National University of Computer and Emerging Sciences

Karachi Campus

Probability and Statistics (MT2005)

Date: 14-05 -2025

Course Instructor(s)

Mr. Moheez Ur Rahim

Final Exam

(Special)

Total Time: 3 Hours

Total Marks: 100

Total Questions: 06

Attempt all the questions.

CLO 1: Describe the fundamental concepts in probability and statistics

Q1: [6 + 4 marks]

(a) Consider the following dataset of employee records. Fill the missing values with the appropriate measure of central tendency (mean, media or mode)

Employee Id	1	2	3	4	5	6	7
Age	32	25	?	41	38	?	2
Salary	45000	?	52000	480000	?	51000	46000
Gender	Male	Female	Female	Male	?	Female	Female
Performance Rating (1-5)	4	5	?	3	4	2	?

(b) A machine learning model is trained to classify emails as spam (S) or not spam (¬S) based on the presence of the word "free" in the email. The known probabilities are: The probability that any email is spam: P(S) = 0.3, If an email is spam, the probability it contains the word "free" is 0.8, and If an email is not spam, the probability it contains the word "free" is 0.1. Compute the probability that an email is spam given that it contains the word "free"

CLO 2: Analyze the data and produce probabilistic models for different problems

Q2: [2+2+2+4 marks]

Two sensors (X=power consumption in watts, Y=Temperature in centigrade) in a robot collect continuous measurements. The Joint PDF is given by:

$$f(x,y) = \begin{cases} K(x^2y + xy^2) & if\ 0 \leq x \leq 1, and\ 0 \leq y \leq 1 \\ 0 & otherwise \end{cases}$$

(i) Find the constant K that makes $f(x,y)$ a valid Joint PDF

(ii) Find the marginal PDFs of X and Y

(iii) Compute $P(X = 0.5|Y < 0.1)$

(iv) Find the covariance between X and Y if they are dependent

CLO 3: Apply the rules and algorithm of probability and statistics to relevant problems

Q3: **[5+7+3 Marks]**

(a) A CS researcher tests two sorting algorithms (QuickSort and MergeSort) on the same set of 5 input arrays to compare their runtime (in milliseconds). The results are below:

Dataset ID	1	2	3	4	5
QuickSort (X_i)	12.4	8.7	21.5	15.2	10.9
MergeSort (Y_i)	14.1	9.3	22.0	16.8	11.5

Estimate the 95% confidence interval for true mean difference ($\mu_y - \mu_x$). Assuming that population is Normal

(b) A developer compares the response times (ms) of two web servers, NodeJS (X) and Go (Y), with small independent samples: NodeJS (45, 52, 48, 50) and Go (38, 42, 35, 40, 36). Assuming unequal variances and normality, compute the 90% confidence interval for the mean difference (μ_x - μ_y).

(c) A cloud service claims its average API response time is 50 ms with a known population standard deviation of 8 ms. A sample of 64 requests shows a mean of 53 ms. Calculate the upper bound of the 95% confidence interval.

CLO 3: Apply the rules and algorithm of probability and statistics to relevant problems

Q4: **[7+5+3 Marks]**

(a) A psychologist wants to investigate whether there is a significant difference in the average stress levels between students who study in the morning and those who study at night. She records stress level scores from two independent samples: 10 students in the morning group with scores 12, 14, 15, 13, 16, 14, 13, 15, 14, 13 and 12 students in the night group with scores 17, 16, 15, 18, 17, 19, 16, 17, 18, 16, 17, 18. Assuming that the populations are normal with equal variances, conduct a two-sample t-test at the 5% significance level to determine if there is a significant difference in the population means.

(b) A tech company wants to evaluate whether the average response times of two different server architectures differ by more than 50 milliseconds under load. A random sample of 150 response times from Server A shows a mean of 620 ms with a standard deviation of 80 ms, while a sample of 170 response times from Server B shows a mean of 560 ms with a standard deviation of 95 ms. Assuming the samples are independent and the population variances are unequal, test at the 5% significance level whether the true difference in mean response times between the two servers is significantly different from 50 milliseconds.

(c) A computer science student claims that a new algorithm reduces the average execution time of a sorting task to less than 2.5 seconds. To verify this, a professor records the execution times from a sample of 8 runs: 2.4, 2.6, 2.3, 2.5, 2.2, 2.4, 2.1, and 2.3 seconds. Assuming the execution times are normally distributed and the population standard deviation is unknown, conduct test at the 5% significance level to determine whether the true mean execution time is significantly less than 2.5 seconds.

LO 3: Apply the rules and algorithm of probability and statistics to relevant problems

Q5: **[30 + 10 Marks]**

(a) A researcher wants to study the relationship between the number of hours computer science students spend coding per week and their scores on a programming assessment. The data below was collected from 10 students: **[7+4+2+2+4+6+5 Marks]**

Hours Spent Coding (X):	10	12	14	16	18	20	22	24	26	28
Assessment Score (Y):	65	68	72	74	78	80	83	85	88	90

i. Compute $\sum x, \sum y, \sum xy, \sum x^2, \sum y^2$ and estimated regression coefficients b_0 and b_1

ii. Compute S_{xx}, S_{yy}, S_{xy} and $s^2 = S_{yy} - b_1 S_{xy}$

iii. Estimate the regression line $\hat{y} = b_0 + b_1 x$

iv. Predict Assessment score for student who spent 11 hours on coding

v. Compute sample coefficient of correlation and determination and comment on it

vi. Compute the 95% confidence interval for true regression coefficients β_0 and β_1

vii. Test the correlation co-efficient $\rho = 0$ against $\rho \neq 0$ at 5% level of significance.

(b) Assume, you are working as a data scientist at a company that monitors network request latency in a distributed system. The company has collected the following data points: Request Size (in kilobytes, x) and Latency (in milliseconds, y): (50, 100), (100, 150), and (150, 200). Compute the regression coefficients using gradient descent for the equation $\hat{y} = b_0 + b_1 x$, with a learning rate $\alpha = 0.01$. Perform 3 iterations of gradient descent, compute the updated values of b_0 and b_1. Calculate the Mean Squared Error (MSE) or loss after each iteration using the below formula:

$MSE = Loss = L = \frac{1}{2n}\sum(\hat{y} - y)^2$, where n is the number of data points.

CLO 3: Apply the rules and algorithm of probability and statistics to relevant problems

Q6: **[10 Marks]**

A software development company has collected data on the time (in hours) taken by 6 developers to complete a task using five programming languages: Python, JavaScript, Ruby, Java and C#. Data is below:

Phyton	JavaScript	Ruby	Java	C#
12	10	14	11	9
14	11	15	12	10
13	12	13	13	9
15	13	16	14	11
14	12	14	13	10
16	9	17	12	8

At the 0.05 significance level, determine if there is a significant difference in the average completion time across the languages using a one-way ANOVA test, with the null hypothesis stating no difference and the alternative hypothesis suggesting at least one difference.

---The End---