DO NOT OPEN THIS TEST BOOKLET UNTIL YOU ARE ASKED TO DO SO

23-0009-AZ

TEST BOOKLET

STATISTICS

PAPER – I

Time Allowed: 3 hours

Maximum Marks: 300

INSTRUCTIONS TO CANDIDATES

Read the instructions carefully before answering the questions: -

1. This Test Booklet consists of 16(Sixteen) pages and has 75 (seventy-five) items (questions).
2. IMMEDIATELY AFTER THE COMMENCEMENT OF THE EXAMINATION, YOU SHOULD CHECK THAT THIS BOOKLET ***DOES NOT*** HAVE ANY UNPRINTED OR TORN OR MISSING PAGES OR ITEMS ETC. IF SO, GET IT REPLACED BY A COMPLETE TEST BOOKLET.
3. **Please note that it is the candidate's responsibility to fill in the Roll Number and other required details carefully and without any omission or discrepancy at the appropriate places in the OMR Answer Sheet and the Separate Answer Booklet. Any omission/discrepancy will render the OMR Answer Sheet and the Separate Answer Booklet liable for rejection.**
4. Do not write anything else on the OMR Answer Sheet except the required information. Before you proceed to mark in the OMR Answer Sheet, please ensure that you have filled in the required particulars as per given instructions.
5. Use **only Black Ball Point Pen** to fill the OMR Answer Sheet.
6. This Test Booklet is divided into 4 (four) parts – **Part – I, Part – II, Part - III** and **Part – IV**.
7. **All three parts are Compulsory.**
8. **Part-I consists of Multiple Choice-based Questions.** The answers to these questions have to be marked in the **OMR Answer Sheet** provided to you.
9. **Part-II, Part-III** and **Part-IV consist of Conventional Essay-type Questions.** The answers to these questions have to be written in the separate **Answer Booklet** provided to you.
10. In Part-I, each item (question) comprises of 04 (four) responses (answers). You are required to select the response which you want to mark on the OMR Answer Sheet. In case you feel that there is more than one correct response, mark the response which you consider the best. In any case, choose ***ONLY ONE*** response for each item.
11. After you have completed filling in all your responses on the OMR Answer Sheet and the Answer Booklet(s) and the examination has concluded, you should hand over to the Invigilator ***only the OMR Answer Sheet and the Answer Booklet(s)***. You are permitted to take the Test Booklet with you.
12. **Penalty for wrong answers in Multiple Choice-based Questions:**

THERE WIL BE **PENALTY** FOR WRONG ANSWERS MARKED BY A CANDIDATE.

(i) There are four alternatives for the answer to every question. For each question for which a wrong answer has been given by the candidate, one-third of the marks assigned to the question will be deducted as penalty.

(ii) If a candidate gives more than one answer, it will be treated as a wrong answer even if one of the given answers happens to be correct and there will be same penalty as above to the question.

(iii) If a question is left blank. i.e., no answer is given by the candidate, there will be no penalty for that question.

DO NOT OPEN THIS TEST BOOKLET UNTIL YOU ARE ASKED TO DO SO

PART - I
(Multiple Choice-based Questions)

Instructions for Questions 1 to 50:

- *Choose the correct answers for the following questions.*
- *Each question carries 3 marks.* *[3x50=150]*

1. A bag contains 10 red balls and 5 blue balls. What is the probability of drawing a red ball from the bag?
   (a) 1/2
   (b) 2/3
   (c) 3/5
   (d) 5/8

2. What is the BLUE property of the Gauss-Markov theorem?
   (a) It states that the least squares estimator is unbiased.
   (b) It states that the least squares estimator has minimum variance among all unbiased linear estimators.
   (c) It states that the least squares estimator is consistent.
   (d) It states that the least squares estimator is efficient.

