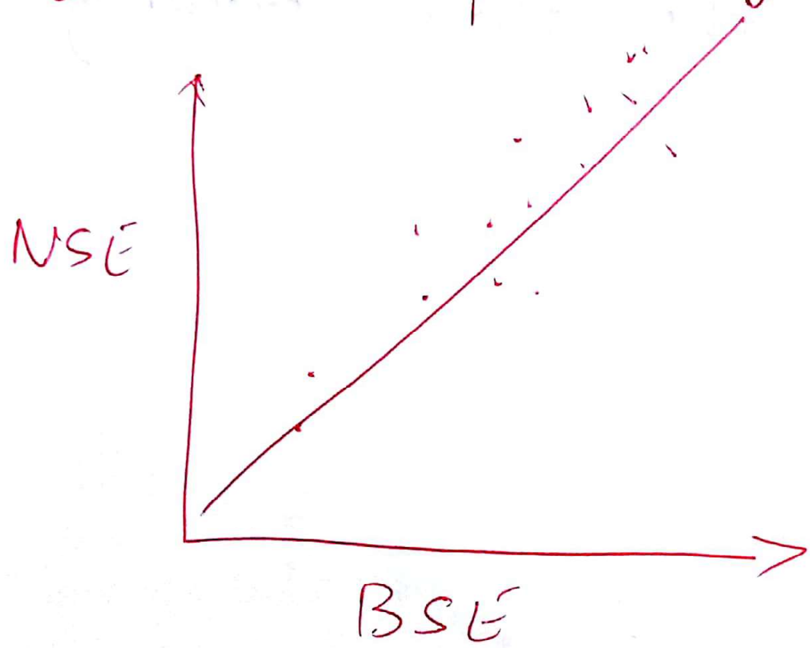
Covariance and correlation

→ They form a part of bivariate relationship analysis.



→ Look at the data points
↳ Pattern looks like linear relationship

→ When one rises, other rises in the same way.

→ In this case, relationship is strong.

In other words, there is a strong positive relationship.

We can quantify this relationship via covariance i.e how they co-vary.

Note: Covariance is only one measure to quantify this idea.

Others include: correlation.

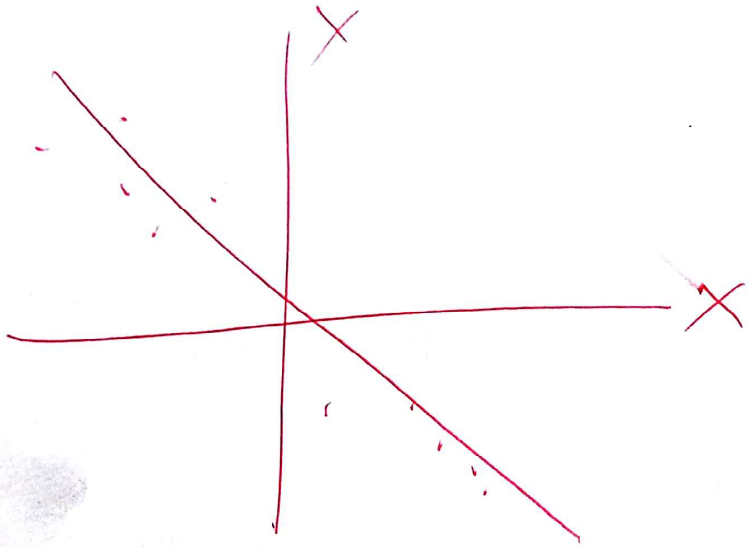
Covariance: A descriptive measure of the relationship b/w two variables.

Emphasis is on <u>direction</u>.

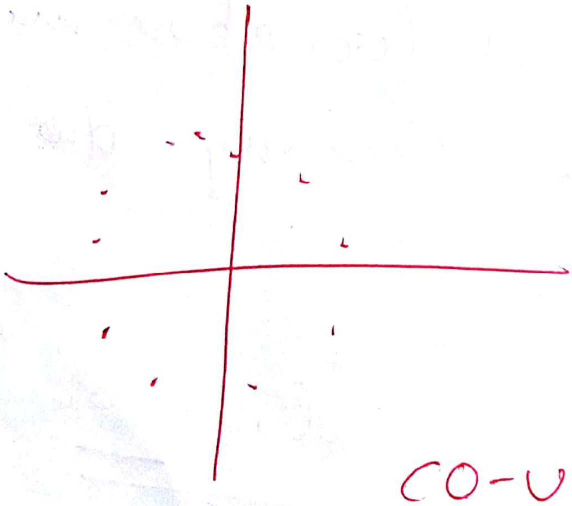
+ve indicates increasing R/p

-ve " decreasing R/p

Note Covariance is used to identify linear relationship



-ve covariance



co-variance = 0

$$cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Co-variance:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

for sample

$$\sigma_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \cancel{\mu_x})(y_i - \mu_y)}{n}$$

For population.

