Welcome back everyone

We will be having an extra class

on May 6

Monday

7 pm – 10 pm

First 1 hr ⇒ doubt clearing

then teaching.

Measure of Central Tendency

A measure of central tendency is a descriptive statistic that describe the average or typical value for set of observ. / events / data.

eg. Age of country population
Placement stats of college.

Mean

mean is defined as

⇒ arithematic average of all observation

$\Sigma X_i / N$

N = no. of observation

Σx_i — sum of all observation.

mean of sample $\bar{x}$, X

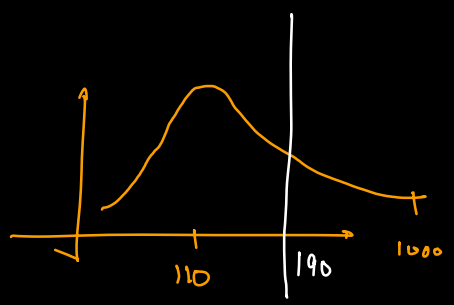
mean of population μ

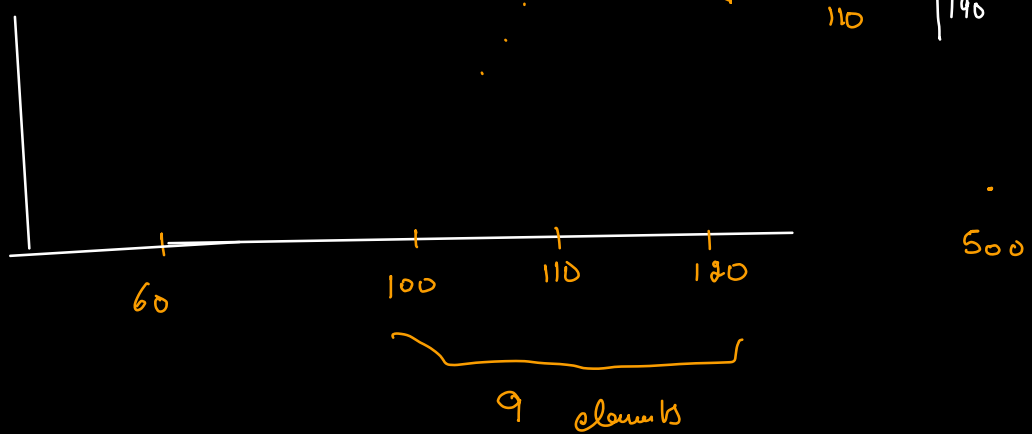
⇒ 100, 102, 105, 109, 110, 111, 115, 118, 120,

1600

mean = 110 ⟶ 190

Skewness in data





-ve skewness

+ve skewness

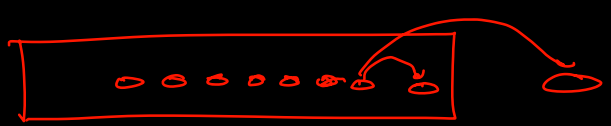
⇒ mean is sensitive to every data point

⇒ If you change one point in data, the mean will also change.

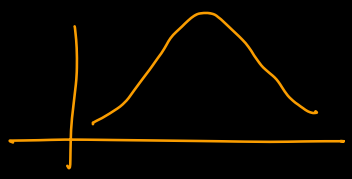
⇒ We should avoid using mean on skewed data.

