Indian Statistical Institute
Second Semestral Examination : (2005-2006) B.Stat(Hons.) III
Statistical Inference II
Maximum Marks: 100 Date: 9.5.06 Duration :- $3\frac{1}{2}$ hours

This paper contains questions worth a total of 110 points. Answer as much as you can. The maximum you can get is 100 points.

1.(a) Suppose $X_1, \dots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ where both μ and σ are unknown. Find the $100(1-\alpha)\%$ confidence interval based on a suitable pivot T which has the shortest **expected length** among all $100(1-\alpha)\%$ confidence intervals based on T. Clearly argue why this interval has the shortest expected length.
(b) State and prove the Ghosh-Pratt equation connecting the expected length of a confidence interval and the probabilities of false coverage. [7+7=14]

2. Let X_i, $i = 1, 2, \dots$ be i.i.d. Bernoulli(θ), where $0 < \theta < 1$. Consider the SPRT for testing $H_0 : \theta = \frac{1}{3}$ against $H_1 : \theta = \frac{1}{2}$ where the boundaries satisfy $0 < B < 1 < A < \infty$. Show, **without using Stein's lemma**, that the SPRT terminates with probability one under $\theta = \frac{2}{3}$. [10]

3.(a) Consider two small positive numbers α and β with $0 < \alpha + \beta < 1$. Consider Wald's SPRT for simple hypotheses with target strength (α, β), in the case of i.i.d. observations. Find approximate expressions for average sample number (ASN) under H_0 and H_1 using Wald's approximations.
(b) Suppose X_i, $i = 1, 2, \dots$ are i.i.d. observations having exponential distribution with density

$$f_\sigma(x) = \frac{1}{\sigma} e^{-\frac{x}{\sigma}},\ x > 0.$$

Consider the SPRT with target strength $(\alpha = 0.05, \beta = 0.05)$ for testing $H_0 : \sigma = 1$ against $H_1 : \sigma = 2$, where Wald's approximations for the boundaries are used. Compare the approximate average sample numbers under H_0 and H_1 with the minimum sample size required by the MP test for testing H_0 against H_1 with error probabilities at most α and β. (You can use (i) normal approximation to the chi-square distribution, (ii) $P(Z \le 1.645) \approx 0.95$ where $Z \sim N(0,1)$, (iii) $\ln(19) \approx 2.94$ and (iv) $\ln(2) \approx 0.70$.) [5+12=17]

4. Let $\{X_i\}_{i \ge 1}$ be a sequence of random variables and the joint density of $(X_1, \dots, X_m)$ be $p_{jm}(x_1, \dots, x_m)$ under hypothesis H_j, $j = 0, 1$. Let $0 < B < 1 < A < \infty$ and consider the SPRT of strength (α, β) that stops first time $\lambda_m = \frac{p_{1m}}{p_{0m}} \ge A$ or $\le B$. H_0 is rejected if $\lambda_n \ge A$, accepted if $\lambda_n \le B$ and no decision is made if $n = \infty$, where n is the stopping time of the test. Assume $\alpha > 0$ and $\beta > 0$.

Will the Wald inequalities connecting A, B, α and β remain true ? Justify your answer. (Note that here we are not assuming that the X_i's are i.i.d. and it is not guaranteed that the SPRT terminates with probability 1 under H_0 or H_1.) [10]

5. Let $X_1, X_2, \dots,$ be i.i.d $N(\mu, \sigma^2)$, where both μ and σ are unknown. Derive a confidence interval of confidence coefficient at least $1-\alpha$ for μ, of

P.T.O

pre-specified length $2l$ using Stein's two-stage procedure. [10]

6. Let $X_1, X_2, \ldots, X_n$ be i.i.d. F, where F is a continuous distribution. Consider the Kolmogorov-Smirnov statistic $D_n = \sup_x |F_n(x) - F(x)|$, where $F_n(.)$ is the empirical distribution function of the sample.
(a) Show that the distribution of D_n is independent of F, i.e. it remains the same if the sample is generated from any other continuous distribution.
(b) For $n = 2$, find $P(D_n < \frac{1}{4} + v)$ for $-\infty < v < \infty$. [6+8=14]

7. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. F where F is a bivariate distribution and the random variables X and Y are continuous and independent under F. Derive the asymptotic distribution of (suitably normalized) Kendall's sample tau coefficient. [12]

8. Two judges A and B rank 5 competitors as $\{5, 3, 1, 2, 4\}$ and $\{5, 2, 3, 1, 4\}$ respectively, where the i-th element in each vector refers to the ranking of the i-th competitor by the respective judge.
Compute 2 nonparametric measures of association between the two judges' rankings. [10]

9. let $X_1, \ldots, X_n$ be i.i.d. $F(x - \theta)$ where F is a continuous distribution having a continuous density f which is symmetric around 0 and $f(0) > 0$ and $\text{Var}_F(X_i) = {\sigma_F}^2 < \infty$. We want to test $H_0 : \theta = 0$ versus $H_1 : \theta > 0$. Derive a general expression for the asymptotic relative efficiency of the sign test relative to the t test in this problem and evaluate it when (i) $F = N(0, 1)$ and (ii) $F =$ Double exponential with density $f(x) = \frac{1}{2}e^{-|x|}$, $-\infty < x < \infty$. [9+4=13]

INDIAN STATISTICAL INSTITUTE

2005-06 I Semester
B.Stat. (Hons.) 3rd year

Linear Statistical Models

Mid-semestral Examination

/

Maximum time: 2 hours | September 9, 2005 | Maximum marks: 100

Answer as much as you can. The entire question paper is worth 110 marks, but the maximum you can score is 100.

1. Consider the piecewise linear model

$$y = \begin{cases} \alpha_0 + \alpha_1 x + \epsilon & \text{if } x \le x_0, \\ \beta_0 + \beta_1 x + \epsilon & \text{if } x > x_0. \end{cases}$$

Show that if x_0 is known, this model can be rewritten as a linear model — with a suitable choice of explanatory variables. Find a restriction on the parameters of this model which ensures that the expectation of y given x is continuous in x. Suggest another linear model with no constraint, which would be equivalent to the above model with continuity constraint on the parameters. [5+5+5=15]

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

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

where β_1, β_2 and β_3 are unknown parameters and $\epsilon_1, \ldots, \epsilon_6$ are uncorrelated errors with mean 0 and variance σ^2.

(a) Describe the set of all estimable linear functions of β_1, β_2 and β_3.
(b) Describe the set of all linear zero functions of the model.
(c) Describe the set of linear functions of $y_1, \ldots, y_6$ which are best linear unbiased estimators of their respective expectations.
(d) Identify the best linear unbiased estimator of $E(y_1)$.
(e) What is the variance of the best linear unbiased estimator of $E(y_1)$?
(f) Find an unbiased estimator of σ^2. [3+6+6+3+3+4=25]

3. Consider the hypothesis $\beta \propto b$, where b is a specified vector. Show that it can be reformulated as a *linear* hypothesis. Construct the ANOVA table and explain very clearly how the generalized likelihood ratio test statistic can be computed using a standard statistical package for linear regression. [8+5+7=20]

P.T.O

4. Consider the the linear model $(y, X\beta, \sigma^2 I)$ having n observations and k components of the parameter β.

   (a) Determine the average value of the n leverages.

   (b) If the vector $\mathbf{1}$ is included in $C(X)$, show that the ith leverage can be written as $1/n$ plus a quantity that can be interpreted as a squared distance of the ith row of X from the average of all the rows of X. ~~Interpret~~ this result. & Explain the conse

   (c) If h_i and e_i are the ith leverage and ith residual, respectively, show that

$$0 \leq h_i + e_i^2/e'e \leq 1.$$

   Interpret this result. [2+(5+1)+(5+2)=15]

5. Consider the two-way classified data model $(y, X\beta, \sigma^2 I)$ with

$$X_{40\times 5} = \begin{pmatrix} 1_{10\times1} & 1_{10\times1} & 0_{10\times1} & 1_{10\times1} & 0_{10\times1} \\ 1_{10\times1} & 0_{10\times1} & 1_{10\times1} & 1_{10\times1} & 0_{10\times1} \\ 1_{10\times1} & 1_{10\times1} & 0_{10\times1} & 0_{10\times1} & 1_{10\times1} \\ 1_{10\times1} & 0_{10\times1} & 1_{10\times1} & 0_{10\times1} & 1_{10\times1} \end{pmatrix}, \quad \beta_{5\times1} = \begin{pmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \tau_1 \\ \tau_2 \end{pmatrix},$$

   where the parameter μ represents a general effect, β_1 and β_2 represent blocks effects and τ_1 and τ_2 represent the respective effects of two treatments. Consider the hypothesis $\mathcal{H}_0 : \tau_1 = \tau_2 = 0$.

   (a) Show that the hypothesis $\mathcal{H}_0$ is only partially testable.

   (b) Suggest two hypothesis $\mathcal{H}_{01}$ and $\mathcal{H}_{02}$ such that (i) their intersection is equivalent to $\mathcal{H}_0$, (ii) $\mathcal{H}_{01}$ is testable and (iii) $\mathcal{H}_{02}$ is completely untestable. Prove that (i), (ii) and (iii) hold in this case.

   (c) A statistician obtains the error sum of squares of the given model (R_0^2) and the error sum of squares (R_H^2) of the simplified model where the parameters τ_1 and τ_2 and the corresponding columns of X have been eliminated. He seeks to test $\mathcal{H}_0$ by using the test statistic $[(R_H^2 - R_0^2)/2]/[R_0^2/3]$ with null distribution $F_{2,3}$. Is this procedure correct? If it is, justify the procedure. If it is not, explain and suggest a correct procedure. [3+9+8=20]

6. You are given data of the type (x_i, y_i), $i = 1, \ldots, n$ on the response variable y and the explanatory variable x. Using the simple linear regression model of y on x with normal errors, give explicit expressions for the following. (Make sure that your expressions are simplified as much as possible for the specific model at hand.)

   (a) Simultaneous confidence intervals for the two regression coefficients including intercept with minimum coverage probability .95.

   (b) A confidence region for the 'true' regression line with minimum coverage probability .95.

   (c) Simultaneous prediction intervals for two future observations corresponding to x-values x_{n+1} and x_{n+2}, respectively, with minimum coverage probability .95.

   (d) Tolerance interval of observations $y_{n+1}, \ldots, y_{n+100}$ corresponding to the common x-value x_0, with the target of capturing 90% of the observations, and minimum coverage probability .95. [3+4+4+4=15]

INDIAN STATISTICAL INSTITUTE

Mid-semester Examination: (2005-2006)

Name of the Course: Introduction to Sociology

B. Stat.-III yr.

Date: 12.9.05 Maximum Marks: 30 Duration: 1½ hours

Instruction: Write answers for group–A & Group-B in separate answer sheets.

