BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
Work Integrated Learning Programmes Division
Second Semester 2024-2025
Mid-Semester Test
(EC-2 Make Up)

Course No.	:	AIMLCZC418
Course Title	:	Introduction to Statistical Methods
Nature of Exam	:	Closed Book
Weightage	:	30%
Duration	:	2 Hours
Date of Exam	:	28.06.2025_FN

No. of Pages = 4
No. of Questions = 6

Note to Students:
1. *Please follow all the Instructions to Candidates given on the cover page of the answer book.*
2. *All parts of a question should be answered consecutively. Each answer should start from a fresh page.*
3. *Assumptions made if any, should be stated clearly at the beginning of your answer.*

Answer all the questions:

Q1) Consider marks of the student out of 30:
28, 12, 10, 22, 18, 18, 17, 15, 16, 15, 14, 15, 16, 17, 18
Determine the mean and five-number summary. Based on the results obtained, comment on the symmetry of the data. [3Marks]

Solution:

Sorted Data: 10, 12, 14, 15, 15, 15, 16, 16, 17, 17, 18, 18, 18, 22, 28 [0.5 Mark]
Mean= 16.73 [0.5Mark]

Five-number Summary: Min = 10, Q1 = 15, Median = 16, Q3 = 18, Max = 28. [1Mark]

As Mean > Median the data is slightly positively skewed. [1Mark]

b) The average height of adult males in a certain city is 70 inches, with a standard deviation of 3 inches. If a random sample of 49 adult males is selected, find the probability that the mean height of the sample is between 69 and 71 inches. **[2 Marks]**

Step 1: Calculate the Standard Error of the Mean

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{49}} = \frac{3}{7} \approx 0.43 \text{ inches}$$

Step 2: Standardize the Sample Mean Range

- For $\bar{x} = 69$:

$$z_1 = \frac{69 - 70}{0.43} = \frac{-1}{0.43} \approx -2.33$$

- For $\bar{x} = 71$:

$$z_2 = \frac{71 - 70}{0.43} = \frac{1}{0.43} \approx 2.33$$

Step 3: Find the Probability Using the Standard Normal Distribution

- From standard normal tables or a calculator:
  - $P(Z < 2.33) \approx 0.9901$
  - $P(Z < -2.33) \approx 0.0099$
- The probability between these z -scores is:

$$P(-2.33 < Z < 2.33) = P(Z < 2.33) - P(Z < -2.33) = 0.9901 - 0.0099 = 0.9802$$

Q2 a) You are given a very small dataset of text messages, and your task is to determine whether a new message is spam or not spam (ham) using Naive Bayes.

Message	Label
Win money	Spam
Free prize	Spam
See you soon	Ham
Call me now	Ham

Test the new message "Win free money" by using the Naive Bayes to classify as Spam or Ham. **[5 Marks]**

Solution:

Spam Class (2 messages):

- Words: `{"win", "money", "free", "prize"}`
- `count("win" in Spam)` = 1
- `count("money" in Spam)` = 1
- `count("free" in Spam)` = 1
- `count("prize" in Spam)` = 1
- `Total words in Spam (N_spam)` = 4

Ham Class (2 messages):

- Words: `{"see", "you", "soon", "call", "me", "now"}`
- `count("see" in Ham)` = 1
- `count("you" in Ham)` = 1
- `count("soon" in Ham)` = 1
- `count("call" in Ham)` = 1
- `count("me" in Ham)` = 1
- `count("now" in Ham)` = 1
- `Total words in Ham (N_ham)` = 6

Spam Class (2 messages):

- Words: `{"win", "money", "free", "prize"}`
- `count("win" in Spam)` = 1
- `count("money" in Spam)` = 1
- `count("free" in Spam)` = 1
- `count("prize" in Spam)` = 1
- `Total words in Spam (N_spam)` = 4

Step 2: Calculate Prior Probabilities

- Total number of messages = 4
- Number of Spam messages = 2
- Number of Ham messages = 2
- $P(\text{Spam}) = \frac{2}{4} = 0.5$
- $P(\text{Ham}) = \frac{2}{4} = 0.5$

For Ham Class ($N_{ham} = 6$, $|V| = 10$):