3. Which of the following statements is true about the convergence of a sequence of random variables in probability?
   (a) It implies almost sure convergence
   (b) It implies convergence in mean squared
   (c) It is equivalent to convergence in distribution for continuous random variable
   (d) It implies convergence in distribution

4. Which of the following statements is true about the central limit theorem?
   (a) It states that a sequence of random variables converges to a normal distribution in probability
   (b) It states that a sequence of random variables converges to a normal distribution in distribution
   (c) It states that a sequence of independent and identically distributed random variables with finite variance converges in distribution to a normal distribution
   (d) It states that a sequence of independent and identically distributed random variables with finite mean converges in distribution to a normal distribution.

5. Which of the following is true about the standard error of the mean?
   (a) It increases as the sample size increases
   (b) It decreases as the sample size increases
   (c) It is equal to the sample size
   (d) It is equal to the sample standard deviation

6. Which of the following is a nonparametric test?
   (a) t-test
   (b) ANOVA
   (c) Wilcoxon rank-sum test
   (d) F-test

7. What will be the value of $P(\bar{E})$ if P(E) = 0.07?
   (a) 0.90
   (b) 0.007
   (c) 0.93
   (d) 93

8. In tossing three coins at a time the probability of getting at the most one head is?
   (a) 3/8
   (b) 7/8
   (c) 1/2
   (d) 1/8

9. In a class, 30% of the students study Hindi, 45% study Math, and 15% study both Hindi and Math. If a student is randomly selected, what is the probability that he/she studies Hindi or Math?
   (a) 1/5
   (b) 3/5
   (c) 2/5
   (d) 2/7

10. In binomial distribution successive trials are -
   (a) Mutually exclusive
   (b) Dependent
   (c) Independent
   (d) None of the above

11. A probability density function is of the form $p(x)=Ke^{-\alpha|x|}$, x □$(-\infty, \infty)$, the value of K is
   (a) 0.5
   (b) 1
   (c) 0.5a
   (d) A

12. Which of the following is a type of ANOVA that is used to analyse one variable only?
   (a) One-way ANOVA
   (b) Two-way ANOVA
   (c) Three-way ANOVA
   (d) N-way ANOVA

13. Ratio estimation involves estimating the population parameter by multiplying the sample statistic with a known _____ and dividing it by the same value for the population.
   (a) Mean
   (b) Variance
   (c) Total
   (d) None of the above

14. In a popular shopping centre, the waiting time for an ABSA ATM machine is found to be uniformly distributed between 1 and 5 minutes. What is the probability of waiting between 2 and 4 minutes to use the ATM?
    (a) 0.25
    (b) 0.50
    (c) 0.75
    (d) 0.20

15. The mean of an exponential distribution happens to be equal to the mean of the Poisson distribution when the mean of the former distribution happens to be equal to:
    (a) 2.0
    (b) 0.25
    (c) 0.5
    (d) 1.0

16. A larger standard deviation for a normal distribution with an unchanged mean indicates that the distribution becomes
    (a) Narrower and more peaked
    (b) Flatter and wider
    (c) More skewed to the right
    (d) More skewed to the left

17. The boundaries of a confidence interval are called -
    (a) Confidence levels
    (b) The test statistics
    (c) The degrees of confidence
    (d) The confidence limits

18. What value of z would you use to calculate the 80% confidence interval for a population mean, given that you know the population standard deviation, the sample size and the sample mean of your sample?
    (a) $z = 1.96$
    (b) $z = 2.58$
    (c) $z = 0.84$
    (d) $z = 1.28$

19. In the formula for the confidence interval, $z_{\alpha/2}$ is part of the formula. What does the subscript $\alpha/2$ refer to?
    (a) the level of confidence
    (b) the level of significance
    (c) the probability that the confidence interval will contain the population mean
    (d) the area in the lower tail or upper tail of the sampling distribution of the sample mean

20. Least square estimators of the parameters of a linear model are
    (a) Unbiased
    (b) BLUE
    (c) UMVU
    (d) All of the above

