

Unit 6 — Correlation & Regression

Correlation & types · Karl Pearson's coefficient · Spearman's rank correlation (with repeated ranks) · Regression lines X on Y and Y on X · Regression coefficients · Angle between regression lines

Exam slot: Part-B Q11 (10 M) — OR with Unit 5

Part-A: 1j (3 M)

Written in ASD-STE100 style

PYQ TREND (2020 – 2026): WHAT THE EXAM ASKS FROM THIS UNIT

- **Part A 1j, 3 M:** "If $x = 2y + 3$, $y = \frac{x}{8} + 6$ are the regression lines, find r " — Apr-2026, Nov-2024, Mar-2024. Answer: $r = 0.5$.
- **Spearman's rank correlation with repeated ranks** (C++ vs Statistics; fathers vs sons; economics vs statistics): Apr-2026, May-2025, Nov-2024.
- **Two regression lines + prediction** (internal vs university marks; predict Y at $X = 67.5$).
- **Wrongly copied pairs** — find the correct r and regression lines: Oct-2023, Apr-2023, Mar-2022.

6.1 Correlation and its Types

CORRELATION

Correlation expresses the relationship (interdependence) between two sets of variables. One variable is the **independent** (subject) variable. The other is the **dependent** (relative) variable. The **correlation coefficient r** measures the strength and direction of the **linear** relationship.

Types of correlation Oct-2023 · 3M

1. Positive and negative

Positive: the two variables move in the **same** direction. *Examples:* height and weight; rainfall and crop yield; price and supply.

Negative: they move in **opposite** directions.
Example: price and demand.

2. Simple and multiple

Simple: we study only **two** variables.

Multiple: we study **more than two** variables at the same time.

3. Partial and total

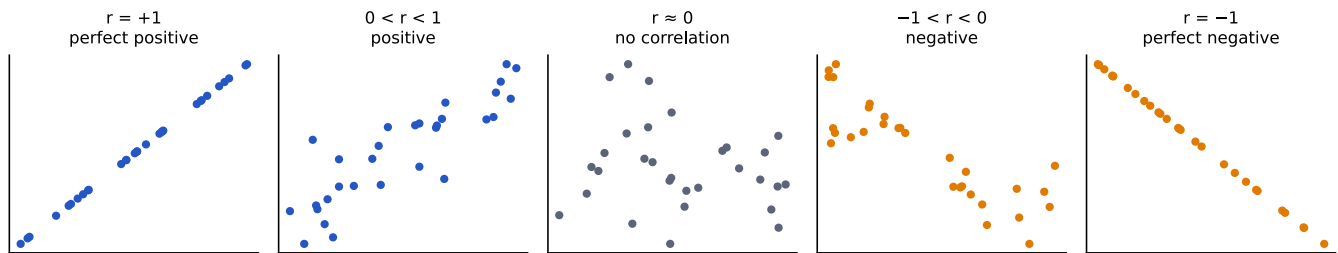
Partial: we study two variables and keep the effect of other variables constant (excluded).

Total: we take all the facts into account.

4. Linear and non-linear

Linear: the ratio of change between the two variables is uniform (points near a straight line).

Non-linear (curvilinear): the ratio of change is not uniform.

Fig 6.1 — Scatter diagrams for different values of r .

6.2 Karl Pearson's Coefficient of Correlation

DEFINITION AND FORMULAS APR/MAY-2023 · 3M

$$r = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y} = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}, \quad X = x - \bar{x}, Y = y - \bar{y}$$

Direct (raw-sum) form:
$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$$

Properties: (1) $-1 \leq r \leq 1$. (2) r has no units. (3) r does not change with a change of origin or scale. (4) $r = \pm 1$: perfect correlation; $r = 0$: no linear correlation.

CAUTION

$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$ is correct **only** when x, y are deviations from the mean. With raw sums, use the direct form.

PYQ. Calculate Karl Pearson's correlation coefficient: X: 10, 20, 30, 40, 50; Y: 12, 21, 32, 43, 54.