⇒ We normally/mostly use mean for Quantative variable.

every point ← outlier is also a point

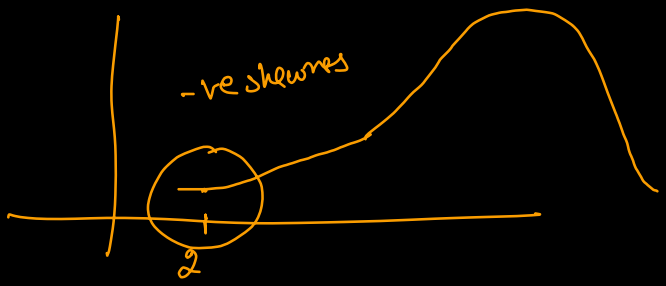


-ve shewness

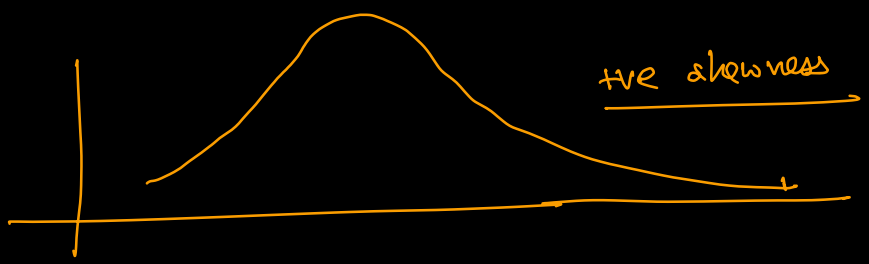


+ve shewness

2 22, 23, 25, 27, 28



80



Median (50th percentile)

middle most value

half the point towards left | half the point towards right

100 102 105 108 110 112 115 118 120 1000

1 2 3 4 5 6 7 8 9 10

N = 10

$\frac{X_5 + X_6}{2} = 111$

1st] Get all the values

2nd] Arrange in to ascending / descending order

3rd] If no. of values odd ⇒ middle most

even ⇒ $\frac{X_{N/2} + X_{N/2+1}}{2}$

1 2 3 4 N = 4

$\frac{2+3}{2} = 2.5$ $\frac{X_2 + X_3}{2}$

110 ⟶ 111 , median change

↘ 90 mean change

⇒ median is not sensitive to each & every point.

& it is often used when our data has outliers.

& can be used when data is +vely or -vely skewed.

& few really small value [-ve] or few very large values [+ve] will not overly influence the median.

So, whenever our data is skewed or have outliers, we will consider median.

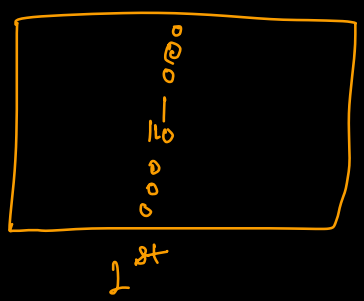
Mode

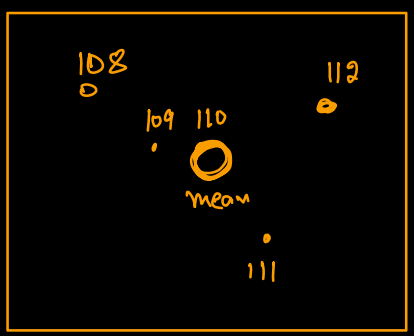
most occuring value / most frequent value

used mainly for categorical data.

Variance

⇒ Variance measures how far a set of values are spread out out from their mean.





⇒ Variance is calculated by taking diff b/w each point in the data set with its mean, squaring the diff. [to make it +ve] & then dividing it by total no. of values.

