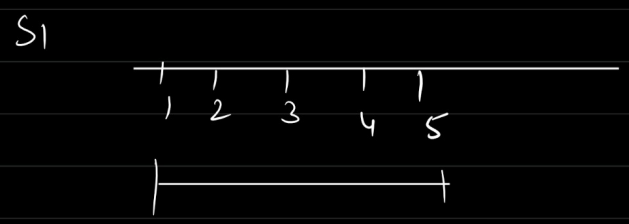
* Measures of Dispersion

S_1 – 1, 2, 3, 4, 5

mean/median – 3

S_2 = 3, 3, 3, 3, 3

mean/median = 3



* How the data is spread?

→ Range
→ Percentage and percentile
→ Quartiles (Boxplot)
→ Variance
→ Standard deviation.

* Range — difference between maximum and minimum value.

{1, 2, 3, 4, 5}

Range = 5 – 1 = 4

{1, 2, 3, 4, 1000}

Range = 1000 – 1 = 999

* Outlier affects the range.

* Percentage and percentiles

1, 2, 3, 4, 5

What is the percentage of nos that are odd?

$$\frac{3}{\cancel{5}} \times \frac{20}{\cancel{100}} = 60\%$$

* Percentile

Defⁿ – A percentile is a value below which a certain percentage of observation lie.

{1, 2, 3, 4, 4, 6, 7, 7, 8, 10}

What is the percentile rank of 3 ?

$$\text{Percentile rank of a no} = \frac{\text{No. of values below that no}}{\text{Total nos } (n)} \times 100$$

$$= \frac{2}{10} \times 100 = 20^{th} \text{ percentile}$$

what value exists at 75^{th} percentile?

$$\text{Value} = \frac{\text{Percentile}}{100} \times (n+1)$$

$$= \frac{75}{100} \times (10+1)$$

$$= \frac{3}{4} \times 11 = 8.25^{th} \text{ number}$$

$\Downarrow$

8^{th} number

(8.5^{th})

$\Downarrow$

avg of 8^{th} & 9^{th} no

75^{th} percentile – 7.

→ 75^{th} percentile is 8

↳ It means 75% of the no in the data is equals to or below 8.

* Quartile

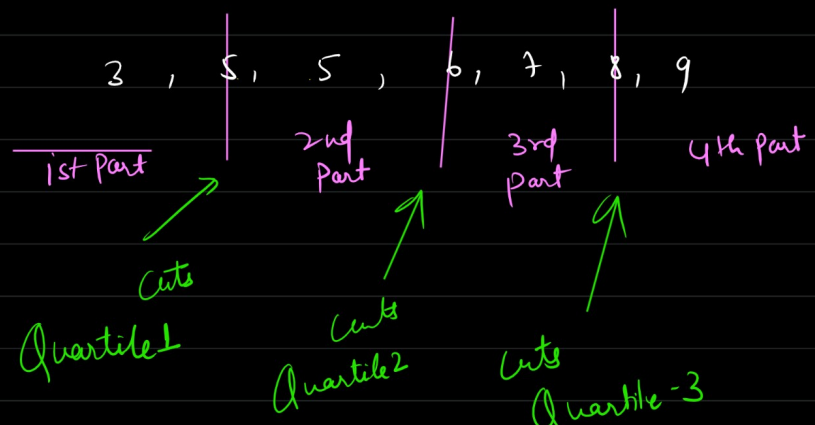
→ Quartiles are values that divides a list of numbers into quarters.

* Put the no in order
* then cut the number into 4 equal parts
* The quartiles are at the cut.

Ex. 6, 8, 5, 5, 7, 3, 9

order - 3, 5, 5, 6, 7, 8, 9

Cut the no into quarters.



$Q1 \rightarrow 5$

$Q2 \rightarrow 6$

$Q3 \rightarrow 8$

Ex - 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 4

total - 11 nos. Q_1 Q_2 Q_3

if total no is odd

- $Q_1 = \frac{n+1}{4}^{th} = \frac{11+1}{4}^{th} = \frac{12}{4}^{th} =$ 3rd no. $Q_1 = 3$
- $Q_3 = \frac{3(n+1)}{4}^{th} = \frac{3(11+1)}{4}^{th} = \frac{3\times12}{4} = 9^{th}$ no $Q_3 = 3$
- $Q_2 = \left(\frac{n+1}{2}\right)^{th} = \frac{11+1}{2}^{th} = 6^{th}$ no. $Q_2 = 2$ (median)

