National University of Computer and Emerging Sciences

Karachi Campus

Probability and Statistics (MT2005)

Date: 29-05 -2025

Course Instructor(s)

Dr. S.M.Fahad Riaz, Dr. Khusro Mian, Mr. Nadeem Arif, Ms. Asma Masood, Mr. Moheez Ur Rahim

Final Exam

Total Time: 3 Hours

Total Marks: 100

Total Questions: 06

_____ _____ _____

Roll No Section Student Signature

Attempt all the questions.

CLO 1: Describe the fundamental concepts in probability and statistics

Q1: [6 + 4 marks]

(a) In a university's Programming Fundamentals course, the exam scores (out of 100) of male and female students were recorded to analyze performance based on gender. The scores for male students are: 62, 65, 67, 70, 72, 75, 77, 80, 82, 85, 87, 90, 92, 94, 95. The scores for female students are: 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88. Draw two box plots, one for male and one for female student, on the same scale. Then, based on the box plots, comment on the performance of both groups in terms of spread, median, and any patterns you observe.

Step 1: Organize Data
Male Scores:
62, 65, 67, 70, 72, 75, 77, 80, 82, 85, 87, 90, 92, 94, 95 (1 Mark)

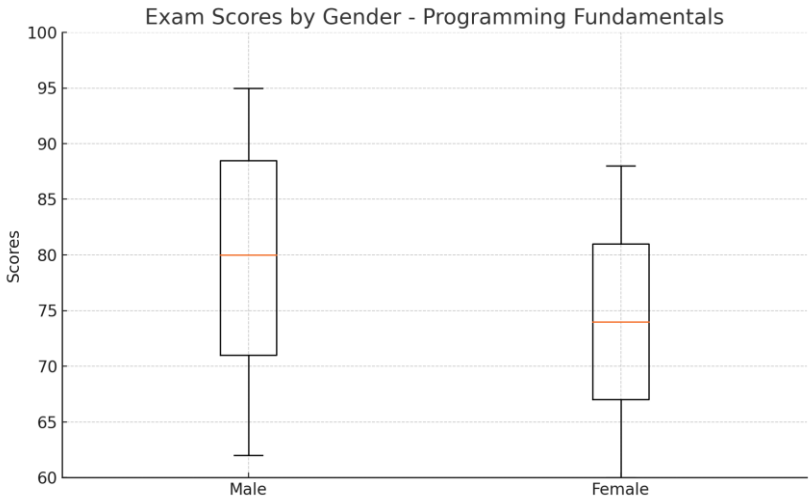
Female Scores:
60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88 (1 Mark)

Step 2: Five-Number Summary

Male Students: (1 Mark)
Min = 62
Q1 = 70
Median = 80
Q3 = 90
Max = 95

Female Students: (1 Mark)
Min = 60
Q1 = 66
Median = 74
Q3 = 82
Max = 88

Step 3: Box Plot (1 Mark)



Step 4: Comments Based on Box Plots (1 Mark)

Scores of male student skewed slightly higher, wider upper range whereas scores of female students more concentrated, slightly lower overall.

(b) A computer program checks if a website is a phishing site (a fake website that tries to steal your personal information) by looking for the word "login" in the website link. The chance that any website is a phishing site is 25%. If a website is a phishing site, there is a 90% chance that the word "login" appears in its link. On the other hand, if the website is not a phishing site, there is a 20% chance that the word "login" appears. Based on this information, what is the probability that a website is a phishing site if its link contains the word "login"?

Solution:

(1 mark) Correctly identify and state the relevant probabilities:

- $P(P) = 0.25, P(L|P) = 0.90, P(L|P^c) = 0.20, P(P^c) = 0.75.$

(1 mark) Write down the formula for total probability of L:

$$P(L) = P(L|P)P(P) + P(L|P^c)P(P^c)$$

(1 mark) Calculate $P(L)$ correctly:

$$P(L) = 0.225 + 0.15 = 0.375$$

(1 mark) Use Bayes' theorem correctly and compute $P(P|L)$:

$$P(P|L) = \frac{P(L|P)P(P)}{P(L)} = \frac{0.225}{0.375} = 0.6$$

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

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

A tech company studies how hours spent on code optimization (X) and testing (Y) affect software performance, with their joint density given by

$$f(x,y) = \begin{cases} C(x^2+y) & if\ -1 \le x \le 1, and\ 0 \le y \le 1 \\ 0 & otherwise \end{cases}$$