Group-A

Q. 1) Identify which of the following statements are true or false: ½ x 10 = 5

a) Talcott Parsons attempted to conceptually integrate the personality system with cultural system, but failed to incorporate personality system with social system.

b) The 'positive' society, according to Auguste Comte, denotes the industrial society.

c) L. Coser is one of the main proponents of 'middle range theories'.

d) 'Mechanical solidarity', according to Emile Durkheim, denotes the solidarity in primitive societies.

e) Max Weber has a great contribution in sociological methodology.

f) The rate of 'Egoistic suicide' increases, according to Emile Durkheim, because of over-attachment of individual to society.

g) Ralph Dahrendorf's conflict theory finds close affinity with Karl Marx's theoretical propositions on conflict in society.

h) Legal system, according to Karl Marx's proposition, is a good example of 'super-structure' in society.

i) Herbert Spencer was a great proponent of 'individualism'.

j) 'Social fact' can only be comprehended through 'sympathetic introspection'.

Q. 2) Answer any two:

a) How would you define 'social action'? What according to Max Weber, are the classifications of social actions? 2+3 = 5

P.T.O

b) Which sociological variables, according to Emile Durkheim, are responsible for the variation in the rate of suicide? Does the possibility of 'altruistic' suicide remains high in tradition bound societies? – Explain briefly. 2+3 =5

c) What, according to Talcott Parsons, are the 'functional requisites' of a social system? What are the meanings of those functional requisites? 2+3 =5

Group-B

Q.1. Answer any two of the following:

(a) What is sociological research? Point out the scope of sociological research? 2+3 = 5

(b) What are the essentials steps in a research procedure in sociological study? Explain briefly. 3+2 = 5

(c) What is a questionnaire? What are its different forms? If you want to survey on illiterate group of persons, what kind of questionnaire would you use? 2+1+2 = 5

Q.2. Write short notes on: 2½ + 2½ = 5

(a) Observation.

(b) Experiment.

B. Stat. III years
Introduction to Anthropology and Human Genetics
Semester I (2005- 2006)

Time 2 hours Date: 12.9.05 Full marks 30

(A) Attain any five of the following questions. 5X2= 10

1. What is diastema?
2. What do you mean by knuckle- walking?
3. What are the different varieties of *Australopithecine* fossils?
4. What are the major changes in human brain in course of evolution?
5. What are the probable causes of human genetic diversity of Indian population?
6. What is dermatoglyphics trait?
7. Define Hardy- Weinberg Law?

(B) Attain any two of the following questions.

1. What is Anthropology? What are the different branches of Anthropology? How are they interrelated to the study of human beings? (1+1+3= 5)
2. What do you mean by genetic polymorphism? What are the different types of genetic polymorphism in human population? Why 'HbS' is called a balanced polymorphic trait? (1+2+2= 5)
3. What are the major changes occurred in human lower limb due to erectus posture? (5)

(C) Attain all of the following questions.

1. What do you mean by coefficient of relationship? How it's differing from coefficient of inbreeding? Estimates coefficient of relationship and coefficient of inbreeding with a hypothetical diagram. (1+1+2= 4)
2. In a sample, the Shabar tribal population (n= 175), observed ABO blood group types were A= 50; B= 30; AB= 20 and O= 75. Estimates the gene frequencies. (4)
3. The following numbers of human M- N blood groups were recorded in a sample of American Whites.

M	MN	N
1787	3039	1303

What are the gene frequencies? (2)

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination : (2005 –2006)

B. Stat (Hons) III Year

Economics III

Date: 12 September 2005 Maximum Marks: 30 Duration: 2 Hours

Note: This paper carries 32 marks. Answer as many questions as you like.
The maximum you can score is 30.

1. What are the important properties of the (i) minimum norm, (ii) least squares and (iii) Moore - Penrose G - inverses?

   Compute the G - inverse for $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 3 & 3 & 3 \end{pmatrix}$. [4 + 4 = 8]

2. (a) Define unbiased estimator, minimum variance estimator and consistent estimator.

   (b) Show that if $X_1, X_2, \ldots, X_n$ are i.i.d. observations coming from $F(\mu, \sigma^2)$ with $E(X_i) = \mu$ and $V(X_i) = \sigma^2$ then $T_n = \overline{X}_n + 1/\sqrt{n}$ is a consistent estimator for μ.

   Compute $E(T_n)$ and $V(T_n)$. [3 + 3 + 2 = 8]

3. State all the assumptions of 'classical linear regression model'. Define the goodness of fit measure, R^2. Show that corr $(Y, \hat{Y}) = \sqrt{R^2}$, $Y = X\beta + \varepsilon$, $\hat{Y} = X\hat{\beta}_{ols}$. All notations are self-explanatory. [3 + 1 + 4 = 8]

4. What is non-spherical disturbance? In such context what is the need for GLS? Analytically show that GLS is more efficient than OLS. [2 + 2 + 4 = 8]

INDIAN STATISTICAL INSTITUTE

MID-SEMESTRAL EXAMINATION (2005-2006)

B. Stat. III Year

GEOLOGY

Date: 12 September 2005 Duration: 3 hours Total Marks: 100

Answer question no. 1 and any four from the following questions

1. a) Discuss briefly a coherent hypothesis about the origin of the solar system.
   b) Point out the merits and demerits of this theory with respect to the "Collisional Hypothesis".
   c) What would be approximate age of the solar system?

10+8+2=20

2. a) What is a mineral? Name one important mineral from each of the following group: i) oxide, ii) silicate, iii) sulphide and iv) carbonate.
   b) What is the cleavage of a mineral? What is the difference in the cleavage sets of feldspar and amphiboles?
   c) Arrange the following minerals according to Moh's Scale of Hardness: Feldspar, Gypsum, Flourite, Corundum, Calcite, Diamond.
   d) Describe with figures and example the different silicate structures which can be identified in the minerals.

4+3+3+10=20

3. a) What is a fossil?
   b) In which kind of rock one usually gets fossils?
   c) Why mostly the hard parts of the organisms are preserved as fossils?
   d) "Minerals present in a fossilized skeleton generally differ from that of the original skeleton". Explain.
   e) How would you qualify a fossil as a "guide" fossil?
   f) Why "Formation" is mostly used in lithostratigraphy as working unit?
   g) What is a biozone? What are the different types zones used in biostratigraphy?

2+1+2+4+4+2+5=20

4. Write short notes on any <u>four</u> of the following:

a) Big Bang theory
b) Law of Superposition
c) Rock cycle
d) Isostasy
e) Bowen's reaction series
g) Radioactive dating of rocks

4x5=20

P.T.O

5. a) What is a rock?
b) What are the basic differences between granite and basalt?
c) What do you mean by porphyritic texture? Give a suitable sketch.
d) Give examples of a coarse grained ultramafic rock and a fine-grained acid rock.
e) Name one clastic and one non-clastic sedimentary rocks.
f) What are the major textural elements of a clastic sedimentary rock?
g) What is the most common igneous rock of the continent?
h) Name the metamorphic equivalent of the following rocks: i) sandstone, ii) granite

2+4+2+2+2+3+1+4=20

6. a) How can you explain the presence of earth's magnetic field?
b) Does moon and have a magnetic field?
c) How can you use the presence of earth's magnetic field to show that the continents have changed their position in the past?

8+2+10

7. a) What causes an earthquake?
b) What are the different kinds of earthquake waves?
c) How can you infer different types of motion along the fault surfaces from the study of the seismogram?
d) What geological evidences suggest that mountains have underlying roots?

4+3+5+8=20

INDIAN STATISTICAL INSTITUTE
Mid- Semester Examination: 2005-06
B. Stat. III Year
Differential Equations

Date: 14.9.05 Maximum Marks: 50 Duration: 3 Hours

STATE ANY RESULT THAT YOU USE.

1. Let $p(x)$ and $q(x)$ be continuous on $[a,b]$ and $y_1(x), y_2(x)$ be any two solutions of $\frac{d^2y}{dx^2}+p(x)\frac{dy}{dx}+q(x)y=0$ on $[a,b]$. Prove that the Wronskian of $y_1(x)$ *and* $y_2(x)$ is either identically zero or is never zero on the interval $[a,b]$.

[10]

2. Verify exactness of

$(Sin\,x\;Sin\,y - xe^{y})dy = (e^{x} + Cos\,x\;Cos\,y)dx$ and hence solve the equation.

[8]

3. Find the general solution of

$y'' - 3y' + 2y = 14\,Sin\,2x - 18\,Cos\,2x$ by the method of undetermined coefficients.

[8]

4. Verify that $y_1 = e^{x}$ is a solution of $xy'' - (2x+1)y' + (x+1)y = 0$, hence find the general solution.

[8]

5. Find a particular solution of $y'' - y' - 6y = e^{-x}$ by variation of parameters method.

[8]

6. Find the general solution $y(x) = \sum_{n\geq 0} a_n x^n$ of $y'' + xy' + y = 0$ of the form $y(x) = a_0 y_1(x) + a_1 y_2(x)$, where $y_1(x)$ and $y_2(x)$ are power series.

What is the radius of convergence for the solution?

[8]

***

INDIAN STATISTICAL INSTITUTE
Mid-semester Examination : 2005-2006
B.Stat. (Hons.) III Year
Sample Surveys

Date : 16. 09 .2005 Maximum Marks : 100 Duration : 3 Hours

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

1. After the decision to take an SRSWOR sample had been made , it was realized that y_1 would be unusually low and y_N would be unusually high . Consider the following estimator of $\overline{Y}$

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

where c is a constant .

(a) Show that $\hat{\overline{Y}}_S$ is unbiased for $\overline{Y}$ with

$$\text{Var}(\hat{\overline{Y}}_S) = (1\text{-f})\left[\frac{S^2}{n} - \frac{2c}{(N-1)}(y_N - y_1 - nc)\right]$$

where $f = \frac{n}{N}$ and S^2 is the population variance with divisor (N-1) .

(b) Show that

$$\text{Var}(\hat{\overline{Y}}_S) < \text{Var}(\bar{y}) \text{ if } 0 < c < (y_N - y_1)/n.$$

(5 + 15 + 5) = [25]

2. Two dentists A and B make a survey of the state of the teeth of 200 children in a village . Dr.A selects a simple random sample of 20 children and counts the number of decayed teeth for each child , with the following results .

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

Dr.B using the same dental techniques , examines all 200 children , recording merely those who have no decayed teeth . He finds 60 children with no decayed teeth .

P.T.O.

Estimate the total number of decayed teeth in the village children , (a) using A's results only , (b) using both A's and B's results . (c) Are the estimates unbiased ? (d) Which estimate do you expect to be more precious ?

(10 + 10 +3 + 2) = [25]

3. If the cost function is of the form $C = C_0 + \sum t_h \sqrt{n_h}$ where C_0 and t_h are known numbers , show that in order to minimize Var ($\bar{y}_{st}$) for fixed total cost n_h must be proportional to

$$\left(\frac{W_h^2 S_h^2}{t_h} \right)^{2/3} .$$

Find the n_h for a sample of size 1000 under the following conditions .