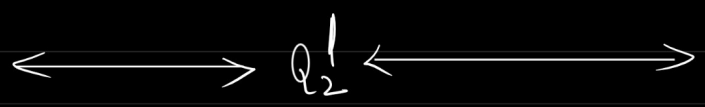
if total no is even

$Q_1 = \frac{n}{4}^{th}$ no

$Q_3 = \frac{3n}{4}^{th}$ no

$Q_2 = \frac{\frac{n}{2}^{th} + \left(\frac{n}{2}^{th}+1\right)^{th}}{2}$

1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 4



Median - the point that divides the data into two equal parts

* even no. of nos

1, 2, 3, 3, 4, 4 $\quad N = 6$

$Q_1 \rightarrow \frac{6}{4}^{th} = \frac{3}{2} \Rightarrow 1.5^{th} \Rightarrow \frac{1+2}{2} = \frac{3}{2} = 1.5$ (Q_1)

$Q_3 \rightarrow \frac{3n}{4}^{th} = \frac{3 \times 6}{4}^{th} = \frac{18}{4}^{th} = 4.5^{th}$

4^{th} & 5^{th} no

$Q_3 = \Rightarrow \frac{3+4}{2} = 3.5$

Q_2 (median when the total nos are even)

$\text{avg}\left(\frac{n}{2}^{th} \text{ no}, \frac{n}{2}^{th} + 1\right)$

$\text{avg}\left(\frac{6}{2}^{th}, \frac{6}{2}^{th} + 1\right)$

$\text{avg}\left(3^{rd}, 4^{th}\right)$

$= \frac{3+3}{2} \Rightarrow 3 \quad Q_2 = 3$

Five point summary

Q_0 (min) $- 1 \longrightarrow 0\% - 0^{th}$ Percentile

$Q_1 - 1 \longrightarrow 25\% - 25^{th}$ percentile

(median) $Q_2 \longrightarrow 2 \longrightarrow 50\% - 50^{th}$ percentile

$Q_3 - 3 \longrightarrow 75\% - 75^{th}$ percentile

Q_4 (maximum) $- 4 \longrightarrow 100\% - 100\%$ Percentile

transaction amount

1000
2000
3000
.
.
.

min = 1000 (Q_0)

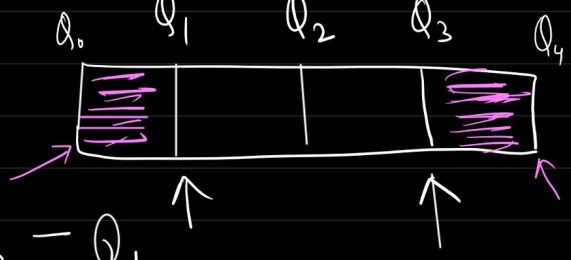
$Q_1 = 25^{th}$ percentile

↳ 5000 → 25% of transaction amount is equals to or below 5000 in the data

$Q_2 = 10000$ → 50% of transaction is equals to or below 10000

Q_3 – 75th percentile
Q_4 – max no.

Inter Quartile range



$IQR \Rightarrow Q_3 - Q_1$

2, 4, 4, 5, 6, 7, 8

Q_1 = 4, Q_3 = 7

$IQR = 7 - 4 = 3$

{2, 3, 3, 3, 3, 4, 4, 5, 5, 5, 6, 6, 6, 7, 8, 99} → Outlier.

Q_1, Q_3, $N = 16$

$Q_1 = 25^{th}$ percentile $= \frac{25}{100} \times 16^{th} = \frac{1}{4} \times 16^{th} = 4^{th}$ no

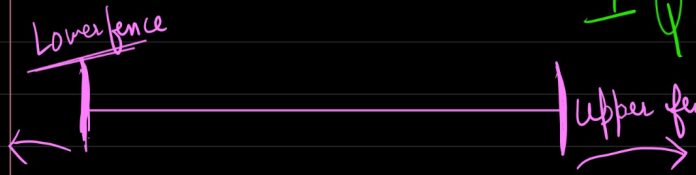
$Q_3 = 75^{th}$ percentile $= \frac{75}{100} \times 16^{th} = \frac{3}{4} \times 16^{th} = 12^{th}$ no.

