INDIAN STATISTICAL INSTITUTE
Back Paper Examination: 2013-2014
M. Stat. II Year: Semester-II

Survival Analysis

Date: 28·7·14 Full Marks: 100 Duration: 3 hours

Note: Answer all questions

1. Suppose the mean residual life of a continuous survival time T is given by $\text{MRL}(t) = t + 10$. Find the hazard function $\lambda(t)$.

[9]

2. Consider random right censoring with failure time $T \sim F$ and censoring time $C \sim G$ with $\bar{G} = [\bar{F}]^{\theta}$, $\theta > 0$. Assume both F and G are continuous. Prove that $X = \min(T, C)$ and $\delta = I\{T \leq C\}$ are independent.

[10]

3. Suppose for a life time T, we have for a given covariate Z, $\log T = \beta^T Z + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. Then verify whether the distribution of T follows Cox's proportional hazards assumption or not.

[5]

4. Consider a Type-II censoring scheme, with $r = 10$ and $n = 25$. Observed failure times are

$$0.004, 0.029, 0.135, 0.147, 0.177, 0.194, 0.264, 0.349, 0.353, 0.387.$$

Perform the modified Kolmogorov-Smirnov test to check whether the data are from *Exponential*(1) or not. For truncation proportion 0.4, upper alpha value $U_{0.4} = 1.975$.

[5+2+2+1=10]

5. The data below show remission times, in weeks, for 24 leukemia patients randomly assigned to two treatments A and B. Asterisks denote censoring times.

Treatment A: 1, 2, 3, 6*, 7, 10, 12*, 14, 15*, 18, 20*, 22.
Treatment B: 1, 1, 2, 2*, 3, 4, 5, 8, 8, 9*, 11, 12.

(a) Compute Kaplan-Meier estimate of survival probabilities in both the groups.

(b) Suppose that failure times from two groups of individuals are exponential with failure rates λ and λe^{β}, respectively. Derive a test for the homogeneity of the two groups giving full details. What is your conclusion about the homogeneity of the two groups A and B?

[8 + 12 = 20]

6. Suppose the survival prospect of two groups of patients are to be compared using a covariate Z. Group 0, indicated by $Z= 0$, has survival function $S_0(t)$ which is unspecified. Group 1, indicated by $Z= 1$, has survival function $S_1(t) = [S_0(t)]^{\psi}$.

(a) Write down this problem as Cox proportional hazard model.

(b) Suppose you have censored and tied data. Derive a score test for testing the equality of two survival functions.

[2+8=10]

7. Consider the proportional hazard model $\lambda(t; Z) = \lambda_0(t)\exp(\beta^T Z)$, where $\lambda_0(t)$ is unspecified and arbitrary. Derive Breslow's estimator for $\Lambda_0(t)$, the cumulative baseline hazard function. Show that Breslow's method also gives partial likelihood for estimating β. [10 + 2 = 12]

8. Consider a continuous covariate z_i, equal to the body mass index of individual i. We have 9 individuals who have suffered a heart attack. Let t_i be the time of death in days following the attack and δ_i the censoring indicator. The data are given below.

i	t_i	δ_i	z_i
1	6	1	31.4
2	98	0	21.5
3	189	1	27.1
4	374	1	22.7
5	1002	0	35.7
6	1205	1	30.7
7	2065	1	26.5
8	2201	0	28.3
9	2421	1	27.9

Consider Cox proportional hazard model to study the effect of covariate. Construct partial likelihood.

[7]

9. Consider the accelerated failure time model

$$Y = \ln T = \mu + \beta^T Z + \sigma e,$$

where e is error random variable. Show that the probability of observed rank statistic under the null hypothesis of no covariate effect based on censored data is $\prod_{i=1}^{k} n_i^{-1}$, where k is the number of uncensored observations and n_i denotes the number at risk just prior to ith ordered failure time.

[10]

10. Describe a simple illness-death model where direct death is possible. Write down the observations and likelihood function for estimating the hazard rates corresponding to the different states.

[3+4=7]

INDIAN STATISTICAL INSTITUTE

Backpaper Examination, Second Semester: 2013-14
M.Stat. II Year (AS)

Actuarial Models

Date: 18·7·14, 2014 Maximum marks: 100 Duration: 3 hours

Answer all questions. Standard actuarial notations are followed.

1. Define, in the context of stochastic processes, (i) a mixed process and (ii) a counting process. Give an example application of each type of process. [4]

2. The price of a stock can either take a value above a certain point (state A), or take a value below that point (state B). Assume that the evolution of the stock price in time can be modelled by a two-state Markov jump process with homogeneous transition rates $\sigma_{AB} = \sigma$, $\sigma_{BA} = \rho$. The process starts in state A at $t = 0$ and time is measured in weeks.

   (a) Write down the generator matrix of the Markov jump process. [1]

   (b) State the distribution of the holding time in each of states A and B. [1]

   (c) If $\sigma = 3$, find the value of t such that the probability that no transition to state B has occurred until time t is 0.2. [2]

   (d) Assuming all the information about the price of the stock is available for a time interval $[0, T]$, explain how the model parameters σ and ρ can be estimated from the available data. [3]

   (e) State what you would test to determine whether the data support the assumption of a two-state Markov jump process model for the stock price. [1]

   [Total 8]

3. (a) Explain the difference between a time-homogeneous and a time inhomogeneous Poisson process. [1]

   An insurance company assumes that the arrival of motor insurance claims follows an inhomogeneous Poisson process. Data on claim arrival times are available for several consecutive years.

   (b) Describe a test that can be used to test the validity of the assumption. [3]

   The company concludes that an inhomogeneous Poisson process with rate $\lambda(t) = 3\cos(2\pi t)$ is a suitable fit to the claim data (where t is measured in years).

   (c) Comment on the suitability of this transition rate for motor insurance claims. [2]

   (d) Write down the Kolmogorov forward equations for $P_{0j}(s,t)$. [2]

   (e) Verify that these equations are satisfied by:

   $$P_{0j}(s,t) = \frac{(f(s,t))^j \exp(-f(s,t))}{j!}$$

   for some $f(s,t)$ which you should identify. [2]

   (f) Comment on the form of the solution compared with the case where λ is constant. [2]

   [Total 12]

4. A motor insurance company wishes to estimate the proportion of policyholders who make at least one claim within a year. From historical data, the company believes that the probability a policyholder makes a claim in any given year depends on the number of claims the policyholder made in the previous two years. In particular:

- the probability that a policyholder who had claims in both previous years will make a claim in the current year is 0.25;
- the probability that a policyholder who had claims in one of the previous two years will make a claim in the current year is 0.15; and
- the probability that a policyholder who had no claims in the previous two years will make a claim in the current year is 0.1.

(a) Construct this as a Markov chain model, identifying clearly the states of the chain. [2]

(b) Write down the transition matrix of the chain. [1]

(c) Explain why this Markov chain will converge to a stationary distribution. [2]

(d) Calculate the proportion of policyholders who, in the long run, make at least one claim at a given year. [4]

[Total 9]

5. A manufacturer uses a test rig to estimate the failure rate in a batch of electronic components. The rig holds 100 components and is designed to detect when a component fails, at which point it immediately replaces the component with another from the same batch. The following are recorded for each of the n components used in the test $(i = 1, 2, \dots, n)$:

$$\begin{aligned} s_i &= \text{time at which component } i \text{ is placed on the rig,} \\ t_i &= \text{time at which component } i \text{ is removed from rig,} \\ f_i &= \begin{cases} 1 & \text{if component is removed due to failure,} \\ 0 & \text{if component is working at end of test period.} \end{cases} \end{aligned}$$

The test rig was fully loaded and was run for two years continuously. You should assume that the force of failure, μ, of a component is constant and component failures are independent.

(a) Show that the contribution to the likelihood from component i is:

$$\exp\left(-\mu\left(t_i - s_i\right)\right) \cdot \mu^{f_i}. \qquad [2]$$

(b) Derive the maximum likelihood estimator for μ. [4]

[Total 6]

6. (a) Assume that the force of mortality between consecutive integer ages, y and $y+1$, is constant and takes the value μ_y. Let T_x be the future lifetime after age x $(x \leq y)$ and $S_x(t)$ be the survival function of T_x. Show that

$$\mu_y = \log[S_x(y - x)] - \log[S_x(y + 1 - x)]. \qquad [4]$$

(b) An investigation was carried out into the mortality of male life office policyholders. Each life was observed from his 50th birthday until the first of three possible events occurred: his 55th birthday, his death, or the lapsing of his policy. For those policyholders who died or allowed their policies to lapse, the exact age at exit was recorded. Using the result from part (a) or otherwise, describe how the data arising from this investigation could be used to estimate μ_{50} and ${}_5q_{50}$. [4]

[Total 8]

7. A national mortality investigation is carried out over the calendar years 2002, 2003 and 2004. Data are collected from a number of insurance companies. Deaths during the period of the investigation, θ_x, are classified by age nearest at death. Each insurance company provides details of the number of in-force policies on 1 January 2002, 2003, 2004 and 2005, where policyholders are classified by age nearest birthday, $P_x(t)$.

   (a) State the rate year implied by the classification of deaths. [1]

   (b) State the ages of the lives at the start of the rate interval. [2]

   (c) Derive an expression for the exposed to risk, in terms of $P_x(t)$, which may be used to estimate the force of mortality in year t at each age. State any assumptions you make. [3]

   (d) Describe how your answer to part (c) would change if the census information provided by some companies was $P_x^*(t)$, the number of in-force policies on 1 January each year, where policyholders are classified by age last birthday. [3]

[Total 9]

8. An investigation took place into the mortality of persons between exact ages 60 and 61 years. The table below gives an extract from the results. For each person it gives the age at which they were first observed, the age at which they ceased to be observed and the reason for their departure from observation.

Person	*Age at entry*		*Age at exit*		*Reason for exit*
	years	*months*	*years*	*months*	
1	60	0	60	6	withdrew
2	60	1	61	0	survived to 61
3	60	1	60	3	died
4	60	2	61	0	survived to 61
5	60	3	60	9	died
6	60	4	61	0	survived to 61
7	60	5	60	11	died
8	60	7	61	0	survived to 61
9	60	8	60	10	died
10	60	9	61	0	survived to 61

   (a) Estimate q_{60} using the Binomial model. [5]

   (b) List the strengths and weaknesses of the Binomial model for the estimation of empirical mortality rates, compared with the Poisson and two-state models. [3]

[Total 8]

9. An investigation was undertaken of the mortality of persons aged between 40 and 75 years who are known to be suffering from a degenerative disease. It is suggested that the crude estimates be graduated using the formula:

$$\overset{\circ}{\mu}_{x+1/2} = \exp\left[b_0 + b_1\left(x + \frac{1}{2}\right) + b_2\left(x + \frac{1}{2}\right)^2\right].$$

(a) Explain why this might be a sensible formula to choose for this class of lives. [1]

(b) Suggest two techniques which can be used to perform the graduation. [2]

(c) The table below shows the crude and graduated mortality rates for part of the relevant age range, together with the exposed to risk at each age and the standardized deviation at each age.

Age last birthday	*Graduated force of mortality*	*Crude force of mortality*	*Exposed to risk*	*Standardized deviation*
x	$\overset{\circ}{\mu}_{x+1/2}$	$\hat{\mu}_{x+1/2}$	E_x^c	$z_x = \dfrac{E_x^c(\overset{\circ}{\mu}_{x+1/2} - \hat{\mu}_{x+1/2})}{(E_x^c\overset{\circ}{\mu}_{x+1/2})^{1/2}}$
50	0.08127	0.07941	340	–0.12031
51	0.08770	0.08438	320	–0.20055
52	0.09439	0.09000	300	–0.24749
53	0.10133	0.10345	290	0.11341
54	0.10853	0.09200	250	–0.79336
55	0.11600	0.10000	200	–0.66436
56	0.12373	0.11176	170	–0.44369
57	0.13175	0.12222	180	–0.35225

Test this graduation for (i) overall goodness-of-fit, (ii) bias, and (iii) the existence of individual ages at which the graduated rates depart to a substantial degree from the observed rates. [9]

[Total 12]

10. A life insurance company has carried out a mortality investigation. It followed a sample of independent policyholders aged between 50 and 55 years. Policyholders were followed from their 50th birthday until they died, they withdrew from the investigation while still alive, or they celebrated their 55th birthday (whichever of these events occurred first).

(a) Describe the censoring that is present in this investigation. [2]

An extract from the data for 12 policyholders is shown in the table below.

Policyholder	*Last age at which policyholder was observed (years and months)*	*Outcome*
1	50 years 3 months	Died
2	50 years 6 months	Withdrew
3	51 years 0 months	Died
4	51 years 0 months	Withdrew
5	52 years 3 months	Withdrew
6	52 years 9 months	Died
7	53 years 0 months	Withdrew
8	53 years 6 months	Withdrew
9	54 years 3 months	Withdrew
10	54 years 3 months	Died
11	55 years 0 months	Still alive
12	55 years 0 months	Still alive

(b) Calculate the Nelson-Aalen estimate of the survival function. [6]

(c) Sketch on a suitably labelled graph the Nelson-Aalen estimate of the survival function. [3]

[Total 11]

11. An investigation was undertaken into the effect of a new treatment on the survival times of cancer patients. Two groups of patients were identified. One group was given the new treatment and the other an existing treatment. The following model was considered:

$$h_i(t) = h_0(t)\exp\left(\beta^T z\right),$$

where $h_i(t)$ is the hazard at time t, where t is the time since the start of treatment

$h + 0(t)$ is the baseline hazard at time t

z is a vector of covariates such that:

$z_1 =$ sex (a categorical variable with $0 =$ female, $1 =$ male)

$z_2 =$ treatment (a categorical variable with $0 =$ existing treatment, $1 =$ new treatment), and

β is a vector of parameters, (β_1, β_2).

The results of the investigation showed that, if the model is correct, then (i) the risk of death for a male patient is 1.02 times that of a female patient, and (ii) the risk of death for a patient given the existing treatment is 1.05 times that for a patient given the new treatment.

(a) Estimate the value of the parameters β_1 and β_2. [4]

(b) Estimate the ratio by which the risk of death for a male patient who has been given the new treatment is greater or less than that for a female patient given the existing treatment. [2]

(c) Determine, in terms of the baseline hazard only, the probability that a male patient will die within 3 years of receiving the new treatment. [2]

[Total 8]

INDIAN STATISTICAL INSTITUTE

Backpaper Examination

M. Stat. II Year Semester II : 2013-2014

Stochastic Processes I

Date: 28·7·14 Total Marks : 100 Time : 3 Hours

1. Let (S, ρ) be a separable metric space and let S^∞ be the usual infinite cartesian product of S (with itself), that is, $S^\infty = \{x = (x_1, x_2, \ldots) : x_i \in S \; \forall \; i \geq 1\}$.
(a) Show that $\bar{\rho}(x, y) = \sum\limits_i 2^{-i}(\rho(x_i, y_i) \wedge 1)$, for $x, y \in S^\infty$, defines a metric on S^∞ and that S^∞ is separable under this metric.
(b) Denote $\mathcal{S}$ to be the borel σ-field on S and $\mathcal{S}^\infty$ to be the product σ-field on S^∞. Show that $\mathcal{S}^\infty$ is the borel σ-field on S^∞.
(c) Show that the finite-dimensional measurable cylinders forms a convergence determining class on $(S^\infty, \mathcal{S}^\infty)$. [For this, you may use result(s) proved in class but state clearly what you use.] (8+8+9)=[25]

2. Let $\{B_t, t \in [0, \infty)\}$ be a SBM starting at 0 with natural filtration $\mathcal{F}_t$, $t \geq 0$.
(a) Show that the following processes are martingales with respect to $\{\mathcal{F}_t\}$:
(i) $\{X_t = B_t^3 - 3tB_t, \; t \geq 0\}$ and (ii) $\{Y_t = \exp\{\alpha B_t - \frac{1}{2}\alpha^2 t\}, \; t \geq 0\}$ for any real α.
(b) Using the martingale (ii) above or otherwise, show that $\dfrac{B_t}{t} \to 0$ as $t \to \infty$ with probability one.
(c) Fix $0 \leq a < b < \infty$. For a partition $\pi = \{a = t_0 < t_1 < \cdots < t_k = b\}$ of $[a, b]$, let $Q(\pi)$ denote the sum $\sum\limits_{i=1}^{k} [B(t_i) - B(t_{i-1})]^2$. Show that if $\{\pi_n, n \geq 1\}$ is any sequence of finite partitions of $[a, b]$ with $\|\pi_n\| \to 0$, then the sequence $\{Q(\pi_n), n \geq 1\}$ converges in L_2 to $(b - a)$. Show also that, in case $\sum_n \|\pi_n\| < \infty$, one has almost sure convergence as well. ((4+4)+9+(5+3))=[25]

3. Let $\{B_t, t \in [0, \infty)\}$ be a SBM starting at 0 with natural filtration $\mathcal{F}_t$, $t \geq 0$.
(a) Show that for every $t \geq 0$, the map φ on $[0, t] \times \Omega$ defined as $\varphi(s, \omega) = B(s, \omega)$ is measurablr with respect to $\mathcal{B}[0, t] \otimes \mathcal{F}_t$.
(b) Using (a) or otherwise, show that for any $\{\mathcal{F}_t\}$-stopping time τ, the random variable $B_{\tau \wedge t}$ is $\mathcal{F}_t$-measurable and hence deduce that for any finite stopping time τ, the random variable B_τ is $\mathcal{F}_\tau$-measurable.
(c) For $t > 0$, denote $M_t = \max\{B_s : 0 \leq s \leq t\}$. Use strong markov property to show that for any $a > 0$, $P[M_t > a] = 2P[B_t > a]$. ·(9+8+8)=[25]

4. Consider a Feller markov process $\{X_t\}$ on $(\Omega_c, \mathcal{F}, \mathcal{F}_t, \{P_x, x \in S\})$.
(a) Show that for $f \in C_b(S)$, the function $\int_0^t T_s f(\cdot) ds$ belongs to $C_b(S)$ for any $t > 0$ and $R_\lambda(\int_0^t T_s f(x) ds) = \int_0^t R_\lambda(T_s f)(x) ds = \int_0^t T_s(R_\lambda f)(x) ds$ for any $\lambda > 0$.
(b) Let $f \in C_b(S)$ be in the domain of A. Show that both $T_t f$ and $\int_0^t T_s f ds$ are in the domain of A and that $A(T_t f) = T_t(Af)$ and $A(\int_0^t T_s f(x) ds) = \int_0^t A(T_s f)(x) ds$.
(c) Show that for $f \in C_b(S)$ in the domain of A, $E_x[\int_0^t Af(X_s) ds] = T_t f(x) - f(x)$ for any $t > 0$, $x \in S$. Deduce that for such an f, $M_t = f(X_t) - \int_0^t Af(X_s) ds$, $t \geq 0$, is a martingale with respect to $\{\mathcal{F}_t\}$, under any P_x. (7+8+(5+5))=[25]

INDIAN STATISTICAL INSTITUTE

Compensatory Back Paper Examination (2013–2014)

M. STAT. II (MSP)

Functional Analysis

Date : 29.08.2014 Maximum Marks : 75 Time : 2 hrs 30 minutes.

Precisely justify all your steps. Carefully state all the results you are using.

1. Let $\mathcal{H}$ be a Hilbert space and $\{T_n\}_{n\in\mathbb{N}}$ be a sequence of bounded operators from $\mathcal{H}$ to $\mathcal{H}$ such that $\langle T_n(\xi),\eta\rangle \underset{n\to\infty}{\longrightarrow} 0$ $\forall \xi,\eta\in\mathcal{H}$. Show that $\sup\limits_{n\in\mathbb{N}}\|T_n\| < \infty$. [10]

2. Define $T : L^2([0,1]) \to L^2([0,1])$ by $(Tf)(x) = xf(x)$ for all $f \in L^2([0,1])$ and $x \in [0,1]$. Show that T is a bounded linear operator. Is T invertible? Justify your answer. [15]

3. Consider the following subspaces of $c_0(\mathbb{N})$:

$$M_1 = \{\{x_n\}_{n\in\mathbb{N}} : x_1 = 0\} \quad\text{and}\quad M_2 = \{\{x_n\}_{n\in\mathbb{N}} : x_1 = x_2 = 0\}$$

Prove that M_1 is isometrically isomorphic to M_2 but the corresponding quotient spaces c_0/M_1 and c_0/M_2 are not isomorphic. [10]

4. Let $\mathcal{H}$ be a Hilbert space with orthonormal basis $\{e_n\}_{n=1}^{\infty}$. Define $T : \mathcal{H} \to \mathcal{H}$ by

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

for $x \in \mathcal{H}$. Show that T is a compact operator and find T^*. [15]

5. Let $\{e_n\}$ be an orthonormal basis of an infinite dimensional separable Hilbert space. For each n, let $f_n = e_{n+1} - e_n$. Show that the vector subspace generated by the sequence $\{f_n\}$ is dense. [15]

6. Let Y be a closed linear subspace of a normed linear space X. Show that

$$Y = \bigcap\{\ker f : f \in X^*, Y \subseteq \ker f\}.$$

[10]

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination: (2014–2015)

M. Stat 2nd Year

Statistical Computing

Date:08.09.14 Marks: ..30.. Duration: .2 hours.

Attempt all questions

1. Let the lifetimes of electric light bulbs of a certain type have uniform distribution in the interval $(0, \theta]$, where $\theta > 0$ is unknown. A total of $n + m$ bulbs are tested in two independent experiments. The observed data consists of $\mathbf{y} = (y_1, \ldots, y_n)'$ and $\mathbf{z}^* = (z^*_{n+1}, \ldots, z^*_{n+m})'$, where $\mathbf{y}$ are exact lifetimes of a random sample of n bulbs, and $\mathbf{z}^*$ are indicator observations on a random sample of m bulbs; that is, for $i = n + 1, \ldots, n + m$.

$$\begin{aligned} z^*_i &= 1 \text{ if bulb } i \text{ is still burning at a fixed time point } T > 0 \\ &= 0 \text{ if expired.} \end{aligned}$$

Let $\mathbf{z} = (z_1, \ldots, z_m)' = (y_{n+1}, \ldots, y_{n+m})'$ denote the missing data. Also, let $s = \sum_{i=n+1}^{n+m} z^*_i \geq 1$.

   (a) Then directly obtain the maximum likelihood estimator (MLE) of θ.

   (b) Derive an EM algorithm to obtain the MLE of θ.

   (c) Does the result of the EM algorithm match the directly estimated MLE? Justify your answer.

Marks: 2+3+5=10

2. (i) Consider obtaining the ordinary linear regression estimate of $\boldsymbol{\theta}$ by minimizing the least squares criterion $\sum_{i=1}^{m}(y_i - \mathbf{x}_i'\boldsymbol{\theta})^2$. Construct an MM algorithm for estimating the minimizer $\hat{\boldsymbol{\theta}}$. Is there any advantage of your MM algorithm over the usual procedure of exact minimization?

(ii) Consider the function

$$f(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2.$$

This function has global minima at $x = \pm 1$ and a local maximum at $x = 0$.

(a) Show that the function

$$g(x|x_n) = \frac{1}{4}x^4 + \frac{1}{2}x_n^2 - xx_n$$

majorizes $f(x)$ at x_n and leads to the MM update $x_{n+1} = x_n^{1/3}$.

(b) Prove that the alternative update $x_{n+1} = -x_n^{1/3}$ leads to the same value of $f(x)$, but the first update $(x_{n+1} = x_n^{1/3})$ always converges while the second $(x_{n+1} = -x_n^{1/3})$ oscillates in sign and has two converging subsequences.

Marks: 4+3+3=10

3. For nodes $x_0 < x_1 < \cdots < x_n$ and function values $f_i = f(x_i)$, develop a quadratic interpolating spline $s(x)$ satisfying

(a) $s(x)$ is a quadratic polynomial on each interval $[x_i, x_{i+1}]$,

(b) $s(x_i) = f_i$ at each node x_i,

(c) the first derivative $s'(x)$ exists and is continuous throughout the entire interval $[x_0, x_n]$.

Do you require any additional information to completely determine the spline? Justify your answer. **Marks**: 10

Theory of Finance I

Midsem. Exam. / Semester I 2014-15

Time - 2 hours/ Maximum Score - 30

09·09·14

1. (3+3+6=12 marks)

   (a) Let $\{R_i\}$ be a sequence of i.i.d. random variables with mean zero, variance $2/\lambda^4$ and moment generating function $\Psi(t) = \frac{1}{\lambda^2 - t^2}$ where μ and σ^2 are the mean and the variance, respectively. Let $X_n = R_1 + \cdots + R_n$. Define a filtration $\{\mathcal{F}_n\} = \sigma\{R_1, \ldots, R_n\}$. Are the sequences below follow martingale property w.r.t. $\{\mathcal{F}_n\}$ under some conditions on λ? Justify.

   (i) $Z_n = X_n^2 - n/2$; (ii) $W_n = e^{X_n}$.

   (b) Let the daily price of a stock move according to $\{\exp(X_n)\}$ in (a) above with an appropriate $\lambda > 1$, so that it is a martingale. Define, $T = \inf\{n \geq 0 : S_n \geq b,\ or,\ X_n \leq a\}$, with $a < 0 < b$. Find the probability that starting at 1, the price of the stock reaches a before reaching b Calculate the expected number of days it takes to reach the upper or lower boundary starting from 1.

   State clearly any theorem that you may be using. In calculating your answer you may assume the process hits the boundary exactly, as in continuous time case.

2. (6+5=11 marks) (a) Assume the market has only a Bond (which grows at the interest rate r) and an asset whose price process is given by S. Further assume that in the next time period it can go up or down (i.e, a one step Binomial model). Suppose your contingent claim is the Call payoff at time T with strike price K. Can you give a self financing strategy (sfs) to replicate the claim.

