SECI2143: PROBABILITY & STATISTICAL DATA ANALYSIS

2020/2021 – SEMESTER 2

ASSIGNMENT 1

INSTRUCTIONS:

1. This assignment must be conducted in a group (3 or 4 students). Please clearly write the group members name & matric number in the front-page of the submission.
2. Solutions for each question must be readable and neatly written on plain A4 paper. Every step or calculation should be properly shown. Failure to do so will result in rejection of the submission of assignment.
3. For submission, scan and combine all answer/solution sheets as one PDF file. Then only ONE group member needs to submit on behalf of the group via e-learning (Due date: **17th April 2021, 11.59pm**)
4. This assignment has 6 questions (100 marks), which contribute 5% of overall course marks.

QUESTION 1 [6 marks]

Read the following case study.

1. A Malayan sports journalist intends to write a book about football clubs in Malaysiana. He will analyse all the Malaysiana Super League matches in the season. He records for each match whether it is a home win, an away win or a draw. He also records for each match the total number of goals scored and the amount of time played before a goal is scored. Old newspapers articles showed that in the previous season, the mean number of goals per game was 3.08. On the first Saturday of the season, he recorded the number of goals scored in each match and calculated the mean number of goals per match as 2.97.

2. All students at a school in JayBie must undergo a medical examination during their first year at the school. The data recorded for each pupil include place of birth, gender, age (in years), height and weight. A summary of the data collected is available on request. A class of statistics students decides to collect data on the weight of second year pupils and compare them with the data on first year pupils. It is agreed that the data will be collected one lunchtime. Each member of the class will be provided with a set of bathroom scales and will weigh as many second-year pupils as possible. At the end of the lunchtime, they will each report the number of pupils weighed and the mean of the weights recorded.

For each for the case study, identify an example of:

(a) the population,
(b) the sample,
(c) a discrete variable,
(d) a continuous variable,
(e) primary data,
(f) secondary data.

(*one example for each element – ½ mark for each answer*)

QUESTION 2 [34 marks]

A small, explorative survey was done among 20 respondents in the AERON car park to determine what factors were important to buyers when buying a car. The four most important factors considered by the buyers were price, condition of the car, fuel efficiency, and car depreciation.
Buyers were given a questionnaire that had a 4-point interval scale on which they could rate their preference. One on the scale meant very important and four very unimportant. In between points on the scale were intended to allow for degrees of preference between the polar extremes. Table 1 shows the results as follows:

Table 1: Surveys on Car Buying Factors

Respondents	Factors			
	Price	**Condition of the Car**	**Fuel Efficiency**	**Car Depreciation**
A	1	2	2	2
B	2	2	3	1
C	1	3	3	2
D	2	1	4	2
E	1	2	3	2
F	1	3	3	1
G	2	3	2	3
H	1	3	2	1
I	1	1	2	1
J	2	1	2	2
K	1	2	3	3
L	2	3	3	2
M	1	3	1	2
N	1	3	2	2
O	1	3	2	4
P	1	3	2	2
R	2	2	2	1
S	1	3	2	2
T	1	2	3	2
U	1	3	2	1

1. Summarize Table 1 in a frequency table as below. *(10 marks)*

Scales / Factors	Frequency				Total
	1	2	3	4	
Price					
Condition of the Car					
Fuel Efficiency					
Car Depreciation					

2. Summarize all the factors (price, condition of the car, fuel efficiency, and car depreciation) from Table 1 in a frequency distribution. For each of the factor, use the table as below:
(1 table for a factor = 5 marks, 4 tables = 20 marks)

Scale	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
1				
2				
3				
4				
Total				

3. Draw the results from (b) using a comparative bar chart. *(4 marks)*

QUESTION 3 [15 marks]

The following data shows the number of days taken by 18 contestants to finish developing a mobile app for a competition (the duration given was 90 days).