Q_2 = avg of 8^{th} & 9^{th} no $= 5$

$IQR = 6 - 3 = 3$

$(Q_3) - (Q_1)$

Outliers are extremes value



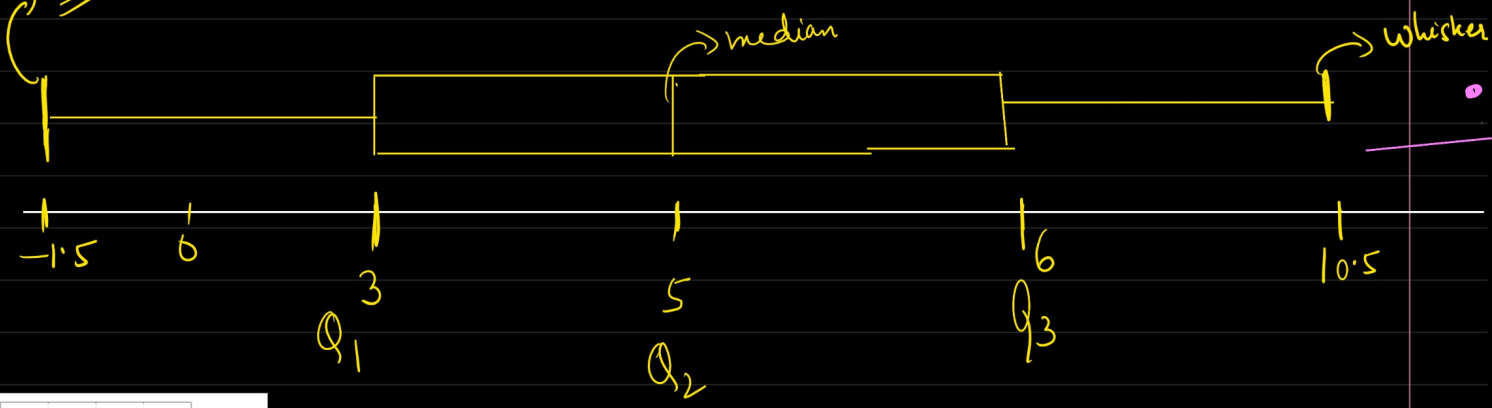
$$\text{Lower fence} = Q_1 - 1.5 \times IQR$$
$$\text{Upper fence} = Q_3 + 1.5 \times IQR$$

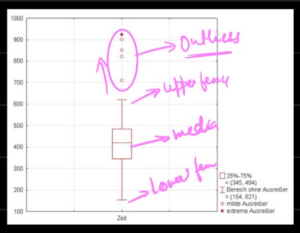
$L.F = 3 - 1.5 \times 3 = -1.5$

$U.F = 6 + 1.5 \times 3 = 4.5 + 6 = 10.5$

Box – whisker plot

{2, 3, 3, 3, 3, 4, 4, 5, 5, 5, 6, 6, 6, 7, 8, 9}

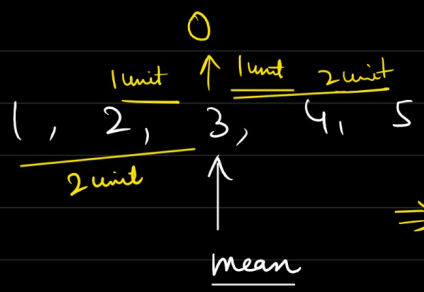




Measures of Spread / dispersion

1. Variance
2. Standard deviation

* Mean- deviation



$$\Rightarrow \frac{2+1+0+1+2}{5}$$

$$\Rightarrow \frac{6}{5} = 1.2$$

→ On an avg each of the data is 1.2 units away from mean value

* Variance — the average of the squared differences from the mean.

Population Variance

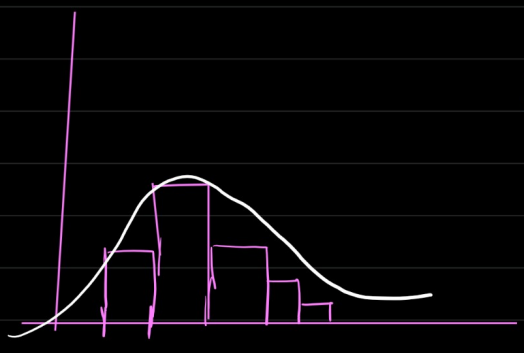
$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

