing the hypothesis of exponential quadratic growth curve model (EPQGM), i.e., to test

$$H_0 : \theta(t) = e^{b_0 + b_1 t + b_2 t^2} \ \text{ag.}\ \ H_1 : \text{not } H_0$$

, based on the data structure as described in the question number 1.(a).

(a) Using the approximate expression for expectation and variance of the logarithm of ratio of size variables for two consecutive time points describe two testing procedures and critical regions in testing the null hypothesis of EPQGM.

(b) Also suggest required modifications of test statistics when the time spacings are unequal and the errors are non-normal.

[9 + (4 + 3) = 16]

or

2. (a) Define Fisher's Relative Growth Rate (RGR) based on the size measurements at t specific time points. Comments on its extension and growth law non-invariant form. Derive t expression for this extended RGR metric for the logistic growth law based on the size measur ments at three consecutive time points. How this extended metric is affected through reading measurement errors under logistic law ?

(b) Show that the bias for Fisher's RGR is always negative under first order of approximati (where order is defined by the power of the "difference of relative errors" at two time points). Al obtain the expression for MSE under same approximation.

[(1 + 3 + 4 + 3) + (2 + 3) = 16]

3. (a) Define quasi-equilibrium probabilities of a general birth death process.

(b) Consider the Von Bertalanffy growth equation for a single species population dynamics a follows:

$$\frac{dx}{dt} = ax^{\frac{2}{3}} - bx,$$

where parameters have their usual interpretations.

Derive the quasi equilibrium probabilities and determine the expression for the approximat mean and variance by incorporating random perturbation in the above model. You may assum that the growth variable x to be bounbed by twice the carrying capacity.

[2 + (4 + 7) = 13]

4. (a) Find the analytical solution of the growth curve governed by the following growth equatio

$$\frac{1}{x(t)}\frac{dx(t)}{dt} = bt^c exp(-at) \qquad (*),$$

where, $a, b > 0$ and c is an integer. Compare the point of inflexions of RGR curves for $c = 0$ and 1 separately with proper interpretations.

[5 + 6 = 11]

or

(a) Let us rewrite eqn. (*) as

$$R_t = b\ t^c e^{-at} + \epsilon_t \qquad (**),$$

where, R_t is the empirical estimate of RGR at time t. Show that nonlinear least square estimates derived from (**) are consistent and asymptotically normal.

[4 + 7 = 11]

Group B

1. Note the following assumptions (a) - (d):

(a) susceptible individuals are recruited either by birth or immigration into the population at a constant rate

(b) the disease transmission is of standard incidence type

(c) both the population is affected by the natural death rate (you may consider the natural death rate as constant)

(d) the population is also affected by recovery rate and the disease related death rate

(i) Formulate an epidemiological model of SIS type.

(ii) Prove that the solutions of the above system are uniformly bounded.

(iii) Find out the disease free and endemic steady states. Hence, derive the condition(s) for which the disease will spread.

(iv) By constructing a suitable Liapunov function, show that the endemic steady state is globally asymptotically stable.

[2 + 2 + 2 + 4 = 10]

2. (a) Stating clearly the assumptions, formulate Kermack and McKendrick general epidemic model.

(b) Derive Kermack and McKendric threshold phenomenon.

(c) Show that for the general epidemic model, the spread of the disease will not stop due to lack of susceptible.

(d) A survey of freshman students at Yale University found that 25% were susceptible to rubella at the beginning of the year and 9.65% were susceptible at the end of the year. What fraction would have had to be immunized to avoid the spread of rubella?

[4 + 3 + 5 + 5 = 17]

3. (a) State the basic assumptions of pure birth and death process. Hence formulate the basic stochastic differential equation of the process.

(b) Comment on the behavior of the system by computing mean and variance of the process.

[5 + 8 = 13]

4. Define reproductive ratio (R_0). Find out R_0 of malaria model by using next generation matrix approach.

[2 + 8 = 10]

INDIAN STATISTICAL INSTITUTE

M.Stat./M.Math. II Year
Semestral Examination : Semester II : 2015-2016
STOCHASTIC PROCESSES I

Date : 26.04.2016 **Maximum Score : 60** Time : $3\frac{1}{2}$ Hours

Note : This paper carries questions worth a **total** of **72 marks**. Answer as much as you can. The **maximum** you can score is **60**.

1. Let $\{B_t, t \in [0,\infty)\}$ be a Standard Brownian Motion on some probability space. Fix $0 \le a < b$ and let $Q(\pi; a, b)$ denote the quadratic variation of $\{B_t\}$ along a finite partition of $[a, b]$.
(a) Show that $E[Q(\pi; a, b) - (b-a)]^2 \le 3(b-a)\|\pi\|$, where $\|\pi\|$ denotes the 'norm' of the partition π.
(b) Deduce that if $\{\pi_n\}$ is a sequence of partitions with $\sum_n \|\pi_n\| < \infty$, then $Q(\pi_n; a, b)$ converges to $(b-a)$ in L_2 and also almost surely. (5+5) = **[10]**

2. Let $\{B_t, t \in [0,\infty)\}$ be a Standard Brownian Motion on some probability space. For $0 \le a < b$, let M_a^b denote $\sup_{a \le t \le b} B_t$.
(a) Show that for $0 \le a < b < c < d$, the random variables $(M_a^b - B_b), (B_c - B_b)$ and $(M_c^d - B_c)$ are mutually independent and hence show that P-almost surely $M_a^b \ne M_c^d$.
(b) Is the above result true in case $a < b = c < d$? Justify your answer. (5+5) = **[10]**

3. Let $\{B_t, t \in [0,\infty)\}$ be a Standard Brownian Motion on some probability space. Given a real-valued continuous function f on $[0,\infty)$, show that for almost every ω, the function $B(\cdot,\omega) - f(\cdot)$ on $[0,\infty)$ is nowhere differentiable. **[10]**

4. Let $\{B_t, t \in [0,\infty)\}$ be a Standard Brownian Motion on some probability space.
(a) State the law of iterated logarithm as $t \to 0$ and show that if $\tau = \inf\{t > 0 : B_t = 0\}$, then $\tau = 0$ almost surely.
(b) Fix an $r > 0$ and let $\tau_r = \inf\{t \ge r : B_t = 0\}$. Show that $\tau_r = \inf\{t > \tau_r : B_t = 0\}$ almost surely.
(c) Show that, for almost every ω, the set $Z_\omega = \{t \ge 0 : B_t(\omega) = 0\}$ is a closed unbounded set of Lebesgue measure zero with no isolated points. ((2+3)+5+5) = **[15]**

