2

Mid- semestral Examination
Advanced Design of Experi ments
M.Stat- 2nd Year : 2011-12

Full Marks: 70 Time : 2 hrs Date: 1.09.2011

Answer all questions.

Q1. Define A- , D- and E- optimality criteria, along with their statistical interpretations for the non singular inferential problem $\pi : \underline{\eta} = L\underline{\tau}$. (4+6+5=15)

Q2. State and prove sufficient conditions for a design to be A-, D- and E- optimal for estimating a full set of orthonormal contrasts of treatment effects. Hence show that for given v, b, and k, a BIBD, whenever exists is A-, D- and E- optimal for the above inferential problem. (9+9= 18)

Q3. Consider the following 4x4 row-column design d with treatment symbols A, B, C, D . Check the estimability of a full set of independent elementary treatment contrasts, providing one unbiased estimator, whenever exists ,for each such contrast.

d:

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	D

(12)

Q4. Construct a GDD with parameters v= 12, b= 9, k=4, r=3. Characterize the category of this GDD, with proper justification. (8+2=10)

Q5. Define an OA(N, k, s, t) and construct OA(16,5,4,2). (3+12=15)

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination: 2011-12 (First Semester)
Master of Statistics (M. Stat.) II Year
Advanced Probability I

Instructor: Antar Bandyopadhyay and Parthanil Roy

Date: 02/09/2011 **Maximum Marks: 30** **Duration: 10:00 am - 12:30 am**

Note:

- Please write your name and roll number on top of your answer paper.
- There are five problems each carrying 8 points with a total of 40 points. Solve as many as you can. Show all your works and write explanations when needed. Maximum you can score is 30 points.
- This is an open note examination. You are allowed to use your own hand written notes (such as class notes, your homework solutions, list of theorems, formulas etc). Please note that no printed materials or photo copies are allowed, in particular you are not allowed to use books, photocopied class notes etc.

1. Let $(\Omega, \mathcal{F})$ be a measurable space and μ be a signed measure on it. Show that there exist a probability measure $\mathbf{P}$ on $(\Omega, \mathcal{F})$ and an integrable random variable X such that for any $A \in \mathcal{F}$,
$$\mu(A) = \mathbf{E}[X\mathbf{1}_A].$$
[8]

2. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $\mathcal{G}$ and $\mathcal{H}$ be two sub-σ-algebras. Let $X : \Omega \to \mathbb{R}$ be a random variable with finite second moment. Suppose X is independent of $\mathcal{G}$ but $\mathbf{E}\left[X \middle| \mathcal{H}\right]$ is $\mathcal{G}$-measurable. Then show that $\mathbf{E}\left[X \middle| \mathcal{H}\right] = \mathbf{E}[X]$ a.s. [8]

3. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be an integrable random variable.
[4 + 4 = 8]

   (a) For $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra show that
   $$\mathbf{E}\left[X \middle| \mathcal{G}\right] = \int_0^\infty \mathbf{P}\left(X > t \middle| \mathcal{G}\right) dt,$$
   a.s with respect to $[\mathbf{P}]$ provided X is a.s. non-negative.

   (b) Let $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$ be an increasing sequence of sub-σ-algebras. Assume X is square integrable then show that
   $$\mathbf{E}\left[\mathbf{Var}\left(X \middle| \mathcal{G}\right)\right] \leq \mathbf{E}\left[\mathbf{Var}\left(X \middle| \mathcal{H}\right)\right].$$

[P.T.O.]

4. **Consider the probability space $\left([0,1]^2, \mathcal{B}_{[0,1]^2}, \lambda_2\right)$ where λ_2 is the *Lebesgue measure (area measure)* on $[0,1]^2$. Let $M : [0,1]^2 \to [0,1]$ be the function $M(x,y) = \max(x,y)$. Find a *regular conditional probability (RCP)* given the random variable M.** [8]

5. **Let T be an uncountable index set and $\mathcal{F} := \otimes_{t\in T}\mathcal{B}$ be the product σ-field on $\mathbb{R}^T$ (here $\mathcal{B}$ denotes the Borel σ-field on $\mathbb{R}$). Show that if $B \in \mathcal{F}$, then there exists a countable set $T_0 \subset T$ and a set $B_0 \in \otimes_{t\in T_0}\mathcal{B}$ such that $\omega \in B$ if and only if $(\omega(t),\, t \in T_0) \in B_0$.** [8]

Wish you all the best

Date: 2.9.2011
Time: 2 hours

INDIAN STATISTICAL INSTITUTE

Statistical Methods in Genetics – I
M-Stat (2nd Year) 2011-2012
Mid-Semester Examination

This paper carries 40 marks.

1. Consider the following genotype data at a triallelic locus on 500 randomly chosen individuals in a population:

Genotype	Frequency
AA	182
AB	170
AC	66
BB	48
BC	34

Test whether the locus is in Hardy-Weinberg Equilibrium. [8]

2. Consider a recessive disorder controlled by an autosomal biallelic locus. If an individual is affected, determine, in each of the following pairs of relationship with the individual, who is more likely to be affected:

(a) father and sib
(b) grandfather and uncle [5 + 5]

3. Consider data on genotypes at a biallelic locus for a random set of individuals in an inbred population. Obtain the m.l.e.s of the allele frequencies and the inbreeding coefficient. Show that the variances of the m.l.e.s of the allele frequencies increase with increase in the inbreeding coefficient. [10]

4. Conder a X-linked biallelic locus with alleles (A,a). Suppose the selection coefficients corresponding to the allele A is s and that for the allele a is t where $s \neq t$. The fitness coefficients of the genotypes follow a multiplicative model (i.e., it is s^2 for AA, st for Aa, etc.). Examine whether the allele frequencies reach non-trivial equilibrium values separately for males and females. [12]

INDIAN STATISTICAL INSTITUTE

M. STAT. SECOND YEAR

Topology and Set Theory

Date: 05.09.11 Mid. Semestral Examination Time : 3 hrs.

This paper carries 70 marks.The maximum you can score is 60.

1. Let X, Y be Hausdorff spaces. If $f, g : X \to Y$ are continuous functions, show that $\{x : f(x) = g(x)\}$ is a closed subset of X. [8]

2. Let $A \subset X$ where X is a topological space.If A' is the set of limit points of X, show that A' is closed [8]

3. Show that a one to one continuous function from a compact Hausdorff space X onto a Hausdorff space Y is a homeomorphism. Show this is not neccessarily true if X is not compact. [10]

4. Let X be a completely regular space , $K \subset X$ compact and $A \subset X$ closed. Show that there is a continuous function f from X into $[0, 1]$ such that $f(x) = 0$ if $x \in K$ and 1 if $x \in A$. [14]

5. Let $\mathcal{B}$ be the family of subsets of the form $\{[a, b) : a < b\}$ of the reals.

   (a)Show that $\mathcal{B}$ is a base for some topology on the reals. [2]

   (b)What are the compact sets in this topology? [14]

   (c)What are the connected sets in this topology? [14]

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination: 2011-12 (First Semester)
M. Stat. II Year

STATISTICAL COMPUTING

Date: 7 September 2011 Maximum Marks: 50 Duration: 2 hr

1.

a. Explain how a generator of exponential random variables can be used to simulate from the gamma distribution with probability density function

$$\mathrm{Ga}(\alpha,\beta)=\frac{\beta^{\alpha}}{\Gamma\alpha}x^{\alpha-1}e^{-\beta x},\ x>0,$$

when α is known to be a positive integer.

b. Consider the following two-parameter Bayesian model from which n observations $y_1, y_2, \ldots y_n$ are available:

$$\begin{aligned} y_i &\sim \mathrm{N}(\mu,\tau^{-1}),\\ \mu &\sim \mathrm{N}(0,1),\\ \tau &\sim \mathrm{Ga}(2,1), \end{aligned}$$

where $\mathrm{Ga}(a,b)$ is as in part (a) of this question, and $\mathrm{N}(a,b)$ denotes a normal distribution with mean a and variance b. It is assumed that $y_1, y_2, \ldots y_n$ are conditionally independent given μ and τ, and μ and τ are themselves independent.

i. Show that the full conditionals of μ and τ given $y_1, y_2, \ldots y_n$ are respectively the normal and gamma distributions with parameters to be determined by you.

ii. Given generators of the standard normal and exponential random variables, set up the Gibbs sampler for simulating from the posterior distribution of (μ,τ) given $y_1, y_2, \ldots, y_n$, when n is assumed to be an even integer.

[5+(5+5)=15]

(Please Turn Over)

2. Consider the problem of obtaining Monte Carlo estimates of

$$\theta = \int_0^1 e^{x^2} dx.$$

a. What is the raw simulation estimate of θ, based on 2 independent realizations U_1 and U_2 of the random variable U distributed uniformly over (0,1)? What is the variance of this estimate?

b. If, in place of U_2, you use $1-U_1$ in the estimate you have proposed in part (a) of this problem, what will be the variance of the resulting estimate? Show that it is smaller than the variance of the raw estimate. What is this approach to variance reduction in Monte Carlo simulation referred to as?

c. Use any other variance reduction technique to the Monte Carlo estimation of θ, and establish that variance is indeed reduced as a result.

d. Is it possible to combine the variance reduction techniques in (b) and (c) to arrive at another estimate with an even smaller variance? Justify your answer.

[(2+3)+(3+1+1)+2+3=15]

3. Sixteen mice were randomly assigned to a treatment group or a control group, and those in the former group received an experimental treatment that was intended to prolong survival. The survival time in days of mice in both groups following the treatment is given below:

Group	Size	Data
Treatment	7	94 197 16 38 99 141 23
Control	9	52 104 146 10 50 31 40 27 46

Test whether the treatment is successful in prolonging survival, by computing the achieved significance level of a permutation test based on 5 permutation replications of an appropriate test statistic. Specify all relevant details clearly.

[10]

4. Suppose that the observed data vector of frequencies $y = (y_1, y_2, y_3, y_4, y_5)'$ with $\sum_{i=1}^{5} y_i = n$, where n is fixed, is assumed to arise from a multinomial distribution with five cells and corresponding cell probabilities $\frac{1}{2}, \frac{1}{4}\theta, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta)$ and $\frac{1}{4}\theta$, where θ is to be estimated on the basis of y. If the observation y_1 is missing, then show how the EM algorithm can be used to obtain the maximum likelihood estimate of θ, clearly specifying the expectation and maximization steps, and final iterates for the estimate.

[10]

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination: (2011-2012)

MS(QE) I & MSTAT II

Microeconomic Theory I

Date: 07.09.2011 **Maximum Marks:** 100 **Duration:** $3\frac{1}{2}$ hrs.

Note: Answer all questions.

(1) Show that if R on X is rational, then we have the following:
   (a) P is both irreflexive and transitive.
   (b) I is reflexive, transitive and symmetric.
   (c) If $xPyRz$, then xPz. Similarly, if $xRyPz$, then xPz.
   (4+6+4=14)

(2) Consider a rational preference relation R on X. Show that if $u(x) = u(y)$ implies that xIy and $u(x) > u(y)$ implies that xPy, then $u(.)$ is a utility function representing R. **(6)**

(3) Suppose that X is finite and R is a rational preference defined on X. Consider the function $u^* : X \to \Re$ such that $\forall\, x \in X$, $u^*(x) = |X| - |\{z \in X \mid zPx\}|$. Is the function $u^*(.)$ a valid utility representation of the preference relation R on X? Justify your answer. **(10)**

(4) Show that a choice structure $(\mathcal{B}, C(.))$ for which a rationalizing preference relation exists, satisfies the path-invariance property: For every pair $B_1, B_2 \in \mathcal{B}$ such that $B_1 \cup B_2 \in \mathcal{B}$ and $C(B_1) \cup C(B_2) \in \mathcal{B}$, we have $C(B_1 \cup B_2) = C(C(B_1) \cup C(B_2))$. **(18)**

(5) Define the weak axiom of revealed preference for the market economy. Show that if the Walrasian demand function $x(p, w)$ is homogeneous of degree zero and satisfies Walras' law, then the weak axiom of revealed preference holds if and only if it holds for all compensated price changes. **(2+16=18)**

(6) Show that if the Walrasian demand function $x(p, w)$ is generated by a rational preference relation, then it must satisfy the weak axiom of revealed preference. **(10)**

(7) If $u(.)$ is a continuous utility function representing R on X, then show that R must be continuous. **(9)**

(8) Let R be a preference relation defined on X and let $u(.)$ be a utility function representing it. Define convexity of R and quasi-concavity of $u(.)$. Show that R is convex if and only if the utility function $u(.)$ representing it is quasi-concave. **(1+2+12=15)**

INDIAN STATISTICAL INSTITUTE

Mid-semester Examination : (2011-2012)

M.Stat. 2nd Year

TOPICS IN BAYESIAN INFERENCE

Date: 9 September, 2011 Max. Marks: 90 Duration: $2\frac{1}{2}$ Hours

Answer as many questions as you can. Maximum you can score is 90.

1. What is a conjugate prior? Give an example to show that a conjugate prior can be interpreted as additional data.

[5]

2. Describe how the posterior distribution can be used for estimation of a real parameter. How do you measure the accuracy of an estimate?

[6]

3. Consider i.i.d. observations $X_1, \ldots, X_n \sim f(x|\theta) = \exp\{c(\theta) + \theta t(x)\}h(x)$ (one parameter exponential family). It is given that the usual regularity conditions hold and $c(\theta)$ is sufficiently smooth. A statistically natural parameter is $\mu = E_\theta t(X)$. Show that μ is a one-one function of θ. Find MLE of μ. Show that for a conjugate prior $\pi(\theta) = c\exp\{mc(\theta) + \theta s\}$, the posterior mean of μ is a weighted average of prior estimate and classical estimate (MLE). (assume $\pi(a) = \pi(b) = 0$.)

[13]

4. (a) Define a highest posterior density (HPD) credible region for an unknown parameter.

(b) Consider i.i.d. observations with a common distribution involving an unknown real parameter θ. Assuming that the usual regularity conditions hold, find a large sample approximation to a $100(1-\alpha)$ % HPD credible interval for θ.

[3+5=8]

5. Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta, \sigma^2)$ variables.

(a) Consider a standard noninformative prior for (θ, σ^2) and find the posterior distribution of θ. Also find the $100(1-\alpha)\%$ HPD credible set for θ.

(b) Assume that σ^2 is known and consider a conjugate prior for θ. Find the posterior distribution of θ and the posterior predictive distribution of a future observation X_{n+1}.

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

6. (a) Let $X_1, \ldots, X_n$ be i.i.d. with a common density $f(x|\theta)$ where $\theta \in R$. State the result on asymptotic normality of posterior distribution of suitably normalized and centered θ under suitable conditions on the density $f(\cdot|\theta)$ and the prior distribution. Prove only the part for the tail of the posterior distribution (without the normalizer).

(c) Use a stronger version of the result on asymptotic normality of posterior (to be stated by you) to prove that $\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n) \to 0$ with probability one, where $\tilde{\theta}_n$ and $\hat{\theta}_n$ denote respectively the posterior mean and MLE.

[17+7=24]

7. Show that the result on asymptotic normality of the posterior distribution of $\sqrt{n}(\theta - \hat{\theta}_n)$, proved in the class, implies consistency of the posterior distribution of θ at θ_0 [10]

8. Suppose X has density $e^{-(x-\theta)}I(x > \theta)$ and the prior density of θ is $\pi(\theta) = [\pi(1+\theta^2)]^{-1}$. Consider a loss function $L(\theta, a) = I(|\theta - a| > \delta)$ and find the Bayes estimate of θ.

[11]

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination : Semester I (2011-12)

M. Stat. II Year

Actuarial Methods

Date: 09.09.2011 Maximum marks: 50 Time: 2 hours

This test is open notes. Calculator and RMM table can be used. Books cannot be used and notes cannot be exchanged. Refer to your notes properly, but do not reproduce derivations from there. Answer as many as you can. Total mark is 53.

1. Two taxi drivers, A and B, are waiting at the bus stand for two regular passengers, C and D, who will arrive by different buses. The amount which a taxi driver can earn by driving passenger C to his home is Rs 100, while the corresponding amount for passenger D is Rs 200. The bus carrying passenger C is scheduled to arrive at 6:00 PM, but it may be delayed by X minutes following an *exponential* distribution with mean 30. The bus carrying passenger D is scheduled to arrive at 6:15 PM, but it may be delayed by Y minutes following an *exponential* distribution with mean 15. The two taxi drivers will have to serve the two passengers, and there is no other passenger or taxi driver. Driver B gives driver A the choice of serving the passenger who arrives first, or waiting for the passenger arriving second, but insists that driver A makes his choice before 6:00 PM.

   (a) Describe the decision problem of driver A as a game clearly identifying the nature's choices.

   (b) Write down the loss matrix for driver A.

   (c) Calculate the probabilities of the events corresponding to nature's choices.

   (d) What is the Bayes solution to the decision problem of driver A?

   [2+2+3+3=10]

2. Ten IID observation from a $Poisson(\lambda)$ distribution are 3, 4, 3, 1, 5, 5, 2, 3, 3, 2. Assuming an $Exp(0.2)$ prior distribution for λ, find the Bayes estimator of λ under squared error loss. Express this estimator as an weighted average of the sample mean and the prior mean. [5+1=6]

3. (a) If claim amounts follow a $N(500, 400)$ distribution and there is a retention limit of $M = 550$, find the mean amount paid by the re-insurer on all claims.

   (b) An insurer has the following two options for reinsurance arrangement: (1) a quota share arrangement with 75% retained proportion, and (2) an XOL arrangement with retention limit Rs 25,000/-. Assuming that the claim amounts follow a lognormal distribution with parameters $\mu = 8.5$ and $\sigma^2 = 0.64$, find the mean of the amount to be paid by the insurer on a claim (a) without reinsurance, (b) with option 1, and (c) with option 2.

P.T.O.

[Note: For a lognormal distribution with parameters μ and σ^2, given by the density $f(x)$,

$$\int_L^U x^k f(x)dx = e^{k\mu+\frac{1}{2}k^2\sigma^2}\left[\Phi(u_k) - \Phi(l_k)\right],$$

where $l_k = \frac{\log L-\mu}{\sigma} - k\sigma$ and $u_k = \frac{\log U-\mu}{\sigma} - k\sigma$, and Φ is the standard normal cdf.]

[5 + (2 + 3 + 3) = 13]

4. Consider an XOL reinsurance arrangement with retention limit M (fixed) and annual inflation factor k on the claim amount.

   (a) Will the mean insurer's pay-off be inflated by the factor k. Justify your answer.

   (b) If the original claim amount follows a Pareto distribution, derive the distribution of the claim amount after inflation (and before re-insurance).

   (c) Assuming Pareto claim distribution, derive the mean insurer's pay-off before and after inflation.

[3+3+(7+1)=14]

5. The annual aggregate claim amount from a risk has a compound Poisson distribution with Poisson parameter 10. Individual claim amounts are uniformly distributed on $(0, 2000)$. The insurer of this risk has effected an excess of loss reinsurance with retention level 1600. Calculate the mean and variance of both the insurer's and the reinsurer's aggregate claim amounts under this reinsurance arrangement. Derive the same measures for the aggregate claim before reinsurance and check for their additive property with comments. [3+5+2=10]

INDIAN STATISTICAL INSTITUTE

Mid-semestral exam. (Semester I: 2011-2012)

Course Name: M. Stat. 2nd year

Subject Name: Analysis of discrete data

Date: ~~29.08.11~~ 12.09.11 , Maximum Marks: 35. Duration: 1 hr. 30 min.

