STATISTICS

Paper – I

Time Allowed : **Three** Hours

Maximum Marks : **200**

Question Paper Specific Instructions

Please read each of the following instructions carefully before attempting questions :

*There are **EIGHT** questions in all, out of which **FIVE** are to be attempted.*

*Questions no. **1** and **5** are **compulsory**. Out of the remaining **SIX** questions, **THREE** are to be attempted selecting at least **ONE** question from each of the two Sections A and B.*

Attempts of questions shall be counted in sequential order. Unless struck off, attempt of a question shall be counted even if attempted partly. Any page or portion of the page left blank in the Question-cum-Answer Booklet must be clearly struck off.

All questions carry equal marks. The number of marks carried by a question/part is indicated against it.

*Answers must be written in **ENGLISH** only.*

Unless otherwise mentioned, symbols and notations have their usual standard meanings.

Assume suitable data, if necessary and indicate the same clearly.

SECTION A

Q1. (a) The random variable X has the exponential probability density function (pdf) given by

$$f(x) = \lambda \exp(-\lambda x), \quad x \geq 0, \lambda > 0.$$

Show that, for any $c > 0$, $P(X > c) = \exp(-\lambda c)$.

Hence show that, for any $x > c$, $P(X > x \mid X > c) = \exp(-\lambda(x - c))$.

Deduce the conditional pdf of X given that $X > c$, and comment briefly. *2+3+3*

(b) 12·5% of the candidates in a specific examination of a certain year are known to have a score of at least 70% in Statistics Paper I, while another 18·1% have a score of at most 38%. Assuming the underlying distribution to be normal, estimate the probability that in a random sample of 5 such candidates, 2 will have a score of 60% or more. [You may use the following information : For a standard normal variate X, $P\{X \leq K = 0{\cdot}637, 0{\cdot}875$ and $0{\cdot}919$ for $K = 0{\cdot}35, 1{\cdot}15$ and $1{\cdot}40$ respectively] *8*

(c) The random variable Y has geometric distribution with parameter $p(0 < p < 1)$, i.e.,

$$P(Y = y) = (1 - p)^y . p \quad \text{for } y = 0, 1, 2, \ldots,$$

(i) Find the probability generating function of Y, and hence find the mean and variance of this distribution. *8*

(ii) The random variables $Y_1, Y_2, \ldots, Y_n$ constitute a random sample from this distribution. Define $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Find an unbiased estimator of $\frac{1}{p}$ (to be shown), and check for its consistency. *6+2*

(d) Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with probability density function

$$f(x) = \beta(1 - x)^{\beta - 1}, \; 0 < x < 1,$$

where β (> 0) is an unknown parameter.

(i) Find the maximum likelihood estimator, $\hat{\beta}$, of β. *5*

(ii) Suppose that the values of $X_1, X_2, \ldots, X_n$ are not known, but you do know Y, the number of X_i less than 0·5. State the distribution of Y. *3*

Q2. (a) The random variables X and Y are jointly distributed with probability density function (pdf)

$$f(x, y) = \begin{cases} \dfrac{1}{3\log 2}\left(\dfrac{x}{y} + \dfrac{y}{x}\right), & 1 \le x \le 2,\ 1 \le y \le 2 \\ 0, & \text{otherwise.} \end{cases}$$

(i) Find the marginal pdf of X. *4*

(ii) Find the conditional pdf $f(y/x)$, for $1 \le x \le 2$, $1 \le y \le 2$, and hence evaluate $P[\,Y < 1{\cdot}5 \mid X = 1]$. *4+2*

(b) Suppose X and Y are independent random variables having Poisson distributions with respective means $\lambda\ (>0)$ and $\mu\ (>0)$.

(i) Show that $X + Y$ also follows a Poisson distribution. *5*

(ii) Find $P(X = k \mid X + Y = n)$, where k and n are integers with $0 \le k \le n$. For given $n > 0$, name the distribution you have obtained. *5*

(c) If a certain team loses one of its matches, then it has probability 0·5 of losing the next match and probability 0·4 of drawing it. If the team draws a match, then it has probability 0·3 of losing the next match and probability 0·4 of drawing it. If the team wins a match, then it has probability 0·2 of losing the next match and probability 0·4 of drawing it.

(i) Model this as a Markov chain, and write down its transition matrix. *5*

(ii) If the team loses its first game of the season, find the probability that it wins its third game. *5*

(d) Suppose (X, Y) follows bivariate normal distribution $N_2(0, 0, 1, 1, \rho)$. Then show that *10*

$$\rho = \cos q\pi, \text{ where } q = P\{XY < 0\}.$$

Q3. (a) (i) State the Central Limit Theorem (CLT) for sums of independently and identically distributed random variables. *4*

(ii) Examine if CLT holds for the sequence of random variables $\{X_K\}$ with $P\left\{X_K = \mp\sqrt{2K-1}\right\} = \dfrac{1}{2}$, K = 1, 2, ... *6*

(b) The probability that a certain tomato seed germinates is θ. A gardener sows a set of n such seeds and finds that x of them have germinated. It can be assumed that seeds germinate independently of one another. Find the posterior distribution of θ, assuming that its prior distribution is beta with parameters α and β. *10*

(c) (i) Define a family of UMAU confidence sets with a given confidence coefficient $1 - \alpha$ for a parameter θ.