5. Let S be a separable metric space with $\{x_1, x_2, \ldots\}$ a countable dense subset. Let P be a probability on the Borel σ-field on S.
(a) Show that for each $n \ge 1$, there exist an integer $k_n \ge 1$ and disjoint Borel sets $A_1, \ldots, A_{k_n}$ with $A_i \subset B\left(x_i, \frac{1}{n}\right)$, for $1 \le i \le k_n$ and $\sum_{i=1}^{k_n} P(A_i) > 1 - \frac{1}{n}$.
(b) For each $n \ge 1$, choose rational numbers $s_{i,n} \in [0,1]$, $1 \le i \le k_n$, in such a way that $\sum_{i=1}^{k_n} |P(A_i) - s_{i,n}| < \frac{1}{n}$. Show that $\alpha_n = \sum_{i=1}^{k_n} s_{i,n}$ satisfies $1 - \frac{2}{n} < \alpha_n < 1 + \frac{1}{n}$.
(c) For each $n \ge 1$, let P_n be the probability supported on the finite set $\{x_1, \ldots, x_{k_n}\}$ with $P_n(\{x_i\}) = \frac{s_{i,n}}{\alpha_n}$, $1 \le i \le k_n$. Show that $P_n \xrightarrow{w} P$. (3+2+7) = **[12]**

6. Consider the Markov Process defined by the family of probabilities $\{P_x, x \in \mathbb{R}\}$ on $\Omega_c = \Omega_c(\mathbb{R})$, such that, under P_x, the co-ordinate Process $\{X_t, t \ge 0\}$ on Ω_c is distributed like the process $\{x + t + B_t, t \ge 0\}$, where $\{B_t, t \ge 0\}$ is a Standard Brownian Motion.
(a) Write down the transition probabilities and show that the Markov Process has the Feller property.
(b) For $f \in C_b(\mathbb{R})$ and $\lambda > 0$, show that the function $u = R_\lambda f$ is twice continuously differentiable and satisfies $u'' = 2\lambda u - 2u' - 2f$.
[You may use the fact that $\int_0^\infty e^{-(\alpha u - \beta/u)^2} du = \frac{\sqrt{2\pi}}{2\alpha}$ for any $\alpha > 0, \beta > 0$.]
(c) Using $C_b(\mathbb{R})$ as the domain of the underlying semigroup, describe the generator of the above Markov Process. (3+10+2) = **[15]**

INDIAN STATISTICAL INSTITUTE

Second Semestral Examination: 2015-16

M. Stat. II Year

APPLIED MULTIVARIATE ANALYSIS

Date: April 27, 2016 Maximum Marks: 100 Duration: 3 hr

Note: *Answer as many questions as you can. The maximum you can score is 100.*

1.

a) Formulate the problem of clustering a set of N observations on a p-variate random variable $\boldsymbol{X}$ into K clusters, as that of the decomposition of a finite, identifiable mixture of probability densities.

b) If the components of this mixture density are $\mathcal{N}_p(\mu_i, \Sigma_i)$, for i=1,2,..., K, all the parameters including the mixing proportions π_i ($i = 1, 2, \cdots, K$) being unknown, then derive an iterative algorithm for the maximum likelihood estimation of all the parameters.

[5+15=20]

2. Answer **<u>any two</u>** of the following:

a) Describe the metric multidimensional scaling (MDS) problem and describe the main steps of the classical solution to it.

b) What are Multivariate Adaptive Regression Spline (MARS)? Describe how such models are fit to data.

c) For finding the decision boundary in a two-category discrimination problem, formulate the optimization problem under the Support Vector Machine approach when the classes are linearly separable.

[10+10=20]

3. Consider two bivariate random variables $\boldsymbol{X} = (X_1, X_2)'$ and $\boldsymbol{Y} = (Y_1, Y_2)'$. The correlation matrix of $(X_1, X_2, Y_1, Y_2)'$ is

$$\begin{pmatrix} 1 & 0.2 & 0.3 & 0.3 \\ 0.2 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.5 \\ 0.3 & 0.3 & 0.5 & 1 \end{pmatrix}.$$

a. Show that the first canonical correlation coefficient of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is 0.4472.

b. Hence deduce the first canonical correlation vectors for $\boldsymbol{X}$ and $\boldsymbol{Y}$.

[10+(5+5)=20]

(Please Turn Over)

4. Consider the problem of discrimination between two populations Π_1, Π_2 on the basis of observations on a random variable X. It is assumed that in Π_i, i=1,2, X is distributed as

$$p(x|\Pi_1) = \frac{1}{2\pi} \cdot \frac{1}{1 + \left(\frac{x-3}{2}\right)^2},$$
$$p(x|\Pi_2) = \frac{1}{2\pi} \cdot \frac{1}{1 + \left(\frac{x-6}{2}\right)^2}.$$

Assume that $\pi_1 = \pi_2 = \frac{1}{2}$, where π_i denotes the prior probability for the population Π_i, $i = 1,2$.

a. Deduce the explicit form of the Bayes rule for discriminating between Π_i and Π_i under zero-one loss.
b. Show that the overall error probability for this rule is

$$\frac{1}{2} - \frac{1}{\pi}\tan^{-1}\frac{3}{4}.$$

c. Write down the modified Bayes discriminant rule if, instead of zero-one loss, the following loss function is used:

$$l(i,j) = \begin{cases} 0 & \text{if } i = j, \\ 2 & \text{if } i = 1, j = 2, \\ 1 & \text{if } i = 2, j = 1. \end{cases}$$

Here $l(i,j)$ denotes the loss incurred when an observation from Π_i is allocated to Π_j, i,j = 1,2. How does the overall error probability change under this loss? Explain.

