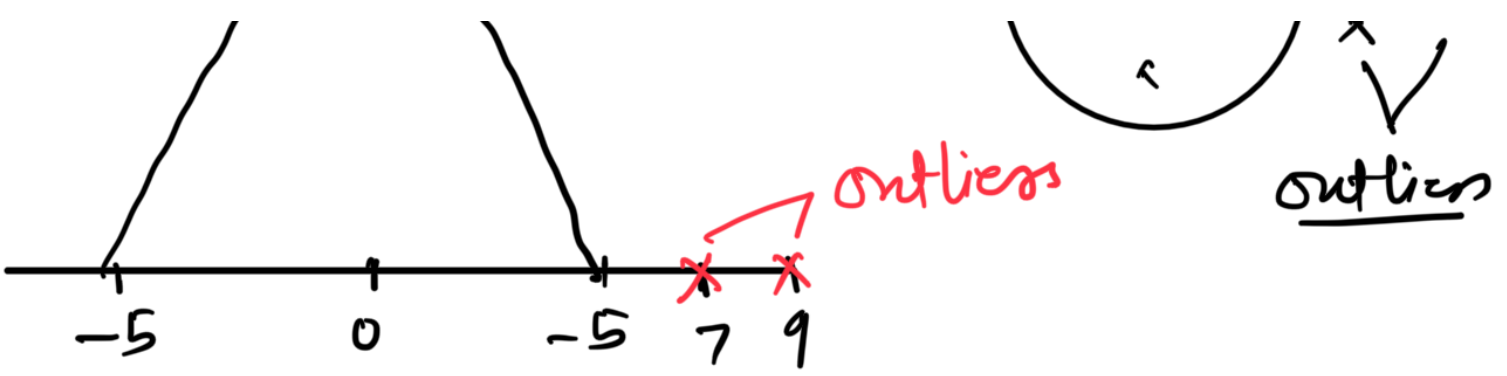
Descriptive Analysis / Statistics

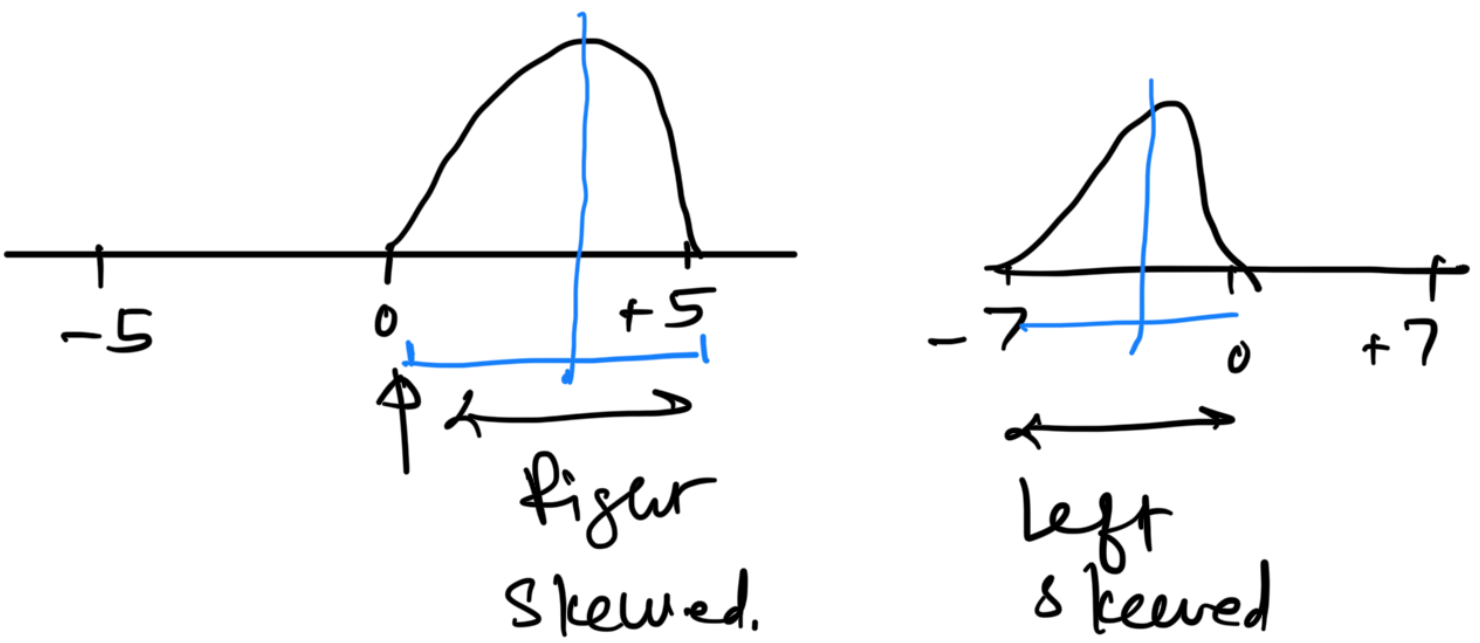
- Descriptive statistics summarize & describe main features of a dataset.
- provides simple summaries about the data and measures.
- Measures → mean
  median
  mode
  variance
  Std deviation