   Is this market complete? Justify. If the answer is yes, find the unique risk neutral probability, p, for going up.

   (b) Assume the process above is trinomial (i.e., asset can take 3 possible values) then find an sfs to replicate the claim of the call payoff.

   Is this market complete. Justify.

3. (4+3+5=12 marks)

   A trader buys an European Call option on a stock (S) for $Rs.C$ with a maturity in T months. The initial stock price was $Rs.S_0$ and the strike price was $Rs.K$. Assume the price of the option follows risk-neutral valuation and the stock follows the Black-Scholes model with the risk-free interest rate r and volatility σ

   (a) Find the probability that the option will be exercised. Would the probability increase if (i) σ increases? (ii) T increases? Justify your answers.

   (b) Calculate the expected value of S_T when the option is exercised.

   (c) Can you find the Call price in terms of **this expected value** and the **probability of exercise the option**, found above? Should the Call price increase if (i) S_0 increases? (ii) r increases? Justify your answer.

All the best.

Indian Statistical Institute
Analysis of Discrete Data
M-II, Midsem

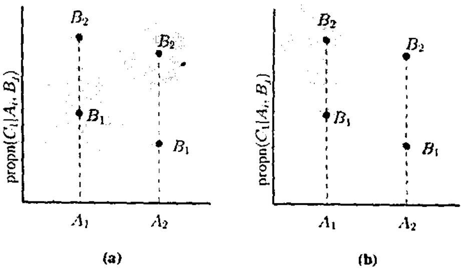
Date: Sep 10, 2014

Duration: 2hrs.

Attempt all questions. The maximum you can score is 40. Justify all your steps. This is a closed book examination. You may use your own calculator. You are allowed to use your own "cheat sheet" containing formulae.

If copying is detected in the solution for any problem, all the students involved in the copying will get 0 for that problem. Also an additional penalty of 5 will be subtracted from the overall aggregate of each of these students.

1. *(i)* We have two 2 × 2 × 2 contingency tables with the same set of factors A, B and C. The levels of any factor F are denoted by F_1, F_2. Consider the following two diagrams that show the proportions of C_1 for different (A_i, B_j) combinations. The radius of each circle is proportional to the frequency of the (A_i, B_j)-th cell. Which one is more likely to have Simpson's paradox?



(ii) While analysing a 2 × 2 × 2 table for Gender×Medium×Performance we get the following output.

```
>  dat = read.csv('simpms.csv',head=T)
> names(dat)
[1] "Gender"       "Medium"      "Performance"
> oddsratio(xtabs(~Gender+Medium+Performance,dat))
     Bad       Good
-4.576833 -3.734559
> oddsratio(xtabs(~Gender+Performance+Medium,dat))
 Bengali  English
1.449116 2.335157
> oddsratio(xtabs(~Medium+Performance+Gender,dat))
   Female       Male
0.06252036 0.84255142
> oddsratio(xtabs(~Gender+Medium,dat))
[1] -4.465384
> oddsratio(xtabs(~Gender+Performance,dat))
[1] 1.704546
> oddsratio(xtabs(~Medium+Performance,dat))
[1] -0.3689405
```

Does this show Simpson's paradox? 5+5

2. A social scientist has constructed the following 2 × 2 contingency table:

	Tuition	No tuition
Admitted	45	7
Not admitted	40	9

The data were collected as follows. A simple random sample of size 52 was drawn with replacement from the (large) population of all students admitted to various top ranking institutes in India. Another independent sample

101 students were asked if they had taken any special tuition for these entrance exams. The scientist wants you to estimate the chance that a randomly selected student taking tuitions will get admitted to some top ranking institute. What will be your response? Justify your answer in a layman's language. (Social scientists are not always brilliant in maths!) [5]

3. Data are collected about number of traffic rule violations (over a period of one hour) from different junctions in a big city using CCTV footage. The junctions are either **Manned** (traffic police present) or **Unmanned.** The data were collected at different time points classified as either **Busy** hour or **Off** hour.

	Manned	Unmanned
Busy	7	10
Off	2	5

Model the data using a Poisson generalised linear model using log link where the systematic component is a 2-way ANOVA without interaction. Write down the log-likelihood function. Clearly write down how you will compute the MLEs of the parameters. Also write down the corresponding saturated model. [10+5

4. For the following contingency table find the γ coefficient.

	< primary	Primary	Secondary	HS	≥ College
Low income	3	6	23	12	2
Middle income	0	0	2	43	128
High income	0	0	0	0	34

A statistician wants to test the presence of positive association between income level and education level using γ. But with so many empty cells she is not comfortable about using asymptotics. She wants to perform a simulation test instead. Could you please help her with this. No need to write any R code. Just write down the steps in plain English assuming that you can draw random numbers from any standard distribution. [5+10]

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination: Semester I (2014-15)
M. Stat. II Year

ACTUARIAL METHODS

Date: September 10, 2014 Maximum Marks: 50 Duration: 2 hr

*Note: Answer **any five** of the following questions.*

1. The number of claims on a portfolio of insurance policies follows a Poisson distribution with parameter 25. Individual claims may be regarded as realizations of a random variable Y=200X, where X has the distribution with probability density function

$$f(x) = \begin{cases} \frac{2}{25}(5-x), & 0 < x < 5, \\ 0, & \text{otherwise.} \end{cases}$$

In addition, for each claim, there is a 25% chance that an additional fixed expense of 500 will be incurred.

   a) Calculate the mean and variance of the total individual claim amounts.
   b) Hence compute the mean and variance of the aggregate claims on the portfolio.

[4+6=10]

2. For a general insurance policy, the settled amount of a particular claim (denoted by X) has uniform distribution over the interval $(0, \theta)$, where θ is the maximum claim size permissible for that risk. An analyst, who has knowledge of X but does not know θ, wants to guess the value of θ from a single observed value of X. The analyst prefers to use the absolute error loss function for this purpose.
   a) If the prior distribution of θ is $f(\theta) = \theta e^{-\theta}$ for $\theta > 0$, derive the Bayes estimator of θ with respect to the absolute error loss function.
   b) Suppose no prior is used, and that the analyst uses the estimator kX, where k is an appropriate constant. Determine the value of k he should use so that the estimator has the smallest mean absolute error for any fixed value of θ.

[5+5=10]

3. A health insurance company has option of three products to sell in the market and must decide which product to sell in the coming year. There are three possible choices: *basic, lean* or *rich* product, each with different related costs based on the complexity of the product. The manufacturer has fixed overheads of Rs. 1,500,000.
The revenue and cost for each product are as follows.

Policy	Cost	Revenue per policy
Basic	500,000	1500
Lean	300,000	1000
Rich	1,000,000	2000

The insurer had sold 2,100 policies last year, and is preparing forecasts of profitability for the coming year based on three scenarios: *Low sales* (80% of last year's level), *Medium sales* (same as last year's level) and *High sales* (20% higher than last year's level).

a) Determine the annual profit in rupees for each product under each scenario.

b) Determine the minimax solution to the choice problem.

c) Determine the Bayes solution, if there is 20% chance of *Low sales,* 60% chance of *Medium sales* and 20% chance of *High sales.*

[3+3+4=10]

4. The table below shows aggregate annual claim statistics for four different general insurance products over a period of five years. Annual aggregate claims for product i in year j are denoted by $X_{ij}, i = 1,2,3,4; j = 1,2,\cdots,5$.

Product (i)	$\bar{X}_i = \frac{1}{5}\sum_{j=1}^{5} X_{ij}$	$\frac{1}{4}\sum_{j=1}^{5}(X_{ij} - \bar{X}_i)^2$
1	125	300
2	85	60
3	140	35
4	175	100

a) Write down the assumptions made in the specification of EBCT Model 1.

b) Calculate the credibility premium of product no. 3 under this model.

c) Explain why the credibility factor is relatively high in this case.

[4+4+2=10]

5. A reinsurer believes that claims from a catastrophic event insurance policy follow a Pareto distribution

$$f(x;\,\alpha,\lambda) = \frac{\alpha\lambda^{\alpha}}{(x+\alpha)^{\alpha+1}}, \qquad x > 0,$$

with parameters α = 3 and λ= 500. The reinsurer wishes to draft an excess-of-loss agreement such that 50% of the losses result in no claim on the reinsurer.

a) Calculate the size of the deductible (that is, the retention limit).

b) Calculate the average claim amount net of the deductible, in respect of those losses that result in some amount of claim on the reinsurer.

c) In case a direct insurer agrees to the above reinsurance arrangement, calculate the average claim amount for that insurer.

[2+4+4=10]

6. The monthly number of claims (N) from a certain insurance policy has a Poisson distribution with parameterλ, where λ itself is a random variable having an exponential distribution with parameter 3. Individual claims are assumed to have a gamma distribution with mean 2 and variance 2.

a) Find the distribution of N.

b) Determine the moment-generating function of the monthly aggregate claims under this policy.

c) In case of excess-of-loss reinsurance with a retention level M, deduce the distribution and moment-generating function of the aggregate claims paid by the reinsurer based on the actual number of claims paid by him.

[2+4+4=10]

INDIAN STATISTICAL INSTITUTE

First-Semestral Examination : (2014-2015)

M. Stat 2nd Year

Pattern Recognition and Image Processing

Date: September 12, 2014 Maximum marks: 60 Time: 2 hours.

Note: Attempt all questions. Maximum you can score is 60. Answer Group A and Group B questions in separate answerscripts.

Group A

1. In a two-class classification problem with two features, the density of the observations from the first class is $N_2(\mathbf{0}, \mathbf{I})$ while that for the second class is $N_2(\mathbf{0}, 4\,\mathbf{I})$.

   (a) Assuming equal prior probabilities and costs of misclassification, find the Bayes optimal error for this problem. [8]

   (b) If the prior probabilities are unknown, give a rough sketch of the boundary produced by any admissible rule. [4]

   (c) What will be the form of the minimax classification rule? [6]

2. Suppose based on n independent observations $X_1, \ldots, X_n$ from the univariate standard normal density, we want to estimate the density.

   (a) Find the bias of the naive density estimate. [8]

   (b) Find the bias of the kernel density estimate using a gaussian kernel. [8]

Group B

1. Consider the task of segmenting a gray level image into the foreground and the background. Assume that the gray level values in the image are random samples from a finite normal mixture distribution.

   (a) Define such a mixture model for the task above indicating the mixture parameters to be estimated. [3]

   (b) Derive the estimates of these parameters on the basis of the gray level values in the image. [12]

   (c) Explain how a pixel is classified into the foreground and the background on the basis these estimates. [3]

2. Let f be a real valued function defined on the interval [0,7]. Suppose the maximum value of f occurs only at 5. Suppose you have only the crossover operation in a Genetic Algorithm framework with the probability of crossover being $0 < pc < 1$. If the string length of the chromosome is 3 and the initial population is {000, 111}, then

   (a) what is the probability that the maximum value of f will be attained in the second generation ? [4]

   (b) what is the probability that the maximum value of f will be attained in the third generation ? [4]

3. Consider a set S of 8 points in the plane where 7 of the points lie on the circumference of a circle so that the Euclidean distance between any two adjacent points is the same and the other point is the center of the circle. If we want to partition S into two clusters by single linkage clustering using Euclidean distance, what will be the clusters? Justify your answer. [8]

INDIAN STATISTICAL INSTITUTE
Mid Semestral Examination: (2014-2015)
MS (Q.E.) II Year
International Economics I

Date: 12·09·14 Maximum Marks 40 Duration 3 hours

Answer all questions

1. Consider an economy with two goods x and y and three types of individuals. Each type 1 individual is endowed with 1 unit of x and 0 units of y. Each type 2 is endowed with ½ units of x and ½ units of y. Each type 3 is endowed with 0 units of x and 1 unit of y. All individuals have the same utility function $U = x_c^{1/2} y_c^{1/2}$ where x_c, y_c are consumption levels of the two goods. There are 100 individuals of each type.
   (a) Find the equilibrium relative price under autarky.
   (b) Find the levels of sales and purchase of each type of individual under autarky.
   (c) Now suppose the possibility of international trade opens up and our economy is a small price-taking economy in the world market. The relative price in the international market is 1. Find the levels of exports and imports of the country.
   (d) Find the levels of sales and purchases of each type of individual in trade equilibrium.
   (e) Check for each type of individual whether there is a gain or loss from trade.

   [2x5=10]

2. Using the structure of a two country, two good specific factors model, show that if one of the sectors is demand constrained, then the country exporting the demand constrained good unambiguously gains and the country importing the demand constraint good unambiguously loses in free trade equilibrium.

   [10]

3. Show that in a three-agent setting, a transfer paradox might occur even when the equilibrium is Walras stable. In this context discuss the role of substitution effects in ensuring normal results. [10]

4. Consider a 2 country, 2 commodity trading world, with perfectly competitive markets and show that imposing an ad- valorem export tax is the same as imposing an ad-

valorem import tariff when the government redistributes all tax and tariff revenues lump sum. Also derive the optimal export tax.

[Note: An export tax on good i means the following: $p_i(1+\tau) = p_i^*$, where p_i is the domestic price and p_i^* is the international price of good i and τ is the ad valorem export tax rate.] [10]

Indian Statistical Institute

MStat II Year   MidSemester Examination   15·09·2014

Advanced Design of Experiments

Keep your answers BRIEF and to the point

Answer all questions. Total marks: 40   **Time: One and a half hours**

1. (a) Obtain the elements of GF(7), i.e., a Galois field of order 7. Show the steps starting from the cyclotomic polynomial to obtaining the primitive element.

   (b) What are the quadratic residues in GF(7)?

   (d) Show that a Hadamard matrix of order N, denoted by H_N, say, can exist only if N is divisible by 4.

   (e) Using the elements in (a) and (b) above, construct H_8 (only first 3 columns of H_8 need be shown), showing the relevant steps in this construction. [4 × 4 = 16]

2. (a) Define an orthogonal array $OA(N, k, s, t)$.

   (b) Will any $k' \times N$ subarray ($k' < k$) of an $OA(N, k, s, t)$ also be an orthogonal array? Justify your answer.

   (c) Will an orthogonal array $OA(24, 12, 2, 3)$ exist? Justify your answer, actual construction not needed. [2+2+4=8]

3. (a) Consider a 3 × 2 × 3 factorial experiment in 3 factors F_1, F_2, F_3. Write down explicitly the full set of orthonormal contrasts belonging to main effect F_3 and interaction effect F_1F_2.

   (b) Prove that the first set of contrasts obtained by you in (a) above do actually (i) belong to main effect F_3, and they (ii) give a full set of such contrasts.

   (c) In an experiment there are 4 factors, each at level 2. Construct a Resolution (1,2) fractional factorial plan in 8 runs.

   [4 + 4 + 6 = 16]

INDIAN STATISTICAL INSTITUTE

203 B. T. Road, Kolkata 700 108

Master of Statistics (M.Stat.) IInd Year
Advanced Probability I

Academic Year 2014 - 2015: Semester I

Mid Semester Examination

Instructor: Antar Bandyopadhyay

Date: September 16, 2014 **Total Points: 50**
Time: 10:30 AM - 01:30 PM **Duration: 3 Hours**

Note:

- **Please write your roll number on top of your answer paper and DO NOT write your name.**
- **There are five problems carrying a total of 60 points. Solve as many as you can. Show all your works and write explanations when needed. Maximum you can score is 50 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). However, 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***. Write brief reasons supporting your answers. For each **correct guess you will get $+1$ point** but for each **wrong guess you get -2 point**. If your guess is **correct and your reasoning is also correct then you will get an additional $+4$ points**. However, if you give a **wrong reasoning then you will receive additional -3 points**.
[$(1+4) \times 3 = 15$]

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

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

   (c) If $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ are two *martingales* defined on the same probability space with their respective *natural filtrations*, then $(X_n + Y_n)_{n\geq 0}$ is also a martingale with respect to its natural filtration.

2. Let T be an uncountable index set and $\mathcal{F} := \bigotimes_{t\in T} \mathcal{B}_{\mathbb{R}}$ be the product σ-field on $\mathbb{R}^T$. Show that, for any $B \in \mathcal{F}$, there exists a countable set $T_0 \subset T$ and a set $B_0 \in \bigotimes_{t\in T_0} \mathcal{B}_{\mathbb{R}}$, such that $\omega := (\omega(t))_{t\in T} \in B$, if and only if, $(\omega(t))_{t\in T_0} \in B_0$.
[10]

[P.T.O.]

3. 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. [10]

4. Consider $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \mathbf{P})$ where $\mathbf{P}$ is a probability such that $\mathbf{P}(B) = \mathbf{P}(-B)$ for every $B \in \mathcal{B}_{\mathbb{R}}$. Let $X : \mathbb{R} \to [0, \infty)$ be defined as $X(x) = x^2$. Find a *regular conditional probability (RCP)* on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ given the random variable X. Is it a *proper* RCP? [8 + 2 = 10]

5. Let $X_1, X_2, \cdots$ be independent random variables with

$$X_n = \begin{cases} 1 & \text{with probability } 1/(2n); \\ 0 & \text{with probability } 1 - 1/n; \text{ and} \\ -1 & \text{with probability } 1/(2n). \end{cases}$$

Let $Y_1 = X_1$ and for $n \geq 2$ recursively define

$$Y_n := \begin{cases} X_n & \text{if } Y_{n-1} = 0; \text{ and} \\ n\, Y_{n-1}\, |X_n| & \text{otherwise.} \end{cases}$$

Let $\mathcal{F}_n := \sigma(X_1, X_2, \ldots, X_n)$, $n \geq 1$.

(a) Show that $(Y_n, \mathcal{F}_n)_{n \geq 1}$ is a martingale. [4]

(b) Show that

$$\lim_{n \to \infty} \left(\mathbf{E}[|Y_n|] - \log n\right) = \gamma - \frac{\pi^2}{6}.$$

where γ is the *Euler Constant.* [4]

(c) Show that $Y_n \xrightarrow{\mathbf{P}} 0$ as $n \to \infty$. [2]

(d) Show that $(Y_n)_{n \geq 1}$ does not converge almost surely. [5]

Good Luck

Indian Statistical Institute
Midterm Examination
First Semester, 2014-2015 Academic Year
M.Stat. 2nd Year
Topics in Bayesian Inference

Date: 17 September, 2014 Total Marks : 40 Duration: $2\frac{1}{2}$ Hours

Answer all questions

1. What is the difference between the classical paradigm and Bayesian inference with respect to evaluation of performance of a statistical procedure? Give an example of an inference problem where classical inference gives a paradoxical answer while Bayesian inference gives reasonable answer. [3+5=8]

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

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

   (b) Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples, respectively, from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, where σ^2 is known. Construct a $100(1-\alpha)\%$ HPD credible interval for $(\mu_1 - \mu_2)$ assuming a uniform prior for (μ_1, μ_2). [4]

   (c) Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ 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. Construct a $100(1-\alpha)\%$ HPD credible interval for the common mean μ assuming a uniform prior. Compare this with the frequentist $100(1-\alpha)\%$ confidence interval for μ. [4]

4. Let $X_1, \ldots, X_n$ be i.i.d. with a common density $f(x|\theta)$ where $\theta \in R$.

   (a) 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. [2+6=8]

(b) Show that the above result implies consistency of the posterior distribution of θ at θ_0. [3]

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

Mid-semester Exam-2014, Indian Statistical Institute, Kolkata.

M.Stat. II year, Statistical Methods in Genetics I DATE : 18.09.14

Answer all questions. Show your works to get full credit.

1. In the ITO representation system, show that the transition probability matrix for grandparent-grandchild relationship is T^2. Hence show that $T^n \to 0$ as $n \to \infty$, implying that the relationship disappears after many generations. **5+5**

2. (a) The ability to taste PTC is due to a single dominate allele "T". You sampled 215 individuals in biology, and determined that 150 could detect the bitter taste of PTC and 65 could not. Calculate all of the genotype frequencies under HWE. **5**
(b) What allelic frequency will generate twice as many recessive homozygotes as heterozygotes under HWE? **5**

3. Consider the problem related to sex linked genes. Suppose for a particular sex linked gene, the initial population (generation-0) is given by $(0.30, 0.70) \times (0.20, 0.50, 0.30)$. Without assuming equilibrium at any stage, give the genotype combinations for males and females in generations 1, 2 and 3. **10**

4. Consider a particular gene for which there are three genotypes AA, Aa and aa with probabilities 0.4, 0.4 and 0.2 respectively. Assume that the population under consideration has attained Hardy-Weinberg equilibrium. We consider a family where Jim has two children, Alex and Trevor. Prince is the son of Trevor.
(a) Given that Jim is Aa, what is the probability that Prince is aa? **5**
(b) What is the probability that Prince is Aa and Alex is AA? **5**

5. Consider the multiple alleles for blood types. Note that there are three alleles A, B and O with probabilities p, q and r respectively. Assuming HWE, construct the transition probability matrix for parent-child. (You just have to provide the 6×6 matrix). **10**

6. Consider the linkage analysis for an F_2 design. Suppose we have two genes with alleles A, a and B, b respectively and r be the recombination fraction. For a sample of n subjects, let $\hat{r}$ be the mle of r.
Show that $var(\hat{r})$ can be estimated as $\frac{\hat{r}(1-\hat{r})(1-2\hat{r}+2\hat{r}^2)}{2(1-3\hat{r}+3\hat{r}^2)n}$. **10**

INDIAN STATISTICAL INSTITUTE

Mid-term Examination, First Semester: 2014-15

M.Stat. II Year (AS)

Life Contingencies

Date: 18 September, 2014 Time: 2:30 pm

Duration: 2 hours Maximum Marks: 60

Note: (i) Desk calculators are allowed; (ii) Actuarial tables are allowed; (iii) Symbols and notations have their usual meaning. The entire question paper is for 67 marks.

1. If $\mu_x = \frac{3}{100-x} - \frac{10}{250-x}$ for $40 < x < 100$,
   (a) calculate ${}_{40}p_{50}$, [3]
   (b) obtain a simple numerical equation from which the median of T_{40} can be calculated. [3]

2. A mortality table is formed by computing one-year conditional death probability, q_x (for different ages x), from a complicated distribution. Another mortality table is formed similarly from a distribution with force of mortality twice as high as the first one. Will the one-year conditional death probability at age x in this table be smaller than q_x, equal to q_x or larger than q_x? [4]

3. (a) If $p_{65} = 0.97$ and $p_{66} = 0.96$, calculate the probability that (65) will die between ages $65\frac{3}{4}$ and $66\frac{3}{4}$ under the assumption of uniform distribution of death (UDD) within each integer year of age. [4]
   (b) If you want to calculate ${}_{\frac{3}{4}|}q_{[65]+1}$ by assuming UDD, which of the following will you look for in a select-and-ultimate life table with select period of two years: (i) $p_{[65]+1}$ and $p_{[66]+1}$, (ii) $p_{[65]+1}$ and $p_{[65]+2}$, (iii) $p_{[65]+1}$ and p_{67}, (iv) p_{66} and p_{67}? [2]

4. Prove and interpret the relation $A_x^{(m)} + d^{(m)}\ddot{a}_x^{(m)} = 1$. [4+2]

5. Arrange in increasing order of values: $P(\bar{A}_x)$, $P^1_{x:\overline{n|}}$, $P(\bar{A}_{x:\overline{n|}})$, $P(\bar{A}^1_{x:\overline{n|}})$, $\bar{P}(\bar{A}_{x:\overline{n|}})$, $P^{(2)}(\bar{A}_{x:\overline{n|}})$. [5]

6. (a) Prove that ${}_{k|}A_x = {}_kE_x A_{x+k}$. [3]

   (b) By taking expectations of suitable 'present value' random variables, prove the relation $A_x = A^1_{x:\overline{k|}} + {}_{k|}A_x$. [2]

   (c) Interpret the relation proved in part (b). [2]

   (d) By using the relations of parts (a) and (b), calculate $A^1_{30:\overline{30|}}$ from the AM92 tables, assuming $i = 0.04$. [3]

   (e) Calculate $A^1_{30:\overline{30|}}$ from the AM92 tables, assuming $i = 0.06$. [2]

7. Calculate the following quantities from the AM92 tables, assuming $i = 0.04$:

   (a) ${}_{3|2}q_{50}$; [2]

   (b) ${}_{4|}q_{[60]+1}$; [2]

   (c) Variance of the present value of an insurance benefit, where the mean of that random variable is $A_{[40]}$; [2]

   (d) ${}_{3|}\ddot{a}_{45}$. [2]

8. An assurance contract provides a death benefit of Rs. 1,000 payable immediately on death and a survival benefit of Rs. 500 payable on every fifth anniversary of the inception of the policy. The force of mortality is $\mu_x = 0.05$ for all x, and the force of interest is $\delta = 0.04$. Calculate the level premium payable annually in advance for life. [7]

9. A life office issued 750 identical 25-year term assurance policies to lives aged 30 exact, each with a sum assured of Rs. 75,000 payable at the end of year of death. Premiums are payable annually in advance for 20 years or until earlier death.

   (a) Show that the annual net premium for each policy is approximately equal to Rs. 104 using the basis given below. [3]

   (b) Calculate the net premium reserve per policy at the ends of the 19th and 20th year of the policy. [5]

   (c) Calculate the mortality profit or loss to the life office during the 20th year if twelve policyholders die during the first 19 years of the policies and two policyholders die during the 20th year. [5]

   Basis: Mortality: AM92 Ultimate; Interest: 4% per annum.

INDIAN STATISTICAL INSTITUTE

First-semestral Examination : (2014-2015)

M. Stat 2nd Year

Pattern Recognition and Image Processing

Date: November 17, 2014 Maximum marks: 100 Time: 3 hours.

Note: Attempt all questions. Maximum you can score is 100. Answer Group A and Group B questions in separate answerscripts.