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

5. Consider a two-category discrimination problem based on four binary variables X_1, X_2, X_3 and X_4 Suppose the following observations (in the form of 4-tuples of bits) are available from the two categories, where the i-th bit in each 4-tuple represents the observation on $X_i, i = 1,2,3,4$:

Category 1	0110	1010	0011	1111
Category 2	1011	0000	0100	1110

Use the misclassification impurity measure to create an unpruned binary classification tree from this data. [20]

6. Assignment [10]

INDIAN STATISTICAL INSTITUTE

Semestral Examination : Semester II (2015-16)

M. Stat. II Year

Survival Analysis

)ate: 29.04.2016 Maximum marks: 100 Time: 3 hours

Answer as many as you can. Total mark is 86. Maximum mark is 80.

1. State if the following statements are true or false giving suitable reasons.

   (a) The vector $\underset{\sim}{\lambda} = (0.25, 0.25, 0.25, 0.25)$ represents the discrete hazards for a discrete lifetime variable with four mass points.

   (b) The discrete time regression model, obtained by grouping the lifetime following the Cox proportional hazards model, does have proportional hazards.

   (c) Starting with n individuals subject to random right censoring, having independent exponential life time and exponential censoring time distribution with mean 50 days and 75 days, respectively, the expected number of failures is $n \times 50/125$.

   (d) The integrated hazard $\Lambda(t, z)$ for the accelerated failure time model $\lambda(t, z) = \lambda_0(te^{z\beta})e^{z\beta}$ is given by $\Lambda(t, z) = \Lambda_0(te^{z\beta})$.

   (e) A series system is constructed with two independent components having identical life distributions, the survival function of which is estimated by $\hat{S}(t)$ with standard error $s(t)$. The standard error of the estimated system survival function is $2s(t)$.

   (f) Consider the multiple decrement model, with covariate z, given by $Q(t_1, \cdots, t_m; z) = [Q(t_1, \cdots, t_m)]^{\exp(z\beta)}$, where $Q(t_1, \cdots, t_m)$ denotes the unknown and arbitrary 'baseline multiple decrement function'. The corresponding cause specific hazard rates are of proportional hazards form. [3 × 6 = 18]

2. Consider type II censored data from Weibull$(\lambda, p = 2)$ life distribution. Develop an exact test for $\lambda = \lambda_0$ against $\lambda \neq \lambda_0$. Suggest a graphical test for the validity of the model.

   [5+3=8]

3. Describe the life table estimates of a survival function with grouped survival data including withdrawals. Mention one of its limitations. [6+1=7]

4. (a) Suppose that data has been generated from two groups of individuals by the Cox's proportional hazards model and that we observe the following values:

   $$(U_i, \delta_i, z_i) = (16, 1, 1), (13, 1, 0), (21, 1, 1), (11, 1, 0), (12, 1, 1), (14, 0, 1), (24, 0, 0) ,$$

   where U_i denotes the observation time, δ_i is the censoring indicator (1 if uncensored and 0 if censored) and z_i is the group indicator (0 for Group I and 1 for Group II). Write down the expression for the Cox's partial likelihood given the above data. How will the partial likelihood change if the observation times U_i's (measured in months) are recorded as $30 \times U_i$'s (measured in days)?

   (b) Suggest one method for testing goodness of fit of the Cox's Proportional Hazards model. [(10+1)+2=13]

5. (a) Describe the Buckley-James estimator for the regression parameter in the framework of accelerated failure time model.

(b) Noting that the locally most powerful rank test statistic for the significance testing of the regression parameter based on only uncensored data, in the framework of accelerated failure time model, has the form $\sum_{i=1}^{n} c_i z_{(i)}$ with $\sum_{i=1}^{n} c_i = 0$, write down the form of c_i explaining all the notation and the constraint. Obtain c_i when the error variable follows extreme value distribution.

[8+(5+3)=16]

6. (a) Consider the competing risks model given by the cause-specific hazards

$$\lambda_j(t,z) = \lambda_0(t)e^{z\beta_j},\ j = 1,\cdots,m,$$

with $\lambda_0(t)$ being unknown and arbitrary. Obtain an appropriate partial likelihood to estimate the regression parameters β_j's.

(b) Consider the competing risks model given by the cause-specific hazards

$$\lambda_j(t,z) = \lambda_0(t)e^{\gamma_j+z\beta_j},\ j = 1,\cdots,m,$$

with $\gamma_1 = 0$ and $\lambda_0(t)$ being unknown and arbitrary. Obtain an appropriate partial likelihood to estimate the parameters γ_j's and β_j's. Find $P[J = j; z]$ for a fixed covariate z, where J denotes the cause of death, and hence prove that the life time T and cause of death J are independent.

(c) Consider a parallel system with two independent components having exponential life distributions with failure rates λ_1 and λ_2, respectively. If the jth component fails ($j = 1$ or 2), it is immediately repaired after which its failure rate changes to α_j. The second failure of a component is beyond repair. Describe the life time of this system as a multiple failure time model by identifying the various cause-specific hazard rates.

[6+(6+3+3)+6=24]

INDIAN STATISTICAL INSTITUTE

Final Examination : Semester II (2015-2016)

Course Name : BSDA (M. Stat. 2nd year)

Subject Name : Statistical Methods in Biomedical Research

Date : 02.05.2016, Maximum Marks : 60. Duration : 3 hrs.

1. Explain how to estimate the relative potency of a test drug with respect to a standard one using direct bioassay. Derive the variance of your estimator. [4+2]

2. Describe the Continual Reassessment Method of dose finding in a phase I clinical trial. [6]

3. Define type I error spending function $\alpha^*(t)$. Suppose for all s, t such that $0 \leq s < t \leq 1$, the type I error in the interval $(s, t]$ will be proportional to $(t-s)^\delta$ for some real δ. Find all possible values of δ and interpret. [2+6]

4. Suppose 60 patients are to be treated in a clinical trial for comparing two treatments A and B. Out of the first 18 patients, 10 are treated by A and 8 are treated by B. What will be the allocation probability for the 19th patient to treatment A for

   (i) Random allocation rule (RAR);

   (ii) Truncated binomial design (TBD);

   (iii) Permuted block design ($B = 6$, $b = 10$);

   (iv) Permuted block design ($B = 10$, $b = 6$);

   (v) Eforn's biased coin design with $p = 3/4$;

   (vi) Big stick rule ($c = 2$);

   (vii) Biased coin design with imbalance tolerance (BCDWIT(p,$c = 3$));

   (viii) Friedman-Wei's urn design ($\alpha = \beta$);

   (ix) Ehrenfest urn design ($w = 5$);

   (x) Generalized biased coin design (GBCD($\rho = 5$))?

