STATISTICS
Paper III

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 divided under **TWO** Sections.

Candidate has to attempt **FIVE** questions in all.

Both the questions in Section **'A'** are compulsory.

Out of the **SIX** questions in Section **'B'**, any **THREE** questions are to be attempted.

The number of marks carried by a question/part is indicated against it.

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

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

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 (QCA) Booklet must be clearly struck off.

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

SECTION 'A'

1.(a) Consider a population of size N. Let S_1 be a simple random sample of size n_1 drawn without replacement. Another simple random sample S_2 of size n_2 was also drawn without replacement from the remaining population.

(i) Find the probability of obtaining the combined sample $S_1 \cup S_2$ from the population.

(ii) Define $\hat{\bar{Y}}_\alpha = \alpha\hat{\bar{Y}}_1 + (1-\alpha)\hat{\bar{Y}}_2$, $0<\alpha<1$. Show that $\hat{\bar{Y}}_\alpha$ is an unbiased estimator for the population mean. Here $\hat{\bar{Y}}_i$ is the mean of sample S_i, $i = 1, 2$. (7+8)

1.(b) Consider the multiple regression model with a set of linear equality restrictions binding the regression coefficients.

(i) Derive the restricted regression estimator by minimizing the residual sum of squares under the set of restrictions.

(ii) Obtain the bias of the restricted regression estimator when the restrictions may not be true. Show that the estimator is unbiased and satisfies linear restrictions provided the restrictions are true. (6+4)

1.(c) (i) Explain the steps in constructing a consumer price index and discuss a method of its construction.

(ii) Indicate the precautions required while using the consumer price index numbers. (10+5)

2.(a) (i) Describe a ratio estimator. Obtain the bias of a ratio estimator.

(ii) Under standard notations it is given
$N = 10{,}000$, $n = 100$, $\bar{X} = 50$, $\hat{\bar{Y}} = 4500$, $\hat{\bar{X}} = 45$, $s_y^2 = 25$, $s_x^2 = 16$, and the sample correlation coefficient between x and y, $\hat{\rho} = 0{\cdot}8$.
Estimate the population mean using ratio estimator and its variance. (4+6)

2.(b) Consider the multiple linear regression model $y = X\beta + u$ where $y : n \times 1$, $X : n \times k$, $\beta : k \times 1$, $u : n \times 1$.

(i) How are the properties of ordinary least squares estimator affected when X and u are correlated and plim $\left(\frac{1}{n} X'u\right)$ is also not equal to zero ?

(ii) Derive the instrumental variables (IV) estimator of β when the number of instrumental variables is greater than the number of explanatory variables. Give its interpretation as a two-stage least squares estimator. (5+10)

2.(c) (i) Define exponentially weighted moving average (EWMA) for smoothing a time series and how can we use it for adoptive forecasting.

(ii) When do we use the triple exponential smoothing of Holt and Winters ? Describe it for multiplicative seasonality. (7+8)

SECTION 'B'

3.(a) (i) Distinguish between sampling and non-sampling errors.

(ii) Describe the effect of non-response on the bias of the sample estimate of the population mean under simple random sampling. (5+5)

3.(b) We are interested in the average age of a large population of employees in a particular service sector. The population is stratified based on the information of their age. A simple random sample was then taken from each strata with a total size 100. Under standard notations, using the information given in the table and ignoring fpc

Stratum age	$\frac{N_h}{N}$	$\bar{Y}_h$	S_h^2	n_h
< 40	50%	25	16	40
$40 - 50$	30%	45	10	20
> 50	20%	58	20	40

(i) Give the stratified estimator of the population mean. Is this estimator different from simple mean calculated over all the sample ?

(ii) Give the variance of the estimator in (i).

(iii) What would have been the gain in precision had you resorted to proportional allocation ? Obtain the strata sample size using proportional allocation.

(3+4+8)

3.(c) For estimating mean of the population, show that systematic sampling is more effective in removing the effect of a suspected linear trend in the population than simple random sampling. 15

4.(a) Consider the following two variables linear model with observations taken as deviation from mean :

$y_t = \beta x_t + u_t$; $t = 1, 2, ..., n$

where u_t's follow AR(1) process

$u_t = \rho u_{t-1} + \epsilon_t$, $|\rho| < 1$,

ϵ_t's are iid with mean O and variance σ_ϵ^2.

(i) Obtain the OLS and GLS estimators of β, say b and $\hat{\beta}$ respectively.

