BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
Work Integrated Learning Programmes Division
First Semester 2025-2026
Mid-Semester Test
(EC-2 Regular)

Course No.	:	DSECLZC418
Course Title	:	Introduction to Statistical Methods
Nature of Exam	:	Closed Book
Weightage	:	30%
Duration	:	2 Hours
Date of Exam	:	18.01.2025 FN

No. of Pages = 2
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. a) A sample of eleven weights (in kg) is given: 60, 72, 65, 68, 70,100, 62, 75, 78, 80, 83.

i) Calculate Q1, Q2, Q3, and IQR.
ii) Determine potential outliers using the 1.5 × IQR rule. [3Marks]

SOL:

Sorted Data: 60, 62, 65, 68, 70, 72, 75, 78, 80, 83, 100. [0.5M]
Q1 = 65, Q2= 72, Q3 = 80 [0.5M]
IQR= Q3-Q1=15 [0.5M]
Lower bound = Q1-1.5*IQR=65-22.5=42.5 [0.5M]
Upper bound = Q3+1.5*IQR= 80 + 22.5 =102.5 [0.5M]
Interpretation: Here No outliers present in this data [0.5M]

b) The following table shows the time (in minutes) it took for a group of astronauts to complete various training simulations: 18, 25, 19, 32, 27, 22, 29, 30, 24, 26. Find the mean and standard deviation of these times. Interpret what these statistical measures indicate about the variability and central tendency of the simulation completion times. **[2Marks]**

Solution:

Find the Mean
Sum=18+25+19+32+27+22+29+30+24+26=252
n=10
Mean=252/10 =25.2

Find the Squared Differences

Time (x)	$x-\bar{x}$	$(x-\bar{x})2$

Time (x)	$x-\bar{x}$	$(x-\bar{x})2$
18	-7.2	51.84
25	-0.2	0.04
19	-6.2	38.44
32	6.8	46.24
27	1.8	3.24
22	-3.2	10.24
29	3.8	14.44
30	4.8	23.04
24	-1.2	1.44
26	0.8	0.64

Total sum of squared deviations:
$\sum(x-\bar{x})2=189.6$

Step 3: Calculate Sample Variance and Standard Deviation
Variance=10−1189.6=9189.6≈21.07
Standard Deviation=21.07≈4.59

Final Answer:
Mean : 25.2 minutes
Standard Deviation : ≈ 4.59 minutes

Interpretation:
Central Tendency (Mean):
The average time it took the astronauts to complete the simulation tasks was 25.2 minutes . This gives us a central value around which most of the times cluster.
Variability (Standard Deviation):
The standard deviation of approximately 4.59 minutes tells us that, on average, individual completion times deviated from the mean by about 4.59 minutes. A smaller standard deviation suggests that the data points are relatively close to the mean, indicating moderate consistency in performance across the group.

Conclusion:

The typical astronaut took about 25 minutes to finish a simulation.
Most astronauts completed the simulations within roughly ±5 minutes of the mean.
This suggests that the group performed with reasonable consistency, though there were some faster and slower performers (e.g., times ranged from 18 to 32 minutes).

Q2 a) The length of life of an instrument produced by a machine has a normal distribution with a mean of 12 months and a standard deviation of 2 months. Find the probability that an instrument produced by this machine will last **[2 Marks]**

i) less than 7 months.
ii) between 7 and 12 months.

ANSWER:
Given $\mu = 12$, $\sigma = 2$

$$z = \frac{x-\mu}{\sigma}$$

(i) $P(x < 7) = P(z < -2.5) = 0.0062$

(ii) $P(7 < x < 12) = P(-2.5 < z < 0) = 0.4938$

2 b) A biased coin lands heads with probability p=0.15 . The coin is flipped n=40 times. **[3M]**

(i) Apply the binomial distribution to calculate the probability of getting exactly k=5 heads.
(ii) Recall the Poisson approximation to estimate the same probability.
(iii) Compare the two results and analyze how closely the Poisson result matches the binomial result.

Answer :

Binomial Result:

$P(X=5)\approx 0.1692$
Poisson Approximation Result:

$P(X=5)\approx 0.1606$
Absolute Difference=$|0.1606-0.1692|=0.0086$
Analysis:

The Poisson approximation overestimates the probability compared to the exact binomial result.
This discrepancy arises because the Poisson approximation works best when:
n is very large,
p is very small,
and $\lambda=np$ is moderate (say between 0 and 10).
In our case:
n=40 is moderately large,
p=0.15 is not extremely small,
so the Poisson approximation isn't very accurate here.

Conclusion:

While the Poisson approximation is useful for rare events with large n , it doesn't closely match the binomial result here due to the relatively high value of p . For better accuracy, stick to the binomial distribution unless p is much smaller (e.g., < 0.01).