   [10]

5. A three treatment play-the-winner rule (with treatments A, B and C) is defined as follows. The first patient is treated by any of the treatments with equal probability. Thereafter success by a treatment results the next patient to be treated by the same treatment, whereas the next patient is treated by either of the remaining two treatments with equal probabilities for a failure by any treatment. Find the unconditional probability of treating the nth patient by treatment A in terms of a recursion relation. [10]

6. Consider a two-treatment response-adaptive allocation with binary responses. Let p_A and p_B be the success probabilities by the two treatments, say A and B. Suppose for $0 < a < b < 1$, (p_A, p_B) has the following two-point prior:

$$(p_A, p_B) = \begin{cases} (a, b) & \text{with probability } 0.5 \\ (b, a) & \text{with probability } 0.5. \end{cases}$$

The proposed Bayesian response-adaptive design allocates any entering patient to treatment A with probability equal to the posterior probability of the event $\{p_A > p_B\}$. Find the allocation probability for the 26-th patient to treatment A given that out of the first 25 patients 15 are treated by treatment A and the number of failures are 6 for both the treatments so far. [10]

7. In a clinical trial, blood pressure level (systolic, diastolic), cholesterol level (HDL, LDL and Triglyceride) and blood sugar level are measured for 100 patients on a particular day. Based on these 6 covariates, the health status score is measured for each patient in a scale from 0 to 10. For 85 patients, we observe the complete data (i.e. there is no missing observation) but for the others we have missing values for at least 3 variables per subject. Assuming "missing at random" mechanism, suggest a suitable multiple imputation technique for imputing the missing values for a powerful statistical inference. [10]

ASYMPTOTIC THEORY OF INFERENCE

M-Stat II, Second Semester 2015-2016

Semestral Examination

Date: 02.05.2016 Maximum marks: 100 Duration: 3 hours.

Note: Total marks 120; maximum you can score is 100. If necessary conditions are not provided assume what is required, but clearly state them.

1. Suppose that $X_1, \ldots, X_n$ are i.i.d. each with with a density $f_\theta(x)$ with respect to an appropriate dominating measure μ, and $\theta = (\theta_1, \theta_2, \ldots, \theta_s)$, where $s > 1$ is fixed. Show, under suitable conditions to be provided by you, that with probability tending to 1 as $n \to \infty$, there exist solutions $\hat{\theta}_n = (\hat{\theta}_{1n}, \hat{\theta}_{2n}, \ldots, \hat{\theta}_{sn})$ of the likelihood equations such that

   (a) $\hat{\theta}_{jn}$ is consistent for estimating θ_j.

   (b) $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normal with (vector) mean zero and covariance matrix $[I(\theta)]^{-1}$, where $I(\theta)$ is the Fisher information matrix.

   [4+8+8=20]

2. (a) Let X_n be an extended real-valued random variable on $(\Omega_n, \mathcal{A}_n, P_n), n \geq 1$. Show that $P_n(X_n > u_n) \to 0$ for every $\{u_n\}$ such that $-\infty \leq u_n \leq \infty$ and $u_n \to \infty$ if and only if for each $\epsilon > 0$, there exists a real $b > 0$ and an integer $N \geq 1$ satisfying $P_n(X_n > b) < \epsilon$ for all $n \geq N$. [5+5=10]

   (b) Let $X_1, X_2, \ldots, X_n$ be i.i.d. observations from the density $f_\theta(x)$, $\theta = (\theta_1, \theta_2)^T$. We want to test

   $$H_0 : \theta_1 = \theta_{01} \text{ versus } H_1 : \theta_1 \neq \theta_{01},$$

   where θ_{01} is known. (Here θ_2 is an unknown nuisance parameter). Write down the likelihood ratio statistic for the above testing problem. Derive the asymptotic null distribution of the suitably normalized likelihood ratio statistic. Clearly state the assumptions (and notations) you need. [1+(7+2)=10]

3. (a) Define a statistical functional $T(F)$. Give two examples of statistical functionals. [4]

   (b) Under appropriate assumptions and notations, define the von-Mises derivative of a statistical functional. Give an example where the von-Mises derivative does not exist. [6]

   (c) What is the von-Mises expansion of the functional T at F? Show how this is helpful for determining the asymptotic distribution of the statistic $T(F_n)$, where F_n is the empirical distribution function based on an i.i.d. sample. [4]

(d) Let V and W be topological vector spaces and let $L(V, W)$ be the set of all continuous linear transformations from V to W. Let $\mathcal{S}$ be the class of subsets of V such that every subset consisting of a single point belongs to $\mathcal{S}$ and let A be an open subset of V. In this connection define the $\mathcal{S}$-differentiability of a function $T : A \to W$ at $F \in A$. Also differentiate between Hadamard differentiability, Fréchet differentiability and Gateaux differentiability. [6]

4. (a) Define superefficiency, and give an example of a superefficient estimator. Explain why superefficiency is not a statistically relevant concept. [5]

   (b) Construct an example where the maximum likelihood estimator is inconsistent, but the inconsistency may be removed by suitably modifying the conditions. [8]

   (c) Give an example where the maximum likelihood estimator does not exist but there exists a consistent sequence of roots of the likelihood equations. [7]

5. (a) Define the exact slope of a test statistic. [6]

   (b) Under appropriate conditions, derive the exact slope of the likelihood ratio test statistic. How is it linked to the Kullback-Leibler information? [14]

6. (a) Let $X_1, \ldots, X_n$ be i.i.d. observations from a discrete distribution supported on $\chi = \{0, 1, 2, \ldots\}$. Let G be the true distribution function having probability mass function $g(x)$; let $f_\theta(x)$ represent the probability mass function of the model. Also let $d_n(x)$ be the relative frequency at $x \in \chi$. Define the disparity $\rho_C(d_n, f_\theta)$ between d_n and f_θ based on the function $C(\cdot)$. Desribe the properties that the defining function $C(\cdot)$ must have. [4]

   (b) What is the residual adjustment function (RAF) of a disparity? What is a regular RAF? [4]

   (c) Show, under appropriate assumptions, that

$$-n^{1/2}\nabla\rho_C(d_n, f_\theta) = n^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n} u_\theta(X_i)\right] + o_p(1),$$

where $u_\theta(x) = \nabla \log f_\theta(x)$ is the likelihood score function, and ∇ represents derivative with respect to θ. [12]

Semester 2, 2015-2016 FINAL EXAM Date: 4 May 2016

MStatII DIRECTIONAL DATA ANALYSIS Time: 3 hrs.