1. Mean : Average is the sum of all values divided by the number of values. It provides a measure of the central tendency of the data.
   — Use it to normize data
   - helpful in summarizing the data

2. Median : The median is the middle value of a sorted dataset



Correct → Normal distributed
{±5}



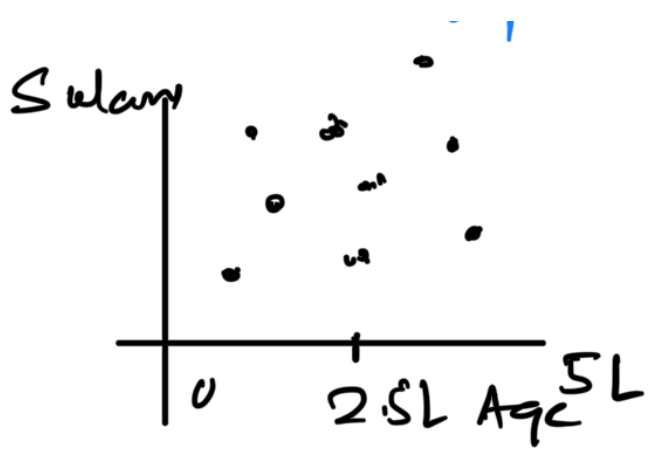
→ Used to handle skewed data

→ Useful in descriptive statistics to understand data distribution.

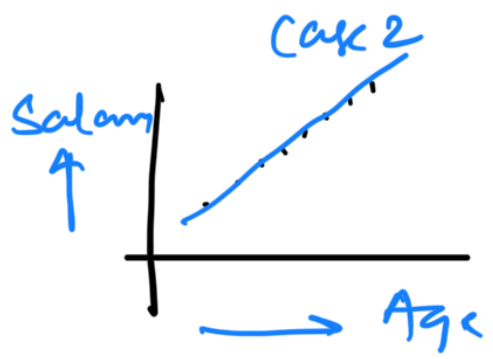
3. Mode – The mode is the most-frequently occuring value in a dataset.

– Use to handle categorical (text) data
– helpful to identify the most common value in the dataset

4. Variance –



Variance measures the dispersion of the data points from the mean



Case 1 – All professions

Case 2 – IT profession

→ Indicates the variability in data

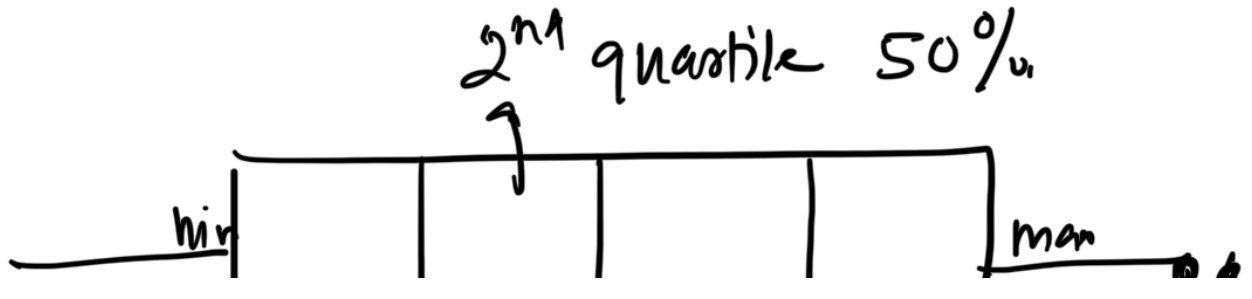
→ Used to understand the spread of features

