

Statistical Modelling for Machine Learning (313307)

Model Question Paper — Set A · Model Answers

Q.1

a) Q_1 and Q_3 of raw data.

2 M

Ascending order ($n = 11$): 15, 17, 19, 23, 25, 27, 28, 31, 35, 38, 42

$Q_1 = 1(n + 1)/4$ th value = $1 \times 12 / 4 = 3$ th value = **19**

$Q_3 = 3(n + 1)/4$ th value = $3 \times 12 / 4 = 9$ th value = **35**

(Arrangement $\frac{1}{2}$ M, Q_1 $\frac{1}{2}$ M, Q_3 1 M)

b) Harmonic mean.

2 M

x	1/x
4	0.25000
6	0.16667
8	0.12500
12	0.08333
24	0.04167
Total	0.66667

H.M. = $n / \Sigma(1/x) = 5 / 0.66667 = 7.5$

(Reciprocals 1 M, H.M. 1 M)

c) Karl Pearson's coefficient of skewness.

2 M

$Sk_p = (\text{Mean} - \text{Mode}) / \sigma = (45 - 41.5) / 7 = 3.5 / 7 = 0.5$

The distribution is positively skewed.

(Formula 1 M, answer 1 M)

d) Correlation coefficient from sums.

2 M

$$\begin{aligned}
 r &= (n\Sigma xy - \Sigma x \cdot \Sigma y) / \sqrt{[(n\Sigma x^2 - (\Sigma x)^2)(n\Sigma y^2 - (\Sigma y)^2)]} \\
 &= (10 \times 430 - 50 \times 80) / \sqrt{[(10 \times 290 - 50^2)(10 \times 700 - 80^2)]} \\
 &= 300 / \sqrt{(400 \times 600)} = 300 / 489.8979 = \mathbf{0.6124}
 \end{aligned}$$

There is a moderate degree of positive correlation.

(Formula 1 M, answer 1 M)

e) Probability — die thrown twice.

2 M

$n(S) = 6 \times 6 = 36$. A: sum is 9 = $\{(3,6), (4,5), (5,4), (6,3)\} \Rightarrow n(A) = 4$. B: doublet = $\{(1,1), (2,2), \dots, (6,6)\} \Rightarrow n(B) = 6$. $A \cap B = \emptyset$ (sum of a doublet is even).

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 4/36 + 6/36 - 0 = 10/36 = \mathbf{5/18}$

(Sample space $\frac{1}{2}$ M, events 1 M, answer $\frac{1}{2}$ M)

f) Forward difference table of a cubic.

2 M

$$f(x) = 2x^3 - x^2 + 3 \Rightarrow f(0) \dots f(4) = 3, 4, 15, 48, 115$$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
0	3				
		1			
1	4		10		
		11		12	
2	15		22		0
		33		12	
3	48		34		
		67			
4	115				

The third differences are all equal to **12**, i.e. constant ($= 2 \times 3! \times h^3 = 12$ for $h = 1$), as expected for a polynomial of degree 3.

(Table 1½ M, conclusion ½ M)

g) Null and alternative hypothesis.

2 M

- **Null hypothesis (H_0):** a definite statement about a population parameter which is tested for possible rejection, assuming it to be true. It usually says there is *no significant difference*, e.g. $H_0: \mu = 50$ (sample has come from a population with mean 50).
- **Alternative hypothesis (H_1):** the statement which is accepted when H_0 is rejected, e.g. $H_1: \mu \neq 50$ (two-tailed) or $\mu > 50$ / $\mu < 50$ (one-tailed).

(Each definition with example 1 M)

Q.2

a) Median of grouped data.

4 M

Class	Frequency f	c.f. (less than)
0–10	5	5
10–20	9	14
20–30	14	28
30–40	20	48
40–50	16	64
50–60	10	74
60–70	6	80
Total	N = 80	

$$N/2 = 80/2 = 40$$

Median class: the c.f. just greater than or equal to 40 is 48 $\Rightarrow$ Median class = **30–40**

$$L = 30, cf = 28 \text{ (c.f. of the class before), } f = 20, h = 10$$

$$\text{Median} = L + ((N/2 - cf) / f) \times h = 30 + ((40 - 28) / 20) \times 10 = \mathbf{36}$$

(c.f. table 1 M, median class 1 M, formula and answer 2 M)

b) Mode — graphical method and formula.

4 M

Graphical method: Draw a histogram. In the highest bar (modal class) join the top-left corner of the modal bar to the top-left corner of the next bar, and the top-right corner of the modal bar to the top-right corner of the previous bar. From the point where these two lines cross, draw a perpendicular to the x-axis; its foot gives the mode.

