Assignment 1

DAIICT Winter 2021

Data Analysis and Visualization
CS306

Questions	Marks
Answer all 7 questions	95 marks
	95 Total

Instructions

- This assignment contributes towards 20% of your total marks for grading. Please submit the answers to the assignment by 25th of March 2021.
- If you took or gave help to others then please do mention in you assignment submission. Both handwritten and typed answers are acceptable, however your plots have to be computer generated. They should come from some codes you wrote to do this assignment. Do submit your computer codes.
- Please do mention your reference materials.
- We have kept 5 marks for presentation and communication. Take care to properly present your answers. Scanning quality is machine and human readable and scan orders are correct.

Probability

Question 1

Suppose that a coin which was unfair was tossed 3 times in a such a way that the over the long run the following relative frequencies where observed.

outcomes (e)	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
Relative Probability Pr(e)	0.13	0.10	0.10	0.17	0.17	0.10	0.10	0.13

We are interested in the following events

- E : More than one heads
- F : All coins are the same
- G : More than and equal to 2 tails
- H : Some coins different

(a) Compute the following probabilities $P(E), P(F), P(E \cap F), P(E \cup F)$ [4 marks]

(b) State and verify the addition rule of probability for events E and F. [2 marks]

(c) Compute the following probabilities $P(G), P(H), P(G \cup H), P(G \cap F)$ [4 marks]

(d) State and verify the independence rule of probability for events G and H. [2 marks]

[Total for Question 1: 12 marks]

Conditional probability

Question 2

I am thinking of buying a used PQR car. Somehow by internet searches, talks etc. I found out that 35% of used cars are faulty. I hired a mechanic who can correctly classified 94% of the faulty cars and 6% of the times he classified faulty cars to be good. He correctly classified 88% of the good cars and 12% of the times he made the mistake of classifying good cars to be faulty.

What is the chance that a car I am thinking of buying has a fault under the following conditions

(a) Before I get advice from mechanic. [2 marks]

(b) When the mechanic has certified the car to be faulty. [4 marks]

(c) Mechanic has certified the car to be good. [4 marks]

[Total for Question 2: 10 marks]

Bayes Theorm

Question 3

From a data base of a clinic customers it is found out that $1/20$ patients have disease A, $1/10$ have disease B, and $17/20$ have neither. Of those with disease A, $19/20$ have headaches; of those with disease B, $1/4$ have headaches and of those with neither disease, $1/10$ have headaches.

A patient in the clinic complains of having headache.

(a) What is the probability he has disease A? [2 marks]

(b) What is the probability he has disease B? [2 marks]

(c) What is the probability he has neither of the above two diseases? [2 marks]

(d) What is the probability he has both of the above two diseases? [2 marks]

[Total for Question 3: 8 marks]

Statistics : confidence interval

Question 4

A random sample of 10 men and 10 women were drawn from a population of IT graduates who are working in different positions for annual income in lakhs of INR. Following is the table for the sample:

Men	11	13	24	14	20	34	25	22	49	8
Women	9	41	10	21	23	25	37	27	15	7

(a) Calculate 95% and 99% confidence interval, and interpret them on how much men's income is higher than women's income. [6 marks]

(b) Is the men's income really different from women's based on the above data sample? Give reasons for your answer. [2 marks]

[Total for Question 4: 8 marks]

Random Data generation and analysis

Question 5

Generate $N = 100$ samples from a 2 dimensional normal distribution specified by mean $\mu = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and variance matrix $\Sigma = \begin{bmatrix} 4 & 4 \\ 4 & 9 \end{bmatrix}$. You will be writing a computer code for this question.

(a) Generate and display a scatter plot of the generated samples. [2 marks]

(b) Compute the sample mean m and sample covariance matrix. [2 marks]

(c) Repeat the above two steps of each with a new set of randomly generated data 10 times. Compute and report the average of the sample mean and sample covariance matrix of the 10 trials. [3 marks]

(d) Repeat the above steps for $N = 20, 40, 60, 80, 100, 200, 300, 400, 500$. Display your results as a table and comment on your results. [5 marks]

[Total for Question 5: 12 marks]

Monte Carlo simulation

Question 6

Simulate 10 observations of weight from a normal population with $\mu = 60\ kg$ and $\sigma = 12\ kg$ call it sample A. The simulate second and third samples B and C.

(a) From the array of 3×10 observations, evaluate the ANOVA table and test whether these three samples are from the same population. [5 marks]

(b) Record the values of F ratio computed above for 200 runs of the above 3×10 observations each begin a new random generation. Use the programming language you are comfortable with. [5 marks]

(c) Visualize the resulting distribution of F. What proportion of F values exceed $F_{0.05} = 3.35$? [5 marks]

(d) Repeat the above experiment by changing the number of observations (5, 10, 15, 25, 30). [8 marks]

(e) What is the meaning of α in the ANOVA test. Comment on how F critical changes when we change the α value. [2 marks]

[Total for Question 6: 25 marks]

Data Analysis : periodicity in data

Question 7

For this question use the data given in file “data-ass1.csv”.

(a) Visualize the data in a graph plot. Comment on the periodicity of the data. [2 marks]

(b) Describe an algorithm discussed to detect the periodicity in this long sequence data . [4 marks]

(c) Write a pseudo code to implement the algorithm you have described. [4 marks]

(d) Implement this algorithm in a language of you choice and state the periodicity of the data. Validate your answer with a plot. This plot is different from the one you did in (a). Repeat the above steps (a) and (d) for the data “data2-ass1.csv”. [10 marks]

[Total for Question 7: 20 marks]

End of assignment