PART C — (1 × 15 = 15 marks)

16. (a) An investigator polls common cold sufferers, asking them to estimate the number of hours of physical discomfort caused by their most recent colds. Assume that their estimates approximate a normal curve with a mean of 83 hours and a standard deviation of 20 hours.

(i) What is the estimated number of hours for the shortest-suffering 5 percent?

(ii) What proportion of sufferers estimate that their colds lasted longer than 48 hours?

(iii) What proportion suffered for fewer than 61 hours?

(iv) What is the estimated number of hours suffered by the extreme 1 percent either above or below the mean?

(v) What proportion suffered for between 1 and 3 days, that is, between 24 and 72 hours?

(vi) What proportion suffered for between 2 and 4 days?

Or

(b) An investigator wishes to determine whether alcohol consumption causes a deterioration in the performance of automobile drivers. Before the driving test, subjects drink a glass of orange juice, which, in the case of the treatment group, is laced with two ounces of vodka. Performance is measured by the number of errors made on a driving simulator. A total of 120 volunteer subjects are randomly assigned, in equal numbers, to the two groups. For subjects in the treatment group, the mean number of errors $\overline{X}_1$ equals 26.4, and for subjects in the control group, the mean number of errors $\overline{X}_2$ equals 18.6. The estimated standard error equals 2.4.

(i) Use t to test the null hypothesis at the .05 level of significance.

(ii) Specify the p-value for this test result.

(iii) If appropriate, construct a 95 percent confidence interval for the true population mean difference and interpret this interval.

(iv) If the test result is statistically significant, use Cohen's d to estimate the effect size, given that the standard deviation, $S_{p,}$ equals 13.15.

Reg. No. : | | | | | | | | | | | | |

Question Paper Code : 30011

B.E./B.Tech. DEGREE EXAMINATIONS, APRIL/MAY 2023.

Third/Fourth Semester

Artificial Intelligence and Data Science

AD 3491 – FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

(Common to : Computer Science and Business Systems)

(Regulations 2021)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. What is brushing and linking in exploratory data analysis?
2. How does confusion matrix define the performance of classification algorithm?
3. Identify the correct distribution reflects the below scenario 'Reading achievement scores for a third-grade class consisting of about equal numbers of regular students and learning-challenged students"
4. How regression toward the mean differs other parameters? Give an example.
5. State the Central Limit Theorem.
6. Compare between population and sample.
7. What is the significance of p-value in hypothesis?
8. Compare between t-test and ANOVA.
9. Why do we need Goodness of Fit?
10. What is survival analysis?

PART B — (5 × 13 = 65 marks)

11. (a) (i) Exemplify in detail about different Facets of data with examples. (7)

(ii) Sketch and outline the step-by-step activities in the data science process. (6)

Or

(b) Explain in detail about Cleansing, integrating, and transforming data with examples.

12. (a) For each of the following pairs of distributions, first decide whether their standard deviations are about the same or different. If their standard deviations are different, indicate which distribution should have the larger standard deviation. Note that the distribution with the more dissimilar set of scores or individuals should produce the larger standard deviation regardless of whether, on average, scores or individuals in one distribution differ from those in the other distribution.

(i) SAT scores for all graduating high school seniors (a1) or all college freshmen (a2)

(ii) Ages of patients in a community hospital (b1) or a children's hospital (b2)

(iii) Motor skill reaction times of professional baseball players (c1) or college students (c2)

(iv) GPAs of students at some university as revealed by a random sample (d1) or a census of the entire student body (d2)

(v) Anxiety scores (on a scale from 0 to 50) of a random sample of college students taken from the senior class (e1) or those who plan to attend an anxiety-reduction clinic (e2)

(vi) Annual incomes of recent college graduates (f1) or of 20-year alumni (f2)

Or

(b) In a survey, a question was asked

"During your lifetime, how often have you changed your permanent residence?" a group of 18 college students replied as follows:

1, 3, 4, 1, 0, 2, 5, 8, 0, 2, 3, 4, 7, 11, 0, 2, 3, 3.

Find the mode, median, mean and standard deviation.

13. (a) (i) Reading achievement scores are obtained for a group of fourth graders. A score of 4.0 indicates a level of achievement appropriate for fourth grade, a score below 4.0 indicates underachievement, and a score above 4.0 indicates overachievement. Assume that the population standard deviation equals 0.4. A random sample of 64 fourth graders reveal a mean achievement score of 3.82. Construct a 95 percent confidence interval for the unknown population mean. (Remember to convert the standard deviation to a standard error.) Interpret this confidence interval; that is, do you find any consistent evidence either of overachievement or of underachievement? (6)

(ii) Illustrate in detail about estimation method and confidence interval. (7)

Or

(b) (i) For the population at large, the Wechsler Adult Intelligence Scale is designed to yield a normal distribution of test scores with a mean of 100 and a standard deviation of 15. School district officials wonder whether, on the average, an IQ score different from 100 describes the intellectual aptitudes of all students in their district. Wechsler IQ scores are obtained for a random sample of 25 of their students, and the mean IQ is found to equal 105. Using the step-by-step procedure, test the null hypothesis at the .05 level of significance. (7)

(ii) Imagine a simple population consisting of only five observations: 2, 4, 6, 8, 10. List all possible samples of size two. Construct a relative frequency table showing the sampling distribution of the mean. (6)

14. (a) (i) A library system lends books for periods of 21 days. This policy is being reevaluated in view of a possible new loan period that could be either longer or shorter than 21 days. To aid in making this decision, book-lending records were consulted to determine the loan periods actually used by the patrons. A random sample of eight records revealed the following loan periods in days: 21, 15, 12, 24, 20, 21, 13, and 16. Test the null hypothesis with *t-test*, using the .05 level of significance. (7)

(ii) A consumers' group randomly samples 10 "one-pound" packages of ground wheat sold by a supermarket. Calculate the mean and the estimated standard error of the mean for this sample, given the following weights in ounces:

16, 15, 14, 15, 14, 15, 16, 14, 14, 14 (6)

Or

(b) (i) Illustrate in detail about one factor ANOVA with example. (7)

(ii) A random sample of 90 College students indicates whether they most desire love, wealth power, health, fame, or family happiness. Using the .05 level of significance and the following results, test the null hypothesis that in the underlying Population, the various desires are equally popular using chi-square test. (6)

DESIRES OF COLLEGE STUDENTS

FREQUENCY	LOVE	WEALTH	POWER	HEALTH	FAME	FAMILY HAP.	TOTAL
Observed (f_o)	25	10	5	25	10	15	90

15. (a) (i) Describe in detail about logistic regression model in predictive analysis. (7)

(ii) Exemplify in detail about multiple regression model with example. (6)

Or

(b) Explain in depth about Time series analysis and its techniques with relevant examples.