Covariance – Quantifying a relationship between variables. Direction of a relationship.

Size	Price
1200 sqft	₹1L
1800 sqft	₹2L
2500 sqft	₹3L

Quantify a relationship between Size and Price.

Size ↑ Price ↑

Size ↓ Price ↓

$$\underset{x\ \ y}{cov(size, price)} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$\text{Variance}(x),\ var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})*(x-\bar{x})$$

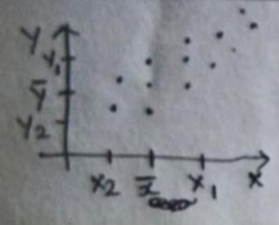
$$Cov(x,x) = var(x)$$

So, through covariance we will get a value,

if $x\uparrow, y\uparrow$ = ☐ +ve. covariance

$x\uparrow, y\downarrow$ = ☐ -ve. covariance

Suppose $x\uparrow\ y\uparrow$.

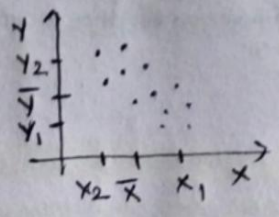


$= (x_1-\bar{x})*(y_1-\bar{y})$
$= (+ve)*(+ve)$
$= +ve.$

$= (y_2-\bar{y})*(x_2-\bar{x})$
$= (-ve)*(-ve)$
$= +ve.$

So, always covariance is +ve when $x\uparrow, y\uparrow$.

Suppose $x\uparrow, y\downarrow$



$= (x_1-\bar{x})*(y_1-\bar{y})$
$= (+ve)*(-ve)$
$= -ve.$

$= (x_2-\bar{x})*(y_2-\bar{y})$
$= (-ve)*(+ve)$
$= -ve.$

So, always covariance is -ve when $x\uparrow\ y\downarrow$.

Covariance find the direction of relationship. But we don't know exact value of strength.

Pearson Correlation Coefficient → Strength of the relationship between variables and direction also of the relationship.

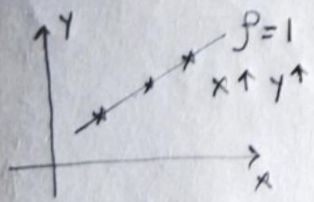
$$P(x,y) = \frac{Cov(x,y)}{\sigma_x\,\sigma_y}$$

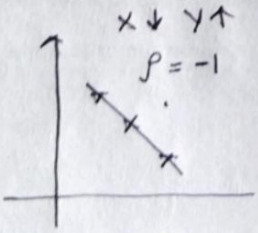
Range is → $-1 \le P(x,y) \le 1$.

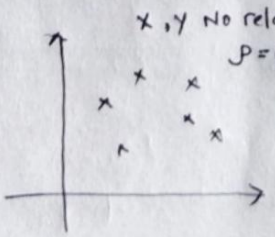
$Cov(x,y)$ → covariance (x,y)

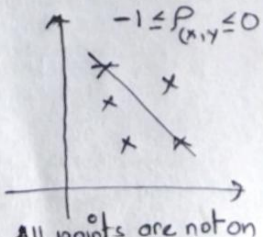
σ_x → Standard deviation of x.

σ_y → SD of y.





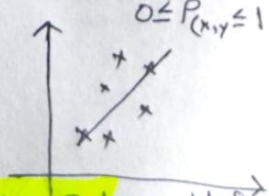




All points are not on straight line.

Suppose we have 3 variables a, b, c.
a and b have pearson correlation coefficient of 1. In short they are same, then we will remove one of the variable/feature.

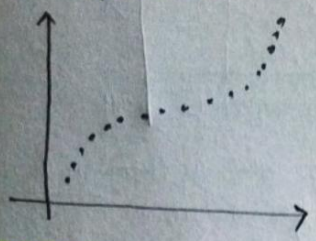
We also use spearman rank Correlation (Refer to Numerical Measure-Mean)



Pearson work only good for linear relationship. But for others we use spearman rank correlation coefficient.
(Refer to page 14)

Spearman's Rank Correlation coefficient

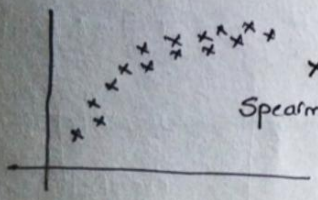
If the plot is non-linear relationship.



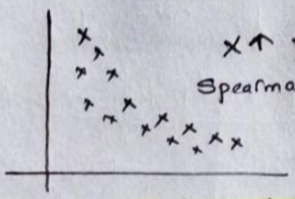
Spearman correlation = 1.
Pearson correlation = 0.88

$$\text{Pearson} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} \qquad \text{Spearman} = \frac{\text{Cov}(\text{rank}_X, \text{rank}_Y)}{\sigma_{\text{rank}_X} \sigma_{\text{rank}_Y}}$$

- Pearson correlation assesses only linear relationship. whereas Spearman correlation assesses monotonic relationship (whether linear or not).
- Spearman correlation is from range -1 to 1.
- If the two variables are very similar, then spearman correlation will be 1.
- If the two variables are dissimilar / fully opposed, then spearman correlation will be -1



$x \uparrow$ $y \uparrow$.
Spearman correlation = 0.92



$x \uparrow$ $y \downarrow$.
Spearman correlation = -0.92

$x \rightarrow$ Independent Var
$y \rightarrow$ Dependent Var.

Sign of Spearman correlation indicates the direction of association between x and y

If Spearman correlation = 0, then there is no tendency of y increase or decrease when x is increased.

Steps in the methods are - i) Sort the data and assigned the rank.
ii) Sort the second column and assign the rank.
iii) Create a difference column based on rank
iv) Create final column based on difference by squaring

IQ	Hours of TV.	Sort → IQ (x_i)	Hours of TV (y_i)	rank (x_i)	rank (y_i)	d_i	d_i^2
106	7	86	0	1	1	0	0
100	27	97	20	2	6	-4	16
86	2	99	28	3	8	-5	25
101	50	100	27	4	7	-3	9
99	28	101	50	5	10	-5	25
103	29	103	29	6	9	-3	9
97	20	106	7	7	3	4	16
113	12	110	17	8	5	3	9
112	6	112	6	9	2	7	49
110	17	113	12	10	4	6	36

$\sum d_i^2 = 194$.

$n = 10$.

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

Shortcut formula.

$$= 1 - \frac{6(194)}{10(10^2-1)}$$

$= -0.175$.

So, $\rho = -0.175$ which is very close to 0. So there is no correlation between IQ and watching TV for hours.