$$\sigma^2 = \text{Variance} = \frac{\Sigma (x - \mu)^2}{N}$$

105 108 110 112 115

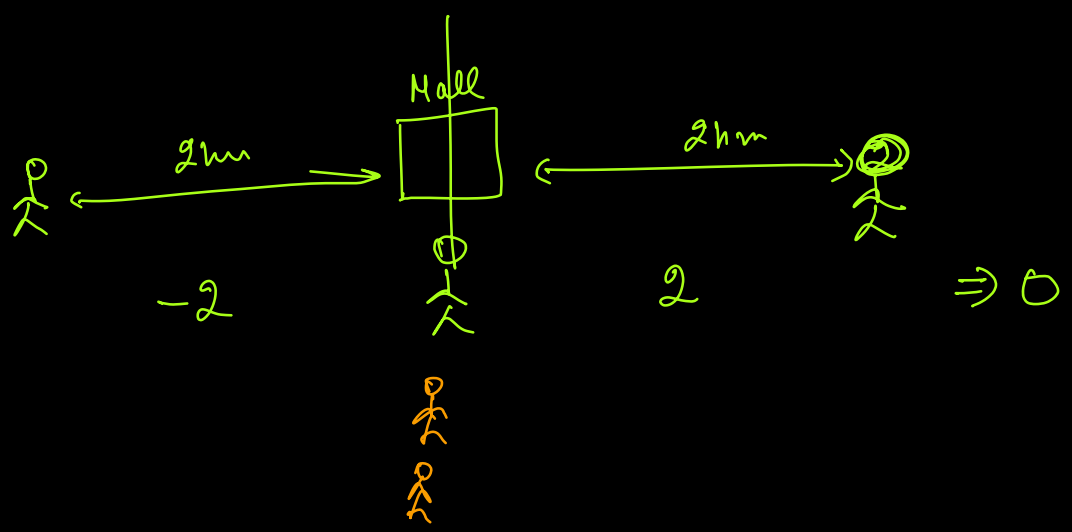
$\bar{x} = 110$

$$\sigma^2 = \frac{[105-110]^2 + [108-110]^2 + [110-110]^2 + [112-110]^2 + [115-110]^2}{5}$$

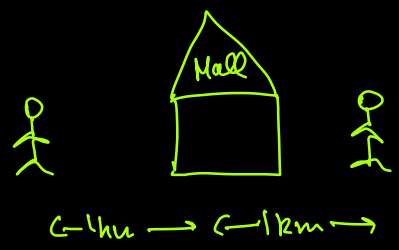
without squaring

$$= \frac{-5 + -2 + 0 + 2 + 5}{5}$$

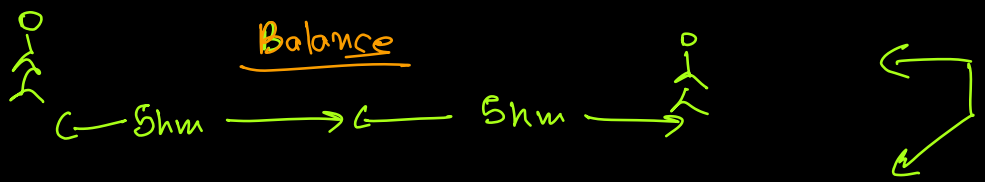
what if each value was 110? $\sigma^2 = 0$