1. Derive the joint asymptotic distribution of log odds ratios in a 2×3 contingency table. Consequently find 95% confidence interval for log odds ratio for the following table.

	Myocardial Infraction	
	Yes	No
Placebo	28	656
Aspirin	18	658

[10+4]

2. Give the geometric interpretation of $\theta = 8$, where θ is the odds ratio in a 2×2 table. [10]

3. Carry out Fisher's exact test for independence against both one-sided and two-sided alternatives for the 2×2 table. Also carry out unconditional tests of independence.

3	0
0	2

[6+5]

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination : 2011-12

M.Stat. II Year

Life Contingencies

Date: 29.09.2011 Full Marks : 50 Duration : 2 Hours

(Attempt all questions. Allotted marks are shown in brackets.)

1. Consider the *stop loss* type insurance contract for the loss random variable x with p.d.f. f:

$$I_d(x) = \begin{cases} 0 & x < d \\ x - d & x \geq d \end{cases}$$

with expected claim = $\int_d^\infty (x-d)f(x)dx = \beta$. Show that given any β, d exists and is unique.

[3 + 3 = 6]

2. With the usual notation, for an *annually increasing whole life insurance payable at the end of year of death*, show that

$$(IA)_x = [vq_x + vp_x A_{x+1}] + vp_x (IA)_{x+1}$$

with usual initial values. [6]

3. With the usual notation, for a *whole life insurance payable at the moment of death*, establish the following differential equation

$$\frac{d}{dx}\overline{A_x} = -\mu(x) + \overline{A_x}[\delta + \mu(x)]$$

and provide a solution with usual initial values. [3 + 4 = 7]

4. With the usual notation, for a *n-year deferred whole life annuity due*, establish the variance formula

$$Var = \frac{2}{d} v^{2n}\, {}_np_x(\ddot{a}_{x+n} - {}^2\ddot{a}_{x+n}) + {}^2_{n|}\ddot{a}_x - ({}_{n|}\ddot{a}_x)^2$$

[6]

5. Write short notes on the following.

   i) Force of mortality
   ii) Central rate of mortality
   iii) Curtate expectation of life

[3+3+4 = 10]

P.T.O

6. In a certain population, the force of mortality equals 0.02 at all ages.

i) Find out the probability that a life aged exactly 8 will die before age 10.
ii) Find out the probability that a life aged exactly 5 will die between ages 8 and 1
iii) Find out complete expectation of the life of a new born baby.
iv) Find out the curtate expectation of life of a new born baby.

[2x4

7. i) Show algebraically that $e_x = p_x (1 + e_{x+1})$.

ii) What does the UDD assumption implicitly assume about μ_x over the year of age?

iii) Prove that ${}_{t-s}q_{x+s} = \{(t-s)\, q_x\}/(1-s.q_x)$

iv) Show that under the UDD assumption over each year of age

$l_{x+t} = (1 - t).l_x - t.l_{x+1}$ for $x = 0, 1, 2, \ldots\ldots$and for $0 \leq t < 1$.

[2 + 1 + 2 + 2=

< END >

INDIAN STATISTICAL INSTITUTE
Mid-Semester Examination: 2011-12 (First Semester)

M. STAT. II YEAR
Mathematical Statistics and Probability (MSP) Specialization
Functional Analysis

OPEN NOTES CLOSED BOOK EXAMINATION
(Handwritten notes are allowed but any printed or photocopied material is not allowed)

Date : 13.09.2011

Duration : $2\frac{1}{2}$ Hours
Total Marks : 40
Maximum Marks : 30

χ_E denotes the characteristic function of a set E.
$\mathcal{B}(X)$ denotes the set of all bounded linear operators on X.
$KerT$ denotes the kernel of a linear operator T.
X^* denotes the space of all bounded linear functionals on X.

1. For $1 \leq p < \infty$, consider $E_p = \left\{ f \in L^p[0,\infty) : \int_0^\infty f(x)dx = 0 \right\}$. Prove that E_p is closed in $L^p[0,\infty)$ if and only if $p = 1$. [5]

2. Let $A_n = \left[\frac{1}{2^n}, \frac{1}{2^{n-1}}\right)$ for $n \in \mathbb{N}$. For $x = (x_1, x_2, \cdots) \in l^\infty$ define an operator $U_x : L^p[0,1] \to L^p[0,1]$ by $U_x(f) = \sum_{n=1}^{\infty} x_n \chi_{A_n} f$. Prove that $U_x \in \mathcal{B}(L^p[0,1])$. Define a map $j : l_\infty \to \mathcal{B}(L^p[0,1])$ by $j(x) = U_x$. Prove that j is an isometry. Hence or otherwise prove that for $1 \leq p < \infty$, the space $\mathcal{B}(L^p[0,1])$ is not separable. [8]

3. Let X and Y be two normed linear spaces and $T : X \to Y$ be a linear continuous and surjective operator. Prove that the operator $\tilde{T} : X/KerT \to Y$ given by $\tilde{T}(x + KerT) = Tx$ for $x \in X$ is a well-defined bijective linear and continuous operator and $\|T\| = \|\tilde{T}\|$. [5]

4. Let X be a Banach space with Schauder basis $\{x_n\}_{n\in\mathbb{N}}$. Let $\{a_n(x)\}_{n\in\mathbb{N}}$ be the coefficients of x in this basis i.e. $x = \sum_{n=1}^{\infty} a_n(x)x_n$. Show that each a_n is a bounded linear functional on X. [7]

[Hint: Consider the space Y consisting of all sequences $a = (a_1, a_2, \cdots)$ for which the series $\sum_{n=1}^{\infty} a_n x_n$ converges in X. Define the norm of such a sequence as $\|a\| = \sup_{n\in\mathbb{N}} \left\| \sum_{j=1}^{n} a_j x_j \right\|$. Show that Y is a Banach space with this norm. Consider the operator $T : Y \to X$ given by $T(a) = \sum_{n=1}^{\infty} a_n x_n$.]

5. Let X be a Banach space and $f_n \in X^*$ for $n \in \mathbb{N}$. Prove that the following are equivalent :

   i. $\sup_{n\in\mathbb{N}} \|f_n\| < \infty$.

   ii. If $\sum_{n\in\mathbb{N}} \|x_n\| < \infty$ then $\sup_{n\in\mathbb{N}} \left| \sum_{k=1}^{n} f_k(x_k) \right| < \infty$. [5]

   [Hint : Define $l_1(X) = \left\{ x = (x_n)_{n\in\mathbb{N}} \text{ where each } x_n \in X : \sum_{n\in\mathbb{N}} \|x_n\| < \infty \right\}$. Use the fact that (you need not prove it) $l_1(X)$ forms a Banach space with the norm $\|x\| = \sum_{n=1}^{\infty} \|x_n\|$ for $x = (x_n)_{n\in\mathbb{N}} \in l^1(X)$.]

6. Show that a Banach space X is reflexive if and only if its dual space X^* is reflexive. [5]

7. Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of scalars. Define $f_n \in l_\infty^*$ by $f_n(x_1, x_2, \cdots) = a_1x_1 + a_2 + \cdots + a_n x_n$ for $x = (x_n)_{n\in\mathbb{N}} \in l^\infty$. Prove that the sequence $(f_n)_{n\in\mathbb{N}}$ is *weak** convergent if and only if $(a_n)_{n\in\mathbb{N}} \in l_1$. [5]

Good luck!

INDIAN STATISTICAL INSTITUTE

First Semestral Examination: 2011-12

M.Stat. II Year

Life Contingencies

Date: 14/11/11 Maximum Marks: 100 Duration: 3 Hours

(Answer all questions. You may use Actuarial Tables.)

1. In the context of *term insurance payable at the end of year of death*, with usual notation, prove the following recursion relation

$$A_{x:\overline{(y-x)}|} = vq_x + vp_x A_{x+1:\overline{(y-(x+1))}|}$$

with usual initial values. Here $A_{x:\overline{(y-x)}|}$ denotes the value of a y year endowment insurance contract starting at age x. [8]

2. How would you evaluate a *3-year deferred 10-year annually decreasing term life insurance?* – Explain with usual notation. [8]

3. What is the difference between *annuity-due* and *annuity-immediate*?
Establish a relationship between *annually* payable and *monthly* payable annuities.
[3 + 8 = 11]

4. What is *apportionable premium*?
In the context of fully discrete annual benefit premiums, how would you adjust for apportionable premium? Clearly state the assumptions you are making. [3 + 8 = 11]

5. Define *joint life* status and *last survivor* status.
What is the covariance between these two status? [3 + 8 = 11]

6. What is the usefulness of a *Copula*? Taking any example of a common Copula, show how the correlation between the different life times is established. [4 + 6 = 10]

7. In the context of actuarial present values of *contingent insurances*, show that

$$A^{1}_{xy} - A_{x\overset{2}{y}} = A_{xy} - A_{y}.$$

Here A^{1}_{xy} denotes the value of contingent insurance contract where (x) dies before (y). $A_{x\overset{2}{y}}$ is similarly defined. [8]

8. (a) Calculate the complete and the curtate expectation of life for an animal subject to a constant force of mortality of 0.05 per annum.

(b) The following table is part of a mortality table used by a life company to calculate survival probabilities for a special type of life insurance policy, where $l_{[x]}$ is expected number of lives who survive to age x with [x] denoting the age of selection.

x	$l_{[x]}$	$l_{[x]+1}$	$l_{[x]+2}$	$l_{[x]+3}$	l_{x+4}
51	1,537	1,517	1,502	1,492	1,483
52	1,532	1,512	1,497	1,487	1,477
53	1,525	1,505	1,490	1,480	1,470
54	1,517	1,499	1,484	1,474	1,462
55	1,512	1,492	1,477	1,467	1,453

(i) Calculate the probability that a policyholder who was accepted for insurance exactly two years ago and is now aged exactly 55 will die at age exactly 57.

(ii) Calculate the corresponding probability for an individual of the same age who has been a policyholder for many years.

(iii) Comment on your answers (i) and (ii).

[5+3+2+2 = 12]

9. A husband and his wife, aged 64 and 60 years respectively, take out a policy under which the benefits are

a) A lump sum of Rs.50,000 payable at the end of the year of the first death provided this occurs within 10 years,

b) An annuity payable annually in advance with the first payment being made 10 years from the date of issue. The annuity will be of Rs.10,000 per annum so long as both husband and wife are still alive or Rs.6,000 while one of them is alive.

Level premiums are payable annually in advance for at most 10 years and will cease on the first death if this occurs earlier. Calculate the amount of the annual premium on the following basis.

Interest: 4% per annum; Mortality: a (55) Males Ultimate; a (55) Females Ultimate.

[8]

10. The premium for a five year combined sickness and death benefit policy is to be calculated using a multiple state model with three states H (healthy), S (sick) and D (dead). At the end of each of the five years, the policy provides a payment of Rs.3,000 if the life is sick at that time, and a premium of Rs.6,000 if the life is dead. No benefits are paid to healthy lives. A level premium is payable at the start of each policy year provided the life is in good health at the time the payment is due. A policy holder starts in state H.

The transition probabilities (which operate independently in each year) are as follows:

P (D at time t+1 | H at time t) = 0.01

P (S at time t+1 | H at time t) = 0.05

P (D at time t+1 | S at time t) = 0.05

P (H at time t+1 | S at time t) = 0.80

(i) Calculate the probabilities that the life in each of the states at the end of policy year t, t = 1,2,....,5.

(ii) Calculate the net premium for a policy issued to a healthy life, assuming an interest rate of 4% per annum.

[6+4 = 10]

11. In a study published by a medical research team which recorded the ages at death of cricketers whose dates of birth and death were known, it was found that the average age at death of left-handed players was six years less than for right-handed players. A statistician refuted the conclusion by pointing out that the effects of selection had been ignored. Which of the following forms of selection would be most likely to explain the misleading impression given by the data? Explain.

(a) Class selection, (b) Self selection, (c) Time selection and (d) Spurious selection.

[3]

< END >

INDIAN STATISTICAL INSTITUTE
Semesteral Examination : 2011-12 (First Semester)

M. Stat. II Year

STATISTICAL COMPUTING

Date: November 14, 2011 Maximum Marks: 100 Duration: 3 hr

Answer as many of the following questions as you can. The maximum you can score is 100.

1. (a) Write down the algorithm for generating a non-homogeneous Poisson process with intensity function

$$\lambda(t) = \begin{cases} \frac{t}{5}, & 0 < t < 5, \\ 1 + 5(t-5), & 5 < t < 10. \end{cases},$$

during the first 10 time units.

(b) Give the specific algorithm for generating, by the rejection method, a realization of a random variable having the the probability density function

$$f(x) = 20x(1-x)^3, \ 0 < x < 1,$$

using the uniform distribution over the interval (0,1) as envelope. What is the expected number of iterations required to get a single realization from $f(x)$ by this method?

(c) Describe succinctly the adaptive rejection sampling scheme for simulating from a log-concave probability density function.

[6+(5+3)+6=20]

2. (a) Let X and Y be random variables jointly distributed in the bivariate normal form with parameters

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

where μ_X, μ_Y and ρ are unknown. 10 independent observations (x_i, y_i), $i = 1, 2, \ldots, 10$ were taken on (X, Y), but scrutiny revealed that that x_6, x_7, y_8, y_9 and y_{10} were corrupted and had to be discarded. Formulate an iterative procedure by the Expectation-Maximization (EM) algorithm for estimating μ_X, μ_Y and ρ, giving all relevant details.

(b) For the setup described in part (a) above, suppose now that only y_{10} was found to be dubious after scrutiny, and was discarded. Treating μ_X, μ_Y and ρ as independent random variables and assuming that $\mu_X \sim N(5,1)$, $\mu_Y \sim N(0,4)$ and $\rho \sim Beta(2,2)$, explain how Prob$(X+Y > 5)$ can be estimated through multiple imputation by data augmentation, mentioning clearly all the steps involved.

[10+10=20]

(Please turn over)

3. (a) For $i = 1, 2, \ldots, 112$, let X_i be a random variable representing the number of coal-mining disasters per year in the year $1850 + i$. The X_i's are assumed to be independently distributed as

$$p_i(x) = \begin{cases} \text{Poisson}(\lambda_1), & i = 1, 2, \ldots, \theta, \\ \text{Poisson}(\lambda_2), & i = \theta + 1, \ldots, 112. \end{cases}$$

Assume $\lambda_i|\alpha \sim \text{Gamma}(3, \alpha)$, $i = 1, 2$, where $\alpha \sim \text{Gamma}(10, 10)$, and also assume that θ has a discrete uniform distribution over $\{1, 2, \ldots, 112\}$.

i. Derive the conditional distributions required to implement the Gibbs sampler for simulating from the joint distribution of λ_1, λ_2 and θ.

ii. Write down the explicit algorithm for implementing the Gibbs sampler for this problem.

iii. Describe briefly any method for assessing the convergence of the sampler.

[9+6+5=20]

4. (a) Let $\{1, 2, 6\}$ be a random sample from a distribution with mean θ.

i. Determine the exact probability distribution of the usual estimate of θ, that is, the sample mean, obtained from all possible (3^3) bootstrap samples.

ii. Obtain a simple one-sided bootstrap confidence interval for θ having confidence coefficient as close to 0.95 as possible.

(b) Describe a typical projection pursuit regression model, and explain how such a model can be fit to data.

(c) Explain the general principle behind the Supersmoother.

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

5. (a) In the context of optimization clustering, what are sihouette plots? Explain how they can be used for finding the optimal number of clusters in a data set.

(b) Consider a set of $n = 2k + 1$ observations on a random variable X, in which k observations are equal to -2, another k are equal to 0, and the remaining observation is equal to $a(> 0)$.

i. Show that the two-cluster partitioning which minimizes the sum-of-squared-errors criterion groups the k observations equal to 0 with the single observation equal to a if $a^2 < 2(k + 1)$.

ii. What is the optimal clustering when $a^2 > 2(k + 1)$? Justify your answer.

[10+(5+5)=20]

6. Describe how a regression tree can be grown by the CART approach, and discuss how an excessively large tree can be pruned by minimizing the cost-complexity criterion.

OR

Describe the multilayer perceptron (MLP) model for regression, and explain how its parameters can be estimated by backpropagation. [10]

Semestral Examination

Advanced Design of Experi ments

M.Stat- 2nd Year : 2011-12

Full Marks: 100 Time : 2:30- 6:00 Date: 18.11.2011

Answer any five questions.

Q1. a). Define Universal Optimality criterion. State and prove a set of sufficient conditions for a design d^* to be universally optimal.

b) Define a Balanced Block design d^* and show that d^* is universally optimal in the class of appropriate block designs for given b, v and k. (3+2+6) +(3+6)=20

Q2. a) Let d be a Latin Square design with 5 treatments. Delete the last row of this design and identify the resultant row column design d^*. Obtain the corresponding C- matrix for the estimation of treatment effects .

b) Consider the following row column design d_0 with 4 rows , 4 columns and 8 treatments. Identify one set of maximum number of estimable independent elementary treatment contrasts, indicating an unbiased estimator of each of the contrasts suggested by you.

$$d_0: \quad \begin{matrix} 1 & 2 & 5 & 6 \\ 3 & 4 & 7 & 8 \\ 8 & 6 & 1 & 3 \\ 7 & 5 & 2 & 4 \end{matrix}$$

(1+5) + 14=20

Q3) a) Construct an Orthogonal Array OA(27, 13, 3, 2). Is this OA saturated? Give reasons.

b) Define a fractional factorial design of resolution d. Can you identify the above OA as a fractional factorial design of appropriate resolution?(Clearly state the result you are using for your claim). (13+2)+(3+2)=20

Q4)a) Define a Second Order Rotatable response surface design (SORD).

b) Derive the necessary and sufficient conditions of the moment matrix of a SORD.

3+17=20

Q5) Consider a k- degree polynomial regression model. State de la Garza (DLG) phenomenon in this context. Using DLG, in the set up of a quadratic regression model, starting with a n-point design, derive a D-optimal design in the class of symmetric designs for the estimation of regression coefficients. 3+17=20

Q6)a) Define a strongly balanced uniform repeated measurements design (SBURMD). Construct one such design with 3 treatments in 6 periods and 9 units.

b) State and prove the optimality property of a SBURMD , in an appropriate class of RMD .

c) Construct a Hadamard matrix of order 12. (2+4)+8+6 =20

INDIAN STATISTICAL INSTITUTE

M. STAT. SECOND YEAR

Topology and Set Theory

Date: 18·11·11 Semestral Examination Time : 3 hrs.

This paper carries 80 marks.The maximum you can score is 70.

1. Let $(X, \Im)$ be a second countable topological space and $\mathbf{B}$ an open base. Show that there is a countable subcollection of $\mathbf{B}$ which is a base. [15]

2. Let D be dense and U open in a topological space. Show that closure$(U \cap D)$ = closure$U$ [10]

3. Let X, Y be Hausdorff spaces, Y compact and $A \subset X \times Y$ closed. Show that the projection of A to X is a closed set. [15]

4. Show that a connected manifold is path connected. [10]

5. On the space of reals $\mathbf{R}$, define $x \sim y$ if $x - y$ is rational.

   (a) Show that $\sim$ is an equivalence relation. [1]

   (b) If $[x] = \{y : x \sim y\}$, show that $[x]$ is dense in $\mathbf{R}$ for any x. [7]

   (c) Show that the quotient topology on $\mathbf{R}/\sim$ equals $\{\phi, \mathbf{R}/\sim\}$ [7]

6. Let $\mathbf{J}$ be the set of irrational numbers with usual topology. Show that the Stone Cech compactification of $\mathbf{J}$ has infinitely many connected components. [15]

21·11·2011

MSTAT II SEMESTER I 2011-2012

ADVANCED MULTIVARIATE ANALYSIS

Attempt all questions. Show your work explicitly. Total Marks: 100 Time: 3 hrs.