Oct-2025 · 5M
Class notes

x	$X = x - 30$	X^2	y	$Y = y - 32.4$	Y^2	XY
10	-20	400	12	-20.4	416.16	408
20	-10	100	21	-11.4	129.96	114
30	0	0	32	-0.4	0.16	0
40	10	100	43	10.6	112.36	106
50	20	400	54	21.6	466.56	432
150		1000	162		1125.2	1060

$\bar{x} = 30, \bar{y} = 32.4. r = \frac{1060}{\sqrt{1000 \times 1125.2}} = 0.9993$. Very strong positive correlation. **$r = 0.9993$**

PYQ. Find Karl Pearson's coefficient: Wages 100, 101, 102, 102, 100, 99, 97, 98, 96, 95; Cost of living 98, 99, 99, 97, 95, 92, 95, 94, 90, 91. Nov-2020 · 7M Class notes

$\bar{x} = 99, \bar{y} = 95. \sum X^2 = 54, \sum Y^2 = 96, \sum XY = 61. r = \frac{61}{\sqrt{54 \times 96}} = \frac{61}{72} = 0.847$.

PYQ. Find a suitable coefficient of correlation: Fertilizer (tons) 15, 18, 20, 24, 30, 35, 40, 50; Productivity (tons) 85, 93, 95, 105, 120, 130, 150, 160. **Nov-2020 · 8M**

Both are measured numbers, so Karl Pearson's r is suitable. $\bar{x} = 29$, $\bar{y} = 117.25$. $\sum X^2 = 1022$, $\sum Y^2 = 5343.5$, $\sum XY = 2317$.

$$r = \frac{2317}{\sqrt{1022 \times 5343.5}} = 0.9915. \text{ (Ranks of } x \text{ and } y \text{ agree completely, so Spearman's } \rho = 1.)$$

Example 1. Heights (57, 59, 62, 63, 64, 65, 55, 58, 57 in) and weights (113, 117, 126, 126, 130, 129, 111, 116, 112 lb). **Class notes**

$$\bar{x} = 60, \bar{y} = 120, \sum X^2 = 102, \sum Y^2 = 472, \sum XY = 216. r = \frac{216}{\sqrt{102 \times 472}} = 0.984.$$

6.3 Spearman's Rank Correlation Coefficient

FORMULA (NO REPEATED RANKS)

$$\rho = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}$$

ρ = rank correlation coefficient, D = difference between the two ranks of an item, n = no. of pairs.

Properties: (1) $-1 \leq \rho \leq 1$. (2) $\rho = 1$: complete agreement in the order of ranks. (3) $\rho = -1$: complete disagreement; the ranks are in opposite order.

FORMULA WITH EQUAL (REPEATED) RANKS

Give each tied value the **average** of the ranks it would occupy. Then

$$\rho = 1 - \frac{6 \left[\sum D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \dots \right]}{n(n^2 - 1)}$$

m_i = number of times a rank repeats. Add one correction term for **each** group of ties, in **both** series.

PYQ. Distinguish between Karl Pearson's and Spearman's rank correlation coefficients.

Oct-2025 · 3M

Karl Pearson's r	Spearman's ρ
Uses the actual values of x and y .	Uses only the ranks.
For quantitative (measured) data.	For qualitative data (beauty, honesty) or ranked data.
$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$	$\rho = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}$
Affected by extreme values.	Less affected by extreme values.
Measures linear relationship only.	Measures any monotonic relationship.

Example 2. Grades of 5 students. Maths: 85, 60, 73, 40, 90. Statistics: 93, 75, 65, 50, 80. **Class notes**

Ranks (highest = 1): Maths 2, 4, 3, 5, 1; Stats 1, 3, 4, 5, 2. D : 1, 1, -1, 0, -1. $\sum D^2 = 4$.

$$\rho = 1 - \frac{6(4)}{5(24)} = 1 - 0.2 = 0.8.$$

Example 3. Ranks of 16 students in Maths and Statistics: (1,1)(2,10)(3,3)(4,4)(5,5)(6,7)(7,2)(8,6)(9,8)(10,11)(11,15)(12,9)(13,14)(14,12)(15,16)(16,13). [Class notes](#)

$$\sum D^2 = 136. \quad \rho = 1 - \frac{6 \times 136}{16(256-1)} = 1 - \frac{816}{4080} = 0.8.$$

The class notes miss the factor 6 in the working but write 0.8. Always write the 6.

