

Statistical Modelling for Machine Learning (313307)

Model Question Paper — Set C · Model Answers

Q.1

a) Median and mode of raw data.

2 M

Arranging in ascending order: 11, 11, 12, 13, 15, 15, 15, 16, 18, 20

$n = 10$ (even) $\Rightarrow$ Median = average of $(n/2)$ th and $(n/2 + 1)$ th values = $(15 + 15) / 2 = 15$

Mode: 15 occurs 3 times (maximum) $\Rightarrow$ **Mode = 15**

(Median 1 M, mode 1 M)

b) Geometric mean.

2 M

x	$\log_{10} x$
3	0.4771
6	0.7782
24	1.3802
48	1.6812
Total	4.3167

G.M. = $\text{antilog}(\Sigma \log x / n) = \text{antilog}(4.3167 / 4) = \text{antilog}(1.0792) = 12.00$

(Directly: G.M. = $(3 \times 6 \times 24 \times 48)^{1/4} = (20736)^{1/4} = 12.$)

(Logs 1 M, G.M. 1 M)

c) Correlation from regression coefficients.

2 M

$r = \pm\sqrt{(b_{yx} \times b_{xy})} = \sqrt{(0.8 \times 0.45)} = \sqrt{0.36} = 0.6$

r takes the sign of the regression coefficients (both positive), so $r = +0.6$.

(Formula 1 M, answer 1 M)

d) Karl Pearson's coefficient of skewness.

2 M

Mode is not given, so the median form is used: $Sk_p = 3(\text{Mean} - \text{Median})/\sigma = 3(50 - 48)/12 = 6/12 = 0.5$ (positive skewness)

(Formula 1 M, answer 1 M)

e) Probability — mutually exclusive events.

2 M

Total = $4 + 6 + 5 = 15$. R and G are mutually exclusive $\Rightarrow P(R \cup G) = P(R) + P(G) = 4/15 + 5/15 = 9/15 = 3/5$

(Addition rule 1 M, answer 1 M)

f) Newton's forward interpolation.

2 M

x	y	Δy	$\Delta^2 y$
10	15		
		10	
20	25		6
		16	
30	41		

Here $x_0 = 10$, $h = 10$, $x = 14 \Rightarrow p = (x - x_0)/h = (14 - 10)/10 = 0.4$

Newton's forward interpolation formula:

$$y = y_0 + p\Delta y_0 + p(p-1)/2! \Delta^2 y_0$$

$$y(14) = 15 + (0.4)(10) + (0.4)(-0.6)(6)/2$$

$$= 15 + 4 - 0.72$$

$$\Rightarrow f(14) = 18.28$$

(Table $\frac{1}{2}$ M, p $\frac{1}{2}$ M, value 1 M)

g) Standard error.

2 M

The **standard error (S.E.)** of a statistic is the standard deviation of its sampling distribution. It measures the reliability of the statistic as an estimate of the population parameter — smaller S.E. means a more reliable estimate.

- S.E. of sample mean: $S.E.(\bar{x}) = \sigma/\sqrt{n}$
- S.E. of sample proportion: $S.E.(p) = \sqrt{(PQ/n)}$, $Q = 1 - P$

(Definition 1 M, formulae 1 M)

Q.2

a) Median — inclusive classes.

4 M

The classes are inclusive, so they are first converted into exclusive classes (class boundaries) by subtracting 0.5 from each lower limit and adding 0.5 to each upper limit.

Class	Class boundaries	Frequency f	c.f. (less than)
20–29	19.5–29.5	5	5
30–39	29.5–39.5	12	17
40–49	39.5–49.5	20	37
50–59	49.5–59.5	28	65
60–69	59.5–69.5	15	80
70–79	69.5–79.5	7	87
Total		N = 87	

$$N/2 = 87/2 = 43.5$$

Median class: the c.f. just greater than or equal to 43.5 is 65 $\Rightarrow$ Median class = **49.5–59.5**

$L = 49.5$, $cf = 37$ (c.f. of the class before), $f = 28$, $h = 10$

$$\text{Median} = L + ((N/2 - cf) / f) \times h = 49.5 + ((43.5 - 37) / 28) \times 10 = \mathbf{51.82}$$

(Boundaries 1 M, c.f. 1 M, median 2 M)

b) Mode by histogram and formula.