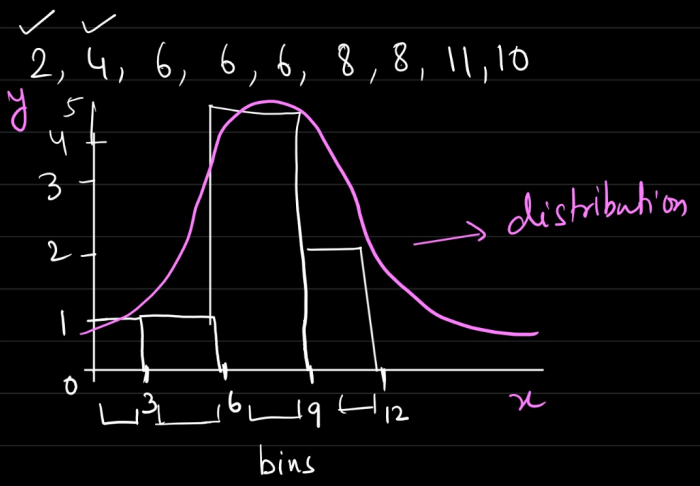
μ → population mean

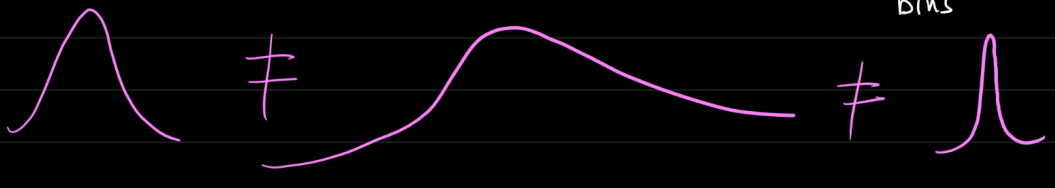
Sample variance

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

$\bar{x}$ → Sample mean







data = {1, 2, 3, 3, 4, 4}

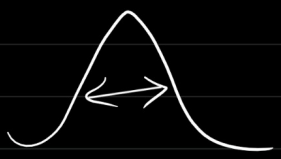
How to Calculate Variance

→ Calculate mean
→ for each no in data, subtract the mean and the no
→ Square of difference
→ Calculate the avg of square of difference.

x	$\bar{x}$	$x-\bar{x}$	$(x-\bar{x})^2$
1	2.83	-1.83	3.34
2	2.83	-0.83	0.68
3	2.83	0.17	0.03
3	2.83	0.17	0.03
4	2.83	1.17	1.37
4	2.83	1.17	1.37
2.83			6.82

$$s^2 = \frac{6.82}{n-1} = \frac{6.82}{5} = 1.37$$

Variance ↑ Spread ↑



→ Spread ↓ Var ↓

* Standard deviation

Standard deviation is a measure of how spread out numbers are.

↳ Square root of Variance

$s = \sqrt{var}$

Std $\Rightarrow \sqrt{1.37} = 1.17$

Std dev of Population $\Rightarrow \sigma = \sqrt{var_p}$

Std dev of Sample $\Rightarrow s = \sqrt{var_s}$

1, 2, 3, 3, 4, 4

→ Var = 1.37 ✓

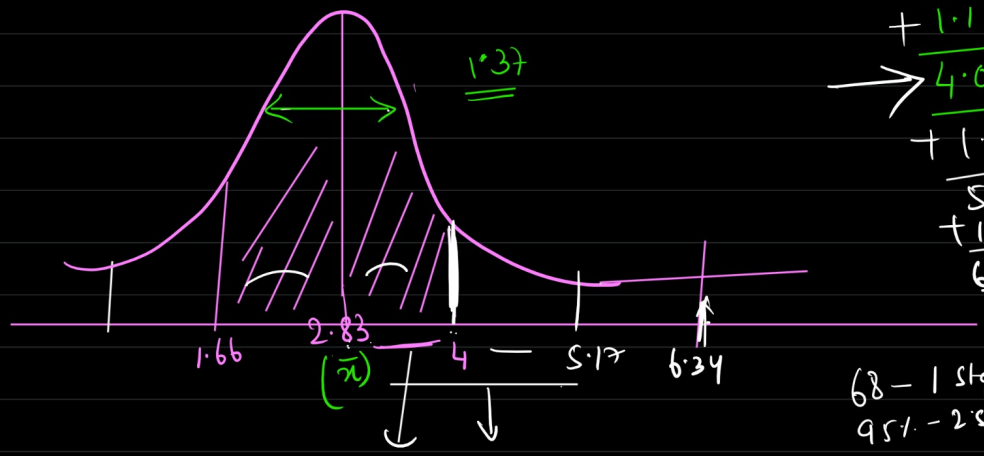
→ Std dev = 1.17 → Standard way of knowing where your data lies

$2.83 - 1.17 = 1.66$

$2.83 + 1.17 = 4.00$

$4.00 + 1.17 = 5.17$

$5.17 + 1.17 = 6.34$



68 – 1 Std

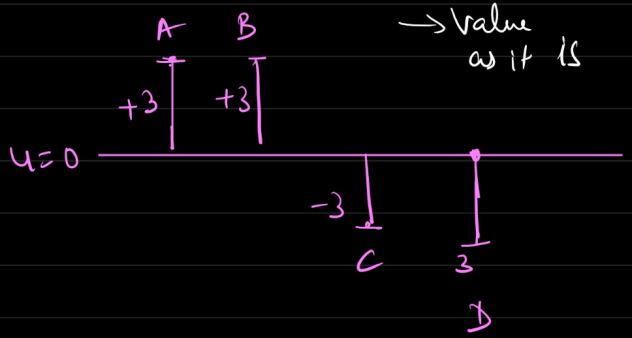
95% – 2 Std

99.7% – 3 Std

* Variance →

$$Var_p = \sigma^2 = \frac{\sum_{i=1}^{n} (x - M)^2}{n}$$

Why Square ?