1. (18 + 7) (a) Derive the characterization of the most general family of bivariate distributions whose conditionals are specified as possibly different members of an exponential family. State explicitly any result you may need to assume.

(b) Derive the bivariate Cauchy conditionals distribution. State explicitly all results you may need to use.

2. (8 + 10 + 7) (a) Characterize the standard bivariate normal distribution $N_2(0, 0; 1, 1; \rho)$ as the Maximum Entropy distribution for a suitably chosen Entropy measure.

(b) Derive the Entropy of the distribution and its Maximum Likelihood estimator under the setup in (a).

(c) Show how you will conduct the Likelihood Ratio test for a given value of the Entropy, say E*, under the setup in (a).

3. (17 + 8) (a) Derive the Likelihood Ratio test for the homogeneity of overall variability of $k(> 2)$ independent multivariate normal populations of possibly different dimensions.

(b) Suggest a generalization of Hartley's F_{max} test for the univariate case to the multivariate situation described in (a) and show how you will implement this test in practice.

4. (12 + 13) (a) Based on the notion of curvature of the power hypersurface, derive the general form of an exact optimal test for testing a simple multiparameter hypothesis against its two-sided alternative.

(b) Let, $f(x) = pN_2(x; 0, 0; 1, 1; \rho) + (1 - p)N_2(x; 0, 0; 1, 1; 0)$. Derive an optimal test for No Mixture against the mixture distribution $f(x)$ when p $(0 < p < .5)$ is known. Suggest an optimal test for the case when p is unknown with $0 \leq p < .5$.

INDIAN STATISTICAL INSTITUTE

Semester Examination : Semester I (2011-12)

M. Stat. II Year

Actuarial Methods

Date: 23.11.2011 Maximum marks: 100 Time: $3\frac{1}{2}$ hours

Calculator and RMM table can be used.

1. The profit per client-hour made by a privately owned health centre depends on the variable cost involved. Variable cost, over which the owner of the health centre has no control, takes one of the three levels θ_1 = high, θ_2 = medium and θ_3 = low. The owner has to decide at what level to set the number of client-hours that can be either $d_1 = 16,000$, $d_2 = 13,400$ or $d_3 = 10,000$. The profit (in Rs.) per client-hour is as follows:

	θ_1	θ_2	θ_3
d_1	85	95	110
d_2	105	115	130
d_3	125	135	150

Determine the minimax solution. Given the probability distribution $p(\theta_1) = 0.1$, $p(\theta_2) = 0.6$, $p(\theta_3) = 0.3$, determine the solution based on the Bayes criterion. [3+5=8]

2. Consider n independent and identically distributed observations on claims following a $N(\mu, 100)$ distribution. Assuming that μ is equally likely to be either 50 or 60, find the Bayes estimator of μ under squared error loss. Find this Bayes estimator of μ, when it is only given that the mean of the sample exceeds 57. [4+4=8]

3. Claim amounts from a portfolio have the distribution with pdf

$$f(x) = 2cxe^{-cx^2},\ x \geq 0,\ c > 0.$$

An XOL reinsurance arrangement is in force with retention limit $M = 3$. A sample of reinsurer's payment amounts gives the following values: $n = 10$, $\sum y_i = 8.7$ and $\sum y_i^2 = 92.3$. Find maximum likelihood estimate of c. [8]

4. (a) Explain the concept of benefits and perils in the context of general insurance by means of examples.

   (b) Describe a credibility estimate, specifying its characteristics.
   The Bayesian approach using quadratic loss always produces an estimate which can be readily expressed in the form of a credibility estimate. Prove or disprove the statement.

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

5. (a) Describe a compound Poisson distribution in the context of aggregate claims. Derive expressions for its mean and variance in terms of those of number of claims and claim size distributions.

   (b) Consider a proportional reinsurance arrangement wherein the direct writer retains a proportion k. Find the mgf of Y, the net individual claim amount paid by the direct writer, in terms of that of X. Hence find expressions for the mgf's of the aggregate claim amounts paid by the direct writer and the reinsurer separately, if the number of claims has a $Poisson(\lambda)$ distribution.

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

6. (a) Describe the difference between collective risk model and individual risk model.

   (b) A portfolio of policies consists of one-year term assurances on 100 lives aged exactly 30 and 200 lives aged exactly 40. The probability of a claim during the year on any one of the lives is 0.0004 for the 30 year olds and 0.001 for the 40 year olds. If the sum assured on a life aged x is uniformly distributed between $1000(x-10)$ and $1000(x+10)$, calculate mean and variance of the aggregate claims from this portfolio during the year. Which of the two risk models, mentioned in part (a), has been used here?

[3+(2+4+1)=10]

7. The aggregate claim process for a particular risk is a compound Poisson process with Poisson parameter $\lambda = 20$ per year. Individual claim amounts are Rs 100 w.p. 1/4, Rs 200 w.p. 1/2 and Rs 250 w.p. 1/4. The initial surplus is Rs 1000. Using a normal approximation, calculate the smallest premium loading factor such that the probability of ruin at year 3 is at most 0.05. [10]

8. (a) Describe a typical No Claim Discount (NCD) system specifying its objective(s).

   (b) An NCD system with three discount categories (0%, 25% and 50%), the rules are as follows: (*i*) following a claim-free year, the discount increases by one level, or remains at category 3, (*ii*) following a year in which exactly one claim has been made, the discount decreases by one level, or remains at category 1, (*iii*) following a year in which more than one claim has been made, the discount returns to category 1.

   The number of claims made in each year under each policy is assumed to follow a *Poisson* distribution with mean λ. Write down the transition matrix of the probabilities p_{ij} that a policyholder in category i in one year will move to category j in the following year. Obtain the proportions of policyholders in different categories under equilibrium. Calculate the average premium amount in equilibrium assuming the basic premium to be c.

[3+(5+4+2)=14]

9. The table below shows the cumulative claims (in Rs '000s) incurred on a particular class of insurance policies, divided by accident year and development year.

Accident Year	Development Year			
	0	1	2	3
1997	502	556	589	600
1998	487	565	593	
1999	608	640		
2000	551			

State the assumptions underlying the basic chain-ladder method. Using the method, estimate the outstanding claims reserve as on 31 December 2000. [2+6=8]

10. (a) Discuss Generalized Linear Models (GLM) in contrast with Linear Models specifying the assumptions clearly. What is canonical link function?

    (b) Identify the process $2X_t = 7X_{t-1} - 9X_{t-2} + 5X_{t-3} - X_{t-4} + e_t - e_{t-2}$ as an ARIMA model.

    (c) Using the Acceptance-Rejection method, generate a discrete random variable from $p(1) = 1/3$ and $p(2) = 2/3$ with an unbiased coin only.

[(3+1)+4+4=12]

Date: 25.11.2011
Time: 3 hours

Statistical Methods in Genetics – I
M-Stat (2nd Year)
Final Examination 1st Semester 2011-12

This paper carries 60 marks. Answer all questions.

1 (a) Define each of the following:
Snyder's Ratio; heterozygote advantage; haplotype phase; segregation ratio; LOD score.

(b) Suppose the test for Hardy-Weinberg Equilibrium is separately carried out in two different samples. If neither of the samples provides any significant evidence of departure from HWE, explain whether the test based on the pooled sample can yield a significant result. [5 + 5]

2 (a) Consider a biallelic X-linked locus with alleles (A,a). If the initial frequencies of the allele A among males and females are p_1 and p_2, respectively, in how many generations will the frequency of A be p^* among males?

(b) Consider data on A-B-O blood groups for a random sample of individuals chosen from an inbred population. Describe an EM algorithm to obtain the m.l.e. of the frequencies of the alleles A, B, O and the inbreeding coefficient. [4 + 6]

3 (a) Consider a dominant disorder controlled by an autosomal biallelic locus. If the prevalence of the disease in the population is 0.19, what is the probability that the marriage between a uncle and his niece will produce an affected offspring?

(b) What is the estimated i.b.d. score for a pair of sibs, both heterozygous at an autosomal biallelic locus and having a parent who is also heterozygous at that locus?

(c) Consider two autosomal biallelic loci with alleles (A,a) and (B,b), respectively. If the frequencies of the haplotypes AB, aB and ab are 0.4, 0.3 and 0.1, respectively, determine the coefficient of linkage disequilibrium between the two loci. [4 + 5 + 3]

4 (a) Explain why linkage disequilibrium exists over much smaller distances on the genome compared to linkage.

(b) Consider the following data from an affected sib-pair study (both sibs are affected and affection status of parents are unknown) on Coronary Artery Disease (CAD). *MTHFR* (methylenetetrahydrofolate reductase) on Chromosome 1 is believed to be a candidate gene for CAD. Twelve sib-pairs comprising all affected siblings (along with their parents) were genotyped at a triallelic marker locus *D1S1012* near this candidate gene. Do these data provide evidence of linkage between *D1S1012* and a locus controlling CAD?

Sibship	Parental genotypes	Genotypes of affected sibs
1	AA,AB	AA,AB
2	AB,BC	AB,BC
3	AB,CC	AC,BC
4	BB,BC	BB,BB
5	AB,AB	AB,AB
6	AC,*	AA,AA
7	AB, BB	AB,BB
8	AA,AB	AA,AB
9	*, *	AC,AC
10	CC,CC	CC,CC
11	AB,AC	AC,BC
12	AC,AC	AA,AC

* denotes missing genotype [3 + 15]

5. Presentation of assigned problems [10]

INDIAN STATISTICAL INSTITUTE
End-Semester Examination : 2011 - 2012 (First Semester)

M. STAT. II YEAR
Functional Analysis

OPEN NOTES CLOSED BOOK EXAMINATION
(Handwritten notes are allowed but any printed or photocopied material is not allowed)

Date : 25.11.11

Duration : $3\frac{1}{2}$ Hours
Total Marks : 60
Maximum Marks : 50

T^* : adjoint of T
$M^\perp$: orthogonal complement of M
$\mathcal{R}(A)$: Range of A
$\mathcal{B}(\mathcal{H})$: set of bounded linear operators on $\mathcal{H}$
$\mathcal{B}(X,Y)$: set of bounded linear operators from X to Y
Y^* : set of bounded linear functionals on Y
$\sigma(T)$: spectrum of T

1. Let E be the space of all sequences of complex numbers $x = (x_1, x_2, x_3, \cdots)$ such that $\sum_{j=1}^{\infty} |x_{2j}| < \infty$ and $\sum_{k=0}^{\infty} |x_{2k+1}|^2 < \infty$. Then it can be proved that

$$\|x\| = \sum_{j=1}^{\infty} |x_{2j}| + \left(\sum_{k=0}^{\infty} |x_{2k+1}|^2 \right)^{1/2}$$

defines a norm on E. Show that E is a Banach space with this norm and that there is no inner product on E such that $\langle x, x \rangle = \|x\|^2$ for all $x \in E$. [4]

2. Let $(u_n)_{n=1}^{\infty}$ and $(v_n)_{n=1}^{\infty}$ be orthonormal bases of a Hilbert space $\mathcal{H}$ and let $(\lambda_n)_{n=1}^{\infty}$ be a bounded sequence of complex numbers. For $x \in \mathcal{H}$, define

$$T(x) = \sum_{n=1}^{\infty} \lambda_n \langle x, u_n \rangle v_n.$$

(i) Prove that T is a bounded linear operator on $\mathcal{H}$. Find $\|T\|$. [2]
(ii) Determine T^* and show that T is unitary if and only if $|\lambda_n| = 1$ for all $1 \leq n < \infty$. [3]

3. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two Hilbert spaces and let $T : \mathcal{H}_1 \to \mathcal{H}_2$ be a bounded linear operator. Let M_1 and M_2 be subsets of $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively such that $T(M_1) \subset M_2$. Prove that $T^*(M_2^\perp) \subset M_1^\perp$. [2]

4. Let $Y = \{x = (x_1, x_2, x_3, \cdots) \in l^2 : x_1 = 0, x_2 + x_3 = 0, x_3 + x_4 = 0\}$ and let $P : l^2 \to l^2$ be the orthogonal projection onto $Y^\perp$. Determine P. [3]

5. Let X be a Banach space and let $x_1, x_2, x_3, \cdots$ be a sequence of points in X. Assume that for each $f \in X^*$ the sequence $f(x_1), f(x_2), f(x_3), \cdots$ is bounded. Prove that the sequence $x_1, x_2, x_3, \cdots$ is bounded. [4]

6. Let $\{e_n\}_{n=1}^\infty$ be an orthonormal basis for a Hilbert space $\mathcal{H}$. Define the operator T by the formula

$$T(x) = \sum_{n=1}^{\infty} \frac{\langle x, e_{n+1} \rangle}{n+1} e_n$$

for $x \in \mathcal{H}$. Show that T is a compact operator and find T^*. [4]

7. Let $\mathcal{H}$ be a Hilbert space and P_1, P_2 be orthogonal projections on $\mathcal{H}$. Show that the following conditions are equivalent:
(i) $\mathcal{R}(P_1) \subset \mathcal{R}(P_2)$
(ii) $P_1 \leq P_2$
(iii) $P_1 P_2 = P_2 P_1 = P_1$
(iv) $\|P_1 x\| \leq \|P_2 x\| \ \forall x \in \mathcal{H}$. [8]

8. Let A be any operator on a Hilbert space $\mathcal{H}$. Let $A = W|A|$ be the polar decompostion of A, where W is a partial isometry and $|A| = (A^*A)^{1/2}$. Show that:
(i) $W^*A = |A|$.
(ii) W is an isometry if and only if A is injective.
(iii) W and $|A|$ commute if and only if A commutes with A^*A. [6]

9. Let $A \in \mathcal{B}(\mathcal{H})$ be such that $AT = TA$ for every compact operator T on $\mathcal{H}$, where $\mathcal{H}$ is a Hilbert space. Prove that $A = \lambda I$ for some $\lambda \in \mathbb{C}$. [4]

10. Let X and Y be Banach spaces and let $T \in \mathcal{B}(X, Y)$. For $f \in Y^*$, define $T^*f \in X^*$ by $T^*f(x) = f(Tx)$ for $x \in X$.
(i) Prove that if T is surjective then T^* is injective. [2]
(ii) Prove that T^* is surjective if and only if T is injective and $T^{-1} \in \mathcal{B}(\mathcal{R}(T), X)$. [5]

11. Let $T_1, T_2, T_3, \cdots$ and $S_1, S_2, S_3, \cdots$ be bounded linear operators on a Banach space X such that $S_n T_m = T_n S_m$ for all $n, m \in \mathbb{N}$. Assume that $\{T_n\}_{n=1}^\infty$ and $\{S_n\}_{n=1}^\infty$ are both strongly operator convergent with limits T and S respectively. Prove that $TS = ST$. [5]

12. Let K be an arbitrary, non-empty compact set in $\mathbb{C}$. Give an example of a bounded linear operator T such that $\sigma(T) = K$. [8]

INDIAN STATISTICAL INSTITUTE

First Semester Examination: 2011-2012

M.Stat. 2nd Year. 1st Semester

Analysis of Discrete Data

Date: November 28, 2011 **Maximum Marks: 65** **Duration: 3 hours**

GROUP A

(Answer all the questions.)

1. Discuss kappa and generalized kappa as measures of agreement. Find their values (setting appropriate weights for generalized kappa) for the following data of diagnoses, and interpret the results.

	Pathologist B			
Pathologist A	1	2	3	4
1	22	2	2	0
2	5	7	14	0
3	0	2	36	0
4	27	12	69	17

[3+4+2 = 9]

2. 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 $\pi_{i+} = \sum_j \pi_{ij}$, $\pi_{+j} = \sum_i \pi_{ij}$.

[2+4 = 6]

GROUP B

(This group carries 53 points. Answer as much as you can.
However, the maximum you can score in this group is 50.)

1. A binomial GLM assumes that $n_i Y_i \sim \text{Bin}(n_i, \pi_i)$, where $\pi_i = \Phi\left(\sum_j \beta_j x_{ij}\right)$ with an arbitrary continuous cdf Φ as its inverse link function. Show that the likelihood equations for estimating β's are given by

$$\sum_i \frac{n_i(y_i - \pi_i)x_{ij}}{\pi_i(1-\pi_i)} \phi\left(\sum_j \beta_j x_{ij}\right) = 0,$$

P.T.O.

where ϕ is the derivative of Φ, the pdf associated with it. [You cannot assume any form of the likelihood equations for a GLM.] Also, find an expression for the estimate of the asymptotic covariance matrix of $\hat{\beta}$, the maximum likelihood estimate of β. [You may assume a relevant result from GLM.] [8+4 = 12]

2. (a) Construct the log-likelihood function for the model $\text{logit}[\pi(x)] = \alpha + \beta x$ with independent binomial outcomes of y_0 successes in n_0 trials at $x = 0$ and y_1 successes in n_1 trials at $x = 1$, where $\pi(x)$ denotes the probability of success. Derive the likelihood equations, and show that $\hat{\beta}$, the maximum likelihood estimate of β, is the sample log odds ratio. [4+3 = 7]

(b) Consider an $I \times 2$ contingency table. Let y_i be the number of outcomes in the first column (successes), out of n_i trials. We consider a logit model given by $\text{logit}(\pi_i) = \alpha + \beta_i$, with π_i denoting probability of success in i-th row.

(1) Given $\{\pi_i > 0\}$, show how to find $\{\beta_i\}$, in terms of $\{\pi_i > 0\}$, satisfying $\beta_I = 0$. [2]

(2) Prove that $\beta_1 = \beta_2 = \cdots = \beta_I$ is the independence model. Find the corresponding likelihood equation, and show that $\hat{\alpha} = \text{logit}[(\sum_i y_i)/(\sum n_i)]$. [1+3+2 = 6]

3. Consider a logit model with N settings of the explanatory variables.

(a) The deviance residual is defined by $\sqrt{d_i} \times \text{sign}(y_i - n_i\hat{\pi}_i)$, $i = 1, \ldots, N$, where d_i is the contribution for observation i. Show that

$$d_i = 2\left(y_i \log \frac{y_i}{n_i\hat{\pi}_i} + (n_i - y_i) \log \frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right). \qquad [9]$$

(b) Write down expressions for Pearson residuals and standardized Pearson residuals, explaining all your notations. [2+2 = 4]

4. (a) Consider a loglinear model for a three-way table. Justify, with adequate justification, the following statement. "*Any model not having the three-factor interaction term has a homogeneous association for each pair of variables.*" [6]

(b) For the loglinear model for an $I \times J$ table, $\log \mu_{ij} = \lambda + \lambda_i^X$, where μ_{ij} is the expected frequency of the (i, j)-th cell, λ is the general effect, and λ_i^X is the effect corresponding to the i-th row. Show that $\hat{\mu}_{ij} = n_{i+}/J$ and residual df $= I(J-1)$. . [4+3 = 7]

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

INDIAN STATISTICAL INSTITUTE

First Semester Examination : 2011-12

M. Stat. II Year

Topics in Bayesian Inference

Date : 28.11.2011 Maximum Marks : 100 Duration : $3\frac{1}{2}$ Hours

Group-A

Answer all questions

1. Consider the linear regression model $y = X\beta + \varepsilon$ where $y = (y_1, \ldots\ldots, y_n)'$ is the vector of observations on the "dependent "variable , $X = (x_{ij})_{nxp}$ is of full rank , x_{ij} being the values of the nonstochastic regressor variables, $\beta = (\beta_1, \ldots\ldots, \beta_p)'$ is the vector of regression coefficients and the components of ε are independent , each following N (O, σ^2). Consider the noninformative prior $\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$, $\beta \in \mathbb{R}^p$, $\sigma^2 > 0$.