4 M

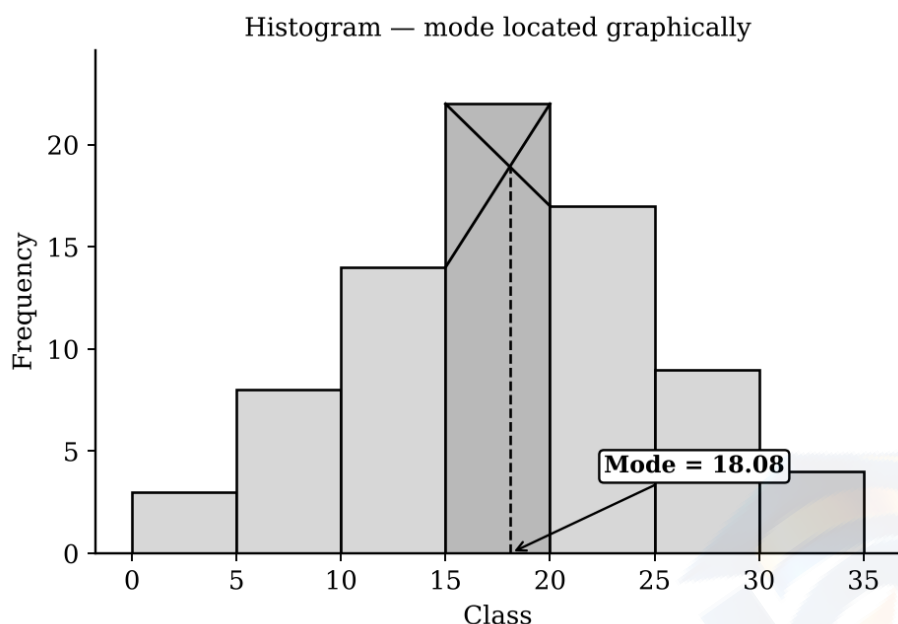


Fig.: Histogram — mode located graphically

Join the top corners of the modal (highest) bar crosswise to the adjacent corners of the neighbouring bars; the perpendicular from the point of intersection to the x-axis gives the mode.

Check by formula:

Maximum frequency = 22 $\Rightarrow$ Modal class = 15–20

$L = 15$, $f_1 = 22$ (modal class), $f_0 = 14$ (class before), $f_2 = 17$ (class after), $h = 5$

Mode = $L + ((f_1 - f_0) / (2f_1 - f_0 - f_2)) \times h = 15 + ((22 - 14) / (2 \times 22 - 14 - 17)) \times 5 = 15 + (8 / 13) \times 5 = 18.08$

(Histogram 2 M, formula 2 M)

c) Quartile deviation from a less-than table.

4 M

Convert the less-than cumulative frequencies into class frequencies (f = difference of successive c.f.):

Class	Frequency f	c.f. (less than)
0–10	6	6
10–20	12	18
20–30	22	40
30–40	30	70
40–50	18	88
50–60	12	100
Total	$N = 100$	

Q_1 : $1N/4 = 1 \times 100/4 = 25$

Q1 class: the c.f. just greater than or equal to 25 is 40 $\Rightarrow$ Q1 class = 20–30

$L = 20$, $cf = 18$ (c.f. of the class before), $f = 22$, $h = 10$

$$Q_1 = L + ((N/4 - cf) / f) \times h = 20 + ((25 - 18) / 22) \times 10 = \mathbf{23.18}$$

$$Q_3: 3N/4 = 3 \times 100/4 = 75$$

Q3 class: the c.f. just greater than or equal to 75 is 88 $\Rightarrow$ Q3 class = **40–50**

$L = 40$, $cf = 70$ (c.f. of the class before), $f = 18$, $h = 10$

$$Q_3 = L + ((3N/4 - cf) / f) \times h = 40 + ((75 - 70) / 18) \times 10 = \mathbf{42.78}$$

$$Q.D. = (Q_3 - Q_1)/2 = (42.78 - 23.18)/2 = \mathbf{9.8}$$

$$\text{Coefficient of Q.D.} = (Q_3 - Q_1)/(Q_3 + Q_1) = 19.596/65.9596 = \mathbf{0.2971}$$

(Frequencies 1 M, Q_1 and Q_3 2 M, Q.D. and coefficient 1 M)

d) Quartiles from Bowley's coefficient.

4 M

$$Sk_B = (Q_3 + Q_1 - 2 \cdot \text{Median}) / (Q_3 - Q_1)$$

$$0.6 = (130 - 2 \times 59) / (Q_3 - Q_1) = 12 / (Q_3 - Q_1) \Rightarrow Q_3 - Q_1 = 12/0.6 = 20$$

$$Q_3 + Q_1 = 130 \text{ and } Q_3 - Q_1 = 20 \Rightarrow 2Q_3 = 150 \Rightarrow \mathbf{Q_3 = 75, Q_1 = 55}$$

$$\text{Coefficient of Q.D.} = (Q_3 - Q_1) / (Q_3 + Q_1) = 20/130 = \mathbf{0.1538}$$

($Q_3 - Q_1$ 1 M, quartiles 2 M, coefficient 1 M)

Q.3

a) Karl Pearson's coefficient of skewness.