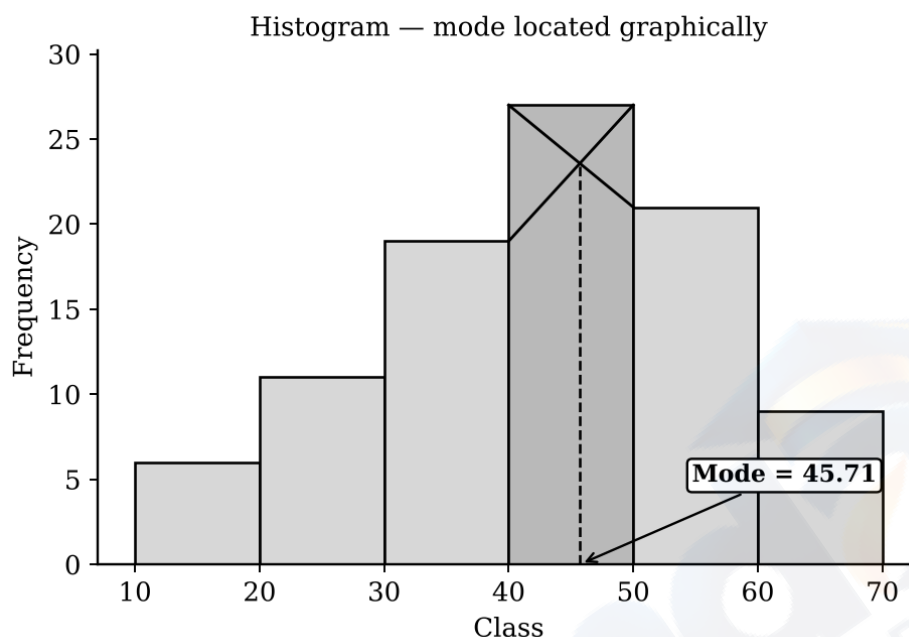


Fig.: Histogram with mode located graphically

Verification by formula:

Maximum frequency = 27 $\Rightarrow$ Modal class = **40–50**

$L = 40$, $f_1 = 27$ (modal class), $f_0 = 19$ (class before), $f_2 = 21$ (class after), $h = 10$

Mode = $L + ((f_1 - f_0) / (2f_1 - f_0 - f_2)) \times h = 40 + ((27 - 19) / (2 \times 27 - 19 - 21)) \times 10 = 40 + (8 / 14) \times 10 = \mathbf{45.71}$

The graphical value agrees with the formula value: **Mode = 45.71.**

(Histogram 2 M, formula verification 2 M)

c) Bowley's coefficient of skewness.

4 M

Class	Frequency f	c.f. (less than)
0–10	5	5
10–20	12	17
20–30	20	37
30–40	18	55
40–50	10	65
50–60	5	70
Total	N = 70	

$Q_1: 1N/4 = 1 \times 70/4 = 17.5$

Q1 class: the c.f. just greater than or equal to 17.5 is 37 $\Rightarrow$ Q1 class = **20–30**

$L = 20$, $cf = 17$ (c.f. of the class before), $f = 20$, $h = 10$

$Q_1 = L + ((N/4 - cf) / f) \times h = 20 + ((17.5 - 17) / 20) \times 10 = \mathbf{20.25}$

Median (Q_2): $N/2 = 70/2 = 35$

Median class: the c.f. just greater than or equal to 35 is 37 $\Rightarrow$ Median class = **20–30**

$L = 20$, $cf = 17$ (c.f. of the class before), $f = 20$, $h = 10$

Median = $L + ((N/2 - cf) / f) \times h = 20 + ((35 - 17) / 20) \times 10 = \mathbf{29}$

Q_3 : $3N/4 = 3 \times 70/4 = 52.5$

Q_3 class: the c.f. just greater than or equal to 52.5 is 55 $\Rightarrow Q_3$ class = **30–40**

$L = 30$, $cf = 37$ (c.f. of the class before), $f = 18$, $h = 10$

$Q_3 = L + ((3N/4 - cf) / f) \times h = 30 + ((52.5 - 37) / 18) \times 10 = \mathbf{38.61}$

$Sk_B = (Q_3 + Q_1 - 2 \cdot \text{Median}) / (Q_3 - Q_1) = (38.61 + 20.25 - 2 \times 29) / (38.61 - 20.25) = 0.8611 / 18.3611 = \mathbf{0.0469}$

Skewness is positive, i.e. the distribution is positively skewed (longer tail to the right; Mean > Median > Mode).

(Q_1 , Q_2 , Q_3 3 M, Sk_B 1 M)

d) Mode and median from Sk , mean and σ .

4 M

$Sk_p = (\text{Mean} - \text{Mode}) / \sigma \Rightarrow -0.4 = (60 - \text{Mode}) / 15 \Rightarrow 60 - \text{Mode} = -6 \Rightarrow \mathbf{\text{Mode} = 66}$

Empirical relation (moderately skewed distribution): $\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$

$66 = 3 \text{ Median} - 2 \times 60 \Rightarrow 3 \text{ Median} = 66 + 120 = 186 \Rightarrow \mathbf{\text{Median} = 62}$

Check: Mean (60) < Median (62) < Mode (66) — consistent with negative skewness.

(Mode 2 M, median 2 M)

Q.3

a) Karl Pearson's coefficient of skewness (grouped data).

4 M

Class	Mid-value x	f	$d = (x - 37.5)/5$	fd	fd^2
20–25	22.5	6	–3	–18	54
25–30	27.5	14	–2	–28	56
30–35	32.5	22	–1	–22	22
35–40	37.5	30	0	0	0
40–45	42.5	16	1	16	16
45–50	47.5	8	2	16	32
50–55	52.5	4	3	12	36
Total		$N = 100$		$\Sigma fd = -24$	$\Sigma fd^2 = 216$

Mean $\bar{x} = A + (\Sigma fd / N) \times h = 37.5 + (-24 / 100) \times 5 = \mathbf{36.3}$

$\sigma = h \times \sqrt{(\Sigma fd^2 / N - (\Sigma fd / N)^2)} = 5 \times \sqrt{(216 / 100 - (-24 / 100)^2)} = 5 \times \sqrt{(2.1600 - 0.0576)} = 5 \times 1.4500 = \mathbf{7.25}$