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

Solution:

(a) Find the constant C that makes $f(x, y)$ a valid PDF. [2 marks]

$$\int_{-1}^{1}\int_{0}^{1} C(x^2 + y)\, dy\, dx = 1$$

Inner integral:

$$\int_{0}^{1} (x^2 + y)dy = x^2 \times 1 + \frac{1^2}{2} = x^2 + \frac{1}{2}$$

Outer integral:

$$\int_{-1}^{1} \left(x^2 + \frac{1}{2}\right) dx = \int_{-1}^{1} x^2 dx + \int_{-1}^{1} \frac{1}{2} dx = \frac{2}{3} + 1 = \frac{5}{3}$$

Therefore,

$$C \times \frac{5}{3} = 1 \implies C = \frac{3}{5}$$

Marking Scheme: Inner Integral = 1 Mark, Outer integral + correct value of c = 1 mark

(ii) Find the marginal PDFs of X and Y

Solution:

(b) Find the marginal PDFs $f_X(x)$ and $f_Y(y)$. [2 marks]

- Marginal of X:

$$f_X(x) = \int_{0}^{1} f(x, y)dy = \int_{0}^{1} C(x^2 + y)dy = C\left(x^2 + \frac{1}{2}\right), \quad -1 \le x \le 1$$

- Marginal of Y:

$$f_Y(y) = \int_{-1}^{1} f(x, y)dx = \int_{-1}^{1} C(x^2 + y)dx = C\left(\frac{2}{3} + 2y\right), \quad 0 \le y \le 1$$

Marking:

- Correct expression for $f_X(x)$ with limits **(1 mark)**
- Correct expression for $f_Y(y)$ with limits **(1 mark)**

(iii) Compute $P(Y < 0.6|X < 0.5)$

Solution:

$$= \frac{\int_{-1}^{0.5}\int_{0}^{0.6} C(x^2 + y)dydx)}{\int_{-1}^{0.5} C\left(x^2 + \frac{1}{2}\right) dx}$$

Step 1: Cancel C (since it's common factor numerator and denominator):

$$= \frac{\int_{-1}^{0.5}\int_{0}^{0.6}(x^2+y)\,dy\,dx}{\int_{-1}^{0.5}\left(x^2+\frac{1}{2}\right)dx}$$

Step 2: Solve numerator integral

Inner integral:

$$\int_0^{0.6}(x^2+y)dy = x^2\times 0.6 + \frac{0.6^2}{2} = 0.6x^2+0.18$$

Outer integral:

$$\int_{-1}^{0.5}(0.6x^2+0.18)dx = 0.6\int_{-1}^{0.5}x^2dx + 0.18\int_{-1}^{0.5}dx$$

Calculate each:

$$\int_{-1}^{0.5}x^2dx = \left[\frac{x^3}{3}\right]_{-1}^{0.5} = \frac{0.125}{3} - \left(-\frac{1}{3}\right) = \frac{1.125}{3} = 0.375$$

$$\int_{-1}^{0.5}dx = 0.5-(-1) = 1.5$$

Numerator:

$$0.6\times 0.375 + 0.18\times 1.5 = 0.225 + 0.27 = 0.495$$

Above working (step 1 and step 2) = 1.5 marks

Step 3: Solve denominator integral

$$\int_{-1}^{0.5}\left(x^2+\frac{1}{2}\right)dx = \int_{-1}^{0.5}x^2dx + \frac{1}{2}\int_{-1}^{0.5}dx = 0.375+0.75 = 1.125$$

Step 4: Calculate conditional probability

$$P(Y<0.6 \mid X<0.5) = \frac{0.495}{1.125} = 0.44$$

Final answer:

$$\boxed{0.44}$$

Above working (step3 and step4) = 0.5 marks

(iv) Are X and Y independent? If no, then compute Covariance between X and Y

Solution:

Step 1: Check independence

$$f(x,y) = C(x^2+y), \quad f_X(x) = C\left(x^2+\frac{1}{2}\right), \quad f_Y(y) = C\left(\frac{2}{3}+2y\right)$$

Check if:

$$f(x,y) \stackrel{?}{=} f_X(x) f_Y(y)$$

Since:

$$C(x^2+y) \neq C^2\left(x^2+\frac{1}{2}\right)\left(\frac{2}{3}+2y\right)$$

Conclusion: Not independent.