Show all your work. Marks are indicated in the margin. Total marks: 100.

Q. 1. [25=12+13] (a) Derive the general Fourier series representation of the probability density function of a circular random variable in terms of its trigonometric moments.

(b) Consider the circular distribution with probability density function given by

$f(\theta) = K.[1 + 2\rho cos(\theta - \mu) + 2\rho^2 cos2(\theta - \mu)]$

(i) Obtain K. (ii) Derive the characteristic function of Θ.

Q. 2. [25=13+12]

(a) (i) Show that the method of stereographic projection gives an unified approach of deriving a circular probability density function from a given linear probability density function and vice-versa.

(ii) For the method in (i), (1) derive the linear probability density function corresponding to the circular von Mises distribution and (2) establish that the circular uniform distribution is the circular distribution corresponding to the Cauchy distribution.

(b) Explain why the Cramer - von Mises functional test for Goodness-of-Fit is not applicable to a circular probability density function and prove that Watson's U^2 test overcomes this problem.

Q. 3. [25=12+13]

3. (a) (i) Derive the Locally Most Powerful test for Isotropy against the family of symmetric wrapped stable distribution.

(ii) Obtain the interval of alternatives under which the test in (a) has a monotone power function.

(b) (i) Describe one model each for Cylindrical and Toroidal regression.

(ii) Prove that for a certain circular distribution, to be presented by you, the method of Maximum Likelihood and that of Circular Least Squares are equivalent in Toroidal regression.

Q. 4. [25] TAKE HOME. (a) For the change-point problem with a von Mises distribution, obtain the threshold values using the method of simulation discussed in class.

INDIAN STATISTICAL INSTIUTE

M Stat. 2nd Year, 2015 – 16

Second Samestral Examination

Subject: Theory of Games and Statistical Decisions

Date: 06.05.2016 Full Marks: 60 Duration: 3 hours

Attempt all questions

1. Let set of States of Nature = { θ_1 , θ_2 }
   $\Upsilon(\theta_i, \delta)$ = risk of decision rule δ, i = 1, 2.
   A = { $(x_1, x_2) : x_i = \Upsilon(\theta_i, \delta)$, i =1, 2 for possible decision rule δ }
   Let A takes the value, for some statistical game, given by
   $$A = \{ (x_1, x_2) : 16(x_1 - 8)^2 + 9(x_2 - 8)^2 \le 144 \}$$
   For this problem,
   (a) So that if δ_0 is minimax decision rule then $\Upsilon(\theta_1, \delta_0) = \Upsilon(\theta_2, \delta_0)$
   (b) Find point P in A corresponding to δ_0
   (c) Find Least Favourable Configuration.
   (d) Show that δ_0 is Bayes with respect to this prior.

   [4 + 3 + 4 + 3]

2. Let { I, v } denote a cooperative game with I = set of players, v = characteristic function.
   (a) Define strategic equivalence of cooperative games.
   (b) Show that strategic equivalence is an equivalence relation.
   (c) Consider all cooperative games { I, v } derived from constant-sum non-cooperative games with 3 players. Show that there are only 2 equivalence classes for such games.

   [3 + 3 + 8]

3. Let $A^t = [a_1, a_2, a_3]$ with
   $a_1 = (1, 0)^t$, $a_2 = (0, 1)^t$, $a_3 = (\frac{1}{2}, 1)^t$
   For the 3 x 2 matrix game A, find
   (a) Value of the game.
   (b) Optimal strategies of player I and Player II.

   [4 + (4 + 4)]

4. (a) State Nash's Theorem in connection to finite non-cooperative games.
   (b) Prove Nash's Theorem stating all lammas you might need in the proof.
   (c) Give an example of a 3 - player non-cooperative game with each player having 3 strategies such that the game has more than one equilibrium situations each with different pay-off for the same player.
   (d) Show that if mixed extension of a matrix game has more than one equilibrium situations, each situation has same set of pay-off.

   [2 + 11 + 4 + 3]

---------- O ----------

Date: May 9, 2016
Time: 3 hours

Statistical Methods in Genetics – II
M-Stat (2nd Year)
Second Semester Examination 2015-16

The questions are divided into two groups: Group A and Group B.

You need to answer Group A within the first 90 minutes. This part is closed notes. You will be given the questions of Group B after you submit your answer script corresponding to Group A. This part is open notes. The paper carries 66 marks with each group carrying 33 marks.

Group A

Answer Question 1 and any four from the remaining questions.

1. Define bioinformatics. [1]

2. Outline the basic steps for any method for predicting gene functions, with necessary equations, using microarray gene expression, protein transitive homologues and KEGG pathway profiles as databases. [8]

3. Write the algorithm and construct the perfect binary phylogenetic tree for the given character state matrix shown below. [8]

	Character					
Object	c_1	c_2	c_3	c_4	c_5	c_6
A	0	0	0	1	1	0
B	1	1	0	0	0	0
C	0	0	0	1	1	1
D	1	0	1	0	0	0
E	0	0	0	1	0	0

4. Find a possible semi-global alignment between sequences "*AGC*" and "*AAAC*" through explaining the necessary steps in bidimensional score matrix. [8]

P.T.O

5. Describe the Jukes-Cantor model to study the evolution of DNA sequences. [8]

6. Briefly describe the BLAST algorithm for aligning two protein sequences with example. [8]

7. Explain the basic steps for constructing the phylogenetic tree from the distance matrix shown below. [8]

	A	B	C	D	E
A	0				
B	12	0			
C	14	12	0		
D	14	12	6	0	
E	15	13	7	3	0

Group B

Answer all questions

8(a) Consider data on a quantitative trait on sib-pairs and genotypes at a marker locus for both parents as well as the sibs. Assuming the classical Haseman-Elston framework with no dominance, show that the correlation of the sib-pair trait values conditioned on their i.b.d. score at the marker locus is a linear function of the i.b.d. scores. How would you test for linkage between a QTL and a marker locus based on the above property?