Mode: Maximum frequency = 30 $\Rightarrow$ Modal class = **35–40**

$L = 35$, $f_1 = 30$ (modal class), $f_0 = 22$ (class before), $f_2 = 16$ (class after), $h = 5$

Mode = $L + ((f_1 - f_0) / (2f_1 - f_0 - f_2)) \times h = 35 + ((30 - 22) / (2 \times 30 - 22 - 16)) \times 5 = 35 + (8 / 22) \times 5 = \mathbf{36.82}$

$Sk_p = (\text{Mean} - \text{Mode}) / \sigma = (36.3 - 36.82) / 7.25 = \mathbf{-0.0715}$

Skewness is negative, i.e. the distribution is negatively skewed (longer tail to the left; Mean < Median < Mode).

(Mean 1 M, S.D. 1 M, mode 1 M, Sk_p 1 M)

b) Central moments, β_1 , β_2 .

4 M

Given moments about A = 4: $\mu_1' = 1$, $\mu_2' = 6$, $\mu_3' = 10$, $\mu_4' = 120$. Mean = $A + \mu_1' = 4 + 1 = 5$.

Central moments (moments about the mean), using $\mu_1' = 1$, $\mu_2' = 6$, $\mu_3' = 10$, $\mu_4' = 120$ (about $x = 4$):

- $\mu_1 = 0$ (always)
- $\mu_2 = \mu_2' - (\mu_1')^2 = 6 - (1)^2 = \mathbf{5}$
- $\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = 10 - 3(6)(1) + 2(1)^3 = \mathbf{-6}$
- $\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4 = 120 - 4(10)(1) + 6(6)(1)^2 - 3(1)^4 = \mathbf{113}$

$\beta_1 = \mu_3^2 / \mu_2^3 = (-6)^2 / (5)^3 = \mathbf{0.2880}$ $\beta_2 = \mu_4 / \mu_2^2 = 113 / (5)^2 = \mathbf{4.5200}$

Comment: The distribution is negatively skewed ($\mu_3 < 0$) and leptokurtic ($\beta_2 > 3$, more peaked than normal).

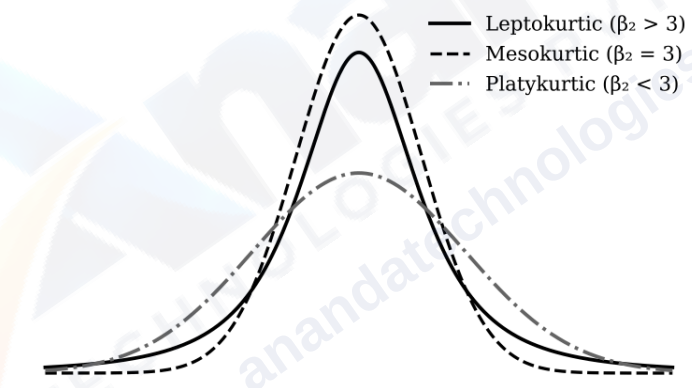


Fig.: Types of kurtosis

(μ_2 , μ_3 , μ_4 2 M, β_1 , β_2 1 M, comment 1 M)

c) Karl Pearson's coefficient of correlation.

4 M

x	y	dx = x - $\bar{x}$	dy = y - $\bar{y}$	dx ²	dy ²	dx·dy
4	12	-5	-6	25	36	30
6	15	-3	-3	9	9	9
8	14	-1	-4	1	16	4
10	20	1	2	1	4	2
12	24	3	6	9	36	18
14	23	5	5	25	25	25
$\Sigma x = 54$	$\Sigma y = 108$			70	126	88

$$\bar{x} = \Sigma x/n = 9, \quad \bar{y} = \Sigma y/n = 18$$

$$r = \Sigma dx \cdot dy / \sqrt{(\Sigma dx^2 \cdot \Sigma dy^2)} = 88 / \sqrt{(70 \times 126)} = 88 / 93.9149 = \mathbf{0.9370}$$

Type: $r = 0.9370 \Rightarrow$ high degree of positive correlation — as x increases, y also tends to increase.

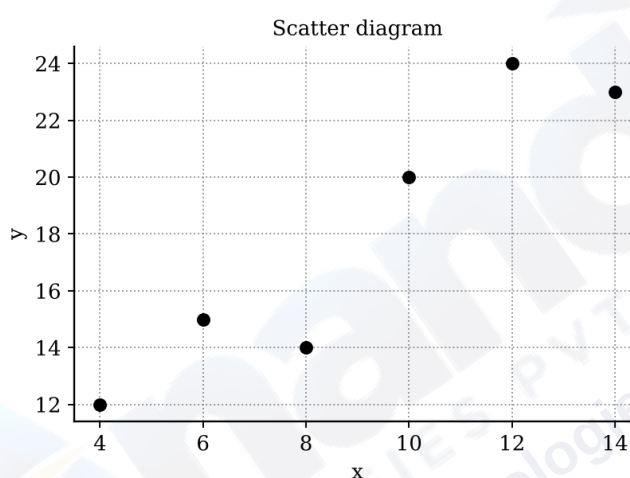


Fig.: Scatter diagram (points rise from left to right)

(Table 2 M, r 1 M, type 1 M)

d) Fitting a straight line by least squares.