Example: x: workers — 10 random samples from new building

y: no of walls painted

$\bar{x} = 21.3$ $\bar{y} = 41.2$

x ($s=648$)	y ($s=1669$)	$x_i - \bar{x}$	$y - \bar{y}$	$\sum(x_i - \bar{x})(y_i - \bar{y})$
12	20	-9.3	-21.2	197.16
30	60	8.7	18.8	
15	27	-6.3	-14.2	
24	50	2.7	8.8	
14	21	-7.3	-20.2	
18	30	-3.3	-11.2	
28	61	6.7	19.8	
26	54	4.7	12.8	
19	32	-2.3	-9.2	
27	57	5.7	15.8	

$\sum = 962.4$

$$Cov(x,y) = S_{xy} = \frac{962.4}{10-1}$$

$$= 106.93$$

Positive linear relationship.

Focus only on sign

Correlation

Covariance provides the direction

Correlation provide direction and strength.

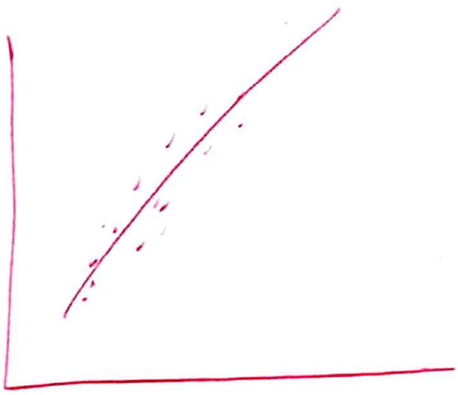
Covariance → No upper and lower bound.

Correlation → Between (-1, 1).

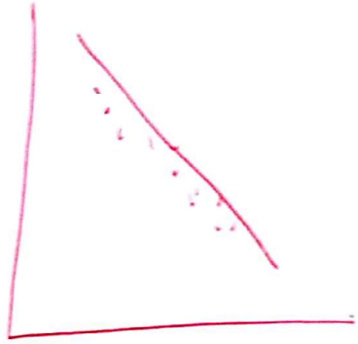
Covariance is dependent on scale

Correlation is independent of scale.

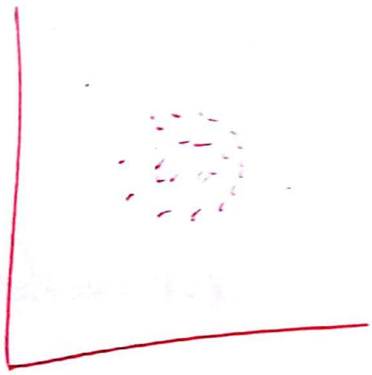
Drawback: Covariance and corr... are applicable to lin... relationships.



Near +1



Close to −1



Close to 0.

Correlation Coefficient (r)

$$r_{xy} \, (\text{or } \rho_{xy}) = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}$$

$$cov(X,Y) = E[(X-\mu)(Y-\mu)]$$

From wikipedia.

If

(x_i, y_i) (for $i = 1$ to n) can be realized with probability p_i, then

$$cov(X,Y) = \sum_{i=1}^{n} p_i (x_i - E[X])(y_i - E[Y])$$

$$cov(X,a) = 0$$

$$cov(X,X) = var(X)$$

$$cov(X,Y) = cov(Y,X).$$

$$cov(aX,bY) = ab\,cov(X,Y)$$

$$cov(X,Y) = E[XY] - E[X]E[Y].$$

$$\rho_{xy} = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

$$= \frac{106.93}{6.48 \times 16.69} = 0.989.$$

<u>Rule of thumb</u>

if $|r| \geq \frac{2}{\sqrt{n}}$.

then relation exists

<u>Not Universal</u>

Important point:

Correlation does not imply causation.