Group A

1. Consider a J-class problem with equal class probabilities based on a single feature x. The density of x for the j-th class is uniform over the union of intervals $(0\,,\,\frac{rJ}{J-1})$ and $(j\,,\,j+1-\frac{rJ}{J-1})$ where r is fixed, and $0<r<\frac{J-1}{J}$. Find out the overall Bayes error probability for this problem. Also find the overall asymptotic error probability of the 1-NN rule and compare with the Bayes error probability in this case. [6 + 6]

2. For CART,

   (a) show that the resubstitution error rate of a tree can not increase when a terminal node is split. [7]

   (b) describe briefly the cost-complexity pruning algorithm for generating the α-sequence and the corresponding minimizing sub-trees. [10]

   (c) show that the α-sequence is strictly non-decreasing. [3]

   (d) show that the minimizing subtree for any α-value is unique, and the resulting sequence of trees is nested. [5]

3. Show the required architecture (with appropriate weights) of a multilayer perceptron which can

   (a) find the minimum of two positive real numbers and identify which one is the minimum. [8]

   (b) solve the two-dimensional two-class classification problem where the first class consists of all points in the first and the third quadrants of the measurement space and the second class is the complement of the first class. [10]

[P. T. O.]

Group B

1. (a) Describe the histogram equalization approach to contrast enhancement in a gray level image.

   (b) Define the morphological operation opening. Illustrate with examples how it can be used to remove noise in a binary image. [5 + (6+5)]

2. (a) Define a connected component in a binary image. Suppose in a binary image there are several connected components among which one is U shaped. Describe a linear time (in terms of the number of pixels) algorithm to assign unique labels to the connected components.

   (b) Describe an algorithm to rotate a binary image of size $n \times n$ around the center of the image by an arbitrary angle. [(3+9) + 5]

3. Describe how a colour image is converted from the Red-Green-Blue space to the Hue-Saturation-Value space. Describe an algorithm to segment a colour image in terms of hue alone. [5 + 6]

4. Define a hidden Markov model. Describe an algorithm to compute the probability of a finite sequence (of length T) of observation vectors generated by a given hidden Markov model with computational complexity not higher than $O(Tn^2)$. [5 + 6]

Indian Statistical Institute

MStat II Year Semester Examination 18 Nov, 2014

Advanced Design of Experiments

Answer all questions. Maximum you can score is: 60 Time: Three hours

1. a) Define the D-optimality criterion and explain its statistical significance.

   b) State and prove the D-optimality of balanced incomplete block designs for estimating full sets of orthonormal treatment contrasts under the usual model.

   c) Construct a D-optimal block design with 7 treatments and 7 blocks, each block being of size 10.

   d) Verify if the design constructed in (c) above will also be universally optimal for treatment effects in the class of all blocks designs with $t = 7$, $b = 7$ and $k = 10$. (You must actually check whether it satisfies the sufficient conditions for universal optimality). [(2+2)+4+4+4=16]

2. a) Define a strongly balanced uniform crossover design. What are the necessary conditions on the design parameters for such a design to exist?

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

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

   c) Let Ω be the class of all crossover designs d with t treatments, n subjects and p periods. For $d \in \Omega$, let the information matrix for estimating direct and carryover effects jointly under the standard model be written as $\begin{pmatrix} C_{d11} & C_{d12} \\ C_{d21} & C_{d22} \end{pmatrix}$.

   Let $d^* \in \Omega$ be a strongly balanced uniform design. Show that $C_{d^*12} = 0$ and C_{d^*11} is completely symmetric. [(2+2) + 3+ 3+ (3 + 3)=16]

3. (a) Does a Hadamard matrix of order 28 exist? Justify your answer. (Actual construction not needed.)

   (b)Does an orthogonal array $OA(81, 10, 9, 2)$ exist? Justify your answer. (Actual construction not needed.

   (c) An optimal main effects plan is required for a factorial experiment in 11 factors with each factor at 2 levels. What is the smallest number of runs required for this plan? Indicate its construction.

   (d) What is meant by a N-run resolution (2,3) plan in the set up of a $2 \times 3 \times 4$ factorial experiment? [3+3+(2+3)+2=13]

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

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

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

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

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

Write

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

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

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

(b) Show that X is nonsingular.

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

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

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

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

(f) On the basis of your findings in (e), will you recommend the use of the fraction d ? Or, can you suggest some other design which will perform even better? Justify your answer. [3+3+2+3+2+5=18]

Indian Statistical Institute
Semestral Examination
First Semester, 2014-2015 Academic Year
M.Stat. 2nd Year
Topics in Bayesian Inference

Date: 21.11.2014 Total Marks : 60 Duration: 3 Hours

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

1. Let $X_1, \ldots, X_n$ be iid $N(\theta, \sigma^2)$ variables.

   (a) Consider a standard noninformative prior for (θ, σ^2) and find the posterior distribution of 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}. [3+(2+3)=8]

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

   (b) Let $X_1, \ldots, X_n$ be iid $N(\theta, 1)$ and suppose θ has a two-point prior distribution π with $\pi(\theta = 1) = \pi(\theta = -1) = \frac{1}{2}$. Supoose the true value of θ is $\theta_0 = \frac{1}{2}$. Show that the posterior mean of θ does not converge to θ_0 almost surely under θ_0. [5+5=10]

3. Define the Jeffreys prior. Describe in a few sentences the justification of thinking of Jeffreys prior as a noninformative/low information prior. [5]

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

(b) Consider the general nested model selection problem and derive the intrinsic prior determining equations. [6+4=10]

5. Consider p independent random samples, each of size n from p normal populations $N(\theta_j, \sigma^2)$, $j = 1, \ldots, p$ where σ^2 is known. Our problem is to estimate $\theta_1, \ldots \theta_p$. We assume that $\theta_1, \ldots, \theta_p$ are iid $N(\eta_1, \eta_2)$, where η_1 and η_2 are unknown constants.

   (a) Explain the parametric empirical Bayes (PEB) approach in this problem and describe how the PEB estimates "borrow strength" from the whole data at hand.

   (b) Explain why a James-Stein type shrinkage estimator might be preferable to the usual vector of sample means as an estimator of $(\theta_1, \ldots, \theta_p)$ where p is large. [4+3=7]

6. Suppose $Y_1, \ldots, Y_{n_1}$ is a random sample of size n_1 from a normal population $N(\theta_1, \sigma^2)$ whereas $Y_1, \cdots, Y_{n_2}$ is an independent random sample of size n_2 from another normal population $N(\theta_2, \sigma^2)$. Here $(\theta_1, \theta_2, \sigma^2)$ are unknown parameters. Assuming the prior $\pi(\theta_1, \theta_2, \sigma^2) \propto \frac{1}{\sigma^2}$, derive the posterior distribution of $\eta = \theta_1 - \theta_2$ and hence find a $100(1-\alpha)\%$ HPD credible set for η. [12]

7. (a) Describe the Metroplois-Hastings and Gibbs Sampling techniques for performing Markov Chain Monte Carlo.

   (b) Suppose one has to sample from the density $\frac{g(\theta)}{K}$ where $\theta \in [0, 1]$ and $K > 0$ is a normalizing constant and $g(\theta)$ is given by

$$g(\theta) = \exp(-2\theta)\theta^{\alpha-1}(1-\theta)^{\beta-1}I\{0 \leq \theta \leq 1\},$$

   $\alpha > 0$, $\beta > 0$ being known constants. Describe the accept-reject Monte Carlo method and the Metropolis-Hastings MCMC algorithm for simulation from the described density. [4+6=10]

Indian Statistical Institute
Analysis of Discrete Data
M-II, Semestral Examination

te: Nov 24, 2014 Duration: 3 hrs.

Attempt all questions. The maximum you can score is 60. Justify all your steps. This is a closed ok examination. You may use your own calculator.

If copying is detected in the solution for any problem, all the students involved in the copying will t 0 for that problem. Also an additional penalty of 5 will be subtracted from the overall aggregate of ch of these students.

1. Explain the concept of polychoric correlation. How is it used in applying SEM in item analysis? Suppose that $(X_1, Y_1), ..., (X_n, Y_n)$ are iid observations from a bivariate normal population with $N(0,1)$ marginals and correlation ρ. Let $U_i = X_i/|X_i|$ and $V_i = Y_i/|Y_i|$. Suggest how you may estimate ρ based on only $(U_1, V_1), ..., (U_n, V_n)$. [4+3+8]

2. Consider the following data set about political ideology and party affiliation:

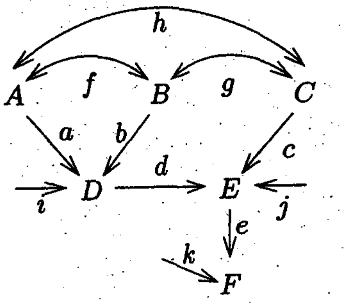
Ideology	Party1	Party2
Very liberal	80	30
Slightly libral	81	46
Moderate	171	148
Slightly conservative	41	84
Very conservative	55	99

Write down the proportional odds cumulative ordinal multicategory logistic model and also the adjacent category ordinal logistic model for this data set. R can only fit the latter, and produces the following output (here `frq` denotes the frequencies in the above table):

```
Call:
polr(formula = as.factor(ideo) ~ party, data = dat, weights = frq)

Coefficients:
 party1
0.9744634

Intercepts:
       1|2         2|3          3|4         4|5
-1.4944179 -0.5000454  1.2116427  2.0440699

Residual Deviance: 2474.985
AIC: 2484.985
```

Fit the former model using this output. If you think that this is impossible, then justify why. [15]

3. Consider the following path model where the labels and arrows have the usual meanings.



Find the correlation between the nodes A and E and also between B and E. [15]

4. Define the odds ratio parameter for a 2×2 contingency table under the binomial sampling scheme as well as under the multinomial sampling scheme. Find the MLE for this parameter under both the sampling schemes. [15]

5. Write down the homogeneous association loglinear model under geometric distribution assumption for an $m \times n$ contingency table. Also write down the log-likelihood function. [2+3]

INDIAN STATISTICAL INSTITUTE

Semestral Examination, First Semester: 2014-15

M.Stat. II Year (AS)

Life Contingencies

Date: November 24, 2014 Maximum marks: 100 Duration: $3\frac{1}{2}$ hours

Students are permitted to use non-programmable calculators and Actuarial Formulae and Tables.

1. Calculate the following quantities from the AM92 table.

   (a) ${}_{2|2}q_{[45]}$. [2]

   (b) ${}_{5}p_{[50]+1}$. [1]

   (c) ${}_{0.5}q_{65.5}$ (assume a constant force of mortality). [3]

   [Total 6]

2. Explain the rationale of using super-compound reversionary bonus in an assurance policy. [2]

3. An endowment insurance contract provides for the following benefits.
   (i) A lump sum of Rs. 500,000 payable immediately on death within 5 years.
   (ii) A lump sum of Rs. 200,000 payable on expiry of 5 years in case of survival.
   Calculate the expected present value and the variance of the aggregate benefit. [9]

   Basis: Force of mortality $\mu_x = 0.02$ throughout; interest rate $i = 4\%$ per annum.

4. A 20 year term assurance is issued to a life aged 45. The benefit amount is Rs. $1,000,000 \times (1.04)^{k-1}$ payable in case of death in the k^{th} year. The benefit is payable immediately on death, and monthly level premium is to be paid in advance.

   The following basis is used for all computations.

Mortality	AM92 Select
Interest	4%
Initial expenses	Rs. 5,000
Initial commission	60% of first 12 monthly premiums
Renewal expenses	Rs. 1,000, payable at the beginning of each renewal year
Renewal commission	2.5% of each premium starting from the 13th month
Claim expense	Rs. 2,000 (fixed) at the time of claim payment

   (a) Show that the gross monthly premium is about Rs. 765. [11]

   (b) Calculate the gross premium prospective reserve at the end of the 19th year. [4]

   [Total 15]

5. (a) Give expressions for the present value random variable, its mean and its variance, for a whole life insurance of amount S payable to a life aged x, at the end of the year of his/her death. [2]

   (b) Give an expression for the net annual premium payable in advance throughout life, in respect of the above insurance. [1]

   (c) Derive an expression for the annual premium payable if the insurer wants to ensure that there is only a 5% probability of making a loss in a portfolio of n similar policies. [5]

   [You may assume that the lives are independent and make a suitable assumption about the distribution of the aggregate loss.]

   (d) Compare the numerical values of the net premium and the premium obtained above, when $x = 50$, $n = 50$, mortality follows AM92 select and the interest rate is 4%. [2]

[Total 10]

6. Explain the four quantities appearing in the central cell of the following table and justify the row and column sums.

	A	B	$A + B$
I	${}_nq^{1}_{xy}$	${}_nq^{2}_{xy}$	${}_nq_x$
II	${}_nq_{x\overset{1}{y}}$	${}_nq_{x\overset{2}{y}}$	${}_nq_y$
$I + II$	${}_nq_{xy}$	${}_nq_{\overline{xy}}$	${}_nq_x + {}_nq_y$

[8]

7. A special annuity-cum-assurance issued to a male life aged 60 and a female life aged 50 provides for the following benefits.

   (i) A joint life annuity of Rs. 15,000 payable monthly in advance.

   (ii) A life annuity of Rs. 10,000 payable monthly in advance to the female life, starting from the month after the death of the male life, provided he dies before her.

   (iii) An annuity of Rs. 5,000 payable monthly in advance to a charity for eternity, starting from the month after the second death.

   (iv) Rs. 100,000 (lump sum) payable to the charity immediately on death of the second life.

   Calculate the total expected present value of all the benefits. [11]

   Basis: Mortality as per PMA92C20 and PFA92C20, interest 4% per annum, expenses nil.

8. (a) In an illness-death model, there are three states: Well (W), Critically ill (C) and Dead (D). The rate of transition from W to D is $\mu = 0.1$, and the rate of transition from W to C is $\sigma = 0.2$. Calculate ${}_t(aq)^D_x$ and ${}_t(aq)^C_x$. [3]

   (b) Let ${}_t(aq)^\alpha_x$, $\alpha = 1, 2, \ldots, k$ be k multiple decrement probabilities and ${}_tq^\alpha_x$, $\alpha = 1, 2, \ldots, k$ be the corresponding single decrement probabilities. Assuming ${}_t(aq)^\alpha_x = t(aq)^\alpha_x$ for $\alpha = 1, 2, \ldots, k$ (i.e., each decrement satisfies the uniform distribution of death assumption), show that the single decrement probabilities satisfy the relation

$$q^\alpha_x = 1 - [1 - (aq)_x]^{\frac{(aq)^\alpha_x}{(aq)_x}}$$

   for $\alpha = 1, 2, \ldots, k$, where $(aq)_x = \sum_{j=1}^{k} (aq)^j_x$. (4)

[Total 7]

9. An 8-year term insurance policy has the profit vector (-28, -15, -6, 7, -5, 10, 8, 15).
   If reserves are set up to zeroise future negative cash flows, determine the reserves required for each year and the revised profit vector. [6]

   Basis: Mortality 0.2% per annum for at each integer age; Interest 4% per annum.

10. (a) Describe two ways in which occupation affects mortality and morbidity. [3]
    (b) Define the crude mortality rate and the directly standardised mortality rate. [4]

    [Total 7]

11. A three-year unit linked endowment insurance is issued to a male life aged 65, with an annual premium of Rs. 25,000 payable in advance. The benefits are as under.

    Surrender: None.
    Mortality: Bid value of units at the end of the year, subject to a minimum of Rs. 50,000.
    Maturity: Bid value of units at the end of the term, subject to a minimum of Rs. 50,000.

    No non-unit reserve is held, while unit reserves are equal to the full bid value of the units. The following assumptions are made.

Non-unit fund interest rate:	5%
Unit fund growth rate:	15%
Risk discount rate:	10%
Mortality:	AM92 Ultimate
Allocated premium:	98% each year
Bid-offer spread:	5%
Management charge:	1%
Initial expenses:	Rs. 1,000
renewal expenses:	Rs. 1,000

    (a) Determine the unit fund balance at the end of each policy year, after management charges. [4]
    (b) Determine the profit made by the insurer at the end of each policy year. [6]
    (c) Calculate the net present value of the profit. [2]

    [Total: 12]

12. (a) Describe two typical methods of defining the final pensionable salary. [2]
    (b) An employee currently aged 50, had joined a pension scheme at the age of 40. He has drawn a salary of Rs. 300,000 in the last year. The pension scheme provides a pension of 1/80th of the final pensionable salary (FPS) for each year of service, on age retirement or ill health retirement. The FPS is the average of three years' salary preceding retirement. Using the Example Pension Scheme table in the Actuarial tables, calculate the expected present value of this employee's total pension. [5]

    [Total: 7]

Indian Statistical Institute

First Semester Exam: 2014-15

M.Stat. II year, Statistical Methods in Genetics I

Maximum Marks: 50, Duration: 2 hours

Answer all questions. Show your works to get full credit.

1. Consider the QTL mapping problem for a Backcross design. The goal is to locate the QTL(s) controlling the observed trait. Suppose there are n subjects in the study and $Y_1, Y_2, \ldots, Y_n$ are response measures. Note that we dont know the QTL genotypes of the subjects under study.

   Define the variables $Z_1, Z_2, \ldots, Z_n$ as follows: Z_i=0, if the i-th subject belongs to the first genotype group (normal population with mean=μ_1, variance=σ^2) and 1 for the second genotype group (normal population with mean=μ_2, variance=σ^2). Note that Z_i's are not observed and thus can be treated as "missing variables". Let π be the proportion (unknown) of individual in the population belonging to the first genotype group.

   (a) Write down the incomplete data likelihood function.
   (b) Write down the complete data likelihood function.
   (c) Give the E-step of the EM algorithm and clearly give the expression for $E(Z_i|$data, initial estimates of the parameters).
   (d) Give the closed form expressions for π, μ_1, μ_2 and σ^2 from M-step after the first iteration. **2+2+2+4**

2. (a) What do you mean by "Double First Cousins"? Show that the transition probability matrix for Double First Cousins is the square of the transition probability matrix for fullsibs. **5**
   (b) How can you establish the relationship between two transition probability matrices in part(a) intuitively? Explain properly. **5**

3. (a) Consider the QTL mapping based on Linkage Disequilibrium. Suppose we observe the marker data for n subjects where M and m are the marker alleles and Q, q are the QTL alleles. Let D be the coefficient of linkage disequilibrium between the marker and the QTL and $Y_1, Y_2, \ldots, Y_n$ be the observed responses from n subjects. Construct a Bootstrap based test procedure for testing the existence of a potential QTL controlling the observed trait. **5**
   (b) Show that the probability that the bootstrap resample does not contain the response from a particular subject will tend to e^{-1} as $n \to \infty$. **5**

4. (a) In GWAS, suppose you observe the genotypes of N subjects for a particular SNP and X_i be the SNP genotype of the i-th subject. Let Y_i be the phenotype of the individual i and Y_i=1 for case and 0 for control.

Construct an appropriate regression model to test whether this SNP is linked to the disease status. **5**

(b) What do you mean by family-wise error rate and per comparison error rate in the context of multiple testing? Define "False Discovery Rate" following Benjamini-Hochberg (1995). **5**

5. High blood pressure is considered as an important biomarker for cardiovascular disease. Suppose the goal of a study is to locate the QTLs controlling the time to get heart disease through high blood pressure.

Suppose we have N subjects from which systolic and diastolic blood pressure are measured at T different evenly spaced time points. Subjects are followed for a period of time and we also observe the time to event (heart attack) data s_i as the following: s_i is the time to event for the uncensored subjects and the last followup time for the censored subjects.

(a) Write an appropriate linear model for the observed longitudinal trait (blood pressure). **3**

(b) Give the expression of the likelihood function for time to event data in terms of the hazard function. **3**

(c) Write a joint model for the trait and the event-time and discuss how you can test the association between the trait and the event-time. **4**

INDIAN STATISTICAL INSTITUTE

203 B. T. Road, Kolkata 700 108

Master of Statistics (M.Stat.) IInd Year
Advanced Probability I

Academic Year 2014 - 2015: Semester I

Final Examination

Date: November 26, 2014 | Total Points: 100
Time: 10:30 AM - 02:30 PM | Duration: 4 Hours

Note:

- **Please write your roll number on top of your answer paper and DO NOT write your name.**
- **There are six problems each carrying 20 points. Solve Problem # 1 and any four from the remaining. If you solve all the six problems and do not indicate which ones to be graded, then only the Problem # 1 and first four from the others will be graded. Show all your works and write explanations when needed.**
- **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). However, 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***. Write brief reasons supporting your answers. For each **correct guess you will get +1 point** but for each **wrong guess you get −2 point**. If your guess is **correct and your reasoning is also correct then you will get an additional +4 points**. However, if you give a **wrong reasoning then you will receive additional −3 points**. $[(1+4)\times 4 = 20]$

   (a) If $(X_k)_{1\le k\le n}$ is a finite sequence of Bernoulli variables which are *exchangeable*, then there exists a probability π on $\left([0,1], \mathcal{B}_{[0,1]}\right)$ such that

   $$\mathbf{P}\left(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\right) = \int_0^1 p^{\sum_{i=1}^n x_i}\left(1-p\right)^{n-\sum_{i=1}^n x_i}\, \pi\left(dp\right)$$

   for any $x_1, x_2, \ldots, x_n \in \{0,1\}$.

   (b) Suppose $\nu \ll \mu$ be two probability measures. If $\{A_n\}_{n\ge1}$ is a sequence of events such that $\mu(A_n) \to 0$ as $n\to\infty$, then $\nu(A_n)\to 0$.

   (c) Let $\{X_n\}_{n\ge1}$ be i.i.d. with ±1 values with equal probabilities. Then the random series $\sum_{n=1}^{\infty}\left(\frac{\log n}{n}\right)X_n$ has a Gaussian distribution.

   (d) A *bi-infinite martingale* $(X_n, \mathcal{F}_n)_{-\infty<n<\infty}$ is necessarily uniformly integrable.

[Please Turn Over]

2. (a) Suppose $(\nu_n)_{n\geq 1}$ and $(\mu_n)_{n\geq 1}$ are two infinite sequences of probability measures defined on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, such that, $\nu_n \sim \mu_n$ for all $n \geq 1$. Show that for all $n \geq 1$, $\bigotimes_{k=1}^{n} \nu_k \sim \bigotimes_{k=1}^{n} \mu_k$ [10]

(b) Let μ_∞ and ν_∞ be the unique probability measures on $(\mathbb{R}^\infty, \mathcal{B}_{\mathbb{R}^\infty})$ such that for all $n \geq 1$, $\mu_\infty \circ \pi_n^{-1} \sim \nu_\infty \circ \pi_n^{-1}$, where $\pi_n : \mathbb{R}^\infty \longrightarrow \mathbb{R}^n$ is the projection map onto the first n coordinates. Conclude whether it is necessary that $\mu_\infty \sim \nu_\infty$. [10]

3. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and X be an integrable random variable defined on it. Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra. Suppose $\mathbf{E}\left[X \,\middle|\, \mathcal{G}\right]$ has same distribution as X.

(a) If $\mathbf{E}\left[X^2\right] < \infty$ then show that X is a.s. $\mathcal{G}$-measurable. [5]

(b) Do you think the conclusion holds in general? Justify your answer. [15]

4. Let $(S_n)_{n\geq 0}$ be an *asymmetric simple random walk* starting at 0 with mean increment $\mu > 0$.

(a) Show that $X_n := (S_n - n\mu)^2 - n\left(1 - \mu^2\right)$, $n \geq 0$ is a martingale. [5]

(b) Let $T_b := \inf\left\{n \geq 0 \,\middle|\, S_n = b\right\}$ for $b \in \mathbb{N}$. Show that T_b has finite second moment and find a formula for the $\mathbf{Var}\,(T_b)$ in terms of μ and b. Do you think T_b has all moments finite?

$[4 + 6 + 5 = 15]$

5. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $(Y_n)_{n\geq 1}$, Y and Z be random variables defined on it. Suppose $Y_n \longrightarrow Y$ a.s. as $n \to \infty$ and $|Y_n| \leq Z$ for all $n \geq 1$, where $\mathbf{E}[Z] < \infty$. Suppose $\mathcal{F}_n \uparrow \mathcal{F}_\infty \subseteq \mathcal{F}$, an increasing sequence of σ-algebras. Then show that as $n \to \infty$,

$$\mathbf{E}\left[Y_n \,\middle|\, \mathcal{F}_n\right] \longrightarrow \mathbf{E}\left[Y \,\middle|\, \mathcal{F}_\infty\right] \quad \text{a.s}$$

[20]

6. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $(M_n, \mathcal{F}_n)_{n\geq 0}$ be a bounded increment martingale sequence defined on it.

(a) Show that $\frac{M_n}{n} \longrightarrow 0$ a.s. as $n \to \infty$. [10]

(b) If the increments are also independent and non-degenerate, then show that there exists a sequence of positive real numbers $(\sigma_n)_{n\geq 0}$ such that $\left(\frac{M_n}{\sigma_n}\right)_{n\geq 0}$ converges weakly to a non-trivial distribution. When do you think the limit will be Gaussian? $[8 + 2 = 10]$

Good Luck

INDIAN STATISTICAL INSTITUTE

203 B. T. Road, Kolkata 700 108

Master of Statistics (M.Stat.) IInd Year
Advanced Probability I

Academic Year 2014 - 2015: Semester I

Final Examination

Date: November ~~DD~~ 26, 2014 Total Points: 100
Time: 10:30 AM - 02:30 PM Duration: 4 Hours

Note:

- **Please write your roll number on top of your answer paper and DO NOT write your name.**
- **There are six problems each carrying 20 points. Solve Problem # 1 and any four from the remaining. If you solve all the six problems and do not indicate which ones to be graded, then only the Problem # 1 and first four from the others will be graded. Show all your works and write explanations when needed.**
- **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). However, 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***. Write brief reasons supporting your answers. For each **correct guess you will get +1 point** but for each **wrong guess you get −2 point**. If your guess is **correct and your reasoning is also correct then you will get an additional +4** points. However, if you give a **wrong reasoning then you will receive additional −3 points**. $[(1+4)\times 4 = 20]$

   (a) If $(X_k)_{1 \le k \le n}$ is a finite sequence of Bernoulli variables which are *exchangeable*, then there exists a probability π on $\left([0,1], \mathcal{B}_{[0,1]}\right)$ such that
   $$\mathbf{P}\left(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\right) = \int_0^1 p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i} \, \pi(dp)$$
   for any $x_1, x_2, \ldots, x_n \in \{0, 1\}$.