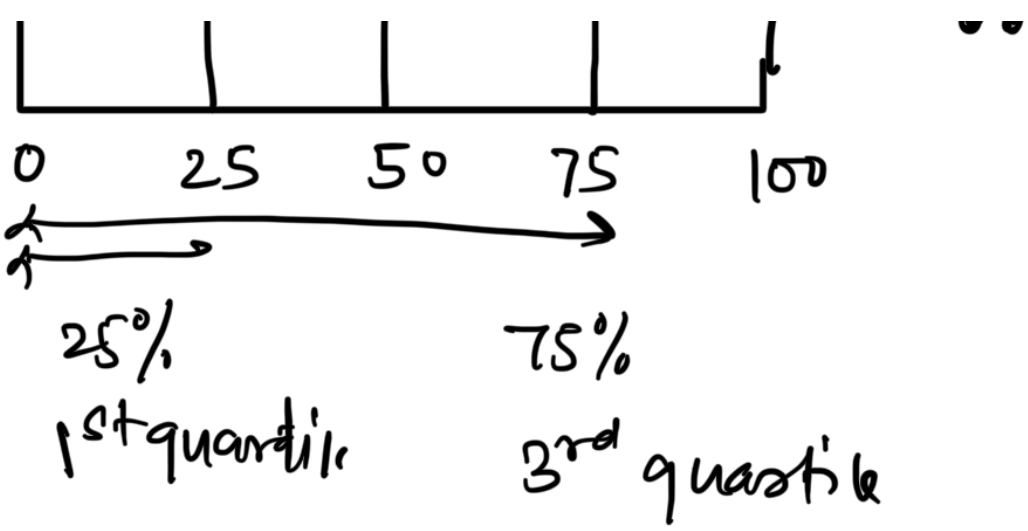
5. Standard Deviation – It is the Square root of the variance, provide the measure of the average distance from the mean.

– Use it to Scale the data

– help us to understand the spread of dataset.

6. Range and Interquartile Range (IQR)





- The range is the difference betn max and min values.
- The IQR is the difference betn the 75th and 25th quartile (percentile)
- helps to measure the spread of the middle 50% of the data

Use

- Identify outliers
- Understanding about distribution

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

$$\text{std deviation } (\sigma) = \sigma = \sqrt{\frac{\Sigma (x-\mu)^2}{n}}$$

$$IQR = Q_3 - Q_1$$

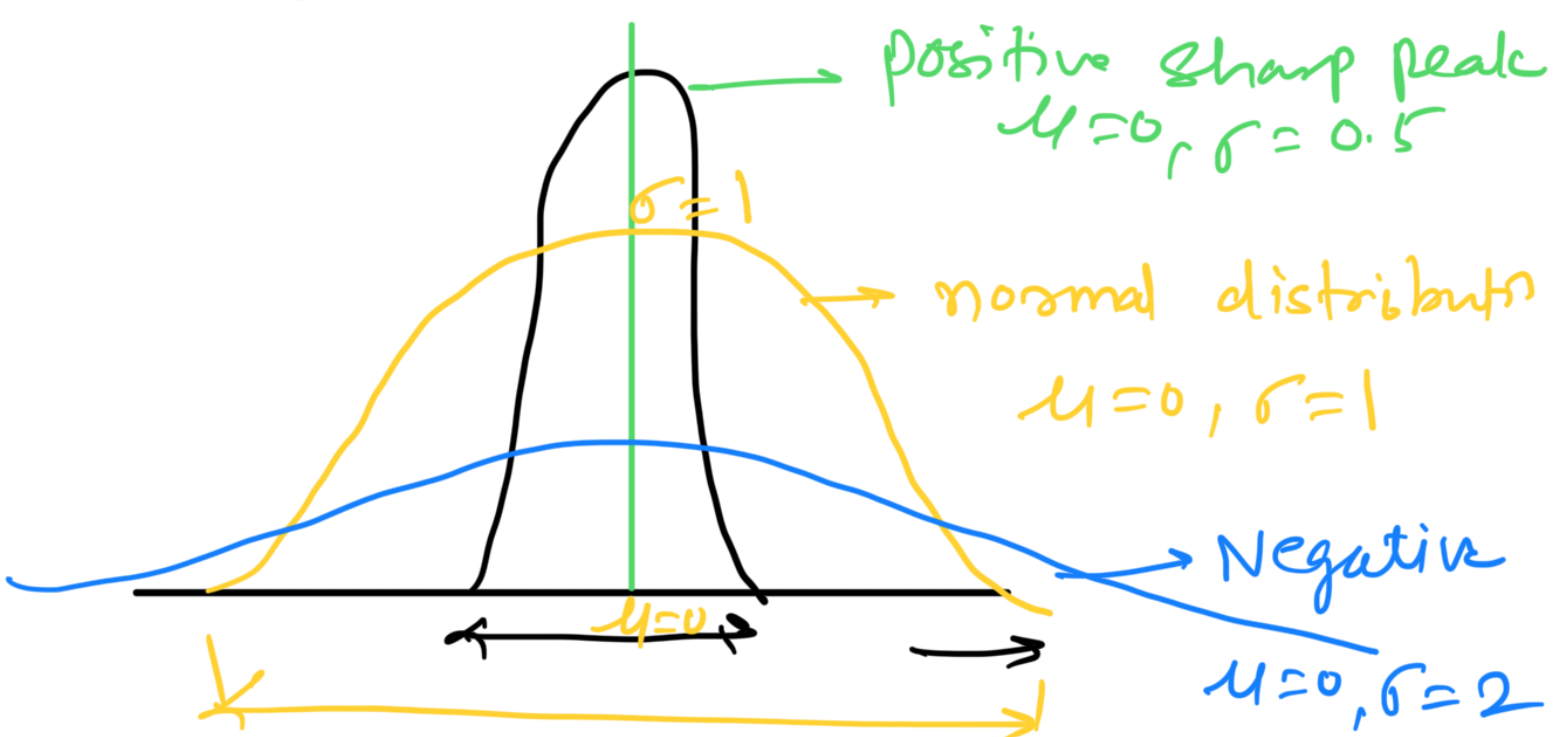
Coefficient of Variance (CV) $= \left(\frac{\sigma}{\mu}\right) \times 100$