PYQ. Marks of 10 students in C++ (x): 6, 5, 8, 8, 7, 6, 10, 4, 9, 7 and Statistics (y): 8, 7, 7, 10, 5, 8, 10, 6, 8, 6. Compute Spearman's rank correlation coefficient. [Apr/May-2026 · 10M](#) [Class notes](#)

x	Rank x	y	Rank y	D	D ²
6	7.5	8	4	3.5	12.25
5	9	7	6.5	2.5	6.25
8	3.5	7	6.5	-3	9
8	3.5	10	1.5	2	4
7	5.5	5	10	-4.5	20.25
6	7.5	8	4	3.5	12.25
10	1	10	1.5	-0.5	0.25
4	10	6	8.5	1.5	2.25
9	2	8	4	-2	4
7	5.5	6	8.5	-3	9
$\sum D^2$					79.5

Step 1 Ties in x: 8 (×2), 7 (×2), 6 (×2) → three groups with $m = 2$. Ties in y: 10 (×2), 8 (×3), 7 (×2), 6 (×2) → $m = 2, 3, 2, 2$.

Step 2 Correction = $\frac{1}{12} [6(2^3 - 2) + (3^3 - 3)] = \frac{1}{12} [36 + 24] = 5$.

Step 3 $\rho = 1 - \frac{6(79.5 + 5)}{10(100 - 1)} = 1 - \frac{507}{990} = 0.488$.

$\rho = 0.488$

PYQ. Heights of 12 fathers (X): 65, 63, 67, 64, 68, 62, 70, 66, 68, 67, 69, 71 and their eldest sons (Y): 68, 66, 68, 65, 69, 66, 68, 65, 71, 67, 68, 70. Compute Spearman's rank correlation coefficient.

[May-2025 · 10M](#) [Class notes](#)

Ranks X: 9, 11, 6.5, 10, 4.5, 12, 2, 8, 4.5, 6.5, 3, 1. Ranks Y: 5.5, 9.5, 5.5, 11.5, 3, 9.5, 5.5, 11.5, 1, 8, 5.5, 2. $\sum D^2 = 72.5$. Ties: X: 67 (×2), 68 (×2). Y: 68 (×4), 66 (×2), 65 (×2).

Correction = $\frac{1}{12} [(8 - 2) + (8 - 2) + (64 - 4) + (8 - 2) + (8 - 2)] = \frac{84}{12} = 7$.

$\rho = 1 - \frac{6(72.5 + 7)}{12(143)} = 1 - \frac{477}{1716} = 0.722$. **$\rho = 0.722$**

PYQ. Marks of 8 students in Economics (x): 78, 56, 36, 66, 25, 75, 82, 62 and Statistics (y): 84, 44, 57, 58, 60, 68, 62, 58. Compute Spearman's rank correlation. [Nov-2024 · 10M](#)

Ranks x: 2, 6, 7, 4, 8, 3, 1, 5. Ranks y: 1, 8, 7, 5.5, 4, 2, 3, 5.5 (58 occurs twice → ranks 5, 6 → 5.5). D : 1, -2, 0, -1.5, 4, 1, -2, -0.5. $\sum D^2 = 1 + 4 + 0 + 2.25 + 16 + 1 + 4 + 0.25 = 28.5$. Correction =

$$\frac{1}{12}(8 - 2) = 0.5.$$

$$\rho = 1 - \frac{6(28.5 + 0.5)}{8(63)} = 1 - \frac{174}{504} = 0.655. \quad \rho = 0.655$$

6.4 Regression

REGRESSION AND REGRESSION EQUATION OCT-2025 · 5M

Regression is the statistical method that estimates the unknown value of one variable from the known value of a related variable. The equation that we use for this is the **regression equation (line of regression)**.

Regression of Y on X: we estimate Y for a given X (X is independent). **Regression of X on Y:** we estimate X for a given Y.

Example: from the marks in an internal test (X), estimate the university marks (Y). This is the regression of Y on X.

LINES OF REGRESSION AND REGRESSION COEFFICIENTS

Y on X: $y - \bar{y} = b_{yx}(x - \bar{x}), \quad b_{yx} = r \frac{\sigma_y}{\sigma_x} = \frac{\sum XY}{\sum X^2}$

X on Y: $x - \bar{x} = b_{xy}(y - \bar{y}), \quad b_{xy} = r \frac{\sigma_x}{\sigma_y} = \frac{\sum XY}{\sum Y^2}$

($X = x - \bar{x}, Y = y - \bar{y}$.) **Key result:** $b_{xy} \times b_{yx} = r^2$, so $r = \pm \sqrt{b_{xy} b_{yx}}$.