   (b) Suppose $\nu \ll \mu$ be two probability measures. If $\{A_n\}_{n \ge 1}$ be a sequence of events such that $\mu(A_n) \to 0$ as $n \to \infty$, then $\nu(A_n) \to 0$.

   (c) Let $\{X_n\}_{n \ge 1}$ be i.i.d. with ± 1 values with equal probabilities. Then the random series $\sum_{n=1}^{\infty} \left(\frac{\log n}{n}\right) X_n$ has a Gaussian distribution.

   (d) A *bi-infinite martingale* $(X_n, \mathcal{F}_n)_{-\infty < n < \infty}$ is necessarily uniformly integrable.

[P.T.O.]

2. (a) Suppose $(\nu_n)_{n\geq 1}$ and $(\mu_n)_{n\geq 1}$ are two infinite sequences of probability measures defined on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, such that, $\nu_n \sim \mu_n$ for all $n \geq 1$. Show that for all $n \geq 1$, $\bigotimes\limits_{k=1}^{n} \nu_k \sim \bigotimes\limits_{k=1}^{n} \mu_k$ [10]

   (b) Let μ_∞ and ν_∞ be the unique probability measures on $(\mathbb{R}^\infty, \mathcal{B}_{\mathbb{R}^\infty})$ such that for all $n \geq 1$, $\mu_\infty \circ \pi_n^{-1} \sim \nu_\infty \circ \pi_n^{-1}$, where $\pi_n : \mathbb{R}^\infty \to \mathbb{R}^n$ is the projection map onto the first n coordinates. Conclude whether it is necessary that $\mu_\infty \sim \nu_\infty$. [10]

3. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and X be an integrable random variable defined on it. Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra. Suppose $\mathbf{E}\left[X \,\middle|\, \mathcal{G}\right]$ has same distribution as X.

   (a) If $\mathbf{E}\left[X^2\right] < \infty$ then show that X is a.s. $\mathcal{G}$-measurable. [5]

   (b) Do you think the conclusion holds in general? Justify your answer. [15]

4. Let $(S_n)_{n\geq 0}$ be an *asymmetric simple random walk* starting at 0 with mean increment $\mu > 0$.

   (a) Show that $X_n := (S_n - n\mu)^2 - n\left(1 - \mu^2\right)$, $n \geq 0$ is a martingale. [5]

   (b) Let $T_b := \inf\left\{n \geq 0 \,\middle|\, S_n = b\right\}$ for $b \in \mathbb{N}$. Show that T_b has finite second moment and find a formula for the $\mathbf{Var}\left(T_b\right)$ in terms of μ and b. Do you think T_b has all moments finite? [4 + 6 + 5 = 15]

5. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $(Y_n)_{n\geq 1}$, Y and Z be random variables defined on it. Suppose $Y_n \longrightarrow Y$ a.s. as $n \to \infty$ and $|Y_n| \leq Z$ for all $n \geq 1$, where $\mathbf{E}\left[Z\right] < \infty$. Suppose $\mathcal{F}_n \uparrow \mathcal{F}_\infty \subseteq \mathcal{F}$, an increasing sequence of σ-algebras. Then show that as $n \to \infty$,

$$\mathbf{E}\left[Y_n \,\middle|\, \mathcal{F}_n\right] \longrightarrow \mathbf{E}\left[Y \,\middle|\, \mathcal{F}_\infty\right] \quad \text{a.s}$$

[20]

6. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $(M_n, \mathcal{F}_n)_{n\geq 0}$ be a bounded increment martingale sequence defined on it.

   (a) Show that $\frac{M_n}{n} \longrightarrow 0$ a.s. as $n \to \infty$. [10]

   (b) If the increments are also independent, then show that there exists a sequence of positive real numbers $(\sigma_n)_{n\geq 0}$ such that $\left(\frac{M_n}{\sigma_n}\right)_{n\geq 0}$ converges weakly to a non-trivial distribution. When do you think the limit will be Gaussian? [8 + 2 = 10]

Good Luck

INDIAN STATISTICAL INSTITUTE
Final Examination: 2014-15

Course name: MSQE II
Subject name: Incentives and Organisations
Date: 26.11.14
Maximum marks: 50
Duration: 3 hours

Answer all questions

Q1. Within the framework proposed by Lazear and Rosen (Journal of Political Economy, 1981), show that with risk-neutral workers, both piece-rate contracts and competitive tournaments yield efficient allocation of resources. Argue rigorously. Make sure to specify the assumptions you are making, and clarify the notation. [15]

Q2. Within the framework proposed by Green and Stokey (Journal of Political Economy, 1983), show that if the production relationship has no common shock, then the principal will always prefer to offer agents individual contracts, rather than propose a tournament. Argue rigorously. Make sure to specify the assumptions you are making, and clarify the notation. [15]

Q3. Within the 2-agent framework proposed by Nalebuff and Stiglitz (Bell Journal of Economics, 1983), suppose an agent wins not only if his output is higher than that of the other agent, but is higher by a discrete amount or a gap. Correspondingly, he loses if his output is lower than that of the other agent by a discrete gap, with the agents tying if the difference between the two outputs is less than the gap. Would introducing a separate reward for a tie in a tournament, instead of there being only two possible rewards, one for winning and the other for losing, improve welfare? Argue rigorously. Make sure to specify the assumptions you are making, and clarify the notation. [20]

INDIAN STATISTICAL INSTITUTE

First Semestral Examination: 2014-15
M. STAT. II YEAR

ACTUARIAL METHODS

Date: November 26, 2014 Maximum Marks: 100 Duration: 3½ hr

Note: *Actuarial Tables and calculators can be used to solve the problems.*

1. A software development company has three versions of a software product on sale in the market and must decide which version to sell in the coming year. The three versions are: *basic*, *intermediate* and advanced, each with different related costs based on the complexity of the product. The company has fixed overheads of Rs. 1,500,000.

   The revenue and cost for each product are as follows.

Version	Cost	Profit per unit
Basic	500,000	1500
Intermediate	600,000	1000
Advanced	1,000,000	2000

   The company had sold 2,300 units last year, and is preparing forecasts of profitability for the coming year based on three scenarios: *Low sales* (80% of last year's level), *Medium sales* (same as last year's level) and *High sales* (20% higher than last year's level).

   a. Determine the annual profit in rupees for each version under each scenario.
   b. Determine the minimax solution to the choice problem.
   c. Determine the Bayes solution, if there is 20% chance of *Low sales*, 60% chance of *Medium sales* and 20% chance of *High sales*.

[3+3+4=10]

2.

   a. $x_1, x_2, \dots, x_n$ are independent observations from a *Gamma*(α, λ) distribution, where α is a known constant and λ has an exponential prior with parameter θ. Find the Bayes estimate of λ under *zero-one error* loss.

   b. In a large portfolio of non-life policies involving a new product, let θ denote the proportion of policies on which claims are made in the first year. The value of θ and is assumed to have a *Beta* prior with parameters α, β and mean μ_0, which is a function of α *and* β. If a random sample of n such policies gives rise to x claims in the first year, show that the posterior mean of θ is given by

$$w_n \mu_0 + (1 - w_n)({}^{x}/_{n}),$$

   where the weight w_n is a function of α, β and n.

[5+5=10]

3.

a. An actuary has reason to believe that claims from a particular type of policy follow the Burr distribution with parameters α= 2, λ=1000 and γ=0.75. Calculate the size of the deductible such that 25% of the losses do not result in any claim to the insurer.

b. Claims occur as a *Pareto* distribution with parameters α= 6 and λ=200. A proportional reinsurance arrangement is in force with a retained proportion of 80%. What would the expected loss to the insurer be if, instead, there is an excess-of-loss reinsurance arrangement with retention level 50.

[3+7=10]

4.

a. Mention two differences between collective and individual risk models.

b. Explain the concept of benefits and perils in the context of general insurance by means of examples.

c. Enumerate two differences between linear models and generalized linear models.

d. Under a group life insurance covering n lives, it is assumed that each insured life is subject to the same mortality rate (p), and that lives are independent with respect to mortality. Derive from first principles the moment-generating function of the aggregate claim amount in a given year.

[2+2+2+4=10]

5. The table below shows the claims paid (in Rs. lakh) over a four-year period by an insurer on three different general insurance products:

Product	Total claims paid (in Rs. lakhs)			
	2010	2011	2012	2013
Fire insurance	20	12	25	36
Motor insurance	100	250	175	200
Household insurance	50	63	70	62

a. Compute the estimate of the credibility premium for the year 2014 of each product under EBCT model 1, together with an estimate of its variance.

b. Enumerate the differences between EBCT models 1 and 2.

[(3+3)+4=10]

a. Consider a portfolio of 100 health insurance policies for which

i. the annual premium per policy is Rs. 5000;

ii. aggregate annual claims per policy follow a normal distribution with mean 3500 and variance 10000;

iii. annual expense per policy is Rs. 200.

If the initial surplus is Rs. 1,00,000, determine the probability of ruin at the end of the first year, if claims are assumed to be independent.

b. State Lundberg's Inequality and use it to obtain a rough approximation to the ruin probability in respect to a general insurance portfolio under which aggregate claims follow a compound Poisson process with Poisson parameter 5 and an exponential claim size distribution with mean 10, given that the initial surplus is 1000.

c. Prove that, for the adjustment coefficient (R) for aggregate claims modeled as a compound Poisson process, the following is true:

$$\frac{1}{M}\log(1+\theta) < R < 2\theta\frac{m_1}{m_2},$$

where m_1 and m_2 are the first and second moments of the claim size distribution, θ is the premium loading, and M is an upper limit to the individual claim size.

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

7. Cumulative payments (in lakhs of rupees) made by an insurance company against claims from its motor insurance portfolio are given in the table below in each development year, for the accident years 2000-13, together with the premium earned (in lakhs of rupees) in each year.

Accident Year	Development Year				Premium earned
	0	1	2	3	
2010	49	102	189	256	507
2011	77	129	191		340
2012	83	135			442
2013	95				397

a. Estimate, by using the chain-ladder method, the reserve to be maintained at the end of 2013.

b. Use the Bornhuetter-Ferguson method to estimate the reserve at the end of 2013 assuming an ultimate loss ratio of 90%.

c. Compare the two estimates and comment on them.

[4+4+2=10]

8.

a. Consider the following ARIMA(p,d,q) model for a time series process, in which $\{\varepsilon_t\}$ is a sequence of uncorrelated random variables with zero means and common variance:

$$X_t = 1 + \frac{11}{6}X_{t-1} - X_{t-2} + \frac{1}{6}X_{t-3} + \varepsilon_t - 4\varepsilon_{t-1}.$$

i. Determine p, d, q.
ii. Is the process stationary? Justify your answer.
iii. What is the mean of X_t?

b. Consider the AR(2) process

$$X_t = -\beta X_{t-1} + \beta^2 X_{t-2} + Z_t,$$

where $\{Z_t\}$ is a zero-mean white noise process with $var(Z_t) = \sigma^2\ \forall t > 0$.

i. What is the range of values of β for which the above process can be stationary? Justify your answer.
ii. Determine the autocovariances γ_0, γ_1 and γ_2 in terms of β and σ^2.

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

9.

a. Residuals are plotted after fitting a ARIMA(2,1,2) model based on a sample of 100 observations from a time series. The graph has 43 turning points. Test independence of the residuals at the 5% level of significance.
b. Indicate clearly how the fitted model can be used for forecasting, using the Box-Jenkins approach.

[4+6=10]

10. Realizations of a random variable which has the probability density function

$$p(x) = \begin{cases} \dfrac{\sqrt{3}}{4\left(1 + {x^2}/{2}\right)^{3/2}} & \text{for } |x| < 2, \\ 0 & \text{otherwise,} \end{cases}$$

are required for a simulation-based study.

a. Show that the acceptance-rejection method based on the standard normal density $\phi(x)$ can be used for this purpose.

b. Assuming that samples from the uniform distribution over [0,1] can be readily generated, and the inverse of the standard normal distribution function is available, describe explicitly the steps required to generate a single sample from $p(x)$ by the method in part (a).

c. Deduce the expected proportion of samples from the standard normal distribution that will be rejected in the process of generating a sample from $p(x)$ by the method.

[4+4+2=10]

MSTAT II - Theory of Finance I

Final Exam. / Semester I 2014-15

Time - 3 hours/ Maximum Score - 50

Date - 28.11.2014

<u>NOTE</u> : **SHOW ALL YOUR WORK. RESULTS USED MUST BE CLEARLY STATED.**

1. Let $X_n = X_0 + R_1 + \cdots + R_n$ and $Y_n = Y_0 + S_1 + \cdots + S_n$, for all $n \geq 1$, where (R_k, S_k) are i.i.d. random variables with $P(R_k = 1, S_k = 1) = 3/10 = P(R_k = -1, S_k = -1)$ and $P(R_k = -1, S_k = 1) = 1/5 = P(R_k = 1, S_k = -1)$.

   (a) (3 marks) Assuming $X_0 \equiv 7 \equiv Y_0$, give brief reason that X_n and Y_n are martingale with respect to appropriate filtering $\{\mathcal{F}_n\}$ after describing this $\{\mathcal{F}_n\}$.

   (b) (5 marks) Find the compensators for $\{X_n^2\}$ and $\{X_nY_n\}$.

   (c) (5 marks) If the price of an asset on the nth day is X_n^2 Rupees, then find the expected time the price will take to hit either 16 or 121 Rupees.

2. Let the market consists of 3 risky assets and a riskless asset r_f. The mean return vector of the risky assets is $\mu = (0.5, 1, 0.8)'$, and the covariance matrix of their returns is

$$\Sigma = \begin{pmatrix} 5 & -2 & 1 \\ -2 & 8 & -3 \\ 1 & -3 & 6 \end{pmatrix}, \quad \text{with} \quad \Sigma^{-1} = \frac{1}{175}\begin{pmatrix} 39 & 9 & -2 \\ 9 & 29 & 13 \\ -2 & 13 & 36 \end{pmatrix}.$$

   (a) (5 marks) Find the the expression for the tangency portfolio in this market when $r_f = 0.6$.

   (b) (8 marks) Suppose you have some amount of wealth to invest in these assets. Can you calculate the best portfolio by giving the portfolio weights, that is chosen among these assets so that your mean return is 2? Explain the meaning of your weights (sell, buy etc for different assets). What is your beta in this case, corresponding to tangency portfolio?

3. A financial institution in India plans to hedge the (Crude) Oil prices for their customers, which is quoted in US dollar. Assume that the price of one unit of Oil at time t is S_t and in the risk-neutral world, it follows the model,

$$dS_t = \mu S_t dt + \sigma S_t dW_t,$$

where $\{W_t\}$ is a standard Brownian motion, $\mu = r_U$ is the risk-free interest rate in the US and σ_U is the constant volatility of the stock during the period considered. Since the company is India-based and so are their customers, they would like to get their profit in rupees only. Let P_t be price of a US dollar in rupees and its risk-neutral model,

$$dP_t = \mu_I P_t dt + \sigma_I P_t dB_t,$$

where $\{B_t\}$ is a standard Brownian motion independent of $\{W_t\}$. Here $\mu_0 = r_I - r_U =$ difference between the risk-free interest rate in India and the risk-free interest rate in the US, whereas σ_I is the constant volatility of the US$ (with respect to the rupees) during the period considered.

(a) (7 marks) Let $U_t = P_t S_t$ be the price of Oil in rupees. Use Itô's formula to show that U_t satisfies,

$$dU_t = \mu_1 U_t dt + \sigma_1 U_t d\hat{W}_t,$$

for some μ_1, σ_1 and a standard Brownian motion $\{\hat{W}_t\}$. Write down μ_1 and σ_1 and $\{\hat{W}_t\}$ in terms of the above parameters and standard Brownian motions $\{B_t\}$ and $\{W_t\}$, justifying your answer.

(b) (7 marks) A derivative on the oils stock pays off $Rs.500$ at maturity time T if $5000 < U_T \leq 6000$, and pays $Rs.1000$ if $U_T < 5000$ and pays nothing otherwise. Draw a diagram for the profit (against U_T). Use the risk-neutral valuation to calculate the price of the derivative at time 0, when $r_I = 9\%$, $r_U = 3\%$, $\sigma = 0.3$, $\sigma_I = 0.45$, $S_0 = 100$ US dollar, $P_0 = Rs.60$ and $T = 3$ months.

4. ($4 \times 5 = 20$) Write TRUE or FALSE and justify your answer (the justification should be Mathematical as much as possible).

(a) If the log price ($\log S$) follow a Standard Brownian motion then expected time for the log price to hit certain barrier $a < \log S_0$ is finite.

(b) ARCH(m) and IGARCH(1,1) both can capture clustering of volatility and both provide structure for positive long run variance.

(c) Suppose two portfolios have equal expected return. Then the portfolio with more volatility is preferable to the portfolio with less volatility, whether the expected return is positive or negative.

(d) Suppose the market consists of N assets (of which at most one riskless asset) whose possible prices in the next unit of time is given by a matrix $D_{N \times M}$ with $M = N$. If D is nonsingular then there exists a unique martingale measure and the market is complete.

All the best.

INDIAN STATISTICAL INSTITUTE

Semestral Examination: (2014–2015)

M. Stat 2nd Year

Statistical Computing

Date:28.11.14 Marks: ..100.. Duration: .3 hours.

Attempt all questions

1. (a) The standard linear regression model can be written in matrix notation as $\mathbf{X} = \mathbf{A}\boldsymbol{\beta} + \mathbf{U}$. Here $\mathbf{X}$ is the $r \times 1$ vector of dependent variables, $\mathbf{A}$ is the $r \times s$ design matrix, $\boldsymbol{\beta}$ is the $s \times 1$ vector of regression coefficients, and $\mathbf{U}$ is the $r \times 1$ normally distributed error vector with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}$. For each $i = 1, \ldots, r$, suppose that there is a constant c_i such that $Y_i = \min\{c_i, X_i\}$ is observed, so that the dependent variables are right censored. Derive an EM algorithm for estimating the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$ in the presence of right censoring.

   (b) Develop an MM algorithm for minimizing the function

$$f(x_1, x_2) = \frac{1}{x_1^3} + \frac{3}{x_1 x_2^2} + x_1 x_2,$$

   where $x_1, x_2 > 0$.

Marks: 10+15=25

2. (a) Discuss using relevant theorems (with proofs) how the Fast Fourier Transform reduces the computational burden involved in the Discrete Fourier Transform.

   (b) Suppose that $f(x)$ is a twice continuously differentiable function and $s(x)$ is the spline interpolating $f(x)$ at the nodes $x_0 < x_1 < \cdots < x_n$. If $h = \max\limits_{0 \le i \le n-1} (x_{i+1} - x_i)$, then prove that

$$\max_{x_0 \le x \le x_n} |f(x) - s(x)| \le h^{\frac{3}{2}} \left[\int_{x_0}^{x_n} \{f''(y)\}^2 dy \right]^2, \text{ and}$$

$$\max_{x_0 \le x \le x_n} |f'(x) - s'(x)| \le h^{\frac{1}{2}} \left[\int_{x_0}^{x_n} \{f''(y)\}^2 dy \right]^2.$$

Marks: 10+15=25

3. (a) Suppose that it is necessary to draw samples from a density f, for which no standard simulation techniques exist. Additionally, evaluation of f at any point x is also computationally burdensome. Assume however, that there exists a density g_{upper}, a function g_{lower}, and a positive constant M, such that, for all x

$$g_{lower}(x) \leq f(x) \leq M g_{upper}(x).$$

and that there exist standard methods to simulate from g_{upper}. Construct an efficient accept-reject algorithm for simulating from f and prove its validity.

(b) Let $X \sim f$, and let $X_1, \ldots, X_N \sim g$, g being an appropriate density. Now consider the importance sampling estimator

$$\hat{E}_f[h(X)] = \frac{1}{N}\sum_{i=1}^{N} \frac{f(X_i)}{g(X_i)} h(X_i)$$

of the quantity

$$E_f[h(X)] = \int h(x) f(x) dx.$$

Show that, the choice of g that minimises the variance of the estimator $\hat{E}_f[h(X)]$ is

$$g(x) \propto |h(x)| f(x).$$

(*Hint: Use Jensen's inequality*).

Marks: 10+15=25

4. Assume that the Markov transition kernel $P(x, A)$ satisfies a minorization condition on a small set C, in addition to the regularity conditions necessary for convergence to the target distribution π. Assume that two different chains are run, one started from an arbitrary fixed starting value x_0, and another started by drawing the initial value y_0 from the invariant distribution.

(a) Derive the coupling inequality

$$\| P^n(x_0, \cdot) - \pi(\cdot) \| = \sup_{A \in \mathcal{B}} |P^n(x_0, A) - \pi(A)| \leq Pr(T > n),$$

where $P^n(x_0, A)$ is the probability of hitting the set A in n iterations, beginning at x_0, T is the time at which coupling occurs and $\mathcal{B}$ is the appropriate Borel σ-algebra.

(b) How would you modify the above coupling inequality if the small set is the entire state space?

(c) Consider an independent Metropolis-Hastings sampler (one dimensional) with target density $\pi(x)$ and proposal density $p(x)$. Further suppose that both the densities are continuous and strictly positive on the entire state space and satisfies

$$\frac{\pi(x)}{p(x)} \leq M \text{ for all } x, \text{ for some } M > 0.$$

Then show that, for any starting value x_0,

$$\| P^n(x_0, \cdot) - \pi(\cdot) \| = \left(1 - \frac{1}{M}\right)^n.$$

Marks: 10+5+10=25

INDIAN STATISTICAL INSTITUTE
Semester Examination
M. Stat II year and Research Course: 2014–2015
Graph Theory and Combinatorics

Date: 12. 12. 2014 (11 AM to 2 PM) Marks: 80 Time: 3 Hours

Please try to write all the part answers of a question at the same place.

1. (a) What is graph isomorphism?

   (b) Prove that the graphs G and H are isomorphic if and only if $\overline{G}$ and $\overline{H}$ are isomorphic.

   (c) What is the worst-case time complexity of a naïve algorithm for testing whether two given graphs of n vertices are isomorphic or not?

   [2 + 6 + 4 = 12]

2. (a) Let G be a k-regular bipartite graph with $k \geq 2$. Show that G has no cut edges.

   (b) Is 3,3,3,3,3,2,2,1 a graphic sequence?

   (c) What is the time complexity of calculating the number of spanning trees of an arbitrary graph by edge contraction algorithm? Is there an alternative method of computing it more efficiently?

   [4 + 4 + (4 + 4) = 16]

3. (a) What is the relation between the maximum size of matching and minimum size of edge-cover for a graph with no isolated vertices?

   (b) Show that out of all possible "stable" marriages, every man is happiest under the Gale-Shapley male-proposal algorithm.

   (c) Why does the augmenting path algorithm for finding the maximum matching in a bipartite graph does not extend to general graphs?

   [2 + 10 + 6 = 18]

4. (a) Prove that both edge-coloring and face-coloring can be reduced to vertex coloring.

   (b) Is k-coloring in Planar graphs NP-complete for $k \geq 3$?

   (c) What is the chromatic polynomial for a tree?

   (d) How the principle of inclusion-exclusion can be used to find the chromatic polynomial of a general graph?

   [(3+3) + 4 + 4 + 12 = 26]

5. In Erdős-Réyni random graph model,

   (a) what is the expected degree of each node?

   (b) what is the probability of a vertex being isolated?

   [4 + 4 = 8]

Indian Statistical Institute

First Semester Exam: 2014-15

M.Stat. II Year, Statistical Methods in Genetics I (Back paper)

Max. Marks: 100, Time: 3 Hours

29|12|2014

1. State and prove Hardy-Weinberg theorem for a single gene with three alleles A, B and C. **2+8**

2. Show that the transition probability matrix for grandparent-grandchild relationship is T^2. Hence show that $T^n \to 0$ as $n \to \infty$, implying that the relationship disappears after many generations. **5+5**

3. (a) Consider two genes with alleles A, a; and B, b respectively. If r is the recombination fraction between the genes, give the joint probability distribution of different genotypes for F_2 design. (You need to provide the 3×3 table with the cell probabilities). **5**
(b) Define the map distance (d) between two genes. Show that $d = -\frac{1}{2}log(1-2r)$, using Poisson process. **5**

4. (a) Consider a sex-linked gene with alleles A and a with respective frequency p and q. Assuming equilibrium, give the transition probability matrix for mother-daughter relation and father-son relation. **5**
(b) What do you mean by joint analysis of longitudinal trait and time to an event? Assuming a Cox PH model, write down the joint likelihood function explicitly. **5**

5. (a) What does GWAS stand for? What is fGWAS? Why is fGWAS more efficient than classical GWAS? **5**
(b) Why multiple testing is so crucial in GWAS? Clearly state BH algorithm for controlling false discovery rate in multiple testing. **5**

6. Consider Interval Mapping problem for an F_2 design. Considering all the marker intervals and assuming that QTL is in the marker interval, give the joint marker-QTL genotype distribution. (You need to provide the 9×3 table with the cell probabilities). **20**

7. Define "broad sense heritability" and give appropriate method for its estimation. What is the importance of "broad sense heritability" in the context of GWAS? **5+5+5=15**

8. Consider interval mapping for Backcross design. There are two markers M_1 and M_2. The recombination fractions between two markers, between M_1 and QTL , between

M_2 and QTL are r, r_1, r_2 respectively. If the two markers are highly linked, show that r is approximately equal to $r_1 + r_2$. **15**

INDIAN STATISTICAL INSTITUTE

First Semestral (Backpaper) Examination: 2014-15
M. STAT. II YEAR

ACTUARIAL METHODS

Date: 30|12|2014 **Maximum Marks: 100** **Duration: 3 hr**

Note: ***Actuarial Tables and calculators can be used to solve the problems.***

1. A health insurance company has option of three products to sell in the market and must decide which product to sell in the coming year. There are three possible choices: *basic*, *intermediate* or *advanced* product, each with different related costs based on the complexity of the product. The manufacturer has fixed overheads of Rs. 1,500,000.
The revenue and cost for each product are as follows.