Measures of Dispersion

Range, σ, σ^2, IQR, CV

Measures of Shape

Kurtosis — Measure the 'tailedness' of the distribution



- kurtosis talks @ the central peak and the dispersion of datapoints
- Types of kurtosis -

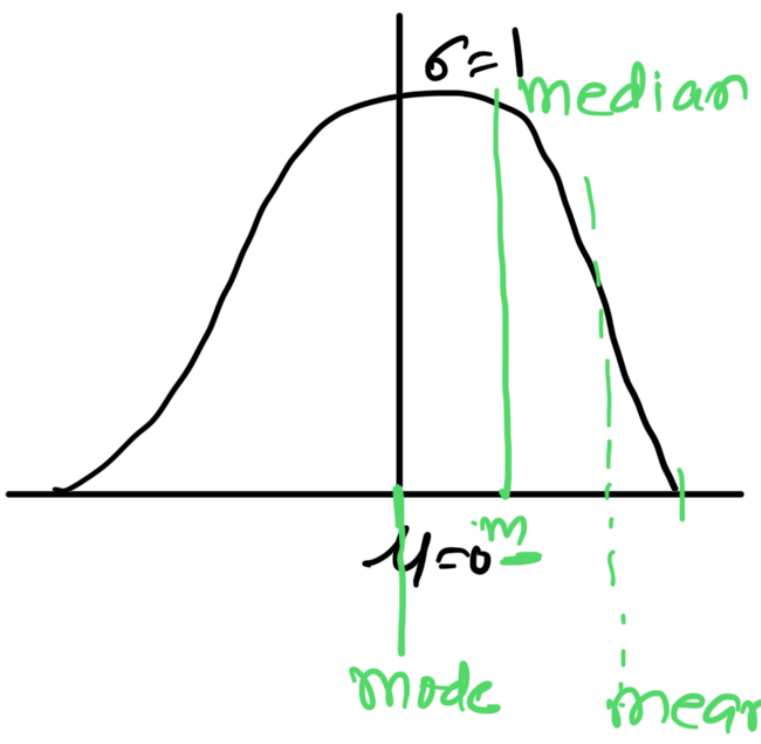
1. Positive ⟶ Sharp peaks and

lighter dispersion

2. Negative → wide peaks and flicker dispersion

3. Normal distribution

Skewness < Positive skewed / Negative skewed

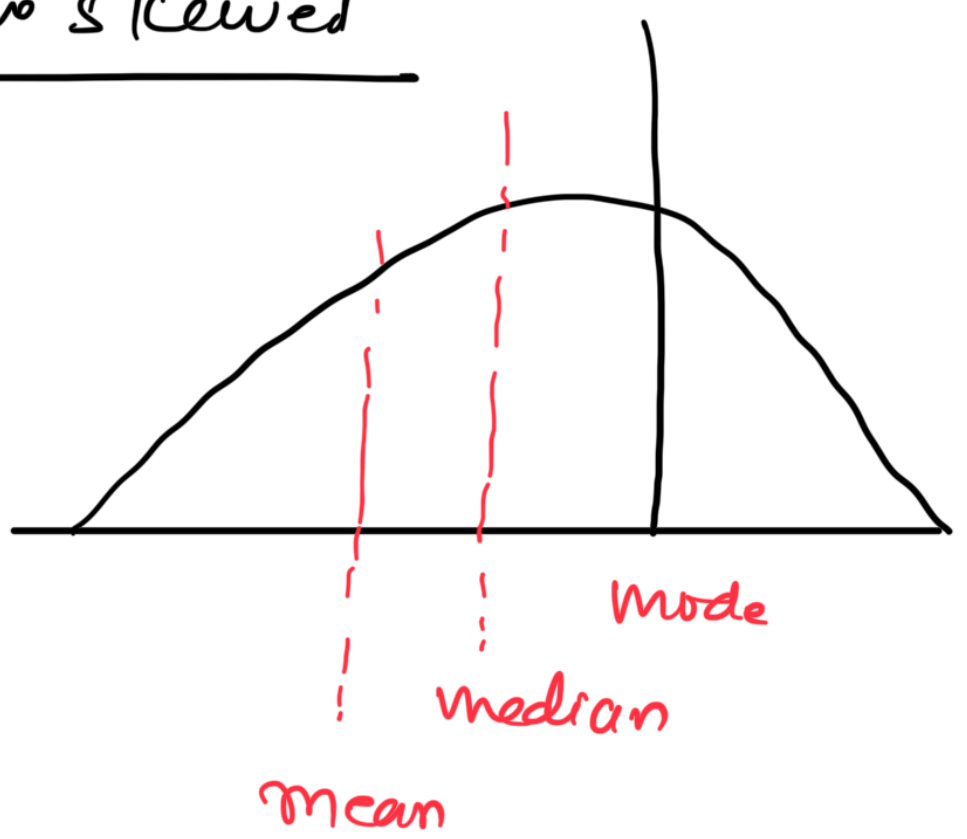


Positive skewed