4 M

Class	Mid-value x	f	d = (x - 57.5)/5	fd	fd ²
40–45	42.5	5	-3	-15	45
45–50	47.5	12	-2	-24	48
50–55	52.5	18	-1	-18	18
55–60	57.5	25	0	0	0
60–65	62.5	20	1	20	20
65–70	67.5	13	2	26	52
70–75	72.5	7	3	21	63
Total		N = 100		$\Sigma fd = 10$	$\Sigma fd^2 = 246$

$$\text{Mean } \bar{x} = A + (\Sigma fd / N) \times h = 57.5 + (10 / 100) \times 5 = \mathbf{58}$$

$$\sigma = h \times \sqrt{(\Sigma fd^2 / N - (\Sigma fd / N)^2)} = 5 \times \sqrt{(246/100 - (10/100)^2)} = 5 \times \sqrt{(2.4600 - 0.0100)} = 5 \times 1.5652 = \mathbf{7.83}$$

Mode: Maximum frequency = 25 $\Rightarrow$ Modal class = **55–60**

$L = 55$, $f_1 = 25$ (modal class), $f_0 = 18$ (class before), $f_2 = 20$ (class after), $h = 5$

$$\text{Mode} = L + ((f_1 - f_0) / (2f_1 - f_0 - f_2)) \times h = 55 + ((25 - 18) / (2 \times 25 - 18 - 20)) \times 5 = 55 + (7 / 12) \times 5 = \mathbf{57.92}$$

$$Sk_p = (\text{Mean} - \text{Mode}) / \sigma = (58 - 57.92) / 7.83 = \mathbf{0.0106}$$

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

(Mean 1 M, σ 1 M, mode 1 M, Sk 1 M)

b) Bowley's coefficient of skewness.

4 M

Class	Frequency f	c.f. (less than)
0–20	8	8
20–40	15	23
40–60	27	50
60–80	25	75
80–100	15	90
100–120	10	100
Total	N = 100	

Q₁: $1N/4 = 1 \times 100/4 = 25$

Q1 class: the c.f. just greater than or equal to 25 is 50 $\Rightarrow$ Q1 class = **40–60**

L = 40, cf = 23 (c.f. of the class before), f = 27, h = 20

$Q_1 = L + ((N/4 - cf) / f) \times h = 40 + ((25 - 23) / 27) \times 20 = \mathbf{41.48}$

Median: $N/2 = 100/2 = 50$

Median class: the c.f. just greater than or equal to 50 is 50 $\Rightarrow$ Median class = **40–60**

L = 40, cf = 23 (c.f. of the class before), f = 27, h = 20

$\text{Median} = L + ((N/2 - cf) / f) \times h = 40 + ((50 - 23) / 27) \times 20 = \mathbf{60}$

Q₃: $3N/4 = 3 \times 100/4 = 75$

Q3 class: the c.f. just greater than or equal to 75 is 75 $\Rightarrow$ Q3 class = **60–80**

L = 60, cf = 50 (c.f. of the class before), f = 25, h = 20

$Q_3 = L + ((3N/4 - cf) / f) \times h = 60 + ((75 - 50) / 25) \times 20 = \mathbf{80}$

$Sk_B = (Q_3 + Q_1 - 2 \cdot \text{Median}) / (Q_3 - Q_1) = (80 + 41.48 - 2 \times 60) / (80 - 41.48) = \mathbf{0.0385}$

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

(Quartiles 3 M, Sk_B 1 M)

c) Central moments, β_1 , β_2 and kurtosis.

4 M

Taking A = 2 (d = x - 2), the raw moments about A are $\mu'_r = \Sigma f d^r / N$:

x	f	d = x - 2	fd	fd ²	fd ³	fd ⁴
0	5	-2	-10	20	-40	80
1	10	-1	-10	10	-10	10
2	20	0	0	0	0	0
3	10	1	10	10	10	10
4	5	2	10	20	40	80
Total	50		0	60	0	180

$\mu'_1 = 0/50 = 0$, $\mu'_2 = 60/50 = 1.2$, $\mu'_3 = 0/50 = 0$, $\mu'_4 = 180/50 = 3.6$

Central moments (moments about the mean), using $\mu_1' = 0$, $\mu_2' = 1.2$, $\mu_3' = 0$, $\mu_4' = 3.6$ (about $x = 2$):

- $\mu_1 = 0$ (always)
- $\mu_2 = \mu_2' - (\mu_1')^2 = 1.2 - (0)^2 = 1.2$
- $\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = 0 - 3(1.2)(0) + 2(0)^3 = 0$
- $\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4 = 3.6 - 4(0)(0) + 6(1.2)(0)^2 - 3(0)^4 = 3.6$

$$\beta_1 = \mu_3^2 / \mu_2^3 = (0)^2 / (1.2)^3 = 0.0000 \quad \beta_2 = \mu_4 / \mu_2^2 = 3.6 / (1.2)^2 = 2.5000$$

Comment: The distribution is symmetric ($\mu_3 = 0$) and platykurtic ($\beta_2 < 3$, flatter than normal).

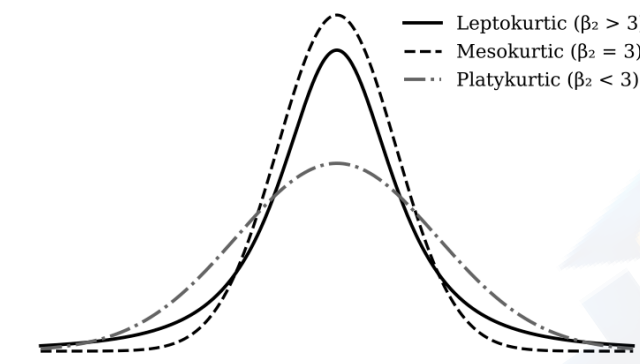


Fig.: Types of kurtosis

(Raw moments 1 M, central moments 1 M, $\beta_1 \beta_2$ 1 M, comment 1 M)

d) Fitting a second-degree parabola.