→ Value as it is

$x - M$

$$\frac{+3 + 3 + (-3) + (-3)}{4} = 0$$

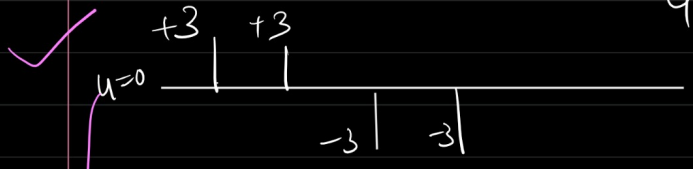
↓

+ve and –ve are negating each other

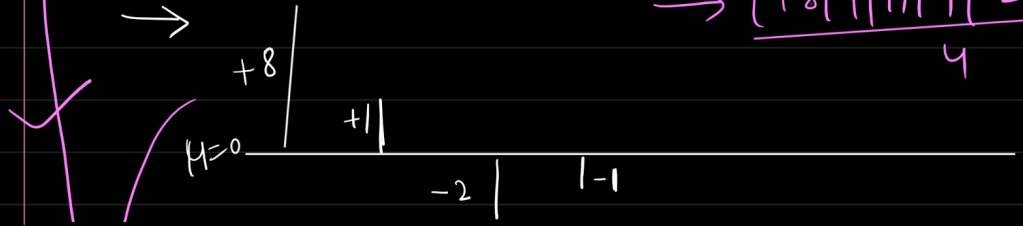
absolute value

→ $$\frac{|3| + (3| + |-3| + |-3|}{4} = \frac{12}{4} = 3$$

↑ mean deviation



→ $$\frac{|+8| + (+1| + |-2) + |-1|}{4} = \frac{12}{4} = 3$$



$$\sqrt{\frac{3^2+3^2+(-3)^2+(-3)^2}{4}} = \sqrt{\frac{36}{4}} = 3$$

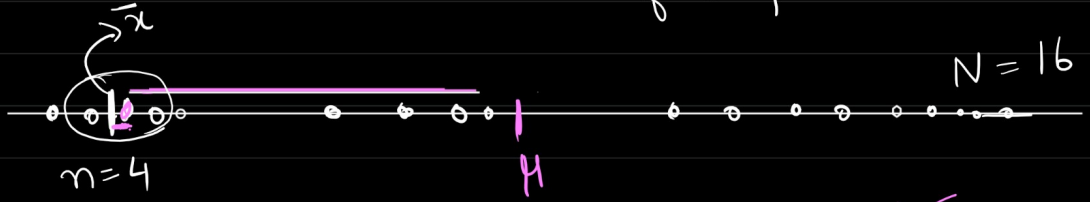
$$\sqrt{\frac{8^2+1^2+(-2)^2+(-1)^2}{4}} = \sqrt{\frac{64+1+4+1}{4}} = \sqrt{\frac{70}{4}} = 4.184$$

* Variance$_{sample}$ $= \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$ — Bessel correction → Why?

We use $n-1$ rather than n is because sample variance will be unbiased estimator

↓

We are estimating variance of population using variance of sample



$(x-\bar{x})^2$ ✓

Population $(x-\mu)^2$ ✓

$(x-\mu)^2 > (x-\bar{x})^2$ = Var.

$n \rightarrow n-1$.

In most of the cases you are reducing numerator

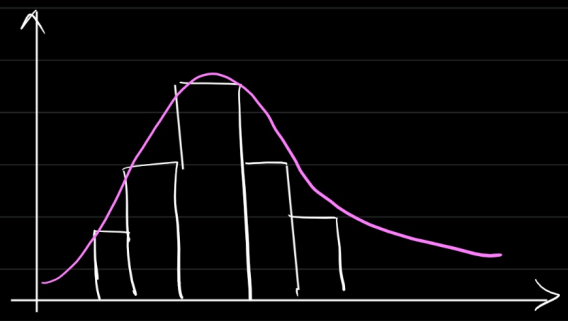
$$\frac{2}{5} = \frac{10}{5} = \frac{8}{4} = 2$$

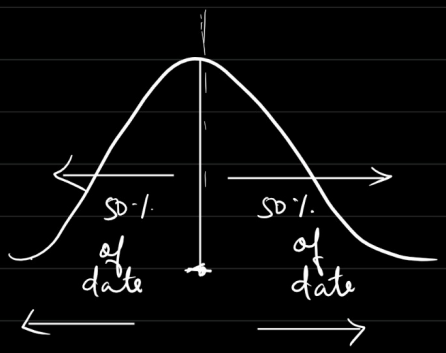
$$\Rightarrow s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

$n-1$ → Small

Measures of Symmetricity

→ If anything is exactly towards the left and right.