mean > median > mode

mean > median > mode

Negative skewed



Mean < median < mode

Summary

Kurtosis – Measure the "tailedness" of the distribution

Skewness – Measure the asymmetry of the distribution

1) Positive skew – Mean > Median

2) Negative skew – Mean < Median

3) Zero skew – Mean = Median

Covariance & Correlation

Covariance : $Cov(x, y)$

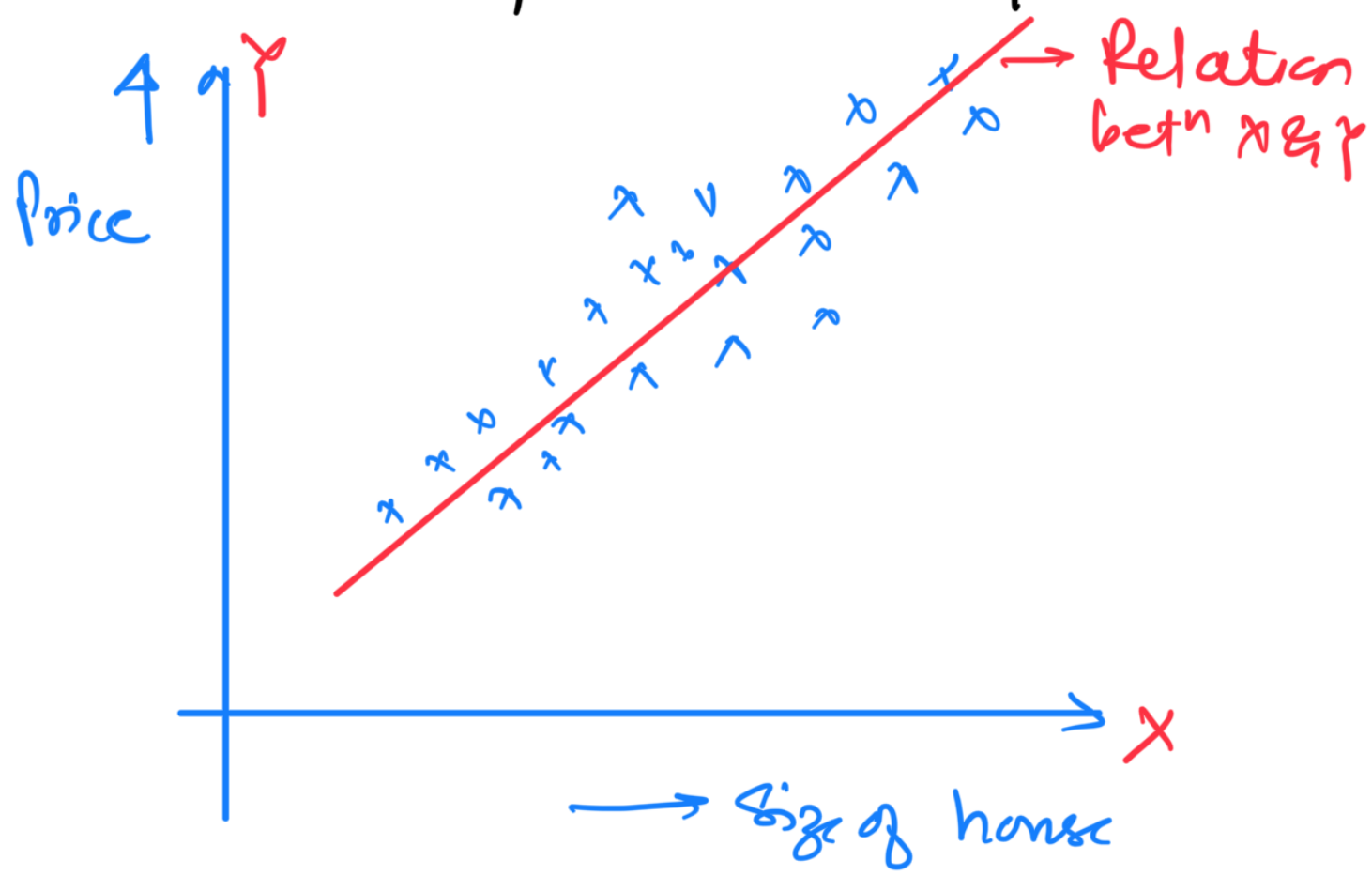
$$Cov(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Correlation : $(\rho(X, Y))$