4 M

x	y	x ²	x ³	x ⁴	xy	x ² y
0	2	0	0	0	0	0
1	3	1	1	1	3	3
2	6	4	8	16	12	24
3	11	9	27	81	33	99
4	19	16	64	256	76	304
10	41	30	100	354	124	430

Normal equations for $y = a + bx + cx^2$:

$$\Sigma y = na + b\Sigma x + c\Sigma x^2$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$$

$$\Sigma x^2y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

Substituting the totals:

$$41 = 5a + 10b + 30c \quad \dots(1)$$

$$124 = 10a + 30b + 100c \quad \dots(2)$$

$$430 = 30a + 100b + 354c \quad \dots(3)$$

Solving (1), (2), (3): $a = 2.0857$, $b = -0.3714$, $c = 1.1429$

Required parabola: $y = 2.0857 - 0.3714x + 1.1429x^2$

At $x = 5$: $y = 2.0857 - 0.3714(5) + 1.1429(25) = 28.8$

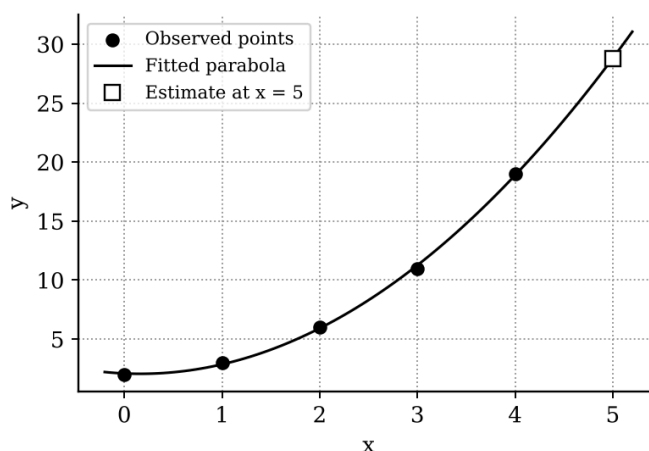


Fig.: Observed points and the fitted parabola

(Table 1 M, normal equations 1 M, solution 1 M, estimate 1 M)

Q.4

a) Regression lines and estimates.

4 M

x	y	$dx = x - \bar{x}$	$dy = y - \bar{y}$	dx^2	dy^2	$dx \cdot dy$
3	2	-4	-2.67	16	7.11	10.67
5	3	-2	-1.67	4	2.78	3.33
6	4	-1	-0.67	1	0.44	0.67
8	6	1	1.33	1	1.78	1.33
9	5	2	0.33	4	0.11	0.67
11	8	4	3.33	16	11.11	13.33
$\Sigma x = 42$	$\Sigma y = 28$			42	23.33	30

$$\bar{x} = 7, \bar{y} = 4.667$$

$$b_{yx} = \Sigma dx \cdot dy / \Sigma dx^2 = 30 / 42 = 0.7143 \quad b_{xy} = \Sigma dx \cdot dy / \Sigma dy^2 = 30 / 23.33 = 1.2857$$

Regression line of y on x: $y - \bar{y} = b_{yx}(x - \bar{x}) \Rightarrow y - 4.667 = 0.7143(x - 7) \Rightarrow$

$$y = 0.7143x - 0.3333$$

Regression line of x on y: $x - \bar{x} = b_{xy}(y - \bar{y}) \Rightarrow x - 7 = 1.2857(y - 4.667) \Rightarrow x = 1.2857y + 1$

y at $x = 10$ (use y on x): $y = 0.7143(10) - 0.3333 = 6.81$

x at $y = 7$ (use x on y): $x = 1.2857(7) + 1 = 10$

(Table 1 M, two lines 2 M, estimates 1 M)

b) Spearman's rank correlation (ranks given).

4 M

R_1	R_2	$d = R_1 - R_2$	d^2
3	6	-3	9
5	4	1	1
8	9	-1	1
4	8	-4	16
7	1	6	36
10	2	8	64
2	3	-1	1
1	10	-9	81
6	5	1	1
9	7	2	4
		$\Sigma d = 0$	$\Sigma d^2 = 214$

$$\rho = 1 - \frac{6 \Sigma d^2}{n(n^2 - 1)} = 1 - \frac{(6 \times 214)}{(10 \times 99)} = 1 - \frac{1284}{990} = -0.2970$$

Comment: there is a low degree of negative correlation — the two judges' opinions do not agree.

(d and d^2 2 M, ρ 1 M, comment 1 M)

c) Scatter diagram and Karl Pearson's r.

