Reg. No.															

Question Paper Code	13282

B.E. / B.Tech. - DEGREE EXAMINATIONS, NOV / DEC 2024

Fifth Semester

Information Technology

20ITPW501 - STATISTICAL ANALYSIS USING R PROGRAMMING WITH LABORATORY

Regulations - 2020

(Use of Statistical Tables is permitted)

Duration: 3 Hours Max. Marks: 100

PART - A (MCQ) (20 × 1 = 20 Marks)

Answer ALL Questions

		Marks	*K – Level*	*CO*

1.	What is the output of the following snippet?
	fun1 = function(fruit = "Apple", veg = "Carrot")
	{
	paste(fruit, veg)
	}
	fun1(veg = "Beans", fruit="Banana")
	
	(a) Banana Beans (b) Beans Banana (c) Apple Carrot (d) Error
		1	*K1*	*CO1*

2.	What will be the output of the following R code? > sqrt(-17) (a) -4.02 (b) 4.02 (c) NaN (d) 3.67	*1*	*K1*	*CO1*

3.	What is the output of following code
	fun1 = function(x,y)
	{
	return (x*y)
	}
	print(fun1(c(1:4),c(3:4)))
	(a) 3 8 9 16 (b) 1 2 3 4 (c) 3 4 (d) Error
		1	*K1*	*CO1*

4.	Which function generates random numbers from an exponential distribution in R? (a) rexp() (b) runif() (c) rnorm() (d) rpois()	*1*	*K1*	*CO2*
5.	Which function is used to create a histogram in R? (a) hist() (b) boxplot() (c) barplot() (d) plot()	*1*	*K1*	*CO2*
6.	Which function is used to create a box plot in R? (a) hist() (b) boxplot() (c) plot() (d) barplot()	*1*	*K1*	*CO2*
7.	In the context of hypothesis testing, which statement is most accurate? (a) A low p-value guarantees the null hypothesis is false (b) A high p-value guarantees the null hypothesis is true (c) P-values help assess the strength of evidence against the null hypothesis (d) P-values indicate the probability of the sample data	*1*	*K1*	*CO3*
8.	A researcher uses the wilcox.test function on two samples and gets a p-value of 0.04. What can they conclude at a 5% significance level? (a) Null hypothesis cannot be rejected (b) Null hypothesis should be rejected (c) Cannot conclude anything (d) High chances of error	*1*	*K1*	*CO3*
9.	Which of the following describes a two-sample T Test? (a) Tests for the mean of one group (b) Compares the means of two independent groups (c) Compares the medians of two related groups (d) Analyzes the variance within one group	*1*	*K1*	*CO3*
10.	Why are residuals important in assessing a regression model? (a) They help compute the slope of the regression line (b) They measure the goodness-of-fit of the model (c) They indicate whether the relationship is nonlinear (d) They determine the correlation between two variables	*1*	*K1*	*CO4*

11.	If the confidence interval for a slope coefficient in regression includes 0, what does it suggest? (a) The independent variable is not statistically significant (b) The relationship between the variables is positive (c) The correlation is high (d) The residuals are normally distributed	*1*	*K1*	*CO4*
12.	Which of the following is a key assumption of Pearson correlation? (a) The variables are nominal (b) The relationship between variables is linear (c) There are outliers in the data (d) The sample size is small	*1*	*K1*	*CO4*
13.	Which of the following best describes the Friedman Test? (a) A parametric test for independent samples (b)A non-parametric test for related samples (c) A test for equal variances among groups (d) A test for differences in proportions	*1*	*K1*	*CO5*
14.	When plotting residuals against fitted values in a multiple regression analysis, what are you primarily checking for? (a) Normality of the residuals (b) Independence of residuals (c) Homoscedasticity (d) Correlation between predictors	*1*	*K1*	*CO5*
15.	Which of the following assumptions must be met for the Kruskal–Wallis Test to be valid? (a) The samples must be independent (b) The samples must have equal variances (c) The data must be normally distributed (d) Both A and C	*1*	*K1*	*CO5*
16.	If an ANOVA table shows a very small p-value (< 0.001) for the model, what action would you take? (a) Reject the model as insignificant (b) Conclude that the model explains a significant amount of variance in the dependent variable (c) Perform a t-test for individual coefficients (d) Gather more data	*1*	*K1*	*CO5*
17.	In polynomial regression, what effect does increasing the degree of the polynomial have on the model? (a) It always improves the model's predictive accuracy (b) It increases the model's flexibility to fit more complex relationships (c) It reduces the model's likelihood of overfitting (d) It always leads to a simpler model	*1*	*K1*	*CO6*
18.	In a two-way ANOVA with replication, what does “replication” mean? (a) Having more than two independent variables (b) Having more than one observation for each combination of factors (c) Repeating the experiment under the same conditions (d) Adding more than two levels for each factor	*1*	*K1*	*CO6*
19.	Which R package provides advanced plotting options for polynomial regression models, especially with ggplot2 syntax? (a) `graphics` (b) `ggplot2` (c) `dplyr` (d) `tidyr`	*1*	*K1*	*CO6*
20.	In R, how to calculate the predicted values from a polynomial regression model? (a) Using `predict()` (b) Using `fitted()` (c) Using `lm.predict()` (d) Both A and B	*1*	*K1*	*CO6*

PART - B (10 × 2 = 20 Marks)