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Regression Analysis

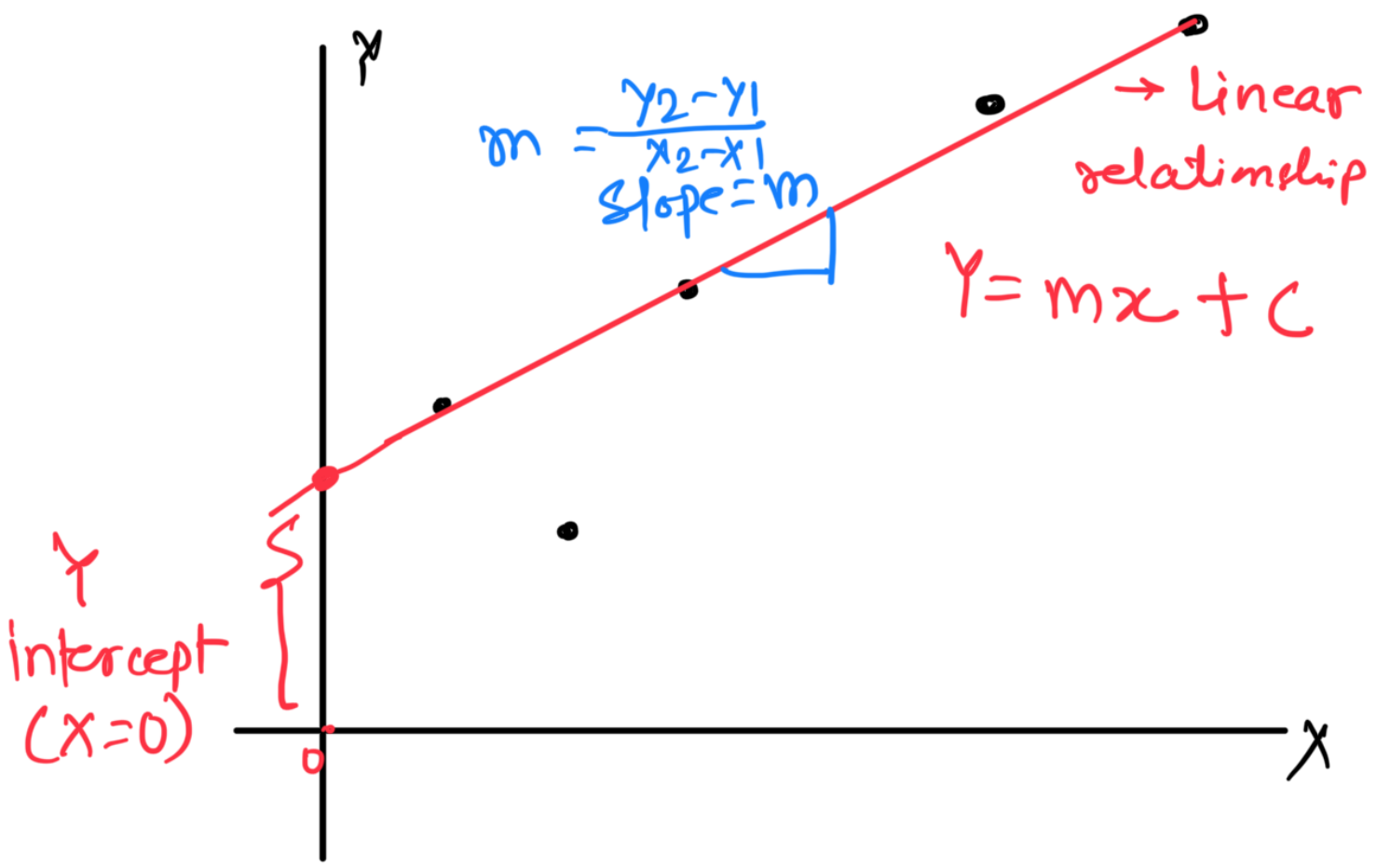
Example : Regression Techniques



Ok

To define relationship we can use Regression Technique

X	1	2	3	4	5
Y	5	3	6	8	10



$$Y = \boxed{m} X + \boxed{C}$$

- Y ← Get the value of Y
- m → Slope
- X → Value of X from table
- C → Constant term