30 43 32 21 65 8 4 18 16 38 9 44 33 23 24 81 42 55

1. Represent the data in a Stem-and-Leaf plot. *(4 marks)*

2. Based on the dataset, find the:
   a. Mean
   b. Mode
   c. Median
   (3 marks)

3. Represent the data in a modified box plot. Show the calculation to get:
   a. 1^{st}, 2^{nd} (median) and 3^{rd} quartiles.
   b. Interquartile range (iqr).
   c. mild and/or extreme outliers in the dataset.
   (8 marks)

QUESTION 4 [5 marks]

The value of 11 houses in a new development area in Secudai are shown in Table 2:

Table 2

Value per House	Number of Houses
RM 175,000	1
RM 250,000	5
RM 500,000	4
RM 700,000	1

Based on the table:

1. Find the mean value of these houses in RM. *(2 marks)*
2. Find the median value of these houses in RM. *(2 marks)*
3. State which measures the central tendency, the mean or the median, “best” represents the values of the 11 houses. Justify your answer. *(1 mark)*

QUESTION 5 [20 marks]

There was little known on how the interaction effect between obesity and current smoking affected the incidence of hypertension. The aim of this study was to investigate how body mass index (BMI) modified the effect of current smoking on the incidence of hypertension. According to the World Health report, smoking is the leading cause of 1.69 million cardiovascular diseases-related deaths. Also, there were approximately 7 million hypertension-related deaths each year. The sample data were collected from 20 patients of the General Hospital. BMI was calculated as weight in kilograms divided by the square of height in meters. Table 3 shows the results as follows:

Table 3: BMI of 20 Patients in General Hospital

Subject ID	Age	Current Smoker*	Body Mass Index (BMI)	Hypertension*
1	63	0	29.6	1
2	74	1	26.4	0
3	75	1	24.5	0
4	74	0	31.9	1
5	70	0	22.8	0
6	72	0	19.8	0
7	81	0	27.6	1
8	68	1	26.8	1
9	67	0	24.7	1
10	77	0	23.0	0
11	65	1	23.5	1
12	73	1	26.2	0
13	83	1	24.8	0
14	66	0	23.9	1
15	64	0	25.6	0
16	76	1	20.8	0
17	62	0	21.6	1
18	84	1	18.9	1
19	66	0	22.5	1
20	78	0	28.1	0

**Note: 0 indicates No; 1 indicates Yes.*

1. Identify the level of measurements (nominal, ordinal, interval, ratio) used in Table 3.
   (2 marks)

Age	
Current Smoker	
Body Mass Index (BMI)	
Hypertension	

2. Consider the column Body Mass Index (BMI) from Table 3 as a univariate dataset, summarize this data into a frequency table according to the conditions below:
   (2 marks)

   If BMI is less than 18.5, it falls within the underweight range.
   If BMI is 18.5 to <25, it falls within the normal.
   If BMI is 25.0 to <30, it falls within the overweight range.
   If BMI is 30.0 or higher, it falls within the obesity range.

CLASS INTERVAL	CATEGORY	FREQUENCY
0<BMI<18.5	Underweight	
18.5<=BMI<25.0	Normal	
25.0<=BMI<30.0	Overweight	
30.0<=BMI	Obesity	

3. Find the class boundaries, class midpoint, and frequency distribution for the **AGE** of Table 3.
   (12 marks – ½ mark for each answer in cell)

CLASS INTERVAL	CLASS BOUNDARIES	CLASS MIDPOINT	FREQUENCY	CUMULATIVE FREQUENCY
60-64				
65-69				
70-74				
75-79				
80-84				
Total				

4. Draw a histogram for AGE based on results in (3).
   (4 marks)

QUESTION 6 [20 marks]

A flight delay is when an airline flight takes off and/or lands later than its scheduled time. The Federal Aviation Administration (FAA) considers a flight to be delayed when it is 15 minutes later than its scheduled time. The following distribution shows a sample of 100 flights with delay times (rounded to the nearest minute) recorded in ABC Airport.

Table 4: Flight Delay Time

TIME (MINUTES)	FREQUENCIES
16-30	3
31-45	13
46-60	30
61-75	25
76-90	14
91-105	8
106-120	4
121-135	2
136-150	1

1. Find the MEDIAN, MEAN and MODE for the values above. Show the calculation to get each answer. *(12 marks)*

2. Based on your answer in (1):

   a. Draw the distribution graph. *(4 marks)*
   b. What can you tell on the shape of the distribution (skewness/kurtosis)? *(2 marks)*
   c. What does this tell you about the data? *(2 marks)*