Stratum	W_h	S_h	t_h
1	0.4	4	1
2	0.3	5	2
3	0.2	6	4

(15 + 10) = [25]

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

Compare the precision of a 1-in-8 systematic sample with a simple random sample of the same size for estimating the proportion of households in which the head is nonwhite . [25]

5. The analysis of variance of a population of 340 villages divided into 4 unequal strata is given in the following table . Calculate the efficiency of stratification with proportional allocation and with SRSWR in each stratum as compared to unstratified sampling for estimating the area under wheat .

Source of variation	Degrees of freedom	Sum of squares	Mean square
Between strata	3	$\sum_h N_h(\bar{Y}_h - \bar{Y})^2$	5400
Within strata	336	$\sum_h \sum_i (y_{hi} - \bar{Y}_h)^2$	24
Total	339	$\sum_h \sum_i (y_{hi} - \bar{Y})^2$	71.58

[25]

6. (a) Describe how you would select a PPSWR sample in n draws by Lahiri's method . Show that a sample selected according to this method is really a PPS sample .

(b) If N is not a multiple of n , what are the shortcomings of linear systematic sampling ? Describe how you can either modify the sampling procedure or suggest a suitable method of estimation so as to get rid of the shortcomings in case N is not a multiple of n .

(13 + 12) = [25]

INDIAN STATISTICAL INSTITUTE

STATISTICAL INFERENCE I

Mid Semester Exam, BStat 3rd Yr, 2005-2006

Total points 112. Maximum you can score is 100.

Date: September 5, 2005 Time: 3 hours

1. Let a real valued random variable X have a one parameter exponential family distribution where the density function $f_\theta(x)$ can be written as

$$f_\theta(x) = h(x)exp\{T(x)\eta(\theta) - B(\theta)\}.$$

(a) Suppose a random sample $X_1, X_2, \ldots, X_n$ of size n is drawn from this distribution. Find the minimal sufficient statistic for the above probability model.

(b) Show that the minimal sufficient statistic itself has a probability distribution which belongs to the exponential family. [It is fine if you just prove it for the discrete case where the proof is simpler. The result, however, is equally true for a continuous distribution].

(c) Suppose $X_1, X_2, \ldots, X_n$ is a sample from a population with density $f_\theta(x) = (x/\theta^2)exp(-x^2/2\theta^2)$; $x > 0, \theta > 0$. This is called the Rayleigh distribution. Find the mean and variance of $\sum_{i=1}^n X_i^2$. [Hint: Use part (b) with other exponential family results]. [4+8+10=22 points]

2. Let $\boldsymbol{\theta} = (\theta_1, \theta_2)$ be a bivariate parameter. Given a random sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, suppose that $(T_1(\boldsymbol{X}),\ T_2(\boldsymbol{X}))$ are two statistics. Find an example where $(T_1(\boldsymbol{X}), T_2(\boldsymbol{X}))$ are jointly sufficient for $\boldsymbol{\theta}$, $T_1(\boldsymbol{X})$ is sufficient for θ_1 whenever θ_2 is fixed and known, but $T_2(\boldsymbol{X})$ is not sufficient for θ_2 when θ_1 is fixed and known. [18 points]

3. Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution with success probability p. Find the minimum varianced unbiased estimator for p. [Hint: Use the Rao-Blackwell theorem]. [12 points].

4. Suppose that n integers are drawn at random and with replacement from the integers 1, 2, $\ldots$, N. Here N is unknown, but you can use the fact that the variance of the largest sampled

P.T.O

.teger is approximately

$$\frac{nN^2}{(n+1)^2(n+2)}.$$

(a) Find the methods of moment estimator $\hat{N}_1$ for N.

(b) Find $E(\hat{N}_1)$ and $Var(\hat{N}_1)$.

(c) Find the MLE $\hat{N}_2$ of N.

(d) Show that $E(\hat{N}_2)$ is approximately $[n/(n+1)]N$. Adjust $\hat{N}_2$ to form an estimator $\hat{N}_3$ whi is approximately unbiased for N.

(e) Find an approximate value for $Var(\hat{N}_3)$

(f) Show that for $n > 1$ and large N, $Var(\hat{N}_3)$ is expected to be smaller than $Var(\hat{N}_1$

[5+5+5+5+5+5=30 points]

5. Let θ be a real valued parameter. We shall say that a loss function is convex if $L(\theta, \alpha a_0 + (1-\alpha)a_1) \leq \alpha L(\theta, a_0) + (1-\alpha)L(\theta, a_1)$, for any a_0, a_1, θ. Suppose that there is an unbiased estimator δ for θ, and let $T(\boldsymbol{X})$ be sufficient. Show that if $L(\theta, a)$ is convex, and if $\delta^*(\boldsymbol{X}) = E(\delta(\boldsymbol{X})|T(\boldsymbol{X}))$, then $R(\theta, \delta) \geq R(\theta, \delta^*)$, where $R(\theta, \delta)$ represents the risk function for $L(\theta, \delta)$. [20 points]

INDIAN STATISTICAL INSTITUTE
Mid-Semester Examination: November 2005
B. Stat. III Year
Linear Statistical Models

Date: 22.11.2005 Full Marks: 100 Time: 3 hrs.

Answer all the Questions.

1. Let y be a random variable and x be a random vector with $E\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$ and $D\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} V_{xx} & v_{xy} \\ v'_{xy} & v_{yy} \end{pmatrix}$. Then show that the best linear predictor $\hat{E}(y|x)$ satisfies the following :

a) $\hat{E}(y|x) = \mu_y + v'_{xy} V^{-}_{xx}(x - \mu_x)$

b) $y - \hat{E}(y|x)$ is uncorrelated with every linear function of x.

c) $E\left[\hat{E}(y|x)\right] = \mu_y$.

[3×14=42]

2.(a) In the linear model $(y, X\beta, \sigma^2 I)$ show that $p'\beta$ is estimable if and only if it is identifiable.

(b) Show that $X = X(X'X)^{-}X'X$, where $(X'X)^{-}$ is a g-inverse of $(X'X)$.

[2×14=28]

3. In a small scale regression study, the following data were obtained.

i	1	2	3	4	5	6
x_{i1}	7	4	16	3	21	8
x_{i2}	33	41	7	49	5	31
y_i	42	33	75	28	91	55

Assuming a multiple linear regression model with independent normal error terms, obtain

(a) estimates of the regression coefficients and their standard errors ;

(b) Perform analysis of variance test to test the significance of the regression coefficients ;

(c) $\hat{y}$ at $x_1 = 10$ and $x_2 = 30$ and its standard error.

[16+8+6=30]

***

INDIAN STATISTICAL INSTITUTE
First Semester Examination: 2005-06
B. Stat. (Hons.) III Year
Linear Statistical Models

Date: 24.11.2005 | Maximum Marks: 100 | Duration: $3\frac{1}{2}$ Hours

Answer as many Questions as you can. Maximum you can score is 100.

1. Show that in the linear model $\begin{pmatrix} y, & X\beta, & \sigma^2 I \\ n\times 1 & n\times k\ k\times 1 & k\times k \end{pmatrix}$ the linear statistic $\ell' y$ is

a) a linear unbiased estimator of the linear parametric function $p'\beta$ iff $X'\ell = p$,

b) a linear zero function iff $X'\ell = 0$ i.e. ℓ is of the form $(I - P_X)\, m$ for some vector m, where $P_X = X(X'X)^- X'$.

[5+5=10]

2. Let $A\beta = \xi$ be a consistent restriction on the model $(y, X\beta, \sigma^2 I)$. Then show that

a) all estimable linear parametric functions of the unrestricted model are also estimable under the restricted model ;

b) all linear zero functions of the unrestricted model are linear zero functions under the restricted model ;

c) the restriction can only reduce the dispersion of the BLUE of $X\beta$;

d) the restriction can only increase the error sum of squares.

[5x4=20]

3. Show that in the linear model $(y, X\beta, \sigma^2 I)$, if $A\beta$ is estimable and $\hat{\beta}$ is any least squares estimator of β, then the BLUE of $A\beta$ is given by $A\hat{\beta}$.

[12]

4. Show that the matrix $I - X(X'X)^- X'$ is idempotent where $(X'X)^-$ is a g inverse of $(X'X)$.

[8]

5. Explain the following terms

a) Michaelis – Menten model
b) Collinearity between columns of a design matrix
c) Logistic regression
d) Probit regression.

[3x4=12]

P.T.O

6. Let y_1, y_2, y_3, y_4, y_5 five observations for which we assume the following linear model :

$$y_1 = \theta_1 - \theta_2 + \theta_3 + e_1 ,$$
$$y_2 = \theta_1 + 2\theta_3 + e_2 ,$$
$$y_3 = \theta_2 + \theta_3 + e_3 ,$$
$$y_4 = 2\theta_1 - \theta_2 + 3\theta_3 + e_4 ,$$
$$y_5 = \theta_2 + \theta_3 + e_5 ,$$

where $\theta_1, \theta_2, \theta_3$ are fixed but unknown effects, and e_i's are random errors assumed to be i.i.d. $N(0, \sigma^2)$.

a) Show that $C_1\theta_1 + C_2\theta_2 + C_3\theta_3$ is estimable if and only if $2C_1 + C_2 - C_3 = 0$.

b) Obtain a set of least squares estimators for θ_1, θ_2 and θ_3 and use them to obtain an unbiased estimator for σ^2.

c) Show that $H_0 : \theta_2 + \theta_3 = 0$ is testable. Obtain an appropriate test statistic in terms of $\{y_i\}$ for testing this hypothesis.

[5 X 3 =15]

7. a) Explain when and how you would use analysis of covariance to fit a linear model to a data set.

b) Develop analysis of covariance for one way classified data with one concomitant variable. Give the analysis of variance and covariance table indicating the details of the computations involved.

[5+10=15]

8. The following data shows the observations from 12 days' operation of a plant for the oxidation of ammonia, which is used for producing nitric acid. The response variable (y) is the 'stack loss' defined as the percentage of ingoing ammonia that escapes unabsorbed. The explanatory variables are cooling water inlet temperature (x_1) and acid concentration in percentage (x_2). Using a homoscedastic linear regression of y on x_1 and x_2 obtain the BLUE's of the regression coefficients alongwith their standard errors. Also give a measure of the goodness of your fit.

Obs.	1	2	3	4	5	6	7	8	9	10	11	12
x_1	27	27	25	24	22	23	24	24	23	18	18	17
x_2	58.9	58.8	59.0	58.7	58.7	58.7	59.3	59.3	58.7	58.0	58.9	58.8
y	4.2	3.7	3.7	2.8	1.8	1.8	1.9	2.0	1.5	1.4	1.4	1.3

[20]

*****

INDIAN STATISTICAL INSTITUTE
First Semester Examination: 2005-06
B. Stat. III Year
Differential Equations

Date: 28.11.05 Maximum Marks: **50** Duration: 3 Hours

Answers all the questions.