(b) What is the expected value of the classical TDT statistic under the null hypothesis of no linkage or no association? [7 + 3]

9(a) Consider a dominant disorder controlled by an autosomal biallelic locus. Suppose that a marker locus is in linkage disequilibrium with the disease locus. For what value of the disease allele frequency will the difference between the marker allele frequencies among cases and controls be the maximum?

(b) Consider genotype data at a marker locus on parent-offspring pairs where the offspring is affected with a recessive disorder. Show that the test for linkage disequilibrium between the marker locus and the disease locus is equivalent to a test for Hardy-Weinberg Equilibrium of the parental genotypes. Explain whether this test is protected against population stratification. [6 + 9]

10. Consider the following data at a marker locus in an affected sib-pair study:

Sib-pair	Parental marker genotypes	Sib-pair marker genotypes
1	AC-BC	AC-AC
2	AB-AB	BB-BB
3	AB-BC	AB-AC
4	AC- *	AA-AB
5	AB-AC	AB-AB

(* denotes missing genotype)

Using Holmans' Triangle approach, test for linkage between the marker locus and a disease locus. [8]

5. Describe the Jukes-Cantor model to study the evolution of DNA sequences. [8]

6. Briefly describe the BLAST algorithm for aligning two protein sequences with example. [8]

7. Explain the basic steps for constructing the phylogenetic tree from the distance matrix shown below. [8]

	A	B	C	D	E
A	0				
B	12	0			
C	14	12	0		
D	14	12	6	0	
E	15	13	7	3	0

Group B

Answer all questions

8(a) Consider data on a quantitative trait on sib-pairs and genotypes at a marker locus for both parents as well as the sibs. Assuming the classical Haseman-Elston framework with no dominance, show that the correlation of the sib-pair trait values conditioned on their i.b.d. score at the marker locus is a linear function of the i.b.d. scores. How would you test for linkage between a QTL and a marker locus based on the above property?

(b) What is the expected value of the classical TDT statistic under the null hypothesis of no linkage or no association? [7 + 3]

9(a) Consider a dominant disorder controlled by an autosomal biallelic locus. Suppose that a marker locus is in linkage disequilibrium with the disease locus. For what value of the disease allele frequency will the difference between the marker allele frequencies among cases and controls be the maximum?

(b) Consider genotype data at a marker locus on parent-offspring pairs where the offspring is affected with a recessive disorder. Show that the test for linkage disequilibrium between the marker locus and the disease locus is equivalent to a test for Hardy-Weinberg Equilibrium of the parental genotypes. Explain whether this test is protected against population stratification. [6 + 9]

10. Consider the following data at a marker locus in an affected sib-pair study:

Sib-pair	Parental marker genotypes	Sib-pair marker genotypes
1	AC-BC	AC-AC
2	AB-AB	BB-BB
3	AB-BC	AB-AC
4	AC- *	AA-AB
5	AB-AC	AB-AB

(* denotes missing genotype)

Using Holmans' Triangle approach, test for linkage between the marker locus and a disease locus. [8]

Indian Statistical Institute
Second Semestral Examination 2015-16
M. Stat II year
Statistical Inference II

Date: May 10, 2016 | Maximum marks: 100 | Duration: 3 hrs.

Answer all Questions. *Paper carries 120 points.*

1 (a) Define defficiency between two statistical experiments $\mathcal{E}$ and $\mathcal{F}$.

(b) Consider two experiments where n and $(n+1)$ random samples have been drwan repectively from a Poission distribution with unknown mean λ for a given $n \geq 1$. By denoting these experiments by $\mathcal{E}$ and $\mathcal{F}$ respectively compute $\delta(\mathcal{F}||\mathcal{E})$ and $\delta(\mathcal{E}||\mathcal{F})$ respectively. [Note that both answers might not be in eaxct closed form. Well justified approximations/ upper/ lower bounds will get credit too.]

[5 + (10+10) = 25]

2 (a) Let $\mathcal{M} = \{P_\theta\}$ be a statistical model on a measurable space $(\Omega, \mathcal{A})$. Define coherence of $(\Omega, \mathcal{A}, \mathcal{M})$.

(a) Give an example of a model which is not coherent. Justify your answer.

[5+15 = 20]

3 (a) Define notions of specific sufficiency, θ-oriented statistic and partial sufficiency in a model $(\Omega, \mathcal{A}, \mathcal{M})$ where $\mathcal{M} = \{P_{\theta,\phi}\}$ with ϕ being a nuisance parameter.

(b) Consider the family $\{P_{\theta,\phi}\} = \text{Unif}(\phi\theta, \theta/\phi)$ where $\theta > 0$ and $0 < \phi < 1$ respectively. Let $X_1, X_2, \ldots, X_n$ be iid samples from $\{P_{\theta,\phi}\}$. Do there exist specific sufficient, oriented and partial sufficient statistics for θ and ϕ respectively?

[10+ (10+10) =30]

4 (a) State the three principles of foundation of inference, namely: sufficiency, conditionality and likelihood principles (stating the context clearly).

(b) Show that in discrete cases sufficiency and conditionality principles imply and are implied by the likelihood principle.

[10+ 15 =25]

... P.T.O

5 (a) In a multiple testing problem briefly discuss the notion of False discovery rate (FDR) and the Benjamini-Hochberg algorithm for thresholding. State the main theorem regarding the BH algorithm from the flagship article mentioned the class.

(b) Define the Pitman estimator for a sample $X_1, X_2, \ldots, X_n$ coming from a density with location parameter θ. Compute the Pitman estimator when the data are iid from Unif$(\theta - 1/2b, \theta + 1/2b)$, with b known. Show that Pitman estimator is Minimum Risk Equivariant (MRE) in this example.

[10 +10 =20]

INDIAN STATISTICAL INSTITUTE

Second Semestral Examination : 2015-2016

M. Stat. II Year

Advance Sample Survey

Date : 12/05/2016 **Time : 3 Hrs.**

Group A together with Group B

Group – A

Answers to Group A and Group B must be given in separate answer books.
Assignments Records to be submitted carry 5 marks.
Answer any 2 questions each carrying 10 marks

Maximum Marks : 25 for Group A

1 Show how Taylor Series method of linear expansion is useful in estimating (1) a finite population correlation coefficient between two variables and (2) a small domain total by synthetic generalized regression method, both from unequal probability samples, along with suitable methods of measuring the accuracy levels.