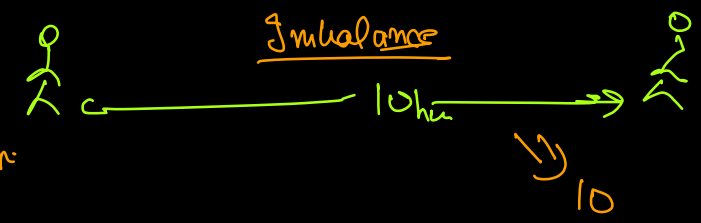


1st example

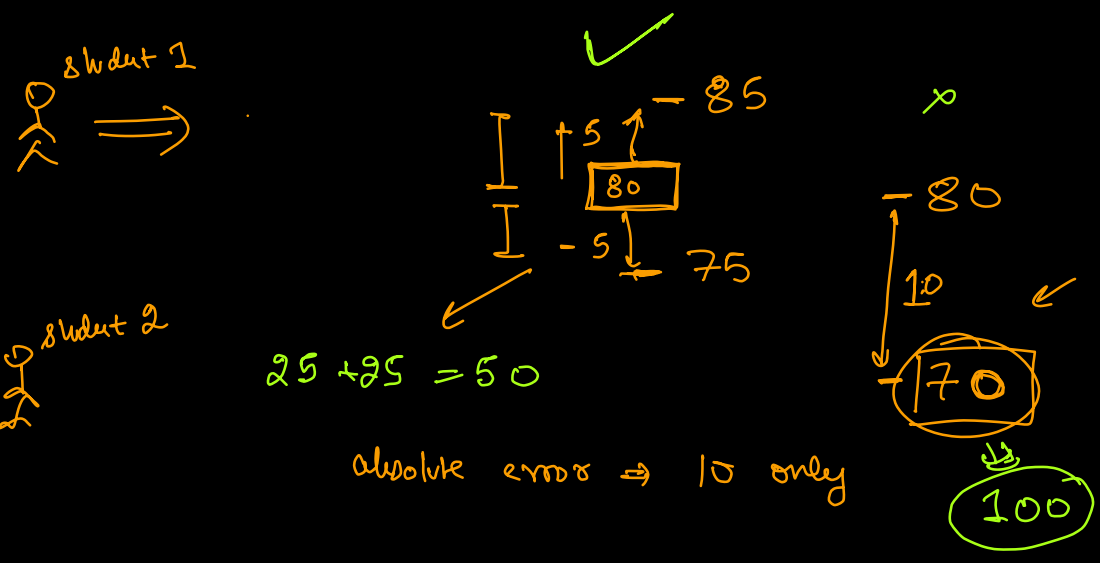


⇒ 10



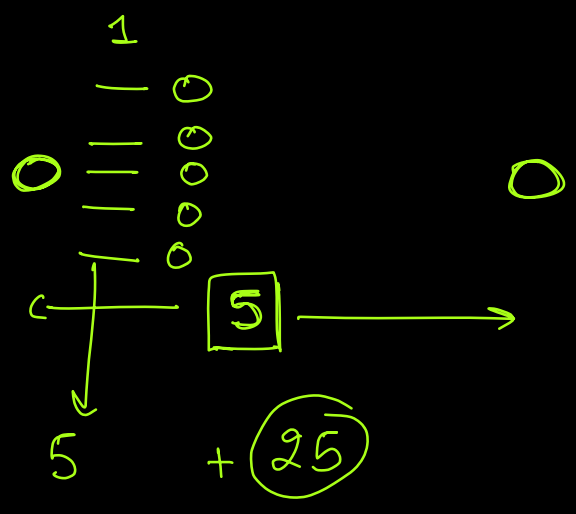


good example.



50 90

50 10


1
5
5
+ 25

Whether in our data, we should have more variance or less variance.

1st	2nd
100	10 →
100	80
100	90
100	100
100	110
100	180
100	190

If we have high variance in our data, we have more information in our data

more information is required

$$\sigma^2 = \frac{\Sigma (x_i - \bar{x})^2}{N-1} \quad \{\text{sample data}\}$$

⇒

homework

Why when we work on sample data we sometime take $n-1$ instead of n in variance formula?

Standard Deviation [σ]

It is the square root of variance

$$\sigma = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{N-1}}$$

n x

N μ

tells us how much data deviates from mean.

low σ $\Rightarrow$ data points are closed to mean

high σ $\Rightarrow$ data points are spread away from mean.

Why both σ^2 and σ are needed?

Variance and s.d. both serve similar/ some purpose.

both are measure of spread / dispersion

but they differ in scale?