   (a) Show that the marginal posterior distribution of β is a multivariate t distribution.
   (b) Find a 100(1-α) % HPD credible set for β.

   [7+10=17]

2. (a) What are the difficulties with improper noninformative priors in Bayes testing? Describe intrinsic Bayes factor (IBF) and fractional Bayes factor (FBF) as solutions to this problem with improper priors.

   (b) What is an intrinsic prior in the context of non-subjective Bayes testing? Consider the nested case and find the intrinsic prior determining equations corresponding to AIBF. Show that the solution suggested by Berger and Pericchi satisfies the intrinsic prior determining equations.

   [(3+9) + 14=26]

3. Consider the model where we have p independent random samples from p normal populations:

   X_{ij}, j=1,2,.....,n are i.i.d. N(θ_i, σ^2), i=1,2,...., p.

   Assume that θ_i are i.i.d. N(η_1, η_2) and σ^2 is known. Our problem is to make inference about $\theta_1, \ldots\ldots, \theta_p$. Describe the hierarchical Bayes approach and the parametric empirical Bayes (PEB) approach in this context. Derive James-Stein estimate as a PEB estimate.

   [12+5 = 17]

GROUP – B

Answer as many questions as you can.

Maximum you can score is 40.

4. Let X follow N(θ,1) where θ is known to be nonnegative. Find the Bayes estimate of θ (in its simplest form) for squared error loss using the standard noninformative prior.

[5]

5. Suppose we are studying the distribution of the number of defectives X in the daily production of a product. Consider the model $(X|Y,\theta)$ ~ Bin(Y,θ) where Y, a day's production, follows Poisson (λ). The difficulty is that Y is not observable and inference has to be made on the basis of X only.

Consider a Beta (α,β) prior for θ, and show how Gibbs sampling can be used to sample from the posterior distribution of θ. Find all the required conditional distributions.

[15]

6. Let $X_1,.....,X_n$ be i.i.d. with a common density $f(x \mid \theta) = e^{-(x-\theta)}$, $x > \theta$. Consider the problem of comparing the models $M_0 : \theta = 0$ and $M_1 : \theta > 0$ with the prior $\pi(\theta) = 1$ on $\theta > 0$. Find the fractional Bayes factor (FBF) with a fraction b and show that FBF ≥1 for any 0<b<1 and all possible data.

[7+5 = 12]

7. Let $X_1,.....,X_n$ be i.i.d. N(θ,1) variables. Consider the problem of comparing the models $M_0 : \theta = 0$ and $M_1 : \theta \in IR$ with the prior $\pi(\theta) \equiv 1$ for θ under M_1. It is well known that the intrinsic prior corresponding to the AIBF based on training samples of size 1 is an N(0, 2) prior. Find the intrinsic prior corresponding to the AIBF based on training samples of size 2 (You may use the result of Berger and Pericchi for the nested case). Which of these two intrinsic priors would you recommend for calculating a nonsubjective Bayes factor? Give reasons for your answer.

[12=2 = 14]

****

INDIAN STATISTICAL INSTITUTE

Back paper Examination (Semester I) : 2011-12

M. Stat. II Year

Topics in Bayesian Inference

Date : 22.12.2011 Maximum Marks : 100 Duration : 3 Hours

1. (a) Define posterior predictive distribution of a future observation.

   (b) Let $X_1, \ldots\ldots, X_n$ be i.i.d. observations each following a Bernoulli(θ) distribution, $0<\theta<1$. Consider the uniform prior distribution for θ and find the posterior predictive distribution of a future observation X_{n+1}.

   (c) Let $X_1, \ldots\ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown.
   (i) Suggest a suitable Bayes test for $H_0 : \mu=0$ versus $H_1 : \mu\neq 0$.
   (ii) Consider the noninformative prior $\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$. Find the marginal posterior distributions of σ^2 and μ. Also find the $100(1-\alpha)\%$ credible set for μ.

   [2+9+(5+6+6+4)=32]

2. Let $X_1, \ldots\ldots, X_m$ and $Y_1, \ldots\ldots, Y_n$ be independent random samples, respectively from $N(\mu, \sigma_1^2)$ and $N(\mu, \sigma_2^2)$ where both σ_1^2 and σ_2^2 are known. Consider a uniform prior for the common mean μ. Find the posterior distribution of μ and the Bayes estimate of μ for the squared error loss.

   [7]

3. (a) Consider the problem of model selection with two competing models. Define Bayes factor and show how it is related to the posterior probabilities of the models.

   (b) What are the difficulties with improper noninformative priors in Bayesian testing? Describe intrinsic Bayes factor and fractional Bayes factor as solutions to this problem with inproper priors.

   (c) What is an intrinsic prior in the context of nonsubjective Bayes testing? Suppose we have observations $X_1, \ldots\ldots, X_n$. Under model M_0, X_i are i.i.d. $N(0, 1)$ and under model M_1, X_i are i.i.d. $N(\theta,1)$, $\theta \in \mathbb{R}$. Consider the noninformative prior $g_1(\theta) \equiv 1$ for θ under M_1. Find the intrinsic prior for θ and show that the ratio of the AIBF and the BF with this instrinsic prior tends to one as n tends to infinity.

   [5+(4+9)+(3+11)=32]

P.T.O.

4. (a) Describe the Metropolis-Hastings algorithm and the Gibbs sampling method for Bayesian computation.

(b) Consider the hierarchical Bayesian model where we have k independent random samp $(Y_{i1}, \ldots\ldots, Y_{in_i})$, i = 1,2, ...,k, from k normal populations :

$Y_{ij} \sim N(\theta_i, \sigma_i^2)$, i=1, ..., k, j = 1, ..., n_i

θ_i are i.i.d. $N(\mu_\pi, \sigma_\pi^2)$

σ_i^2 are i.i.d. Inverse-Gamma (a_1, b_1),

$(\theta_1, \ldots\ldots, \theta_k)$ and $(\sigma_1^2, \ldots\ldots, \sigma_k^2)$ are independent and the second stage prior on μ_π and σ_π^2 a

$\mu_\pi \sim N(\mu_0, \sigma_0^2)$ and $\sigma_\pi^2 \sim$ Inverse-Gamma (a_2, b_2).

Assume that a_1, a_2, b_1, b_2, μ_0, and σ_0^2 are specified constants. Describe how you can fi estimates of $\theta_1, \ldots\ldots, \theta_k$ using Gibbs sampler. Derive the required full conditional distributions

[8+21 =

****

INDIAN STATISTICAL INSTITUTE
Back-Paper Examination: 2011-2012 (First Semester)

M. STAT. II YEAR
Functional Analysis

Date : 23.12.11

Duration : 3 Hours
Total Marks: 100

$M^{\perp}$: orthogonal complement of M
dim M : dimension of M
codim M : codimension of M = dimension of X/M
$\mathcal{B}(\mathcal{H})$: set of bounded linear operators on $\mathcal{H}$
$\mathcal{B}(X, Y)$: set of bounded linear operators from X to Y
Y^* : set of bounded linear functionals on Y

1. Let M be a closed subspace of a Banach space X. Show that codim M = dim $M^{\perp}$ in the sense that either both sides are infinite or they are finite and equal. [10]

2. Prove that every infinite dimensional Banach space X contains a vector subspace that is algebraically isomorphic to l^{∞}. [10]

3. Let $T \in \mathcal{B}(\mathcal{H})$, where $\mathcal{H}$ is a Hilbert space. Show that the following statements are equivalent :
(i) T is compact.
(ii) There is a sequence of finite rank operators $\{T_n\}_{n=1}^{\infty}$ such that $\|T - T_n\| \to 0$.
(iii) T^* is compact. [15]

4. Let X and Y be Banach spaces. If $\mathcal{A} \subset \mathcal{B}(X, Y)$ is such that for every $x \in X$ and $g \in Y^*$, we have $\sup_{A \in \mathcal{A}} |g(Ax)| < \infty$ then prove that $\sup_{A \in \mathcal{A}} \|A\| < \infty$. [10]

5. Show that an operator $A \in \mathcal{B}(\mathcal{H})$ is diagonalizable if and only if there is an orthonormal basis of $\mathcal{H}$ consisting of eigenvectors for A, where $\mathcal{H}$ is a Hilbert space. [10]

6. Let $K : L^2[0, 1] \to L^2[0, 1]$ be given by the formula

$$Kf(t) = \int_0^1 tsf(s)ds$$

for $f \in L^2[0, 1]$. Show that the range of K is one-dimensional. Deduce from this that K has only one non-zero eigenvalue and find it. [10]

7. Let $c = \{x = (x_1, x_2, x_3, \cdots) \in l^\infty : \lim_{j\to\infty} x_j \text{ exists and is finite}\}$. We say that $x \in l^\infty$ stabilizes if there exists N such that for all $j \geq N$ we have $x_j = x_N$. Let $M = \{x \in c : x \text{ stabilizes}\}$. Show that c is the closure of M. [15]

8. Let $\mathcal{H}$ be a complex Hilbert space. Let P be an orthogonal projection on $\mathcal{H}$ and let S be a unitary operator on $\mathcal{H}$. Prove that the operator $Q = S^{-1}PS$ is an orthogonal projection. [10]

9. Let X and Y be normed spaces and $T : X \to Y$ be a linear operator. The graph of T is the set $\mathcal{G}(T) = \{(x, y) \in X \times Y : y = Tx\}$. We will treat $X \times Y$ as a normed space with the norm $\|(x, y)\| = \|x\| + \|y\|$. Let $T_1, T_2 : X \to Y$ be two linear operators. Show that if $\mathcal{G}(T_1)$ is closed in $X \times Y$ and T_2 is bounded, then $\mathcal{G}(T_1 + T_2)$ is closed in $X \times Y$. [10]

INDIAN STATISTICAL INSTITUTE

Backpaper Examination: 2011-12 (First Semester)
Master of Statistics (M. Stat.) II Year
Advanced Probability I

Instructor: Antar Bandyopadhyay and Parthanil Roy

Date: 23/12/2011 **Total Points: 100** **Duration: 3 Hours**

Note:

- Please write your roll number on top of your answer paper.
- Show all your works and write explanations when needed.
- Answer all questions.
- This is an open note examination. You are allowed to use your own hand written notes (such as class notes, your homework solutions, list of theorems, formulas etc). Please note that no printed materials or photo copies are allowed, in particular you are not allowed to use books, photocopied class notes etc.

1. State whether the following statements are *true* or *false* and provide detailed reasons supporting your answers.

   (a) (10 points) Let $\nu \ll \mu$ be two probability measures on $(\Omega, \mathcal{F})$ and let $\{A_n\}_{n\geq 1}$ be a sequence of events such that $\mu(A_n) \to 0$ as $n \to \infty$, then $\nu(A_n) \to 0$ as $n \to \infty$.

   (b) (10 points) Let μ be a σ-finite measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ which is *absolutely continuous* with respect to the *Lebesgue measure*, then $\mu(K) < \infty$ for any *compact subset* K of $\mathbb{R}$.

2. (20 points) Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and X and Y be two i.i.d. Normal $(0,1)$ random variables defined on it. Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra such that $\mathbf{E}\left[X \,\middle|\, \mathcal{G}\right]$ is $\sigma(Y)$-measurable. Show that $\mathbf{E}\left[X \,\middle|\, \mathcal{G}\right]$ is almost surely constant and find the constant.

3. Let $\{X_n\}_{n\geq 1}$ be a sequence of i.i.d. continuous random variables. We call X_k a *record value* for the sequence if $X_k > X_l$ for $1 \leq l < k$. Let I_k be the indicator function for the event that X_k is a record value. Let R_n is the number of record values in the first n observations $X_1, X_2, \ldots X_n$.

P.T.O.

(a) (10 points) Show that

$$\sum_{k=1}^{\infty} \frac{I_k - k^{-1}}{\log k} < \infty$$

almost surely.

(b) (10 points) Using (a), show that

$$\frac{R_n}{\log n} \to 1$$

almost surely.

4. (25 points) Let $\{X_n\}$ be a mean zero L^2 martingale. Show that for $x > 0$,

$$P\Big(\max_{1 \le k \le n} X_k > x\Big) \le \frac{E(X_n^2)}{x^2 + E(X_n^2)}$$

5. Let $X_1, X_2, \ldots$ be independent random variables with $P(X_n = 1) = P(X_n = -1) = 1/2n$ and $P(X_n = 0) = 1 - 1/n$ for all $n \ge 1$. Define $Y_1 = X_1$ and for all $n \ge 2$,

$$Y_n = \begin{cases} X_n & \text{if } Y_{n-1} = 0, \\ nY_{n-1}|X_n| & \text{if } Y_{n-1} \neq 0. \end{cases}$$

(a) (3 points) Show that $\{Y_n\}$ is a martingale with respect to its natural filtration.

(b) (4 points) Show that $Y_n \to 0$ in probability.

(c) (5 points) Show that Y_n does not converge almost everywhere.

(d) (3 points) Explain why the martingale convergence theorem fails in this case.

Wish you all the best

INDIAN STATISTICAL INSTITUTE

M. STAT. SECOND YEAR

Topology and Set Theory Date: 30/12/11 Backpaper Examination Time : 3 hrs.

Maximum Marks-100

1. Let $X \subseteq \mathbf{R}^2$ be the set $\{(x, y) : x \text{ is rational and } y \text{ is irrational}\}$.
   Is X completely metrizable ? Justify your answer. [15].

2. Let X be a compact metric space and $f : X \longrightarrow X$ an isometry. Show that f is onto. [15]

3. Let X be a non compact metric space. Show that the Stone Cech compactification of X is not metrizable. [15]

4. Show that the product of countably many manifolds need not be a manifold. [15]

5. Let $\{A_\alpha : \alpha \varepsilon I\}$ be a locally finite family of closed sets. Let f be a real valued function on X such that f is continuous on each A_α. Show that f is continuous on X. [15]

6. Let $X = \{(x, y) : x^2 + y^2 = 1\}$. Define $\sim$ on X by $(x, y) \sim (-x, -y)$ and $(x, y) \sim (x, y)$. Show that $X/\sim$ is homeomorphic to X. [15]

7. Let X, Y be topological spaces and $f : X \longrightarrow Y$ a continuous function. Let $G = \{(x, f(x)) : x \varepsilon X\}$. Show that G is homeomorphic to X. [10]

INDIAN STATISTICAL INSTITUTE

Mid-semester Examination : Semester II (2011-2012)

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

Subject Name : Statistical Methods in Biomedical Research

Date : 20.02.12 , Maximum Marks : 30. Duration : 1 hr. 30 min.

1. (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%? [2]

   (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? [2]

2. Describe Efron's biased coin design for allocation with two treatments, say A and B. Find the expectation of the proportion of allocation by A. [3+3]

3. If there are only three groups in a group sequential study, give one form of type I error spending function which will spend the total type I error in a 1:7:19 fashion in the three groups (assuming equal time interval for each group). [4]

4. Suppose three treatments A, B and C are being compared in a clinical trial. The first patient is treated randomly by choosing any treatment with same probability. For any subsequent patient i, if the response of the $(i-1)$th patient is a success, we treat the ith patient by the same treatment as the $(i-1)$th patient. If the response of the $(i-1)$th patient is a failure, we treat the ith patient by any of the remaining two treatments by tossing a fair coin. If the success probabilities of the three treatments are 0.7, 0.4 and 0.3 respectively, find the probabilities that the 10th patient is treated by A and the 10th patient results in a success. [10]

INDIAN STATISTICAL INSTITUTE

M. Stat. II year, 2011 – 12

Mid-Semester Examination

Pattern Recognition and Image Processing

Date: 20.02.12 **Maximum marks: 60** **Duration: 150 minutes.**

Note: Answer all the questions

1. State the Bayes decision rule for three-class classification problem and show that it minimizes the probability of misclassification. [2+10=12]

2. Let there be two classes C_1 and C_2 with prior probabilities P_1 and P_2, and the corresponding prob. density functions p_1 and p_2 where
   $p_k(x) = k\exp(-kx) \quad x > 0 \; and \quad k = 1,2.$
   (i) Find the Bayes decision rule for classification and its probability of misclassification.
   (ii) Find the probability of misclassification of the following decision rule. Put x in class 1 if $x < \log 3$. Otherwise, put it in class2. [10+5=15]

3. (i) State the minimum within cluster distance criterion.
   (ii) Describe the k-means algorithm for clustering.
   (iii) Give an example of a data set and two initial sets of seed points so that the two final results of the k-means algorithm on the data are different.
   [5+5+5=15]

4. Describe the Perceptron method for two class classification problem. [10]

5. (i) State the k-nearest neighbor method of estimating a probability density function. (ii) Derive the k-nearest neighbor decision rule from the density estimation procedure. [3+5=8]

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination: 2011-12 (Second Semester)
M. Stat. II Year

ACTUARIAL MODELS

Date: 22 February 2012 Maximum Marks: 50 Duration: 2 hr

1. A company assesses the credit-worthiness of various companies every quarter, the ratings A, B, C and D being in the order of decreasing merit. From past experience, it is known that credit rating of a company evolves as a time-homogeneous Markov chain with transition matrix P given below:

$$P = \begin{pmatrix} 1-\alpha-\alpha^2 & \alpha & \alpha^2 & 0 \\ \alpha & 1-2\alpha-\alpha^2 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1-2\alpha-\alpha^2 & \alpha \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

   a. Draw the transition graph of the chain.
   b. Determine the range of values of α for which P is a valid transition matrix.
   c. Is the chain irreducible and aperiodic? Justify your answer.
   d. Deduce a stationary distribution for the chain. Is it unique? Justify your answer.
   e. If α=0.1, calculate the probability that a company's rating in the third quarter is D, given that the company's rating in the first quarter was B.

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

2. The total number of claims received by an insurance company can be described by a Poisson process with parameter λ.

   a. Stating clearly the assumptions you make, derive a differential-difference equation for $p_j(t)$, the probability that the company receives j claims by time t.
   b. Define the first holding time T_0 for the process, and deduce its distribution.

[5+5=10]

3. The levels of discount in a no-claims discount scheme for car insurance are 10%, 20%, and 40%. If a driver does not make a claim in any given year, in the following year he moves up one level, or stays at the same level if he is at the highest level. On the other hand, if he makes a claim in a given year, he moves down one level or, if he already at the lowest level, he stays there. A reckless driver has a probability of 0.3 of making a claim in any year.

(PLEASE TURN OVER)

a. Suggest an appropriate stochastic model for the scheme for this type of driver, and specify the model completely.
b. Deduce the long-term properties of this model.
c. If the full premium is Rs. 10000 and the driver is currently at the 0% level, determine the average premium he will have to pay in the long run.

[4+4+2=10]

4. A particular machine is in constant use. Regardless of how long it has been since the last repair, it tends to break down once a day, and on an average, it takes the repairman 6 hours to fix it. Model the machine's status as a time-homogeneous Markov process with two states: ***being repaired*** (denoted by 0) and ***working*** (denoted by 1).

a. Write down the Kolmogorov forward differential equation for $P_{00}(t)$, the probability that the machine is being repaired at time t, given that it is being repaired at time 0.

