

MATH 9102 - Probability and Statistical Inference

Assignment 2

Due 6pm, Sunday 10th March, 2024

15 marks

Submission guidelines:

- You will need to upload only single R markdown (.Rmd) file.
- File name of your Rmd file must be **idnumber_CA2**
- Do not upload given datasets (if any).
- Make use of R built in datasets (if mentioned in the question). If you have considered external dataset instead of R built in, upload the dataset without zipping it.
- Use the following statement if installing any package:
`if(!require(packageName))install.packages("packageName")`

General Instructions:

- Read the questions carefully and answer all parts to secure full marks.
- Do not ask for direct solutions. This is part of your assessment.
- Assignment will be penalized if you miss any of the submission guidelines.
- Please complete assignment individually and avoid plagiarism as it will lead to penalties and negatively affect your overall grade.
- Including comments/markup in your code to explain what you did and provide answers to all questions.

1. A study was done in which the high daily temperature and the number of traffic accidents within the city were recorded. These sample data are shown as follows.

```
temp <- c(91,56,75,68,50,39,98)
accidents <- c(2,9,7,6,6,10,1)
accTemp <- data.frame(temp, accidents)
```

Identify whether the data is positively correlated or negatively correlated using a scatter plot. **[1 mark]**

2. Following data gives the readings of sugar level of 5 diabetic patient before and after taking insulin. Using 1% level of significance test

whether insulin has reduced sugar level. **[1 mark]**

```
before <- c(350,400,250,200,180)
after <- c(200,300,200,150,120)
```

3. The following dataset includes the score of students in mathematics and statistics. Justify with R code suitable correlation test. [Hint: Check for normality and test the data for correlation] **[2 marks]**

```
maths <- c(50,54,56,59,60,62,61,65,67,71,71,74)
stat <- c(22,25,34,28,26,30,32,30,28,34,36,40)
scoreSet <- data.frame(maths, stat)
```

[Note: Save the read dataset in a variable called *heartdisease*. Do not change the variable name.

For ex: `heartdisease <- read.csv("heartdisease.data")`]

4. Consider given dataset *heartdisease.csv*. Write code to do the following.

- a) Assess the following variables for normality. **[2 marks]**

- Cholesterol (Chol)
- Blood Pressure (RestBP)
- MaxHR

- b) Choose the test you think is correct based on your assessment of these variables - choose either Pearson, Spearman/Kendall to investigate the following: **[3 marks]**

- Relationship between cholesterol and blood pressure
- Relationship between cholesterol and old peak

5. Investigate whether there is a difference in the people who have hepatitis and those who did not, by considering the following variables. **[2 marks]**

- BILIRUBIN
- SGOT

[Note: Save the read dataset in a variable called *hepatitis*. Do not change the variable name.

For ex: `hepatitis <- read.csv("hepatitis.data")`.

Dataset is available <https://archive.ics.uci.edu/ml/datasets/Hepatitis>. Use the code below to add header.]

#To add headers

```
colnames(hepatitis) <- c("Class", "AGE", "SEX",
"STEROID", "ANTIVIRALS", "FATIGUE", "MALAISE",
```

"ANOREXIA", "LIVER BIG", "LIVER FIRM", "SPLEN
PALPABLE", "SPIDERS", "ASCITES",
"VARICES", "BILIRUBIN", "ALK PHOSPHATE",
"SGOT", "ALBMIN", "PROTIME", "HISTOLOGY")

6. Formulate a hypothesis by considering ALK PHOSPHATE levels and hepatitis histology by considering hepatitis dataset. Mention whether you accept or reject the hypothesis. **[2 marks]**
7. Investigate the following questions by considering the Hepatitis dataset. **[2 marks]**
 - Does Bilirubin level impact the Liver Firm?
 - Are there any differences in steroid level and hepatitis histology?