Answer ALL Questions

21.	Difference between a data frame and a matrix in R Program.	*2*	*K2*	*CO1*
22.	Write an R Program to Compute the mean of the square root of a vector of 100 random numbers.	*2*	*K2*	*CO1*
23.	State Boxplots.	*2*	*K1*	*CO2*
24.	Write an R Program to generate Random numbers.	*2*	*K2*	*CO2*
25.	List the differences between one sample t test and two sample t test.	*2*	*K2*	*CO3*
26.	Illustrate Wilcoxon signed rank test.	*2*	*K2*	*CO3*
27.	State the difference between cor() and cor.test() in R Program.	*2*	*K2*	*CO4*
28.	Differentiate between linear regression and multiple regression.	*2*	*K2*	*CO4*
29.	Write the difference between one-way ANOVA and two-way ANOVA.	*2*	*K2*	*CO5*
30.	Sate how can polynomial regression be used to model non-linear relationships.	*2*	*K2*	*CO6*

PART - C (6 × 10 = 60 Marks)

Answer ALL Questions

31. a) Please obtain the transpose matrix of B named tB.
Consider A=matrix(c(2,0,1,3), ncol=2) and B=matrix(c(5,2,4,-1), ncol=2).
(i) Find A + B — *5 K2 CO1*
(ii) Find A – B — *5 K2 CO1*

OR

b) i) Write a R program to create a factor corresponding to height of women data set, which contains height and weights for a sample of women. — *5 K2 CO1*

ii) Write a R program to print the numbers from 1 to 100 and print "Fizz" for multiples of 3, print "Buzz" for multiples of 5, and print "FizzBuzz" for multiples of both. — *5 K2 CO1*

32. a) Illustrate strip charts and histograms with examples and explain its importance with appropriate R codes. — *10 K2 CO2*

OR

b) i) Write R program to create pie chart for the following data. — *5 K2 CO2*
Housing --600, Food --300, Clothes -150, Entertainment—100,Others---200

ii) Explain how to plot multiple curves in the same graph for table data and explain with an example. — *5 K2 CO2*

33. a) What is a One Sample T Test? Explain with an example and how To Calculate a Test Statistic and accept or reject the null hypothesis with an example program. — *10 K2 CO3*

OR

b) Explain how to perform the two sample Wilcoxin test for any given data and write the appropriate R code. — *10 K2 CO3*

34. a) A researcher wants to determine if there is a significant monotonic relationship between hours spent studying and scores on a test. However, the data does not meet the assumption of normality, so the researcher decides to use Spearman's correlation test. The dataset contains the following values: — *10 K3 CO4*

Hours	2	4	3	5	6	7	8
Test Score	50	55	53	70	65	78	80

Find Spearman Correlation test statistics for x and y. Write the appropriate R program.

OR

b) A psychologist is studying the relationship between stress levels and sleep quality in a group of 10 individuals. The psychologist measures stress levels (on a scale of 1 to 10) and sleep quality (also on a scale of 1 to 10) but finds that the data contain tied ranks, so they decide to use Kendall's Tau correlation test. The dataset is as follows: — *10 K3 CO4*

Stress	4	7	5	9	6	3	8	5	6	2
Sleep	7	6	7	5	6	8	4	7	6	9

Do a rank correlation test of x and y using Kendall test. Write the appropriate R code.

35. a) — *10 K3 CO5*

		OVEN TEMPERATURE		
		325	350	400
TYPE OF SUGAR	WHITE SUGAR	10.75	8.75	4.00
		9.50	8.25	5.50
		10.00	9.00	4.75
		10.00	8.00	4.00
		9.25	8.25	5.00
	WHITE AND BROWN SUGAR	12.00	10.25	7.00
		10.00	9.00	7.25
		10.50	8.50	6.50
		11.25	10.50	5.00
		11.00	9.75	8.00

Given the weight of the cookies for different oven temperatures and different sugar types. Is there difference in weight of the cookies for differing sugar type and operating temperature levels? Use two way ANOVA to test the significance level. $f_{(1, 24),0.05} = 4.26$ (for Sugar), $f_{(2, 24),0.05} = 3.403$ (for temperature), $f_{(2, 24),0.05} = 3.403$ (for interaction).

OR

b) Given the survey results for 7 online stores for the last year find the equation of the straight line that fits the data best. Write an R program for the analysis. *10 K3 CO5*

Online Store	Monthly E-commerce Sales (in 1000 s)	Online Advertising Dollars (1000 s)
1	368	1.7
2	340	1.5
3	665	2.8
4	954	5
5	331	1.3
6	556	2.2
7	376	1.3

36. a) Given two samples, each comparing life expectancy vs. smoking for males and the second for females, determine whether there is any significant difference in the slopes for these two populations. Write the R code for comparing Regression Lines and Data Visualization *10 K3 CO6*

Men		Women	
Cig(x)	Life Exp(y)	Cig(x)	Life Exp(y)
5	80	22	88
23	78	7	95
25	60	20	86
48	53	23	60
17	85	15	82
8	84	34	75
4	73	4	80
26	79	40	68
11	81	8	93
19	75	16	77
14	68	11	72
35	72	52	67
29	58	3	90
4	92	31	66
23	65	18	72
		8	78

OR

b) Fit polynomial regression for the below data set using R. Write appropriate R code for data Visualization, Regression Fit and Draw the graphical representation *10 K3 CO6*

Position	Level	Salary
Business Analyst	1	45000
Junior Consultant	2	50000
Senior Consultant	3	60000
Manager	4	80000
Country Manager	5	110000
Region Manager	6	150000
Partner	7	200000
Senior Partner	8	300000
C-level	9	500000
CEO	10	1000000