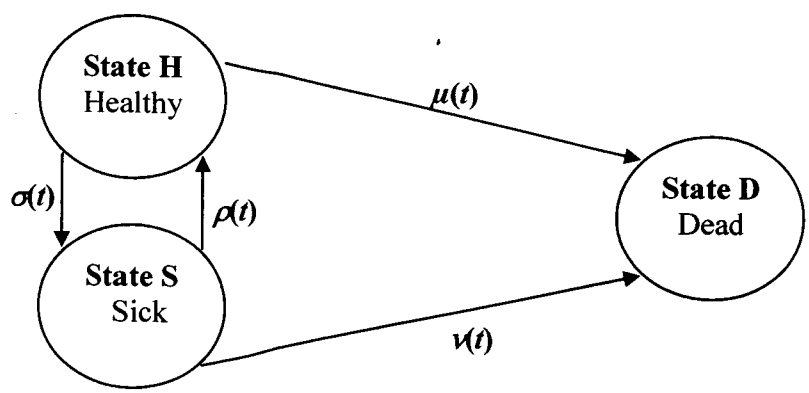
b. Hence show that

$$P_{00}(t) = \frac{1}{5} + \frac{4}{5}e^{-5t}.$$

c. Determine $P_{\overline{11}}(t)$.

[2+3+5=10]

5. The following diagram represents a three-state time-inhomogeneous Markov jump process with transition rates indicated against respective transitions:



a. Write down the transition matrix for this model.
b. Interpret the transition rates in terms of transition probabilities for the model.
c. From first principles, deduce an expression for the probability $P_{\overline{HH}}(s,t)$ that, starting from state **H** at time s, an individual remains in state **H** till time t.

[1+4+5=10]

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination: 2011-12 (Second Semester)
Master of Statistics (M. Stat.) II Year
Stochastic Processes I

Instructor: Parthanil Roy

Date: 23/02/2012 **Maximum Marks: 40** **Duration: 2:30 - 4:30 pm**

Note:

- **Please write your roll number on top of your answer paper.**
- **There are four problems each carrying 10 points with a total of 40 points. Solve as many as you can. Show all your works and write explanations when needed. Maximum you can score is 40 points.**
- **You are NOT allowed to use class notes, books, homework solutions, list of theorems, formulas etc. If you are caught using any, you will obtain a zero grade in the course.**

1. Let S and T be complete separable metric spaces and $\mu_n \Rightarrow \mu$ in $\mathcal{P}(S)$. Let $\phi : S \to T$ be a Borel measurable map and D be the set of points $s \in S$ where ϕ is not continuous. Show that D is a Borel set in S. If $\mu(D) = 0$, then show that $\mu_n \circ \phi^{-1} \Rightarrow \mu \circ \phi^{-1}$ in $\mathcal{P}(T)$. [4 + 6 = 10]

2. X is a Polish space and $P_n \Rightarrow P$ in $\mathcal{P}(X)$. If F is a uniformly bounded and equicontinuous family of real valued functions defined on X, then show that $\int f dP_n \to \int f dP$ uniformly in $f \in F$. [10]

3. Show that the Prokhorov's metric stated in class metrizes weak convergence of probability measures on Polish spaces. (You do not have to prove that it is indeed a metric.) [10]

4. Let $\{B_t\}_{0 \le t \le 1}$ be a standard Browian motion starting at zero and define $\tilde{B}_t := B_{1-t} - B_1$ for all $t \in [0, 1]$. Show that $\{\tilde{B}_t\}_{0 \le t \le 1}$ is also a standard Brownian motion starting at zero. [10]

Wish you all the best

INDIAN STATISTICAL INSTITUTE

Mid-semestral Examination : (2011-12)

Course Name : M.Stat. (BSDA)
Subject Name : Statistical Methods in Public Health

(Maximum Score: 50)

Date : 23·02-12 Duration : 1hr.30 min.

1. Explain briefly the following.

a) Latency.
b) Tertiary prevention.
c) Germ theory.
d) Deficiency in the postulates of Robert Koch.
e) Differences among risk factor, promoter, detection factor and prognostic factor.
f) Stable population.
g) Directionality.
h) Advantages and limitations of quasi experiments.

(1+1+2+1.5+4+1.5+2+2 = 15)

2. Stating the basic assumptions clearly write down simple epidemic model. Find out the solution of the deterministic model and hence define epidemic curve. Show that the solution thus obtained is not same that of its stochastic counterpart.

(3+7+3+7 = 20)

3. (a) I am stating the following assumptions for you:
   i) Prey population follow the logistic growth law;
   ii) predation follow the simple law of mass action principle;
   iii) predator death rate is density dependent.

   Taking into consideration the above assumptions, write down the predator prey model. Find out the condition(s) for which both the species will coexist. Interpret your result in ecological sense.

   (b)Write down the assumption(s) by which you can reduce the above predator-prey model into a simple SI (Susceptible-Infected) model. Do you think the assumption which you made to obtain SI model is a realistic one?

(3+6+2+2+2 = 15)

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination : 2011 – 12

M. Stat (2nd Year)

Theory of Finance I

Date: 24 February 2012 Maximum Marks: 20 Duration: 1½Hours

1. Explain the following terms with examples [2 + 1½ + 1½ = 5]
   a) Equivalent Martingale Measure
   b) Futures contract on a Stock
   c) Measure of Relative Risk Aversion

2. Demonstrate or contradict that the following processes are Martingales:
   a) $y_t = \epsilon_t + \alpha\epsilon_{t-1}$ where $\epsilon_t \sim N(0,1)$.
   b) $y_t = \alpha y_{t-1} + \epsilon_t$ where $\epsilon_t \sim N(0,t)$.
   c) $(y_t^2 - t)$ where $y_t \sim N(0,\sigma^2 t)$. [2 + 2 + 3 = 7]

3. Find the probability distribution of the random variable $\int_0^t W_s^2 dW_s$. [4]

4. Suppose $\succcurlyeq$ is a preference relation on the set of all lotteries $\mathcal{P}$ satisfying:

   (i) $\forall\, p,q,r \in \mathcal{P}$ and $a \in (0,1], p \succ q \Rightarrow ap + (1-a)r \succ aq + (1-a)r$.

   (ii) $\forall\, p,q,r \in \mathcal{P}$ if $p \succ q \succ r$ then $\exists\, a,b \in (0,1)$ such that
   $ap + (1-a)r \succ q \succ bp + (1-b)r$.

   Then show that $p \succcurlyeq q \succcurlyeq r$ and $p \succ r \Rightarrow \exists$ unique $a^* \in [0,1]$ such that $q \sim a^*p + (1-a^*)r$. [4]

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination : Semester II (2011-12)

M. Stat. II Year

Survival Analysis

Date: 27.2.2010 Maximum marks: 50 Time: 105 minutes

(This test is open notes. Total mark is 54.)

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 type II censored data from *exponential* (λ) life distribution. Prove that the maximum likelihood estimate of expected life time is unbiased.

   (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. Let $X \sim$ Weibull(λ, p) and $T = [X]$, the largest integer less than or equal to X. Find the p.m.f. and discrete hazard of T and investigate the IFR/DFR property of its distribution. [2+2+2=6]

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. Consider right censored life time data on failure time of a parallel system with two independent components each having life distribution given by the survival function S. Based on n such data, find nonparametric maximum likelihood estimate of $S(t)$ and an estimate of its variance. [3+4=7]

5. Consider the illness-death model, where an elderly subject can move from the 'walking' state (W) to the 'terminally bed-ridden' state (B), with no possibility of recovery, and from either of these states to the 'dead' state (D). A group of n subjects are observed till death, and the data are of the type (B_i, D_i), $i = 1, \ldots, n$, where D_i is the time till death of the ith subject, B_i is 0 if the ith subject is never terminally bed-ridden, and is otherwise equal to the time till this subject becomes terminally bed-ridden.

P.T.O.

(a) Specify the Nelson-Aalen estimator of the cumulative hazard rate of time till death for the subjects, based on $D_1, \ldots, D_n$ alone.

(b) Specify a consistent estimator of the variance of this estimator.

(c) Specify the Nelson-Aalen estimators of the cumulative intensities of transition directly from W to D (without being 'terminally bed-ridden') and from B to D.

(d) Describe a test of the hypothesis of equality of the intensities of transition directly from W to D (without being 'terminally bed-ridden') and from B to D, in terms of the given data.

[3+3+(2+2)+4=14]

Indian Statistical Institute, Kolkata

Midsemestral Examinations : M.Math.II year & M.Stat.II year

Ergodic Theory

Maximum marks : 30 February 29, 2012 Time : 3 hours

Answer all questions

1. Let T be a measure preserving invertible transformation of the probability space $(X, \mathcal{B}, m)$. Say that T has a countable Lebesgue spectrum if there exists $f_0, f_1, f_2, \cdots \in L^2(X, \mathcal{B}, m)$ such that f_0 is the constant function 1 and the family $\{f_0, U_T^k f_j : k = 0, \mp 1, \mp 2, \cdots, j = 1, 2, 3, \cdots\}$ is an orthonormal basis of $L^2(X, \mathcal{B}, m)$.

   (a) Let $\mathcal{P}$ be the spectral measure corresponding to U_T. Show that if T has a countable Lebesgue spectrum then for any $f \in L^2(X, \mathcal{B}, m), < f, 1 >= 0, < f, f >= 1$, the measure $< \mathcal{P}(.)f, f >$ is the Lebesgue measure on $\mathbb{T}$. [4]

   (b) If T has countable Lebesgue spectrum, show that T is mixing. [4]

2. Let G be a compact, connected metric abelian group. Let $Tx = aAx$ be an affine transformation of G (i.e. $a \in G$ and A is a continuous homomorphism of G onto itself.) Suppose that for $x_0 \in X$, the orbit $\{T^n x_0, n \geq 0\}$ is dense in G. Show that if for some character γ of G, and positive integer k, $\gamma \circ A^k = \gamma$, then $\gamma \circ A = \gamma$. [5]

3. Let T be a measurable transformation of the measurable space $(X, \mathcal{B})$ and $\mathcal{M}_T$ the space of all probability measures μ on $\mathcal{B}$ such that μ is $T-$invariant. $\mathcal{M}_T$ is a convex subset of the space of probability measures on $\mathcal{B}$.

   Show that

   (a) if $\mu, \nu \in \mathcal{M}_T, \nu << \mu$, then the Radon-Nikodym derivative $\frac{d\nu}{d\mu}$ is a $T-$invariant function *a.e.* [2]

   (b) if, in addition to the hypothesis in (a), μ is given to be ergodic, then $\nu = \mu$. [2]

   (c) $\mu \in \mathcal{M}_T$ is ergodic if and only if μ is an extreme point of $\mathcal{M}_T$. [3]

4. Let T be the transformation on $[0, 1)$ defined by $Tx =< \frac{1}{x} >$, if $x \neq 0$ and $T0 = 0$, called the continued fraction map and μ the measure given by $\mu(a, b) = \int_0^1 \frac{1}{1+x} dx$, called the Gauss measure.

   (a) Show that T preserves μ. [5]

   (b) Assume the facts (i) and (ii):

   (i) For $x \in (0, 1)$ irrational, $x = \frac{1}{a_1 + \frac{1}{a_2 + \cdots}}$ is the (simple) continued fraction representation of x where $a_1 = [\frac{1}{x}], a_2 = [\frac{1}{Tx}], a_3 = [\frac{1}{T^2 x}], \ldots$.

(ii) T is ergodic for μ.

Now Show that for $a.e.x$, the asymptotic proportion of 1's among the $a_1, a_2, a_3, \ldots$ is a constant. [5]

Applied Multivariate Analysis
Mid-Semestral Examination
M.Stat. II Year, 2011-12
Full Marks - 69+1=70
Time - 3 hrs.

DATE: 29.2.12

Indian Statistical Institute
Kolkata 700 108, INDIA

Attempt all questions:

1. (a) Initially an experiment is set up with one toxic (*Dinophysis sp.*) and two non-toxic phytoplankton (*Chaetocerous gracilis* and *Biddulphia regia*) species. All the three phytoplankton growth profiles are monitored with regular recorded biomass counts of 10 samples on each of the 8 experimental days. The phytoplankton are collected from the deltaic region of river Subarnarekha ($87^o31'$E and $21^o37'$N) and the isolation is done in the laboratory. Species culture are maintained at optimal conditions in the laboratory although the species might have a negligible amount of genetic variation. As per the literature in general, the ratio of the final average abundance of the toxic species to the first non-toxic species is $2K_1$ where as the similar ratio for the toxic species to the second non-toxic species is $3K_2$, where $K_1, K_2 \geq 1$. How can he verify these claims based on a statistical test under two different conditions? Can you suggest any alternative test procedure for the same hypothesis ?

(b) Construct an appropriate multivariate testing procedure based on an appropriate sampling scheme to judge whether the excessive T-20 and ODI matches lead an impact upon the performance of Indian cricket team.

$12 + 6 + 8 = 26$

2. What is principal component analysis ? Give the geometrical interpretation and shortcomings of such technique. Let, a random vector has the covariance matrix $\sum$ with non-distinct eigen values and $corr(X_i, X_k) = \rho$, $i \neq k$. Suggest a suitable testing procedure (excluding LR test) in testing $H_0 : \rho = \rho_0$ against $H_1 : \rho \neq \rho_0$. Can you suggest any result to identify whether X_1 and X_2 are equally important to it's 1st principal component ?

$4 + 6 + 4 + 4 = 18$

3. (a) Define Relative Growth Rate (RGR). Assuming Gaussion set up among size variables for the above experiment (as stated in question number (1)) with proper Koopman's structure in the covariance matrix suggest two estimators of RGR based on a sample of size n. Find asymptotic mean and variance for any one of the two estimators. Suggest a suitable test procedure for testing the equality of expected RGR for three species at final time point.

(b) Responses of swell diameters (ranging from 0 - 8 mm) are recorded on a skin prick test data of 100 atopic patients to judge the patients typical allergenic responses to 21 allergens. Not all allergens are sensitive to all the patients. Suggest a suitable procedure to construct a distance matrix based on a dissimilarity measure of all pair of allergens. Also describe a suitable method in

identifying similar allergens using the above distance matrix so that the information can be used in immunotherapy treatment in reducing the dose of vaccination.

$$2 + 3 + 7 + 4 + 4 + 5 = 25$$

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination : 2011 – 12

M. Stat II year : MSP

Statistical Inference – II

Date : 02.03.2012 Maximum Marks : 15 + 30 = 45 Duration : 2 Hours 30 Minutes

Note : Answer Part I and Part II in separate Answerscripts.

Part – I

Answer all questions. Maximum you can score is 15.

A. Let $X_1, \ldots\ldots, X_n$ be iid $U(0, 1)$ variables. Let $k > 0$ be a real number, and $n \geq 2$. Find

$$E\left(\frac{X_1 + X_2}{X_{(n)}}\right)^k$$

where $X_{(n)} = \max(X_1, \ldots\ldots, X_n)$. Simplify as far as possible.

[7]

B. Let f be a measurable function from $(X, \mathcal{A})$ into $(Y, \mathcal{B})$. Let

$$\sigma(f) = \{f^{-1}(B) : B \in \mathcal{B}\}.$$

Let g be a real-valued measurable function on $(X, \mathcal{A})$. Show that g is $\sigma(f)$ - measurable iff there is a real-valued measurable function ϕ on $(Y, \mathcal{B})$ such that $g = \phi(f)$; in case g is also non-negative, ϕ can be chosen to be non-negative, while for a bounded g, ϕ can be chosen to be bounded.

[5 + 2 = 7]

C. Define clearly a sufficient σ - field. Show that if $\mathcal{B}$ is sufficient for M and $\mathcal{B}_1 = \mathcal{B}[M]$, then $\mathcal{B}_1$ is also sufficient for M.

[1 + 6 = 7]

PART II

This part carries 30 points with five problems carrying 6 points each. Show all working and write clearly. All the best!

1) State the de Finetti's theorem.

2) Let $x_1, \ldots, x_n$ be i.i.d. with a Student's t distribution $t(\alpha, \mu, \sigma^2)$, α being a known constant and μ, σ being unknown parameters. Set the prior $\pi(\mu, \sigma^2) \propto \sigma^{-2}$ on the parameters.
a) Show that the t distribution can be expressed as a scale mixture of Normals.
b) Use this fact to develop a Gibbs sampling algorithm to sample from the posterior of the parameters.

3) A total of $(m+n)$ light bulbs are tested in two independent experiments. In the first experiment involving n bulbs, the exact life times $y_1, \ldots, y_n$ of all the bulbs are recorded. In the second involving m bulbs, the only information available is whether these bulbs are still burning at some fixed time $t > 0$. Assume that the distribution of lifetime is exponential with mean $1/\theta$ and use $\pi(\theta) \propto \theta^{-1}$. Find the posterior mode using the E-M algorithm.

4) Consider k independent responses $y_1, \ldots, y_k$ and k predictors $x_1, \ldots, x_k$ in $\Re$. Let $y_i \sim \text{Bin}(n_i, p_i)$ with $p_i = H(\beta_0 + \beta_1 x_i)$, H being the logit function and $\beta = (\beta_0, \beta_1)$ are unknown parameters.
a) What is a sufficient statistic for β?
b) Using a bivariate Normal prior on β, derive an approximate Normal posterior for β. Justify your construction.
c) Construct a M-H algorithm to sample from the exact posterior of β.

5) Consider k independent responses $y_1, \ldots, y_k$ in $\Re$ and k predictors $x_1, \ldots, x_k$ in $\Re^p$. Let $y_i \sim \text{Bin}(n_i, p_i)$ with $p_i = H(\beta' x_i)$, H being the standard normal cdf. while $\beta \in \Re^p$ are unknown parameters. Set a multivariate Normal prior $N_p(\mu, \Sigma)$ on β.
a) By introducing suitable latent variables, develop a Gibbs sampler to sample from the posterior of β.
b) Employ Rao-Blackwellization to improve your estimator.

INDIAN STATISTICAL INSTITUTE

Semestral Examination: 2011-12 (Second Semester)
Master of Statistics (M. Stat.) II Year
Stochastic Processes I

Instructor: Parthanil Roy

Date : 23.04.2012

Total Points: 55 **Duration: 3 hours**

- **Please write your roll number on top of your answer paper.**
- **There are five problems in this exam. Problem 1 is worth 5 points and will count towards your Assignments score. Problems 2 - 5 (worth 50 points in total) will count towards your Semestral Exam score.**
- **Show all your works and write explanations when needed.**
- **You are NOT allowed to use class notes, books, homework solutions, list of theorems, formulas etc. If you are caught using any, you will get a zero grade in the course.**

1. (5 points) Let (X, d) be a Polish space and $\{Y_n\}_{n\geq 1}$ and $\{Z_n\}_{n\geq 1}$ be two sequences of X-valued random variables defined on the same sample space. Suppose $\{Y_n\}_{n\geq 1}$ is a tight family and $d(Y_n, Z_n) \to 0$ in probability as $n \to \infty$. Show that $\{Z_n\}_{n\geq 1}$ is also a tight family.

2. (10 points) Suppose (X, d) is a Polish space. Recall that a function $f : X \to \mathbb{R}$ is called upper semicontinuous at a point $x_0 \in X$ if for each $\epsilon > 0$, there exists $\delta > 0$ such that $d(x_0, y) < \delta$ implies $f(y) < f(x_0) + \epsilon$. Let $\mathcal{P}(X)$ be the class of probability measures on X and "$\Rightarrow$" denotes the weak convergence in $\mathcal{P}(X)$. Suppose $P_n \Rightarrow P$ in $\mathcal{P}(X)$ and $f : X \to \mathbb{R}$ is a bounded function, which is upper semicontinuous at every $x \in X$. Show that

$$\limsup_{n\to\infty} \int f dP_n \leq \int f dP.$$

3. Let $\mathcal{B}$ be the Borel σ-field on $C[0, \infty)$ and $Z : C[0, \infty) \to \mathbb{R}$ be a bounded and $\mathcal{B}$-measurable function.