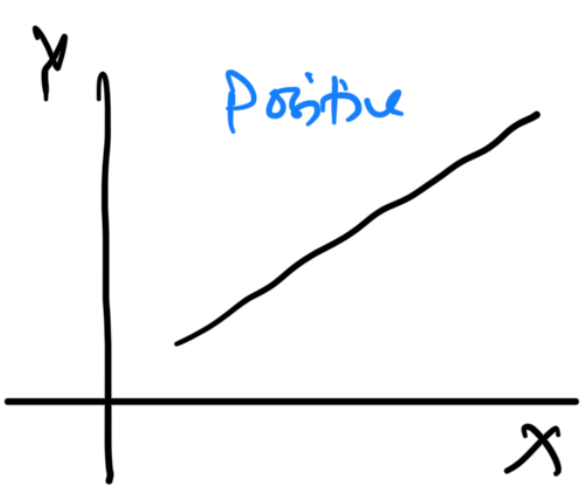
$X = 10$

$X = 20$

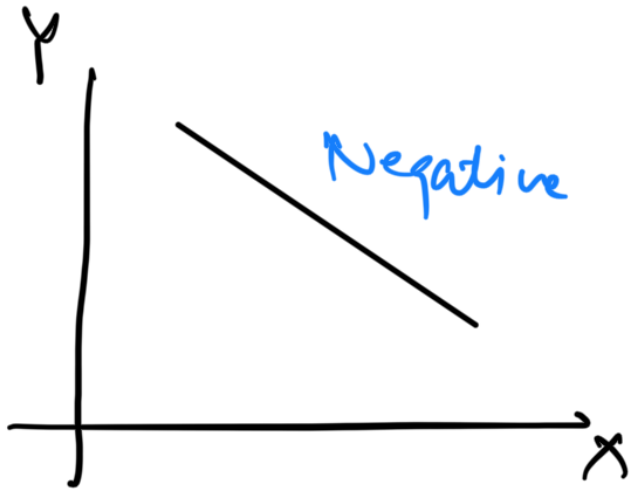
Linear Regression Model

↓

Identify the relationship bet^n 2 features / columns of the dataset



positive linear Relationship



Negative linear Relationship

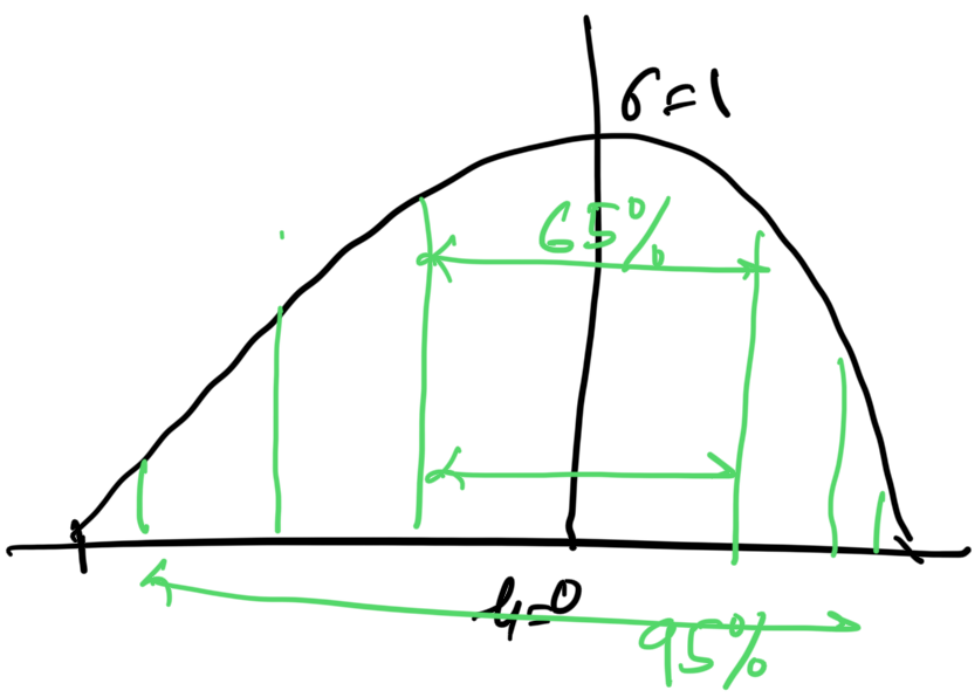
positive linear Relationship	Negative linear Relationship
X → increase	X → increase
Y → increase	Y → decrease
X → decrease	X → decrease
Y → decrease	Y - decrease

Probability & Distribution

Probability distribution - Understanding probability distribution is crucial for many machine learning algorithms.