(Mark: 1 mark for correct conclusion)

Step 2: Calculate $E[X]$

$$E[X] = \int_{-1}^{1} x f_X(x) dx = C\int_{-1}^{1} x\left(x^2+\frac{1}{2}\right) dx = C\int_{-1}^{1} (x^3+\frac{x}{2}) dx$$

Since integrals of odd functions over symmetric interval are zero:

$$E[X] = 0$$

(Mark: 1 mark for correct $E[X]$ calculation)

Step 3: Calculate $E[Y]$

$$E[Y] = C\int_{0}^{1} y\left(\frac{2}{3}+2y\right) dy = C\int_{0}^{1} \left(\frac{2}{3}y+2y^2\right) dy$$

Calculate integrals:

$$\int_{0}^{1} y dy = \frac{1}{2}, \quad \int_{0}^{1} y^2 dy = \frac{1}{3}$$

Substitute back:

$$E[Y] = C\left(\frac{2}{3}\times\frac{1}{2} + 2\times\frac{1}{3}\right) = C\times 1 = 0.6$$

(Mark: 1 mark for correct $E[Y]$ calculation)

Step 4: Calculate $E[XY]$

$$E[XY] = C\int_{-1}^{1} x \int_{0}^{1} y(x^2+y)dydx = C\int_{-1}^{1} x\left(\frac{x^2}{2}+\frac{1}{3}\right)dx = C\int_{-1}^{1}\left(\frac{x^3}{2}+\frac{x}{3}\right)dx = 0$$

(Mark: 1 mark for correct $E[XY]$ calculation and covariance)

Step 5: Calculate covariance

$$\text{Cov}(X,Y) = E[XY] - E[X]E[Y] = 0 - 0 \times 0.6 = 0$$

Final answer:

- X and Y are **not independent.**
- $\text{Cov}(X,Y) = 0$

So, there is dependence (non-linear dependence)**,** but no linear correlation**,** which results in zero covariance.

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

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

(a) A tech company wants to estimate the difference in average screen time (in hours per day) between employees working in-office and those working remotely. A sample of 10 in-office employees reported screen times of 6.2, 6.5, 6.8, 7.0, 6.4, 6.6, 6.7, 6.9, 6.3, and 6.5 hours, while a sample of 12 remote employees reported 7.5, 7.2, 7.8, 7.6, 7.4, 7.3, 7.7, 7.5, 7.6, 7.8, 7.4, and 7.3 hours. Assuming that screen times are normally distributed with equal population variances, construct a 95% confidence interval for the difference in population means ($\mu_{office} - \mu_{remote}$) ($critical\ value = \pm 2.086$)

Solution:

Step 1: State hypotheses (1 mark)

$$H_0 : \mu_{office} = \mu_{remote} \quad vs. \quad H_a : \mu_{office} \neq \mu_{remote}$$

Step 2: Calculate sample means and variances (2 marks)

- $\bar{x}_1 = 6.59,\ s_1^2 = 0.0677$ (Office, $n_1 = 10$)
- $\bar{x}_2 = 7.51,\ s_2^2 = 0.0390$ (Remote, $n_2 = 12$)

Step 3: Calculate pooled variance s_p^2 (1 mark)

$$s_p^2 = \frac{9 \times 0.0677 + 11 \times 0.0390}{20} = 0.0519$$

Step 4: Calculate standard error (1 mark)

$$SE = \sqrt{0.0519 \times \left(\frac{1}{10}+\frac{1}{12}\right)} = 0.0976$$

Step 5: Construct 95% confidence interval (2 marks)

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \times SE = -0.92 \pm 2.086 \times 0.0976 = (-1.12, -0.71)$$

Interpretation:

The average screen time for office workers is estimated to be between **0.71 and 1.12 hours less** than remote workers, with 95% confidence.