   (a) (7 points) Show that for all $x \in \mathbb{R}$ and for all $s \geq 0$,

   $$E_x(Z|\mathcal{F}_s^+) = E_x(Z|\mathcal{F}_s).$$

   (b) (3 points) Fix $x_0 \in \mathbb{R}$ and $s_0 \geq 0$. Using (a) or otherwise, show that for all $A \in \mathcal{F}_{s_0}^+$, there exists $B \in \mathcal{F}_{s_0}$ such that

   $$P_{x_0}(A \Delta B) = 0.$$

   Here the notations are as introduced in the class.

4. Let $\{B_t\}_{0\leq t\leq 1}$ be a Wiener process defined on a sample space $(\Omega, \mathcal{F}, P)$. For each sample point $\omega \in \Omega$, define a number

   $$V(\omega) = \int_0^1 B_s(\omega) ds.$$

   (a) (5 points) Show that V is well-defined and it is a random variable.

   (b) (10 points) Calculate the characteristic function of V. **[P.T.O]**

5. (15 points) Consider the probability space $(C[0,1], \mathcal{C}, W)$, where $\mathcal{C}$ is the Borel σ-field on $C[0,1]$ and W is the Wiener measure. Define a random variable ξ on this space as

$$\xi(\omega) = \omega(1), \quad \omega \in C[0,1].$$

Find a regular conditional probability given the random variable ξ. Justify your answer.

INDIAN STATISTICAL INSTITUTE

Second Semestral Examination: (2011 – 2012)

M.Stat. (BSDA)

Course : Statistical Methods in Public Health

Date: 23.04.2012 Maximum Marks : 100 Duration: 3 Hrs

(Answer Part A and Part B in two separate answer booklets.)

PART – A

(Answer all questions. Maximum allowable score is 60)

1. Indicate whether each of the following computed indices is an incidence density, a cumulative incidence, a point prevalence, a period prevalence or an odds for the disease of interest.

a) The number of adult men in Mumbai identified with psychiatric problems during 1970, divided by the total number of adult men in Mumbai in 1990.

b) The number of sudden infant deaths in Sundarban in 1995, divided by the number of live births in Sundarban in 1995.

c) The number of children born with congenital heart defects in Kanpur in 1992, divided by the number of live births in Kanpur in 1992.

d) The number of persons who resided in Chennai on Jan. 1, 1990 and who developed colon cancer during 1990, divided by the total number of disease-free persons who were Chennai residents on January 1, 1990.

e) The number of 60-64 year old Kolkata residents who had a stroke in 2000, divided by the total number of 60-64 year old Kolkata residents on July 1, 2000.

f) The number of legally deaf males in Delhi in 1995, divided by the number of males who were not legally deaf in Delhi in 1995.

[1x6 = 6]

2. (a) Justify the following statements:

(i) If the observations are based on a longitudinal study, the cross-sectional least squares estimates of the regression parameters will be biased.

(ii)The linear regression analysis of longitudinal data tends to yield more efficient regression parameter estimates than the regression analysis of cross-sectional data.

(b) Develop the exponential correlation model for longitudinal data analysis.

(4+5+4 = 13)

3. (a) Which distribution function is useful for nearest neighbourhood analysis?

(b) Justify its use, stating clearly the principles and the assumptions behind.

(c) What do you mean by "transformation of map"?

(d) Why is such transformation necessary before doing the nearest neighbourhood analysis?

(1+8+3+2 = 14)

4. (a) The figures with question marks have been found doubtful. Do you agree with that? If so, change the figures appropriately. (The symbols have their usual significance.)

j			No. of Deaths	Actuarial Method			Density Method
	N_{0j}	PT_j	$(D_x+D_{xy})_j$	W_j	MD_j	R_j	R_j
1	12	11.5	0	1	0.000	0.000	0.000
2	11	9.5	1	3?	0.105	0.100	0.100?
3	8	5.5	1?	4	0.172?	0.152?	0.182?
4	3	3.5?	0	0?	0.000	0.000	0.000
5	3	2.0?	0	2	0.000	0.000	0.000
Total	-	31.5	2	10	0.058?	-	-

(b) Calculate $R_{(t0,t5)}$ by Density Method.

[7+2=9]

5. (a) Under steady state, show that $P \approx ID(\bar{T})$ for a rare disease, where prevalence of the disease, the incidence density and the mean duration of the disease are represented by P, ID and $\bar{T}$ respectively.

(b) Consider a heterogeneous population, stratified into several groups. Prove that the variance of the estimated number of deaths in the population will be larger if the stratification of the population is ignored.

(4+4 = 8)

6. Define relative risk. With Poisson variable assumption for individual risk, develop a test procedure relative risk.

(2+8 = 10)

PART – B

(Answer all questions. Maximum allowable score is 40)

1. 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 evaluating mean and variance of the process. (3 + 7 == 10)

2. a) Write down the basic model of General epidemic and hence find out the basic reproductive number.
   b) Show that for this model the disease dies out for the lack of infectives and not the lack of susceptibles. (4 + 6 =10)

3. State and prove Bendixon Negative Criterion to show there are no periodic orbits of any 2-D autonomous differential equation. Illustrate the above criterion in a Simple Epidemic Model. (10)

4. a) Define basic Reproduction Number.
   b) Write down a simple Malaria Model. Hence find out the basic reproduction number by using next generation matrix method. (2 + 8 =10)

-END-

INDIAN STATISTICAL INSTITUTE

Semestral Examination : Semester II (2011-12)

M. Stat. II Year

Survival Analysis

Date: 26.4.2012 Maximum marks: 100 Time: $3\frac{1}{2}$ hours

This test is open notes. Books cannot be used and notes cannot be exchanged. Refer to your notes properly, but do not reproduce derivations from there. Answer as many as you can. Total mark is 105.

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 estimate (MLE) of the hazard rate based on random right censored data (measured in weeks) from an exponential distribution is 0.04. The estimate remains the same if the same data are measured in days.

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

   (d) A series system is constructed with two independent components having identical life distribution 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 $2s(t)$.

   (e) For the semiparametric model $\lambda(t, z(t)) = \lambda_0(t)e^{z(t)\beta}$ with time dependent covariate $z(t)$, the partial likelihood for right censored data does not depend on the exact times of failure.

   (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. Prove that the mean residual life of a unit surviving upto time t_0 is

$$\frac{1}{\bar{F}(t_0)} \int_{t_0}^{\infty} \bar{F}(u)du,$$

   where $\bar{F}(\cdot)$ denotes the survival function of the unit. Then obtain the same for a product that has a constant hazard α up to time t_1 (known positive constant), after which the hazard changes to β. [5+5=10]

3. Consider type II censored data from Weibull$(\lambda, p = 2)$ life distribution. Find an exact 95% confidence interval for λ. Suggest a graphical test for the validity of the model. [5+3=8]

4. Suppose the life times of a product, being manufactured in two factories (A and B) and also in day and night shifts (D and N), have exponential life distribution with parameters λ_{ij}, for $i = A, B$ and $j = D, N$. Based on right censored life time data in each factory and each shift, it is of interest to investigate if there is any factory effect.

We wish to use the score test for this purpose. Write down a reparametrization and hence the null hypothesis. Then, obtain the score test based on this model.

[3+2+10=15]

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, $i = 1, \cdots, K$, and $\lambda_1(t)$ is totally unknown and arbitrary. Suggest a graphical test for the model $\mathcal{M}$. Develop a test for homogeneity giving full details. If $\lambda_1(t)$ is a known function involving a constant parameter θ, indicate how to test for homogeneity. [2+5+2=9]

6. Consider n uncensored data from accelerated failure time model given by the error distribution $f(\epsilon) = \frac{1}{2}e^{-|\epsilon|}$ (double exponential density). Assume that the locally most powerful rank test for the covariate effect is of the form $\sum_{i=1}^{n} c_i z_{(i)}$, where $z_{(i)}$ is the covariate value corresponding to the ith ordered observation. Prove that, using the approximation $E[g(u)] \approx g(E[u])$, $c_i = sgn[2i - (n+1)]$, for $i = 1, \cdots, n$, where *sgn* denotes the *sign* function. [10]

7. 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. [6+3+3=12]

8. Suppose that you want to decide whether the hazard rates in two populations are proportional to one another. You have randomly right censored samples from the first population. The lifetimes in the second population are *known* to have the exponential distribution with mean 100.

   (a) Using the Nelson-Aalen estimator of the cumulative hazard function of the first population, suggest a plot that can be used to determine graphically whether the hazards are proportional.

   (b) Indicate, with reasons, the expected shape of the plot when the true hazard rate of the first population is monotonically increasing.

   (c) Using the Nelson-Aalen estimator of the cumulative hazard function of the first population, suggest an analytical test for this problem, mentioning explicitly the test statistic, an estimator of its variance, its asymptotic distribution, and the requisite decision rule. [3+3+8=14]

9. Consider a software having N hidden bugs. The times to detect these bugs are assumed to be independent and exponentially distributed with failure rates $\lambda_1, \cdots, \lambda_N$. As soon as a bug is detected, it is immediately corrected. Write this as a multiple failure time model with competing risks by specifying the different cause-specific hazard rates. What is the probability that the first failure is due to jth bug? What is the conditional probability that there is no further failure up to time $t+s$ given the first failure at time s due to jth bug? Note that the 'failure' here means detection of a bug. [4+2+3=9]

INDIAN STATISTICAL INSTITUTE
Second Semestral Examination: 2011-12

MSTAT II

Directional Data Analysis

Date: 26-04-12 Maximum marks: 100 Time: 3 hrs.

Problem 5 is compulsory. Attempt any three from Problems 1 - 4. Show all your work.

1. [(10)+(5+10)] (a) Obtain the characteristic function of the symmetric wrapped stable family of distributions, SWS (μ, ρ, α).
(b) Identify the wrapped Cauchy (WC) distribution as a member of SWS (μ, ρ, α). Show that the p.d.f. of the WC distribution can be written in the form of a single term involving the circular random variable θ, as compared to the infinite sum representation for the SWS density.

2. [5+15+5] Let, $f(\theta) = K. [1 + 2 \rho \cos(\theta - \mu) + 2\rho^2 \cos 2(\theta - \mu)]$, where K is the normalizing constant, $\theta \in [0, 2\pi)$.
(a) Obtain K. (b) Derive an optimal invariant test for Isotropy vs. the underlying density as $f(\theta)$. (c) Show that this test is also an optimality robust test against a wide class, to be identified by you, of distributions.

3. [6+10+9] Let n i.i.d. observations be available from the circular normal distribution CN (μ, κ).
(a) Obtain the MLEs of (μ, κ) and verify whether these coincide with the corresponding TMMEs. (b) Obtain an unconditional asymptotically optimal test for $H_0: \mu = 0$ vs $H_1: \mu > 0$, when κ is unknown. (c) Let $\kappa = 1$. Obtain a test for the change-point problem that the underlying distribution has a changed mean direction μ_1 from the original mean direction μ_0 at some unknown point between 1 to n.

4. [12 + 13] (a) Describe two methods of constructing distributions on the torus, giving one example of each. State explicitly the theorems and results you need (proofs are not needed).
(b) Consider the three parameter Bivariate circular normals conditionals (BCNC) distribution. Obtain consistent estimators of the parameters, without explicitly deriving the normalizing constant.

5. (Take Home Problem) [10+5 +6+4] (a) Derive explicitly an asymptotically optimal test that a given set of i.i.d. observations come from a circular normal distribution vs. these come from an asymmetric circular normal (ACN)distribution, including all the coefficients arising in the test statistic. (b) Obtain an equivalent test statistic that avoids explicit derivation of the coefficients in (a). (c) Illustrate your tests by two real-life examples – one supplied to you and another of your own choice. (d) Conduct Goodness-of-Fit tests on the data sets to validate the choice of ACN as a model for the data sets.

****

INDIAN STATISTICAL INSTITUTE
KOLKATA - 700108
Semestral Examination : M. Stat. II – BSDA Sem II 2011-12
Statistical Methods in Biomedical Research

Group B

Full Marks : 40

Note : Answer any two questions. Marks allotted to a question are indicated in brackets [ ] at the end.

1. **What are direct assays? How do you estimate relative potency from such assays? State and prove the Fieller's theorem for fiducial limits and use it to obtain the fiducial limits for the relative potency estimate from direct assays. Discuss the experimental designs that you would recommend for such assays with justifications.**

   [2+2+8+5+3=20]

2. **Suppose a 2k-point symmetrical parallel line assay is to be conducted in n randomized complete blocks. Develop the expressions for the relative potency estimate and its fiducial limits as well for such an assay. Also develop the analysis of variance for the validity tests for this design, indicating clearly the various expressions for the sum of squares involved.**

   [(5+7)+8=20]

3. **Describe a change-over trial, and give a construction of a balanced change-over design for v treatments in v periods. Develop for such designs the appropriate analysis of variance, indicating clearly how the various sums of squares are to be computed.**

   [(2+6)+12=20]

INDIAN STATISTICAL INSTITUTE

Final Examination : Semester II (2011-2012)

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

Subject Name : Statistical Methods in Biomedical Research

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

Group A

1. Which of the following two type I error spending functions is stricter for early stopping and why?

$$\begin{aligned}(i)\ \alpha_1^*(t) &= \alpha t,\\ (ii)\ \alpha_2^*(t) &= \alpha t^{3/2}.\end{aligned}$$

[4]

2. In the Zelen's play-the-winner rule, if the two treatments under consideration have success probabilities 0.7 and 0.4 respectively, find the unconditional probability that the 10-th patient will result in a success. Modify this probability if it is known that the 4-th patient is treated by the first treatment. How the probability changes if, in addition, it is known that the first patient was treated by the second treatment? [4+3+2=9]

3. Construct an example to illustrate how the imbalance with respect to prognostic factors induces allocation bias. How can Pocock's minimization rule be implemented to get rid of this difficulty? [4+3=7]

INDIAN STATISTICAL INSTITUTE

Second Semestral Examination : 2011 – 12

Course Name: M. Stat (2nd Year)

Subject Name: Theory of Finance I

Date: 30 April 2012 Maximum Marks: 100 Duration: 3Hours

Note: This paper carries 105 marks. Answer as many as you can. The maximum you can score is 100

1. Discuss the following statements:
   a) Speculation in futures markets is pure gambling. It is not in the public interest to allow speculators to trade on a futures exchange.
   b) If the minimum variance hedge ratio is calculated as 1.0, the hedge must be perfect.
   c) The duration measure of a bond tells us about the sensitivity of the bond price to interest rates.
   d) If most of the call options on a stock are in the money, it is likely that the stock price has risen rapidly in the last few months.
   e) Early exercise of an American Put is a trade-off between the time value of money and the insurance value of a put. [5 X 5 = 25]

2. Suppose that F_1 and F_2 are two futures contracts on the same commodity with times to maturity, t_1 and t_2, $t_2 > t_1$. Prove that $F_2 \leq F_1 e^{r(t_2 - t_1)}$. [8]

3. Consider an exchange traded call option to buy 100 shares with a strike price of Rs. 400 and maturity in two months. Explain the effect on the contract of the following:
(i) a 10% cash dividend, (ii) a 20% stock dividend. [5 + 5 = 10]

4. Suppose that c_1, c_2 and c_3 are the prices of European call options on the same underlying asset with strike prices K_1, K_2 and K_3, respectively; $K_3 > K_2 > K_1$ and $K_3 - K_2 = K_2 - K_1$. All options have the same maturity. Show that $c_2 \leq 0.5\,(c_1 + c_3)$. [10]

5. Show that to match the volatility of a stock, σ, with the variance of the one step Binomial tree model; we need to have the up step, $u = e^{\sigma\sqrt{\Delta t}}$ and the down step, d= $e^{-\sigma\sqrt{\Delta t}}$. [10]

6. Consider options on a non dividend paying stock when the stock price is Rs. 300, the strike price is Rs. 290, the risk free interest rate is 5%, volatility is 25% per annum and the time to maturity is three months. Find the price of European call & put and American call. [5 + 4 + 3 = 12]

7. (a) In the mean-variance portfolio choice problem, define the efficient portfolio frontier and the minimum variance portfolio.
(b) In the standard Capital Asset Pricing Model, borrowing and lending rates are equal. What is the consequence if borrowing rate is strictly larger than the lending rate? [10 + 10 = 20]

8. Consider the nonlinear stochastic differential equation

$$dX_t = rX_t\,(K - X_t)\,dt + \beta X_t\,dW_t\,;\; X_0 = x > 0$$

Verify that $X_t = \dfrac{\exp\{\left(rK - \frac{1}{2}\beta^2\right)t + \beta W_t\}}{\frac{1}{x} + r\int_0^t \exp\{\left(rK - \frac{1}{2}\beta^2\right)s + \beta W_s\}ds}$; $t \geq 0$ is the unique (strong) solution. [10]

INDIAN STATISTICAL INSTITUTE

Second Semestral Examination : (2011-12)

M.Stat. 2nd Year

ASYMPTOTIC THEORY OF INFERENCE

Date: 30 April, 2012 Maximum Marks: 100 Duration: $3\frac{1}{2}$ Hours

Answer as many questions as you can. The Maximum you can score is 100.

1. (a) Describe Fisher's notion of asymptotic efficiency of estimators and comment on this in the light of Hodges' example. Write a short note on the Hajek-Le Cam theory of efficient estimation.

(b) Let $\{P_\theta^n, \theta \in R\}$, $n \geq 1$, be a sequence of statistical experiments such that for a fixed $\theta \in R$ and for all $u \in R$,

$$\log \frac{dP^n_{\theta+u/\sqrt{n}}}{dP^n_\theta} = u\Delta_n - \frac{1}{2}u^2 I + o_p(1)$$

where I is a finite positive number and Δ_n is a sequence of random variables converging in distribution to some random variable Δ (under θ). Show that for all $u \in R$, $\{P^n_{\theta+u/\sqrt{n}}\}$ is contiguous to $\{P^n_\theta\}$ if and only if Δ follows $N(0, I)$.

[14+12=26]

2. (a) Define a regular sequence of estimators. State the Hajek-Inagaki convolution theorem and prove it only for the special case when the asymptotic distribution of the (normalized) estimator is normal.

(b) Use the convolution theorem to find, under LAN condition, a (non-trivial) lower bound to the limiting risk of a regular estimator for a subconvex loss.

[14+12=26]

3. Let $X_1, X_2, \ldots, X_n$ be i.i.d. with a common distribution function F. Consider the U-statistic U_n based on $X_1, \ldots, X_n$ and a kernel $h(x_1, \ldots, x_m)$, $m \leq n$, for estimation of $T(F) = E_F h(X_1, \ldots, X_m)$. Let $V_n = T(F_n)$ be the corresponding V-statistic. Show, under suitable conditions on the moments of $h(X_{i_1}, \ldots, X_{i_m})$ where $i_1, \ldots, i_m \in \{1, 2, \ldots, n\}$, that $\sqrt{n}(U_n - V_n) \xrightarrow{p} 0$.

[12]

4. Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with distribution function $F \in \mathcal{F}$ where $\mathcal{F}$ is the class of all distribution functions on R. Consider a real valued functional $T(\cdot)$ on $\mathcal{F}$.

(a) Define the k-th order Gateaux differential of $T(\cdot)$ at F in the direction of some fixed $G \in \mathcal{F}$.