4 M

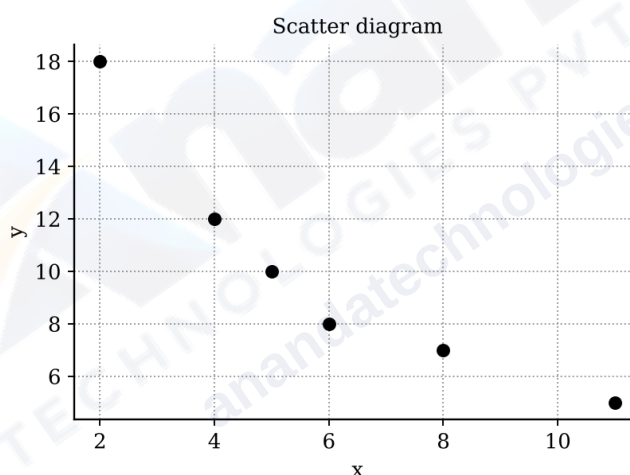


Fig.: Scatter diagram — points fall from left to right (negative correlation)

x	y	$dx = x - \bar{x}$	$dy = y - \bar{y}$	dx^2	dy^2	$dx \cdot dy$
2	18	-4	8	16	64	-32
4	12	-2	2	4	4	-4
5	10	-1	0	1	0	0
6	8	0	-2	0	4	0
8	7	2	-3	4	9	-6
11	5	5	-5	25	25	-25
$\Sigma x = 36$	$\Sigma y = 60$			50	106	-67

$$\bar{x} = \Sigma x/n = 6, \quad \bar{y} = \Sigma y/n = 10$$

$$r = \Sigma dx \cdot dy / \sqrt{(\Sigma dx^2 \cdot \Sigma dy^2)} = -67 / \sqrt{(50 \times 106)} = -67 / 72.8011 = -0.9203$$

There is a high degree of negative correlation.

(Scatter diagram 1 M, table 2 M, r 1 M)

d) Cards — addition theorem; without replacement.

4 M

(i) A = black card (26), B = king (4), $A \cap B$ = black kings (2).

$$P(A \cup B) = 26/52 + 4/52 - 2/52 = 28/52 = 7/13$$

$$(ii) P(H_1 \cap H_2) = P(H_1) \cdot P(H_2 | H_1) = (13/52) \times (12/51) = 156/2652 = 1/17 = 0.0588$$

(Each part 2 M)

e) Bayes' theorem — medical test.

4 M

Let E_1 = person has the disease, E_2 = person is healthy, T = test is positive.

$$P(E_1) = 0.02, P(E_2) = 0.98; P(T | E_1) = 0.95, P(T | E_2) = 0.04$$

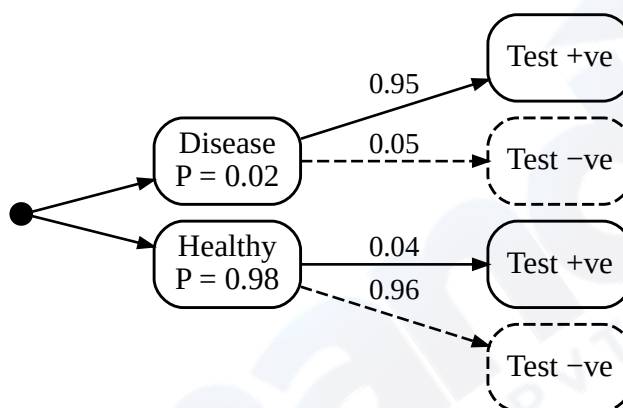


Fig.: Probability tree

$$P(T) = (0.02)(0.95) + (0.98)(0.04) = 0.019 + 0.0392 = 0.0582$$

$$P(E_1 | T) = P(E_1)P(T | E_1)/P(T) = 0.019/0.0582 = 0.3265 \approx 32.6\%$$

Even after a positive test, the chance of really having the disease is only about one-third, because the disease is rare (false positives from the large healthy group).

(Notation 1 M, P(T) 1 M, Bayes 2 M)

Q.5

a) Lagrange's interpolation.

6 M

(i)

Lagrange's interpolation formula:

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

Here $x_0, x_1, \dots = 2, 3, 5, 6$ and $y_0, y_1, \dots = 3, 7, 21, 31$. Putting $x = 4$:

$$y(4) = [(4 - 3)(4 - 5)(4 - 6)] / [(2 - 3)(2 - 5)(2 - 6)] \times 3 + [(4 - 2)(4 - 5)(4 - 6)] / [(3 - 2)(3 - 5)(3 - 6)] \times 7$$

$$\begin{aligned}
 &+ [(4-2)(4-3)(4-6)] / [(5-2)(5-3)(5-6)] \times 21 \\
 &+ [(4-2)(4-3)(4-5)] / [(6-2)(6-3)(6-5)] \times 31 \\
 &= (2/-12) \times 3 + (4/6) \times 7 + (-4/-6) \times 21 + (-2/12) \times 31 \\
 &= -0.5 + 4.6667 + 14 - 5.1667
 \end{aligned}$$

$$\Rightarrow f(4) = 13$$

$$(ii) x_0 = -1, x_1 = 0, x_2 = 2; y_0 = 2, y_1 = 1, y_2 = 5$$

$$f(x) = [(x-0)(x-2)]/[(-1-0)(-1-2)] \times 2 + [(x+1)(x-2)]/[(0+1)(0-2)] \times 1 + [(x+1)(x-0)]/[(2+1)(2-0)] \times 5$$

$$= (2/3)(x^2 - 2x) - (1/2)(x^2 - x - 2) + (5/6)(x^2 + x)$$

$$= x^2(4 - 3 + 5)/6 + x(-8 + 3 + 5)/6 + 1 = x^2 + 0 \cdot x + 1$$

$$\Rightarrow \boxed{f(x) = x^2 + 1} \quad f(1) = 1 + 1 = 2$$

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

b) Newton's backward interpolation; operator relation.