(b) A performance analyst compares the processing times (in milliseconds) of two database systems, System A and System B, using independent samples. The recorded times for System A are 58, 62, 55, 60, 63, 59, 61, and 57, while for System B they are 50, 48, 52, 47, 49, 51, 53, and 46. Assuming the processing times are normally distributed with unequal variances, compute the 90% confidence interval for the difference in mean processing times ($\mu_x - \mu_y$). ($critical\ value = \pm 1.761$)

Solution:

Step 1: Calculate sample means and variances (2 marks)

- $\bar{x} = 59.375, \quad s_x^2 = 7.12$
- $\bar{y} = 49.5, \quad s_y^2 = 6$

Step 2: Calculate standard error and degrees of freedom (1.5 marks)

- Standard error:

$$SE = \sqrt{\frac{7.12}{8} + \frac{6}{8}} = 1.28$$

- Degrees of freedom (Welch's approximation):

$$df \approx 14$$

Step 3: Calculate confidence interval and interpret (1.5 marks)

$$(\bar{x} - \bar{y}) \pm t^* \times SE = 9.875 \pm 1.761 \times 1.28 = (7.625, 12.125)$$

Interpretation: We are 90% confident that System A's mean processing time is between 7.625 ms and 12.125 ms longer than System B.

(c) A developer wants to estimate the minimum average load time of a new app feature. From 10 trials, the load times (in seconds) recorded are: 1.8, 2.0, 1.9, 2.1, 1.7, 1.9, 2.0, 1.8, 1.9, and 2.2. Assuming load times are normally distributed with unknown population standard deviation, construct a 95% confidence lower bound for the true mean load time. ($critical\ value = 1.8333$)

Solution:

Step 1: Calculate sample mean $\bar{x}$ and sample standard deviation s

(1 mark)

$$\bar{x} = \frac{1.8 + 2.0 + \cdots + 2.2}{10} = \frac{19.3}{10} = 1.93$$

Calculate s:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{0.221}{9}} = \sqrt{0.02456} = 0.1567$$

Step 2: Calculate standard error (SE)

(1 mark)

$$SE = \frac{s}{\sqrt{n}} = \frac{0.1567}{\sqrt{10}} = \frac{0.1567}{3.162} = 0.0495$$

Step 3: Calculate 95% lower confidence bound for mean

(1 mark)

Since it's a **lower bound**, formula is:

$$\bar{x} - t_\alpha \times SE = 1.93 - 1.8333 \times 0.0495 = 1.93 - 0.0907 = 1.8393$$

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

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

(a) A software engineer compares the runtimes (in milliseconds) of two image compression algorithms, Algorithm A and Algorithm B, on the same set of 5 test images. The recorded runtimes are:

Image ID	1	2	3	4	5
Algorithm A (X_i)	120.5	98.7	150.3	130.2	110.4
Algorithm B (Y_i)	125.8	101.2	155.6	135.7	115.0

At a 5% significance level, test whether there is a significant difference in the mean runtimes ($\mu_y - \mu_x \neq 0$) of the two algorithms. Assume the runtimes are normally distributed. ($critical\ value = \pm 2.776$)

Solution:

Step 1: State hypotheses

(1 mark)

$$H_0 : \mu_d = 0, \quad H_a : \mu_d \neq 0$$

Step 2: Calculate differences and their mean and SD

Differences $d_i = Y_i - X_i$:

$$d = [5.3, 2.5, 5.3, 5.5, 4.6]$$

Given (direct):

$$\bar{d} = 4.64, \quad s_d = 1.243, \quad n = 5$$

(2 marks)

Step 3: Calculate standard error (SE)

$$SE = \frac{s_d}{\sqrt{n}} = \frac{1.243}{\sqrt{5}} = 0.556$$

(1 mark)

Step 4: Calculate test statistic

$$t = \frac{\bar{d} - 0}{SE} = \frac{4.64}{0.556} = 8.347$$

(1 mark)

Step 5: Compare with critical value and conclude

Critical value: ± 2.776 (df=4)

Since $8.347 > 2.776$, reject H_0.

Conclusion: There is a significant difference in mean runtimes at 5% significance level.