(b) Find the k-th order Gateaux differential of the functional $T(\cdot)$ considered in Question 3 and hence express the first order Gateaux differential of $T(\cdot)$ at F in the direction of the sample distribution function F_n as an average of i.i.d. random variables.

Show, under suitable assumptions, that $\sqrt{n}(T(F_n) - T(F))$ is asymptotically normal. Use the asymptotic equivalence of a U-statistic and its projection and that of a U-statistic and a V-statistic to prove this result.

(c) Suppose $T(\cdot)$ is Frechet differentiable at F with respect to some norm. Describe how one can prove asymptotic normality of $\sqrt{n}(T(F_n) - T(F))$ in this case using suitable results on F_n.

[2+12+6=20]

5. Let $X_1, X_2, \ldots, X_n$ be a random sample from an $N(\theta, 1)$ population. Consider the problem of testing $H_0 : \theta = 0$ vs $H_1 : \theta = \delta/\sqrt{n}$, $\delta > 0$ (fixed). Find the limiting power of the corresponding sign test.

[14]

6. **Answer either (a) or (b)**

(a) Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta, \sigma^2)$. Let $T_n = \bar{X}/s$ where $\bar{X} = (X_1 + \cdots + X_n)/n$ and $ns^2 = \sum(X_i - \bar{X})^2$. Find the Bahadur slope of T_n when the testing problem is

$$H_0 : \theta = 0 \text{ vs } H_1 : \theta > 0.$$

State and prove the result(s) you are using.

[7+(1+4)=12]

(b)

(i) Let $X_1, X_2, \ldots$ be i.i.d. $Bin(1, \theta)$. Find

$$\lim_{n \to \infty} n^{-1} \log P(X_1 + \cdots + X_n \geq n a_n)$$

where $\lim a_n = a$, $\theta < a < 1$.

(ii) Let $X_1, X_2, \ldots$ be i.i.d. random variables following $U(0,1)$. Let $D_n = \sup\{|F_n(t) - t| : 0 \leq t \leq 1\}$ where F_n is the sample distribution function based on $X_1, \ldots, X_n$. Find

$$\lim_{n \to \infty} n^{-1} \log P(D_n \geq a), \ 0 < a < 1.$$

You may assume that the limit is a continuous function of a.

[5+7=12]

INDIAN STATISTICAL INSTITUTE
M. Stat. II Year, 2011-12
Semestral Examination
Pattern Recognition and Image Processing

Date: 02·05·12 Maximum Marks: 100 Duration: 195 minutes

Note: This paper carries 106 marks. Answer as much as you can.

1. Draw the solution tree for branch and bound feature selection algorithm if 3 features are to be selected from 7 features. [6]

2. Let $\underline{X}^t = (X_1, X_2, X_3)$ be a random vector with dispersion matrix $\begin{pmatrix} 2 & 0 & -1 \\ 0 & 8 & 0 \\ -1 & 0 & 2 \end{pmatrix}$. Find the first two principal components of $\underline{X}$. [10]

3. Let there be 2 classes. Let $A_1 = \{(0,0,0),(0,0,1)\}$ *and* $A_2 = \{(1,-1,0),(-2,1,-1)\}$. Let the elements in A_i be from the i-th class for all i. Let the initial hyperplane under consideration be $x + y + z + 1 = 0$ and the learning rate be 0.05. Find the separating hyperplane between A_1 and A_2 with the help of perceptron algorithm. [12]

4. Describe any two feature selection algorithms. [8]

5. Describe the Canny edge detection algorithm. [12]

6. Describe the algorithm for detecting line segments in a binary image using Hough transform. [12]

7. Suppose there are n labeled observations and a multilayer perceptron (MLP) having 3 nodes in the input layer, 3 nodes in the hidden layer, and 2 nodes in the output layer. Let the node transfer function be the sigmoid function, and the way of learning be the batch mode learning.

(a) Draw the network for the said MLP.

(b) Write the expression for error at the output nodes..

© Write the expression for weight update using back propagation method for the connections joining the hidden layer with the output layer for the above error expression. [2+3+7]

(P.T.O)

8. A white object is represented in a binary image as shown below with `x' denoting a white pixel.

(a) Describe the 8-connected boundary of the object shown in the image by its chain code.

(b) Describe a thinning method for an object, represented by a binary image.

©Provide the final output of applying the thinning algorithm on the figure below.

[3+6+8]

		x			
	x	x	x		
	x	x	x	x	
		x	x	x	
		x	x	x	

9. (a) Define VC dimension of a set of functions.

(b) Define `` support vectors" for a two class classification problem. [4+3]

10. Apply histogram equalization on the following frequency distribution of gray values and provide the final histogram.

Gray value	Frequency of occurrence
0	1
1	5
2	0
3	7
4	9
5	5
6	2
7	1

[10]

INDIAN STATISTICAL INSTITUTE

Second Semester Examination: (2011-2012)

MS(QE) I & MSTAT II

Microeconomics II

Date: 03.05.2012 Maximum Marks: 60 Duration: 3 hrs.

Note: Answer Group A and Group B in separate answerscripts.

Group A

Note: Answer questions 1 and 2 and either 3 or 4.

(1) (a) Suppose there are two securities whose returns are specified as follows:

$$r^1 = (4, 4, 4, 0)$$
$$r^2 = (0, 0, 0, 2)$$

The price of security 1 is 2.3 and the price of security 2 is 0.2. Also, one can borrow or lend any amount at a risk free interest rate of 20%. Are the asset prices arbitrage free? **(5)**

(b) Suppose there are two securities whose returns are specified as follows:

$$r^1 = (1, 2)$$
$$r^2 = (2, 1)$$

It is known that the price of security 1 is 2 and that of security 2 is 3. Find the state prices μ_1, μ_2. **(5)**

(2) Consider an Arrow-Debreu economy with two goods, two periods and two states. Realization of endowments and consumption take place in the second period. In the first period, contingent markets open. Let p_{ls} represent the Arrow-Debreu price of commodity l in state s. Suppose these prices are given by $p_{11} = 1$, $p_{21} = 2$, $p_{12} = 2$, $p_{22} = 3$. Consider a particular agent with an endowment vector $(w_{11}, w_{21}, w_{12}, w_{22}) = (1, 2, 2, 1)$ and a consumption vector $(x_{11}, x_{21}, x_{12}, x_{22}) = (1.5, 1.5, 1.5, 1.5)$.

(a) Show that the consumption vector is feasible.

(b) Suppose there are two securities which pay $(1, 2)$ and $(2, 1)$ units of good 1 in the two states. Find the portfolio which supports the consumption vector. [Fractional purchase of securities is allowed.] **(5+5=10)**

(3) Define arbitrage free security prices. Suppose security prices are arbitrage free and there are three securities with return vectors r^1,r^2,r^3 and prices q^1,q^2,q^3. Show that if $r^1 = \alpha r^2 + \beta r^3$ then $q^1 = \alpha q^2 + \beta q^3$. **(3+7=10)**

(4) Define upper semi continuity and lower semi continuity. Stating your assumptions prove that the budget correspondence is both upper semi continuous and lower semi continuous. **(1+1+4+4=10)**

Group B

Note: Answer all the questions.

(1) Show that in any sub-game perfect Nash equilibrium of the screening game with unknown worker types, the low ability worker accepts $(\theta_L, 0)$ and the high ability worker accepts $(\theta_H, t^{(1)})$, where $t^{(1)}$, the task level assigned to the high type, satisfies $\theta_H - c(t^{(1)}, \theta_L) = \theta_L - c(0, \theta_L)$. Here the marginal (average) productivity of a worker is $\theta \in \{\theta_L, \theta_H\}$ with $0 < \theta_L < \theta_H < \infty$, the probability that a worker is of high type is $\gamma \in (0, 1)$ and the opportunity cost of accepting employment to each type of worker is zero. **(20)**

(2) Set up the problem of regulating a monopolist and derive the mechanism (or contract) when the constant marginal cost and fixed cost of the monopolist are common knowledge. **(10)**

Applied Multivariate Analysis

End-Semester Examination

M.Stat II year (2011 - 2012) DATE – 04.05.12

Full Marks - 100

Time - 3 hours 30 mins.

1. (a) Let $g(x)$ be a given integrable function on the real line. Show that the integral $\int_R g(x)dx$ is minimised with respect to R for $R = \{x : g(x) < 0\}$.

   (b) Suppose we have a single population divided into two mutually exclusive and exhaustive subgroups P_1 and P_2. Suppose further that a proportion π_i of individuals belong to group P_i and $x \sim f_i(x)$, if x is a member of P_i. The range of x is partitioned into two disjoint sets R_1 and R_2, and we consider the rule where x is assigned to P_1 if $x \in R_1$ and to P_2 otherwise.

   (i) Derive the total probability of misclassification under this rule in terms of π_1, R_1, f_1 and f_2.

   (ii) Using the result proved in (a) above, derive the optimum classification rule that minimises the total probability of misclassification.

   (c) Suppose we have $\pi_1 = \pi_2 = 1/2$ and two bivariate normal groups with common covariance matrix $\Sigma = \begin{bmatrix} 6 & 2 \\ 2 & 1 \end{bmatrix}$ and respective means vectors $\mu_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and $\mu_2 = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$. Derive a discriminant function for assigning a new observation x to one of these two groups.

   Find an expression for the probability of misclassifying an individual using this rule when $c(1|2) = c(2|1) = 1$, where $c(i|j)$ is the associated cost for misclassifying an individual in group j to group i.

   If $x = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, to which group would you assign this individual?

   [5 + 5 + 5 + 4 + 4 + 2 = 25]

2. To predict the average yield of plant, the data on three commonly known predictors such as phosphorus concentration, soil pH level and soil moisture content are recorded. Two additional predictors, such as Sulphur concentration and soil microbial load are also available for

the same experiment. Suppose we wish to compare the relative fit of two models that assume the same covariance structure (Autoregressive order 1 (AR(1)), but have different vectors of fixed effects. The first model involves a vector β_f with both commonly known and additional predictors, denoting as full model while the reduced model involves a vector β_r with commonly known predictors alone. Assuming a general linear model set up $Y = X\beta + e$, where the vector of residual errors e follows multivariate normal with mean zero and covariance matrix V, construct a likelihood ratio test. Show that the test statistic has the form,

$$LR = \frac{1}{(1-\widehat{\rho}^2)^2}[(y-\widehat{y_r})'[\widehat{\rho}^2 C_1 + \widehat{\rho} C_2 + I_p](y-\widehat{y_r}) \\ -(y-\widehat{y_f})'[\widehat{\rho}^2 C_1 + \widehat{\rho} C_2 + I_p](y-\widehat{y_f})]$$

where $\widehat{y_f} = X_f\widehat{\beta_f}$, $\widehat{y_r} = X_r\widehat{\beta_r}$ are the predicted means and β_f, β_r are the estimated regression coefficients vectors under the full and reduced models respectively and $\widehat{\rho}$ is a suitable estimate of ρ. $C_1 = diag(0, 1 \ldots 1, 0)$ and C_2 is a tridiagonal matrix with 0 on the diagonal and and 1 on the first superdiagonal and on the first subdiagonal.

Also suggest a large sample distribution and critical region of the test statistic.

$[4 + 6 + 2 = 12]$

3. Let us define $X(t)$= size, $R(t) = \frac{1}{X(t)}\frac{dX(t)}{dt}$= relative growth rate (RGR) at time point t. We assume $(R(1), \ldots, R(q))' \sim N_q(\theta, \Sigma)$. where and $E(R(t)) = \theta(t) = f(\phi, t)$, a suitable rate profile, $t = 1(1)q$. Suppose that we are interested in testing the hypothesis of Gompertz growth curve model(GGCM). i.e., to test

$$H_0 : \theta(t) = ae^{-bt} \quad ag. \quad H_1 : \ not \ H_0.$$

Using the approximate expression for expectation and variance of the logarithm of ratio of RGR for two consecutive time points, describe two testing procedures and critical regions in testing the null hypothesis of GGCM. Also suggest required modifications of test statistics when the errors are non-normal.

$[15 + 5 = 20]$

4. (a) Consider a matrix function

$$f(\Sigma) = log|\Sigma| + tr[\Sigma^{-1}A]$$

where, Σ is the variance-covariance matrix of a multivariate normal distribution. If $A > 0$ then subject to $\Sigma > 0$, $f(\Sigma)$ is minimized uniquely at $\Sigma = A$.

(b) Define canonical correlation briefly. Let $X_1, \ldots, X_n$ denote a random sample from a $N_p(\mu, \Sigma)$ distribution. Develop a test procedure for retaining the first $k(k < p)$ pairs of canonical variables.

(c) Let $X = [X_1, X_2, X_3, X_4]'$ be a random variable with covariance matrix Σ given by

$$\Sigma = \begin{bmatrix} 1 & r & r & r \\ r & 1 & r & r \\ r & r & 1 & r \\ r & r & r & 1 \end{bmatrix}, \text{ where } 0 < r < 1.$$

Find eigenvalues of Σ hence find the first principal component, Y_1.

[4+(2+5)+6 = 17]

5. Suppose the observable random vector X with p components has mean vector μ and covariance matrix Σ. Describe the factor model with factors $F_1, F_2, \ldots, F_m$ as unobservable random variables and p additional sources of variation $\epsilon_1, \epsilon_2, \ldots, \epsilon_p$, called the errors. Briefly mention the steps of Maximum Likelihood Estimation procedure for estimating factor model parameters. Show that for factor model the loading matrix may not be unique.

[4 + 6 + 3 = 13]

6. (a) Suggest a generalized distance for discrete variables which is analogous to the Mahalanobis generalized distance. How this concept can be extended to more than two populations in the form of a heterogeneity distance.

(b) In a soil experiment we have recorded the data for soil pH, soil phosphorus concentration, absence or presence of soil microbes and soil colours (blackish or greyish) for the two qualities of soil.

(i) Suppose that we have available a set of n_1 samples known to have come from the first soil population and a set of n_2 samples come from the second soil population. Set up a discriminating rule between the two soils on the basis of a vector $\mathbf{x}$ of 2 binary variables and a vector $\mathbf{y}$ of 2 continuous variables observed for each sample.

(c) Distinguish single linkage and complete linkage methods in the context cluster analysis problem.

$[6 + 4 + 3 = 13]$

INDIAN STATISTICAL INSTITUTE
Semester Examination: 2011-2012, Second Semester
M-Stat II (MSP)and M-Math II
Ergodic Theory

Date: 04.05.12 **Max. Marks 70** **Duration: 3 Hours**

Note: Answer all questions.
All the measures considered are probability measures unless otherwise stated.

1. a) Prove Poincaré's recurrence theorem.

   b) Let $(X, \mathcal{B}, m, T)$ be a measure-preserving dynamical system. Show that T is ergodic if and only if $m(\cup_{n=1}^{\infty} T^{-n}A) = 1$ for all $A \in \mathcal{B}$ with $m(A) > 0$.

   c) Let T be measure-preserving ergodic and invertible on $(X, \mathcal{B}, m)$. Let $A \in \mathcal{B}$ and $m(A) > 0$. Prove that $\int_A R_A dm = 1$ where

   $$R_A(x) = \inf\{n \geq 1 : T^n(x) \in A\}.$$

   [8+7+8]

2. Let $K = \{z \in \mathcal{C} : |z| = 1\}$ with Borel σ-field and Lebesgue measure. Let $T : K \to K$ be defined as $Tz = az$ where a is not a root of unity. Show that
   a) T is ergodic
   b) $T \times T$ on $K \times K$ is not ergodic. [5+5]

3. a) Let $T : (X, \mathcal{B}, m) \to (X, \mathcal{B}, m)$ be measure-preserving and ergodic. Suppose that f and g are eigenfunctions of T corresponding to the eigenvalue λ. Show that $f = cg$ almost everywhere for some constant c.

   b) If T_1 and T_2 are invertible measure-preserving transformation on $(X, \mathcal{B}, m)$ such that $T_1T_2 = T_2T_1$, T_1 is ergodic, T_2 is weak mixing. Then show that T_1 is weak mixing.

   [6+6]

4. a) If T is an invertible ergodic measure-preserving transformation with discrete spectrum then show that T and T^{-1} are conjugate.

   b) Let $K = \{z \in \mathcal{C} : |z| = 1\}$ with Lebesgue measure. Let $Tz = z^2$. Does T have discrete spectrum? Justify your answer.

   [6 + 6]

5. a) Let $(K, \mathcal{B}, \lambda, T)$ be the dynamical system as given in 4.(b). Find the value of $h(T)$, the entropy of the system.

   b) Let $(X, \mathcal{B}, m, T)$ be a measure-preserving dynamical system. Let $\mathcal{A}$ be a countable measurable partition with finite entropy. Let $I^*(x) = \sup_{n \geq 1} I_{\mathcal{A}|\bigvee_{i=1}^{n} T^{-i}\mathcal{A}}(x)$ where $I_{\cdot|\cdot}$ denotes the conditional information function. Show that for all $A \in \mathcal{A}$, $m(x \in A : I^*(x) > \gamma)$ converges to zero at exponential rate as $\gamma \to \infty$.

   [5+10]

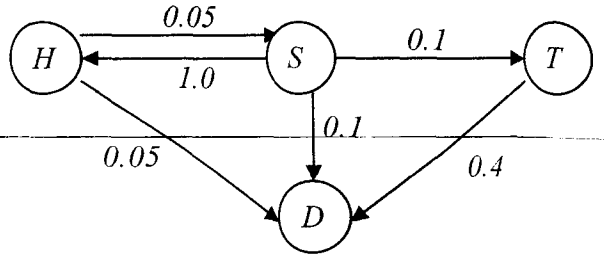
INDIAN STATISTICAL INSTITUTE

Second Semestral Examination: 2011-12
M. Stat. II Year

ACTUARIAL MODELS

Date: May 8, 2012 Maximum Marks: 100 Duration: 3 hr

1. Consider the following continuous-time Markov sickness-death model, having four states, H (healthy), S (sick), T (terminally ill) and D (dead), the time unit used being 1 year. The numbers proximal to the arrows (representing transitions) denote the respective transition intensities.



a. Deduce from first principles the probability of an individual remaining healthy for at least 10 uninterrupted years, given that he/she is healthy now, and hence calculate this probability.

b. Let d_j denote the probability that a life currently in state j, $j \in \{H, S\}$, will never suffer a terminal illness. By considering the first transition from state H, show that

$$d_H = \frac{1}{2} + \frac{1}{2} d_S.$$

Similarly, deduce that $d_S = \frac{1}{12} + \frac{5}{6} d_H$, and hence evaluate d_H and d_S.

c. Determine the expected duration of a terminal illness, starting from the moment of first transition into state T.

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

2. A simple No Claims Discount (NCD) system has four levels of discount: 0%, 20%, 40% and 60%. At the end of each policy year, policyholders changes levels according to the following rules:

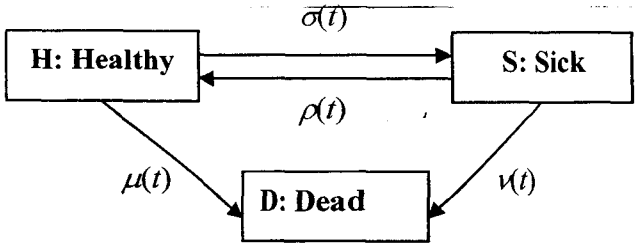
   i. At the end of a claim-free year, a policyholder moves up one level, or remains at the maximum discount level.
   ii. At the end of a year in which exactly one claim was made, a policyholder drops back one level, or remains at 0%.
   iii. At the end of a year in which more than one claim was made, a policyholder drops back to 0% discount level.