- $P(\text{"win"}|\text{Ham}) = \frac{0+1}{6+10} = \frac{1}{16} = 0.0625$
- $P(\text{"free"}|\text{Ham}) = \frac{0+1}{6+10} = \frac{1}{16} = 0.0625$
- $P(\text{"money"}|\text{Ham}) = \frac{0+1}{6+10} = \frac{1}{16} = 0.0625$
- For any word in Ham training messages (e.g., "see", "you"):
  $P(\text{"see"}|\text{Ham}) = \frac{1+1}{6+10} = \frac{2}{16} = 0.125$

Calculate Probability for Ham:

$$
\begin{aligned}
&P(\text{"Win free money"}|\text{Ham}) \times P(\text{Ham}) \\
&= P(\text{"win"}|\text{Ham}) \times P(\text{"free"}|\text{Ham}) \times P(\text{"money"}|\text{Ham}) \times P(\text{Ham}) \\
&= \left(\tfrac{1}{16}\right) \times \left(\tfrac{1}{16}\right) \times \left(\tfrac{1}{16}\right) \times 0.5 \\
&= \left(\tfrac{1}{4096}\right) \times 0.5 \\
&= 0.00024414 \times 0.5 \\
&\approx 0.000122
\end{aligned}
$$

Step 3: Calculate Likelihoods with Laplace Smoothing

Formula: $P(\text{Word}|\text{Class}) = \frac{\text{Count(Word in Class)}+1}{\text{Total words in Class}+|V|}$

For Spam Class ($N_{spam} = 4$, $|V| = 10$):

- $P(\text{"win"}|\text{Spam}) = \frac{1+1}{4+10} = \frac{2}{14} \approx 0.142857$
- $P(\text{"free"}|\text{Spam}) = \frac{1+1}{4+10} = \frac{2}{14} \approx 0.142857$
- $P(\text{"money"}|\text{Spam}) = \frac{1+1}{4+10} = \frac{2}{14} \approx 0.142857$
- For any word not in the Spam training messages (e.g., "you", "call"):
  $P(\text{Word not in Spam}|\text{Spam}) = \frac{0+1}{4+10} = \frac{1}{14} \approx 0.071428$

Calculate Probability for Spam:

$$
\begin{aligned}
&P(\text{"Win free money"}|\text{Spam}) \times P(\text{Spam}) \\
&= P(\text{"win"}|\text{Spam}) \times P(\text{"free"}|\text{Spam}) \times P(\text{"money"}|\text{Spam}) \times P(\text{Spam}) \\
&= \left(\tfrac{2}{14}\right) \times \left(\tfrac{2}{14}\right) \times \left(\tfrac{2}{14}\right) \times 0.5 \\
&= \left(\tfrac{8}{2744}\right) \times 0.5 \\
&= 0.00291545 \times 0.5 \\
&\approx 0.001458
\end{aligned}
$$

Since $P(\text{Spam}|\text{"Win free money"}) > P(\text{Ham}|\text{"Win free money"})$, the new message "Win free money" is classified as Spam.

Q 3. a) A random sample of 49 workers was taken at the company canteen. The average amount of time the workers in the sample stayed in the canteen was 45 minutes with a standard deviation of 14 minutes. With a 0.99 probability, how large of a sample would have to be taken to provide a margin of error of 2.5 minutes or less?

Solution: **[2 Marks]**

We use the formula for margin of error:

$$ME = z \cdot \frac{\sigma}{\sqrt{n}}$$

We don't know the population standard deviation σ , but since we're planning sample size and assuming similar variability, we use the sample standard deviation $s = 14$ as an estimate.

Solve for n :

$$ME = z \cdot \frac{s}{\sqrt{n}} \Rightarrow \sqrt{n} = \frac{z \cdot s}{ME} \Rightarrow n = \left(\frac{z \cdot s}{ME}\right)^2$$

Plug in values:

- $z = 2.576$
- $s = 14$
- $ME = 2.5$

$$n = \left(\frac{2.576 \cdot 14}{2.5}\right)^2 = \left(\frac{36.064}{2.5}\right)^2 = (14.4256)^2 \approx 208.1$$

Round up (always round up when determining sample size):

b) Let X be a discrete random variable with the following probability mass function (PMF):

$$P(X=x) = \begin{cases} kx^2 & if\ x = -1,0,1 \\ 0 & otherwise \end{cases}$$

Find the value of the constant k such that this is a valid PMF. [1Mark]
Calculate the expected value (mean) E[X] . [1Mark]
Compute the variance Var(X). [1Mark]

Step 1: Find the value of K:

A valid PMF must satisfy:

$$\sum P(X = x) = 1$$

Use the PMF values for $x = -1, 0, 1$:

$$P(X = -1) = k(-1)^2 = k \quad P(X = 0) = k(0)^2 = 0 \quad P(X = 1) = k(1)^2 = k$$

$$\Rightarrow \sum P(X = x) = k + 0 + k = 2k = 1 \Rightarrow k = \frac{1}{2}$$

Step 2: Compute E[X]

$$E[X] = \sum x \cdot P(X = x) = (-1)(k) + (0)(0) + (1)(k) = -k + 0 + k = 0$$

Step 3: Find the Variance

We use the formula:

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

We already have $E[X] = 0$, so:

$$\text{Var}(X) = E[X^2] = \sum x^2 \cdot P(X = x) = (-1)^2 \cdot k + (0)^2 \cdot 0 + (1)^2 \cdot k = k + 0 + k = 2k$$

Substitute $k = \frac{1}{2}$:

$$\text{Var}(X) = 2 \cdot \frac{1}{2} = 1$$

4Q.A suspect's DNA matches a sample found at a crime scene. The chance of a random person matching the DNA is 1 in 1,000,000.
There are 5,000,000 people in the region.
Assuming the crime was committed by someone in the region, what's the probability the suspect is guilty based on the DNA match alone?

Solution:
Let:
- G: suspect is guilty
- M: DNA matches

Prior:
- P(G) = 1 / 5,000,000
- P(M|G) = 1 (perfect match)
- P(M|¬G) = 1 / 1,000,000 **[1 Mark]**

Using Bayes' Theorem:

$$P(\text{Guilty} \mid \text{Match}) = \frac{P(\text{Match} \mid \text{Guilty}) \cdot P(\text{Guilty})}{P(\text{Match})}$$

By law of total probability: **[2 Marks]**

$$P(\text{Match}) = P(\text{Match} \mid \text{Guilty}) \cdot P(\text{Guilty}) + P(\text{Match} \mid \text{Innocent}) \cdot P(\text{Innocent})$$

Therefore, by Bayes' Theorem:

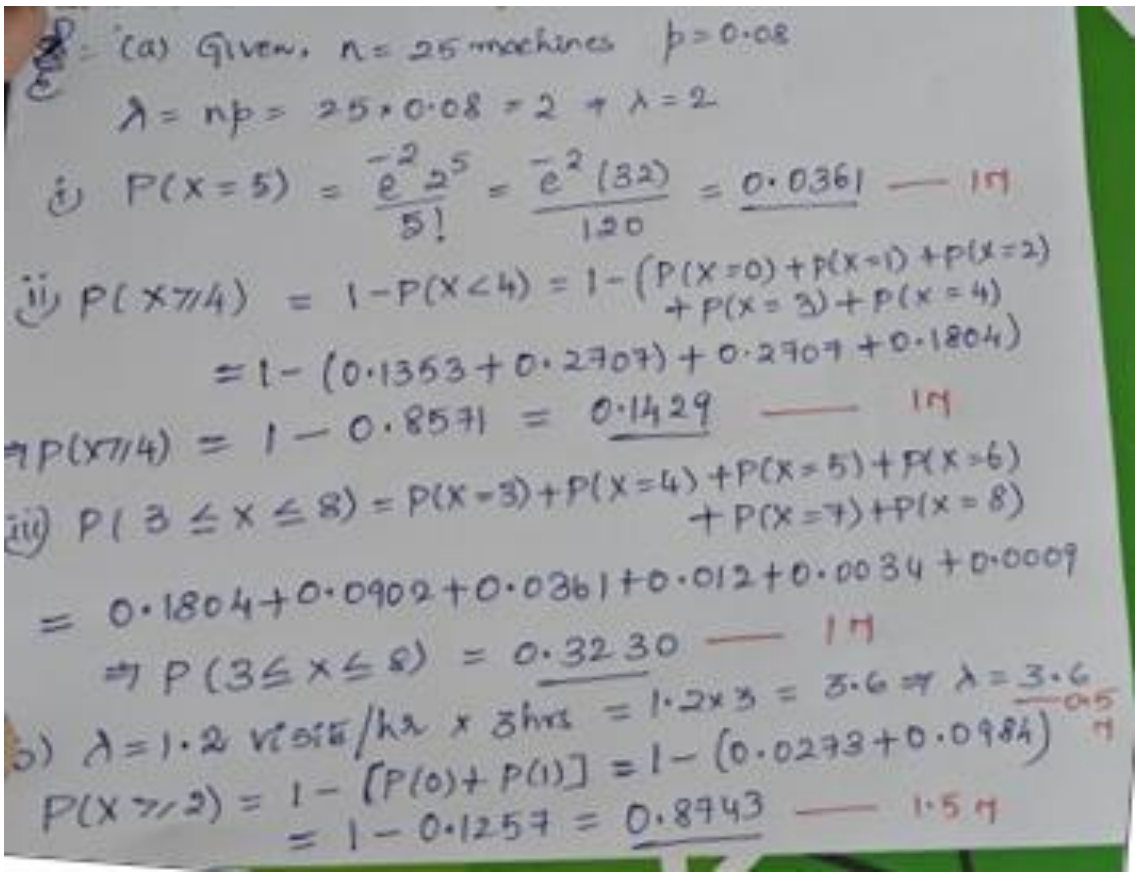
P(G|M) = (1 * 1/5,000,000) / (1 * 1/5,000,000 + 1/1,000,000 * 4,999,999/5,000,000)
≈ 1 / (1 + 5) = 1/6 ≈ 16.7% **[2 Marks]**