Q 3. A company wants to build an automated spam email filter to classify incoming emails as either "Spam" or "Not Spam" (Ham). Given the huge volume of emails and the need for real-time performance, they choose Naïve Bayes due to its simplicity and speed. Use the Naïve Bayes classifier to predict the label of the following new email:

"Congratulations! You have won money quickly" . **[5 Marks]**

Sample Dataset:

Email ID	Text (Bag of Words)	Label
1	Congratulations, you have won a free prize	Spam
2	Monthly meeting scheduled for Monday	Not Spam
3	Earn money quickly with this simple trick	Spam
4	Project report attached. Review by Friday	Not Spam

Soln:

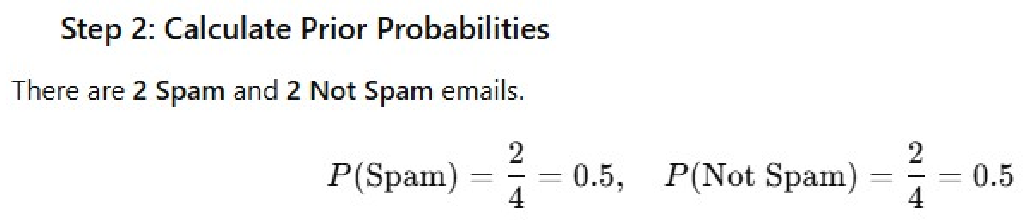
Step 1: Create Vocabulary

From the training data:

1. **Spam Emails**:
    - "Congratulations, you have won a free prize"
    - "Earn money quickly with this simple trick"
2. **Not Spam Emails**:
    - "Monthly meeting scheduled for Monday"
    - "Project report attached. Review by Friday"

Create a vocabulary of all **unique words**, after basic preprocessing (lowercasing and tokenizing):

Vocabulary: [
"congratulations", "you", "have", "won", "a", "free", "prize",
"earn", "money", "quickly", "with", "this", "simple", "trick",
"monthly", "meeting", "scheduled", "for", "monday",
"project", "report", "attached", "review", "by", "friday"
]



Step 2: Calculate Prior Probabilities

There are **2 Spam** and **2 Not Spam** emails.

$$P(\text{Spam}) = \frac{2}{4} = 0.5, \quad P(\text{Not Spam}) = \frac{2}{4} = 0.5$$

Step 3: Calculate Likelihood with Laplace Smoothing

Spam Class Word Counts:

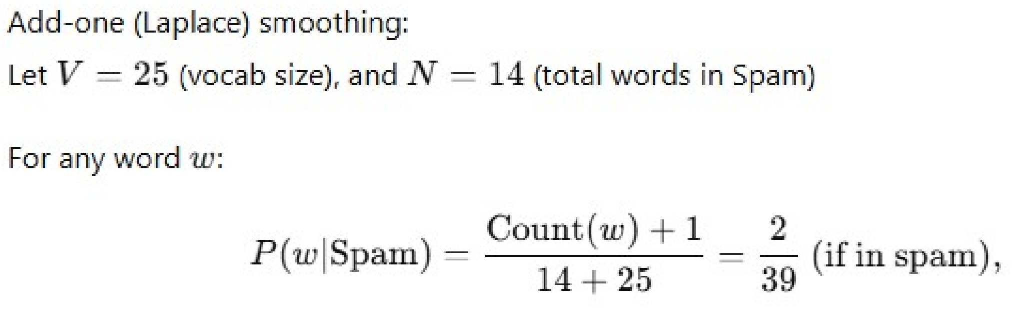
Spam emails contain:

- "Congratulations", "you", "have", "won", "a", "free", "prize"
- "Earn", "money", "quickly", "with", "this", "simple", "trick"

Total words in Spam = 7 + 7 = **14**

Word frequencies in Spam:

Word	**Count**
congratulations	1
you	1
have	1
won	1
a	1
free	1
prize	1
earn	1
money	1
quickly	1
with	1
this	1
simple	1
trick	1



Add-one (Laplace) smoothing:

Let $V = 25$ (vocab size), and $N = 14$ (total words in Spam)

For any word w:

$$P(w|\text{Spam}) = \frac{\text{Count}(w) + 1}{14 + 25} = \frac{2}{39} \text{ (if in spam)},$$

P(w/not Spam)= 1/36

P(Spam/w)=$\frac{P(Spam)X(w/Spam)}{P(w)}$= 9.09 x 10^{-9}

P(NotSpam/w)=$\frac{P(NotSpam)X(w/NotSpam)}{P(w)}$= 2.296 x 10^{-10}

Not Spam

None of the words in the test email appear in Not Spam class. So each term becomes:

P(email | Not\ Spam) = 5.635 × 10^{-10}

Compare Probabilities

P(email | Spam) ≈ 1.685 × 10^{-8}
P(email | Not\ Spam) ≈ 5.635 × 10^{-10}

P(email | Spam) > P(email | Not\ Spam)

The email "Congratulations! You have won money quickly" is classified as: SPAM

1. Tokenized and counted words.
2. Applied Laplace smoothing.
3. Calculated likelihoods for both classes.
4. Compared posterior probabilities.
5. Classified as **Spam** due to higher probability.