4 M

x	y	x ²	xy
1	3	1	3
2	7	4	14
3	8	9	24
4	12	16	48
5	14	25	70
6	17	36	102
$\Sigma x = 21$	$\Sigma y = 61$	$\Sigma x^2 = 91$	$\Sigma xy = 261$

Normal equations for $y = a + bx$:

$$\Sigma y = na + b\Sigma x \quad \dots(1)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \dots(2)$$

Substituting: $61 = 6a + 21b \dots(1)$ $261 = 21a + 91b \dots(2)$

Solving (1) and (2) simultaneously: $b = (n\sum xy - \sum x \sum y) / (n\sum x^2 - (\sum x)^2) = (6 \times 261 - 21 \times 61) / (6 \times 91 - 21^2) = 285 / 105 = 2.7143$

$a = (\sum y - b \sum x) / n = (61 - 2.7143 \times 21) / 6 = 0.6667$

Required line: $y = 0.6667 + 2.7143x$

At $x = 7$: $y = 0.6667 + 2.7143 \times 7 = 19.67$

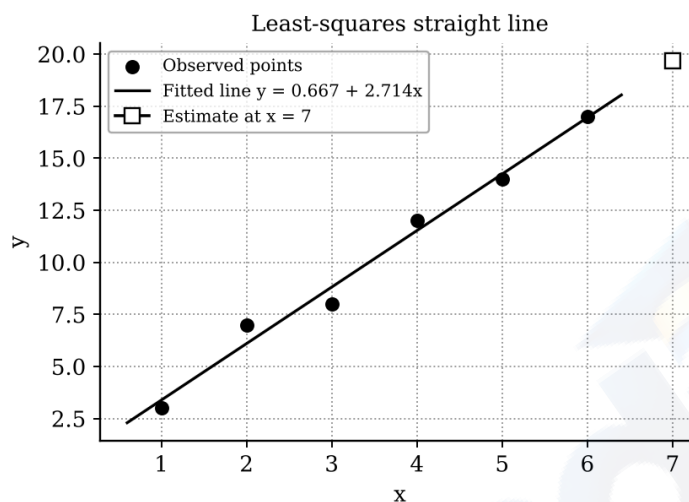


Fig.: Observed points and the fitted line

(Table 1 M, normal equations 1 M, a and b 1 M, estimate 1 M)

Q.4

a) Spearman's rank correlation.

4 M

Ranks are given in descending order (highest mark = rank 1).

x	y	Rank R_1 (x)	Rank R_2 (y)	$d = R_1 - R_2$	d^2
45	52	7	6	1	1
62	60	4	4	0	0
38	41	8	8	0	0
70	68	2	3	-1	1
55	49	5	7	-2	4
49	58	6	5	1	1
81	75	1	1	0	0
66	72	3	2	1	1
				$\sum d = 0$	$\sum d^2 = 8$

$\rho = 1 - 6\sum d^2 / (n(n^2 - 1)) = 1 - (6 \times 8) / (8 \times (8^2 - 1)) = 1 - 48/504 = 0.9048$

There is a high degree of positive correlation between the marks in the two subjects.

(Ranks 1 M, d^2 1 M, formula and ρ 2 M)

b) Regression lines.

4 M

x	y	dx = x - $\bar{x}$	dy = y - $\bar{y}$	dx ²	dy ²	dx·dy
1	15	-4	5	16	25	-20
3	12	-2	2	4	4	-4
5	10	0	0	0	0	0
7	8	2	-2	4	4	-4
9	5	4	-5	16	25	-20
$\Sigma x = 25$	$\Sigma y = 50$			40	58	-48

$$\bar{x} = 5, \bar{y} = 10$$

$$b_{yx} = \Sigma dx \cdot dy / \Sigma dx^2 = -48 / 40 = -1.2 \quad b_{xy} = \Sigma dx \cdot dy / \Sigma dy^2 = -48 / 58 = -0.8276$$

$$\text{Regression line of y on x: } y - \bar{y} = b_{yx}(x - \bar{x}) \Rightarrow y - 10 = -1.2(x - 5) \Rightarrow \boxed{y = -1.2x + 16}$$

$$\text{Regression line of x on y: } x - \bar{x} = b_{xy}(y - \bar{y}) \Rightarrow x - 5 = -0.8276(y - 10) \Rightarrow$$

$$\boxed{x = -0.8276y + 13.2759}$$

Estimate: at $x = 6$, $y = -1.2 \times 6 + 16 = 8.8$

Check: $r = -\sqrt{(b_{yx} \cdot b_{xy})} = -\sqrt{(-1.2 \times -0.8276)} = -0.9965$ (both coefficients negative).

(Table 1 M, each line 1½ M)

c) Probability — playing cards (addition theorem).

4 M

$n(S) = 52$. Addition theorem: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

(i) A = king (4 cards), B = heart (13 cards), $A \cap B$ = king of hearts (1 card).

$$P(\text{king or heart}) = 4/52 + 13/52 - 1/52 = 16/52 = \mathbf{4/13}$$

(ii) A = red card (26), B = face card (J, Q, K: 12), $A \cap B$ = red face cards (6).

$$P(\text{red or face}) = 26/52 + 12/52 - 6/52 = 32/52 = \mathbf{8/13}$$

(Each part 2 M)

d) Drawing without replacement.