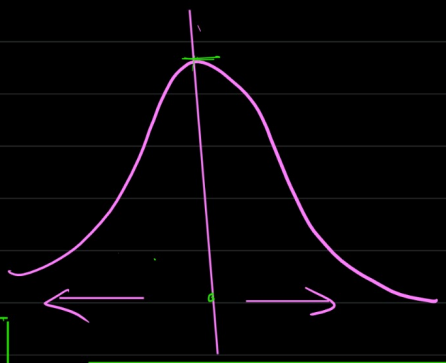




* Skewness → measure of data symmetry

① No Skewness

Skewness = 0



$Q_3 - Q_2 \approx Q_2 - Q_1$

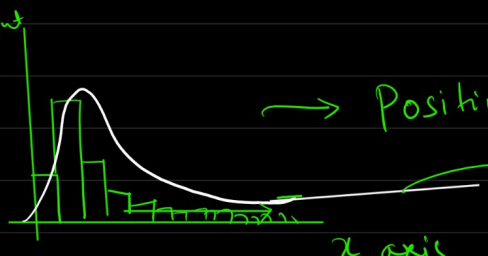
$$\text{Skewness} = \frac{x_i - \mu}{n\sigma^3}$$

mean = median = mode

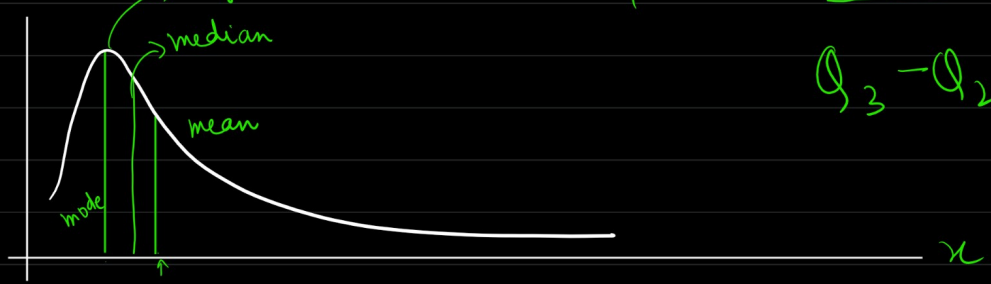
1, 2, 3, 4, 5

② → Positive Skewed

→ tail is on right side of distribution



$Q_3 - Q_2 > Q_2 - Q_1$



mean ≥ median ≥ mode

③ Left skewed distribution | neg skewed distribution

$Q_2 - Q_1 \geq Q_3 - Q_2$

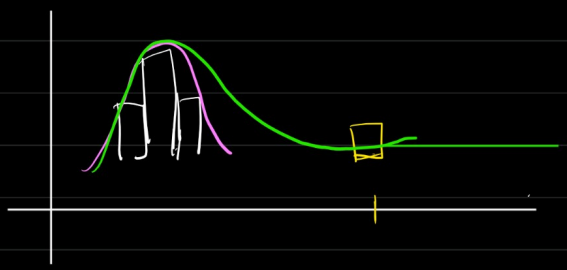
Relationship

mode ≥ median ≥ mean

Transformation

- Log transform
- Box-cox tranformation
- Exponential tran
- reciprocal "
- Outlier treatment

*



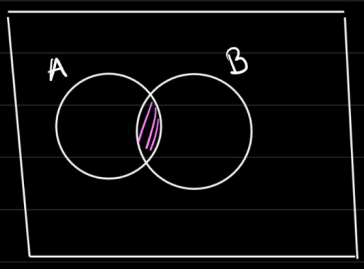
Set

$A = \{1, 2, 4, 5, 7\}$

$B = \{2, 4, 5\}$

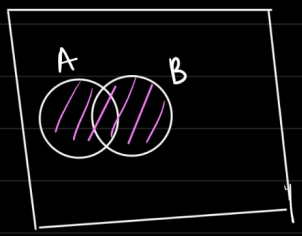
① Intersection (Common elements)

$A \cap B = \{2, 4, 5\}$



② Union (all distinct elements from both the sample)

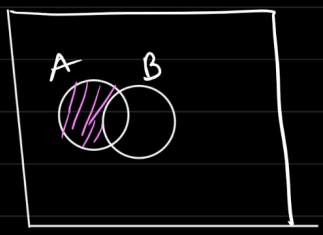
$A \cup B = \{1, 2, 4, 5, 7\}$



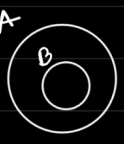
③ Difference (items which are present only in first set)

$A - B = \{1, 7\}$

A-B



④ Subset (all the element of B is present in A, then we say B is subset of A)



$B \rightarrow A$ —— True

(B is subset of A)