Policy	Cost	Revenue per policy
Basic	500,000	1500
Intermediate	600,000	1000
Advanced	1,000,000	2000

The insurer had sold 2,100 policies last year, and is preparing forecasts of profitability for the coming year based on three scenarios: *Low sales* (80% of last year's level), *Medium sales* (same as last year's level) and *High sales* (20% higher than last year's level).

a. Determine the annual profit in rupees for each product under each scenario.
b. Determine the minimax solution to the choice problem.
c. Determine the Bayes solution, if there is 20% chance of *Low sales*, 60% chance of *Medium sales* and 20% chance of *High sales*.

[3+3+4=10]

2. The number of claims (*N*) per year from a certain insurance policy has a negative binomial distribution with probability mass function

$$P(N=n)=\binom{k+n-1}{n}p^k(1-p)^n,\ n=0,1,2,\ldots$$

where $k=3$ and $p=0.6$. It is known from past experience that in a given year, 70% of the claims arising out of such policies are for Rs. 10,000, 25% are for Rs. 20,000 and 5% of the claims are for Rs. 50,000

a. Determine the cumulant-generating function of the annual aggregate claims (*S*) under this policy.
b. Hence deduce the expected value and the variance of *S*.
c. In case of excess-of-loss reinsurance with a retention level of Rs. 15,000, deduce the expected aggregate claims paid by the insurer and the reinsurer.

[2+4+4=10]

3. An insurance company, starting with an initial surplus of Rs. 10 lakhs for a particular type insurance policy, sells 100 policies at the beginning of a given year in respect of identical risks, an annual premium of Rs. 5000 each. Aggregate claims arising out of such a policy in a particu year are assumed to have a compound Poisson distribution with parameter λ=5, and individual cla sizes to be uniformly distributed over the interval [0,20000]. Suppose that 150 more such policies sold at the beginning of the second year, and the earlier policies continue to remain in for Ignoring administrative expenses and interest earned on the surplus,

   a. specify the surplus process under this setup, for the first two years;
   b. define ruin probabilities in finite and infinite time;
   c. calculate the probability that the company will be insolvent in respect of this type of policy
      i. at the end of the first year,
      ii. at the end of the second year.
   d. If the initial surplus is reduced by 10%, by what percentage will the probability of ruin in th first year change?

[2+2+4+2=1

4. Aggregate claims under a policy occur according to a compound Poisson process with parameter and individual claim sizes are independently distributed in the exponential form with parameter λ.

   a. Write down the equation for the adjustment coefficient and prove that it has a uniqu positive root.
   b. Determine the adjustment coefficient, assuming a premium loading factor of 20%, taking μ=7 and λ=15.
   c. If the premium loading factor is increased, how would you expect the value of the adjustmen coefficient to change? Justify your answer.
   d. How is the infinite-time ruin probability expected to be affected by an increase in th premium loading factor, and why?

[4+2+2+2=10

5.

   a. The number of claims in a given year for a certain risk is assumed to have a Poisso distribution with parameter μ. μ in turn is assumed to have a gamma distribution wit parameters 12 and 3. Suppose $x_1, x_2, \ldots, x_n$ are observed values of the number of claims in different years in the past.
      i. Derive the posterior distribution of μ.
      ii. What is the Bayesian estimate of μ under quadratic loss?
      iii. Show that the Bayesian estimate of μ is a credibility estimate for the expecte number of claims per year, and give the value of the credibility factor.

[3+1+2=6]

b. The table below shows aggregate annual claim statistics for four different general insurance products over a period of five years. Annual aggregate claims for product i in year j are denoted by $X_{ij}, i = 1,2,3,4; j = 1,2,\cdots,5$.

Product (i)	$\bar{X}_i = \frac{1}{5}\sum_{j=1}^{5} X_{ij}$	$\frac{1}{4}\sum_{j=1}^{5}(X_{ij} - \bar{X}_i)^2$
1	125	300
2	85	60
3	140	35
4	175	100

Calculate the credibility premium of the first three products under EBCT model 1.

[6]

6. The table below gives the cumulative payments (in lakhs of rupees) made by an insurance company against claims from its motor insurance portfolio, in each development year, for the accident years 2004-07.

Accident	**Development Year**			
Year	**0**	**1**	**2**	**3**
2004	49	102	189	256
2005	77	129	191	
2006	83	135		
2007	95			

a. Use the basic chain-ladder method to estimate the ultimate outstanding claims at the end of 2007.

b. If the rate of claims inflation, measured over the 12-month period to the middle of the year, increased by 2% each year, in the period 2004-2007 and is expected to show the same behavior in the next few years, obtain the inflation-adjusted estimate of the outstanding claims at the end of 2007.

[4+8=12]

7.

a. Determine ρ_k, the autocorrelation at lag k(>0), of the stationary ARMA(1,1) time series process

$$Y_n = aY_{n-1} + \varepsilon_n + b\,\varepsilon_{n-1},$$

where $|a| < 1$, and ε_n denotes the white noise process with variance σ^2.

b. Hence deduce the autocorrelation functions of the AR(1) time series

$$Y_n = aY_{n-1} + \varepsilon_n,$$

and the MA(1) time series

$$Y_n = \varepsilon_n + b\varepsilon_{n-1}.$$

c. When is an ARMA(p,q) process said to be invertible?

[6+4+2=12]

8.

a. Describe the Box-Muller method for generating a pair of independent realizations from th standard normal distribution, providing appropriate justification.

b. Explain how you can generate realizations of the central χ^2 random variable with 24 degre of freedom, using a generator for uniform random variates.

c. Describe any method for generating realizations of a random variable X with probabili density function

$$p(x) = \alpha\beta \frac{x^{\alpha\beta-1}}{(1+x^{\alpha})^{\beta+1}},\ x > 0.$$

d. Describe the acceptance sampling method for simulating from a probability density $p_X(x)$ a show that its output is indeed a realization from $p_X(x)$.

[4+2+2+4=1

9.

a. A generalized linear model (GLM) has independent responses $Y_1, Y_2, \ldots, Y_n$ with probabili density function

$$f(y;\beta,\mu) = \frac{\beta^{\beta}}{\mu^{\beta}\,\Gamma(\beta)}\, y^{\beta-1}\, e^{-\beta y/\mu},\quad y > 0.$$

i. Show that Y_i belongs to the exponential family of random variables.

ii. Hence determine the expectation and variance for this distribution.

iii. Identify the natural parameter and hence the canonical link function for the GLM.

[2+2+2=

b. Let $Y_1, Y_2, \ldots, Y_n$ be independent Poisson random variables with parameters μ_i, where

$$\log \mu_i = \begin{cases} \alpha & i = 1,2,\ldots,k \\ \beta & i = k+1,\ldots,m \\ \gamma & i = m+1,\ldots,n \end{cases}$$

If $k = 6, m = 10, n = 15$, and for observations $y_1, y_2, \ldots, y_n$ on the Y_i's,

$$\sum_{i=1}^{k} y_i = 37,\ \sum_{i=k+1}^{m} y_i = 32,\ \sum_{i=m+1}^{n} y_i = 41,$$

i. compute the maximum likelihood estimate of α and β, given that $\gamma = \beta$;

ii. compare the adequacy of the model with $\beta = \gamma$, relative to the one given above.

[2+4=6

Indian Statistical Institute
Analysis of Discrete Data
M-II, Backpaper Examination

Date: 31·12·2014 Duration: 3hrs.

Attempt all questions. The maximum you can score is 45. Justify all your steps. This is a closed book examination. You may use your own calculator.

If copying is detected in the solution for any problem, all the students involved in the copying will get 0 for that problem. Also an additional penalty of 5 will be subtracted from the overall aggregate of each of these students.

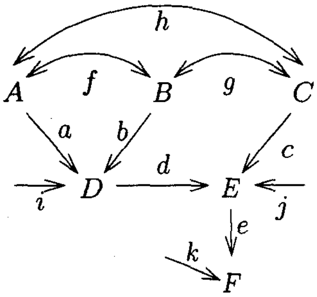
1. What do you understand by tetrachoric correlation. Suppose that $(X_1, Y_1), ..., (X_n, Y_n)$ are iid from a bivariate normal population with $N(0,1)$ marginals and correlation ρ. Let $U_i = sign(X_i)$ and $V_i = sign(Y_i)$. Suggest how you may estimate ρ based on only $(U_1, V_1), ..., (U_n, V_n)$. [3+7]

2. Derive a logistic model for ordinal multicategory response Y and a single continuous explanatory variable x such that for any given response category i and any two given values x_1, x_2 of X

$$\frac{P(Y \leq i|X = x_1)/P(Y > i|X = x_1)}{P(Y \leq i|X = x_2)/P(Y > i|X = x_2)} \propto x_1 - x_2.$$

[15]

3. Consider the following path model where the labels and arrows have the usual meanings.



Find the correlation between the nodes C and D and also between C and E. [5+10]

4. Define the odds ratio parameter for a 2 × 2 contingency table under the multinomial sampling scheme let the probability of the (i,j)-th cell be π_{ij}. Now write down which of the following two statements is/are true/false:

   (a) $\frac{n_{11}}{n_{11}+n_{12}}$ is the MLE of $P(C = 1|R = 1)$ under both the multinomial sampling scheme.

   (b) $\frac{n_{11}}{n_{11}+n_{12}}$ is the MLE of $P(C = 1|R = 1)$ under the indeoendent binomial sampling scheme with fixed row marginals.

   (c) $\frac{n_{11}}{n_{11}+n_{12}}$ is the MLE of $P(C = 1|R = 1)$ under the indeoendent binomial sampling scheme with fixed column marginals.

   (d) Sample odds ratio is the MLE of population odds ratio under the independent binomial sampling scheme with fixed column marginals.

   (e) Sample odds ratio is the MLE of population odds ratio under the multinomial sampling scheme.

[5+5+5+5+5]

5. Suppose that a 2 × 2 × 2 contingency table comes from a loglinear model

$$E(n_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ}.$$

It is given that the odds ratio between X and Y for each particular value of Z is free of the value of Z. Show that $\lambda_{ijk}^{XYZ} = 0$ for all i, j, k. [15]

6. What is meant by Simpson's paradox for a $2 \times 2 \times 2$ contingency table? Give one *real life* example where you expect to see Simpson's paradox. [5+5]

7. For the following contingency table find the γ coefficient.

	< primary	Primary	Secondary	HS	≥ College
Low income	3	6	23	12	2
Middle income	0	0	2	43	128
High income	0	0	0	0	34

A statistician wants to test the presence of positive association between income level and education level using γ. But with so many empty cells she is not comfortable about using asymptotics. She wants to perform a simulation test instead. Could you please help her with this. No need to write any R code. Just write down the steps in plain English assuming that you can draw random numbers from any standard distribution. [10]

INDIAN STATISTICAL INSTITUTE

203 B. T. Road, Kolkata 700 108

Master of Statistics (M.Stat.) IInd Year
Advanced Probability I

Academic Year 2014 - 2015: Semester I

Back Paper Examination

Date: January 01, 2015 **Total Points: 100**
Time: 02:30 PM - 06:30 PM **Duration: 4 Hours**

Note:

- **Please write your roll number on top of your answer paper and DO NOT write your name.**
- **There are six problems each carrying 20 points. Solve any five of them. If you solve all the six problems and do not indicate which ones to be graded, then only the first five will be graded. Maximum you can score is 45. Show all your works and write explanations when needed.**
- **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). However, note that no printed materials or photo copies are allowed, in particular you are not allowed to use books, photocopied class notes etc.**

1. Suppose $X_1, X_2, \ldots, X_n, \ldots$ is an *exchangeable* sequence of random variables with finite second moments defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$.
   (a) Show that $\text{Corr}\,(X_i, X_j) \geq 0$ for any $i, j \geq 1$. [5]
   (b) Suppose that all the finite dimensional distributions are Gaussian. Then show that $\mathcal{T} = \mathcal{E}$ a.s., where $\mathcal{T}$ and $\mathcal{E}$ are respectively the *tail* and the *exchangeable* σ-algebras. [15]

2. Consider the function $\phi(x) = 1/x$ from $(0, \infty)$ onto $(0, \infty)$. Let λ be the Lebesgue measure on the Borel σ-algebra $\mathcal{B}_{(0,\infty)}$ on $(0, \infty)$ and let $\mu = \lambda \circ \phi^{-1}$. Show that $\mu \ll \lambda$ and find $\frac{d\mu}{d\lambda}$. [10 + 10 = 20]

3. Suppose $(X_n, \mathcal{F}_n)_{n \geq 0}$ is bounded increment martingale sequence. Let $a_n \uparrow \infty$ be such that $\sum\limits_{n=1}^{\infty} \frac{1}{a_n^2} < \infty$, then show that $\frac{X_n}{a_n} \longrightarrow 0$ a.s. [20]

4. Let $(X_n, \mathcal{F}_n)_{n \geq 0}$ be a square integrable martingale. For $n \geq 1$, define $\sigma_n^2 := \mathbf{E}\left[(X_n - X_{n-1})^2 \,\middle|\, \mathcal{F}_{n-1}\right]$. Suppose there is a sequence of positive real numbers $\{s_n^2\}_{n \geq 1}$ and some $\delta > 0$ such that

$$\frac{1}{s_n^2} \sum_{k=1}^{n} \sigma_k^2 \xrightarrow{\mathbf{P}} 1 \quad \text{and} \quad \frac{1}{s_n^{2+\delta}} \sum_{k=1}^{n} \mathbf{E}\left[(X_k - X_{k-1})^{2+\delta}\right] \to 0.$$

Then show that

$$\frac{X_n}{s_n} \xrightarrow{d} \text{Normal}\,(0,1)\ .$$

[20]

[Please Turn Over]

5. Suppose $(\xi_i)_{i\geq 1}$ are i.i.d. taking values in the set $\{-1, 0, 1, 2, 3, \ldots\}$ and $\mu := \mathbf{E}[\xi_1] < 0$. Let $S_n := \xi_1 + \xi_2 + \cdots + \xi_n$ for $n \geq 1$ and $S_0 = 0$. For any $z \in \mathbb{Z}$ define $T_z := \inf\left\{n \geq 0 \,\middle|\, S_n = z\right\}$.

   (a) Show that $T_{-a} < \infty$ a.s. for any $a \in \mathbb{N}$. [5]

   (b) Determine whether $\mathbf{E}[T_{-1}] < \infty$ and $\mathbf{Var}(T_{-1})$ exists. [5 + 10 = 15]

6. Consider the measurable space $\left(\{0,1\}^{\mathbb{N}_0} \times \{0,1\}^{\mathbb{N}_0}, \mathcal{B}_{\{0,1\}^{\mathbb{N}_0}\times\{0,1\}^{\mathbb{N}_0}}\right)$ where $\mathbb{N}_0 := \{0, 1, 2, \cdots\}$. Fix $0 < p < 1$.

   (a) Show that there exists a unique probability $\mathbf{P}$ on the above space such that the coordinate variables $\left((\xi_i)_{i\geq 0}, (X_i)_{i\geq 0}\right)$ satisfy the following:

      (i) $(\xi_i)_{i\geq 0}$ are i.i.d. Bernoulli (p);

      (ii) $X_i \sim$ Bernoulli $\left(\frac{1}{2}\right)$ for all $i \geq 0$;

      (iii) $X_i = X_{i+1} + \xi_i \pmod 2$ for all $i \geq 0$ a.s.;

      (iv) $(\xi_i)_{1\leq i\leq n-1}$ is independent of $(X_i)_{i\geq n}$.

   [5]

   (b) Let $\mathcal{F}_n := \sigma(\xi_0, \xi_1, \cdots, \xi_n)$ for $n \geq 0$ and $\mathcal{F} := \sigma\left((\xi)_{i\geq 0}\right)$. Compute $\mathbf{E}\left[X_0 \,\middle|\, \mathcal{F}_n\right]$, and conclude that X_0 is not $\mathcal{F}$-measurable. [3 + 1 = 4]

   (c) Let $\mathcal{T}_n := \sigma(X_n, X_{n+1}, \cdots)$ for $n \geq 0$. Compute $\mathbf{E}\left[X_0 \,\middle|\, \mathcal{T}_n\right]$. [4]

   (d) If $\mathcal{T}$ is the tail σ-algebra of $(X_n)_{n\geq 0}$ then show that X_0 is independent of $\mathcal{T}$. [4]

   (e) Conclude that $\mathcal{T}$ is a trivial σ-algebra. [3]

Good Luck

INDIAN STATISTICAL INSTITUTE

Back-paper Examination, First Semester: 2014-15

M.Stat. II Year (AS)

Life Contingencies

Date: 02·01·2015 Maximum marks: 100 Duration: 3 hours

Students are permitted to use non-programmable calculators and Actuarial Formulae and Tables.

1. If the force of interest is twice the force of mortality, and both rates are constant, calculate the standard deviation of the present value of a continuous whole life annuity, assuming that its expected value is 12.5. [4]

2. A select table (select period 5 years) is based on the following rates of mortality: $q_{[x]+t} = \frac{0.02}{1.02}$ for all values of x and values of $t < 5$; $q_x = \frac{0.03}{1.03}$. If $l_{[10]} = 100,000$, calculate (a) l_{45}; and (b) $l_{[40]+1}$. [2+1=3]

3. In a situation where the force of interest is a constant (δ) per year, it is claimed that one rupee today can be used to pay a continuous whole life annuity to (x) at the rate Rs. δ per year, as well as to provide a whole life assurance to x with benefit of Re. 1. Express this statement in standard actuarial notation, and prove or disprove it from first principles. [5]

4. One thousand independent lives aged 25 are offered a 20-year deferred 15-year endowment assurance of Rs. 10 lakhs payable at the end of the year of death or at maturity. The premium is to be paid annually in advance for 20 years. The lives follow AM92 Select mortality and the interest rate is 6% per annum.

   (a) Using the equivalence principle, calculate the net premium provision per policy that is required at the end of the premium-paying period. [2]

   (b) Using the equivalence principle, calculate the net premium provision per policy that is required at the end of the first year of the benefit period. [2]

   (c) Calculate the death strain at risk, per policy in force at the beginning of the benefit period, during the first year of benefit period. [1]

   (d) At the end of 20 years, 976 policies were in force. There were 6 deaths in the following year. Calculate the mortality profit. [2]

   (e) Comment on the mortality profit or loss. [2]

   [Total 9]

5. An annuity-due makes quarterly payments to a life aged 60 exact where each payment is 1.0094797 times greater than the one immediately preceding. The first quarterly amount is Rs. 1,000. Calculate the actuarial present value of the annuity.

   Basis: Mortality: AM92 ultimate with UDD; Interest: 8% per annum. [5]

6. If $P_x = .017$, $i = 0.03$, $q_x = 0.02$, $q_{x+1} = 0.022$ and $q_{x+2} = 0.024$, calculate ${}_3V_x$. [4]

7. The future lifetimes of two individuals aged x and y are independent, and are subject to constant forces of mortality of 0.02 and 0.03 respectively. Assuming interest rate $i = 3\%$, find the actuarial present value of a whole life annuity-due of Rs. 100,000 per year, payable monthly to the last survivor, commencing at the end of the month of termination of the joint life. [7]

8. A male aged x is due to retire in n years. A benefit scheme provides for a sum of Rs. 100,000 payable to his wife aged y at the end of the year of his death, provided that she is alive at that time. Further, if he dies in service and his wife is alive at that time, she will receive a reversionary whole life annuity of Rs. 30,000 per year payable annually in arrear. Derive a concise expression for the net present value of this benefit, using appropriate actuarial notations. [6]

9. An impaired life aged 35 experiences 5 times the force of mortality of a life of the same age subject to standard mortality. A two year term assurance policy is sold to this impaired life. The policy has a sum assured of Rs. 2,50,000 payable at the end of the year of death.

   (a) Show that the probability of survival of an impaired life aged x for a further year is p_x^5, where p_x is the corresponding probability of a non-impaired life. [2]

   (b) Calculate the expected present value of the benefits payable to each life under the above policy, assuming that standard mortality is AM92 Ultimate and interest is 4% pa. [4]

[Total 6]

10. Assume that m single-decrement tables are given, each representing a different cause of decrement.

   (a) Explain in detail how to obtain a multiple-decrement table embodying these m causes of decrement by proving the result $\mu_x^{(T)} = \sum_{k=1}^{m} \mu_x^{(k)}$. [7]

   (b) Consider a special case of two decrements to prove the following:

$$d_x^{(k)} = \frac{q_x^{(k)}}{q_x^{(1)} + q_x^{(2)}} . d_x^{(T)}, \quad k = 1, 2. \qquad [3]$$

[Total 10]

11. A life insurance company sells 5-year-term, single-premium, unit-linked bonds, each for a single premium of Rs. 15,000. There is no bid/offer spread and the allocation percentage is 100%.

   (a) Assuming that the only charge is a 3% annual management charge and assuming unit growth of 9% pa, calculate the unit provision at the start and the end of each year and the management charge each year. [3]

   (b) Calculate the net present value of the contract assuming:

   - Commission of 5% of the premium,
   - Initial expenses of Rs. 150,
   - Annual renewal expenses of Rs. 50 in the 1st year, inflating at 5% pa,
   - Independent rate of mortality is a constant 0.5%,
   - Independent rate of surrender is 5% pa,
   - Non-unit fund interest rate is 9% pa,
   - Risk discount rate 12% pa.

   The company holds unit provisions equal to the full value of the units and zero non-unit provisions. You may assume that expenses are incurred at the start of the year and that death and surrender payments are made at the end of the year. [4]

[Total 7]

12. The death in service benefit of a pension scheme is four times the member's salary on the date of death. Normal Pension Age is 60. Give an expression for the present value of this benefit to a life aged 35 exact with salary of Rs. 25,000, who has just received a salary increase. Define all the symbols used, and do not use commutation functions.

[5]

13. A 22-year endowment policy has to be issued to a life aged 33 exact, with mortality following the AM92 Ultimate table. The benefit amount B is to be paid at the end of the year of death or at maturity. Level premiums of amount G are paid annually in advance during the 22 years of the contract. The interest rate is 4% per annum.

The insurer incurs the following fixed and variable expenses.

Time of expense	Expense
Policy issue	0.2% of B plus 40% of G
Payment of subsequent premium	0.02% of B plus 2% of G
Payment of death benefit	0.03% of B
Payment of maturity benefit	0.02% of B

Using the equivalence principle, calculate the gross premium G as a multiple of the benefit amount B. [7]

14. Consider the three-state continuous-time Illness-Death Markov model, with states *Healthy*, *Critically ill* and *Dead*, with the following constant forces of transition:

(i) σ for transition from *Healthy* to *Critically ill*,
(ii) μ for transition from *Healthy* to *Dead* and
(iii) ν for transition from *Critically ill* to *Dead*.

No transition from *Critically ill* to *Healthy* is possible. A life insurance company uses this model to price its stand-alone critical illness policies. Under these policies, a lump sum benefit is payable on the occasion that a life becomes critically ill during a specified policy term. No other benefits are payable.

A 20-year policy with sum assured Rs. 200,000 is issued to a healthy life aged 40 exact.

(a) Write down a formula, in integral terms, for the expected present value of benefits under this policy. [2]

(b) Calculate the expected present value at the outset for this policy.
Basis: $\mu = 0.01$, $\sigma = 0.02$, $\nu = 3\mu$, Interest is 8% pa. [3]

[Total 5]

15. A ten-year unit-linked policy is issued to a life aged 18. Assuming no non-unit reserves, the year-end cash-flows turn out to be as follows.

705, −293, 433, 497, 101, 297, −350, 127, 256, 100.

The company wants to hold the minimum non-unit reserves at the end of policy years 1,2,. . .,9, which are needed to make the cash-flows in the following years non-negative. Using the mortality law ${}_tp_x = \frac{x^2}{(x+t)^2}$ and interest rate 5% per annum, calculate the sequence of reserves needed. [8]

16. A 20-year endowment policy for a life aged 40 has a sum assured of Rs. 5 lakhs. There is a bonus of 10% of original sum assured and an additional bonus of 1.92308% on all previously announced bonuses, both of which vest at the end of the year. Net premium is collected annually in advance for the duration of the policy.

(a) Calculate the net premium. [6]

(b) At the end of 10 years, and after bonus have vested as per the above plan, the insurer wishes to convert the with-profit contract into that of a fixed benefit contract, without changing the rate of premium. Calculate the benefit amount that would be equivalent to the benefit committed under the existing contract.

[3]

Basis: Mortality: AM92 Ultimate; Interest: 6%; Expenses: Ignore.

[Total 9]

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination, Second Semester: 2014-15

M.Stat. II Year (AS)

Survival Analysis

Date: March 4, 2015 Maximum marks: 60 Duration: 3 hours

Answer any combination of questions worth 60 marks. The entire paper is worth 70 marks.