21. The p-value of a test is the -
    (a) smallest significance level at which the null hypothesis cannot be rejected
    (b) largest significance level at which the null hypothesis cannot be rejected
    (c) smallest significance level at which the null hypothesis can be rejected
    (d) largest significance level at which the null hypothesis can be rejected

22. A measure of the variability of the treatment effects in an experimental design -
    (a) Mean square error (MSE)
    (b) Mean square treatment (MST)
    (c) Mean square block (MSB)
    (d) Mean square interaction (MSI)

23. In cluster sampling, the population is divided into _____ and a random sample of clusters is selected.
    (a) Homogeneous subgroups
    (b) Strata
    (c) Clusters
    (d) None of the above

24. A researcher fits a linear regression model to a dataset with 50 observations. The model has one predictor variable and an intercept term. The residual sum of squares (RSS) is 300 and the total sum of squares (TSS) is 500. What is the R-squared value of the model?
    (a) 0.4
    (b) 0.6
    (c) 0.8
    (d) 1.0

25. A company wants to test whether the mean weight of a product is different from 10 pounds. The company collects 25 samples and calculates the sample mean and sample standard deviation. The null hypothesis is that the mean weight is 10 pounds. If the test statistic is -2.5, what is the p-value?
    (a) 0.01
    (b) 0.02
    (c) 0.05
    (d) 0.10

26. Mahalanobis' D2 statistic is used to:
    (a) Test for normality of the error term in a regression model
    (b) Test for homoscedasticity of the error term in a regression model
    (c) Measure the distance between a point and the mean of a multivariate normal distribution
    (d) Test for the equality of two covariance matrices

27. What is Wald's fundamental identity?
    (a) A mathematical formula that relates the expectation of a random variable to the conditional expectation of the same variable.
    (b) A statistical method for analyzing data in a regression model.
    (c) An economic theory that explains consumer behaviour.
    (d) None of these

28. What is the formula for Wald's fundamental identity?
    (a) $E[X] = E[E[X|Y]] + E[\varepsilon]$
    (b) $E[X] = E[X|Y] + E[\varepsilon]$
    (c) $E[X] = E[X] + E[\varepsilon]$
    (d) $E[X] = E[E[X|Y]] - E[\varepsilon]$

29. Which of the following is a measure of the precision of the entire regression model, rather than just an individual regression coefficient?
    (a) The f-statistic.
    (b) The t-statistic.
    (c) The p-value.
    (d) The coefficient of determination (R-squared).

30. What is the interpretation of the standard error of the least squares estimate of a regression coefficient?
    (a) It is a measure of the spread of the predictor variable.
    (b) It is a measure of the precision of the estimated regression coefficient.
    (c) It is a measure of the strength of the relationship between the predictor and response variables.
    (d) It is a measure of the bias in the estimated regression coefficient.

31. In linear regression the least squares estimate of the regression coefficients are obtained by:
    (a) Minimizing the sum of the absolute residuals.
    (b) Minimizing the sum of the squared residuals.
    (c) Maximizing the correlation coefficient between the predictor and response variables.
    (d) Maximizing the R-squared value.

32. A linear regression model is fitted to a dataset with 50 observations. The least squares estimate of the regression coefficient for the predictor variable is 0.6. What is the standard error of the estimate for the regression coefficient?
    (a) 0.16
    (b) 0.20
    (c) 0.25
    (d) 0.30

33. What is the assumption of homoscedasticity in the Gauss-Markov theorem?
    (a) The errors have a normal distribution.
    (b) The errors have constant variance.
    (c) The errors are independent.
    (d) The errors have zero mean.