4 M

Total balls = 5 + 8 = 13. Since there is no replacement, the second draw is from 12 balls (multiplication theorem $P(A \cap B) = P(A) \cdot P(B|A)$).

$$(i) P(\text{both red}) = P(R_1) \cdot P(R_2 | R_1) = (8/13) \times (7/12) = 56/156 = \mathbf{14/39} = 0.3590$$

$$(ii) P(\text{one white, one red}) = P(W_1 R_2) + P(R_1 W_2) = (5/13)(8/12) + (8/13)(5/12) = 40/156 + 40/156 = \mathbf{20/39} = 0.5128$$

$$(\text{Same by combinations: (ii)} = {}^5C_1 \cdot {}^8C_1 / {}^{13}C_2 = 40/78 = 20/39.)$$

(Each part 2 M)

e) Bayes' theorem — machines.

4 M

Let E_1, E_2, E_3 = item produced by M_1, M_2, M_3 and D = item is defective.

$$P(E_1) = 0.30, P(E_2) = 0.45, P(E_3) = 0.25; \quad P(D|E_1) = 0.03, P(D|E_2) = 0.02, P(D|E_3) = 0.04$$

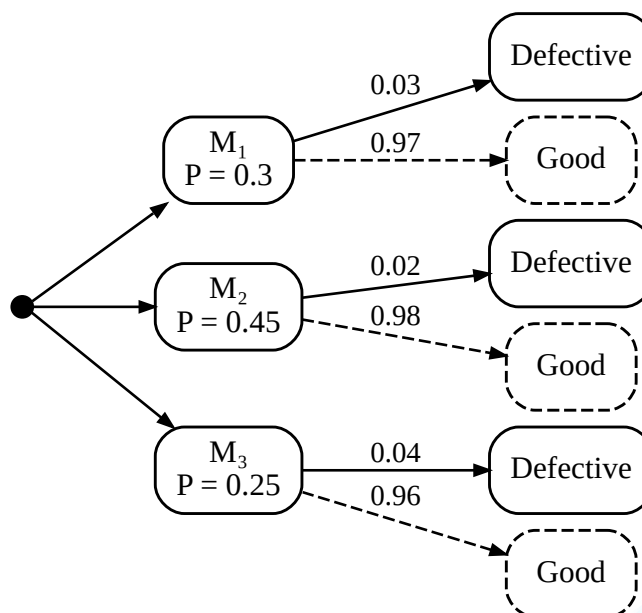


Fig.: Probability tree

$$P(D) = \sum P(E_i)P(D|E_i) = (0.30)(0.03) + (0.45)(0.02) + (0.25)(0.04) = 0.009 + 0.009 + 0.01 = 0.028$$

By Bayes' theorem: $P(E_3|D) = P(E_3)P(D|E_3) / P(D) = 0.01 / 0.028 = 5/14 = 0.3571$

(Notation 1 M, total probability 1 M, Bayes 2 M)

Q.5

a) Lagrange's interpolation.

6 M

(i)

Lagrange's interpolation formula:

$$y = \frac{(x - x_1)(x - x_2)(x - x_3)}{[(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)]} \cdot y_0 + \frac{(x - x_0)(x - x_2)(x - x_3)}{[(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)]} \cdot y_1 + \frac{(x - x_0)(x - x_1)(x - x_3)}{[(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)]} \cdot y_2 + \frac{(x - x_0)(x - x_1)(x - x_2)}{[(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)]} \cdot y_3$$

Here $x_0, x_1, \dots = 1, 3, 4, 6$ and $y_0, y_1, \dots = 5, 13, 22, 50$. Putting $x = 5$:

$$\begin{aligned} y(5) &= \frac{(5 - 3)(5 - 4)(5 - 6)}{[(1 - 3)(1 - 4)(1 - 6)]} \times 5 \\ &+ \frac{(5 - 1)(5 - 4)(5 - 6)}{[(3 - 1)(3 - 4)(3 - 6)]} \times 13 \\ &+ \frac{(5 - 1)(5 - 3)(5 - 6)}{[(4 - 1)(4 - 3)(4 - 6)]} \times 22 \\ &+ \frac{(5 - 1)(5 - 3)(5 - 4)}{[(6 - 1)(6 - 3)(6 - 4)]} \times 50 \\ &= (-2/-30) \times 5 + (-4/6) \times 13 + (-8/-6) \times 22 + (8/30) \times 50 \\ &= 0.3333 - 8.6667 + 29.3333 + 13.3333 \end{aligned}$$

$$\Rightarrow f(5) = 34.3333$$

(ii) $x_0 = 0, x_1 = 1, x_2 = 3; y_0 = 2, y_1 = 3, y_2 = 11$

$$f(x) = \frac{(x - 1)(x - 3)}{[(0 - 1)(0 - 3)]} \times 2 + \frac{(x - 0)(x - 3)}{[(1 - 0)(1 - 3)]} \times 3 + \frac{(x - 0)(x - 1)}{[(3 - 0)(3 - 1)]} \times 11$$

$$= (2/3)(x^2 - 4x + 3) - (3/2)(x^2 - 3x) + (11/6)(x^2 - x)$$

$$= x^2(2/3 - 3/2 + 11/6) + x(-8/3 + 9/2 - 11/6) + 2 = x^2(4 - 9 + 11)/6 + x(-16 + 27 - 11)/6 + 2$$

$$\Rightarrow \boxed{f(x) = x^2 + 2} \quad f(2) = 2^2 + 2 = 6$$

((i) 3 M, (ii) polynomial 2 M, f(2) 1 M)

b) Newton's forward and backward interpolation.