(2 marks)

(b) A cloud engineer wants to verify whether the average API response time of a new micro service exceeds the industry standard of 50 milliseconds. The population standard deviation is known to be 8 milliseconds. From a random sample of 64 requests, the observed mean response time is 53 milliseconds. At the 5% significance level, test whether the true mean response time is significantly greater than 50 milliseconds. ($critical\ value = 1.645$)

Solution:

Step 1: State hypotheses

(1 mark)

$$H_0 : \mu = 50 \quad (\text{mean response time} = \text{industry standard})$$

$$H_a : \mu > 50 \quad (\text{mean response time exceeds standard})$$

Step 2: Calculate test statistic Z

(1 mark)

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{53 - 50}{8/\sqrt{64}} = \frac{3}{1} = 3$$

Step 3: Compare with critical value and conclude

(1 mark)

- Since $Z = 3 > 1.645$, reject H_0.
- There is sufficient evidence at 5% significance level that the mean API response time exceeds 50 ms.

(c) A hardware engineer compares the average battery life (in hours) of two types of laptop batteries. For Battery A, a sample of 40 units shows a mean battery life of 12.5 hours, with a known population standard deviation of 1.8 hours. For Battery B, a sample of 35 units shows a mean of 10.3 hours, with a known population standard deviation of 2.0 hours. At the 5% significance level, test whether Battery A has a significantly longer average battery life than Battery B. ($critical\ value = 1.645$)
Solution:

Step 1: State hypotheses

(1 mark)

$$H_0 : \mu_1 \leq \mu_2 \quad \text{(Battery A is not longer)}$$

$$H_a : \mu_1 > \mu_2 \quad \text{(Battery A has longer battery life)}$$

Step 2: Calculate test statistic Z

(2 marks)

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{12.5 - 10.3}{\sqrt{\frac{1.8^2}{40} + \frac{2.0^2}{35}}}$$

Calculate denominator:

$$\sqrt{\frac{3.24}{40} + \frac{4.00}{35}} = \sqrt{0.081 + 0.114} = \sqrt{0.195} = 0.441$$

Calculate Z:

$$Z = \frac{2.2}{0.441} = 4.99$$

Step 3: Decision and conclusion

(2 marks)

- Since $Z = 4.99 > 1.645$, reject H_0.
- Conclusion: Battery A has significantly longer average battery life than Battery B at the 5% significance level.

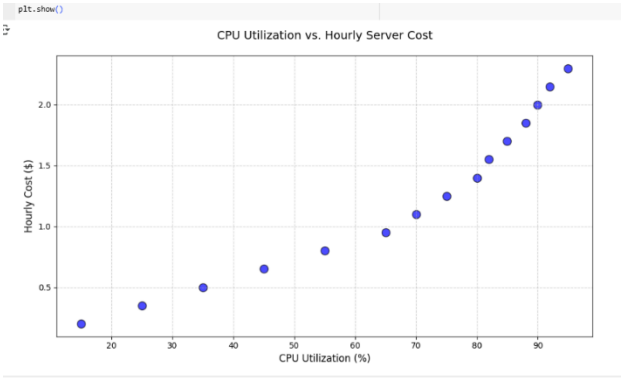
CLO 3: Apply the rules and algorithm of probability and statistics to relevant problems

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

(a) A cloud computing company wants to model the relationship between CPU utilization (%) and hourly server cost ($) for their virtual machine (VM) instances. The DevOps team collected usage data from 15 different VM instances running various workloads. **[3+4+3+3+2+5+5+5 Marks]**

CPU Utilization % (X)	15	25	35	45	55	65	70	75	80	82	85	88	90	92	95
Hourly Cost $ (Y)	0.2	0.35	0.5	0.65	0.8	0.95	1.1	1.25	1.4	1.55	1.7	1.85	2	2.15	2.3

i. Create a scatter plot of CPU Utilization vs. Hourly Cost and comment on the apparent relationship.
Solution:



Graph= (2 marks)

Comment: Strong positive linear relationship. (1 Mark)

ii. Compute the regression line equation ($\hat{Y} = a + bX$). Also, interpret the slope coefficient in terms of cloud pricing and explain the practical meaning of the y-intercept.