$X = x_1 \; x_2 \; x_3$

$cov(X) \rightarrow$ covariance matrix

	X_1	X_2	X_3
X_1	$var(X_1)$	$cov(X_1, X_2)$	$(cov\, X_1, X_3)$
X_2	$cov(X_2, X_1)$	$Var(X_1)$	$cov(X_2, X_3)$
X_3	$cov(X_3, X_1)$	$cov(X_3, X_2)$	$Var(X_1)$

Suppose we have a random variable $X \sim N(\mu, \sigma^2)$.

To convert this normal distribution into standard normal, we follow the standardizing procedure. i.e.

$$Z = \frac{X-\mu}{\sigma}.$$

Here $Z \sim N(0,1)$.

E.g. $X \sim N(162.2, \sigma = 6.8)$

Say what is the prob. that an average girl is taller than 170.5

$P(X > 170.5)$

$$P\left(\frac{X-162.5}{6.8} > \frac{170-162.5}{6.8}\right)$$

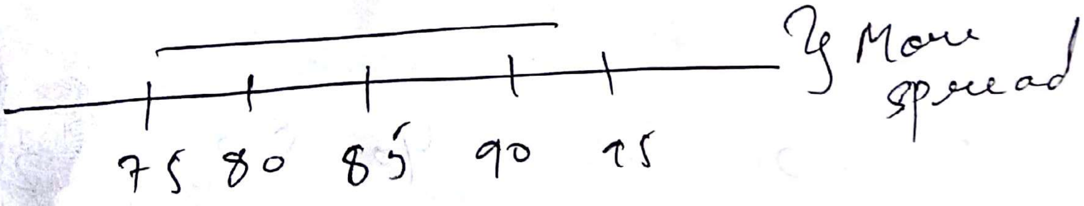
$P(Z > 1.22)$.

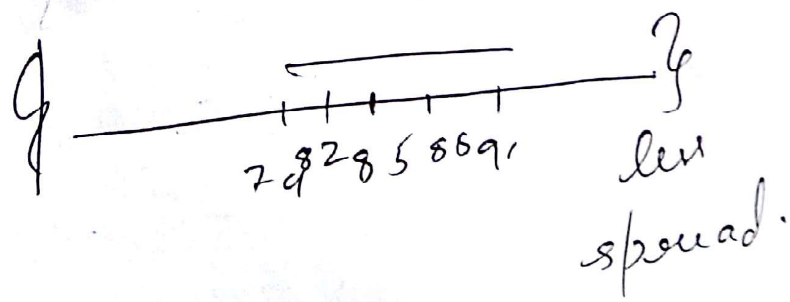
Discussion on scaling z-table

z-score.

Suppose, we have the following data

85	88
95	79
75	91
80	85
90	82
$\mu = 85$	$\mu = 85$





z-score is a statistical measure that describes how far a given data point is from the mean.
↳ It describes the distance in term standard deviation.

Q) Suppose a data sample has mean 50 and σ of 5. A value of 55 will have z-score of 1. and 40 will have z-score of -2

$$z = \frac{x - \mu}{\sigma}$$

$$= \frac{55 - 50}{5} = 1$$

$$z = \frac{x - \mu}{\sigma} = \frac{40 - 50}{5} = -2.$$

Q) Suppose in a test a person scored ~~500~~ 630 marks. Mean marks is 500, sd is 150.

$$z = \frac{x-\mu}{\sigma} = 630$$

$$= \frac{630-500}{150} = 0.83$$

Q) ~~US~~ ICSE board a person ~~had~~ got 87% with others having on average 80% with S.D. of 5.

CBSE board a person got 82% with class avg. 73 and s.d. of 6

$$z_1 = \frac{87-80}{5} \quad ; \quad z_2 = \frac{82-73}{6}$$

$$z_1 = 1.4 \quad ; \quad z_2 = 1.5$$

Scanned by CamScanner

Suppose IQ of people follow normal distribution with $\mu = 100$ and $\sigma = 15$.

What % of people are i) stupid

ii) smart

iii) average

stupid : ~~8~~50 or less

smart : 140 or more

avg : 50~~[illegible]~~ - 140

~~&~~ Stupid $= \dfrac{50-100}{15} = -\cancel{4}3.333$

$\cancel{0.005}\%$ 0.004

$= \cancel{0.5}\%$ 0.4%

smart : $\dfrac{140-100}{15} = \dfrac{40}{15} = \cancel{3}2.66$

$= .9961$

$1 - 0.9961$

$.0039$

0.4%

Scanned by CamScanner