1. State and prove Sturm Separation theorem. [8]

2. Prove that positive zeros of $J_p(x)$ and $J_{p+1}(x)$ occur alternately, where $J_p(x)$ is the Bessel function

$$J_p(x) = \sum_{n=0}^{\infty} (-1)^n \frac{\left(\frac{x}{2}\right)^{2n+p}}{n!(p+n)!}, \quad p > 0, x > 0$$

[8]

3. Let $\{a_n\}$ be the positive zeros of $J_p(x)$ for some fixed $p > 0$. Prove that $\{\sqrt{x}J_p(a_n x)\}$ is an orthogonal family. [10]

4. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a function that satisfies $f(0) = 1, f'(0) = 0$ and $f''(x) = (x^2 - 1)f(x)$. Show that $f(x) \to 0$ as $x \to \infty$. [6]

5. Solve the initial value problem $y'' + 2y' + 3y = 3e^{-x}\sin x, y(0) = 0, y'(0) = 3$ by the method of Laplace transform. [6]

6. Suppose $\psi_1(t) = t, \psi_2(t) = t + e^t$ and $\psi_3(t) = 1 + t + e^t$ are three solutions of a second order non-homogeneous linear differential equation. Find the general solution. [6]

7. Solve : $(x^3 + xy^4)dx + 2y^3 dy = 0$ [6]

***

INDIAN STATISTICAL INSTITUTE

FINAL SEMESTRAL EXAMINATION (2005-2006)

B. Stat. III Year

GEOLOGY

Date: 2.12.2005 Duration: 3 hours Total Marks: 100

Answer any five of the questions

1. Write short notes on <u>any four</u>:
a) Event Stratigraphy
b) Precambrian Life
c) Explosive Volcanism
d) Metamorphic zones
e) Schistosity
f) Silicate Minerals

5x4=20

2. **a**) What is the scientific name of the Jurassic dinosaur which is mounted in the Geology Museum of Indian Statistical Institute? **b**) "Among the innumerable number of organisms on the surface of the earth, only a minute fraction of them are available as fossils" – why? **c**) What are the utilities of studying fossils? **d**) What are the main points of Darwin's theory about the evolution of the organic world? **e**) What is Punctuated Equilibria?

1+5+5+5+4=20

3. **a**) Write the different units of lithostratigraphy in hierarchical order. **b**) What is an unconformity? Define with sketches the different types of unconformities. **c**) Name the different periods of Palaeozoic era. **d**) When did the first bird *Archaeopteryx* arrive on the Earth? **e**) What is adaptive radiation? Give an example.

3+8+4+1+4=20

4. **a**) What are the differences between quartz and calcite in hand specimen? **b**) What are the differences between an inosilicate and a phyllosilicate structure? **c**) What is a dike? How would you differentiate a dike from a sill? **d**) What are the major factors that control the melting temperature of minerals? **e**) Write the names of the different minerals that appeared sequentially in Bowen's Reaction Series (both discontinuous and continuous).

3+4+4+4+5=20

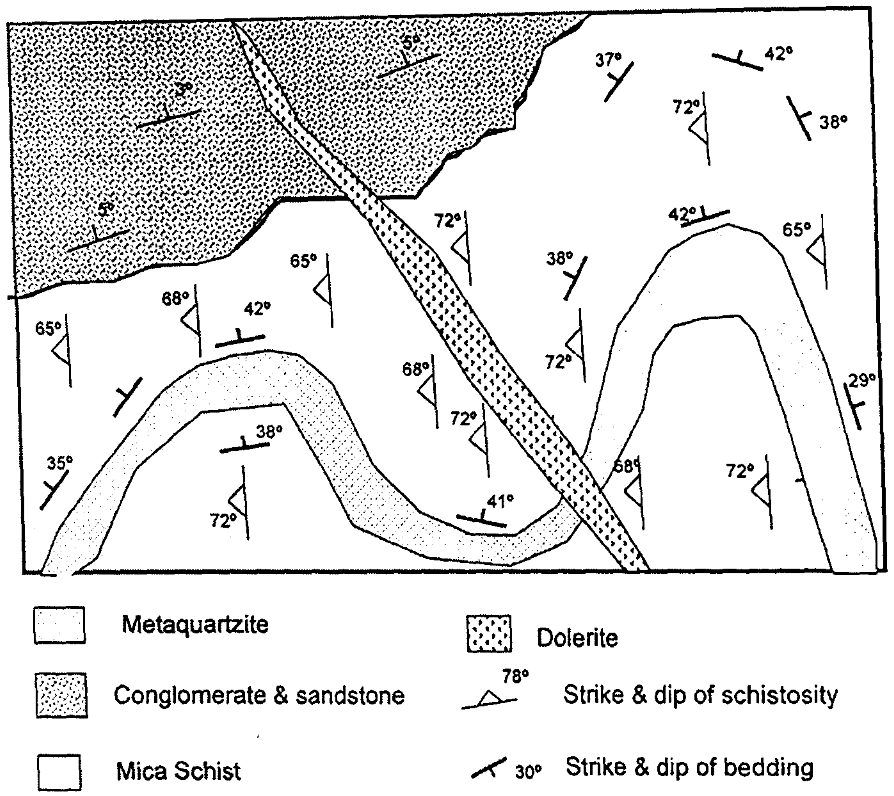
5. **a**) What is a ripple? **b**) Illustrate different parts of a ripple with an appropriate sketch. **c**) How does cross-lamina form from ripple? Discuss giving suitable sketch. **d**) Can ripple form in pure mudstone?

2+8+8+2=20

P.T.O.

6. Describe the major geological features shown in the map below. Infer the relative age of the four rock units shown in the map and justify interpretation.

15+5=20



7. **a)** What is metamorphism? **b)** Name different types of metamorphism. **c)** What is hornfelsic texture? What type of metamorphism is responsible for the development of this type of texture? **d)** Name one typical metamorphic mineral. **e)** What is gneissic texture?

3+5+6+1+5=20

8. **a)** Describe with a suitable block diagram what you understand by dip and strike of bedding. **b)** If a bed is dipping 35° towards south (→ 180°) what is strike direction? **c)** Draw a sketch of the stereographic projection of this plane.

10+5+5=20

9. Illustrate the following features with suitable sketches:
   i) Footwall and hanging wall of a normal fault.
   ii) Upright symmetric fold.
   iii) Isoclinal fold.

10+5+5=20

Introduction to Anthropology and Human Genetics
Semester I (2005-2006), B.Stat. IIIrd Year

Time 2 hours   Date 2. 12. 05   Full Marks 30

1. Answer all questions [5 × 1= 5]

I. Physical Anthropology can best defined as the study of
(a) Human morphology
(b) Primate biology
(c) Human biological variation
(d) Human origins

II. Human races are
(a) Subspecies
(b) Divisions of mankind
(c) Linguistic groups
(d) Religious groups

III. Pedigree analysis is most commonly used by
(a) Human geneticists
(b) Demographers
(c) Statisticians
(d) Zoologists

IV.Gradual changes in average stature over a few decades are called
(a) Secular change
(b) Phylogenic change
(c) Genetic drift
(d) Annual increments

V. Sex in humans is determined by
(a) The number of X-chromosomes
(b) Presence or absence of one Y-chromosome
(c) Ratio of the numbers of autosomes and sex-chromosomes
(d) Ratio of the numbers of X- and Y- chromosomes.

2. What do you mean by the term dermatoglyphics? Describe in brief the main characteristics of dermal ridges.[1+2= 3]

3. Answer all questions [3 × 4= 12]

(a) How do you calculate the frequencies of DZ and MZ twins if we assume that fertilization is at random, p and q are the frequencies of male and female births in the total population?
(b) In a total number of 14114 twins, 2388 are known DZ twins. What will be the frequency of MZ twins?

P. T. O

(c) What is the process of Hybridization or Admixture? How do you measure the extent of admixture if you know the frequencies of particular gene 'qo' in original Black population; 'qm' in the migrant White population and 'qh' in the hybrid population?

OR

What are the types of mutation and the magnitudes of their phenotypic effects? What is the evolutionary significance of mutation? [4+4+4 = 12]

4. What is the difference between acclimatization and adaptation?
Explain Bergman and Allen's rules with suitable example.
Briefly describe the pattern of human nutrition and disease which responsible for Biological adaptation among human population. [2+ 4+ 4=10]

INDIAN STATISTICAL INSTITUTE
First Semestral Examination (2005-2006)
B.Stat.-III
Subject: Introduction to Sociology

Date: 2.12.05 **Maximum Marks: 100** **Duration: Three hours**

(Use separate answer sheets for group –A & group-B)

Group-A

Q.1) Write whether the following statements are true or false: 1x10=10

a) Impersonal structure of organization is reflected in the 'closed' social system.

b) Land does not provide the source of power and control in traditional agrarian social system in India.

c) Dispersed inequality fits with the traditional agrarian social system in India.

d) The growing dissociation between caste and land ownership is major feature in 'modern' rural India.

e) The lesser social distinction between landlords and tenants does not of course mean that all distinctions are disappearing.

f) In contemporary agrarian societies a new class of 'progressive' farmers has emerged as a product of green revolution.

g) One important feature of social change in India is that as the sphere of social relations expands, the referents of caste also changes, so that the unit with which identity is sought is now more and more a group of related castes rather than sub-caste.

h) Personal contact has been playing a vital role in articulating traditional structures such as lineage, caste, and village with the newly emerging structure of production and administration.

i) Marriage is not a social institution.

j) In India, the sociological study on 'agrarian system' precedes the study on 'caste system'.

P.T.O.

Q.2) Answer any three of the following: 10x3=30

1) What is an ethnic group? How would you distinguish it from a minority group? Give an example of a minority group? 3+5+2=10

2) How did Ramkrishna Mukherjee, in his study 'Dynamics of Rural Society', transform the occupational category into class category? What is the sociological justification of this transformation? 5+5=10

3) What do you mean by open social system? What, according to Andre Beteille, are the broad factors responsible for changing the closed system to open system in agrarian societies in India? 4+6=10

4) What is meant by 'cumulative inequality'? How does it differ from 'dispersed inequality'? 5+5=10

5) How would you define social network? How does social network approach differ from the 'conventional variable approach' in sociological study? 4+6=10

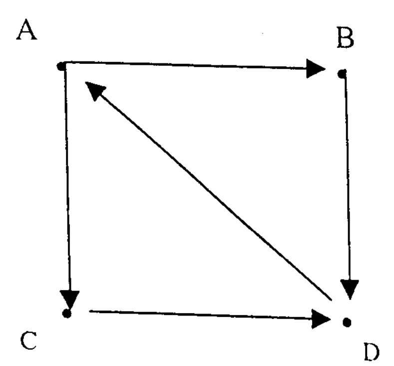
Q.3) Answer the following: 4 x 5=20

a) What is a digraph? Explain diagrammatically all possible digraphs of two vertices A and B. 2½ + 2½ = 5

b) Consider three friends – A, B and C- in a particular social setting. How many possible networks on a help relation can be formed by them if the out-degrees of A, B and C are 2, 1 and 1 respectively. Draw the networks accordingly. 1+4= 5

c) Explain i) Path and ii) Distance in case of a digraph. 2½ + 2½= 5

d) Find distance matrix of the following digraph using matrix algebra: 1 x 5 = 5



Group-B

Q. 4). Answer any three of the following: 10x3=30

a) Define concept. Analyse the relation among concept, proposition and hypothesis. 3+7 =10

b) What is participant observation? How does the participant observation technique differ from merely looking at things? Is not every one a participant observer? 2+5+3 =10

c) What are the important elements of a research design? Discuss the advantages and disadvantages of the descriptive and the explanatory design. 4+6 =10

d) What are the sources of hypothesis? Discuss the qualities of a workable hypothesis. Cite examples. 3+5+2 =10

Q. 5) Write short notes on any two: 2 x 5 =10

a) Content analysis.
b) Strength and weakness of a historical research.
c) Interview.
d) Case study.