Solution:

Given sums:

- $n = 15$
- $\sum X_i = 997$
- $\sum Y_i = 18.75$
- $\sum X_i^2 = 75558$
- $\sum X_i Y_i = 1477.7$

Step 1: Calculate slope b

$$b = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \frac{15 \times 1477.7 - 997 \times 18.75}{15 \times 75558 - 997^2} = \frac{3471.75}{139361} = 0.0249$$

(1 mark)

Step 2: Calculate intercept a

Calculate means first:

$$\bar{X} = \frac{997}{15} = 66.47, \quad \bar{Y} = \frac{18.75}{15} = 1.25$$

Then:

$$a = \bar{Y} - b\bar{X} = 1.25 - 0.0249 \times 66.47 = 1.25 - 1.655 = -0.405$$

(1 mark)

Step 3: Interpret slope b

Slope 0.0249 means:

For each 1% increase in CPU utilization, the hourly cost increases by about 2.49 cents.

(1 mark)

Step 4: Interpret intercept a

Intercept -0.405 means:

At 0% CPU utilization, the predicted hourly cost is -$0.405, which is not practical but a linear model artifact.

(1 mark)

iii. Calculate and interpret the coefficient of determination (R^2).

Solution:

$R^2 = 0.923$ (whole working = 2 marks)

Interpretatin (1 Mark)

About 92.3% of the variability in hourly cost is explained by the linear relationship with CPU utilization.

This indicates a very strong linear relationship

iv. Compute the correlation coefficient (r) and interpret it.

Solution:

$$r = \sqrt{0.915} = 0.957$$

(Since the slope b is positive, r is positive.)

(Whole working = 2 Marks)

Interpretation (1 Mark)

r=0.957 indicates a very strong positive linear relationship between CPU utilization and hourly cost. This means as CPU utilization increases, hourly cost tends to increase strongly and linearly.

v. Estimate the hourly cost for a VM running at 78% CPU utilization.
Solution:
Predicted cost at 78% CPU utilization: $1.537 (2 Marks)

vi. Conduct hypothesis testing for the correlation coefficient $\rho = 0$ against $\rho \neq 0$ at 5% level of significance ($Critical\ value = \pm 2.160$)
Solution:

Step 1: State hypotheses

$$H_0 : \rho = 0 \quad \text{(no linear correlation)}$$

$$H_a : \rho \neq 0 \quad \text{(linear correlation exists)}$$

(Mark: 1 mark)

Step 2: Calculate test statistic t

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Using $r = 0.957, n = 15$:

$$t = 0.957 \times \sqrt{\frac{13}{1-0.957^2}} = 0.957 \times \sqrt{\frac{13}{1-0.916}} = 0.957 \times \sqrt{\frac{13}{0.084}} = 0.957 \times 12.44 = 11.91$$

(Mark: 2 marks)

Step 3: Decision

Degrees of freedom $= n - 2 = 13$
Critical value at 5% significance = ±2.160
Since $t = 11.91 > 2.160$, reject H_0.
(Mark: 1 mark)

Step 4: Conclusion

There is sufficient evidence at 5% level to conclude a significant correlation exists between CPU utilization and hourly cost.
(Mark: 1 mark)

vii. At 5% significance level, check whether the regression slope coefficient is significant ($\beta_1 \neq 0$). ($Critical\ value = \pm 2.160$)
Solution:

Step 1: State hypotheses

$$H_0 : \beta_1 = 0 \quad \text{(slope not significant)}$$

$$H_a : \beta_1 \neq 0 \quad \text{(slope significant)}$$

(1 mark)

Step 2: Calculate test statistic

$$t = \frac{b_1 - \beta_1}{SE_{b_1}} = \frac{b_1 - 0}{SE_{b_1}} = \frac{b_1}{SE_{b_1}}$$

Where:

- $b_1 = 0.0249$ (estimated slope)
- $SE_{b_1} = \frac{s}{\sqrt{S_{XX}}}$ (standard error of slope)

Calculate residual standard error s:

$$SSE = S_{YY} - b_1 \times S_{XY} = 6.30 - 0.0249 \times 232.6 = 0.51$$

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{0.51}{13}} = 0.198$$

Calculate SE_{b_1}:

$$SE_{b_1} = \frac{0.198}{\sqrt{9300}} = \frac{0.198}{96.4} = 0.00205$$

Calculate t:

$$t = \frac{0.0249}{0.00205} = 12.15$$

(3 marks)

Step 3: Decision and conclusion

- Critical value at 5% significance, df=13: ± 2.160
- Since $12.15 > 2.160$, **reject** H_0.
- Conclusion: The slope coefficient is significantly different from zero; CPU utilization significantly predicts hourly cost.