   For a particular policyholder in any given year, the probability of having a claim-free year is 0.7 and the probability of making exactly one claim is 0.2.

   a. Write down the transition matrix for this time-homogeneous Markov chain.
   b. If a policyholder starts at the 0% discount level, determine the probability that he/she is at maximum discount level 4 years later.
   c. If a large number of policyholders having the same claim probabilities take out policies at the same time, what proportion would you expect to be in each discount category after a long time?

   [2+4+4=10]

3. Consider the following three-state time-inhomogeneous Markov jump process model for sickness and death, where the quantities proximal to arrows represent the corresponding transition intensities.



   Let X_t denote the state of the process at time t (>0).

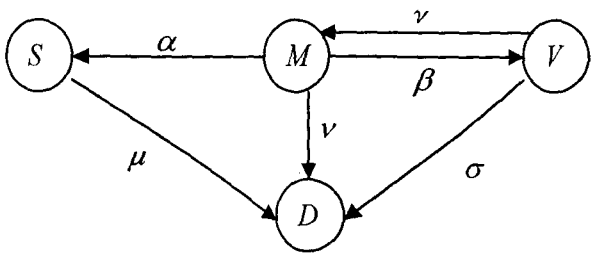
   a. If $p_{\overline{ii}}(s,t)$ denotes the probability of the process remaining in the same state i, $i \in \{H,S,D\}$, from time s to time t, then show that

$$p_{\overline{SS}}(s,t) = \exp\left(-\int_s^t (\rho(u)+\nu(u))du \right).$$

   b. Define the residual holding time R_s at time s. What is the probability density function of R_s given $X_s = H$?
   c. How can the model be modified to take into account the effect of the duration of illness?
   d. Under the modified model, what will be probability of a person being continually sick during a time interval $[s,t]$, given that he was sick for a period of length w prior to time point s?

   [3+2+2+3=10]

4. Consider the following model for marital status, having four states—married (M), single (S), divorced/separated (V) and dead (D), where the quantities proximal to arrows represent the corresponding transition intensities:



If ${}_t p_x^{SM}$ denotes the probability that an individual who is single at age x, is in the married state at age $x+t$, then

a. giving appropriate arguments, deduce the relation

$$\frac{\partial}{\partial t}\,{}_t p_x^{SM} = {}_t p_x^{SM}(\beta+\nu) + \alpha e^{-(\alpha+\mu)t};$$

b. using the boundary condition ${}_0 p_x^{SM} = 0$, solve this differential equation to obtain an explicit expression for ${}_t p_x^{SM}$.

[6+4=10]

5.

a. Deduce the explicit form of the expectation of K_x, that is, the random variable denoting the curtate future lifetime of a individual after age x, in terms of the survival probabilities ${}_k p_x, k = 1,2,\ldots$, at age x.

b. A company pays very high wages but also has a very high "failure rate", both from sackings and through employees leaving. A "life table" for a typical new recruit (with duration measured in years) is as follows:

Duration	No. of employees
0	1000
1	720
2	510
3	360
4	240
5	150
6	100
7	60
8	25
9	0

75 new recruits joined the company on September 1, 2011. Compute the expected number of complete years any of them will complete with the company and his expected "lifetime" with the company.

[5+5=10]

6. A group of 6 lives was observed over a period of time as a part of a mortality investigation. Each of the lives was under observation at all ages from age 55 until they died or were censored. The table below shows the sex, age at exit and reason for exit from the investigation.

Life no.	Sex	Age at exit	Reason for exit
1	M	56	Death
2	F	62	Censored
3	F	63	Death
4	M	66	Death
5	M	67	Censored
6	M	67	Censored

a. Set up a Cox regression model, with sex as a covariate.
b. Write down the explicit partial likelihood equation under this model for the data.
c. Obtain the maximum partial likelihood estimate of the model parameter.

[1+5+4=10]

7. A life insurance company carried out a mortality investigation by following a sample of independent policyholders aged between 50 and 55 years. Policyholders were followed from their 50^{th} birthday till they died, withdrew from the investigation while still alive or reached their 55^{th} birthday (whichever of these events occurred first). The data for 8 policyholders is given in the table below.

Life	Last age at which life was observed (years and months)		Reason for exit
1	50	9	Died
2	51	3	Withdrew
3	51	6	Died
4	51	6	Died
5	51	6	Withdrew
8	54	3	Died
9	54	6	Died
10	55	0	Reached age 55

Calculate the Nelson-Aalen estimate of the survival function from the data, and also an approximate 95% confidence interval for your estimate.

[6+4=10]

8. A pension scheme only allows retirement at exact age 65. An investigation of the mortality of the retired members of the scheme was carried out over the period January 1, 1998 to December 31, 2003. The following data were obtained:

Member	Date of retirement	Date of death (if occurring during the investigation period)
1	1 April 1995	30 April 2002
2	1 August 1997	-
3	1 February 1998	-
4	1 June 1999	31 August 2001
5	1 August 1999	31 December 2003
6	1 March 2001	-
7	1 May 2001	30 November 2003

All months should be assumed to be of equal length.

Calculate the Kaplan-Meier estimate of the survival function $S_{65}(t)$ from these data, stating clearly any additional assumptions that you make.

[10]

9.

a. A mortality investigation was conducted between January 1, 2005 and December 31, 2008. Data on four lives covered under the investigation is given below:

Life no.	Date of birth	Date of entry	Date of exit	Reason for exit
1	11.1.1969	24.3.1998	29.02.2006	Death
2	10.2.1975	10.9.2007	-	-
3	13.4.1988	1.5.2008	-	-
4	23.11.55	10.8.2005	31.5.2007	Surrender

Calculate the number of days of exposure contributed to the central exposed to risk by each life at each age, assuming that only the date of entry and not the date of exit counts in the exposed to risk, and defining age to be the age last birthday.

b. In a mortality investigation, N lives between the ages of x and $x+1$ are observed. Let E_x^c denote the total waiting time (that is, the time spent under observation) for these lives, and assume the force of mortality μ_x to be constant between the ages of x and $x+1$.

Assuming a Poisson model for mortality, derive the maximum likelihood estimate of μ_x, and construct an approximate 95% confidence interval for μ_x when $x=60$, given that 52 deaths were observed between the ages of 60 and 61, and the corresponding total waiting time was 8460 years.

[6+4=10]

10. A study of the leading causes of death in elderly men in the 1970s showed the proportions given in the table below. The table also gives the number of deaths due to the different causes in the year 2005. Using

i. the chi-squared test,
ii. the cumulative deviations test, and
iii. the serial correlations test,

determine whether the pattern has changed from the 1970s to the year 2005. State clearly the underlying assumptions, and comment on the conclusion arrived at, in each case.

Cause of Death	Proportion of Deaths in 1975	Number of Deaths in 2005
Cancer	0.15	809
Heart disease	0.20	1567
Infectious diseases	0.35	345
Respiratory diseases	0.15	123
Other causes	0.15	546

[10]

INDIAN STATISTICAL INSTITUTE

M. Stat - II year (MSP)

Statistical Inference II

Answer Part I and Part II in separate answer books

Date : 08.05.2012 Time : $4\frac{1}{2}$ Hours

Answer Part I and Part II in separate answer books

Part I (Time $2\frac{1}{2}$ hours)

Maximum Marks : 35 Answer both Group A and Group B

Group A

Answer all questions

(a) Let $(\Omega, \mathcal{A}, M)$ be a statistical structure. Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be two sub-σ-fields of $\mathcal{A}$ such that

(i) $\mathcal{A}_2$ is ancillary;

(ii) $\sigma(\mathcal{A}_1 \cup \mathcal{A}_2)$ is sufficient for M; and

(iii) for each $P \in M$, $\mathcal{A}_1$ and $\mathcal{A}_2$ are P- independent.

Show that $\mathcal{A}_1$ is sufficient for M.

What is the relevance of this result to Basu's theorem?

[7+1 = 8]

(b) Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ variables. Let $\bar{X} = (X_1 + \ldots + X_n)/n$, M = a median of $X_1, \ldots, X_n$.

Derive $\text{cov}(\bar{X}, M)$.

[4]

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

[5]

Group B

Answer any two questions.

(a) Let $(\Omega, \mathcal{A}, M)$ be a statistical structure such that $\exists$ a measure μ with the property that $\frac{dP}{d\mu}$ exists $\forall P \in M$.

Show that $(\Omega, \mathcal{A}, M)$ is boundedly complete if $\left\{\frac{dP}{d\mu} : P \in M\right\}$ is complete in $L_1(\Omega, \mathcal{A}, \mu)$, the converse being true if $\mathcal{N}_M = \mathcal{N}_\mu$ and μ is σ-finite.

Give an application of this result.

[3+3+1= 7]

(b) Let $X_1, \ldots, X_n$ be iid $U(0, \theta)$, $0 < \theta < \infty$. Show that $X_{(n)} = \max(X_1, \ldots, X_n)$ is complete.

[2]

4.(a) Define a discrete statistical structure.
Show that in a discrete statistical structure, there is a minimum sufficient σ-field.
(You may assume that if $\mathcal{B}$ is sufficient, then $\mathcal{A}(\prod(\mathcal{B})) = \mathcal{B}$ (using the standard notations), a
that for an inducible σ-field $\mathcal{B}$, $\mathcal{B}$ is sufficient iff "the factorization of the p.d.f." holds).

[1+6=

(b) Let $\Omega = \mathbb{R}^n \setminus \{o\}$, $g_A(x) = Ax$ and $G = \{ g_A : A \text{ is an } n \times n \text{ orthogonal matrix}\}$.

Find a maximal invariant function under the group G.

5.(a) Define a coherent structure.
Let $(\Omega, \mathcal{S}, M)$ be a coherent structure. Show that if M is not connected, then there exists
splitting set.

[1+3=

(b) Define the invariant and almost-invariant σ-fields. What is the relationship between the
Justify.
Give an example where these two σ-fields are unequal; justify.

[(1+1)+3 =

6. Let $X_1, \ldots, X_n$ be independent, and assume that X_i is Bin $(1, p_i)$ for $i = 1, \ldots, n$.
Let $H_0 : p_i \le \frac{1}{2} \forall i$ be tested against $H_1 : p_i > \frac{1}{2} \forall i$. Find the UMP invariant test. Is it m
stringent?; justify. Is it a maximum test?; justify.

[5+2+2 =

*****

Part II: Bayesian Inference
Maximum Marks: 40 Time: 2hours

This part carries 40 points with five problems carrying 8 points each. Answer all questions. Best of luck!

1) Let P_1 and P_2 be two independent probabilities on $\Re$ with $P_i \sim DP(\alpha_i)$, $i = 1, 2$. Let X be a random variable independent of (P_1, P_2) following a $Be(\alpha_1(\Re), \alpha_2(\Re))$ distribution. Derive the distribution of $XP_1 + (1 - X)P_2$.

2) Consider a regression setting with continuous predictors $X_i \in \Re^p$ and discrete responses $Y_i \in \{1, 2, 3\}$, $i = 1, \ldots, n$ such that $\Pr(Y_i = 1|X_i) = \Phi(\beta' X_i)$ and $\Pr(Y_i = 2|X_i) = \Phi(c + \beta' X_i) - \Phi(\beta' X_i)$, where Φ is the standard Normal cdf. The parameters are $\beta = (\beta_1, \ldots, \beta_p)' \in \Re^p$ and $c > 0$. Using a flat prior $\pi(\beta) = 1$ on β and an independent truncated Normal prior on c: $c \sim N(0, 1)I(c > 0)$, develop a MCMC algorithm to sample from the posterior of the parameters.

3) Let $P \sim DP(\alpha)$ and α has the Cauchy density $\frac{1}{\pi}\frac{1}{1+x^2}$. Show that if $X \sim P$, then $E(X|P)$ has the same Cauchy distribution.

4) Let $X_1, \ldots, X_n$ be a iid sample from

$$f(x) = \int \frac{1}{\sigma}\phi(\frac{x - \mu}{\sigma})P(d\mu d\sigma)$$

with parameter P having a $DP(\alpha)$ prior. Under α, $\mu \sim N(0, 1)$ and $\sigma^{-2} \sim Exp(1)$ independently. Devise a Gibbs sampling algorithm to sample from the posterior of P.

5) Consider a six-sided die, with θ the random variable corresponding to the upper face of the die.

a. If π is the distribution of θ, give the maximum entropy prior associated with the information $\mathrm{E}(\theta) = 3.5$.

b. Show that if A is the event "θ is odd", the updated distribution $\pi(.|A)$ is $(1/3, 0, 1/3, 0, 1/3, 0)$.

Part II: Bayesian Inference

Maximum Marks: 50 Time: 1hr 30min

Answer all questions.

1) Let α be a finite measure on $\mathcal{X} = \{1, 2, \ldots, K\}$. Let X be a random variable on $\mathcal{X}$ with distribution $\alpha/\alpha(\mathcal{X})$. Let P be a random probability on $\mathcal{X}$, independent of X with distribution $\text{Dir}(\alpha)$. Let Y be a random variable independent of (X, P) and following a $\text{Be}(1, \alpha(\mathcal{X}))$ distribution. Then derive the distribution of $Y\delta_X + (1-Y)P$. Here δ_X is the random probability that puts all its mass on X. [13]

2) Let $P \sim DP(\alpha)$. Assume $\int x^2 d\alpha < \infty$. Calculate the expected value and variance of $\int x dP$. [13]

3) Let $X_1, \ldots, X_n$ be a iid sample from

$$f(x) = \int \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})P(d\mu)$$

with parameters P and σ. The parameter P follows a $DP(\alpha)$ prior where α is the standard Normal distribution. The parameter σ is independent of P and follows the prior $\pi(\sigma) = \sigma^{-1}$. Devise a Gibbs sampling algorithm to sample from the posterior of (P, σ). [12]

4) For a multinomial with probabilities $p_1, \ldots, p_k$ for k classes, calculate the Jeffreys prior. You may use the result, $\det(A + xx') = \det(A)(1 + x'A^{-1}x)$ for any positive definite A. [12]

INDIAN STATISTICAL INSTITUTE
M. Stat - II year (MSP)

Statistical Inference II

Second Semester back paper Examination (2011 – 2012)

Date : 26.06.2012 Time : 3 Hours

Answer Part I and Part II in separate books
Part I

Maximum Marks : 50

Answer any two questions

1. Show that if $\mathcal{A}_1$ is sufficient and $\mathcal{A}_2$ is countable generated, then $\sigma(\mathcal{A}_1 \cup \mathcal{A}_2)$ is sufficient. State and prove all the results you are using.

[25]

2. State and prove the weak compactness theorem of Banach. Prove all your assertions.

[25]

3. State and prove Hunt – Stein's theorem.

[25]

P.T.O

Applied Multivariate Analysis

End-Semester Examination

(Back Paper)

M.Stat II year (2011 - 2012) DATE- 28.06.12

Full Marks - 100

Time - 3 hours 30 mins.

1. (a) A population is divided into two groups and the prior probability that an individual belongs to group i is π_i, $i = 1, 2$. A measurement, $x \in \mathbb{X}$, is taken on each individual. The probability distribution of x for individuals from group i is p-variate normal with mean μ_i and covariance matrix Σ_i, $i = 1, 2$.

   Assuming the cost of assigning a member of group 1 to group 2 is the same as the cost of assigning a member of group 2 to group 1, find an appropriate decision rule for assigning an individual with measurement x to one of the two groups. Discuss the case when Σ_i' s are equal? Also discuss the distribution of discriminating function in the second case.

   (b) An experiment was carried out to investigate the amount of time that beetles spend on different activities. Define X to be a bivariate vector measurement, X_1 being the amount of time spent for eating and X_2 the amount of time spent at rest.

   The experiment considered two separate equally-common species *fatus* and *apathus* and yielded sample means $\bar{x}_f = \begin{pmatrix} 6 \\ 3 \end{pmatrix}$ and $\bar{x}_a = \begin{pmatrix} 2 \\ 9 \end{pmatrix}$ for the *fatus* and *apathus* respectively, with pooled covariance matrix $S = \begin{bmatrix} 3 & -2 \\ 2 & 4 \end{bmatrix}$.

   Use this data to derive a linear function that will discriminate between the two species of beetle. State any assumptions you make.

   [6 + 4 + 4 + 6 = 20]

2. Let us consider two populations, π_1 and π_2, having p-variate normal distributions with mean μ_1 and μ_2, respectively, a common variance-covariance matrix Σ, and an experimental observation X that has to be classified in one of the two populations. We also assume that $X \sim$

$N(\alpha\mu_1 + \beta\mu_2, \Sigma)$ with $\alpha+\beta = 1$, where μ_1, μ_2, Σ are known. Construct a single test to determine whether the experimental unit is coming from either of the two populations or from a new normal population with expectation equal to a linear combination of the expectations of the $\pi_i, i = 1, 2$.

[12]

3. Let us assume that $(X(1), \ldots, X(q))' \sim N_q(\theta, \Sigma)$, where $X(t)$ = size at time point t, and $E(X(t)) = \theta(t) = f(\phi, t)$, be a suitable growth curve. $t = 1(1)q$. Suppose we are interested in testing the hypothesis of exponential growth curve model(EPGM), i.e., to test

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

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

[4 + 10 + 5 + 4 = 21]

4. (a) Let X be an $(n \times p)$ data matrix in which each row corresponds to a p-variate measurement on one of n individuals. Assuming the p variates are continuous variables, describe two possible measures of dissimilarity of pairs of individuals. Comment on their relative advantages and disadvantages.

(b) What properties must be satisfied for a dissimilarity function to be a metric dissimilarity function?

(c) Let $x = (x_1, \ldots, x_p)$ and $y = (y_1, \ldots, y_p)$, where $x_i, y_i \in \{0, 1\}$ for $i = 1, \ldots, p$. The simple matching coefficient is

$$S_1 = \frac{a+d}{p},$$

where $a = \sum_{i=1}^{p} \mathbf{1}\{x_i = 1, y_i = 1\}$ and $d = \sum_{i=1}^{p} \mathbf{1}\{x_i = 0, y_i = 0\}$. State for each of the following whether it is a metric dissimilarity function, and explain your reasoning:

(i) $d_1 = (1 - S_1)$ (ii) $d_2 = \frac{1}{(1+S_1)}$.

"Any one of the above metrics can be used for standard linkage method of cluster analysis" - Expalin, whether you agree or disagree with the statement.

[6 + 4 + 6 + 3 = 19]

5. (a) What is the purpose of a Principal Components Analysis?

   Let X be a p-variate random variable with covariance matrix Σ. Derive the principal components Y of X and write down the covariance matrix Γ of Y.

   Prove that $\sum_{i=1}^{p} V(Y_i) = \sum_{i=1}^{p} \lambda_i$ where $\lambda_1 > \lambda_2 > \ldots > \lambda_p$ are the eigenvalues of Σ.

   or

   Discuss canonical correlation in the context of judging the quality of association between two variables. Also discuss its computation procedure. Show that canonical correlation remains invariant under any non-singular transformation.

   [3 + 6 + 4 = 14]

6. Suppose that the observable random vector X with p components has mean vector μ and covariance matrix Σ. Describe the factor model with factors $F_1, F_2, \ldots, F_m$ as unobservable random variables and p additional sources of variation $\epsilon_1, \epsilon_2, \ldots, \epsilon_p$, called the errors. Suggest a goodness-of-fit test for such model. Distinguish Principal Component Analysis and Factor Analysis in the context of data reduction technique.

   [4 + 6 + 3 = 13]