34. A multiple linear regression model is fitted to a dataset with 100 observations and 3 predictor variables. The coefficient of determination (R^2) is found to be 0.65. What is the adjusted R-squared value for this model?
    (a) 0.62
    (b) 0.63
    (c) 0.64
    (d) 0.65

35. Which of the following measures is used to assess the goodness of fit of a regression model?
    (a) Mean absolute deviation (MAD)
    (b) Mean squared error (MSE)
    (c) Standard deviation (SD)
    (d) Variance

36. What is the difference between standard error and standard deviation?
    (a) Standard error measures the spread of the sample means, while standard deviation measures the spread of the individual data points.
    (b) Standard error measures the spread of the individual data points, while standard deviation measures the spread of the sample means.
    (c) Standard error is a measure of accuracy, while standard deviation is a measure of variability.
    (d) Standard error is used in regression analysis, while standard deviation is used in hypothesis testing.

37. What is the standard deviation of the residuals for a regression model with a total sum of squares of 100, an explained sum of squares of 60, and a sample size of 20?
    (a) 2.24
    (b) 2.50
    (c) 3.16
    (d) 3.54

38. A box contains 4 red balls, 3 green balls, and 2 blue balls. What is the probability of drawing a green ball without replacement?
    (a) 1/3
    (b) 3/9
    (c) 3/8
    (d) 2/7

39. What does a correlation coefficient of -0.8 indicate?
    (a) A strong negative relationship between the two variables.
    (b) A strong positive relationship between the two variables.
    (c) A weak negative relationship between the two variables.
    (d) A weak positive relationship between the two variables.

40. What is the equation of a simple linear regression line?
    (a) $y = mx + b$
    (b) $y = b + x/m$
    (c) $y = m/x + b$
    (d) $y = b - mx$

41. What is the definition of consistency in the Gauss-Markov theorem?
    (a) The estimator is equal to the true value on average.
    (b) The estimator has a normal distribution.
    (c) The estimator has minimum variance.
    (d) The estimator approaches the true value as the sample size increases.

42. A linear regression model is fitted to a dataset with 100 observations. The coefficient of determination (R^2) is found to be 0.75. What is the correlation coefficient between the predictor and response variables?
    (a) 0.25
    (b) 0.50
    (c) 0.75
    (d) 0.87

43. Factorisation theorem for sufficiency is known as -
    (a) Rao –Blackwell theorem
    (b) Crammer- Rao theorem
    (c) Chapman-Robins theorem
    (d) Fisher-Neyman theorem

44. The probability mass function of a random variable X is given by $p(x) = 0.1x$ for $x = 1, 2, 3, 4, 5$. What is E(X)?
    (a) 1.5
    (b) 2.5
    (c) 3.5
    (d) 4.5

45. What is the standard error of the estimate for a regression model with a mean squared error of 25 and 50 observations?
    (a) 0.5
    (b) 1
    (c) 2
    (d) 5

46. A coin is flipped 3 times. What is the probability of getting exactly 2 heads?
    (a) 1/8
    (b) 3/8
    (c) 1/2
    (d) ¾

47. The probability mass function of a random variable X is given by $p(x) = 0.3$ for $x = 1, 2, 3$ and $p(x) = 0.1$ for $x = 4$. What is $P(X > 2)$?
    (a) 0.3
    (b) 0.4
    (c) 0.7
    (d) 0.9

48. What is the condition for a linear estimator to be efficient?
    (a) The estimator must be unbiased.
    (b) The estimator must have minimum variance.
    (c) The estimator must have a normal distribution.
    (d) The estimator must be consistent.

49. What is the Gauss-Markov theorem's implication for the OLS estimator?
    (a) The OLS estimator is biased.
    (b) The OLS estimator is consistent.
    (c) The OLS estimator is inefficient.
    (d) The OLS estimator is inconsistent.

50. In a class of 30 students, 15 are girls and 15 are boys. If a student is chosen at random, what is the probability that the student being a boy?
    (a) 1/2
    (b) 2/3
    (c) 1/3
    (d) 1/4

PART - II
(Short Answer-type Questions)

Instructions for Questions 51 to 63:

- ***Write the answers in short for any 10 (TEN) out of the thirteen questions.***
- ***Each question carries 5 marks.*** ***[5x10=50]***

51. If the letters of the word 'REGULATIONS' be arranged at random, what is the chance that there will be exactly 4 letters between R and E?

