National University of Computer and Emerging Sciences

FAST School of Computing **Spring-2022** **Islamabad Campus**

MT-2005: Probability and Statistics

Serial No:

Sessional Exam-I

Total Time: 1 Hour

Total Marks: 50

Saturday, 12th March, 2022

_____

Signature of Invigilator

Course Instructors

Dr. Shahzad Saleem

Engr. Sana Saleh

Mr. Waqas Munir

_____ _____

Student Name Roll No. Section Signature

DO NOT OPEN THE QUESTION BOOK OR START UNTIL INSTRUCTED.

Instructions:

1. Attempt on question paper. Attempt all of them. Read the question carefully, understand the question, and then attempt it.
2. No additional sheet will be provided for rough work. Use the back of the last page for rough work.
3. If you need more space write on the back side of the paper and clearly mark question and part number etc.
4. After asked to commence the exam, please verify that you have **nine (9)** different printed pages including this title page. There is a total of **3** questions.
5. Calculator sharing is strictly prohibited.
6. Use permanent ink pens only. Any part done using soft pencil will not be marked and cannot be claimed for rechecking.

	Q-1	Q-2	Q-3	Total
Marks Obtained				
Total Marks	**21**	**15**	**14**	**50**

Question 1 [21 Marks]

a. Mention the sampling method used in the following studies i.e. random, systematic, stratified, cluster. Give reason to support your answer. **[2*5=10]**

1. A farmer separates his apple tree farm into 10 regions. He counts the number of apples produced in just one of the regions and uses that estimate to predict the number of apples produced on the whole farm.

Sampling Method: Cluster Sampling
Reason:

2. To find out what the preferred ice cream flavor is, wait outside an ice cream parlor and ask every 4th person leaving the store to name his or her favorite flavor until you get 25 responses.

Sampling Method: Systematic Sampling
Reason:

3. To study the amount of time students spend doing homework each day, use a random number generator to select 25 students from the student enrollment database to survey.

Sampling Method: Random Sampling
Reason:

4. The school librarian wants to determine how many students use the library on a regular basis. What type of sampling method would she use if she chose to select every 3rd student who enters the library on Tuesday.

Sampling Method: Systematic Sampling
Reason:

5. A large company surveys 100 employees by taking random samples of 10 managers and 90 non-managerial employees.

Sampling Method: Stratified Sampling
Reason:

b. Classify each variable on the level of measurement. Also, mention that the variable type i.e., qualitative, or quantitative. [5]

Variable	Level of Measurement	Variable Type
Hair color (blonde, gray, brown, black, etc.)	**Nominal**	**Qualitative**
Time interval (measured with a stopwatch)	**Ratio**	**Quantitative**
Age (from 0 years to 100+)	**Ratio**	**Quantitative**
Rate service of a restaurant (excellent, average, worst)	**Ordinal**	**Qualitative**
happiness level (measured on a scale of 1-10).	**Ordinal**	**Qualitative**

c. Answer the questions regarding the study mentioned below:

"An experiment was done by medical researchers to determine the association between drinking caffeine and severity of lung cancer. For this purpose, they randomly select 2,00,000 individuals from the USA, who drank on average 3-4 cups of coffee daily. The medical team instructed them to have at least 80 mg of caffeine daily and observed them for 5 years. Results showed that caffeine intake has a strong positive correlation with the severity of lung cancer. Medical team concluded their finding as "Caffeine can adversely affect human health". **[3]**

Questions	Answers
1. Variables mentioned in the study	**Caffeine intake & severity of lungs cancers**
2. Independent Variables	**Caffeine intake**
3. Dependent Variables	**severity of lungs cancers**
4. Potential confusing variable	**Smoking**
5. Type of study (observational/experimental)	**Experimental**
6. The above study falls under descriptive statistics or inferential statistics?	**Inferential Statistics**

d. Answer the following questions and with reason to support your answer: [3]

1. By looking at a single class given as: **"10.5-12.5",** Can you say that it represents class boundaries or class limits? Given reason to support your answer. [1.5]

Answer: We can't say this class represents class limits or boundaries. To comment on class limits or boundaries, we need at least two consecutive classes. If the upper boundary of previous class is same as lower class boundary of the next class and both ends with a 5, then it represent class boundary. As mentioned in the book:

"The basic rule of thumb is that the class limits should have the same decimal place value as the data, but the class boundaries should have one additional place value and end in a 5"

2. State the following data is symmetrical, left skewed or right skewed? [1.5]

A. 1,1,1,2,2,2,2,3,3,3,3,3,3,4,4,4,4,5,5,5 => _____Symmetric_____

B: 6,7,7,7,7,8,8,8,9,10 => _____Right Skewed_____

C: 4,5,6,6,6,7,7,7,7,8 => _____Left Skewed_____

Question 2 [15 Marks]

a) Following data give the number of fuel pumps for SUVs produced per day by manufacturing companies during the month of June in a given year:

72 88 65 68 68 75 87 79 89 79 65 76 81 84 67
82 61 89 85 90 67 68 82 85 79 65 79 74 81 82

Construct a frequency distribution and plot histogram for the above data. Also identify the shape of the distribution. **[6 + 6 +1]**

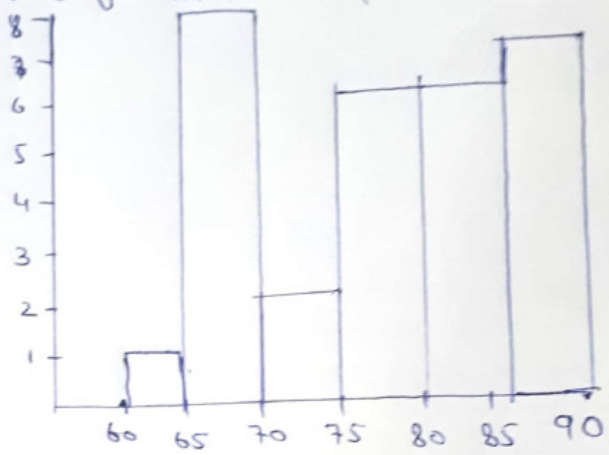
b) State the empirically relationship between mean, median and mode for asymmetrical distributions.

Solution of 2(a)

Frequency distribution table for the survival time of parts.

Class	Frequency
60-65	1
65-70	8
70-75	2
75-80	6
80-85	6
85-90	7
	Σ: 30

Histogram



Here, data follow a negatively skewed distribution. or skewed to the left

b) Epricall relation b/w Mean, Median, and Mode.

For unimodal frequency curves which are moderately skewed (asymmetrical), we have the following empirical relation:

Mean - Mode = 3(Mean - Median).

Question 3 [14 Marks]

The number of annual precipitation days for 25 different cities is listed below.

135 128 136 78 116 77 111 79 44 97 116 123
88 102 26 82 156 133 107 35 112 98 45 122 125

Find

i)	Median	ii)	Mode	iii)	Variance
iv)	Percentile of the value 107	v)	Q_1 and Q_3	vi)	Outliers
vii)	Value at 80th percentile				

[2 × 7 = 14]

Following formulae can be helpful.

$$s^2 = \frac{\sum(X-\bar{X})^2}{s}, \quad s^2 = \frac{n\sum X^2 - (\sum X)^2}{n(n-1)}, \quad c = \frac{n \cdot p}{100}, \quad p = \frac{(\text{no. of values} < X) + 0.5}{n} \cdot 100$$

$$X < Q_1 - 1.5\,(\text{IQR}) \qquad X > Q_3 + 1.5(\text{IQR})$$

Solution

i) Sorting data in ascending order

26 35 44 45 77 78 79 82 88 97 98 102 107 111 112 116 116 122 123 125 128 133 135 136 156

Median : 107

ii) Mode = 116

iii) Variance: $s^2 = \frac{n\sum X^2 - (\sum X)^2}{n(n-1)} = \frac{1}{25(25-1)}[25 \times (26^2 + 35^2 + 44^2 + 45^2 + 77^2 + 78^2 + 79^2 + 82^2 + 88^2 + 97^2 + 98^2 + 102^2 + 107^2 + 111^2 + 112^2 + 116^2 + 116^2 + 122^2 + 123^2 + 125^2 + 128^2 + 133^2 + 135^2 + 136^2 + 156^2) - (26 + 35 + 44 + 45 + 77 + 78 + 79 + 82 + 88 + 97 + 98 + 102 + 107 + 111 + 112 + 116 + 116 + 122 + 123 + 125 + 128 + 133 + 135 + 136 + 156)^2]$

$$s^2 = 1156.7$$

iv) Percentile of the value 107

$$p = \frac{(\text{no. of values} < X) + 0.5}{n} = \frac{12 + 0.5}{25} = 50\%$$

v) Q_1 and Q_3

26 35 44 45 77 78 79 82 88 97 98 102

$$Q_1 = \frac{78 + 79}{2} = 78.5$$

111 112 116 116 122 123 125 128 133 135 136 156

$$Q_3 = \frac{123 + 125}{2} = 124$$

vi) Outliers

$$IQR = Q_3 - Q_1 = 124 - 78.5 = 45.5$$

$$X < Q_3 - 1.5(\text{IQR}) = 124 - 1.5(45.5) = 55.8$$

Any values < 55.8 is an outlier. Therefore, in this case we have $26,35,44,45$ as outliers.

$$X > Q_3 + 1.5(\text{IQR}) = 124 + 1.5(45.5) = 192.3$$

Any values > 192.3 is an outlier. In our case, there is no such values.

Hence, Outliers: $26,35,44,45$.

vii) Value at 80th percentile

$$c = \frac{n \cdot p}{100} = \frac{25 \cdot 80}{100} = 20$$