(ii) Derive a UMAU confidence set for σ^2 with confidence coefficient $1 - \alpha$ in sampling from $N(\mu, \sigma^2)$ with μ unknown. Show that this set is actually an interval. *3+7*

(d) Let $X_1, X_2, ..., X_n$ be a random sample from a uniform distribution

$$P(X_i = K) = \begin{cases} \frac{1}{N}, & K = 1, 2, ..., N \\ 0, & \text{elsewhere.} \end{cases}$$

Find a Uniformly Minimum Variance Unbiased Estimator (UMVUE) (to be shown) of N. *10*

Q4. (a) (i) Define a maximum likelihood estimator for a parameter θ, and state the large sample properties of this type of estimator under regularity conditions, to be stated clearly. *8*

(ii) Suppose that the number of a particular plant species in sampling quadrats follows a Poisson distribution with mean λ and it is required to estimate $\theta = \lambda^2$. A random sample of n such quadrats yields the numbers $X_1, X_2, ..., X_n$. Find the unbiased estimator of θ (to be shown), and also find the Cramer-Rao lower bound for the variance of this unbiased estimator of θ. *6+6*

(b) Discuss how you would estimate the unknown number of fishes of a given size (say, ≥ 1 kg) in a large pond, based on hypergeometric probability model, using catch-recatch method. Indicate how this technique can be used for estimating the unknown number of tigers in a large forest, using pug-mark method.

[*Hint* : Pug-mark can be inflicted on a tiger from a distance when it comes for drinking in a pond. A tiger thus marked can be identified at a later time, again from a distance.] *7+3*

(c) Discuss Wald-Wolfowitz Runs test to examine if two samples of sizes n_1 and n_2 come from an identical population, against the alternative that the two populations from which the two samples have been taken, differ in any respect whatsoever. *10*

SECTION B

Q5. (a) Discuss the randomised response technique for estimating a sensitive parameter, such as the proportion of tax-evaders in a community. *10*

(b) Explain the terms 'Estimable Parametric Function', 'Error Function' and 'Best Linear Unbiased Estimator (BLUE)' in connection with linear estimation. Show that the sample mean in sampling from $N(\mu, \sigma^2)$ is BLUE for μ. *6+4*

(c) The variance of a stratified random sample mean can be written as

$$\mathrm{Var}(\bar{y}_{st}) = \sum_{i=1}^{k} \frac{N_i^2}{N^2}\left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^2 .$$

Explain the (conventional) notation used here. Suppose the total cost of sampling is

$$C = C_0 + \sum_{i=1}^{k} C_i\, n_i,$$

where $C_0, C_1, \ldots, C_k$ are positive constants.

If $\mathrm{Var}(\bar{y}_{st})$ is fixed and the stratum sample sizes are chosen to minimise the total cost of sampling, show that the i^{th} stratum sample size n_i is proportional to

$$\frac{N_i\, S_i \,/ \sqrt{C_i}}{\sum_{i=1}^{k} N_i\, S_i \,/ \sqrt{C_i}}, \quad i = 1, 2, \ldots, k.$$ *10*

(d) (i) State briefly three reasons why an analyst may wish to perform a principal component analysis. *5*

(ii) Under what circumstances would it be sensible to use the variance-covariance matrix instead of the correlation matrix in principal component analysis ? *5*

Q6. (a) Suppose Y_1, Y_2, Y_3, Y_4 are independent with

$$E(Y_1) = E(Y_2) = \theta_1 + \theta_2$$

$$E(Y_3) = E(Y_4) = \theta_1 + \theta_3$$

$$\mathrm{Var}\,(Y_i) = \sigma^2, \; i = 1, 2, 3, 4.$$

Determine the condition of estimability of the parametric function $l'\theta = l_1\theta_1 + l_2\theta_2 + l_3\theta_3$. Obtain a solution of the normal equations and the sum of squares due to error. *3+4+3*

(b) Write in detail about the use of orthogonal polynomials in regression analysis, and how we choose its appropriate degree for a given data set. *7+3*

(c) Show that, with usual notation,

$$1 - r^2_{1.2.3.\dots p} = (1 - r^2_{12})(1 - r^2_{13 \cdot 2}) \dots (1 - r^2_{1p.23\dots p-1}).$$

Discuss the significance of this result. *8+2*

(d) Describe Lahiri's method, with justification, of drawing a sample of size n from a population of size N with probabilities proportional to the sizes of the respective units. *2+8*

Q7. (a) Discuss the advantages of factorial designs. Give the complete analysis, including the layout, of a 3^2-factorial design laid out in r replicates of 3 incomplete blocks each, totally confounding the factorial effect AB^2. *2+8*

(b) Explain the concept of missing-plot technique and the analysis of an $r \times r$ Latin Square design with one missing value, clearly stating the assumptions needed. *10*

(c) Define a symmetrical balanced incomplete block design and show that the number of treatments common between any two blocks of a symmetrical balanced incomplete block design is a constant, say λ. *10*

(d) Show that a randomised block design is an orthogonal design. *10*

Q8. (a) Define one-stage and two-stage cluster sampling. How do cluster sampling and stratified sampling differ, both in construction and use ? Give an example of a survey that uses both stratification and clustering in the sample design. *4+3+3*

(b) Define Hotelling's T^2 statistic, and indicate its applications. Show that

$$T^2(p, m) = \frac{mp}{m - p + 1} F_{p, m-p+1},$$

symbols having their usual significance. *2+3+5*

(c) (i) Discuss the need of ratio estimation and find the bias and the variance of the sample estimator $\hat{R}$, of population ratio R. 7

(ii) Explain briefly the circumstances under which the ratio estimator of a population mean will be less precise than the sample mean of a simple random sample of the same total size. 3

(d) Show that an incomplete block design is connected if and only if the rank of the C-matrix is v – 1, where v denotes the number of treatments. 10