cm^2 $\quad$ cm

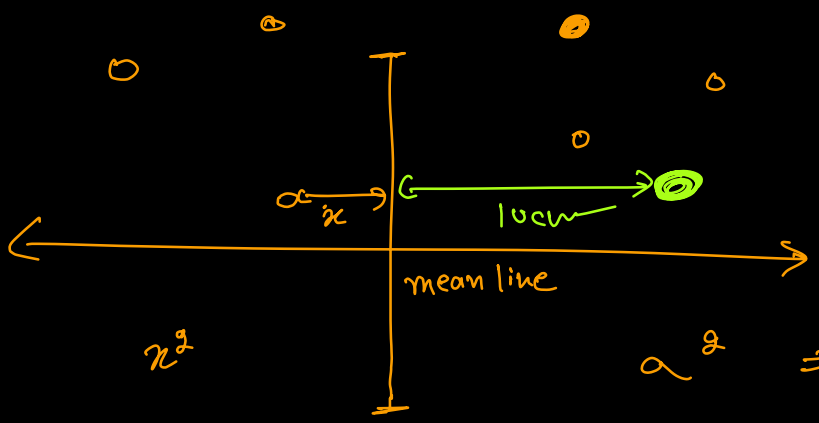
area $\Rightarrow m^2$ $\quad$ m

var $\sigma_A^2 \longleftrightarrow \sigma_B^2 \Rightarrow$ comparison

Class A $\quad$ Class B

σ_A 5cm

7cm



Variance calculating overall spread.

$100\ cm^2$

$10cm = a$ "avg" dist of each student height from avg height

Q :- Find mean, median, std., variance for these columns?

Road show Venue	Sales Team	No of Roadshow	No of new sign ups	*No of rejected application	No of successful sign ups	% of successful Sign ups
In house American Express Branch	A	24	580	19	561	97%
Club house or pubs	A	24	440	50	390	89%
CBD Area Road show	B	24	630	53	577	92%
Door to Door to offices with free gifts	B	24	560	42	518	93%
Shopping malls in town area	C	24	610	280	330	54%
Heartlands malls	C	24	320	140	180	56%
Total	**3 teams**	**144**	**3140**	**584**	**2773**	**80%**

Covariance

⇒ Co-variance of 2 variable x and y in a data sample measures how much 2 variables are linearly related

⇒ is a measure of how much 2 variables change together

⇒ Covariance of 2 var x and y.

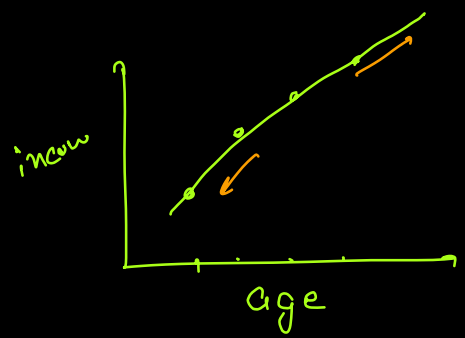
$$Cov(x, y) = E\left[X - \underbrace{E(x)}_{\text{mean}} \right] (Y - E(Y))$$

+ve covariance ⇒ positive relationship

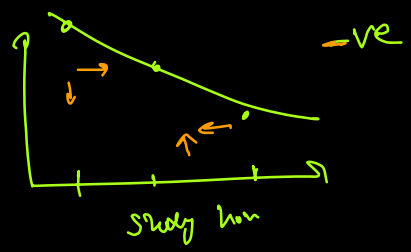
⇒ inc or dec in one would cause inc. or dec. in other.