1. Show that the mean residual life of a random variable is a decreasing function of age if its hazard rate is increasing. Also show that whenever the mean residual life is constant, the hazard rate is also constant. [5+5=10]

2. An electronic system is at continuous risk of failure with a constant hazard of λ events per hour. In addition, power surges occur each hour (i.e., at times 1,2,...) and at each power surge there is a 10% chance that the system will fail immediately. Obtain expressions for the cumulative hazard function and the survival function of the system. Find the mean time to system failure. [2+2+2=6]

3. Consider type-I censored data from the exponential distribution with an unknown parameter, the censoring times being identical. Obtain an expression for the maximum likelihood estimator (MLE) of the survival probability at 1. Also obtain an expression for the asymptotic variance of the estimator. $[3 + 5 = 8]$

4. Describe the notion of generalized maximum likelihood estimator (GMLE) proposed by Kiefer and Wolfowitz. Show that this notion gives rise to the usual MLE in a parametric set-up. Describe (without deriving) the GMLE of a distribution obtained from randomly right censored samples from that distribution. $[2 + 5 + 2 = 9]$

5. Consider a type I censored sample from the exponential distribution with censoring time c common to all n individuals on test. Show that the total number of failures, D, has a binomial distribution with parameters n and $p = 1 - \exp(-\lambda c)$. Compare the asymptotic efficiency of the MLE of λ from the marginal distribution of D to that of (V, D), where V is the total time on test. Under what condition on c is it reasonable to base inferences about λ on D alone? [2+6+2=10]

6. Formulate the problem of testing equality of distributions, using randomly right-censored samples from each of them, into a suitable problem that can be handled through the theory of counting processes (i.e., describe the type of data and redefine the hypothesis under the notations of counting processes). Describe a suitable test based on the Nelson-Aalen estimator, along with an expression for an estimator of its asymptotic variance. Discuss the role of a weight function in this context. [3+3+3+3=12]

7. Explain why the product-limit estimator is called a self-consistent estimator. Also describe the redistribute-to-the-right algorithm of calculating this estimator [3 + 3 = 6]

8. A researcher is reviewing a study, published in a medical journal, into survival after a major operation. The journal only gives the following summary information: (i) the study followed 16 patients from the point of surgery; (ii) the patients were studied until the earliest of 'five years after operation', 'end of the study' and 'withdrawal of the patient from the study'; (iii) the Nelson-Aalen estimate $\hat{S}(t)$ of the survival function was

$$\hat{S}(t) = \begin{cases} 1 & \text{for } 0 \leq t < 1, \\ 0.9355 & \text{for } 1 \leq t < 3, \\ 0.7122 & \text{for } 3 \leq t < 4, \\ 0.6285 & \text{for } 4 \leq t < 5. \end{cases}$$

   (a) Describe the types of censoring which are present in the study.

   (b) Calculate the number of deaths which occurred, at each distinct duration since the operation.

   (c) Calculate the number of patients who were censored.

[2+6+1=9]

INDIAN STATISTICAL INSTITUTE

Back Paper Examination : Semester I (2014-2015)

M. Stat 2nd Year

Pattern Recognition and Image Processing

Date 12|01|2015 Maximum marks: 100 Time: 3 hours.

Note: Attempt all questions. Maximum you can score is 100. Answer Group A and Group B questions in separate answerscripts.

Group A

1. Consider a two-class classification problem in which each component $X_j, j = 1, \dots, M$ of the feature vector $\mathbf{X}$ is either 0 or 1 with conditional probabilities

$$p_j = P\{X_j = 1 \mid \text{Class 1}\}$$

$$q_j = P\{X_j = 1 \mid \text{Class 2}\}$$

Assume that the components of $\mathbf{X}$ are independent and the prior class probabilities as well as the misclassification costs are equal.

(a) Find the Bayes classification rule for the problem. [5]

(b) If, further $p_j = p > \frac{1}{2}$ and $q_j = 1 - p, j = 1, \dots, M$, show that for M odd, the Bayes error probability is given by [5]

$$P_e(M, p) = \sum_{l=0}^{(M-1)/2} \binom{M}{l} p^l (1-p)^{M-l}.$$

2. Let $f(x)$ be uniform from 0 to a and let $K(x) = e^{-x}$ for $x > 0$ and 0 for $x < 0$ be the kernel function. Let h be the window width. Suppose n independent observations $X_1, \dots, X_n$ from f are available. Show that the mean of the kernel density estimate $\hat{f}(x)$ is given by: [10]

$$E(\hat{f}(x)) = \begin{cases} 0 & x < 0 \\ \frac{1}{a}(1 - e^{-\frac{x}{h}}) & 0 \leq x \leq a \\ \frac{1}{a}(e^{\frac{a}{h}} - 1)e^{-\frac{x}{h}} & a < x \end{cases}$$

3. In a two-class classification problem with two features, the density of the observations from the first class is uniform over the circle centered at 0 and radius 1/2 while the density of the observations from the other class is a standard bivariate normal distribution. Assuming equal prior probabilities and costs of misclassifications,

(a) find the Bayes optimal error for this problem. [7]

(b) show the required architecture (with proper justification) of a multilayer perceptron having appropriate weights that can produce the optimal boundary of classification for this problem. [10]

[P. T. O.]

4. In a two-class classification problem, there were 500 observations from each class in a training sample. A binary classification tree T was grown using equal prior probabilities and costs of misclassifications. Let $n_j(t), j = 1, 2$ denote the number of observations present in a node $t \in T$ for the two classes respectively. The stopping rule for splitting a node was that either the resubstitution error of the node was less than 2% or $n_j(t) \leq 5$ for any $j = 1, 2$. It was observed that whenever a node was split, $\lfloor 0.9 * n_1(t) \rfloor$ and $\lfloor 0.1 * n_2(t) \rfloor$ observations went to the left descendant node, and the remaining observations went to the right descendant node, where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x.

   (a) Calculate the number of terminal nodes of the tree. [5]

   (b) Calculate the resubstitution error of T. [3]

   (c) Generate the α-sequence for the cost complexity pruning, and the corresponding minimizing subtrees where α is the cost-complexity parameter. [10]

Group B

1. (a) Describe the Hough transform and explain how it can be used to detect straight line segments in a binary image.

   (b) Describe a morphological operation based approach to skeletonization of a binary image with an illustration. [(3+4) + 9]

2. Define Delaunay triangulation of a set S of n planar points. Describe an algorithm to find the Euclidean minimum spanning tree (EMST) on S from its Delaunay triangulation with computational complexity not as high as $O(n^2)$. Assume that no four points in S are co-circular. Explain how the EMST on S can be used to cluster the points in S. [3+9+5]

3. Define the major axis of an object in a binary image. Express the angle that it makes with the vertical axis in terms of the pixel co-ordinates of the object. Explain how the orientation of the object can be normalized on the basis of the major axis. [3 +3 +5]

4. Define a self organizing map clearly indicating the update equations. Under what conditions. do the weight vectors converge to the desired solution? Explain how a self organizing map can be used for clustering. [4 + 3 + 4]

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination: 2014-15

23-02-2015

Subject Name : **Advanced Cryptography** Maximum Score: 40

Course Name : M.Tech. (CS) II yr. Duration: 3 Hours

Note: Attempt all questions. Marks are given in brackets. Total marks is 50 but you can score maximum 40. Use separate page for each question.

Problem 1 (10). Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be indepedently distributed. Moreover, X_i's and Y_i's are identically distributed. Show that for any distinguisher D, there exists a distinguisher D' such that

$$\Delta_D((X_1, \ldots, X_n); (Y_1, \ldots, Y_n)) \leq n\Delta_{D'}(X_1, Y_1).$$

Problem 2 (16). Let $f : \{0,1\}^n \to \{0,1\}^n$ be an ϵ-one-way permutation. We define $g(x,r) = (f(x), r)$ and $b(x,r) = x \cdot r$ where $x, r \in \{0,1\}^n$. Show that b is hard-core bit for the permutation g. You need to prove for a concrete version by explicitly mentioning the parameters involved.

Problem 3 (8). Describe the one-way amplification construction using a random walk.

Problem 4 (6). Let $f : \{0,1\}^* \to \{0,1\}$ be an efficiently computable function. We define $F(n) = 2^{-n} \sum_{x \in \{0,1\}^n} f(x)$. For any k, construct a PPT A such that

$$Pr[|A(1^n) - F(n)| \leq n^{-k}] \geq 1 - 2^{-n}.$$

Problem 5 (10). Construct an one-bit pseudorandom function from a pseudorandom generator and justify it.

INDIAN STATISTICAL INSTITUTE
Mid-Semester Examination: 2014-2015, Second Semester
M-Stat II (MSP)
Ergodic Theory

Date: 23·02·2015 **Max. Marks 30** **Duration: 2 Hours**

Note: Answer all questions, maximum you can score is 30.

1. Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving dynamical system and suppose μ is a probability measure. Show that the following are consequences of Birkoff's ergodic theorem in this setup:

   (a) Poincaré's recurrence theorem

   (b) Strong Law of Large numbers for iid random variables $\{X_n\}$ with finite expectation. (You may assume that X_n takes only finitely many values.) [7+7]

2. Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving dynamical system on a Polish space X, Borel σ-field $\mathcal{B}$ and probability measure μ. Let $B \in \mathcal{B}$ be such that every point $x \in B$ is recurrent, i.e., $T^n x \in B$ infinitely often $\forall x \in B$. Let $C \subset B$, $C \in \mathcal{B}$.

   (a) Show that every point $x \in C$ is recurrent under T iff it is recurrent under T_B (the induced transformation on B).

   (b) Find mean recurrence time of B if T is ergodic. [4+4]

3. (a) If T is weak mixing, show that T^k is weak mixing $\forall k \geq 1$.

   (b) If T is weak mixing and S is ergodic, then show that $T \times S$ is ergodic. Also, show that the conclusion is not necessarily true if T is not weak mixing but just ergodic.

   (c) Suppose T is a measure preserving transformation such that $T \times S$ is ergodic for all ergodic transformation S. Then show that T is weak mixing.

   [2+6+6]

INDIAN STATISTICAL INSTITUTE

M.Stat Second Year

Mid-semestral Examination, Second Semester, 2014-15

23|02|2015

Time: $2\frac{1}{2}$ Hours

Applied Multivariate Analysis

Full Marks : 60

1. Let $\mathbf{X} = (X_1, X_2, X_3, X_4)'$ be a random vector with the mean $\boldsymbol{\mu} = \mathbf{1}$ and the dispersion matrix $\boldsymbol{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}'$, where $\mathbf{I}$ is the identity matrix, $\mathbf{1} = (1,1,1,1)'$ and $\rho < 0$. Define $Y = \alpha_0 + \sum_{i=1}^{4} \alpha_i X_i$, where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_4)$ is a unit vector.

   (a) Give a choice of $\boldsymbol{\alpha}$ that maximizes the variance of Y.

   (b) If $\alpha_4 > 0$, is this choice of $\boldsymbol{\alpha}$ unique. Justify your answer. [6+2]

2. Let $\mathbf{X} = (X_1, X_2, X_3, X_4)'$ be a random vector with the mean $\mathbf{0}$ and the dispersion matrix $\boldsymbol{\Sigma} = \mathbf{I} + \mathbf{1}\mathbf{1}'$, where $\mathbf{I}$ is the identity matrix and $\mathbf{1} = (1,1,1,1)'$. Define $Y_1 = \alpha_1 X_1 + \alpha_2 X_2$ and $Y_2 = \alpha_3 X_3 + \alpha_4 X_4$, where $\alpha_1, \alpha_2, \alpha_3$ and α_4 are real constants. Show that the correlation coefficient between Y_1 and Y_2 cannot exceed $2/3$. [6]

3. Show that the half space depth of an observation $\mathbf{x}$ with respect to F, a multivariate normal distribution with the mean vector $\boldsymbol{\mu}$ and the dispersion matrix $\boldsymbol{\Sigma}$, is given by

$$HD(\mathbf{x}, F) = 1 - \Phi\left(\sqrt{(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}\right),$$

   where Φ is the cumulative distribution function of the standard normal variate. Hence find the half-space median of the multivariate normal distribution. [6+2]

4. Consider a linear model $\mathbf{Y}_n = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\epsilon}_{p\times 1}$, where $Var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. If $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$ and $\mathbf{m}$ is a unit vector, show that $Var(\mathbf{m}'\widehat{\boldsymbol{\beta}})$ cannot be smaller that σ^2/λ, where λ is the largest eigenvalue of $\mathbf{X}'\mathbf{X}$. [6]

5. Consider a two-class classification problem, where the prior the prior probabilities of the two classes are equal and the density functions of these two classes are given by

$$f_1(x) = \begin{cases} 1 + sin(2k\pi x) & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{and } f_2(x) = \begin{cases} 1 + cos(2k\pi x) & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise,} \end{cases}$$

   where k is a positive integer. Show that the misclassification probability of the Bayes classifier does not depend on the value of k. [10]

6. Consider a classification problem involving two six-dimensional normal distributions $N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$ and $N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$. If the prior probabilities of these two distributions are equal and $\boldsymbol{\Sigma}_2 = 2\boldsymbol{\Sigma}_1$, find the misclassification probability of the Bayes classifier (you can use statistical tables). [10]

7. Prove or disprove the following statements [6 × 2=12]

   (a) If $\mathbf{X}$ follows a p-dimensional multivariate normal distribution, its density function f can be expressed as

$$f(\mathbf{x}) = \prod_{i=1}^{p} f_i(\boldsymbol{\alpha}_i'\mathbf{x}),$$

   where $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_p$ ($\|\boldsymbol{\alpha}_i\| = 1$ for all $i = 1, 2, \ldots, p$) are the principal component directions and $f_1, f_2, \ldots, f_p$ are the density functions of the corresponding principal components.

   (b) Two p-dimensional vectors $\mathbf{X} = (X_, X_2, \ldots, X_p)'$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)'$ are identical if and only if $f_{\mathbf{X}}(t) = \sum_{k=1}^{p} X_k\ sin(kt)$ and $f_{\mathbf{Y}}(t) = \sum_{k=1}^{p} Y_k\ sin(kt)$ coincide for all $t \in [-\pi, \pi)$.

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination : 2013 – 14

MStat (2^{nd} Year)

Actuarial Models

Date: 24 February 2015 Maximum Marks: 40 Duration: 2 Hours

This paper carries 44 marks. Attempt ALL questions. The maximum you can score is 40

1. The two football teams in a particular city are called United and City and there is intense rivalry between them. A researcher has collected the following history on the results of the last 20 matches between the teams from the earliest to the most recent, where:

U indicates a win for United;
C indicates a win for City;
D indicates a draw.

UCCDDUCDCUUDUDCCUDCC

The researcher has assumed that the probability of each result for the next match depends only on the most recent result. He therefore decides to fit a Markov chain to this data.
(i) Estimate the transition probabilities for the Markov chain.
(ii) Estimate the probability that United will win at least two of the next three matches against City.

[4 + 4 = 8]

2. (i) Define a Poisson process.

A bus route in a large town has one bus scheduled every 15 minutes. Traffic conditions in the town are such that the arrival times of buses at a particular bus stop may be assumed to follow a Poisson process. Mr Bean arrives at the bus stop at 12 midday to find no bus at the stop. He intends to get on the first bus to arrive.
(ii) Determine the probability that the first bus will not have arrived by 1.00 pm the same day.

The first bus arrived at 1.10 pm but was full, so Mr Bean was unable to board it.
(iii) Explain how much longer Mr Bean can expect to wait for the second bus to arrive.
(iv) Calculate the probability that at least two more buses will arrive between 1.10 pm and 1.20 pm.

[2 + 2 + 2 + 2 = 8]

3. A motor insurer offers a No Claims Discount scheme which operates as follows. The discount levels are {0%, 25%, 50%, 60%}. Following a claim-free year a policyholder moves up one discount level (or stays at the maximum discount). After a year with one or more claims the policyholder moves down two discount levels (or moves to, or stays in, the 0% discount level). The probability of making at least one claim in any year is 0.2.

(i) Write down the transition matrix of the Markov chain with state space {0%, 25%, 50%, 60%}.

(ii) State, giving reasons, whether the process is:

(a) irreducible.

(b) aperiodic.

(iii) Calculate the proportion of drivers in each discount level in the stationary distribution.

The insurer introduces a "protected" No Claims Discount scheme, such that if the 60% discount is reached the driver remains at that level regardless of how many claims they subsequently make.

(iv) Explain, without any further calculations, how the answers to parts (ii) and (iii) would change as a result of introducing the "protected" No Claims Discount scheme. [2 + 2 + 4 + 2 = 10]

4. Outside an apartment block there is a small car park with three parking spaces. A prospective purchaser of an apartment in the block is concerned about how often he would return in his car to find that there was no empty parking space available. He decides to model the number of parking spaces free at any time using a time homogeneous Markov Jump Process where:

- The probability that a car will arrive seeking a parking space in a short interval dt is

$$A.dt + o(dt).$$

- For each car which is currently parked, the probability that its owner drives the car away in a short interval dt is $$B.dt + o(dt).$$

where $A, B > 0$.

(i) Specify the state space for the above process.

(ii) Draw a transition graph of the process.

(iii) Write down the generator matrix for the process.

(iv) Derive the probability that, given all the parking spaces are full, they will remain full for at least the next two hours.

(v) Explain what is meant by a jump chain.

(vi) Specify the transition matrix for the jump chain associated with this process.

Suppose there are currently two empty parking spaces.

(vii) Determine the probability that all the spaces become full before any cars are driven away.

(viii) Derive the probability that the car park becomes full before the car park becomes empty.

[1 + 2 + 2 +2 + 1 + 2 + 1 + 2 = 13]

5. State three different methods of graduating crude mortality data. Give, for each method, one advantage and one disadvantage. [5]

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination: 2014-15

M. Stat. II Year

Advanced Sample Survey

Date: 25/02/2015 **Maximum Marks: 50** **Duration: 3 Hours**

Answer any 4 questions each carrying 10 marks.

Assignment records to be submitted on the date of examination carry 10 marks.

1. Explain considerations helpful to hit upon a desirable size of a simple random sample to be chosen without replacement from a finite population of a given size, assuming no distributional form for a vector of real values.

2. State and prove Basu's (1971) and Godambe's (1955) non-existence theorems concerning finite population totals. Explain why one of them cannot be deduced from the other.

3. Derive Hartley-Ross Ratio-type unbiased estimator for a finite population mean. Find an unbiased estimator for its variance.

4. From a probability proportional to size with replacement (PPSWR) sample in n (>2) draws present Hansen-Hurwitz's estimator for a finite population total and another one better than it explaining why it is so and give unbiased variance estimators for both.

5. Explain PPS circular systematic sample selection method. How will you employ Horvitz & Thompson estimator for the population total? Discuss how you may proceed to estimate its variance.

*****

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination: (2014-2015)

M. Stat Second Year

Statistical Inference II

Date: 25/2/15 Marks: ..30.. Duration: .2 hours.

Attempt all questions

1. Suppose that $\{X_n\}_{n=1}^{\infty}$ are bounded, exchangeable random variables. Let $\Theta = \lim_{n\to\infty} \sum_{i=1}^{n} X_i/n$, almost surely. Prove that

$$Var\,(\Theta) = Cov\,(X_1, X_2)\,.$$

[8]

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

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

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

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

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

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

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

3. Let $X_1, \ldots, X_n \overset{iid}{\sim} Uniform\,(\alpha, \beta)$.

(a) Obtain a minimal sufficient statistic T for $\boldsymbol{\theta} = (\alpha, \beta)$. Is T also complete? [3]

(b) Obtain $E\left(\bar{X}\,|T\right)$, where $\bar{X} = \sum_{i=1}^{n} X_i/n$. [4]

(c) For any prior π, let $E_\pi\left(\bar{X}|T\right)$ denote the Bayesian conditional expectation with respect to the prior π. Propose a prior π_1 for $\boldsymbol{\theta}$ for which $E_{\pi_1}\left(\bar{X}|T\right) = E\left(\bar{X}|T\right)$, and propose a prior π_2 for which $E_{\pi_2}\left(\bar{X}|T\right) \neq E\left(\bar{X}|T\right)$. [3]

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination

Second semester

M. Stat - Second year 2014-2015

Stochastic Processes I

Date: March 4, 2015 Maximum Marks: 25

Duration: 2 hours 30mins

Anybody caught using unfair means will immediately get 0. Please try to explain every step.

(1) Let (X, d) be a Polish space. Suppose $P_n \Rightarrow P$ and f is a bounded upper semi-continuous function on X. Then show that $\limsup_{n\to\infty} \int f(x)dP_n(x) \leq \int f(x)dP(x)$. [5 points]

[Recall:Upper semi-continuous at a point x means, given $\varepsilon > 0$, there is $\delta > 0$ such that $d(x, y) < \delta$ implies $f(y) < f(x) + \varepsilon$. Upper semi-continuous means upper semi-continuous at all points.]

(2) Let $(W_t^0, 0 \leq t \leq 1)$ be a Brownian Bridge. Define a stochastic process on $[0, \infty)$ by

$$B_t = (1+t)W^0_{\frac{t}{t+1}}, \qquad t \geq 0.$$

(a) Compute the mean and covariance function of B.

(b) Show that it is a continuous sample path Gaussian process with stationary and independent increments. [5+5=10 points]

(3) For any probability measures P and Q on the real line, denote by F_P and F_Q their distribution functions. Define

$$d(P, Q) = \frac{1}{2}\int_{-\infty}^{\infty} |F_{P}(x) - \cancel{G}_{Q}(x)|e^{-|x|}dx.$$

(a) Show that d is a metric on the space of probability measures on real line.

(b) Show that d metrizes weak convergence. [5+5=10 points]

(4) Let X_j be i.i.d. $N(0, 1)$ random variables. Let H be an Hilbert space with orthonormal basis $\{e_j\}_{j\geq 1}$. Let $X = \sum_{j\geq 1} X_j e_j/j$. For any $\varepsilon > 0$, find a compact set K such that, $P(X \in K) > 1 - \varepsilon$. [3 points]

(5) Show that given $(X_n)_{n\in\mathbb{N}}$ a sequence of $\mathbb{R}$-valued random variables $X_n \to X$ in distrn ~~probability~~ if $\mathsf{E}f(X_n) \to \mathsf{E}f(X)$ for any f Lipschitz and bounded. [2 points]

(6) Let C be the collection of bounded real continuous functions f on $(-\infty, \infty)$ which have finite limits at $-\infty$ and $+\infty$. For any probability P on $\mathbb{R}$ define an operator T_P by

$$T_P f(x) = \int f(x-y)P(dy).$$

- Show that T_P maps C into C.
- Show that $P_n \Rightarrow P$ if and only if for every $f \in C$,

$$T_{P_n} f \to T_P f \qquad \text{uniformly .}$$

[3+7=10 points]

Indian Statistical Institute

Second Midsemestral Examination 2014-15

M. Stat. II year

Theory of Games and Statistical Decisions

Date: March 9, 2015 Maximum marks: 60 Duration: 2 hrs.

Answer all Questions. Answers should be brief and to the point.

1. (a) Define a game and discuss rational behavior in game theoretic set up with examples.

   (b) Define utility function. State the conditions on the preference pattern of a player which ensures existence of a utility function. Discuss intutive justifications of these conditions.

   [12 + 8 = 20]

2. Consider the *Lexicographic order* in $\mathbb{R}^2$, that is, $(x_1, y_1) \preceq (x_2, y_2)$ if either $x_1 < x_2$ or, $x_1 = x_2$ and $y_1 \leq y_2$.

   (a) Show that the Lexicographic order is complete, reflexive and transitive.

   (b) Is the Lexicographic ordering continuous? Justify your answer.

   [8 + 12 = 20]

3. (a) Let $X \subseteq \mathbb{R}^n$ for some $n \geq 1$. Define a set valued function $f : X \to X$ whose graph is closed. Give a nontrivial example of such a function.

   (b) State Kakutani's fixed point theorem in $\mathbb{R}^n$.

   (c) Define Nash equilibrium of an n-person game in strategic form. Using 3(b) or otherwise state and prove the existence theorem of Nash equilibrium of an n-person game in strategic form.

   [3+5+12 = 20]

INDIAN STATISTICAL INSTITUTE

SEMESTRAL EXAMINATION

M.Stat (1st Year) 2014-15

Subject: *Time Series Analysis*

Date: 24.04.2015 **Full Marks:** 60 **Duration:** 3 hours

Attempt all questions

1. [2+3+4+(2+2) = 13]

a) Define spectral density of a time series.

b) Give an example of a time series for which spectral density does not exist.

c) Derive spectral density of AR(1) processes with parameters (ϕ, σ^2).

d) For AR(1) process, draw the graphs of spectral densities for (i) $(\phi, \sigma^2) = (+0.5, 1)$ and (ii) $(\phi, \sigma^2) = (-0.5, 1)$.

Discuss the significance of them.

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

a) Write down Bartlett's formula for variance of sample autocorrelation function ($\hat{\varrho}$ (h)), where h≥1 denotes the lag.

b) Derive the expression for Var($\hat{\varrho}$ (h)) in the context of (i) MA(1) processes (ii) AR(1) processes for different values of h≥1.

c) Let for a sample of size 100, $\hat{\varrho}$ (1)= 0.438 and $\hat{\varrho}$ (2)= 0.145.

Construct 95% confidence interval for $\varrho(1)$ and $\varrho(2)$, assuming that data is generated from (i) an AR(1) model, (ii) an MA(1) model.

3. [2+6 = 8]

a) State Wold decomposition theorem.

b) Using the theorem show that a weakly stationary time series is MA(q) iff it is q-correlated.

4. [5 +5 + (2+3+4) = 19]

a) Consider an AR(1) process (assuming suitable parameters). Assume X_1, X_3 are observed and X_2 is missing. Predict X_2 on the basis of X_1 and X_3, using best linear predictor.

b) Consider an AR(p) process (assuming suitable parameters) with mean zero. Derive best linear predictor of X_{n+2} on the basis of X_n, X_{n-1}, ……, X_1, X_0.

c) Define an ARMA(p,q) process with relevant difference equation. Derive suitable necessary and sufficient conditions on the parameters for the process to be(i) stationary, (ii) causal stationary.