***

INDIAN STATISTICAL INSTITUTE

Semestral Examination : (2005 –2006)

B. Stat (Hons) III Year

Economics III

Attempt ALL questions

Date : 2 December 2005 Maximum Marks : 100 Duration : 3 Hours

1. Let the true regression model be $Y = X_1\beta_1 + X_2\beta_2 + u$. However, by mistake, one specifies the model as $Y = X_1\beta_1 + u$. Notation is self-explanatory. The true model satisfies all the assumptions of classical linear regression model. Check the consistency of the OLS estimator of the regression parameters. [8]

2. Consider the classical linear regression equation $y = bx + u$.

   Consider the following class of estimators for b given by $\hat{b} = \frac{\sum x^p y}{\sum x^{1+p}}$, for $p > 0$.

   What is the "best" choice of p in this context? Explain your answer clearly. [12]

3. Consider the regression equation: $Y = X\beta + u$. This equation satisfies all the classical linear regression model assumptions, but $E(u_i^2) = \exp(z_i'\alpha)$, z_i is a vector of known variables with first element as 1, $\alpha = (\alpha_1, \alpha_2, \ldots \alpha_p)$ is a vector of unknown parameters. How will you estimate β efficiently? Is OLS suitable for this situation? Justify your answer. If one applies OLS, what is the appropriate estimate of the standard errors? Discuss an appropriate test for the null hypothesis, $H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_p = 0$. Notation is self-explanatory.

   [8 + 4 + 3 + 10 = 25]

[PTO]

4. (a) Explaining the usual notation, define the order and rank condition for identification of a simultaneous equations system.

(b) Consider the following simultaneous equations system in usual notation.

$y_1 = a_1 + b_1 y_2 + c_1 x_1 + u_1$

$y_2 = a_2 + b_2 y_3 + c_2 x_2 + u_2$

$y_3 = a_3 + b_3 y_1 + u_3$

Discuss the identification of the above system by the order and rank conditions.

(c) Consider the following data on the above system.

Number of Observations: 94

Mean

X1	X2	Y1	Y2	Y3
4.2	42.5	0.8	2.8	0.4

Covariances

	X1	X2	Y1	Y2	Y3
X1	53	35	-1	-2	0
X2	35	118	-5	-14	-2
Y1	-1	-5	19	63	8
Y2	-2	-14	63	203	27
Y3	0	-2	8	27	4

Estimate the simultaneous equations system given in (b) by the 2SLS method.

(d) Explicitly state the 3SLS estimating equations for the system in (b). (You need not solve them with the given data!) [8 + 10 + 25 + 12 = 55]

INDIAN STATISTICAL INSTITUTE
First Semester Examination: 2005-06
B. Stat. II Year
Probability Theory III

Date: 6.12.05 Maximum Marks: 70 Duration: 3 Hours

Answer any six questions. The maximum you may score is 70.

1.(a) If X_1, X_2, X_3 are non-negative random variables having the joint density $e^{-(x_1+x_2+x_3)}$, find the conditional joint density of X_1, X_2 given $U = X_1 + X_2 + X_3 = u$.

(b) Suppose X_1 has the Normal density $N(\mu, \sigma^2)$ and the conditional distribution of $X_2 | X_1 = x_1$ is the normal distribution $N(x_1, \sigma^2)$. Find the joint distribution of (X_1, X_2).

[6+6=12]

2.(a) Let $X_1, X_2, \ldots\ldots$ be a sequence of random variables converging to a random variable X in probability. Show that $X_n \to X$ as $n \to \infty$ in distribution.

(b) For a pair of (proper) distribution functions F,G, let

$$\rho(F, G) = \inf\{\, h > 0 : F(x-h) - h \le G(x) \le F(x+h) + h \text{ for all } x \in \mathbb{R} \,\}.$$

Show that a sequence of distribution functions F_n converges weakly to a (proper) distribution function F iff $\rho(F_n, F) \to 0$ as $n \to \infty$.

(c) Let $0 < p, p' < 1$ and F and F' are the distribution functions attributing weights p and p' to the point 1 and $1-p$ and $1-p'$ to 0 respectively. Calculate $\rho(F, F')$.

[4+5+3=12]

3. Let $A_1, A_2, \ldots\ldots$ be a sequence of events and $N_n = 1_{A_1} + \cdots\cdots + 1_{A_n}$, $n = 1,2,\ldots\ldots$ with $E(N_n) = p_1 + \cdots\cdots + p_n = m_n$. Suppose that $m_n \to \infty$ as $n \to \infty$ and

$$\lim_{n\to\infty} \frac{\sum_{j,k \le n} P(A_j \cap A_k)}{m_n^2} = 1.$$

Find $P\left(\limsup_{n\to\infty} A_n\right)$

[Hint : Obtain a suitable estimate for $P(N_n \le x)$ for any fixed x]

[12]

4.(a) Denote the distribution functions of X_n and Y_n by F_n and G_n respectively, $n = 1,2,\ldots\ldots$ Suppose that $X_n - Y_n \to 0$ in probability and $G_n \to G$ weakly. Show that $F_n \to G$ weakly.

P.T.O.

(b) Let $X_k = \pm 1$ with probability $\frac{1}{2}(1-k^{-2})$ and $X_k = \pm k$ with probability $\frac{1}{2}k^{-2}$, $k = 1,2,\ldots$.

Show that the distributions of $\frac{S_n}{\sqrt{n}}$, where $S_n = X_1 + \cdots + X_n$, converge weakly to the standard normal, but $Var\left(S_n/\sqrt{n}\right) \to 2$ as $n \to \infty$.

(You may use the De Moivre-Laplace theorem without a proof.)

(c) Change the distribution if X_k in (b) so that the first conclusion remains valid, but $Var\left(S_n/\sqrt{n}\right) \to \infty$.

[5+4+3=12]

5.(a) State and prove Kolmogorov's inequality.

(b) Let $X_1,\ldots,X_n,\ldots$ be i.i.d non-negative random variables with $E(X_1) = \infty$. Show that $\frac{S_n}{n} \to \infty$ as $n \to \infty$ with probability one, where $S_n = X_1 + \ldots X_n$.

[6+6=12]

6.(a) Let $\varphi_x(t) = u(t) + iv(t)$ be the characteristic function of a random variable X. Show that

$$u^2(t) \le \frac{1}{2}(1 + u(2t)) \text{ for all } t.$$

(b) If φ_n is a sequence of characteristic functions such that $\varphi_n(t) \to 1$ for $-\delta < t < \delta$, where $\delta > 0$, show that $\varphi_n(t) \to 1$ for all t.

[6+6=12]

7.(a) Show that for any two random variables X and Y, if $\varphi = \varphi_X$ and $\psi = \psi_Y$ are the respective characteristic functions,

$$E(\varphi(Y)) = E(\psi(X))$$

(b) If X has finite variance, show that $\varphi_X(t) = 1 + iE(X)t - E(X^2)\frac{t^2}{2} + \in(t)$

where $\frac{\in(t)}{t^2} \to 0$ as $t \to 0$

[6+6=12]

***

INDIAN STATISTICAL INSTITUTE

B.Stat III, Semester I, 2005-2006

STATISTICAL INFERENCE I

Date: 06.12.05 Semestral Examination Time: 3 hours

Total points 110. Answer as many as you can. Maximum you can score is 100.

1. Let $X_1, \ldots, X_n, X_{n+1}$ be distributed as i.i.d. Bernoulli(p) random variables. Let $h(p)$ be the probability that the sum of the first n observations in the above sample exceeds the $(n+1)$th observation. Find the minimum varianced unbiased estimator of $h(p)$. [15]

2. (a) Let $X_1, \ldots, X_r$ and Y be random variables each having finite second moment, and let $\gamma_i = Cov(X_i, Y), i = 1, \ldots, r$. Let Σ be the covariance matrix of $(X_1, \ldots, X_r)$. Suppose that Σ is positive definite. Then show that

$$\rho^{*2} = \frac{\gamma^T \Sigma^{-1} \gamma}{Var(Y)}$$

where $\gamma = (\gamma_1, \ldots, \gamma_r)^T$, γ^T is the transpose of γ, and ρ^* is the multiple correlation coefficient between $(X_1, \ldots, X_r)$ and Y.

(b) Now suppose that $\mathcal{F} = \{F_{\boldsymbol\theta} : \boldsymbol\theta \in \Theta\}$ be a parametric family of distributions with corresponding densities denoted by $f_{\boldsymbol\theta}$, and $\boldsymbol\theta$ is an r dimensional vector parameter. Assume that the information matrix $I(\boldsymbol\theta)$ is positive definite. Let δ be any statistic with $E_{\boldsymbol\theta}(\delta^2) < \infty$, and suppose that the derivative with respect to θ_i of $\int \delta f_{\boldsymbol\theta}$ exists for each i and can be taken under the integral sign. Then show that

$$Var_{\boldsymbol\theta}(\delta) \ge \boldsymbol\alpha^T I^{-1}(\boldsymbol\theta)\boldsymbol\alpha,$$

where $\alpha_i = \frac{\partial}{\partial \theta_i} E_{\boldsymbol\theta}(\delta)$ and $\boldsymbol\alpha = (\alpha_1, \ldots, \alpha_r)^T$.