2 Postulating a suitable linear regression model through the origin derive, giving detailed proofs, (I) Godambe-Thompson's as well as (II) Brewer- Royall's optimal predictors for a finite population total.

3 Explain the role of 'Model Assisted Approach' in prediction of survey population totals illustrating the derivation of the Brewer's predictor. Show that this predictor is a case of a generalized regression predictor explaining how this helps in deriving a measure of its error.

4. Explain how it is useful to estimate the known population size while estimating a finite population simple correlation coefficient from a sample chosen with unequal probabilities. Explain a problem if any in estimating a finite population 'Spearman's Rank Correlation Coefficient' following the same method of estimating the finite population simple total correlation coefficient from an unequal probability sample.

****

INDIAN STATISTICAL INSTITUTE

M. STAT - II YEAR (2nd Semester Exam)

12.5.16

Group B (Total Marks = 25)

Advance Sample Survey

Answer all questions.

1. Discuss the situation where adaptive sampling technique is suitable and describe its working procedure in detail. Also, state what remedy may be applied to implement it within a limited budget constraint.

(10)

2. Discuss in detail the use of permanent random numbers in SRSWOR and in unequal probability sampling.

(10)

3. Discuss the method of Sitter's mirror-match bootstrap in SRSWOR.

(5)

INDIAN STATISTICAL INSTITUTE

Backpaper Examination

M. Stat./M. Math. II Year Semester II : 2015-2016

Stochastic Processes I

Date: 20/6/2016 **Total Marks : 100** **Time : 3 Hours**

1. (a) Show that for any real-valued function f on a metric space S and any real number t, $\partial\{f > t\} \subset D_f \cup \{f = t\}$. [Here D_f = set of discontinuity points of f.]

   (b) Using (a) or otherwise, show that if $\mathbf{P}_n \xrightarrow{w} \mathbf{P}$ on $(S, \mathcal{S})$, then for any bounded measurable function $f : S \to \mathbb{R}$ with $\mathbf{P}(D_f) = 0$, $\int f\, d\mathbf{P}_n \to \int f\, d\mathbf{P}$. [You may first do it for $0 \le f \le 1$ and then justify.]

   (c) Using (b) or otherwise, prove that if $\mathbf{P}_n \xrightarrow{w} \mathbf{P}$ on $(S, \mathcal{S})$, then for any measurable $h : S \to S'$ with $\mathbf{P}(D_h) = 0$, one has $\mathbf{P}_n h^{-1} \xrightarrow{w} \mathbf{P}h^{-1}$ on $(S', \mathcal{S}')$. (6+7+7)=**[20]**

2. Let $\mathcal{H}$ be an *equi-continuous* family of real-valued functions on a metric space S, meaning that for every $x \in S$ and $\epsilon > 0$, there exists $\delta > 0$ such that $y \in B(x, \delta) \implies |f(y) - f(x)| < \epsilon$ for <u>all</u> $f \in \mathcal{H}$. Assume that S is separable.

   (a) Show that if $\mathbf{P}$ is a probability on $(S, \mathcal{S})$, then for every $\epsilon > 0$, there exists a countable partition $\{A_{k,\epsilon}\}$ of S into $\mathbf{P}$-continuity sets such that for any k, one has $y, z \in A_{k,\epsilon} \implies |f(y) - f(z)| < \epsilon$ for all $f \in \mathcal{H}$. [First get a countable cover of S by $\mathbf{P}$-continuity sets with the same property.]

   (b) Using (a) or otherwise, show that if $\mathcal{H}$ is also uniformly bounded, then $\mathbf{P}_n \xrightarrow{w} \mathbf{P}$ on $(S, \mathcal{S})$ implies that $\int f\, d\mathbf{P}_n \longrightarrow \int f\, d\mathbf{P}$ uniformly in $f \in \mathcal{H}$. (7+7)=**[14]**

3. Let $\{\mathcal{A}_t, t \in [0, \infty)\}$ be a filtration on a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ and $\{X_t, t \in [0, \infty)\}$ be an $\{\mathcal{A}_t\}$-adapted real-valued stochastic process with right-continuous paths.

   (a) Show that (i) if τ is an $\{\mathcal{A}_t\}$-stopping time, then τ is $\mathcal{A}_\tau$-measurable, and (ii) if τ and η are two stopping times, then $\tau \wedge \eta$ is a stopping time and $\mathcal{A}_{\tau \wedge \eta} = \mathcal{A}_\tau \cap \mathcal{A}_\eta$.

   (b) Show that for $t \ge 0$, the map φ on $[0, t] \times \Omega$ defined as $\varphi(s, \omega) = X(s, \omega)$ is measurable with respect to $\mathcal{B}_t \otimes \mathcal{A}_t$. [Here, $\mathcal{B}_t$ = borel σ-field on $[0, t]$.]

   (c) Using the above (or otherwise), show that for any $\{\mathcal{A}_t\}$-stopping time τ, $X_{\tau \wedge t}$ is an $\mathcal{A}_t$-measurable random variable and hence deduce that if τ is a finite $\{\mathcal{A}_t\}$-stopping time, then X_τ is $\mathcal{A}_\tau$-measurable. ((3+(2+3))+7+7)=**[22]**

4. Let $\{B_t, t \in [0, \infty)\}$ be a Standard Brownian Motion on some probability space.

   (a) Fix $s \ge 0, T > 0$ and define $X_t = B_{s+t} - \frac{t}{T} B_{s+T} - (1 - \frac{t}{T}) B_s$, for $t \in [0, T]$. Show that $\{X_u, u \in [0, T]\}$ is a continuous-path gaussian process which is independent of $\{B_u, u \in [0, s] \cup [T, \infty)\}$ and find its covariance kernel.

   (b) Show that for any real α, the process $\{M_t = \exp(\alpha B_t - \frac{1}{2}\alpha^2 t),\ t \ge 0\}$ is a continuous martingale with respect to the natural filtration of $\{B_t\}$.