6 M

Forward difference table (leading diagonal in bold):

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
1971	40				
		8			
1981	48		3		
		11		-1	
1991	59		2		2
		13		1	
2001	72		3		
		16			
2011	88				

(i)

Here $x_0 = 1971$, $h = 10$, $x = 1976 \Rightarrow p = (x - x_0)/h = (1976 - 1971)/10 = 0.5$

Newton's forward interpolation formula:

$$y = y_0 + p\Delta y_0 + \frac{p(p-1)}{2!} \Delta^2 y_0 + \frac{p(p-1)(p-2)}{3!} \Delta^3 y_0 + \frac{p(p-1)(p-2)(p-3)}{4!} \Delta^4 y_0$$

$$y(1976) = 40 + (0.5)(8) + (0.5)(-0.5)(3)/2 + (0.5)(-0.5)(-1.5)(-1)/6 + (0.5)(-0.5)(-1.5)(-2.5)(2)/24$$

$$= 40 + 4 - 0.375 - 0.0625 - 0.07812$$

$\Rightarrow$ Population in 1976 $\approx$ **43.484 thousand** $\approx$ 43,484 persons

(ii) Backward differences are the same numbers; the bottom diagonal ($\nabla y_n, \nabla^2 y_n \dots$) is used:

x	y	∇y	$\nabla^2 y$	$\nabla^3 y$	$\nabla^4 y$
1971	40				
		8			
1981	48		3		
		11		-1	
1991	59		2		2
		13		1	
2001	72		3		
		16			
2011	88				

Here $x_n = 2011$, $h = 10$, $x = 2006 \Rightarrow p = (x - x_n)/h = (2006 - 2011)/10 = -0.5$

Newton's backward interpolation formula:

$$y = y_n + p\nabla y_n + \frac{p(p+1)}{2!} \nabla^2 y_n + \frac{p(p+1)(p+2)}{3!} \nabla^3 y_n + \frac{p(p+1)(p+2)(p+3)}{4!} \nabla^4 y_n$$

$$y(2006) = 88 + (-0.5)(16) + (-0.5)(0.5)(3)/2 + (-0.5)(0.5)(1.5)(1)/6 + (-0.5)(0.5)(1.5)(2.5)(2)/24$$

$$= 88 - 8 - 0.375 - 0.0625 - 0.07812$$

$\Rightarrow$ Population in 2006 $\approx$ **79.484 thousand** $\approx$ 79,484 persons

(Difference table 2 M, forward 2 M, backward 2 M)

c) Extrapolation; relations between operators.

6 M

(i) Difference table:

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
1	2				
		7			
2	9		12		
		19		6	
3	28		18		0
		37		6	
4	65		24		
		61			
5	126				

Third differences are constant (6) and fourth differences are zero, so $f(x)$ is a polynomial of degree 3 and the table can be extended (extrapolated) beyond $x = 5$.

Here $x_0 = 1$, $h = 1$, $x = 7 \Rightarrow p = (x - x_0)/h = (7 - 1)/1 = 6$

Newton's forward interpolation formula:

$$y = y_0 + p\Delta y_0 + \frac{p(p-1)}{2!} \Delta^2 y_0 + \frac{p(p-1)(p-2)}{3!} \Delta^3 y_0$$

$$y(7) = 2 + (6)(7) + \frac{(6)(5)(12)}{2} + \frac{(6)(5)(4)(6)}{6}$$

$$= 2 + 42 + 180 + 120$$

$$\Rightarrow f(7) = 344$$

(Check: the data fit $f(x) = x^3 + 1$; $f(7) = 343 + 1 = 344$.)

(ii)(a) $E = 1 + \Delta$: By definition $\Delta f(x) = f(x + h) - f(x)$ and $Ef(x) = f(x + h)$.

$$\therefore \Delta f(x) = Ef(x) - f(x) = (E - 1)f(x) \Rightarrow \Delta = E - 1 \Rightarrow E = 1 + \Delta.$$

(b) $\nabla = 1 - E^{-1}$: $\nabla f(x) = f(x) - f(x - h)$ and $E^{-1}f(x) = f(x - h)$.

$$\therefore \nabla f(x) = f(x) - E^{-1}f(x) = (1 - E^{-1})f(x) \Rightarrow \nabla = 1 - E^{-1}.$$

((i) table 1 M, extrapolation 2 M; (ii) each proof 1½ M)

Q.6

a) Chi-square test — uniform distribution.

6 M

H_0 : the books are issued uniformly over the six days (no difference between days).

H_1 : the issue of books is not uniform.

Total = 480; under H_0 the expected frequency of each day $E = 480/6 = 80$.

Day	O	E	O - E	(O - E) ²	(O - E) ² /E
Mon	72	80	-8	64	0.8000
Tue	84	80	4	16	0.2000
Wed	66	80	-14	196	2.4500
Thu	90	80	10	100	1.2500
Fri	78	80	-2	4	0.0500
Sat	90	80	10	100	1.2500
Total	480	480			$\chi^2 = 6.0000$

Degrees of freedom = $n - 1 = 6 - 1 = 5$; table value $\chi^2_{0.05} = 11.07$.

Since calculated $\chi^2 = 6.000 < 11.07$, **H₀ is accepted**: the books are issued uniformly over the days.