(c) Use part (2b) to show that

$$I_{ii}^{-1}(\boldsymbol\theta) \le [I^{-1}(\boldsymbol\theta)]_{ii},$$

where the left hand side is the inverse of the scalar quantity $E_{\boldsymbol\theta}\left[\frac{\partial}{\partial \theta_i} \ln f_{\boldsymbol\theta}\right]^2$, while the right hand side is the ith diagonal element of the inverse of the r dimensional Fisher information matrix $I(\boldsymbol\theta)$. [10+5+10=25]

P.T.O

3. Suppose that X has a Poisson distribution with mean parameter μ, and Y has a Poisson distribution (independently of X) with mean parameter λ. We want to test the hypothesis

$$H_0 : \mu = \lambda \quad \text{versus} \quad H_1 : \mu \neq \lambda.$$

Based on the observations X and Y, derive the uniformly most powerful unbiased (UMPU) test of level α for the above hypothesis. [20]

4. Suppose that (X, Y) has a bivariate normal distribution with $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ representing its parameter vector under usual notation. We want to test

$$H_0 : \rho = 0 \quad \text{versus} \quad H_1 : \rho \neq 0.$$

Derive the likelihood ratio test for this hypothesis, and determine the form of the rejection region. Indicate how you can find the exact critical values for the null distribution to perform this test at level α. [15]

5. Let $X_1, \ldots, X_n$ be a random sample from the distribution with density

$$f(x) = \frac{1}{b} exp\left(\frac{x-a}{b}\right), x > a, -\infty < a < \infty, b > 0.$$

This is called the two parameter exponential distribution. Determine a UMP level α test for

$$H_0 : a = a_0 \quad \text{versus} \quad H_1 : a \neq a_0$$

when b is assumed known, if such a test exists. If not, prove that a UMP test does not exist for this scenario. [20]

6. Let the family of densities $f_\theta(x)$, $\theta \in \Theta$, be identifiable and have monotone likelihood ratio in $T(x)$, and suppose that the distribution function $F_\theta(t)$ of $T = T(X)$ is a continuous function in each of the variables t and θ when the other is fixed.

   (a) Show that there exists a uniformly most accurate lower confidence bound $\underline{\theta}$ for θ at each confidence level $1 - \alpha$.

   (b) If x denotes the observed values of X and $t = T(x)$, and if the equation $F_\theta(t) = 1 - \alpha$ has a solution $\theta = \hat{\theta}$ in Θ, then this solution is unique and $\underline{\theta}(x) = \hat{\theta}$. [15]

INDIAN STATISTICAL INSTITUTE
First Semestral Examination : 2005- 2006
B. Stat. (Hons.) III Year
Sample Surveys

Date : 09.12.2005 **Maximum Marks : 100** **Duration : 3 Hours**

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

1. A company intends to interview a simple random sample of employees who have been with it for more than five years . The company has 1000 units of money to spend , and each interview costs 10 units . There is no separate list of employees with more than 5 years service , but a list can be compiled from the files at a cost of 200 units . The company can either (a) compile the list and interview a simple random sample drawn from the eligible employees or (b) draw a simple random sample of all employees , interviewing only those eligible . The cost of rejecting those not eligible in the sample is assumed to be negligible .
   Show that for estimating a total over the population of eligible employees , plan (a) gives a smaller variance than plan (b) only if $C_j < 2\sqrt{Q_j}$, where C_j is the coefficient of variation of the item among eligible employees and Q_j is the proportion of non-eligibles in the company . You may ignore the fpc .
   [25]

2. (a) Describe how would select a sample according to PPSWR sampling scheme in n draws . Show that the sample selected according to that scheme is really a PPS sample .
   (b) Suggest an unbiased estimator of the population total on the basis of a PPSWR sample of n units and derive an expression for its variance . Also obtain an unbiased estimator of the variance of the proposed estimator .
   (5+5+5+5+5)= [25]

3. (a) Show that if fpc be ignored ,

$$V_{ran} \geq V_{prop} \geq V_{opt}$$

   where V_{ran} , V_{prop} and V_{opt} denote respectively the variances of the estimated mean based on unstratified simple random sampling , stratified simple random sampling with proportional allocation and stratified random sampling with optimum allocation for a given total sample size .
   (b) Discuss how one can estimate the gain in efficiency due to stratification on the basis of a stratified simple random sample .
   (12 + 13) = [25]

4. (a) Derive an expression for the sampling variance of the estimated mean in case of linear systematic sampling . How does the sampling variance depend on the intra-class correlation coefficient ? Assume that the population size is an integral multiple of the sample size . **P.T.O.**

(b) What do you mean by interpenetrating network of sub-samples ? Explain how this technique can be applied to estimate the sampling variance of the estimated mean in case of circular systematic sampling .

(10+2+3+10)=[25]

5. (a) Derive an approximate expression for the bias and the mean square error of the ratio estimator based on SRSWOR sampling scheme .
   (b) Obtain a condition under which the ratio estimator of the population total is more efficient than the usual mean per unit estimator both based on SRSWOR sampling scheme .

(10+10+5) = [25]

6. To estimate the total number of words (Y) in an English dictionary , 10 out of 26 letters were selected with PPSWR , size being the number of pages devoted to a letter and for each selected letter , two pages were selected with SRSWOR . The relevant sample data are given in the following table .

Sl. No.	Sample letters	No. of pages devoted	No. of words in sample page	
			1	2
1.	S	131	34	27
2.	C	97	27	26
3.	N	21	44	38
4.	S	131	24	29
5.	F	43	25	32
6.	J	7	42	48
7.	U	18	24	21
8.	P	85	53	24
9.	A	49	47	55
10	D	54	38	57

(Total number of pages in the dictionary is 980)

(a) Estimate unbiasedly Y and obtain an estimate of its RSE .
(b) Estimate also the efficiency of the above method of sampling compared to that of drawing 20 pages from the dictionary using SRSWR .

(10 + 15) = [25]

INDIAN STATISTICAL INSTITUTE

B.Stat III, Semester I, 2005-2006

STATISTICAL INFERENCE I

Date: 13.2.06 Back Paper Examination Time: 3 hours

Total Point 107. Maximum you can score is 45.

1. Suppose that a random sample of length of life measurements $X_1, \ldots, X_n$ is to be taken on components whose length of life has an exponential distribution with mean θ. It is often of interest to estimate $\bar{F}(t) = 1 - F(t) = e^{-t/\theta}$, the reliability at time t of a component of this type. ($F(\cdot)$ is the distribution function of the length of life of the components). Find the uniformly minimum varianced unbiased estimator of $\bar{F}(t)$. [16 points]

2. Prove that an estimator $aX + b$ $(0 \leq a \leq 1)$ of $E_\theta(X)$ is inadmissible (with squared error loss) under each of the following situations:
(i) if $E_\theta(X) \geq 0$ for all θ and $b < 0$.
(i) if $E_\theta(X) \leq k$ for all θ and $ak + b > k$. [6+6=12 points]

3. Let $\mathcal{F}$ be the set of all continuous distributions, and let $T(F) = F^{-1}(1/2)$ represent the median functional for a distribution $F \in \mathcal{F}$. Show that the influence function of the median functional is a bounded function. [16 points]

4. Let $X_1, \ldots, X_n$ be a random sample from a uniform $(0, \theta)$ distribution.
(i) Determine if there exists a UMP test for testing $H_0 : \theta = \theta_0$ against $H_0 : \theta > \theta_0$. If so, is the test unique?
(i) Determine if there exists a UMP test for testing $H_0 : \theta = \theta_0$ against $H_0 : \theta < \theta_0$. If so, is the test unique?
(i) Determine if there exists a UMP test for testing $H_0 : \theta = \theta_0$ against $H_0 : \theta \neq \theta_0$. If so, is the test unique? [5+5+5=15 points]

P.T.O

5. Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, 1)$ distribution. Let $0 < \alpha < 1$, and $\theta_1 < \theta_2 < \theta_3 < \theta_4$. Consider testing the following set of hypotheses:

$$H_0 : \theta \leq \theta_1 \quad \text{or} \quad \theta_2 \leq \theta \leq \theta_3 \quad \text{or} \quad \theta \geq \theta_4$$
$$H_1 : \theta_1 < \theta < \theta_2 \quad \text{or} \quad \theta_3 < \theta < \theta_4.$$

Does there exist a level α UMP test for the above? If yes, derive the test. If not, explain why not. [16 points]

6. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples from exponential distributions with parameters λ_1 and λ_2 respectively. For the hypothesis

$$H_0 : \lambda_1 = \lambda_2 \text{ vs } H_1 : \lambda_1 > \lambda_2$$

derive the uniformly most powerful unbiased (UMPU) test at level α. [16 points]

7. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. Consider testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. Determine the form of the likelihood ratio test for the above hypotheses and show that the final decision can be taken based on the quantiles of an F distribution. [16 points]

INDIAN STATISTICAL INSTITUTE
Backpaper Examination :2005-2006
B.Stat.(Hons.) III Year
Sample Surveys

Date : 14.2.06 **Maximum Marks : 100** **Duration : 3 Hours**

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

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

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

(10 + 15) = [25]

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

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

(13 +12)=[25]

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

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

(15 + 10) = [25]

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

(5+10+10)=[25]

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

P.T.O.

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

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

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

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

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

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

(10 + 15) =[25]

INDIAN STATISTICAL INSTITUTE
First Semester Backpaper Examination: 2005-06

B. Stat. (Hons.) III Year
Linear Statistical Models

Date: 15.02.06 **Maximum Marks : 100** **Duration: 3 Hours**

Answer all the questions.

1. Show that in the linear model $\left(y_{n\times 1},\ x\beta_{n\times k\ k\times 1},\ \sigma^2 I\right)$ every estimable linear parametric function has a unique BLUE.

[20]

2. Consider the linear model $(y,\ x\beta,\sigma^2 I)$. If z is a vector such that its elements constitute a standardized basis set of linear zero functions then show that (a) z has n – Rank (x) elements (b) $z'z$ = e'e.

[20]

3. Explain the following terms :

   (a) Collinearity between columns of a design matrix.
   (b) Autocorrelation.
   (c) Log linear models
   (d) Tolerance interval

.[3 x 4=12]

4. In the linear model $(y,\ x\beta,\sigma^2 I)$ find ℓ such that $\ell'\beta$ is estimable and the ratio of the variance of its least squares estimator to $\ell'\ell$ is a maximum.

[8]

5. Develop analysis of covariance for a two-way classified data with two concomitant variables. Give the analysis of variance and covariance table indicating clearly the details of all the computations involved.

[20]

6. The following data shows the observations from 12 days' operation of a plant for the oxidation of amonia, which is used for producing nitric acid. The response variable (y) is the 'stack loss' defined as the percentage of the ingoing amonia that escapes unabsorbed. The explanatory variables are cooling water inlet temperature in centigrade degree (x_1) and acid concentration in percentage (x_2). Obtain the best linear unbiased estimators of the coefficients of the explanatory variables alongwith their standard errors.