52. In a bolt factory, machines A, B, C manufacture 25%, 35% and 40% respectively of the total. Of their output 5, 4, 2 per cent are known to be defective bolts. A bolt is drawn at random from the product and is found to be defective. What are the probabilities that it was manufactured by (i) Machine A (ii) Machine B or Machine C?

53. Suppose you are given a discrete random variable X with the following probability mass function:

X	1	2	3	4
P(X=x)	0.1	0.2	0.3	0.4

    (a) Find the expected value and variance of X.
    (b) Let $Y = X^2$. Find the probability mass function of Y.

54. Let $X_1, X_2, \ldots$ Be a i.i.d. Poisson variates with parameter λ. Use CLT to estimate P $(120 \leq S_n \leq 160)$, where
$$S_n = X_1 + X_2 + \ldots + X_n \; ; \; \lambda = 2 \; and \; n = 75.$$

55. What is Hotelling's T^2 statistics and its properties.

56. Explain Kolmogorov's test for goodness of fit.

57. The following data represent the weights (in kg) of personal luggage carried on an aircraft by a member of 2 baseball clubs.

Club A	34	39	41	28	33	
Club B	36	40	35	31	39	36

Use Mann-Whitney U test with a = 0.05 to test the hypothesis that the two clubs carry the same amount of luggage on the average.

58. a) A stenographer claims that she can type at the rate of 120 words per minute. Can we reject her claim on the basis of 100 trials in which she demonstrates a mean of 116 words with standard deviation of 15 words? Use 5% level of significance.

b) A random sample of 900 items is taken from a normal population whose mean and the variance is 4. Can the sample with mean 4.5 be regarded as a truly random one and at 1% level of significance? (Table value at 1% level is 2.58).

59. Calculate the Spearman's ranks correlation coefficient for the following data and interpret the result:

X	35	54	80	95	73	73	35	91	83	81
Y	40	60	75	90	70	75	38	95	75	70

60. What is the difference between probability and non-probability sampling? Give an example of each.

61. What are non-sampling errors in sampling and how do they impact the accuracy of sampling estimates? How can they be minimized in sampling designs?

62. Three groups of children contain respectively 3 girls and 1 boy, 2 girls and 2 boys, and 1 girl and 3 boys. One child is selected at random from each group. Show that the chance that the three selected consist of 1 girl and 2 boys is 13/32.

63. In a partially destroyed record the following data are available:
Variance of X = 25.
Regression equation of x on y: 5x-y = 22.
Regression equation of y on x: 64x - 45y = 24.

Find:
(a) mean values of x and y
(b) coefficient of correlation between x and y
(c) Standard deviation of y

PART - III
(Long Answer-type Questions)

Instructions for Questions 64 to 71:

- *Answer any 5 (FIVE) out of the eight questions.*
- *Each question carries 10 marks.* *[10x5=50]*

64. Define Estimation and explain the methods of estimation with their properties.

65. A Computer while calculating correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results:

    $n = 25, \sum X = 125, \sum X^2 = 650, \sum X = 100, \sum X^2 = 460, \sum XY = 508,$

    It was however later discovered at the time of checking that he had copied down two pairs as

X	Y
6	14
8	6

    while the correct values were

X	Y
6	12
8	6

    obtain the correct value of correlation coefficient.

66. Suppose we have a population of 1000 households and we want to estimate the average income of the population. We take a stratified random sample of 200 households, with 100 households from a low-income stratum and 100 households from a high-income stratum. The incomes of the sampled households are as follows:
    Low-income stratum: 5000, 6000, 4000, 5500, 6500, 4500, 5800, 4200, 4700, 5300
    High-income stratum: 8000, 10000, 12000, 9000, 11000, 9500, 8500, 10500, 9500, 9000
    Assume that the strata are defined based on income levels and that each household within a stratum has an equal probability of being selected.