$\times A - B$ —— false

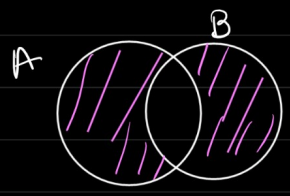
⑤ Superset (A is containing all elements of B, A is superset of B)

$A \rightarrow B \Rightarrow$ True

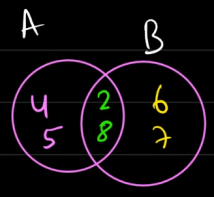
$B \not\rightarrow A. \Rightarrow$ false

⑥ Symmetric difference (opposite of intersection)

The elements that are distinct in both sets excluding intersection



$A \Delta B = \{1, 7\}$



$A = \{4, 5, 2, 8\}$

$B = \{2, 8, 6, 7\}$

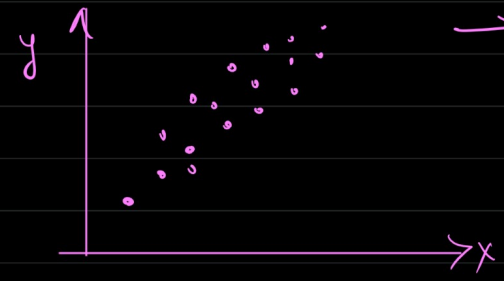
$A \Delta B = \{4, 5, 6, 7\}$

* Covariance and correlation

(X) transaction amount	(Y) transaction count
—	—
—	—
—	—
—	—

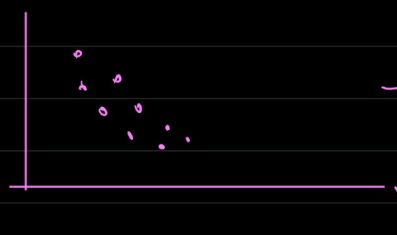
- Y↑ X↑
- Y↑ X↓
- Y↓ X↑
- Y↓ X↓

Understanding the relationship



→ relationship b/w X, Y is direct

X↑ Y↑
X↓ Y↓

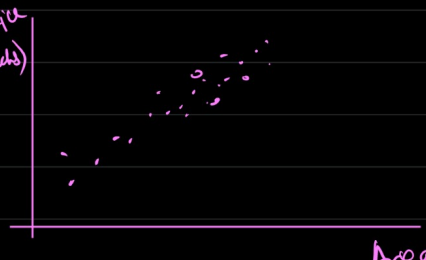


→ indirect relationship

X↓ Y↑
X↑ Y↓

* Example (direct)

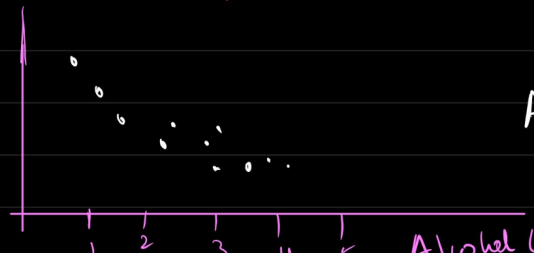
Predict price of house based on area of house



Area of house (sqft)	Price
1100	80
1200	85
—	—
—	—

AH↑ P↑
AH↓ P↓

* Example (Indirect



AC↑ YTL↓
AC↓ YTL↑

① Covariance

✓ $$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Co + Variance

$$Var(x) = \sum_{i=1}^{n} \frac{(x - \bar{x})^2}{n-1} \Rightarrow \sum_{i=1}^{n} \frac{(x-\bar{x})(x-\bar{x})}{n-1}$$

↓

Variance was spread of data ⇒ relationship of a feature with itself

⇓

Cov mean, you are trying to understand the relationship of a feature with respect to other feature.

→ Covariance → relationship b/w two variable.

→

X ↑ Y ↑	or	X ↑ Y ↓
X ↓ Y ↓		X ↓ Y ↑
(+ve) Cov		(−ve Cov)

X	Y
2	3
3	5
6	6
1	8
$\bar{x} = 3$	$\bar{y} = 5.5$

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x-\bar{x})(y-\bar{y})}{n-1}$$

$$\frac{(2-3)(3-5.5) + (3-3)(5-5.5) + (6-3)(6-5.5) + (1-3)(8-5.5)}{4-1}$$

$$\Rightarrow \frac{(-1)(-2.5) + 0 + 3 \times 0.5 + (-2) \times 2.5}{3}$$

$$\Rightarrow \frac{2.5 + 0 + 1.5 - 5}{3} \Rightarrow \frac{-1}{3} = \boxed{-0.33}$$

→ The two features X and Y are negatively related.