PROPERTIES OF REGRESSION COEFFICIENTS NOV/DEC-2022 · 3M

1. $r = \pm \sqrt{b_{xy} b_{yx}}$. r has the **same sign** as the regression coefficients.
2. Both coefficients have the same sign.
3. If one coefficient is greater than 1, the other is less than 1 (because $r^2 \leq 1$).
4. $\frac{b_{xy} + b_{yx}}{2} \geq r$ (arithmetic mean $\geq r$, for positive r).
5. They do not change with a change of origin, but they change with a change of scale.
6. The two regression lines intersect at $(\bar{x}, \bar{y})$.

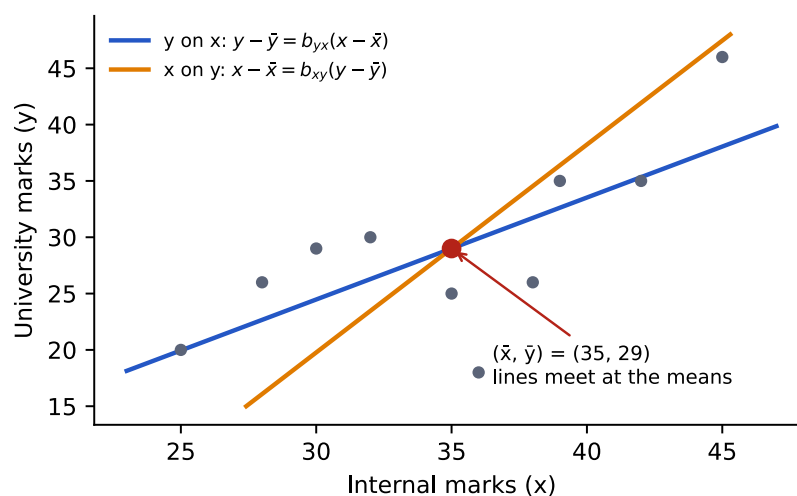


Fig 6.2 — Two regression lines for the internal/university marks PYQ. They meet at the means.

Finding r and the means from two given regression lines

Step 1

Write the line of Y on X in the form $y = a + b_{yx}x$. Write the line of X on Y in the form $x = c + b_{xy}y$.

Step 2 $r = \pm \sqrt{b_{xy}b_{yx}}$. Take the sign of the coefficients.

Step 3 If $b_{xy}b_{yx} > 1$, you labelled the lines wrongly. Swap them.

Step 4 Solve the two equations together. The solution is $(\bar{x}, \bar{y})$.

PYQ. If $x = 2y + 3$ and $y = \frac{x}{8} + 6$ are the regression lines of x on y and y on x respectively, find the correlation coefficient. Apr/May-2026 · 3M Nov-2024 · 3M Mar/Apr-2024 · 3M Class notes

$$b_{xy} = 2, b_{yx} = \frac{1}{8}. r^2 = 2 \times \frac{1}{8} = \frac{1}{4}, r = +\frac{1}{2} \text{ (both coefficients are positive).}$$

$$\text{Means: } x = 2y + 3 \text{ and } x = 8y - 48 \Rightarrow 2y + 3 = 8y - 48 \Rightarrow y = 8.5, x = 20.$$

$$r = 0.5, \bar{x} = 20, \bar{y} = 8.5$$

PYQ. For army personnel, the regression of weight of kidneys (y) on weight of heart (x) is $y = 0.399x + 6.394$, and the regression of x on y is $x = 1.212y - 2.461$. Find r . May-2025 · 3M

$$b_{yx} = 0.399, b_{xy} = 1.212. r = \sqrt{0.399 \times 1.212} = \sqrt{0.4836} = 0.695.$$

Example 4. The regression lines are $7x - 16y + 9 = 0$ and $5y - 4x - 3 = 0$. Find r and the means.

Class notes

Trial 1: First line as x on y : $x = \frac{16}{7}y - \frac{9}{7}$, $b_{xy} = \frac{16}{7}$. Second as y on x : $y = \frac{4}{5}x + \frac{3}{5}$, $b_{yx} = \frac{4}{5}$. Product = $\frac{64}{35} > 1$. Not possible.

Trial 2 (correct): First line as y on x : $y = \frac{7}{16}x + \frac{9}{16}$, $b_{yx} = \frac{7}{16}$. Second as x on y : $x = \frac{5}{4}y - \frac{3}{4}$, $b_{xy} = \frac{5}{4}$. $r^2 = \frac{7}{16} \times \frac{5}{4} = \frac{35}{64}$, $r = \frac{\sqrt{35}}{8} = 0.74$.