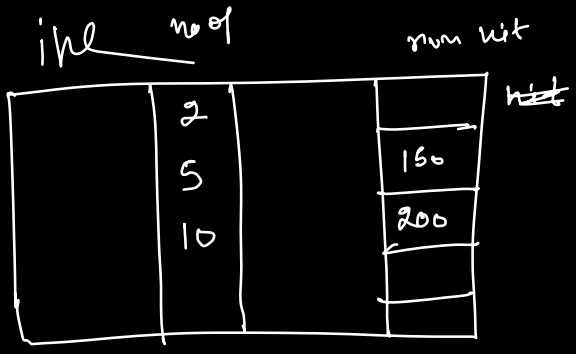
Age & income

+ve covar

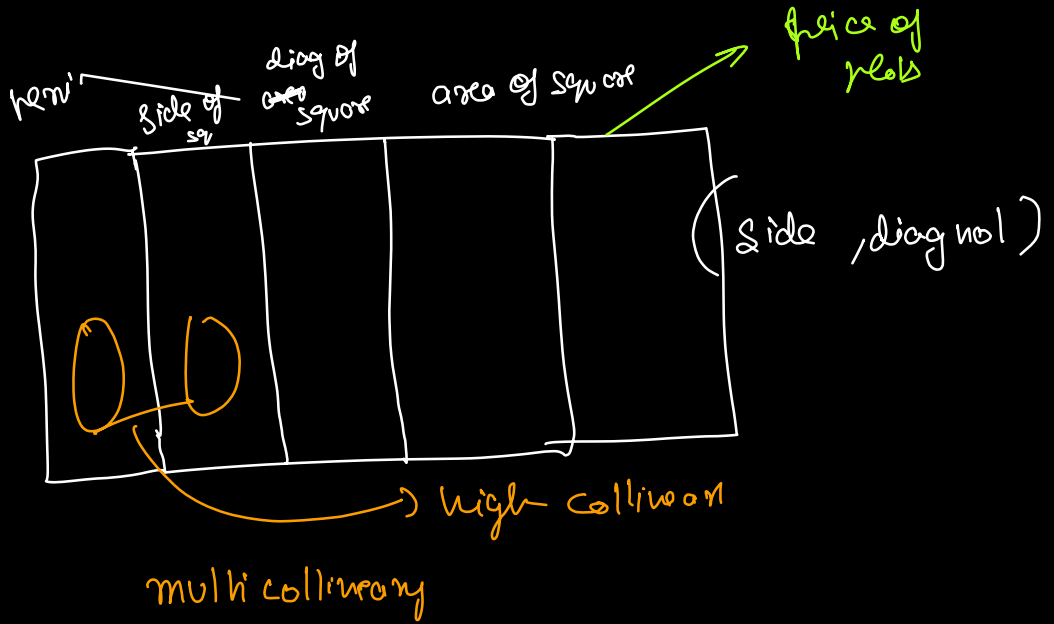


-ve corr hours & chance of fails





hardly we will be using it.



Student	Study Hours (X)	Test Score (Y)
Alice	3	90
Bob	5	85
Charlie	4	88
David	2	92
Emma	1	95

1. Calculate the means of the study hours ($\bar{X}$) and test scores ($\bar{Y}$):

$\bar{X} = \frac{3+5+4+2+1}{5} = \frac{15}{5} = 3$ hours

$\bar{Y} = \frac{90+85+88+92+95}{5} = \frac{450}{5} = 90$ Score

2. Calculate the covariance:

$$\text{Cov}(X,Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

$$\text{Cov}(X,Y) = \frac{(3-3)(90-90)+(5-3)(85-90)+(4-3)(88-90)+(2-3)(92-90)+(1-3)(95-90)}{5}$$

$$\text{Cov}(X,Y) = \frac{0+2*(-5)+1*(-2)+(-1)*2+(-2)*5}{5}$$

$$\text{Cov}(X,Y) = \frac{-10-2-2-10}{5}$$

$$\text{Cov}(X,Y) = \frac{-26}{5}$$

$$\text{Cov}(X,Y) = -5.2$$ hours Score

age (yrs)	income (₹)	Years of expo. (years)
23	10000	3
29	2 L	5
30	~~2~~ 1.5 L	6
35	2.5 L	8

Cov (age, income) =

Cov (age, experience) =

Age (X)	Income (Y)	Years of Experience (Z)
30	$50,000	5
35	$60,000	8
40	$70,000	10
25	$45,000	3
45	$80,000	12

$$\text{Cov}(X,Y) = \frac{(30-35)(50000-61000)+(35-35)(60000-61000)+(40-35)(70000-61000)+(25-35)(45000-61000)+(45-35)(80000-}{5}$$

$$\text{Cov}(X,Y) = \frac{(-5)(-11000)+(0)(-1000)+(5)(9000)+(-10)(-16000)+(10)(19000)}{5}$$