Semester 2, 2014-2015 FINAL EXAM Date: 27 April 2015

MStatII DIRECTIONAL DATA ANALYSIS Time: 3 hrs.

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

Q. 1. [25=12+13] (a) Derive the general representation of the p.d.f of a circular random variable in terms of its trigonometric moments.

(b) Consider the circular distribution with p.d.f. given by
$f(\theta) = [1/2\pi].[1 + 2\rho cos(\theta - \mu) + 2\rho^2 cos2(\theta - \mu)]$

Asssume $\mu = 0$. Derive a closed form consistent (to be proved) estimator of ρ.

Q. 2. [25=12+8+5] (a) Derive the Locally Most Powerful Invariant test for Isotropy against the family of wrapped symmetric stable distributions.

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

(c) Comment on the robustness property and failure of this test against some other families of circular distributions.

Q. 3. [25=12+(10+3)](a) Describe two models for Cylindrical Regression. For any one of these models, demonstrate how you will estimate its parameters.

(b) (i) Starting with linear and circular probability distributions as members of exponential families, show how in general you can derive the distribution of the points in a unit disc. State precisely, without proof, the basic theorem you may need for this derivation. (ii) Give a specific example of a distribution in (i).

Q. 4. [25= 18+7] TAKE HOME. (a) For the given data set on the hyperdisc, test the independence between its linear and spherical component random variables.

(b) For the bivariate wrapped Cauchy distribution defined on the torus as derived in the class, derive the conditions under which both its conditionals have univariate wrapped Cauchy distributions.

Indian Statistical Institute
Second Semestral Examination 2014-15

M. Stat. II

Theory of Games and Statistical Decisions

Date: April 29, 2015 Maximum marks: 100 Duration: 3 hrs.

Answer all Questions. *Paper carries 110 points.*

1 (a) Define a strategic game with n players and the best respose function of a player with the help of a suitable example.

(b) Two players are in dispute over an object. The value of the object to player i is $v_i > 0$, for $i = 1, 2$. The time scale is the set of nonnegative integer variable ($t = 0, 1, 2, \ldots$) and at instant t of the game each player independently decide whether to concede the object or not. If the first player to concede does so at time t, the other player obtains the object at that time. If both players concede the object simultanepusly at the same time point t the object is split equally between them with i th player receving payoff $v_i/2$ respectively. Further assume **there is a penalty for continuing the game** so that until first concession each player looses one unit worth of payoff per instance the game is played.

Formulate the situation as a strategic game (with all components clearly identified) and obtain possible Nash equilibria of the game.

[10+ 15 = 25]

2 (a) Consider a matrix game (zero-sum, with finite possible strategies for both players). Let us assume an $m \times n$ matrix A for this purpose whose entries indicate payoffs for player I. Moreover let $A_{\cdot j}$ and $A_{k\cdot}$ denote the j th row and k th column of A for $1 \leq j \leq m$ and $1 \leq k \leq n$ respectively. If X^* and Y^* denote two mixed strategies for the two players, so that

$$X^{*T}AY^* = \min_X X^TAY^* = \min(X^{*T}A_{1\cdot}, X^{*T}A_{2\cdot}, \ldots, X^{*T}A_{n\cdot})$$

Show that (X^*, Y^*) is a saddle point of the game.

... P.T.O.

(b) Let $v(A)$ denote the value of a matrix game. If $X^{*T} = (x_1^*, x_2^*, \dots x_m^*)$ is an arbitrary optimal strategy for player I and if for some k

$$X^{*T} A_{k\bullet} > v(A),$$

then for any optimal strategy $Y^* = (y_1^*, y_2^*, \dots y_n^*)$, show that $y_k^* = 0$.

[15 + 10 =25]

3 (a) Describe the notion of an extensive form game with in both algebraic and tree (graphical) form with examples. Give an example to show how imperfect information changes (lack of perfect knowledge about each other's moves) can distort set of possible strategies of players. Also, discuss how Nash equilibrium of such a game can be found using backward induction algorithm.

(b) What is a subgame perfect Nash equilibrium? Show that every finite extensive game with perfect information has a subgame perfect equilibrium.

[(4+ 4+ 10) + 12 = 30]

4 (i) Describe a coalitional game and it's characteristic functions.

(ii) Define group and individual rationality in a N-person coalitional game.

(iii) When is a characteristic function of a coalitional game called inessential? What is the core of a coalitional game?

(iv) Show that the core of an essential constant sum game is empty.

(v) Consider the three person game with players I, II and III with two pure strategies each and with the following payoff vectors.

(i) If player I chooses strategy 1, the payoff vectors are given by the following 2-D array

$$\begin{array}{ccc} & III-1 & III-2 \\ II-1 & (0,3,1), & (2,1,1) \\ II-2 & (4,2,3), & (1,0,0) \end{array}$$

(ii) If player I chooses strategy 2, the payoff vectors are given by the following 2-D array

$$\begin{array}{ccc} & III-1 & III-2 \\ II-1 & (1,0,0), & (1,1,1) \\ II-2 & (0,0,1), & (0,1,1) \end{array}$$

Find out $v(\{1\}), v(\{2\})$ and $v(\{1,2,3\})$ respectively when it is assumed that individually the players assume complementary alliances are non-cooperative (making the coalitional subgames zero-sum).

[3+3+3+3+ 18 =30]

INDIAN STATISTICAL INSTITUTE

Semestral Examination: 2014 – 15

MStat (2nd Year)

Actuarial Models

Date: 2 May 2015 Maximum Marks: 100 Duration: 3 Hours

1 (a) In the context of a survival model define right censoring, Type I censoring and Type II censoring. Give an example of a practical situation in which censoring would be informative.

(b) A trial was conducted on the effectiveness of a new cream to treat a skin condition. 100 sufferers applied the cream daily for four weeks or until their symptoms disappeared if this happened sooner. Some of the sufferers left the trial before their symptoms disappeared.
(i) Describe two types of censoring that are present and state to whom they apply.

The following data were collected.

Number of sufferers	*Day symptoms disappeared*	*Number of sufferers*	*Day they left the trial*
2	6	3	2
2	7	2	10
1	10	3	15
2	14		

(ii) Calculate the Nelson-Aalen estimate of the survival function for this trial.
(iii) Sketch the survival function, labeling the axes.
(iv) Estimate the probability that a person using the cream will still have symptoms.

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

2 (a) A graduation of a set of crude mortality rates is tested for goodness-of-fit using a chi-squared test. Discuss the factors to be considered in determining the number of degrees of freedom to use for the test statistic.

(b) A certain non-fatal medical condition affects adults. Adults with the condition suffer frequent episodes of blurred vision. A study was carried out among a group of adults known to have the condition. The study lasted one year, and each participant in the study was asked to record the duration of each episode of blurred vision. All participants remained under observation for the entire year.

The data from the study were analysed using a two-state Markov model with states:

1. not suffering from blurred vision.
2. suffering from blurred vision.

Let the transition rate from state i to state j at time $x+t$ be μ_{x+t}^{ij}, and let the probability that a person in state i at time x will be in state j at time $x+t$ be ${}_{t}p_{x}^{ij}$.

(i) Derive from first principles the Kolmogorov forward equation for the transition from state 1 to state 2.

The results of the study were as follows:

Participant-days in state 1	21,650
Participant-days in state 2	5,200
Number of transitions from state 1 to state 2	4,370
Number of transitions from state 2 to state 1	4,460

Assume the transition intensities are constant over time.

(ii) Calculate the maximum likelihood estimates of the transition intensities from state 1 to state 2 and from state 2 to state 1.

(iii) Estimate the probability that an adult with the condition who is presently not suffering from blurred vision will be suffering from blurred vision in 3 days' time.

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

3. (a) (i) State the form of the hazard function for the Cox Regression Model, defining all the terms used.

(ii)Write down the equation of the Cox proportional hazards model in which the hazard function depends on duration t and a vector of covariates z. You should define all the other terms that you use.

(iii) State two advantages of the Cox Regression Model. Explain why the Cox model is sometimes described as semi-parametric.

(b) Suman is studying for an on-line test. He has collected data on past attempts at the test and has fitted a Cox Regression Model to the success rate using three covariates:

Employment $Z1 = 0$ if an employee, and 1 if self-employed
Attempt $Z2 = 0$ if first attempt, and 1 if subsequent attempt
Study time $Z3 = 0$ if no study time taken, and 1 if study time taken

Having analysed the data Suman estimates the parameters as:

Employment	0.4
Attempt	−0.2
Study time	1.25

Bijan is an employee. He has taken study time and is attempting the test for the second time. Biman is self-employed and is attempting the test for the first time without taking study time.
(i) Calculate how much more or less likely Biman is to pass, compared with Bijan.

Suman subsequently discovers that the effect of the number of attempts is different for employees and the self-employed.
(ii) Explain how the model could be adjusted to take this into account.

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

4. (a) Describe the state space and the time space of the following processes:
General Random Walk, Compound Poisson Process, Counting Process, Poisson Process

(b) A life office is trying to understand the impact of certain factors on the lapse rates of its policies. It has studied the lapse rates on a block of business subdivided by:

- sex of policyholder (Male or Female)
- policy type (Term Assurance or Whole Life)
- sales channel (Internet, Direct Sales Force or Independent Financial Adviser)

The office has fitted a Cox proportional hazards model to the data and has calculated the following regression parameters:

Covariate	*Regression parameter*
Female	0.2
Male	0
Term Assurance	−0.1
Whole Life	0
Internet	0.4
Independent Financial Adviser	−0.25
Direct Sales Force	0

(i) State the sex/sales channel/policy type combination to which the baseline hazard relates.

A Term Assurance is sold to a Female by an Independent Financial Adviser.
(ii) Calculate the probability that this Term Assurance is still in force after five years given that 60% of Whole Life policies bought on the Internet by Males have lapsed by the end of year five.

[8 + 6 + 6 = 20]

5. (a) Describe the differences between deterministic and stochastic models.

(b) An investigation was conducted into the effect marriage has on mortality and a model was constructed with three states: 1 Single, 2 Married and 3 Dead.
It is assumed that transition rates between states are constant.
(i) Sketch a diagram showing the possible transitions between states.
(ii) Write down an expression for the likelihood of the data in terms of transition rates and waiting times, defining all the terms you use.

The following data were collected from information on males and females in their thirties.

Years spent in Married state	40,162
Years spent in Single state	10,218
Number of transitions from Married to Single	1,382
Number of transitions from Single to Dead	12
Number of transitions from Married to Dead	9

(iii) Derive the maximum likelihood estimator of the transition rate from Single to Dead.
(iv) Estimate the constant transition rate from Single to Dead and its variance.

[4 + (6 + 4 + 3 + 3) = 20]

INDIAN STATISTICAL INSTITUTE
Semester Examination: 2014-2015, Second Semester
M-Stat II
Ergodic Theory

Date: 05·05·15 Max. Marks 50 Duration: 3 Hours

Note: 1. Answer all questions.

2. All the measures considered are probability measures

3. Total Marks: 58. Maximum you can score: 50.

1. a) Let $(X, \mathcal{B}, \mu)$ be a probability space, $T : X \to X$ a measure preserving transformation and $f \in L_1(\mu)$. Then show that $a_n(x) := \frac{1}{n}\sum_{i=0}^{n-1} f(T^i(x))$, $n \geq 1$, forms a sequence of uniformly integrable functions. Hence show that, assuming that $\lim_n a_n(x)$ exists $a.s.$, the limit is $E(f|\mathcal{I})$ where. $\mathcal{I}$ is the σ-field consisting of the sets which are invariant under T. [8]

   (b) Consider the probability space $([0,1), \mathcal{B}[0,1), \lambda)$. Show that the asymptotic relative frequency of 8 in the decimal representation of $x \in [0,1)$ is $\frac{1}{10}$ $a.s.$ [7]

2. a) Let $(X, \mathcal{B}, \mu, T)$ be an invertible measure preserving dynamical system. Let $E \in \mathcal{B}$ be such that $\mu(E) > 0$. Show that $\exists F \in \mathcal{B}$ with $\mu(F) > 0$ such that $\{n \in \mathbf{Z} : T^n(x) \in E\}$ has positive upper density $\forall x \in F$. [10]

   b) Let T_1 and T_2 be two invertible measure preserving transformations on a probability space $(X, \mathcal{B}, \mu)$ such that $T_1 \circ T_2 = T_2 \circ T_1$. If T_1 is ergodic and T_2 is weakly mixing, then show that T_1 is weakly mixing. [10]

3. Let $\{x_1, x_2, \ldots, x_k\}$ be a finite sequence of reals, not all equal and

$$f(\lambda) := \frac{1}{\sum_{i=1}^{k} e^{\lambda x_i}} \sum_{i=1}^{k} x_i e^{\lambda x_i}, \ \lambda \in \mathbf{R}.$$

a) i. Prove that f is strictly increasing in λ. [5]

ii. If $x_1 < x_2 < \cdots < x_k$, then show that $\lim\limits_{\lambda \to -\infty} f(\lambda) = x_1$ and $\lim\limits_{\lambda \to \infty} f(\lambda) = x_k$. [5]

iii. If x_i are as in ii. and $\alpha \in \mathbf{R}$ such that $x_1 < \alpha < x_k$ then show that $\exists$ a unique λ with $f(\lambda) = \alpha$. Show that λ is negative or positive according as $\alpha < \frac{1}{k}\sum\limits_{i=1}^{k} x_i$ or, $\alpha > \frac{1}{k}\sum\limits_{i=1}^{k} x_i$ respectively. [5]

(b) Let $x_1 < \alpha < x_k$ (notations are as in (a)). Show that, among all the probability vectors $\mu := (\mu_1, \ldots, \mu_k)$, satisfying $\sum\limits_{i=1}^{k} x_i\mu_i = \alpha$, the one that maximizes entropy with respect to μ, is given by the Boltzmann distribution, viz.,

$$\mu(x_i) := \frac{e^{\lambda x_i}}{\sum\limits_{i=1}^{k} e^{\lambda x_i}}, \quad i = 1, 2, \ldots, k$$

for some unique λ which satisfies the given condition. [8]

INDIAN STATISTICAL INSTITUTE

Semestral Examination, 2014-15

M.Stat Second Year, Second Semester

C 5/05/2015

ime: 3 Hours | Applied Multivariate Analysis | Full Marks: 100

1. Systolic blood pressure (X_1), age (X_2) and weights (X_3) of 10 individuals are given below.

X_1	137	145	153	164	154	168	142	149	159	129
X_2 (in years)	50	58	66	71	63	73	52	60	63	44
X_3 (in pounds)	171	178	193	204	195	212	189	188	203	167

[$\sum X_{1i} = 1500$, $\sum X_{2i} = 600$, $\sum X_{3i} = 1900$, $\sum X_{1i}^2 = 226326$, $\sum X_{1i}X_{2i} = 90985$. $\sum X_{1i}X_{3i} = 286538$, $\sum X_{2i}^2 = 36768$, $\sum X_{2i}X_{3i} = 115102$, $\sum X_{3i}^2 = 362922$.]

(a) Find α_2 and α_3 such that the correlation between X_1 and $\alpha_2 X_2 + \alpha_3 X_3$ is maximum. Calculate that maximum value. [6+2]

(b) Consider a linear model $X_{1i} = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$; $i = 1, 2, \ldots, n$, where the ϵ_i's are independent and identically distributed with the mean 0 and the variance σ^2. Using your result in (a) or otherwise, find $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$, the least square estimates of β_1, β_2 and β_3. If $m_2^2 + m_3^2 = 1$, find m_2 and m_3 such that the variance of $m_2\hat{\beta}_2 + m_3\hat{\beta}_3$ is minimum. [4+4]

2. (a) The following table shows the median diameter (mm) of granules of sand (X) and the gradient of beach slope in degrees (Y) for 7 naturally occurring ocean beaches.

X	0.19	0.22	0.24	0.35	0.42	0.75	0.85
Y	0.70	0.82	1.15	3.00	4.30	9.60	11.30

Using an appropriate distance function and the average linkage method, perform hierarchical clustering on this data set. Draw the corresponding dendogram. [6+2]

(b) Use the Dunn Index or the silhouette index to determine the number of clusters in the data set given in (a). [6]

(c) Write two main advantages of the k-medoids algorithm over the k-means algorithm [2]

3. (a) Write down an orthogonal k-factor model with appropriate assumptions. [3]

(b) Show that if the principal component method is used for factor analysis, the total variance explained by the first k factors is equal to the sum of the largest k eigenvalues of the estimated dispersion matrix. [4]

(c) Describe how you will test whether a two-factor model is appropriate for a data set. [4]

(d) What is varimax rotation ? Considering a two-factor model, describe how it helps in factor analysis. [2+3]

4. Consider a two-class classification problem where each of the competing classes is an equal mixture of two bivariate normal distributions. While class-1 is a mixture of $N_2(1,1,1,1,0)$ and $N_2(-1,-1,1,1,0)$, class-2 is a mixture of $N_2(1,-1,1,1,0)$ and $N_2(-1,1,1,1,0)$. Here $N_2(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$ denotes a bivariate normal distribution with mean $(\mu_1,\mu_2)'$ and the scatter matrix $\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$.

(a) Find the Bayes classifier and compute the Bayes risk [4+4]

(b) Find the mean vectors and the dispersion matrices of these two distributions. [4]

(c) Suppose that we have 100 observations from each of these two competing classes. If one constructs the quadratic discriminant analysis rule based on these 200 observations, how will the resulting classifier perform ? Justify your answer. [4]

5. (a) Prove or disprove- "If the densities of the two competing classes are continuous and the priors are equal, Bayes classifier is unique". [3]

(b) In order to construct a classification tree for a two-class classification problem, one needs to find the best split based on each of the measurement variables. Consider a categorical variable C which has three categories C_1, C_2, C_3. The number of observations in these three categories are given below.

Category	C_1	C_2	C_3	Total
No. of obs. from Class-1	150	200	150	500
No. of obs. from Class-2	150	0	350	500
Total	300	200	500	1000

If the impurity function is concave and symmetric in its arguments, find the best split based on C. Justify your answer. [5]

(c) Consider a kernel discriminant analysis method that uses the Gaussian kernel with the common bandwidth matrix $h^2\mathbf{I}$ for all classes (here $\mathbf{I}$ denotes the identity matrix and h is a positive constant). Suppose that there are J competing classes and their prior probabilities are $p_1.p_2,\ldots,p_J$. Show that

(i) if h is small, it behaves like 1-nearest neighbor classifier. [3]

(ii) if h is large, its probability of correct classification cannot be smaller than $\sum p_i^2$. [5]

6. Assignments [20]

Statistical Methods in Public Health

Semestral Examination

M.Stat. II Year, 2014-2015

Total Marks - 100

Time - 3 hrs. 30 mins.

Indian Statistical Institute
Kolkata 700 108, INDIA

Attempt questions as per the instructions :

Group A

1. (a) Let X be an $(n \times q)$ longitudinal data matrix in which any row corresponds to a q - variate size measurement available at q time points on one of the n individuals. Construct four estimated relative growth rate vectors (each having dimension $(1 \times \overline{q-1})$) using four separate RGR estimates for each of the vectors.

[4]

(b) Fisher's Relative Growth Rate (RGR) is a growth law invariant metric - discuss. Suggest a growth law non-invariant metric in this context with a suitable example. Derive the expressions for bias and MSE for the suggested metric under both first and second order of approximations (where order is defined by the power of "difference of relative errors" at two time points). Write down the expressions for estimated bias and MSE based on the data structure given in 1(a).

[2 + 3 + 8 + 2 = 15]

or

Find the analytical solution of the growth curve governed by the following growth equation

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

where, $a, b > 0$ and c is an integer. Compare the point of inflexions of RGR function for $c = 1$ and 2 respectively, with proper interpretations. Show that the growth curve has an analytic solution for any a, b, and $c > 0$ under Weibull growth function. Define instant maturity rate and show that the instant maturity rate under Weibull structure can capture all monotonicity of RGR.

[6 + 4 + 2 + 3 = 15]

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

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

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

[4 + 6 = 10]

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

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

Using the approximate expression for expectation and variance of the logarithm of the ratio of RGR for two consecutive time points, construct an asymptotic test for the null hypothesis of the extended GGCM. Also, suggest required modifications of the test statistic when the errors are non-normal.

[10 + 3 = 13]

3. Let us consider the stochastic differential equation

$$d(x(t)) = \mu(x(t)) + \sigma(x(t)) * dW(t),$$

where, $x(t)$ is the realization of size variable at time point t for a certain species population and the infinitesimal mean $\mu(x(t))$ specifies the underlying deterministic growth curve, while the infinitesimal variance $\sigma(x(t))$, corresponds to stochastic fluctuations. Let us assume $d(W(t))$ is the differential of Wiener process having mean zero and variance $d(t)$.

(a) Assuming exponential growth and homoscadastic variance structure, find the expression for expected time to extinction with initial population size as $x(0)$.

(b) Derive the limiting growth model when

$$\mu(x(t)) = r_m \left(\frac{x(t)}{a} - 1\right)\left(1 - \left(\frac{x(t)}{k}\right)^{\theta}\right)$$

[5 + 4 = 9]

4. (a) Consider the θ-logistic growth equation for a single species population dynamics as follows:

$$\frac{dx}{dt} = ax - bx^{\theta+1},$$

where parameters have their usual interpretations.

Derive the quasi equilibrium probabilities and determine the expression for the approximate mean and variance while introducing random perturbation in the above model.

[4 + 5 = 9]

Group B

1. (a) State clearly the basic assumptions of pure birth-death process and hence write down the basic stochastic differential equation.

(b) Can we find out the solution of the differential equation? Justify your answer.

(c) Describe the behavior of the process by calculating the mean and the variance of the population size.

(d) Derive the stochastic model of simple epidemics.

[3 + 3 + 2 + 6 + 8 = 22]

2. Some infectious diseases do not confer immunity. Such infections do not have a recovered state and individuals become susceptible again after infection.

(a) Write down the basic mathematical model to represent the above dynamics.

(b) Show that all the solutions of the above system are eventually bounded.

(c) Find out the basic reproduction number and state clearly why this number is so important to public health authority?

(d) Show that the endemic steady state of the system is globally asymptotically stable either by using Dulac criterion or Poincare-Bendixson theorem.

[3 + 3 + 4 + 8 = 18]

INDIAN STATISTICAL INSTITUTE

Semestral Examination: (2014-2015)

M. Stat Second Year

Statistical Inference II

Date: 08/05/15 Full Marks: ..100.. Duration: .3 hours.

Attempt all questions

1. (a) Let Θ be a random variable. Suppose that $\{X_n\}_{n=1}^{\infty}$ are *iid* $N(0,1)$. Let $T(0) = 0$. For each $i > 0$, let $T(i)$ be the first $j > T(i-1)$ such that $X_j \geq \Theta$. Let $Y_i = X_{T(i)}$ for $i = 1, 2, \ldots$. If we use Lebesgue measure as an improper prior distribution for Θ, how many realizations must we observe before the posterior becomes proper?

(b) Let $\mathcal{X} = \{1, 2, \ldots, k\}$ and let $u_1, u_2, \ldots, u_k$ be non-negative integers such that $u = \sum_{i=1}^{k} u_i > 0$. Suppose that an urn contains u_i balls labeled i for $i = 1, \ldots, k$. We draw a ball at random (every ball in the urn has the same probability of being drawn) and record X_1 equal to the label. We then replace the ball and toss in one more ball with the same label. We then draw a ball at random again to get X_2 and repeat the process indefinitely. Prove that the sequence $\{X_i\}_{i=1}^{\infty}$ is exchangeable.

(c) Suppose that an urn has 20 balls, 14 of which are red and 6 of which are blue. Suppose that we draw balls without replacement . Let $X_i = 1$ if the i-th ball is red and 0 if it is blue. Are the X_i exchangeable? Are they conditionally *iid*? Justify.

[10+10+5=25]

2. (a) Suppose that there exists a sufficient statistic of fixed dimension k for all sample sizes. That is, suppose that there exist functions T_n (with image $\mathcal{T} \subseteq \mathbb{R}^k$, for all n), $m_{1,n}$ and $m_{2,n}$ such that

$$f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\theta) = m_{1,n}(x_1, \ldots, x_n) m_{2,n}(T_n(x_1, \ldots, x_n), \theta).$$

Suppose also that for all n and all $t \in \mathcal{T}$,

$$0 < c(t,n) = \int_{\Omega} m_{2,n}(t,\theta) d\lambda(\theta) < \infty,$$

for some measure λ. Then prove that the family of densities with respect to λ, given by

$$\mathcal{P} = \left\{ \frac{m_{2,n}(t,\cdot)}{c(t,n)} : t \in \mathcal{T}, n = 1, 2, \ldots \right\}$$

forms a conjugate family in the sense that the posterior density with respect to λ is a member of this class if the prior is a member of this class.

(b) Suppose that $\mathcal{X} = \mathbb{R}$ and $\mathcal{T}_n = \mathbb{R} \times \mathbb{R}^+$, with $\mathcal{T}_n(x_1, \ldots, x_n) = (\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$, and $r_n(\cdot, (t_1, t_2))$, the conditional distribution of $(X_1, \ldots, X_n)$ given $\mathcal{T}_n(x_1, \ldots, x_n) = (t_1, t_2)$, is the uniform distibution on the surface of the sphere of radius $\sqrt{t_2 - t_1^2/n}$ around $(t_1, \ldots, t_n)/n$. Find the extremal family of distributions.
Hint: If an n-dimensional vector Y is uniformly distributed on the sphere of radius 1 around $\mathbf{0}$, then r_n is the distribution of $t_1/n + Y\sqrt{t_2 - t_1^2/n}$.