X	Y
2	3
4	5
6	7
$\bar{x}=4$	$\bar{y}=5$

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{n-1}$$

$$= \frac{(2-4)(3-5)+(4-4)(5-5)+(6-4)(7-5)}{3-1}$$

$$\Rightarrow \frac{4+0+4}{2} = 4 = +ve$$

→ x & y are having a positive relation.

Advantage

→ Relationship b/w X, Y

+ve or -ve

* Disadvantage

①

X	Y
A	B

$Cov(x,y) = 50$

$Cov(A,B) = 100$

$\Downarrow$

$-\infty$ to ∞

→ No comparison of Strength of relationship in Covariance

→ No any standardized scale to interpret the strength.

② Covariance has dimension

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

x → height → ft

y → wt → kg.

Cov. ⇒ ft. kg

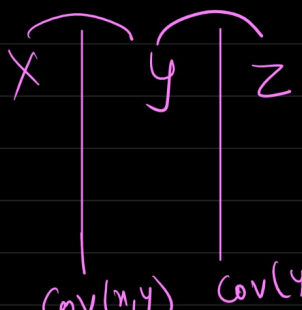
(X) (Y) (Z)

transaction Amount (Rs) | height (ft) | weight (kg)

$Cov(trAmt, ht) \Rightarrow Rs \cdot ft \Rightarrow 450\ Re \cdot ft$

$Cov(height, wt) \Rightarrow ft \cdot kg \Rightarrow 600\ ft \cdot kg$

→ We can not compare two different dimension



Not comparable → different dimension.

* Soln
  - → -1 to 1
  - → dimensionless Quantity

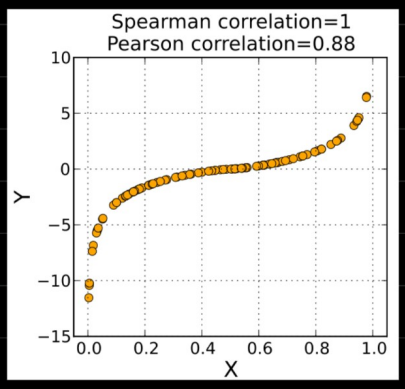
② Pearson Correlation Coefficient [-1 to 1]

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = [-1 \text{ to } 1]$$

$\rho_{x,y} = 0.4$ | $\rho_{A,B} = 0.8$

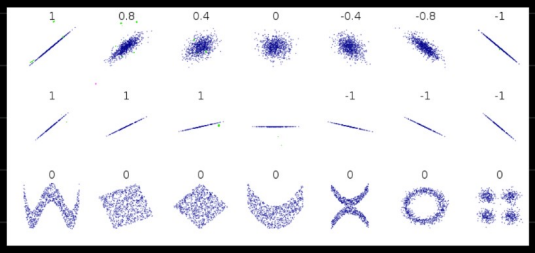
→ feature A,B is highly correlated as compared to x,y.

→ Pearson correlation coefficient always measures the linear relationship



Pearson correlation
↓
linear relationship

$x \uparrow y \uparrow$



What to do for Non linear relationship

Spearman Rank Correlation

$$r_s = \frac{Cov(R(x), R(y))}{\sigma_{R(x)} * \sigma_{R(y)}}$$

$R(x)$ – Rank of x

$R(y)$ – Rank of y

x	y	R(x)	R(y)
5	6	3	1
7	4	2	2
8	3	1	3
1	1	5	5
2	2	4	4

5th 4th 3rd 2nd 1st

$x \rightarrow 5, 7, 8, 1, 2 \longrightarrow$ Sort the value $\rightarrow 1, 2, 5, 7, 8$

Sort the no → Highest no will be rank 1.

Dataset → 1000 feature

o/p (y)

x_1 x_2 x_3 x_4 x_5 x_6 ---- x_{1000} Price

$x_1 - y$

$x_2 - y$

$x_3 - y$

$x_4 - y$

Corr ≈ 0

* Random variable

↳ A set of Possible values from a random experiment.

* Tossing a coin → Experiment is random,

↳ H, T

Outcomes will be random.

↓

Quantify these random values.

$$X = \begin{cases} 0 & \text{—— tail (T)} \\ 1 & \text{—— Head (H)} \end{cases}$$

$X = \{0, 1\}$

↑

random variable

↓

It can take any value from the set of values

* We have an experiment (tossing a coin)
* quantify each event $\{0, 1\}$
* this values is Random variable.

$$x + 5 = 10$$
$$x = 10 - 5 = x = 5$$

In Algebra, a variable value is fixed

$X = \{0, 1, 2, 3\}$

X could be 0, 1, 2, or 3 randomly

$X = \{1, 2, 3, 4, 5, 6\}$ → Sample Space

$P(X=1) = 1/6$

(Tossing a coin) $P(X=1) = 1/2 \longrightarrow \{0, 1\}$ (0 → tail, 1 → Head)

$P(x \leq 4)$