6 M

(i) Backward difference table (h = 2):

x	y	∇y	$\nabla^2 y$	$\nabla^3 y$	$\nabla^4 y$
2015	12				
		3			
2017	15		2		
		5		0	
2019	20		2		0
		7		0	
2021	27		2		
		9			
2023	36				

$$\text{Here } x_n = 2023, h = 2, x = 2022 \Rightarrow p = (x - x_n)/h = (2022 - 2023)/2 = -0.5$$

Newton's backward interpolation formula:

$$y = y_n + p\nabla y_n + p(p+1)/2! \nabla^2 y_n$$

$$y(2022) = 36 + (-0.5)(9) + (-0.5)(0.5)(2)/2$$

$$= 36 - 4.5 - 0.25$$

$$\Rightarrow \text{Sales in 2022} \approx \text{₹ } 31.25 \text{ lakh}$$

(ii) $\Delta f(x) = f(x+h) - f(x)$ and $\nabla f(x) = f(x) - f(x-h)$.

$$\Delta \nabla f(x) = \Delta[f(x) - f(x-h)] = [f(x+h) - f(x)] - [f(x) - f(x-h)] = f(x+h) - 2f(x) + f(x-h)$$

$$\nabla \Delta f(x) = \nabla[f(x+h) - f(x)] = [f(x+h) - f(x)] - [f(x) - f(x-h)] = f(x+h) - 2f(x) + f(x-h)$$

$$(\Delta - \nabla)f(x) = [f(x+h) - f(x)] - [f(x) - f(x-h)] = f(x+h) - 2f(x) + f(x-h)$$

$$\text{All three are equal} \Rightarrow \Delta \nabla = \nabla \Delta = \Delta - \nabla.$$

((i) table 1 M, formula and value 2 M; (ii) proof 3 M)

c) Extrapolation at both ends.

6 M

Difference table:

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
1	3				
		6			
2	9		4		
		10		0	
3	19		4		0
		14		0	
4	33		4		
		18			
5	51				

Second differences are constant (4), third differences are zero $\Rightarrow f(x)$ is a polynomial of degree 2, so values outside the table can be found (extrapolation).

f(0) (before the table — Newton's forward formula with $x_0 = 1$):

Here $x_0 = 1$, $h = 1$, $x = 0 \Rightarrow p = (x - x_0)/h = (0 - 1)/1 = -1$

Newton's forward interpolation formula:

$$y = y_0 + p\Delta y_0 + p(p-1)/2! \Delta^2 y_0$$

$$y(0) = 3 + (-1)(6) + (-1)(-2)(4)/2$$

$$= 3 - 6 + 4$$

$$\Rightarrow f(0) = 1$$

f(7) (after the table — Newton's backward formula with $x_n = 5$):

Here $x_n = 5$, $h = 1$, $x = 7 \Rightarrow p = (x - x_n)/h = (7 - 5)/1 = 2$

Newton's backward interpolation formula:

$$y = y_n + p\nabla y_n + p(p+1)/2! \nabla^2 y_n$$

$$y(7) = 51 + (2)(18) + (2)(3)(4)/2$$

$$= 51 + 36 + 12$$

$$\Rightarrow f(7) = 99$$

(Check: the table fits $f(x) = 2x^2 + 1$; $f(0) = 1$, $f(7) = 99$ ✓)

(Table 2 M, f(0) 2 M, f(7) 2 M)

Q.6

a) Chi-square — is the die unbiased?

6 M

H₀: the die is unbiased (each face has probability 1/6).

H₁: the die is biased.

Expected frequency of each face $E = 120 \times 1/6 = 20$.

Face	O	E	O - E	(O - E) ²	(O - E) ² /E
1	15	20	-5	25	1.2500
2	22	20	2	4	0.2000
3	18	20	-2	4	0.2000
4	26	20	6	36	1.8000
5	17	20	-3	9	0.4500
6	22	20	2	4	0.2000
Total	120	120			$\chi^2 = 4.1000$

d.f. = 6 - 1 = 5; table value $\chi^2_{0.05} = 11.07$. Since $\chi^2 = 4.100 < 11.07$, **H₀ is accepted**: the die may be regarded as unbiased.