(1 mark)

viii. At 5% significance level, check whether the intercept of true regression is significant ($\beta_0 \neq 0$). ($Critical\ value = \pm 2.160$)

Solution:

Step 1: State hypotheses

$$H_0 : \beta_0 = 0 \quad \text{(Intercept is zero — not significant)}$$

$$H_a : \beta_0 \neq 0 \quad \text{(Intercept is not zero — significant)}$$

(Mark: 1 mark)

Step 2: Calculate standard error denominator

$$SE = s \times \sqrt{\frac{1}{n} \times \frac{\sum x_i^2}{nS_{XX}}}$$

Given:

- Residual std error $s = 0.198$
- Sample size $n = 15$
- $\sum x_i^2 = 75558$
- $S_{XX} = 9300$

Calculate inside root:

$$\frac{1}{15} \times \frac{75558}{15 \times 9300} = \frac{1}{15} \times \frac{75558}{139500} = \frac{1}{15} \times 0.5417 = 0.03611$$

Calculate SE:

$$SE = 0.198 \times \sqrt{0.03611} = 0.198 \times 0.19 = 0.0376$$

(Mark: 2 marks)

Step 3: Calculate test statistic

$$T = \frac{b_0 - \beta_0}{SE} = \frac{-0.405 - 0}{0.0376} = -10.77$$

(Mark: 1 mark)

Step 4: Decision and conclusion

- Critical value at 5% (two-tailed), df=13: ±2.160
- Since $|-10.77| > 2.160$, reject H_0.
- **Conclusion:** Intercept is significantly different from zero.
  (Mark: 1 mark)

(b) A digital marketing team wants to predict click-through rates (CTR) based on ad display time (seconds). They have a dataset from an A/B test and want to implement linear regression via gradient descent. The model takes the form: $\hat{y} = w_0 + w_1 x$ (10 Marks)

Display Time (X)	1	2	3
CTR % (Y)	1.0	2.0	2.99

Use the gradient-descent algorithm with a learning rate of 0.2 to estimate the weights w_0 and w_1. Also, compute the loss function, $MSE = \frac{1}{2}\Sigma(\hat{y} - y)^2$ in each iteration. Perform iterations until the loss function converges such that difference of current iteration loss and previous iteration loss is less than 0.1. Take the initial weights as $w_0 = 0, w_1 = 0$.

Solution:

1. Generalized loss function and derivatives

Given training data points (x_i, y_i), $i = 1, \ldots, m$ and the model

$$\hat{y}_i = w_0 + w_1 x_i$$

The loss function is:

$$J(w_0, w_1) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 = \frac{1}{2m}\sum_{i=1}^{m}(w_0 + w_1 x_i - y_i)^2$$

Partial derivatives of J:

$$\frac{\partial J}{\partial w_0} = \frac{1}{m}\sum_{i=1}^{m}(w_0 + w_1 x_i - y_i)$$

$$\frac{\partial J}{\partial w_1} = \frac{1}{m}\sum_{i=1}^{m}(w_0 + w_1 x_i - y_i)x_i$$

(3 Marks)

2. Gradient Descent Update Rules:

Given learning rate α, update equations are:

$$w_0 := w_0 - \alpha\frac{\partial J}{\partial w_0} = w_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(w_0 + w_1 x_i - y_i)$$

$$w_1 := w_1 - \alpha\frac{\partial J}{\partial w_1} = w_1 - \alpha\frac{1}{m}\sum_{i=1}^{m}(w_0 + w_1 x_i - y_i)x_i$$

(1 Mark)

3. Run first iteration with given data:

Data:

$$m = 3, \quad (x, y) = (1, 1.0), (2, 2.0), (3, 2.99)$$

Initialization:

$$w_0 = 0, \quad w_1 = 0, \quad \alpha = 0.2$$

Calculate predictions and errors:

$$\hat{y}_i = w_0 + w_1 x_i = 0 \quad \Rightarrow \quad e_i = \hat{y}_i - y_i = [0 - 1.0, 0 - 2.0, 0 - 2.99] = [-1.0, -2.0, -2.99]$$

Compute gradients:

$$\frac{\partial J}{\partial w_0} = \frac{1}{3}(-1.0 - 2.0 - 2.99) = -1.9967$$

$$\frac{\partial J}{\partial w_1} = \frac{1}{3}(-1.0 \times 1 - 2.0 \times 2 - 2.99 \times 3) = \frac{1}{3}(-1 - 4 - 8.97) = -4.6567$$