① Normal Distribution (Gaussian distribution)

- bell-shaped curve, symmetric around the mean
- Use - many algorithms -
  - eg. - Linear Regression
  - Logistic Regression
- Used in Statistical test



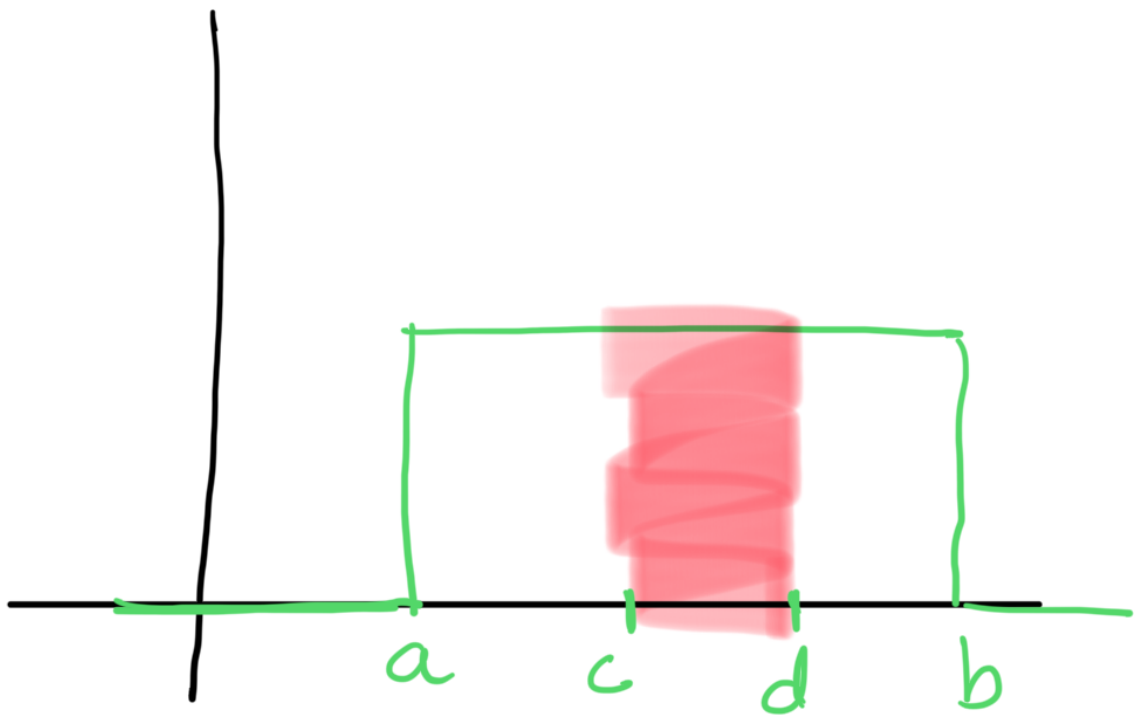
$f(x)$ = pdf = probability distribution function

$$(x-\mu)^2$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)}$$

2. Uniform distribution

- has constant probability
- Used for random sampling
- Useful in simulation



$$f(x) = \frac{1}{b-a}$$

$$P(c \le x \le d) = \frac{d-c}{b-a}$$

$$\mu = \frac{a+b}{2}$$

$$\sigma = \sqrt{\frac{(b-a)^2}{2}}$$

(3) Binomial Distribution

It describes the number of success in a fixed number of trials.

— used in binary classification

- helpful in modeling binary outcomes

$$f(x) = \binom{n}{x} p^x q^{(n-x)}$$

n = no. of trials p = prob of success

x = # of success q = $(1-p)$
= prob of failure

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Hypothesis Testing — is used to make decisions about a population based on sample data. It helps in determining if there is enough evidence to reject a null hypothesis.

(1) t-Test : compares the means

of two groups to see if they are significantly different from each other.

→ feature selection

→ Comparing 2 models

② Chi - Square Test: checks the independence of two categorical variable

- feature selection
- Understanding association in Categorical data

③ Pearson Correlation Coefficient:

measure the linear correlation between two variables.

- quantifying the degree of linear relationship
- Identify highly correlated features.