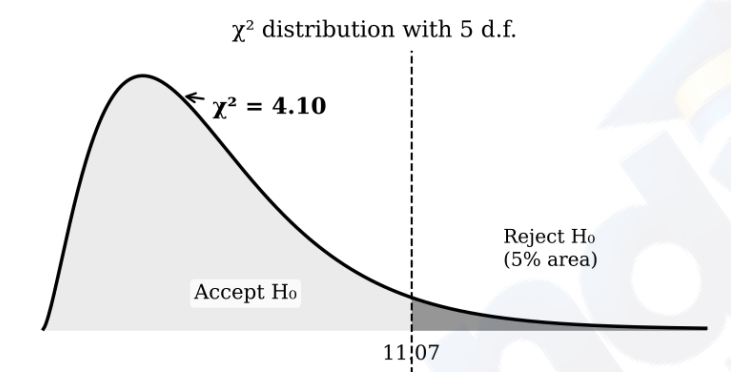


Fig.: χ^2 lies in the acceptance region

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

b) t-test for mean; large-sample test of two means.

6 M

(i)

H₀: $\mu = 100$.

H₁: $\mu \neq 100$.

x	$x - \bar{x}$	$(x - \bar{x})^2$
104	0.5	0.25
98	-5.5	30.25
110	6.5	42.25
95	-8.5	72.25
102	-1.5	2.25
108	4.5	20.25
99	-4.5	20.25
112	8.5	72.25
101	-2.5	6.25
106	2.5	6.25
$\Sigma x = 1035$		$\Sigma(x - \bar{x})^2 = 272.5$

$$\bar{x} = 103.5, s = \sqrt{(272.5/9)} = 5.5025$$

$$t = (\bar{x} - \mu)/(s/\sqrt{n}) = (103.5 - 100)/(5.5025/\sqrt{10}) = \mathbf{2.011}$$

d.f. = 9, $t_{0.05} = 2.262$. Since $|t| = 2.011 < 2.262$, **H_0 is accepted**: the data support a mean IQ of 100.

(ii)

H_0 : $\mu_1 = \mu_2$ (no difference between the colleges).

H_1 : $\mu_1 \neq \mu_2$.

$$S.E. = \sqrt{(s_1^2/n_1 + s_2^2/n_2)} = \sqrt{(64/150 + 100/120)} = \sqrt{(0.4267 + 0.8333)} = 1.1225$$

$$z = (\bar{x}_1 - \bar{x}_2)/S.E. = (62 - 59)/1.1225 = \mathbf{2.673}$$

Since $|z| = 2.673 > 1.96$ (table value), the difference is **significant**. **H_0 is rejected**. The mean marks of the two colleges differ significantly.

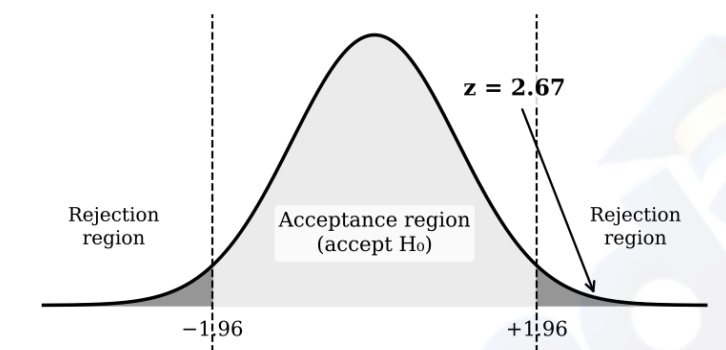


Fig.: z falls in the rejection region (5%, two-tailed)

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

c) Chi-square test of independence (2×3).

6 M

H_0 : preference of language is independent of gender.

H_1 : preference depends on gender.

Cell	O	E = (Row total $\times$ Column total)/N	O - E	(O - E) ²	(O - E) ² /E
Male – Python	60	$100 \times 110/200 = 55$	5	25	0.4545
Male – R	25	$100 \times 55/200 = 27.5$	-2.5	6.25	0.2273
Male – Java	15	$100 \times 35/200 = 17.5$	-2.5	6.25	0.3571
Female – Python	50	$100 \times 110/200 = 55$	-5	25	0.4545
Female – R	30	$100 \times 55/200 = 27.5$	2.5	6.25	0.2273
Female – Java	20	$100 \times 35/200 = 17.5$	2.5	6.25	0.3571
				$\chi^2 = \Sigma$	2.0779

$$d.f. = (r - 1)(c - 1) = (2 - 1)(3 - 1) = 2; \text{ table value} = 5.991.$$

Since $\chi^2 = 2.078 < 5.991$, **H_0 is accepted**: the preferred language is independent of gender.