Q 4. A new test is developed to detect a rare genetic disorder. The disorder affects 0.1% of the population. The test has:
- Sensitivity (true positive rate): 98%
- Specificity (true negative rate): 97%

What is the probability that a person has the disorder if their test result is positive?

Solution:

Let:
- D: person has the disorder
- T^+: positive test result

Given:
- $P(D) = 0.001$
- $P(T^+|D) = 0.98$
- $P(T^+|\neg D) = 0.03$
- $P(\neg D) = 0.999$

Using Bayes' Theorem:

$$P(D|T^+) = (P(T^+|D) * P(D)) / (P(T^+|D) * P(D) + P(T^+|\neg D) * P(\neg D))$$
$$= (0.98 * 0.001) / (0.98 * 0.001 + 0.03 * 0.999)$$
$$= 0.00098 / 0.03095 \approx 0.0316$$

Answer: About 3.16% chance the person has the disorder despite a positive test.

Q 5. Suppose a machine learning model predicts the number of errors in a document using a discrete random variable X, which takes values $\{0,1,2,3\}$. The probability mass function (PMF) of X is given as:

$$P(X=x)=\begin{cases} k & if\ x-0 \\ 2k & if\, x=1 \\ 3k & if\, x=2 \\ 4k & if\, x=3 \\ 0 & otherwise \end{cases}$$

a) Find the value of k , b) Compute Expectation of X c) What does the expected value $E(X)$ mean in this context? c) $Compute\ var(2X+3)$ (5M)

Solution:

a) Since P(X) is a probability function, $\sum P(X) = 1$. Therefore,

$$k + 2k + 3k + 4k + 0 = 1,$$

$$k - \frac{1}{10}$$ (1 M)

b) Expectation of $X = E(X) = \sum x.P(X) = 0.k + 1.2k + 2.3k + 3.4k = 20k$ = 20*1/10

$$E(X) = \frac{20}{10} = 2$$ (1 M)

c) The **expected value** $E(X) = 2$ means that, on average, the model predicts **2 errors** (1 M)

d) $E(X^2) = \sum x^2.P(x) = 0^2.k + 1^2.2k + 2^2.3k + 3^2.4k = 50k = \frac{50}{10} = 5$

Therefore, $var(X) = 5 - (2)^2$

$var(X) = 1$ (1 M)

Hence, Variance of 2X+ 3 = Var(2X+3)

= 2^2 *Var(X) + 0

= 4*Var(X)

= 4*1

$var(2X+3) = 4$ (1 M)

Q 6. A fair coin is tossed three times. The outcome of each toss is recorded as either a head (H) or a tail (T). Let X be random variable which observed the number of tails observed in three tosses and Y be random variable represent the player's winnings(in \$), based on position of first tail according to the following rules :

“The player **wins $3** if the **first tail** appears on the **first toss, $2** if the **first tail** appears on the **second toss**, **wins $1** if the **first tail** appears on the **third toss** and The player **loses $2** if there is **no tail** in all three tosses”

a) Find Joint probability function of X and Y

b) Marginal Probability function for X .

c) E(X) [5 Marks]

Solution :

a) Random variable X assign $\{0,1,2,3\}$ and Y assigns $\{-2,1,2,3\}$ to the sample points. ______________________________(2M)

The joint probability function $P(X,Y)$ is given by:

Y X	-2	1	2	3
0	$\frac{1}{8}$	0	0	0
1	0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
2	0	0	$\frac{1}{8}$	$\frac{2}{8}$
3	0	0	0	$\frac{1}{8}$

____________(2M)

b) Marginal Probability function for X:

$$f_1(0)=\frac{1}{8}+0+0+0=\frac{1}{8}, \qquad f_1(1)=0+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}=\frac{3}{8},$$

$$f_1(2)=0+0+\frac{1}{8}+\frac{2}{8}=\frac{3}{8}, \qquad f_1(3)=0+0+0+\frac{1}{8}$$

c) $E(X)=\sum x f_1(x)=0.\frac{1}{8}+1.\frac{3}{8}+2.\frac{3}{8}+3.\frac{1}{8}=\frac{12}{8}=1.5$ (1 M)

****