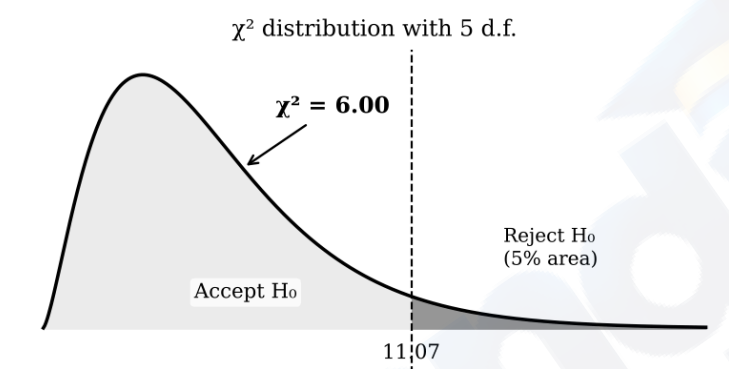


Fig.: Calculated χ^2 lies in the acceptance region

(Hypotheses 1 M, expected values 1 M, χ^2 table 2 M, d.f. and conclusion 2 M)

b) t-test for mean; test of proportion.

6 M

(i)

H₀: $\mu = 47$ (the sample has come from a population with mean 47).

H₁: $\mu \neq 47$ (two-tailed).

x	$x - \bar{x}$	$(x - \bar{x})^2$
48	-1.5	2.25
52	2.5	6.25
47	-2.5	6.25
55	5.5	30.25
51	1.5	2.25
46	-3.5	12.25
49	-0.5	0.25
53	3.5	12.25
44	-5.5	30.25
50	0.5	0.25
$\Sigma x = 495$		$\Sigma(x - \bar{x})^2 = 102.5$

$$\bar{x} = 49.5, \quad s = \sqrt{(\Sigma(x - \bar{x})^2 / (n - 1))} = \sqrt{(102.5 / 9)} = 3.3747$$

$$t = (\bar{x} - \mu) / (s / \sqrt{n}) = (49.5 - 47) / (3.3747 / \sqrt{10}) = 2.50 / 1.0672 = \mathbf{2.343}$$

d.f. = $n - 1 = 9$; $t_{0.05} = 2.262$. Since $|t| = 2.343 > 2.262$, **H_0 is rejected**: the sample cannot be regarded as coming from a population with mean 47.

(ii)

H_0 : the coin is unbiased, $P = 0.5$.

H_1 : $P \neq 0.5$.

$n = 600$, $x = 330$, sample proportion $p = 330/600 = 0.55$, $Q = 1 - P = 0.5$

$$z = (p - P) / \sqrt{(PQ/n)} = (0.55 - 0.5) / \sqrt{(0.5 \times 0.5 / 600)} = 0.05 / 0.02041 = \mathbf{2.449}$$

Since $|z| = 2.449 > 1.96$ (table value), the difference is **significant**. **H_0 is rejected**. The coin is **biased** (at 5% level).

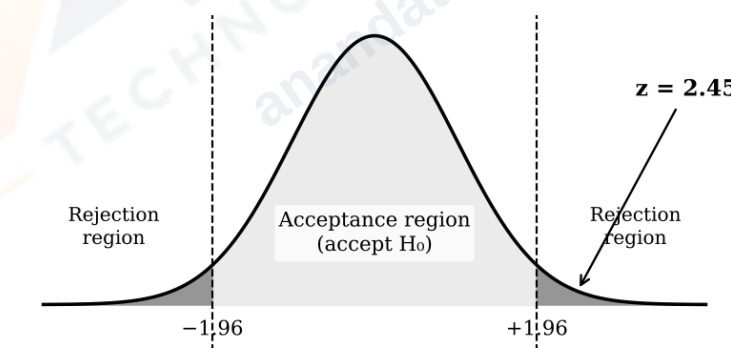


Fig.: z lies in the rejection region (two-tailed test, 5% level)

((i) 3 M, (ii) 3 M)

c) Chi-square test of independence (2×2).

6 M

H_0 : vaccination and attack of the disease are independent (vaccine is not effective).

H_1 : they are associated.

Cell	O	E = (Row total × Column total)/N	O – E	(O – E) ²	(O – E) ² /E
Vaccinated – Attacked	25	$200 \times 70 / 400 = 35$	-10	100	2.8571
Vaccinated – Not attacked	175	$200 \times 330 / 400 = 165$	10	100	0.6061
Not vacc. – Attacked	45	$200 \times 70 / 400 = 35$	10	100	2.8571
Not vacc. – Not attacked	155	$200 \times 330 / 400 = 165$	-10	100	0.6061
				$\chi^2 = \Sigma$	6.9264

Short-cut for a 2 × 2 table: $\chi^2 = N(ad - bc)^2 / [(a + b)(c + d)(a + c)(b + d)] = 400(25 \times 155 - 175 \times 45)^2 / (200 \times 200 \times 70 \times 330) = 6.9264$ (same).

d.f. = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1; table value = 3.84.

Since calculated $\chi^2 = 6.926 > 3.84$, **H₀ is rejected**: vaccination and attack are **not independent** — the vaccine is effective in reducing attacks.