[10+15=25]

3. (a) Suppose that $\{X_n\}_{n=1}^\infty$ are conditionally independent with distribution P given $\mathcal{P} = P$, and P has Dirichlet(α) distribution where α is a finite measure on $(\mathcal{X}, \mathcal{B})$. Let K_n be the number of distinct values amongst $X_1, \ldots, X_n$. Prove that $\lim_{n\to\infty} E(K_n)/\log(n) = \alpha(\mathcal{X})$.

(b) Suppose that person 1 believes that $\{X_n\}_{n=1}^\infty$ are *iid* with a continuous distribution. For each $\theta \in \Omega$, let α_θ be a continuous finite measure with $\alpha_\theta(\mathcal{X}) = c$ for all θ. Suppose that person 2 models the data as conditionally *iid* given $\mathcal{P} = P$ and $\Theta = \theta$ with distribution P and that $\mathcal{P}$ given $\Theta = \theta$ has Dirichlet(α_θ) distribution. Suppose that person 3 models the data as conditionally *iid* given $\Theta = \theta$ with distribution α_θ/c. Assume that $\alpha_\theta \ll \eta$ for all θ. Suppose that person 2 and person 3 use the same prior distribution for Θ. Then prove that person 1 believes that, with probability 1, for every n, person 2 and person 3 will calculate exactly the same posterior distributions for Θ given $X_1, \ldots, X_n$.

[15+10=25]

4. (a) Let $(\mathcal{X}, \mathcal{B})$ be a Borel space, and for each integer $n > 0$, let π_n be a countable partition of $\mathcal{X}$ whose elements are in $\mathcal{B}$. Let $P = \{P(B) : B \in \mathcal{C}\}$ be tailfree with respect to $(\{\pi_n\}_{n=1}^\infty, \{V_{n;B} : n \geq 1, B \in \pi_n\})$, where $\mathcal{C} = \cup_{n=1}^\infty \pi_n$, and $\{V_{n;B} : n \geq 1, B \in \pi_n\}$ is a collection of non-negative random variables. State and prove two necessary and sufficient conditions for P to be a random probability measure.

(b) Prove that Dirichlet processes are tailfree with respect to every sequence of partitions. State and prove any result on Dirichlet distributions that you may use in this regard.

[15+10=25]

INDIAN STATISTICAL INSTITUTE

Semestral Examination: 2014-15 (Second Semester)
Master of Statistics (M. Stat.) II Year
Advanced Probability II

Teacher: Parthanil Roy

Total Marks: 50 **Date: May 11, 2015** **Duration: 2:30 - 6:00 pm**

Note:

- **Please write your name and roll number on top of your answer booklet(s).**
- **This is an open note examination. You are allowed to use your own hand-written notes (such as class notes, exercise solutions, list of theorems, formulas etc.). Please note that no printed or photocopied material is allowed. In particular, you are not allowed to use books, photocopied class notes etc.**

1. Suppose $\{\mu_n\}$ is a sequence of probability measures on $\mathbb{R}^d$ with cumulant generating function $\Lambda_n(\theta) = \log \int e^{\theta . x}\, d\mu_n(x)$, $\theta \in \mathbb{R}^d$. Assume that the (possibly infinite) limit $\Lambda(\theta) := \lim \frac{1}{n}\Lambda_n(\theta)$ exists for every $\theta \in \mathbb{R}^d$, and 0 is an interior point of the set $D_\Lambda := \{\theta \in \mathbb{R}^d : \Lambda(\theta) < \infty\}$. Let Λ^* be the Fenchel-Legendre transform of Λ.

   (a) (6 marks) Show that for each $a \in (0, \infty)$, the set $\{x \in \mathbb{R}^d : \Lambda^*(x) \leq a\}$ is bounded.

   (b) (6 marks) For each $j \in \{1, 2, \ldots, d\}$, let μ_n^j be the j^{th} marginal of μ_n and e_j be the unit vector in the j^{th} direction. For all $j \in \{1, 2, \ldots, d\}$ and for all $\rho, c > 0$, show that

   $$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n^j(-\infty, -\rho) \leq -c\rho + \Lambda(-ce_j).$$

   (c) (6 marks) Using 1(b) or otherwise, show that $\{\mu_n\}$ is exponentially tight.

   (d) (12 marks) Show that for every closed set $F \subseteq \mathbb{R}^d$,

   $$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(F) \leq -\inf_{x\in F} \Lambda^*(x).$$

2. Let $\{\mu_n\}$ be a sequence of probability measures on a Polish space S and C_b be the space of all bounded real-valued continuous functions defined on S.

   (a) (12 marks) If there exists a function $I : S \to [0, \infty]$ such that for every $h \in C_b$,

   $$\liminf_{n\to\infty} \frac{1}{n} \log \int e^{-nh(x)}\, d\mu_n(x) \geq -\inf_{x\in S}(h(x) + I(x)),$$

   then show that for all $u \in S$ and for all $\delta > 0$,

   $$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B(u, \delta)) \geq -I(u).$$

   (b) (8 marks) If for every $h \in C_b$, the (possibly infinite) limit $\Lambda(h) := \lim \frac{1}{n} \log \int e^{-nh(x)}\, d\mu_n(x)$ exists, then show that there exists a lower semicontinuous function $I : S \to [0, \infty]$ such that for all open set $U \subseteq S$,

   $$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(U) \geq -\inf_{x\in U} I(x).$$

INDIAN STATISTICAL INSTITUTE

Semestral Examination, Second Semester: 2014-15

M.Stat. II Year (AS)

Survival Analysis

Date: May 12, 2015 Maximum marks: 100 Duration: $3\frac{1}{2}$ hours

Answer all questions. Standard notations are followed.

1. Obtain an expression for the survival function of the distribution having mean residual life at time t given by $\mu e^{-\theta t}$. Is the failure rate an increasing function? [5+4=9]

2. Derive the generalized maximum likelihood estimator (GMLE, in sense of Kiefer and Wolfowitz) of a distribution function, among the class of all distributions over the positive real line, on the basis of randomly left-censored observations from it. You can assume that the censoring times are samples from another distribution and are independent of the time-to-event of interest. [10]

   [Hint: Consider a monotone decreasing transformation of the original data.]

3. A group of newborn babies undergo health check-up in monthly intervals from birth to two years of age. Along with other health parameters, the status of the baby, indicating whether the first teething has occurred, is recorded. The objective is to estimate the teething time distribution.

   (a) Formulate the model in terms of the underlying random variable, its realizations and the observed data.

   (b) Describe the nature of censoring, and indicate whether the censoring can be said to be independent.

   (c) Obtain the likelihood and express it in terms of the data in summarized form as much as possible.

   (d) Show that, if no parametric family of distributions is assumed, the likelihood does not have a unique maximizer.

   (e) Observing that the likelihood is a function of the values of the distribution at the times of observation, pose the problem of determining the maximum likelihood estimators of these values as a suitable optimization problem. Should there be any constraint? [2+2+4+3+3=13]

4. The Product Limit estimate $\hat{S}(t)$ of the survival function, computed from right-censored survival times of 16 subjects, is

$$\hat{S}(t) = \begin{cases} 1 & \text{for } 0 \leq t < 2, \\ 0.9333 & \text{for } 2 \leq t < 3, \\ 0.7636 & \text{for } 3 \leq t < 4, \\ 0.6788 & \text{for } 4 \leq t < 5. \end{cases}$$

   Calculate the Nelson-Aalen estimator of the cumulative hazard function and the estimator of the survival function based on the Nelson-Aalen estimator. [7+3=10]

5. Consider failure time data with covariates (having no ties) following the relative risk regression model, and assume that the baseline hazard is piecewise constant over pre-determined intervals. Obtain an explicit estimator of the baseline hazard by maximizing the appropriate likelihood, after substituting the regression parameter with Cox's maximum partial likelihood estimator. Comment on the nature of this estimator when the intervals become finer. [7+2=9]

6. A total of n individuals working in a company are observed till they are fired, or up to a fixed time τ_i $(i = 1, \ldots, n)$, whichever is earlier. At the time of joining, the ith individual receives a random number A_i from the company, independently from others, which has the uniform distribution over the interval $[0, \tau]$, τ being a known number. For that individual, the hazard of being fired at time t is

$$\alpha_i(t) = \begin{cases} \alpha & \text{if } t \leq A_i, \\ \beta & \text{if } t > A_i. \end{cases}$$

The fixed times $\tau_1, \ldots, \tau_n$ are smaller than τ.

   (a) Define the counting processes of the number of observed firings by the company till time t as an aggregate of individual level counting processes, with appropriate filtration, stochastic intensity and the appropriate multiplicative intensity model. [Hint: There can be two predictable processes.]

   (b) Compute the expected value of the aggregate counting process.

   (c) Suggest a way of testing for $\alpha = \beta$. [4+8+3=15]

7. Describe three different ways of dealing with tied data while estimating the regression coefficients in the relative risk regression model. [9]

8. In a competing risks set-up with two causes of failure and no censoring, let T_1 and T_2 be the notional times to failure due to cause 1 and cause 2, respectively. Express the cause-specific hazard rate due to cause 1 in terms of the bivariate distribution of T_1 and T_2, and compare it with the marginal hazard of T_1. What happens when T_1 and T_2 are independent? [3+2=5]

9. Consider a cohort of male and female lives, where each life is observed for death due to natural causes (cause 1) and due to unnatural causes (cause 2). Censoring in the form of withdrawal or end of the observation window is also present. There is no covariate. One has to test whether the cause-specific hazard rate for the male is different from the cause-specific hazard rate for the females. Formulate this problem in terms of the relative risk regression model with a suitably chosen covariate, and indicate how you can devise a test using the partial likelihood. Show all the details of the test. [8]

10. Suppose you have uncensored failure time data with covariates, (T_i, Z_i), $i = 1, \ldots, n$, for which you would use the model

$$\log T = \alpha + Z_i'\beta + \sigma\epsilon,$$

where the errors ϵ_i, $i = 1, \ldots, n$, have a common distribution over the real line.

   (a) Explain what is meant by a linear rank test for $\beta = \beta_0$ in this context, and which cut-off can be used.

   (b) Derive the locally most powerful rank test for the hypothesis $\beta = \beta_0$ when the error distribution is $F(\epsilon) = (1 + e^{-\epsilon})^{-1}$.

   (c) Comment on the appropriateness of this test when the actual error distribution is different from the assumed the distribution.

   (d) Indicate how β may be estimated from a linear rank statistic with suitably chosen scores. [3+6+1+2 = 12]

INDIAN STATISTICAL INSTITUTE

Semestral Examination

Second semester

M. Stat - Second year 2014-2015

Stochastic Processes I

Date: May 15, 2015

Maximum Marks: 55

Duration: 3 hours 30 mins

Anybody caught using unfair means will immediately get 0. Please try to explain every step. You can refer to your class notes and exercises.

(1) Show that, with probability one, a standard, one dimensional Brownian motion changes sign infinitely many times in any time interval $[0, \varepsilon]$, $\varepsilon > 0$. [10 points]

(2) (a) Let B be a standard one dimensional Brownian motion and for every n, let $t_i = \frac{i}{n}$, $i = 0, \cdots n$. Prove that for every $p > 0$,

$$n^{\frac{p}{2}-1} \sum_{i=0}^{n-1} |B_{t_{i+1}} - B_{t_i}|^p$$

converges in probability to a constant v_p as $n \to \infty$.

(b) Prove that

$$n^{\frac{(p-1)}{2}} \left(\sum_{i=0}^{n-1} |B_{t_{i+1}} - B_{t_i}|^p - n^{1-\frac{p}{2}} v_p \right)$$

converges in law to a Gaussian random variable. [5+5=10 points]

(3) Let $(B_t)_{t \geq 0}$ be a standard one dimensional Brownian motion on $(\Omega, \mathcal{A}, P)$. Define for fixed $\omega \in \Omega$,

$$Z_\omega = \{0 \leq t < \infty : B_t(\omega) = 0\}.$$

Show that for P- a.e. $\omega \in \Omega$, the set Z_ω

(a) has Lebesgue measure zero.

(b) is closed and unbounded

(c) has an accumulation point at $t = 0$.

(d) has no isolated points in $(0, \infty)$.

[Hint: For (b) and (c) first show that the probability that the Brownian motion returns to origin infinitely often is one] [1+4+1+4=10 points]

(4) Let $(X_t, \mathcal{F}_t)$ be the Weiner process. Let $H_t = \int_0^t h(X_r)dr$ and $u(t,x) = E_x[\exp(H_t)]$ where h is a bounded function. Show that if $0 < s < t$,

$$E_x\left[\exp(H_t)|\mathcal{F}_s^+\right] = \exp(H_s)u(t-s, X_s).$$

[Refer to class notes for notation $\mathcal{F}_s^+$.] [5 points]

(5) Suppose $P_n \Rightarrow P$ on $\mathbb{R}$ and F is an uniformly bounded and equi-continuous family of real valued functions. Show that

$$\int f dP_n \to \int f dP$$

uniformly for all $f \in F$.

[5 points]

(6) Let B_t be a one-dimensional BM. Let $f(t) = 1 + \alpha\sqrt{t}$ where $\alpha > 0$ is fixed. Let $\tau_\alpha = \inf\{t \geq 0 : |B_t| = f(t)\}$. Show that τ_α is finite almost surely, but $E[\tau_\alpha]$ is finite for $\alpha < 1$ and infinite for $\alpha > 1$. [10 points]

(7) Let $(B_t)_{t \geq 0}$ be 2 dimensional Brownian motion starting at 0. Let $\mathcal{B}_t = \sigma(B_s : s \leq t)$. Let $B_t^{(1)}$ be the first coordinate of B_t. Show that $(B_t^{(1)}, \mathcal{B}_t)$ is a martingale. Show $\mathcal{B}_1 \neq \sigma(B_s^{(1)}, s \leq 1)$. [3+2=5 points]

INDIAN STATISTICAL INSTITUTE

Second Semestral Examination: 2014-15

M. Stat. II Year
Advanced Sample Survey

Date: 15.05.15 **Maximum Marks: 50** **Duration: 3 Hours**

Answer any 4 questions each carrying 10 marks.

Assignment records carry 10 marks.

1 Derive a suitable unbiased estimator for the mean square error of a homogeneous linear estimator for a finite population total examining whether it may be uniformly non-negative, developing requisite theorems.

2 Derive an unbiased estimator for the variance of Murthy's unbiased estimator for a finite population total based on a suitable sampling design required to be described by you.

3 Describe a 'generalized regression estimator' for a finite population total discussing its rationale and a method of assessing its accuracy. Show how it may match 'Brewer's predictor' explaining the motivation for its genesis.

4 Explain situations when and how Warner's and Simmons's randomized response techniques may be used to unbiasedly estimate the proportion of people bearing a stigmatizing feature explaining how you may assess the accuracies in estimation, employing appropriate varying probability sampling designs.

5 Explain and illustrate problems in developing Small Domain statistics. Discuss how you may tackle such problems.

*****

STATISTICAL METHODS IN BIOMEDICAL RESEARCH

M Stat 2nd (2014-15)

Semester Examination

Date: 18 .05.2015 Time: 2:30 hours

You can score maximum 70 marks

1. (a) What do you mean by inverse sampling? In the context of testing a binomial proportion, inverse sampling leads to a left tailed test for two-sided alternative – justify your answer clearly.
(b) If θ is the parameter measuring treatment difference, suggest a fixed-width confidence interval with proper justification on its optimality. How do you construct such an interval in case of partial sequential sampling?
(4+6+5+5=20)

2. (a) What do you mean by response-adaptive design for clinical trials?
(b) In the context of clinical trials, briefly explain the randomized play-the-winner (RPW) design.
(c) Based on dichotomous treatment assignments and dichotomous outcomes, propose a permutation statistic using RPW(α,1) rule clearly explaining the underlying steps. Mention the asymptotic distribution of the statistic with appropriate assumptions. (3+4+5+3=15)

3. (a) Consider a clinical trial in which patients are accrued sequentially and immediately are assigned to treatment *A* or treatment *B*. What is the main purpose for adaptive design in such a scenario?
(b) In the situation in (a), under randomized play-the-winner rule, show that the probability (p$_n$) that the n-th patient is assigned to treatment *A* is given by $p_n = \frac{1+S_A(n-1)+F_B(n-1)}{n+1}$, where $S_A(k)$ denotes the number of patients with outcome *S* on treatment *A* among the first *k* patients, and $F_B(k)$ denotes the corresponding number of patients with outcome *F* on treatment *B*. Also show that p_n converges to $\frac{q_B}{q_A+q_B}$ in probability as $n \to \infty$, where $q_A = P(Outcome\ F\ on\ treatment\ A)$ and $q_B = P(Outcome\ F\ on\ treatment\ B)$ where *S* and *F* are two kinds of outcome.
(c) Under (a) and (b) justify the statement: 'if treatment *A* is doing well relative to *B* early in the trial, more patients will tend to be placed on treatment *A*, and vice versa'. (4+6+5=15)

4. (a) Let $(X_1^A, X_1^B), (X_2^A, X_2^B), \ldots$ be independent where X_i^A and X_i^B are potential responses of the i-th patient to treatments *A* and *B* respectively. Assume that $X_i^A \sim N(\theta^A, 1)$ and $X_i^B \sim N(\theta^B, 1)$ for $i \geq 1$. It is desired to test

whether A is superior to B, i.e. whether $\theta = \theta^A - \theta^B$ is greater than 0. Then for $k \geq 1$ and $a, b \in (0, \infty)$, define,

$$\hat{\theta}_k = \frac{1}{m_k} \sum_{j=1}^{k} \delta_j X_j^A - \frac{1}{n_k} \sum_{j=1}^{k} (1 - \delta_j) X_j^B, \quad Z_k = \frac{m_k n_k}{k} \hat{\theta}_k, T_{a,b} = \inf \{k \geq 1 : Z_k > a \ or \ Z_k < -b\}$$

Derive a sequential probability ratio test of $\theta > 0$ based on T and Z_T.

(b) Assuming that there is one unit cost to administer either treatment and an additional ethical cost of $g(\theta)$ for assigning a patient to inferior treatment, allocate proportionally the patients to treatment A and B.

(c) Write down δ_k in Efron allocation scheme and rewrite the above test statistic clearly. (8+6+6=20)

5. Describe in brief Zelen's design in the context of clinical trials. Discuss its advantages and disadvantages. (4+6=10)

INDIAN STATISTICAL INSTITUTE

Backpaper Examination, Second Semester: 2014-15

M.Stat. II Year (AS)

Survival Analysis

Date: 12. 06.2015 Maximum marks: 100 Duration: 3 hours

Answer all questions. Standard notations are followed.

1. Show that for randomly right censored data arising from an underlying combination of notional failure and censoring times, the observed time (censored or uncensored) and the censoring indicator are independent if and only if the notional failure and censoring times have proportional hazards [10]

2. Given type-I censored data from the Weibull distribution, derive the score test for the hypothesis that the underlying distribution is, in fact, exponential. [10]

3. If Cox's regression model is postulated for the *reciprocal* of the time-to-event, a different regression model is obtained.

   (a) When does this model coincide with the accelerated failure time model? Explain.

   (b) Show that the usual theory of partial likelihood can be used for semiparametric regression analysis under this model, with left-censored data.

   (c) How will the partial likelihood change when another decreasing transformation of the times (instead of the reciprocal transformation) is used? Discuss the implications of your answer on inference based on the partial likelihood. [5+3+4=12]

4. For the relative risk regression model, explain how the counting process theory can be used to express the score vector as a special case of a stochastic integral of a predictable process with respect to a martingale, in an appropriate set-up. State the ensuing convergence result for the score process at a particular time, without proof, with all the requisite expressions. [12]

5. Let (t_i, Z_i), $i = 1, \ldots, n$ be independent observations following the exponential regression model $\lambda(t; Z) = \lambda e^{\beta' Z}$.

   (a) Show that the estimation problem for β is invariant under the group of scale transformations on the survival time, and that $t_2/t_1, \ldots, t_n/t_1$ are jointly marginally sufficient for β.

   (b) Obtain the marginal likelihood for β and compare it with the full likelihood partially maximized with respect to λ. [5+5=10]

6. Consider the illness-death model in continuous time with three states: Alive (A), Terminally ill (T) and Dead (D). Only three types of transition are possible: A to T, A to D and T to D. You want to compare the hazards of death in the A state for two groups of subjects. Describe a suitable mathematical set-up where this comparison can be made through a formal statistical test, the appropriate data and the form of the test. [3+2+2=7]

7. Consider possibly tied failure times data with covariates, for which you wish to estimate the baseline cumulative hazard function under the relative risk regression model.

   (a) Show that, for fixed values of the regression coefficients, the likelihood is maximized by putting probability masses of the baseline distribution at the distinct times of observed failure only.

   (b) Using the result of part (a) and the product integral representation

$$S(t;Z) = \mathcal{P}_0^t \, [1 - d\Lambda_0(u)]^{e^{\beta' Z(u)}} ,$$

   where

$$d\Lambda_0(u) = \begin{cases} -\dfrac{dS_0(u)}{S_0(u)} & \text{at points of continuity,} \\ 1 - \dfrac{S_0(u)}{S_0(u-)} & \text{at mass points,} \end{cases}$$

   reduce the problem to a finite dimensional optimization problem.

   (c) If Cox's maximum partial likelihood estimator is substituted in the objective function of part (b), and there are no ties, show that there is an explicit solution for the remaining parameter values.

[3+4+4=11]

8. Suppose you have uncensored failure time data with a single binary covariate, (T_i, Z_i), $i = 1, \ldots, n$, for which you would use the model

$$\log T = \alpha + Z_i'\beta + \sigma\epsilon,$$

   where the errors ϵ_i, $i = 1, \ldots, n$, have a common distribution over the real line.

   (a) For which distribution of the errors would the locally most powerful rank test for the hypothesis $\beta = 0$ reduce to the log rank test? Explain.

   (b) State, without giving reasons, how you would adjust the scores of the LMPRT if the data are randomly right censored.

   (c) Does the modified test of part (c) still correspond to the log rank test?

[6+2+2 = 10]

9. Consider a cohort of male and female lives, where each life is observed for death due to natural causes (cause 1) and due to unnatural causes (cause 2). Censoring in the form of withdrawal or end of the observation window is also present. There is no covariate. One has to test whether the cause-specific hazard rate for the male is different from the cause-specific hazard rate for the females. Formulate this problem in counting process set-up, and indicate how you can devise a test. Show all the details of the test. [10]

10. An investigation was undertaken into the effect of a new treatment on the survival times of cancer patients. Two groups of patients were identified. One group was given the new treatment and the other an existing treatment. The following model was considered:

$$h_i(t) = h_0(t) \exp\left(\beta^T z\right),$$

where $h_i(t)$ is the hazard at time t, where t is the time since the start of treatment, for the ith group ($i = 1, 2$)
$h_0(t)$ is the baseline hazard at time t
z is a vector of covariates such that:
$z_1 =$ sex (a categorical variable with $0 =$ female, $1 =$ male)
$z_2 =$ treatment (a categorical variable with $0 =$ existing treatment, $1 =$ new treatment), and
β is a vector of parameters, (β_1, β_2).

The results of the investigation showed that, if the model is correct, then (i) the risk of death for a male patient is 1.02 times that of a female patient, and (ii) the risk of death for a patient given the existing treatment is 1.05 times that for a patient given the new treatment.

(a) Estimate the value of the parameters β_1 and β_2.

(b) Estimate the ratio by which the risk of death for a male patient who has been given the new treatment is greater or less than that for a female patient given the existing treatment.

(c) Determine, in terms of the baseline hazard only, the probability that a male patient will die within 3 years of receiving the new treatment. [4+2+2=8]

STATISTICAL METHODS IN BIOMEDICAL RESEARCH

M Stat 2nd (2014-15)

Semester Examination (Back paper)

Date: 23/07/2015 Full marks: 100 Time: 2:30 hours

1. What do you mean by partial sequential sampling procedure? Discus Wolfe's procedure in this context. What is the statistic proposed by Wolfe? Explain the logic behind it. What is the asymptotic distribution of the test statistic under null hypothesis? (4+6+5+5=20)

2. When do you say that the sequence of test statistics has canonical joint distribution? Assuming large samples, show that, at the k-th stage for a group sequential testing method in a two-sample problem, the test statistic has canonical joint distribution. (8+8=16)

3. (a) Consider a clinical trial in which patients are accrued sequentially and immediately are assigned to treatment *A* or treatment *B*. What is the main purpose for adaptive design in such a scenario?

   (b) In the situation in (a), under randomized play-the-winner rule, show that the probability (p_n) that the n-th patient is assigned to treatment *A* is given by $p_n = \frac{1+S_A(n-1)+F_B(n-1)}{n+1}$, where $S_A(k)$ denotes the number of patients with outcome *S* on treatment *A* among the first *k* patients, and $F_B(k)$ denotes the corresponding number of patients with outcome *F* on treatment *B*. Also show that p_n converges to $\frac{q_B}{q_A+q_B}$ in probability as $n \to \infty$, where $q_A = P(Outcome\ F\ on\ treatment\ A)$ and $q_B = P(Outcome\ F\ on\ treatment\ B)$ where *S* and *F* are two kinds of outcome.

   (c) Under (a) and (b) justify the statement: 'if treatment *A* is doing well relative to *B* early in the trial, more patients will tend to be placed on treatment *A*, and vice versa'. (8+12+8=28)

4. (a) What would be the sample size to achieve 80% power to detect a difference in efficacy of mirtazapine (new drug) and sertraline (standard drug) for the treatment of resistant depression in 6-week treatment duration? Assume that outcome measure if dichotomous. Cleary derive all steps.

   (b) Assuming the treatment measure is continuous, calculate the sample size to get 80% power in detecting significant efficacy of ACE II antagonist (new drug) and ACE inhibitor (standard drug) for the treatment of primary hypertension. Note the fact that change of sitting blood pressure (SDBP, mmHg) is the primary measurement compared to baseline. Other assumptions must be mentioned clearly and all steps must be derived clearly. (18+18=36)