Means: solve $7x - 16y = -9$ and $-4x + 5y = 3$: $\bar{y} = \frac{15}{29}$, $\bar{x} = -\frac{3}{29}$.

CORRECTION TO CLASS NOTES

The class notes take trial 1 and get $r = \pm \frac{8}{\sqrt{35}}$, which is greater than 1. That is impossible. Also $\bar{x} = -\frac{3}{29}$, not $+\frac{3}{29}$.

PYQ. Marks of 10 students. Internal: 25, 28, 30, 32, 35, 36, 38, 39, 42, 45. University: 20, 26, 29, 30, 25, 18, 26, 35, 35, 46. Obtain the two regression equations and find (a) the most likely internal marks for university marks of 25 (b) the most likely university marks for internal marks of 30.

Mar/Apr-2024 · 10M

Class notes

x	X	X^2	y	Y	Y^2	XY
25	-10	100	20	-9	81	90
28	-7	49	26	-3	9	21
30	-5	25	29	0	0	0
32	-3	9	30	1	1	-3
35	0	0	25	-4	16	0
36	1	1	18	-11	121	-11
38	3	9	26	-3	9	-9
39	4	16	35	6	36	24
42	7	49	35	6	36	42
45	10	100	46	17	289	170
350		358	290		598	324

$$\bar{x} = 35, \bar{y} = 29. b_{yx} = \frac{324}{358} = 0.905, b_{xy} = \frac{324}{598} = 0.542.$$

Step 1 Y on X: $y - 29 = 0.905(x - 35)$, i.e. $y = 0.905x - 2.68$.

Step 2 X on Y: $x - 35 = 0.542(y - 29)$, i.e. $x = 0.542y + 19.29$.

Step 3 (a) $y = 25$: $x = 35 + 0.542(25 - 29) = 32.83 \approx 33$.

Step 4 (b) $x = 30$: $y = 29 + 0.905(30 - 35) = 24.47 \approx 24$.

$$r = \sqrt{0.905 \times 0.542} = 0.700.$$

The class notes use 32 for the 5th university mark (the PYQ has 25). With 32: $\bar{y} = 29.7$, $b_{xy} = 0.553$, answers 33 and 25.

PYQ. Find the correlation coefficient and the lines of regression, and predict Y if X = 67.5. X: 65, 66, 67, 67, 68, 69, 71, 73; Y: 67, 68, 64, 68, 72, 70, 69, 70. Jul-2021 · 15M

Class notes

$$\bar{x} = 68.25, \bar{y} = 68.5, \sum X^2 = 49.5, \sum Y^2 = 40, \sum XY = 21.$$

$$r = \frac{21}{\sqrt{49.5 \times 40}} = 0.472. b_{yx} = \frac{21}{49.5} = 0.4242, b_{xy} = \frac{21}{40} = 0.525.$$

$$Y \text{ on } X: y - 68.5 = 0.4242(x - 68.25). X \text{ on } Y: x - 68.25 = 0.525(y - 68.5).$$

$$\text{At } x = 67.5: y = 68.5 + 0.4242(-0.75) = 68.18. \quad \mathbf{Y \approx 68.18}$$

PYQ. Find the coefficient of correlation and the regression lines: x: 6, 2, 10, 4, 8; y: 9, 11, 5, 8, 7.

Nov/Dec-2022 · 10M

$$\bar{x} = 6, \bar{y} = 8. X: 0, -4, 4, -2, 2; Y: 1, 3, -3, 0, -1. \sum X^2 = 40, \sum Y^2 = 20, \sum XY = 0 - 12 - 12 + 0 - 2 = -26.$$

$$r = \frac{-26}{\sqrt{800}} = -0.919. b_{yx} = -0.65, b_{xy} = -1.3.$$

$$Y \text{ on } X: y - 8 = -0.65(x - 6) \Rightarrow y = -0.65x + 11.9. X \text{ on } Y: x - 6 = -1.3(y - 8) \Rightarrow x = -1.3y + 16.4.$$

PYQ. Find the regression line of Y on X and the correlation coefficient: x: 1, 5, 3, 2, 1, 1, 7, 3; y: 6, 1, 0, 0, 1, 2, 1, 5. Oct/Nov-2021 · 15M Class notes

$$\bar{x} = 2.875, \bar{y} = 2. \sum X^2 = 32.875, \sum Y^2 = 36, \sum XY = -10.$$

$$b_{yx} = \frac{-10}{32.875} = -0.304. \text{ Y on X: } y - 2 = -0.304(x - 2.875) \Rightarrow y = -0.304x + 2.875.$$

$$r = \frac{-10}{\sqrt{32.875 \times 36}} = -0.291.$$

Correction: the class notes get $\sum XY = -5.75$ because of two sign slips (row 2: $2.125 \times (-1) = -2.125$; row 5: $(-1.875)(-1) = +1.875$). The correct total is -10 .

