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

Both the questions are compulsory

1. *(a)* Describe the advantages of sampling versus complete enumeration. Write the circumstances under which complete enumeration is preferred to sampling. 10

(b) In the regression model $y_i = \alpha + \beta x_i + u_i$, $i = 1, 2, \cdots, n$, if the sample mean $\bar{x}$ of x is zero, show that $\text{cov}(\hat{\alpha}, \hat{\beta}) = 0$, where $\hat{\alpha}$ and $\hat{\beta}$ are the least square estimators of α and β. Assume that $u_1, u_2, \cdots, u_n$ are independent and are from $N(0, \sigma^2)$ distribution. 15

(c) Define stationary time series process and autocovariance function. Show that autocovariance function, denoted by $\gamma(h)$, is an even function, positive semi-definite and uniformly continuous if it is continuous at $h = 0$. 15

2. *(a)* Assume that the correlation function of a continuous parameter process is given by $\rho(t) = ae^{-b|t|}$, $a, b > 0$. Find the spectral density function for the process. Comment on the density curve with respect to its parameter. (Here, t is the time lag) 10

(b) Define simple random sampling with replacement and without replacement. In simple random sampling without replacement, show that the sample mean $\bar{y}$ is unbiased estimate of $\bar{Y}$, the population mean. Derive its variance. Also, find the standard error of the estimate of the population total. 15

(c) Describe the problem of heteroscedasticity. How does one detect it? What happens to OLS estimators if we introduce heteroscedasticity? 15

SECTION—B

Answer any THREE questions of the SIX questions given below

3. *(a)* Explain the concept of interpenetrating sub-sampling. 10

(b) Define ratio estimator for the population ratio, R. Obtain its bias. In SRSWOR, for large n, derive its variance. Deduce the variance of the estimator for population total. 15

(c) Define Des Raj ordered estimator for population mean for the case when the sample size is 2. Show that it is unbiased. Derive its variance. 15

4. *(a)* Consider an econometric model where two variables Y and X are jointly determined by the following equations :

$$Y = \beta_1 + \beta_2 X + \beta_3 Z + U$$

$$X = \alpha_1 + \alpha_2 Y + V$$

The Greek letters denote unknown parameters, U and V are model errors, mutually uncorrelated, with zero mean, and Z is an exogenous variable, independent of the errors. Show that OLS estimator of α_2 is asymptotically biased in general. 10

(b) Give some practical examples for reasons of lags in distributed lag models and explain the Koyck approach to distributed lag models. 15

(c) Consider a regression function $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u$ with $u \sim IN(0, \sigma^2)$, which relates annual salary (y) of employees for a large firm in terms of their years of education (x_1) and years of experience (x_2). The following data are obtained from a sample of size $n = 23$:

$\bar{x}_1 = 10$	$\bar{x}_2 = 5$	$\bar{y} = 12$
$s_{11} = 12$	$s_{12} = 8$	$s_{22} = 12$
$s_{1y} = 10$	$s_{2y} = 8$	$s_{yy} = 10$

(i) Compute $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$. Present the regression equation.

(ii) Determine residual sum of squares and multiple coefficient of determination.

(iii) Test the hypothesis $\beta_2 = 0$ at the 5% significance level.

(Symbols used have their usual interpretation) 15

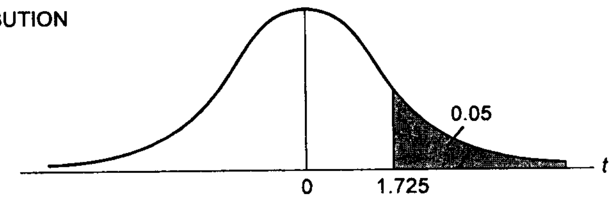
PERCENTAGE POINTS OF THE *t*-DISTRIBUTION

Example :

