

THE KAVERY ENGINEERING COLLEGE*(An Autonomous Institution)***B.E/B.TECH DEGREE END SEMESTER THEORY EXAMINATIONS, NOV/DEC 2025***Fourth Semester**Artificial Intelligence and Data Science***AD3491 - Fundamentals of Data Science and Analytics***Regulations - 2021**Duration : 3 Hrs**Maximum : 100 Marks***PART - A (ANSWER ALL QUESTIONS) (Marks 10*2=20)**

<i>Q.No</i>	<i>Questions</i>	<i>BLT</i>	<i>CO</i>	<i>Marks</i>
1.	What is the need for data science in modern industries?	L1	1	2
2.	Summarize the role of exploratory data analysis (EDA) in the data science process.	L2	1	2
3.	Outline about how does the interquartile range help in assessing the variability of a data set.	L2	2	2
4.	What is the significance of the standard error of estimate in regression analysis?	L2	2	2
5.	Critically evaluate the statement: "A larger sample size always leads to a more accurate representation of the population parameter." Under what specific conditions might this not hold true?	L5	3	2
6.	Explain, in your own words, the difference between a population and a sample. Provide a simple example to illustrate this difference.	L2	3	2
7.	In a two-factor ANOVA, what are the two main types of effects that are typically investigated?	L1	4	2
8.	What does the p-value represent in the context of hypothesis testing?	L2	4	2
9.	Define serial correlation (or autocorrelation) in time series data.	L1	5	2
10.	What does a high value of R ² (R-squared) indicate about the goodness of fit of a linear regression model?	L2	5	2

PART - B (ANSWER ALL QUESTIONS) (Marks 5*16 = 80)

<i>Q.No</i>	<i>Questions</i>	<i>BLT</i>	<i>CO</i>	<i>Marks</i>
11.	a Dissect the entire data preparation process in data science, focusing on data retrieval, cleansing, integration, and transformation. How do these steps contribute to building effective data models?	L4	1	16
	Or			
	b Analyze the benefits and uses of data science in various fields such as healthcare, finance, and marketing. Provide real-world examples to support your points.	L4	1	16
12.	a Evaluate the concept of least squares regression line. Derive its equation and explain its significance in predicting the dependent variable. How do you interpret the R-squared value in the context of model performance?	L5	2	16
	Or			
	b Critically assess the impact of outliers on the normality of a distribution. How do outliers influence measures of central tendency and variability, and what methods can be used to handle them?	L5	2	16

- | | | | | | |
|-----|---|--|----|---|----|
| 13. | a | <p>Consider two independent research studies investigating the same population mean. Study A uses a sample size of $n=50$, while Study B uses a sample size of $n=200$. Both studies calculate 90% confidence intervals for the population mean based on their respective samples.</p> <p>a) Explain why the sampling distribution of the mean will have a smaller standard error in Study B compared to Study A.</p> <p>b) Discuss how this difference in standard error will likely affect the width of the 90% confidence intervals calculated in the two studies. Justify your answer.</p> <p>c) If both studies obtain the same sample mean, which confidence interval (from Study A or Study B) would provide a more precise estimate of the population mean? Explain your reasoning.</p> <p>d) Explain how the concepts of the Central Limit Theorem and the Law of Large Numbers underpin the relationship between sample size and the precision of statistical estimates.</p> <p>e) Critically evaluate the practical implications of using a very large sample size in research. Are there any potential drawbacks or diminishing returns associated with extremely large samples?</p> | L3 | 3 | 16 |
|-----|---|--|----|---|----|

Or

- | | | | | | |
|-----|---|---|----|---|----|
| 13. | b | <p>A researcher is investigating the average study time of undergraduate students in a large university. They randomly select two independent samples of students: one from the arts faculty ($n_1 = 60$, sample mean = 15.2 hours/week, sample standard deviation = 4.5 hours/week) and another from the science faculty ($n_2 = 75$, sample mean = 16.8 hours/week, sample standard deviation = 5.2 hours/week).</p> <p>a) Formulate the null and alternative hypotheses to test if there is a significant difference in the average study time between the two faculties. Justify your choice of a one-tailed or two-tailed test.</p> <p>b) Explain the concept of the sampling distribution of the difference between two means and its relevance to this problem.</p> <p>c) Calculate the appropriate test statistic and determine the p-value. Show all your steps clearly.</p> <p>d) Using a significance level of $\alpha=0.05$, make a decision regarding the null hypothesis and interpret the results in the context of the research question.</p> <p>e) Discuss potential limitations of this study and suggest ways to improve the research design to enhance the validity of the findings.</p> | L3 | 3 | 16 |
|-----|---|---|----|---|----|
-
- | | | | | | |
|-----|---|---|----|---|----|
| 14. | a | <p>Provide a comprehensive theoretical introduction to chi-square tests. Discuss the fundamental principles underlying the chi-square goodness-of-fit test and the chi-square test of independence. Elaborate on the types of data suitable for chi-square analysis and the theoretical rationale behind comparing observed and expected frequencies. Discuss the theoretical differences between chi-square tests and t-tests or ANOVA in terms of the type of data analyzed and the research questions addressed.</p> | L3 | 4 | 16 |
|-----|---|---|----|---|----|

Or

	b	A marketing team wants to assess the effectiveness of a new advertising campaign. They measure customer satisfaction scores (on a scale of 1 to 10) before and after the campaign for a sample of 15 customers. Before campaign : 6, 7, 5, 8, 6, 9, 7, 5, 4, 6, 8, 7, 6, 5, 7 After campaign : 8, 9, 6, 9, 7, 9, 8, 7, 6, 8, 9, 8, 7, 6, 8 a) Explain why a paired-samples t-test (t-test for two related samples) is the appropriate statistical test for this situation. b) Calculate the mean and standard deviation of the difference scores. c) Calculate the t-test statistic for the paired samples. d) Determine the degrees of freedom and find the p-value for a two-tailed test. e) At a significance level of $\alpha=0.05$, can the marketing team conclude that the advertising campaign had a statistically significant impact on customer satisfaction? Interpret your findings.	L3	4	16
15.	a	Explain the concept of weighted resampling techniques in the context of regression analysis. Discuss the scenarios where weighted resampling might be necessary or beneficial. Describe at least one specific weighted resampling method (e.g., bootstrapping with weights) and elaborate on how it can be implemented to address issues such as heteroscedasticity or influential outliers. Critically evaluate the advantages and potential drawbacks of using weighted resampling compared to traditional least squares estimation.	L4	5	16
		Or			
	b	Explain the mathematical formulation of the linear least squares method for fitting a simple linear regression model. Detail the process of deriving the normal equations and discuss how these equations are used to estimate the slope and intercept of the best-fit line.	L4	5	16