Question Paper

Exam Date & Time: 30-Nov-2023 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

END SEMESTER EXAMINATION
Answer all the questions
Data missing if any, may be suitably assumed

DATA SCIENCE - PART II [CRA 4061]

Marks: 50 **Duration: 180 mins.**

Descriptive

Answer all the questions.

For Questions having subparts, clearly indicate the subpart before answering

1A) The following data represents the trunk width and tree height of 10 randomly chosen maple trees from Leominster State Forest. (5)

Width (in)	12	15	5	17	8	10	14	16	16	9
Height (ft)	26.6	29.3	10.2	34.7	15.8	22.1	27.6	24.9	32.6	22

a. Enter the data into R and Visualize with a plot.

b. Construct a Least-Squares Regression Model.

c. Calculate the Correlation and infer the relationship.

d. With a ggplot visualize the equation of the least-squares regression line.

e. Construct a Residuals Plot and infer the results.

1B) Write a R program to compute the equation of the fitted regression line for the relationship between hours studied and exam scores based on the data provided below. (3)

df < - data.frame(hours=c(1, 2, 4, 5, 5, 6, 6, 7, 8, 10, 11, 11, 12, 12, 14),

score=c(64, 66, 76, 73, 74, 81, 83, 82, 80, 88, 84, 82, 91, 93, 89))

Further based on the above data estimate the scores of student M if he studied for 5 hours.

1C) Suppose you track your commute times for two weeks (10 days) and you find the following times in minutes. (2)

17 16 20 24 22 15 21 15 17 22

a. Enter this into R. Use the function max to find the longest commute time, average, and minimum commute time

b. Replace the value 24 with 18 and recompute the new average.

c. How many times was your commute 20 minutes or more?

d. What percent of your commutes are less than 17 minutes?

2A) Differentiate between lm and cor. Analyse the relationship between the salary of a group of employees in an organization, the number of years of experience, and the age of the employees by (5)

employing both the parameters and interpret your observations.

2B) Write a R program that performs the following operations. (3)

a. Load the mtcars dataset.

b. Fit a linear regression with miles per gallon as the outcome and horsepower as the predictor. Plot horsepower versus the residuals.

c. Compute the R squared for this model.

d. Estimate the Residual Variance

DataSet:

mpg cyl disp hp drat wt qsec vs am gear carb

Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4

Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4

Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1

Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1

Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2

Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

2C) How does the concept of regression to the mean influence the interpretation of statistical analysis, particularly in the context of regression models, and what practical implications does it have in various fields such as economics, sports, or healthcare? (2)

3A) (5)

```
# Load necessary library

library(ggplot2)

# Set seed for reproducibility

set.seed(123)

# Generate a hypothetical dataset

data < - data.frame(

age = rnorm(100, mean = 35, sd = 5),

health_condition = factor(sample(c("Good", "Poor"), 100, replace = TRUE), levels = c("Good", "Poor"))

)

# Create a binary outcome variable based on health_condition

data$outcome < - ifelse(data$health_condition == "Poor", 1, 0)

# Scatter plot to visualize the data

ggplot(data, aes(x = age, y = outcome, color = health_condition)) +

geom_point() +

labs(title = "Scatter Plot of Age vs. Health Outcome",

x = "Age",

y = "Health Outcome",
```

color = "Health Condition")

For the data set generated using the above code, write a R Code to interpret the relationship between age and the likelihood of having a Poor health condition. Considering the nature of the outcome variable, why might linear regression not be the best fit for modeling this relationship? What advantages does logistic regression offer in this context?

3B) Consider a scenario where events occur at a fixed average rate. In the context of a Poisson distribution: (3)

a. Explain the key characteristics of the Poisson distribution and under what conditions it is applicable.

b. Write a R command to generate a random sample from a Poisson distribution with a given mean (λ).

c. How does the shape of the Poisson distribution change as the mean (λ) varies?

3C) Your friend claims that changing the font to comic sans will result in more ad revenue on your web sites. When presented in random order, 9 pages out of 10 had more revenue when the font was set to comic sans. If it was really a coin flip for these 10 sites, what is the probability of getting 9 or 10 out of 10 with more revenue for the new font? (2)

4A) A pharmaceutical company is conducting a study to assess the effectiveness of a potential blood pressure-lowering medication. The blood pressure measurements (in mmHg) for 10 subjects were recorded at baseline and two weeks later. The data is presented in the Table 1: (5)

Table1:BP Assessment Data

Subject	Baseline (mmHg)	Week 2 (mmHg)
1	130	125
2	142	140
3	155	152
4	150	147
5	134	130
6	138	135
7	145	142
8	132	128
9	148	145
10	136	133

Test the hypothesis that there is a mean reduction in blood pressure after two weeks of medication.

a. Formulate the null and the alternative hypothesis.

b. Perform a paired-sample t-test on the given data to obtain the t-statistic. Determine the degrees of freedom (df) and find the two-sided p-value. Use a significance level of 0.05.

c. Make a conclusion based on the p-value and the chosen significance level.

4B) In the realm of hypothesis testing: (3)

a. Define and distinguish between Type I and Type II errors.

b. How do factors like sample size, significance level, and effect size influence the likelihood of committing Type I and Type II errors?

4C) In a court of law, all things being equal, if via policy you require a lower standard of evidence to convict people then which of the following statements is true and why? (2)

1. Less guilty people will be convicted.
2. More innocent people will be convicted.
3. More Innocent people will be not convicted

5A) Consider the following R commands: (5)

```
nosim < - 1000

n < - 10

result < - sd(apply(matrix(sample(0:1, nosim * n, replace = TRUE), nosim), 1, mean))
```

a. Explain the purpose of these commands, including the role of `nosim` and `n`.

b. Interpret the value of `result` (0.1587) in the context of the simulation.

c. How does the expression `1 / (2 * sqrt(n))` relate to the result, and what does it represent in the context of the simulation?

d. Discuss the significance of the `sample` function with arguments `0:1`, `nosim * n`, and `replace = TRUE`.

5B) In the context of statistical hypothesis testing: (3)

a. Discuss the factors that influence statistical power.

b. Write a R command to calculate the statistical power for a given hypothesis test.

c. Explain the interpretation of the output from the R command and how it informs the decision-making process in hypothesis testing.

5C) In R, how would you calculate and interpret the 95th percentile of a normal distribution with a specified standard deviation (σ) and mean (μ)? (2)

-----End-----