Correcting wrongly copied data

METHOD

Correct sum = Given sum - (wrong values) + (correct values). Do this for $\sum x, \sum y, \sum x^2, \sum y^2, \sum xy$.

Then use the **raw-sum** formulas: $S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$, $S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$, $S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \quad b_{yx} = \frac{S_{xy}}{S_{xx}}, \quad b_{xy} = \frac{S_{xy}}{S_{yy}}.$$

PYQ. For 25 pairs: $\sum x = 127, \sum y = 100, \sum x^2 = 760, \sum y^2 = 449, \sum xy = 500$. Later, two pairs were found copied as (8, 14) and (8, 6) instead of (8, 12) and (6, 8). Find the correct r and the correct regression lines. Oct-2023 · 10M Apr/May-2023 · 10M Class notes

Step 1 $\sum x = 127 - 8 - 8 + 8 + 6 = 125; \sum y = 100 - 14 - 6 + 12 + 8 = 100.$

Step 2 $\sum x^2 = 760 - 64 - 64 + 64 + 36 = 732; \sum y^2 = 449 - 196 - 36 + 144 + 64 = 425.$

Step 3 $\sum xy = 500 - 112 - 48 + 96 + 48 = 484.$

Step 4 $\bar{x} = 5, \bar{y} = 4. S_{xx} = 732 - \frac{125^2}{25} = 107; S_{yy} = 425 - 400 = 25; S_{xy} = 484 - \frac{125 \times 100}{25} = -16.$

Step 5 $r = \frac{-16}{\sqrt{107 \times 25}} = -0.309.$

Step 6 $b_{yx} = \frac{-16}{107} = -0.1495, b_{xy} = \frac{-16}{25} = -0.64.$
Y on X: $y - 4 = -0.1495(x - 5) \Rightarrow y = -0.1495x + 4.748.$
X on Y: $x - 5 = -0.64(y - 4) \Rightarrow x = -0.64y + 7.56.$

r = -0.309

CORRECTION TO CLASS NOTES

The class notes put the raw sums into $r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{484}{\sqrt{732 \times 425}} = 0.868$. That formula needs deviations from the mean. Subtract $\frac{\sum x \sum y}{n}$ etc. first, as in step 4.

PYQ. For 10 pairs: $\sum x = 66, \sum y = 69, \sum x^2 = 476, \sum y^2 = 521, \sum xy = 485$. Two pairs were copied as (2, 3), (7, 5) instead of (4, 6), (9, 8). Find the correct r. Mar/Apr-2022 · 15M Class notes

Correct: $\sum x = 70, \sum y = 75, \sum x^2 = 520, \sum y^2 = 587, \sum xy = 540.$

$S_{xx} = 520 - 490 = 30, S_{yy} = 587 - 562.5 = 24.5, S_{xy} = 540 - 525 = 15.$

$r = \frac{15}{\sqrt{30 \times 24.5}} = 0.553. b_{yx} = 0.5, b_{xy} = 0.612. \text{ Y on X: } y - 7.5 = 0.5(x - 7); \text{ X on Y: } x - 7 = 0.612(y - 7.5).$

Regression from means, s.d.s and r

Example 5. Class notes problems. Class notes

- Rainfall ($\bar{x} = 30, \sigma = 5$), production ($\bar{y} = 500$ kg, $\sigma = 100$), $r = 0.8$. Production for rainfall 40: $b_{yx} = 0.8 \cdot \frac{100}{5} = 16$; $y = 500 + 16(10) = 660$ kg.
- $\bar{x} = 7.6, \sigma_x = 3.6$; $\bar{y} = 14.8, \sigma_y = 2.5$; $r = 0.99$. y at $x = 12$: $b_{yx} = 0.6875$; $y = 14.8 + 0.6875(4.4) = 17.825$.
- Judges P and Q: $\bar{x} = 43, \bar{y} = 38, b_{yx} = \frac{21}{28} = 0.75$. Q's mark when P gives 37: $y = 38 + 0.75(-6) = 33.5$.
- Price (10, 12, 13, 12, 16, 15) and demand (40, 38, 43, 45, 37, 43): $b_{yx} = \frac{-6}{24} = -0.25$; demand at price 20: $41 - 0.25(7) = 39.25$.