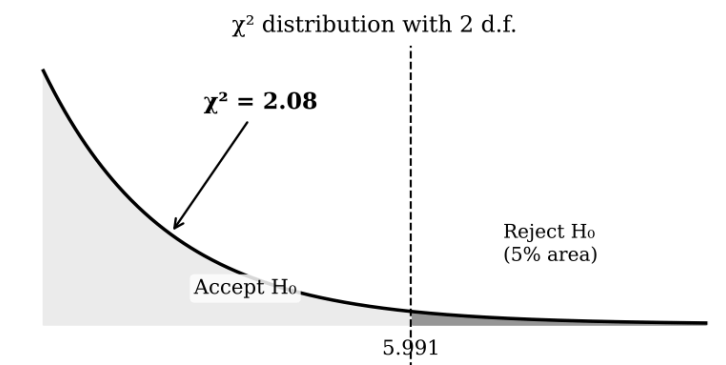


Fig.: χ^2 with 2 d.f. — calculated value in the acceptance region

(Hypotheses 1 M, expected 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** / **sympy** functions. Students can run the same check (this is not part of the exam answer):

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

d = [12, 15, 11, 15, 18, 20, 15, 13, 11, 16]
print("Q1 a) median =", np.median(d), " mode =", stats.mode(d).mode)
print("Q1 b) G.M. =", round(float(stats.gmean([3, 6, 24, 48])), 4))

# Q3 c) central moments of a frequency distribution
x = np.array([0, 1, 2, 3, 4]); f = np.array([5, 10, 20, 10, 5])
m = np.average(x, weights=f)
mu = [np.average((x - m) ** k, weights=f) for k in (2, 3, 4)]
print("Q3 c) mu2, mu3, mu4 =", [round(float(v), 4) for v in mu], " beta2 = %.4f"
% (mu[2] / mu[0] ** 2))
c2, c1, c0 = np.polyfit([0, 1, 2, 3, 4], [2, 3, 6, 11, 19], 2)
print("Q3 d) y = %.4f %+.4f x %+.4f x^2 ; y(5) = %.2f" % (c0, c1, c2, c0 + 5 *
c1 + 25 * c2))

r1 = stats.linregress([3, 5, 6, 8, 9, 11], [2, 3, 4, 6, 5, 8]); r2 =
stats.linregress([2, 3, 4, 6, 5, 8], [3, 5, 6, 8, 9, 11])
print("Q4 a) y = %.4f x %+.4f ; x = %.4f y %+.4f" % (r1.slope, r1.intercept,
r2.slope, r2.intercept))
print("Q4 b) rho =", round(stats.spearmanr([3, 5, 8, 4, 7, 10, 2, 1, 6, 9], [6,
4, 9, 8, 1, 2, 3, 10, 5, 7]).statistic, 4))
print("Q4 c) r =", round(stats.pearsonr([2, 4, 5, 6, 8, 11], [18, 12, 10, 8, 7,
5]).statistic, 4))
print("Q4 e) P(disease | +ve) = %.4f" % (0.02 * 0.95 / (0.02 * 0.95 + 0.98 *
0.04)))

X = sp.symbols('x')
print("Q5 a) f(4) =", sp.interpolate([(2, 3), (3, 7), (5, 21), (6, 31)],
X).subs(X, 4),
      "; polynomial:", sp.expand(sp.interpolate([(-1, 2), (0, 1), (2, 5)], X)))
u = (np.array([2015, 2017, 2021, 2023]) - 2015) / 2
print("Q5 b) sales 2022 = %.4f" % np.polyval(np.polyfit(u, [12, 15, 20, 27, 36],
```

```
4), 3.5))
p = sp.expand(sp.interpolate([(1, 3), (2, 9), (3, 19), (4, 33), (5, 51)], X))
print("Q5 c) f(x) =", p, "; f(0) =", p.subs(X, 0), "; f(7) =", p.subs(X, 7))

r = stats.chisquare([15, 22, 18, 26, 17, 22])
print("Q6 a) chi2 = %.4f, p-value = %.4f" % (r.statistic, r.pvalue))
r = stats.ttest_1samp([104, 98, 110, 95, 102, 108, 99, 112, 101, 106], 100)
print("Q6 b)(i) t = %.4f, p-value = %.4f" % (r.statistic, r.pvalue))
print("Q6 b)(ii) z = %.4f" % ((62 - 59) / np.sqrt(8 ** 2 / 150 + 10 ** 2 / 120)))
r = stats.chi2_contingency([[60, 25, 15], [50, 30, 20]], correction=False)
print("Q6 c) chi2 = %.4f, dof = %d, p-value = %.4f" % (r.statistic, r.dof, r.pvalue))
```

Output:

```
Q1 a) median = 15.0   mode = 15
Q1 b) G.M. = 12.0
Q3 c) mu2, mu3, mu4 = [1.2, 0.0, 3.6]   beta2 = 2.5000
Q3 d) y = 2.0857 -0.3714 x +1.1429 x^2 ; y(5) = 28.80
Q4 a) y = 0.7143 x -0.3333 ; x = 1.2857 y +1.0000
Q4 b) rho = -0.297
Q4 c) r = -0.9203
Q4 e) P(disease | +ve) = 0.3265
Q5 a) f(4) = 13 ; polynomial: x**2 + 1
Q5 b) sales 2022 = 31.2500
Q5 c) f(x) = 2*x**2 + 1 ; f(0) = 1 ; f(7) = 99
Q6 a) chi2 = 4.1000, p-value = 0.5351
Q6 b)(i) t = 2.0114, p-value = 0.0752
Q6 b)(ii) z = 2.6726
Q6 c) chi2 = 2.0779, dof = 2, p-value = 0.3538
```

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