(ii) If r_1 is the sample autocorrelation of $\{x_t, t = 1, ..., n\}$ of order one and terms of order $O(\rho^k r_l)$ are ignored for all $(k+l) \geq 3$, then obtain $\frac{\text{Variance}(\hat{\beta})}{\text{Variance}(b)}$.

Comment on the relative efficiency of GLS estimator over OLS estimator. (7+8)

4.(b) (i) Define finite and infinite distributed lag models.

(ii) How can we estimate parameters and lag length of finite distributed lag models.

(iii) Consider geometric distributed lag model as a special case of infinite distributed lag model and obtain its mean lag and median lag. (3+4+8)

4.(c) (i) Explain the method of principal component regression for handling the multicollinearity problem.

(ii) How can we select the number of principal components to be omitted using the scree plot and variance explained criterion ? (6+4)

5.(a) Explain the technique of 'concentration curves' assuming Pareto's form of income distribution. 15

5.(b) Define price elasticity of demand. Given that demand function is $p = 7 - 8x^2$, then for what value of x, the elasticity of demand will be unity ? [Here x is the quantity demanded and p is the price]. 10

5.(c) (i) Describe the role of CSO in Indian Statistical System.

(ii) Distinguish between 'Income Method' and 'Expenditure Method' of estimating the national income in India. (8+7)

6.(a) (i) Define a Horvitz-Thompson estimator.

(ii) For simple random sampling without replacement, show that the HT estimator is unbiased for the population mean and hence deduce its variance. (4+6)

6.(b) (i) Describe the purpose of double sampling in ratio method of estimation.

(ii) A simple random sample of 15 orchards was selected from 200 orchards in the region to estimate the average yield of apples in the region. In each of the selected 15 orchards, the average yield during the previous session was obtained. For a subsample of 8 orchards from this 15 orchards, the average yield of apples in this season was recorded. Estimate the average yield of apples from all 200 orchards using all the information given in the table below :

Orchard No.	1	2	3	4	5	6	7	8
Yield : previous session	140	119	109	139	127	145	146	188
Yield : present session	128	108	101	125	118	135	133	175
Orchard No.	9	10	11	12	13	14	15	
Yield : previous session	121	144	105	155	115	142	171	

(5+10)

6.(c) (i) Describe cluster sampling. State the sampling variance of the unbiased estimator of population mean in terms of the intra cluster correlation.

(ii) A survey is carried out from a simple random sample of 90 clusters each of 40 households. The clusters are selected using simple random sampling at the rate $f = 1/300$. To improve accuracy of results, a statistician proposes to reduce by half the size of the clusters by selecting twice as many of them. Show that the relative efficiency is

$$\frac{1+19\rho}{1+39\rho}$$

assuming that the inter cluster correlation ρ and other quantities do not change. (7+8)

7.(a) (i) Define endogeneous variables, exogeneous variables and predetermined variables.

(ii) Consider the general structural form of simultaneous equations model and explain the problem of identification using its likelihood function.

(iii) Determine the rank and order conditions for the identification of a particular equation. (2+6+7)

7.(b) Consider the following system of equations :

(i) $y_{1t} = \beta_{10} + \beta_{11} y_{2t} + \gamma_{11} x_{1t} + \gamma_{12} x_{2t} + u_{1t}$

(ii) $y_{2t} = \beta_{20} + \beta_{21} y_{1t} + u_{2t}$

where y's are endogeneous and x's are predetermined variables. Check the identifiability of both the equations. If any of these two equations is identifiable, explain the method of two-stage least squares for estimating the parameters of that equation. 10

7.(c) (i) For estimating the parameters of a general structural model, describe the methods of indirect least squares (ILS) and two stage least squares (2-SLS).

(ii) Show that for a just identified equation, the two estimators are identical. (10+5)

8.(a) (i) Define a moving average process of order q(MA(q) process) and derive its autocorrelation function.

(ii) Stating the underlying conditions, write an MA(1) process as an autoregressive process of infinite order.

(iii) How can we identify and select the order of autoregressive and moving average processes. (8+3+4)

8.(b) (i) Define spectral density function and normalized form of spectral density function of a stationary process. Prove that the normalized spectral density function is a Fourier transform of autocorrelation function (ACF).

(ii) Derive the spectral density function of an AR(1) process. (10+5)

8.(c) (i) Establish the relationship between periodogram and sample autocovariance function (ACF).

(ii) Show that the periodogram provides an estimate of the spectral density function. What kind of preprocessing of data is required before performing the periodogram analysis ? (5+5)