Pr (t > 2.086) = 0.025

Pr (t > 1.725) = 0.05 for df = 20

Pr ($|t|$ > 1.725) = 0.10



df \ Pr	0.25 0.50	0.10 0.20	0.05 0.10	0.025 0.05	0.01 0.02	0.005 0.010	0.001 0.002
1	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	0.816	1.886	2.920	4.303	6.965	9.925	22.327
3	0.765	1.638	2.353	3.182	4.541	5.841	10.214
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	0.686	1.323	1.721	2.080	2.518	2.831	3.527
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505
23	0.685	1.319	1.714	2.069	2.500	2.807	3.485
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.684	1.316	1.708	2.060	2.485	2.787	3.450
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232
120	0.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.090

Note : The smaller probability shown at the head of each column is the area in one tail; the larger probability is the area in both tails.

5. *(a)* The prices and quantities of four commodities produced by a manufacturing firm during the period 2015 and 2016 are given below :

Commodity	*Quantity* (in kg) *during*		*Price* (in ₹) *during*	
	2015	*2016*	*2015*	*2016*
A	175	201	1,540	1,030
B	32	46	1,270	1,490
C	48	43	2,760	2,490
D	65	66	2,190	2,070

Calculate—

(i) Laspeyres' index number;

(ii) Paasche's index number;

(iii) Fisher's index number

for 2016 assuming 2015 as the base year. 10

(b) What is meant by correlogram in time series analysis? Obtain the general expressions for correlograms of AR(1) and MA(q) models. 15

(c) State ergodicity property of a stationary time series process. If $\{X_t, t \in T\}$ is a stationary time series process with $E(X_t) = \mu$ and $\text{var}(X_t) = \sigma^2$ for all t, define ergodicity of μ. Also, show that the process whose covariance function $\gamma(h) \to 0$ as $h \to \infty$ is ergodic for μ. 15

6. *(a)* Distinguish between two-stage sampling and two-phase sampling procedures. State the situations in which they are used. 10

(b) Describe systematic sampling procedure when the population size is an integer multiple of sample size. Mention its advantages. Compare systematic sampling with stratified sampling and cluster sampling. 15

(c) Show that in a stratified sampling with a linear cost function of the form $C = c_o + \sum_{h=1}^{L} c_h n_h$ the variance of the estimated mean $\bar{y}_{st}$ is a minimum for a specified cost C and the cost is a minimum for a specified variance $V(\bar{y}_{st})$, when $n_h \propto W_h S_h / \sqrt{c_h}$.

(Here, c_o is the overhead cost, L is the number of strata, c_h is the cost per unit for stratum h, n_h is the sample size of stratum h, W_h is the weight of stratum h and S_h^2 is the true variance of stratum h) 15

7. (a) Define multicollinearity for the general linear model. What are the methods of detecting it? Also, state the effect on OLS estimators of the regression coefficients and their variances, if there is perfect multicollinearity between two explanatory variables of the model $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + U$. 10

(b) Explain the problem of identification. Does it have any connection with multicollinearity? Develop the rank and order conditions of identifiability. 15

(c) Estimate the coefficients of the model $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i$ with the restriction on the coefficients as $\beta_2 + \beta_3 = 1$, based on a sample of 25 observations, which yielded the following information :

$$X'X = \begin{pmatrix} 20 & 0 \\ 0 & 40 \end{pmatrix} \text{ and } X'\underline{Y} = \begin{pmatrix} 15 \\ 25 \end{pmatrix}$$

[Assume that $U_i \sim IN(0, \sigma^2)$] 15

8. (a) State Pareto's law of income distribution. Using Pareto's law, derive the probability density of income variable X, where $a < X < b$, with a and b being the lowest and highest income, respectively. 10

(b) State the general form of spectral density function for a stationary time series process. Derive the expressions for spectral density functions of MA(1) and AR(1) processes. 15

(c) Enunciate, in detail, the principal stages involved in setting up a Box-Jenkins forecasting model. Also, write a note on Box-Pierce Q-statistic involved in autocorrelation tests. 15

★ ★ ★