Topic → Covariance & Correlation

eg $x = \{5, 2, 8, 4, 6\}$ eg of $y = \{36, 24, 9, 5\}$

Step-1: compute $\bar{x} = \frac{25}{5} = 5$

compute $\bar{y} = \frac{54}{6} = 48$

$(x - \bar{x})^2 \times (y - \bar{y})^2$

Step-2: Difference from mean

$(x - \bar{x}) = [0, -3, 3, -1, 1]$

$(y - \bar{y}) = [-12, 12, -39, 33, 0.2]$

Step-3: Multiply the diff

$(x - \bar{x}) \times (y - \bar{y}) = [0, -3, \ldots]$

Step-4: Summation

$\Sigma (x - \bar{x}) \times (y - \bar{y}) = \ldots$

Step-5: $\text{Cov}(x, y) = \frac{-15}{5} = -3$

unit Dependent

long → $\boxed{\text{o o}}$ → metric

Eng → Inch → Kg

Covariance	Correlation
① Direction	① Direction
② $\sigma_x \sigma_y$	② magnitude
③ $-\infty$ to $+\infty$	③ -1 to $+1$
	-1 & 0 → negative
	0 & 1 → positive
	$+1 \to 0$ & 1
	mean closer 1 → strong
	" " 0 → weak

$\rho = 0.95$

Correlation

empirical formula

$$\rho(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \times \sigma_y}$$

Standardization (μ & σ → 0 to 1, 0 → 1)

then, Correlation which lie b/w -1 to +1

$$\text{correlation}_{(x,y)} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Step-1: $\text{Var}(x)$

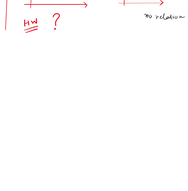
$\text{Var}(x) = [(x_1 - \bar{x})^2] = \ldots \approx 29$

$\text{Var}(y) = [\ldots] = \ldots \approx 26$

Step-2: SD (σ_x & σ_y)

Step-3: compute $\text{Corr}(x, y)$

$= \frac{-3}{\sqrt{29} \times \sqrt{26}} = \simeq -0.1$



unexpect → Weakly negatively correlated.

Correlation

method-1: Pearson Correlation

method-2: Spearman Correlation

$$\rho(x, y) = \text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\sigma_x \times \sigma_y}$$

Pearson Corr

Spearman Corr

$$r = \frac{\text{cov}(\xi_x, \xi_y)}{\sigma_{\xi_x} \times \sigma_{\xi_y}}$$

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$ → applicable for original elements i.e x & y

Students	marks	rank
A	90	2
B	80	3
C	95	1
D	70	4
E	65	5

Value doesn't matter much

Order/Rank

eg $x = \{12, 23, 17, 21, 32\}$

$y = \{113, 123, 117, 115, 122\}$

x	y	rank$_x$	rank$_y$	d diff(x, y)	d^2
12	113	1	1	0	0
23	123	3	3	0	0
17	117	2	4	-2	4
21	115	4	2	2	4
32	122	5	5	0	0

Step-1: Sort x / find rank$_x$

Step-2: Find rank$_y$

Step-3: Difference b/w rank$_x$ & rank$_y$

Step-4: Square of diff

Step-5: Put into empirical formula

$r = 1 - \frac{6 \times 8}{n^3 - n} = \checkmark$

$n = 5$ (no of samples)

Lower value → Lower Rank

higher " → higher "

eg Find Spearman Corr coeff

$x = \{10, 12, 18, 18, 15, 17, 40\}$

$y = \{15, 19, 25, 30, 20, 26, 30\}$

x	y	Rank$_x$	Rank$_y$	rank$_y$	d	d^2
10	15	1	1	1	0	⋮
12	19	2	2	2	0	⋮
18	25	3 (5.5)	5.5	5	0.5	⋮
18	30	4	6.5	6	0.5	⋮
15	20	3	3	4	4	⋮
17	26	4	4	5	1	⋮
40	30	7	7	6.5	0.5	

$$r = 1 - \frac{6 \left[\sum d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3)\right]}{n^3 - n}$$ → application → there is a repition of elements.

m_1, m_2, m_3

Q) Which one to use → Pearson Vs Spearman

V. imp

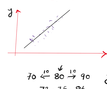
Pearson → Quantitative data (numerical)

Spearman → Qualitative data (Categorical)

(eg. Beautiful Vs intelligent, Sweet Vs sour)

Outliers in data → ? → Spearman's formula is preferred Pearson

Probable Error of Correlation Coefficient



F_1, F_{1000}

Standard error

$$S.E = \frac{1 - r^2}{\sqrt{n}}$$

r = correlation coefficient

n = no of samples

70 ← 80 → 90

72, 75, 84

$\bar{x}_{\text{actual}} = 80$

$\bar{x}_{\text{estimate}} = 85$

$[0.902, \ldots]$

correlation coeff

S.E

significance (p-value)

① A	② B
17	200
97	100
30	150

σ_1, σ_2

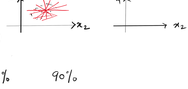
80 → 87

$(87 - 80) = 7$

acceptable / reject

ML

$\mu_{\text{ML}} = 10$



$(\bar{y} - \bar{y})$

95% 90%

87 ← 95 → 87

87 → X

5%

90-100 87X significance A B