    (a) Calculate the Hansen-Hurwitz estimator for the average income of the population.
    (b) Calculate the Horvitz-Thompson estimator for the average income of the population
    (c) What is the difference between Regression and Correlation?

67. If X, Y are standardized random variables, and

$$r(aX + bY, bX + aY) = \frac{1+2ab}{a^2+b^2}$$

    Find r(X, Y), the coefficient of correlation between X and Y.

68. Urn I contain 2 white, 4 red and 6 black balls, urn Il contains 4 white, 2 red and 8 black balls and urn III contains 6 white, 6 red and 4 black balls. Two balanced dice are thrown. If the sum of face points is equal to 5 or less; urn I is selected. If the sum is 6 or 7 the urn III is selected, otherwise, urn II will be selected for drawing a ball. Given that the ball drawn is of white colour, what is the probability that it is drawn from urn II?

69. A company sells a product in four different regions of the country. The sales in each region are recorded, along with the advertising expenses for that region. The following table shows the data:

Region	Sales	Advertising Expense
1		20
2	150	25
3	180	30
4	200	35

( ) Write down the [illegible] equation for the linear regression model that relates sales to advertising expenses.

(b) Use the Gauss-Markov theorem to show that the least squares estimator for the regression coefficients is unbiased.

(c) Compute the least squares estimates for the regression coefficients β_0 and β_1.

(d) Interpret the estimated slope coefficient, β_1, in the context of the problem.

(e) Predict the sales for a region with advertising expenses of 40.

(f) Compute a 95% confidence interval for the true mean sales for a region with advertising expenses of 40.

70. What is an UMP test and how is it different from other hypothesis tests? Describe the steps involved in conducting an UMP test for a given hypothesis.? What are the advantages and limitations of the UMP test?

71. Define CRD, RBD, and LSD designs and explain analyses of LSD.

PART - IV
(Essay-type Questions)

Instructions for Questions 72 to 75:

- *Answer any 2 (TWO) out of the four questions.*
- *Each question carries 25 marks.* *[25x2= 50]*

72. State and prove Gauss Markoff theorem on linear estimation. How do these results modify when the variables are correlated?

73. Suppose that the time in minutes that a person has to wait at a certain station for a train is found to be a random phenomenon, a probability function specified by the distribution functions.

$F(x) = 0$, for $x \leq 0$

$= \frac{x}{2}$, for $0 \leq x < 1$

$= \frac{1}{2}$, for $1 \leq x < 2$

$= \frac{x}{4}$, for $2 \leq x < 4$

$= 1$, for $x \geq 4$

(a) Is the distribution function continuous? If so, give the formula for its probability function?

(b) What is the probability that a person will have to wait (i) more than 3 minutes, (ii) less than 3 minutes, and (iii) between 1 and 3 minutes?

(c) What is the conditional probability that the person will have to wait for a train for (i) more than 3 minutes, given that it is more than I minute, (ii) less than 3 minutes given that it is more than I minute?

74. a) Explain the Mann-Whitney U Test with the help of an example. Why is it used?

b) The following are the weight gains (in pounds) of two random samples of young boys fed on two different diets but otherwise kept under identical conditions:

Diet I	16.3	10.1	10.7	13.5	14.9	11.8	14.3	10.2
	12.0	14.7	23.6	15.1	14.5	18.4	13.2	14.0
Diet II	21.3	23.8	15.4	19.6	12.0	13.9	18.8	19.2
	15.3	20.1	14.8	18.9	20.7	21.1	15.8	16.2

Use the U test at 0.01 level of significance to test the null hypothesis that the two population samples are identical against the alternative hypothesis that on the average the second diet produces a greater gain in weight.

75. (A) Define systematic sampling and give its merits and demerits.
(B) Compare systematic sampling with stratified sampling for estimation of a population mean.