Case	1	2	3	4	5	6	7	8	9	10	11	12
x_1	18	18	17	18	19	18	18	19	19	20	20	20
x_2	58.0	58.9	58.8	58.2	59.3	58.5	57.2	57.9	58.0	58.2	59.1	58.9
y	1.4	1.4	1.3	1.1	1.2	0.8	0.7	0.8	0.8	0.9	1.5	1.5

[20]

*****

Introduction to Anthropology and Human Genetics

Semester I (2005-2006),B.Stat. IIIrd Year

BACK PAPER

Time 3 hours Date: 16.2.06 Full Marks 100

Attempt all questions

1. Define dermatoglyphics and mention its unique functions. [10]

2. How do you define Anthropometry? What are the main objectives of this subject? [2+8=10]

3. What do you mean by population genetics and variation? Describe the purpose of its study and the process of genetic change in population? [3+3+2+2= 10]

4. What are the causes of plural births? Describe various kinds of human Twins. [5+5=10]

5. Answers the following questions [2+4+4=10]
(a) What is the difference between acclimatization and adaptation?
(b) Explain Bergman and Allen's rules with suitable example.
(c) Describe briefly the factors responsible for biological adaptation among human population.

6. Describe the following finger and palmar variables (any five) with definition and procedure of measuring it with the help of diagram. [5×2=10]
(a) Whorl
(b) Loop
(c) Simple Arch
(d) Tented Arch
(e) Angle atd
(f) Ridge count a-b, b-c and c-d
(g) TFRC

7.Describe the following measurements (any five) with definition instrument used definition of landmark and abbreviation. [5×2=10]
(a) Maximum Head Length
(b) Least Frontal Breadth
(c) Bigonial Breadth
(d) Bizygomatic Breadth
(e) Nasal Breadth
(f) Ear Length
(g) Ear Breadth

8. M- N blood groups were recorded: M (1787), MN (3039) and N (1303) in a sample of American Whites. What are the gene frequencies? [5]

P.T.O

9. Describe the methods of estimating the coefficient of relationship and inbreeding with the help of hypothetical diagram. [5]

10. Write short notes on (any five) [5×4=20]
(a) Mutation
(b) Hybridization
(c) Kind of Twins
(d) Contemporary human population groups
(e) Genetic Drift versus natural selection
(f) Bottleneck effect
(g) Causes of plural births (human)

Indian Statistical Institute
Mid-semester Examination : (2005-2006) B.Stat(hons.) III
Statistical Inference II
Maximum Marks: 40   Date: 20.2.06   Duration :- 3 hours

This paper contains questions worth a total of 48 points. Answer as much as you can. The maximum you can get is 40 points.

1. Let $X_1, \ldots, X_n$ be i.i.d. observations from some unknown continuous distribution.
(a) Show that the power $P_F(D_n \geq k_\alpha)$ of the one-sample Kolmogorov-Smirnov test is bounded below by $\sup_x P_F(|F_n(x) - F_0(x)| \geq k_\alpha)$, where $F_0(.)$ is the distribution specified by the null hypothesis.
(b) Now consider the problem where under the null hypothesis the sample is from a $N(\mu, \sigma^2)$ distribution where both μ and σ are unknown. Consider the one-sided Kolmogorov-Smirnov test statistic with μ and σ estimated by $\bar{X}$ and S respectively, i.e consider

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

Show that the null distribution of $\hat{D}_n^+$ does not depend on μ and σ.
[2+3=5]

2. Consider the one-sample problem : $X_1, X_2, \ldots, X_n$ are i.i.d. F where F is continuous with unknown median M. We want to test $H_0 : M = 0$ vs $H_1 : M > 0$. Show that the sign test is consistent.
[6]

3. Consider the one sample location problem.
(a) Describe the Wilcoxon Signed Rank test for the one sample problem, clearly stating all assumptions made about the distribution from which the sample is drawn.
(b) Find a $100(1-\alpha)\%$ (lower) one-sided confidence interval for the unknown median of the population in this problem.
[2+3=5]

4. Consider now the two sample problem.
(a) Describe the Wald-Wolfowitz Runs Test for testing $H_0 : F(x) = G(x)$ for all x vs. $H_1 : F(x) \neq G(x)$ for at least some x, where $X_1, \ldots, X_m$ are i.i.d. with distribution F and $Y_1, \ldots, Y_n$ are i.i.d. with distribution G, both F and G being continuous.
(b) Explain, with an example of an ordered arrangement of $X's$ and $Y's$, why this test statistic may not be suitable where the alternative hypothesis is one sided (say, Y is stochastically larger than X).
[2+2=4]

5. Consider the two-sample Kolmogorov-Smirnov Test. Assume both F and G are continuous.
(a) Under the assumption that $F(x) = G(x)$ for all x, show that for all m, n and c

$$P(D_{mn}^+ < c) = P(D_{nm}^+ < c) = P(D_{mn}^- < c) = P(D_{nm}^- < c).$$

P.T.O

(b) Show that the test based on D_{mn} is consistent for testing $H_0 : F(x) = G(x)$ for all x against the alternative $H_1 : F(x) \neq G(x)$ for at least some x. [4+4=8]

6. Let $X_1, \ldots, X_m$ be i.i.d. with distribution F and $Y_1, \ldots, Y_n$ be iid with distribution G where both F and G are strictly increasing.
(a) Show that in the case Y is stochastically smaller than X, one can write $G(t) = F(t - \Delta(t))$ for some function $\Delta(.)$, such that $\Delta(t) \leq 0$ for all t. Find an explicit expression for $\Delta(t)$.
(b) Let c_α be such that $P_{F=G}(U \leq c_\alpha) = \alpha$ where

$$U = \{ \text{ number of pairs}(i,j) \text{ such that } Y_j < X_i \}.$$

Then using part (a) show that if Y is stochastically smaller than X, then $P(U \leq c_\alpha) \leq \alpha$. [3+5=8]

7. Consider the one sample problem : $X_1, \ldots, X_n$ are i.i.d. observations with distribution F where $F \in \mathcal{F}$, $\mathcal{F}$ being some class of distribution functions.
(a) Define the U-statistics U_n for unbiased estimation of $\theta = \theta(F)$ based on a symmetric kernel $h(x_1, \ldots, x_m)$.
(b) Prove that $nVar(U_n) \to m^2\sigma_1^2$ as $n \to \infty$, where symbols have usual meanings. You may assume that the total contribution to $nVar(U_n)$ from terms involving σ_c^2 for $c > 1$ tends to zero as $n \to \infty$.
(c) Write down the formula for $\hat{U}_n$, the projection of U_n. Show that $\sqrt{n}(U_n - \hat{U}_n) \to 0$ in probability, when $h(x_1, x_2) = \frac{(x_1 - x_2)^2}{2}$, assuming that $E_F(X_1)) = 0$.
(d) State the theorem proved in class for the asymptotic normality of a suitably centered and rescaled variant of U_n. Consider $\mathcal{F} = \{\text{Bernoulli}(\theta), 0 < \theta < 1\}$. Find those values of $\theta \in (0,1)$ so that you can apply this theorem for finding the asymptotic distribution of s^2 (suitable centered and rescaled), the sample variance. [1+3+5+3=12]

Statistics Comprehensive
BIII
Full Marks : 100
Time: 3 hrs.
Books & Classnotes are allowed.
Date: 23 February, 2006

1. An unbiased die is rolled once. Let the score be N $\epsilon\{1, 2, \cdots 6\}$. The die is then rolled N times. Let X be the maximum of these N scores. Find the probability of the event (X=6). [10]

2. Find Simplicial Area Median of the following 5 bivariate observations: (0,0), (0,1), (1,0), (1,1), $(\frac{1}{2}, \frac{3}{4})$. Show each step of your computation. [15]

3. Get 3 sequences: $X = \{X_1, X_2, \cdots X_n\}, Y = \{Y_1, Y_2, \cdots Y_n\}$ and $Z = \{Z_1, Z_2, \cdots Z_n\}, n \geq 5$ such that
Spearman Rank Correlation $(X, Y) >$ Spearman Rank Correlation (X, Z) but Kendal's $\tau(X, Y) <$ Kendal's $\tau(X, Z)$. [15]

4. Suppose you have an unbiased coin & you are allowed to toss it as many times as you like. Suggest a method of choosing one person from a group of 6 persons so that each person has probability of selection $\frac{1}{6}$. Prove that your method achieves the said objective. [15]

5. Suppose $X_1, X_2, \cdots, X_n$ are i.i.d. observations from uniform distribution $[\theta, 2\theta]$. Let $X_{(1)}, X_{(2)}, \cdots X(n)$ be the order statistics.
(a) Are $X_{(1)}, X_{(n)} - X_{(1)}$ and $\frac{X_{(n)}}{2}$ unbiased estimates of θ ? Justify.
(b) Of the above 3 estimates which is the 'best'? Justify.
(c) Can you suggest any estimate which is better than all the 3 above? [3+6+6=15]

6. Consider the following linear model

$$Y_{ij} = \alpha_i + \beta_j + e_{ij}, i = 1, 2, j = 1, 2, 3;$$

a) What is the rank of the error space? Justify.
b) Write down with justification any linear function of observations that belongs to (i) estimation space, (ii) error space.
c) Write down a parametric function that is not estimable. Justify your answer. [3+ 3+3 +6 =15]

P.T.O

7. Consider SRSWR of size n from a population of N units. Suppose the value of Y_1 is unusually low whereas that of Y_N is very high. Consider the following estimator of $\overline{Y}$, the population mean;

$$\hat{\overline{Y}} = \begin{cases} \overline{y} + c \text{ if the sample contains unit 1, but not unit } N \\ \overline{y} - c \text{ if the sample contains unit } N\text{, but not unit 1} \\ \overline{y} \quad \text{otherwise} \end{cases}$$

Where $\overline{y}$ is the sample mean & c is a constant.

a) Show that $\hat{\overline{Y}}$ is unbiased.

b) Can you suggest a "good" choice of c explaining clearly your idea of "good".

(8+7=15)

BIII - Introduction to Stochastic Processes
Midsem. Exam. / Semester II 2006

February 27, 2006 Maximum score: 45 Time - $2\frac{1}{2}$ hours

1. [4 + 5 + 6 = 16 points]

Let there are two urns, A and B, containing N balls in total. A game is continuing as follows: First one chooses an urn with probability proportional to the number of ball it contains and then move a ball from the chosen urn to the other. Let X_n be the number of balls in the urn A after the nth draw.

(a) Argue briefly that it is a homogeneous Markov chain. Is this chain irreducible? Write down its transition probability. Determine the period of state 0.

(b) Is this chain positive recurrent or null recurrent? Why? If positive recurrent find its unique invariant probability.

(c) If the probability of choosing an urn is proportional to the square of the number of balls it contains write down the transition probability. Find the unique invariant probability if it exists, for $N = 5$.

2. [5 + (6 + 5) = 16 points]

Let $\{X_n\}$ be a Markov chain on $S = \{0, 1, 2, \ldots\}$ with transition probability $p_{i,i+1} = p_i > 0$ and $p_{i,0} = q_i = 1 - p_i > 0$ for each $i \geq 0$.