   (c) Use the martingale in (b) and an appropriate maximal inequality to prove that, with probablity one, $\frac{1}{t} B_t \to 0$ as $t \to \infty$. ((3+5+3)+4+6) = **[21]**

5. Let S be a complete separable metric space and $\Omega_c = \{\omega \mid \omega : [0, \infty) \to S,\ \omega \text{ continuous}\}$. Assuming usual notations, let the coordinate process $\{X_t\}$ on $(\Omega_c, \mathcal{F}, \mathcal{F}_t, \{P_x, x \in S\})$ be a Markov process satisfying the Feller property.

   (a) Clearly define what is meant by the generator A of the Markov process.

   (b) Show that if $f \in C_b(S)$, then so is $T_t f$ for every $t \ge 0$ and that $R_\lambda T_t f = T_t R_\lambda f$ for all $t \ge 0$ and $\lambda > 0$.

   (c) Show that if $f \in C_b(S)$ is in the domain of A, then so are $T_t f$ and $\int_0^t T_s f ds$ for all $t \ge 0$ and also that $A(T_t f) = T_t(Af)$ and $A(\int_0^t T_s f(x) ds) = \int_0^t A(T_s f)(x) ds$.

   (d) Show that for $f \in C_b(S)$ in the domain of A, $E_x[\int_0^t Af(X_s) ds] = T_t f(x) - f(x)$ for any $t > 0$, $x \in S$. Hence deduce that if $f \in C_b(S)$ is in the domain of A, then the process $\{M_t = f(X_t) - \int_0^t Af(X_s) ds,\ t \ge 0\}$ is, under any P_x, a continuous martingale with respect to the filtration $\{\mathcal{F}_t\}$. (2+(2+2)+(4+4)+(5+4))=**[23]**

INDIAN STATISTICAL INSTITUTE

Semestral Examination (Backpaper): Semester II (2015-16)

M. Stat. II Year

Survival Analysis

Date: 25.07.2016 | Maximum marks: 100 | Time: 3 hours

1. State if the following statements are **true** or **false** giving suitable reasons.

   (a) The hazard rate for life time T, following a mixture distribution $F(t) = \sum_{i=1}^{k} p_i F_i(t)$ with $\sum_{i=1}^{k} p_i = 1$, is $\lambda(t) = \sum_{i=1}^{k} p_i \lambda_i(t)$, where $\lambda_i(t)$ is the hazard rate corresponding to the ith component distribution $F_i(t)$.

   (b) Starting with n individuals subject to random right censoring, having independent exponential life time and censoring time distribution with mean 50 days and 75 days, respectively, the expected number of failures is $n \times 75/125$.

   (c) Under the Cox's proportional hazards model, when there are ties at some failure times in the context of random right censored data such that the individuals with tied failure times have the same covariate value, Efron's approximation to the Cox likelihood becomes exact.

   (d) In a competing risks problem with two failure types ($j = 1, 2$) and covariate z, the cause specific hazard rates are

   $$\lambda_j(t, z) = \lambda(t) e^{(j-1)\gamma + z\beta} ,$$

   where $\lambda(t)$ is unknown and arbitrary. The probability that an individual with covariate z will fail due to type j does not depend on z.

   (e) The discrete time regression model, obtained by grouping the lifetime following the Cox proportional hazards model, has proportional hazards.

   [$4 \times 5 = 20$]

2. (a) Derive the hazard rate for log-normal distribution.

   (b) Consider randomly right censored data from a Geometric distribution. Find the maximum likelihood estimator of the mean life time.

   [5+5=10]

3. Consider the two sample ($z = 0, 1$) problem in which the hazard is $\lambda(t, z) = \lambda_0(t) e^{z\beta}$, where $\lambda_0(t) = \lambda_i$ if $t \in I_i = [a_{i-1}, a_i)$, for $i = 1, \cdots, K$ with $0 = a_0 < a_1 < \cdots < a_{K-1} < a_K = \infty$. Develop a score test for homogeneity based on right censored life time data in both the groups. Obtain the corresponding variance estimate and give the rejection criterion. [12+8+2=22]

P.T.O.

4. Consider the two sample problem with survival functions $S_1(t)$ and $S_2(t)$, respectively, with $S_2(t) = S_1(\alpha t)$, for some $\alpha > 0$, where $S_1(\cdot)$ is unknown and arbitrary. Obtain a relationship between the hazard rates in the two populations and discuss in detail one method of estimating α based on random right censored data in each population.

[3+7=10]

5. Consider the K-sample problem with the model $\mathcal{M} : \lambda_i(t) = a_i \lambda_1(t)$, for $i = 2, \cdots, K$, where $\lambda_i(t)$ denotes the hazard for the ith population, $i = 1, \cdots, K$, $\lambda_1(t)$ is totally unknown and arbitrary, and $a_i > 0$, for $i = 1, \cdots, K$. There is random right censored data from each population. Suggest a graphical test for the model $\mathcal{M}$.. Write down an appropriate partial likelihood and, hence, develop a score test for homogeneity (Give only the test statistic; derivation of variance is not needed). If $\lambda_1(t)$ is a known function involving a constant parameter θ, indicate how to test for homogeneity. [3+6+4+4=17]

6. (a) Consider random right censored data from the accelerated failure time model $Y = \log T = \alpha + z\beta + \epsilon$ with extreme-value error distribution. Develop the corresponding parametric score test for significance of the regression coefficient (Give only the test statistics). Compare with the corresponding linear rank test and comment.

   (b) Consider the competing risks model given by the cause-specific hazard rates $\lambda_j(t, z) = \lambda_j e^{z\beta}$, for $j = 1, \cdots, m$, where z is the covariate value (scalar). Develop a score test for $\mathcal{H}_0 : \beta = 0$ based on competing risks data (Give only the test statistic). Compare with the same in (a) above and comment.

   (c) Consider modeling dependence of a multiple decrement function on a covariate z by
$$Q(t_1, \cdots, t_m; z) = (Q(t_1, \cdots, t_m))^{g(z\beta)} ,$$
   where $Q(t_1, \cdots, t_m)$ denotes the 'baseline multiple decrement function' for $z = 0$ and $g(\cdot)$ is a one-to-one function. Write down the corresponding cause-specific hazard rates for an individual with covariate z. For unknown and arbitrary $Q(t_1, \cdots, t_m)$, construct an appropriate partial likelihood to estimate β based on right censored competing risks data with covariates.

[(6+2)+(5+2)+(2+4)=21]