6.5 Angle between Two Regression Lines

FORMULA (FROM CLASS NOTES)

Slopes: Y on X: $m_1 = r \frac{\sigma_y}{\sigma_x}$. X on Y written as y: $m_2 = \frac{1}{r} \frac{\sigma_y}{\sigma_x}$. Using $\tan \theta = \left| \frac{m_1 - m_2}{1 + m_1 m_2} \right|$:

$$\tan \theta = \frac{1 - r^2}{|r|} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

If $r = 0$: $\theta = 90^\circ$ (lines perpendicular). If $r = \pm 1$: $\theta = 0$ (lines coincide).

Example 6. (a) $\sigma_y = 2\sigma_x, r = 0.25$: find $\tan \theta$. (b) $\sigma_x = \sigma_y, \theta = \tan^{-1} \frac{4}{3}$: find r . (c) $\tan \theta = 0.6, \sigma_x = \frac{1}{2}\sigma_y$: find r . Class notes

(a) $\tan \theta = \frac{1 - 0.0625}{0.25} \cdot \frac{2\sigma_x^2}{5\sigma_x^2} = 3.75 \times 0.4 = 1.5$.

(b) $\frac{4}{3} = \frac{1 - r^2}{r} \cdot \frac{1}{2} \Rightarrow 3r^2 + 8r - 3 = 0 \Rightarrow (3r - 1)(r + 3) = 0 \Rightarrow r = \frac{1}{3}$.

(c) $0.6 = \frac{1 - r^2}{r} \cdot \frac{2}{5} \Rightarrow 2r^2 + 3r - 2 = 0 \Rightarrow r = \frac{1}{2}$.

6.6 Quick Revision Sheet

Pearson

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}} \text{ (deviations)}$$

$$r = \frac{n \Sigma xy - \Sigma x \Sigma y}{\sqrt{n \Sigma x^2 - (\Sigma x)^2} \sqrt{n \Sigma y^2 - (\Sigma y)^2}}$$

Spearman

$$\rho = 1 - \frac{6[\Sigma D^2 + \Sigma \frac{m^3 - m}{12}]}{n(n^2 - 1)}$$

Regression lines

$$y - \bar{y} = b_{yx}(x - \bar{x}), b_{yx} = \frac{\Sigma XY}{\Sigma X^2} = r \frac{\sigma_y}{\sigma_x}$$

$$x - \bar{x} = b_{xy}(y - \bar{y}), b_{xy} = \frac{\Sigma XY}{\Sigma Y^2} = r \frac{\sigma_x}{\sigma_y}$$

Links

$$r^2 = b_{xy} b_{yx}; \text{ lines meet at } (\bar{x}, \bar{y})$$

$$\tan \theta = \frac{1 - r^2}{|r|} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

★ Assignment — Unit 6 (5 questions)

1. Define correlation. Explain the types of correlation with examples.

[5M]

2. Calculate Karl Pearson's coefficient: x: 1, 2, 3, 4, 5, 6, 7; y: 9, 8, 10, 12, 11, 13, 14.

[5M]

3. Ten competitors are ranked by three judges. A: 1, 6, 5, 10, 3, 2, 4, 9, 7, 8; B: 3, 5, 8, 4, 7, 10, 2, 1, 6, 9. Find the rank correlation between judges A and B. [5M]
4. The regression lines are $8x - 10y + 66 = 0$ and $40x - 18y = 214$. Find (i) the means of x and y (ii) r (iii) σ_y if $\sigma_x^2 = 9$. [10M]
5. Obtain both regression lines and r: x: 1, 2, 3, 4, 5; y: 2, 5, 3, 8, 7. Estimate y when x = 6. [10M]

Answers for self-check: Q2: $r = 0.929$. Q3: $\sum D^2 = 200$, $\rho = -0.212$. Q4: $\bar{x} = 13$, $\bar{y} = 17$, $r = 0.6$, $\sigma_y = 4$. Q5: $y = 1.1 + 1.3x$, $x = 0.5 + 0.5y$, $r = 0.806$, $y(6) = 8.9$.