$$\text{Cov}(X,Y) = \frac{55000+0+45000+160000+190000}{5}$$

$$\text{Cov}(X,Y) = \frac{450000}{5} = 90000$$

$$\text{Cov}(X,Z) = \frac{\sum (X_i-\bar{X})(Z_i-\bar{Z})}{n}$$

$$\text{Cov}(X,Z) = \frac{(-5)(-2.6)+(0)(0.4)+(5)(2.4)+(-10)(-4.6)+(10)(4.4)}{5}$$

$$\text{Cov}(X,Z) = \frac{13+0+12+46+44}{5}$$

$$\text{Cov}(X,Z) = \frac{115}{5} = 23$$

Covariance takes in the unit of my variable.

Co-relation ⭐⭐

Statistical quantity which tells us how strongly 2 variables are related to each other

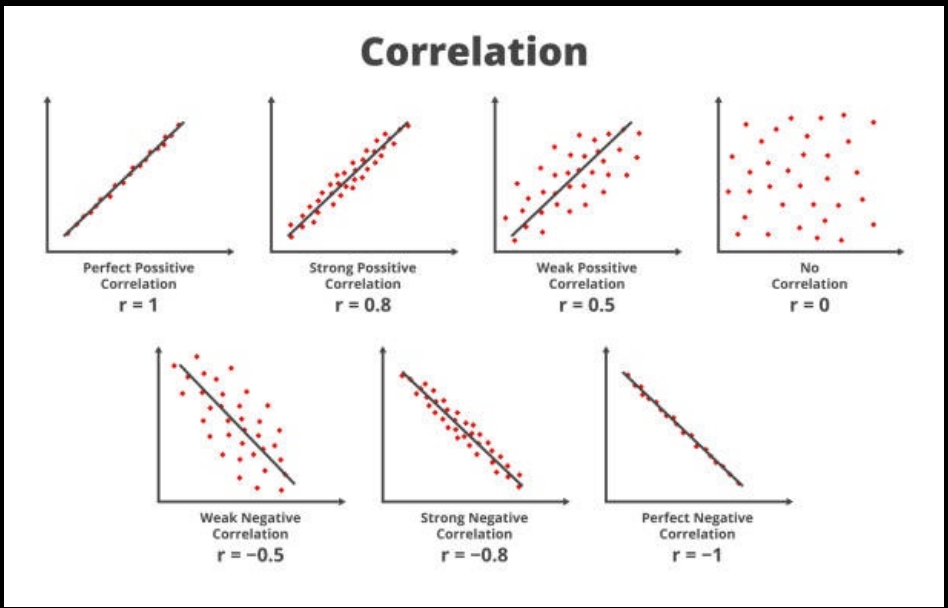
$$Corr(X,Y) = \frac{Cov(X,Y)}{S.d.(X)\ S.d.(Y)}$$

$$-1 < Corr(X,Y) < 1$$

It is a scaled down version of covariance.

+ve corr = increase in x ⇒ inc in y

-ve corr = inc. in x ⇒ dec in y


Correlation
Perfect Possitive Correlation
r = 1
Strong Possitive Correlation
r = 0.8
Weak Possitive Correlation
r = 0.5
No Correlation
r = 0
Weak Negative Correlation
r = −0.5
Strong Negative Correlation
r = −0.8
Perfect Negative Correlation
r = −1

Q = Calculate Co-relL & co-vas.
b/w X, Y, Z and stress.

Participant	Study Hours (X)	Exercise Hours (Y)	Socializing Hours (Z)	Stress Level
1	3	2	1	7
2	5	1	2	6
3	4	3	3	8
4	2	4	2	7
5	6	2	1	5
6	3	1	4	9
7	4	2	2	6
8	5	3	3	8
9	2	4	1	7
10	6	1	3	5
11	3	3	2	7
12	4	2	1	6
13	5	4	4	8
14	2	1	2	6
15	6	3	3	8
16	3	2	1	7
17	4	4	2	5
18	5	1	3	9
19	2	3	2	6
20	6	2	4	7