(a) Show that, the Markov chain is recurrent if and only if

$$\Pi_{i=1}^{n} p_i \to 0 \quad \text{as } n \to \infty.$$

(b) Determine whether the chain is positive or null recurrent, in the following cases, calculating the expected time to return to zero. In the case of positive recurrence find the corresponding invariant distribution.

(i) $p_i = (i+1)/(i+2)$ and (ii) $p_i = (i+1)^2/(i+2)^2$ for each $i \geq 0$.

3. [8 + 8 = 16 points]

Suppose in a maize experiment, with rats learning about reward and punishment (as food and electric shock respectively), a rat moves from one chamber to any of the neihboring chambers with equal probability.

food		×	×	×
×			×	×
×	×	RAT		×
×	×			shock

(a) Calculate the probability that the rat starting at a middle chamber riching the food first before getting shocked.

(b) Find the expected time that the rat will reach one of food or shock chamber.

All the best.

INDIAN STATISTICAL INSTITUTE

Date: 1.3.06

Mid-semester Examination: (2005-2006)

B.Stat 3rd Year, Design of Experiments

This paper carries 33 marks, Maximum you can score is: 30 Duration: 2 hours.

Answer all questions. Marks will be deducted for unnecessarily lengthy answers.

1. Answer true or false:(no justification needed)

   [a] A quantitative factor which is not of primary interest may be used either as a covariate or as a blocking factor.

   [b] In a balanced design, all treatment contrasts are estimable with the same variance.

   [c] A disconnected design may allow the estimation of some treatment contrasts. $[3 \times 1 = 3]$

2. Give the layout of any design in 6 treatments (labeled $1, 2, \ldots, 6$) and 7 blocks such that the design is as follows: (no justification needed) $[4 \times 2 = 8]$

   [a] is orthogonal and balanced

   [b] allows the estimation of all contrasts between treatments 1, 2 and 3 but does not allow the estimation of any contrast between treatments 1 and 4.

   [c] is nonorthogonal

   [d] has replication vector $r = (2, 4, 3, 6, 5)'$, block-size vector $k = (6, 3, 2, 3, 6)'$ and is orthogonal.

3. An experiment was carried out to compare the effects of 4 formulations of a chemical ($say, 0, 1, 2, 3$) on the reaction time. The experiment was done by 4 chemists and the allocation of the formulations to the chemists are as shown below:

Chemist	*Formulation*
1	$0, 1, 2$
2	$0, 1, 3$
3	$0, 2, 3$
4	$1, 2, 3$

   [a] What would you randomize when using this design? [1]

   [b] State a model you would use to analyze data from this experiment. Clearly state your assumptions. [1+1=2]

   [c]Obtain the C matrix for the above design. Hence check if all treatment contrasts are estimable. [1+1=2]

   [d] Is this design variance balanced? [1]

4. (a)State any two alternative definitions of connectedness and show that they are equivalent. [3]

   [b] Under the usual fixed effects, additive model for a block design, derive the expression for adjusted sum of squares for treatments. Write down(without proof) the simpler expression for this sum of squares when the design is an RBD. [3+1=4]

   [c] Give an example of an experiment where a randomised block design will be appropriate. An experiment was conducted as a Randomised Block design with 4 treatments and 7 blocks. The observation from the experimental unit in block 2 which received treatment 3, is lost. Obtain an expression for estimating this missing observation. [1+3=4]

   [d](i)State and prove a sufficient condition for a connected, proper, equireplicate, binary design to be orthogonal. (ii)Suppose d is an orthogonal design in v treatments and b blocks. A new design d^* is constructed whose jth block consists of all treatments in the jth block of d repeated twice, $j = 1, 2, \ldots, b$ Will d^* be orthogonal? Justify your answer. [2+3=5]

INDIAN STATISTICAL INSTITUTE

Mid Semestral Examination: (2005 - 2006)

Course Name: B. Stat **Year: 3rd year**

Subject Name: Database Management Systems

Date: 3.3.06 **Maximum Marks: 50** **Duration: 2 hrs 30 min**

Answer all the questions

1 a) Discuss some of the advantages of using the DBMS approach over the file processing approach. (5)

b) Explain the concepts of logical and physical data independence (5)

2 Consider an insurance company database. This database could be modeled as the entity sets *EMPLOYEES*, with attributes *E#, name, and salary; SALESMEN*, with attributes *E#, name, and salary; MANAGERS*, with attributes *E#, name, and salary; and POLICIES*, with attributes *P#, policy-name, beneficiary* and *amount*. *SALESMEN* and *MANAGERS* are subsets of *EMPLOYEES*. Design an E-R diagram illustrating the relationships among these entity sets. Show the keys and the mapping cardinalities properly along with your assumptions. (10)

3 Consider the EMP-PROJ relation:
EMP-PROJ (SSN, PROJ-NUM, HOURS, EMP-NAME, PROJ-NAME, PROJ-LOCATION)
that records the number of hours an employee works on a project.
(SSN – Social Security Number, PROJ-NUM – the project number, HOURS – the hours that an employee works in the project, EMP-NAME - name of the employee, PROJ-NAME – name of the poject, PROJ-LOCATION – where the project is located)

a) Identify the functional dependencies in this relation and also find all the candidate keys. (7)
b) Identify the normal form of the above relation. Give reasons for your answer. (3)

4 a) Explain the distinction between a key and a superkey? (2)

b) Consider the following relations:
Sailor (<u>s-id</u>, s-name, s-address);
Boat(<u>b-name</u>, color, rent);
Reserves(<u>s-id, b-name</u>, date);

Write the following queries in relational algebra and relational calculus:
"Give the names (s-name) of all sailors who reserved a green boat for 1/1/06." (4+4)

5 Write short notes on any two: (5x2)
a) Safe expression.
b) relationally complete language
c) Foreign key

INDIAN STATISTICAL INSTITUTE

End Semestral Examination: (2005 - 2006)

Course Name: B. Stat. (Hons.) **Year: 3rd year**

Subject Name: Database Management Systems

Date: 12. 5. 06 **Maximum Marks: 60** **Duration: 2 hrs 30 min**

Answer as many questions as possible.

1. Consider a student database system which must contain the following

   Teacher: Employee roll, name, class in which teaching, subject taken in the class.
   Student information: Roll number, name, address, class of the student, subjects, and the marks in the respective subjects.
   Class Information: Class, calendar year, class teacher, subjects.

   (a) Identify the relations and their attributes that you will use in your database. Also mention the dependencies in each, and the candidate keys. Note that the relations must atleast be in 3NF, so normalize the above relations if necessary. (14)

   (b) Write the following queries in SQL:

   i. Retrieve the names and Employee roll numbers of all the teachers who taught in 2005. (4)
   i. Retrieve the Employee roll numbers of all the teachers who teach in classes with more than 10 students. (6)
   ii. Retrieve the roll numbers, the names of the students and their marks in DBMS who were taught this subject by Prof. XYZ in the calendar year 2006. (6)

2. (a) State and explain dynamic programming algorithm for join order optimization.

   (b) Consider the relations $r_1(A,B,C)$, $r_2(C,D,E)$ and $r_3(E,F)$, with primary keys A, C and E, respectively. Assume that r_1 has 1000 tuples, r_2 has 1500 tuples, and r_3 has 750 tuples. Estimate the join size of $r_1 |X| r_2 |X| r_3$, and give an efficient strategy for computing the join. (8+7 = 15)

3. a) Explain the desirable properties of transactions. (4)
   b) What is the utility of a system log, and what are its typical entries? (6)
   c) What is checkpointing? Explain its importance. (5)

3. Write short notes on any two (3x2=6)
   (a) Hierarchical Data Model
   (b) File Indexing
   (c) Database Security

BIII - Introduction to Stochastic Processes

Final Exam. / Semester II 2006

Maximum score: 55 / Time: 3 hours

Date: May 16, 2006

1. [5 + 8 + 7 = 20 points]

   Let $\{X(n)\}$ be a discrete parameter Markov chain on $S = \{0, 1, 2, \ldots\}$ with $p_{0,0} = 1 - p_0 \geq 0$, $p_{i,i+1} = p_i \geq 0$ and $p_{i,i-1} = 1 - p_i \geq 0$, for each $i \geq 0$.

   (a) Assume $p_0 > 0$. Fix $m \geq 1$ an integer and let $\tau = \min\{n \geq 0 : X(n) = m, \text{ or } X(n) = 0\}$. Find the probability $P(X(\tau) = m)$.

   (b) Deriving a criterion using part (a) determine whether $X(\cdot)$ is transient or recurrent in the following cases:

   (i) $p_i = (i+1)/(i^3+1)$ and (ii) $p_i = (i^2 - i + 1)/(i^2 + 1)$.

   (c) For $p_i = (i^2)/(i^2 + i + 1)$. Find the probability of absorption and expected time to be absorbed at 0, starting at some integer $k \geq 1$.

2. [7 + 2 + 6 = 15 points]

   Let $\{X(t)\}_{t\geq 0}$ and $\{Y(t)\}_{t\geq 0}$ be two independent bacterial colony following pure birth processes. For $i \geq 0$, $-q^x_{i,i} = q^x_{i,i+1} = \alpha_i > 0$ and $-q^y_{i,i} = q^y_{i,i+1} = \beta_i > 0$ be parameter of the X and Y processes, repectively. Assume initial size of both the colony is one.

   (a) Write the backward equation for $p^x_{1,j}(t)$. Find the expected size of X when Y has the first offspring.

   (b) Define a non-explosiveness of a stochastic process.

   (c) Show that the process X is non-explosive iff

   $$\sum_{i\geq 1} \frac{1}{\alpha_i} = \infty.$$

3. [5 + 5 + 5 = 15 points]

   Suppose that the life time of a light bulb X is a random variable with hazard rate

   $$h(t) := \frac{f(t)}{P(X > t)} = \beta t, \; t > 0 \quad \text{where } f \text{ is the density of } X.$$

   Each failed light bulb is replaced immediately with a new one.

P. T. O

(a) Find the expected number of light bulbs failed by time t, approximately, for large t.

(b) Find the asymptotic distribution of excess life of the light bulb at time t.

(c) Determine an asymptotic expression for the mean age of the light bulb in service at time t

4. [$4 + 7 + 4 = 15$ points]

A factory has N no. of machines with $K(\leq N)$ no. of people who can run a machine. Each machine can run independently and its failure time is exponentially distributed with parameter λ. The factory has a repair shop which has capacity $L(\leq N)$. Assume that the repair time of a machine is exponentially distributed with parameter μ.

Let $X(t)$ be the no. of machines operating at time t. Observe that it is a Markov process.

(a) Had the repair time been uniformly distributed over $(0, T)$, for some $T > 0$, would X be Markov? Had the repair time been deterministic would X be Markov? Say, yes or no and justify your answer briefly, in both the cases.

(b) Write down its Q matrix. Find the stationary distribution.

(c) Calculate the long run proportion of idle time at the repair shop. Determine the long run proportion of the time that the factory is running in full capacity

All the best.