Update weights:

$$w_0 := 0 - 0.2 \times (-1.9967) = 0 + 0.3993 = 0.3993$$

$$w_1 := 0 - 0.2 \times (-4.65 \downarrow = 0 + 0.9313 = 0.9313$$

Summary of first iteration:

- Updated weights after iteration 1:

$$w_0 = 0.3993, \quad w_1 = 0.9313$$

(2 Marks)

Iteration	w_0	w_1	Predictions $\hat{y}$	Errors $e = \hat{y} - y$	Loss $J = \frac{1}{2m}\sum e_i^2$
0	0	0	[0, 0, 0]	[-1.0, -2.0, -2.99]	2.3233
1	0.3993	0.9313	[1.33, 2.26, 3.19]	[0.33, 0.26, 0.20]	0.0365
2	0.3463	0.8337	[1.18, 2.01, 2.85]	[0.18, 0.01, -0.14]	0.0087

(2 Marks for 2nd iteration)
(2 Marks for third iteration)

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

Q6: **[10 Marks]**

The Pakistan Meteorological Department (PMD) is analyzing whether mean monsoon rainfall (July-September, in mm) differs significantly across four key regions: Northern Highlands (Gilgit, Murree), Pothohar

Plateau (Islamabad, Rawalpindi), Punjab Plains (Lahore, Faisalabad) and Sindh (Karachi, Hyderabad). They collected rainfall data from 5 stations per region during the 2023 monsoon:

Region	Monsoon Rainfall (mm)
Northern Highlands	450, 480, 420, 510, 440
Pothohar Plateau	320, 350, 310, 340, 330
Punjab Plains	280, 300, 270, 290, 310
Sindh	150, 170, 140, 160, 180

Construct ANOVA Table and test the mean monsoon rainfall in all regions are same or not at 5% level of significance. (Critical value= 3.24)

Solution:

Hypotheses (1 mark)

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{(All regions have equal mean rainfall)}$$

$$H_a : \text{At least one region has different mean rainfall}$$

Data Summary

Region	Sum T_j	Observations n_j	Mean $\bar{X}_j$
Northern Highlands	2300	5	460
Pothohar Plateau	1650	5	330
Punjab Plains	1450	5	290
Sindh	800	5	160

Total $N = 20$, Total sum $= 6200$, Overall mean $\bar{X} = 310$

Step 1: Calculate Sum of Squares (6 marks)

- Total sum of squares (SST):

$$SST = \sum x_i^2 - \frac{(\sum x_i)^2}{N} = 2,159,000 - \frac{6200^2}{20} = 237,000$$

- Treatment sum of squares (SSTR):

$$SSTR = \sum \frac{T_j^2}{n_j} - \frac{(\sum x_i)^2}{N} = \frac{2300^2}{5} + \frac{1650^2}{5} + \frac{1450^2}{5} + \frac{800^2}{5} - \frac{6200^2}{20} = 229,000$$

- Error sum of squares (SSE):

$$SSE = SST - SSTR = 237,000 - 229,000 = 8,000$$

Step 2: Degrees of Freedom (1 mark)

Source	df
Treatment	$k - 1 = 3$
Error	$N - k = 16$
Total	$N - 1 = 19$

Step 3: Calculate Mean Squares (1 mark)

$$MS_{Treatment} = \frac{SSTR}{df_{treatment}} = \frac{229,000}{3} = 76,333.33$$

$$MS_{Error} = \frac{SSE}{df_{error}} = \frac{8,000}{16} = 500$$

Step 4: Calculate F-statistic and Conclusion (2 marks)

$$F = \frac{MS_{Treatment}}{MS_{Error}} = \frac{76,333.33}{500} = 152.67$$

Critical value at 5% level, df1=3, df2=16 = 3.24

Since $F = 152.67 > 3.24$, **reject** H_0.

Conclusion: There is a significant difference in mean monsoon rainfall among the four regions.

ANOVA Table

Source	SS	df	MS	F	Critical F (5%)
Treatment	229,000	3	76,333.33	152.67	3.24
Error	8,000	16	500		
Total	237,000	19			