Answer: Only a 16.7% chance of guilt from the DNA match alone.

Q.5 In a factory, 8% of all machines break down at least once a year. Use the Poisson approximation to the binomial distribution to determine the probabilities that among 25 machines (randomly chosen in the factory)

i) 5 will break down at least once a year; **[1 Mark]**
ii) at least 4 will break down once a year: **[1 Mark]**
iii) anywhere from 3 to 8, inclusive, will break down at least once a year. **[1 Mark]**

b) A website receives on average 1.2 visits per hour. Find the probability if, in a 3-hour interval, there will be at least 2 visits. **[2 Marks]**



6) In a machine learning application, the random variable X represents the **rating** of an item in a recommendation system. The rating can range from 0 to 2, and follows a piecewise probability density function (PDF) defined as:

$$f(x) = \begin{cases} c\,x^2 & 0 \le x \le 1 \\ c\,(2-x) & 1 < x \le 2 \\ 0 & otherwise \end{cases}$$

Find a) constant c b) $P(\frac{1}{2} < X < \frac{3}{2})$ c) $E(x)$

Solution:

a) Since $f(x)$ is density function, this gives $\int_{-\infty}^{\infty} f(x)dx = 1$.

$$\therefore \int_0^1 cx^2\,dx + \int_1^2 c(2-x)dx = 1$$

$$\left[\frac{cx^3}{3}\right]_0^1 + \left[c\left(2x - \frac{x^2}{2}\right)\right]_1^2 = 1$$

$$\frac{c}{3} + c\left(2 - \frac{3}{2}\right) = 1$$

$$\frac{c}{3} + \frac{c}{2} = 1$$

$$\frac{5c}{6} = 1$$

$$c = \frac{6}{5}$$

_____**[1Mark]**

b) $P\left(\frac{1}{2} < X < \frac{3}{2}\right) = \int_{\frac{1}{2}}^{\frac{3}{2}} f(x)dx = \int_{\frac{1}{2}}^{1} cx^2dx + \int_1^{3/2} c(2-x)dx$

$$= \left[c\frac{x^3}{3}\right]_{\frac{1}{2}}^{1} + \left[c\left(2x - \frac{x^2}{2}\right)\right]_1^{\frac{3}{2}} = c\left(\frac{1}{3} - \frac{1}{24}\right) + c\left(\left(3 - \frac{9}{8}\right) - \left(2 - \frac{1}{2}\right)\right)$$

$$= c\left(\frac{7}{24} + \left(\frac{15}{8} - \frac{3}{2}\right)\right) = c\left(\frac{7}{24} + \frac{3}{8}\right) = c\left(\frac{16}{24}\right) = \frac{6}{5} \times \frac{16}{24} = \frac{4}{5}$$

_____**[3Marks]**

c) $E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^1 cx^3dx + \int_1^2 c(2x - x^2)dx$

$$E(X) = c\left(\frac{x^4}{4}\right)_0^1 + c\left(x^2 - \frac{x^3}{3}\right)_1^2 = c\left(\frac{1}{4} + \frac{2}{3}\right) = c\frac{11}{12} = \frac{6}{5} \times \frac{11}{12} = \frac{11}{10}$$

E(X)= 1.1 _____**[1Mark]**

---------- End ----------