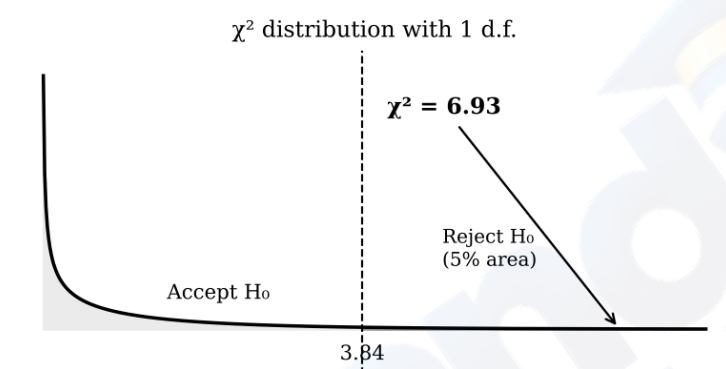


Fig.: Calculated χ^2 compared with the 5% table value

(Hypotheses 1 M, expected frequencies 2 M, χ^2 2 M, conclusion 1 M)

Check your answers with Python (optional)

The numerical answers above were computed by program and cross-checked with the standard **numpy** / **scipy** functions. Students can run the same check (this is not part of the exam answer):

```
# Set A: verify key numerical answers with numpy / scipy
import numpy as np
from scipy import stats

# Q1 a) quartiles by (n+1)p positional method
data = [23, 15, 31, 27, 19, 42, 35, 28, 17, 38, 25]
print("Q1 a) Q1, Q3 =", np.percentile(data, [25, 75],
method='weibull').tolist())
print("Q1 b) H.M. =", round(stats.hmean([4, 6, 8, 12, 24]), 4))

# Q3 c) Karl Pearson's r
x = [4, 6, 8, 10, 12, 14]; y = [12, 15, 14, 20, 24, 23]
print("Q3 c) r =", round(stats.pearsonr(x, y).statistic, 4))

# Q3 d) least-squares line y = a + bx
b, a = np.polyfit([1, 2, 3, 4, 5, 6], [3, 7, 8, 12, 14, 17], 1)
print("Q3 d) a =", round(a, 4), " b =", round(b, 4), " y(7) =", round(a + b * 7,
2))
```

```
# Q4 a) Spearman's rank correlation
print("Q4 a) rho =", round(stats.spearmanr([45, 62, 38, 70, 55, 49, 81, 66],
                                           [52, 60, 41, 68, 49, 58, 75,
                                           72])).statistic, 4))
# Q4 b) regression lines
r1 = stats.linregress([1, 3, 5, 7, 9], [15, 12, 10, 8, 5])
r2 = stats.linregress([15, 12, 10, 8, 5], [1, 3, 5, 7, 9])
print("Q4 b) y on x: y = %.4f x %+.4f ; x on y: x = %.4f y %+.4f" % (r1.slope,
r1.intercept, r2.slope, r2.intercept))

# Q5 interpolation: polynomial through the points (same value as Lagrange /
Newton)
def interp(xs, ys, x0):
    c = np.polyfit(xs, ys, len(xs) - 1)
    return np.polyval(c, x0)
print("Q5 a) f(5) =", round(interp([1, 3, 4, 6], [5, 13, 22, 50], 5), 4))
yrs = np.array([1971, 1981, 1991, 2001, 2011]); pop = [40, 48, 59, 72, 88]
u = (yrs - 1971) / 10 # scale x to avoid large powers
print("Q5 b) 1976:", round(interp(u, pop, 0.5), 4), " 2006:", round(interp(u,
pop, 3.5), 4))
print("Q5 c) f(7) =", round(interp([1, 2, 3, 4, 5], [2, 9, 28, 65, 126], 7), 4))

# Q6 a) chi-square goodness of fit (uniform)
r = stats.chisquare([72, 84, 66, 90, 78, 90])
print("Q6 a) chi2 = %.4f, p-value = %.4f" % (r.statistic, r.pvalue))
# Q6 b) one-sample t-test and z for proportion
r = stats.ttest_1samp([48, 52, 47, 55, 51, 46, 49, 53, 44, 50], 47)
print("Q6 b)(i) t = %.4f, p-value = %.4f" % (r.statistic, r.pvalue))
print("Q6 b)(ii) z =", round((330 / 600 - 0.5) / np.sqrt(0.25 / 600), 4))
# Q6 c) chi-square test of independence (no Yates correction, as in MSBTE
answers)
res = stats.chi2_contingency([[25, 175], [45, 155]], correction=False)
print("Q6 c) chi2 = %.4f, dof = %d, p = %.4f" % (res.statistic, res.dof,
res.pvalue))
```

Output:

```
Q1 a) Q1, Q3 = [19.0, 35.0]
Q1 b) H.M. = 7.5
Q3 c) r = 0.937
Q3 d) a = 0.6667 b = 2.7143 y(7) = 19.67
Q4 a) rho = 0.9048
Q4 b) y on x: y = -1.2000 x +16.0000 ; x on y: x = -0.8276 y +13.2759
Q5 a) f(5) = 34.3333
Q5 b) 1976: 43.4844 2006: 79.4844
Q5 c) f(7) = 344.0
Q6 a) chi2 = 6.0000, p-value = 0.3062
Q6 b)(i) t = 2.3426, p-value = 0.0438
Q6 b)(ii) z = 2.4495
Q6 c) chi2 = 6.9264, dof = 1, p = 0.0085
```

Note: This is a **model question paper prepared by Ananda Technologies** only for practice and exam preparation, based on the pattern of previous MSBTE question papers and the K-scheme syllabus. It is **not an official MSBTE paper**, and it does **not** mean that these questions will come in the exam. Answers follow standard textbook points; marks shown are indicative. Free for students — anandatechnologies.in