

## Statistics

### a) Descriptive

- Analyzing, summarizing, organising data in the form of graphs i.e visualisation.
- Bar plot, pdf, diff distribution
- Measure of central tendency (Mean, median, mode)
- Measure of variance, sd variance

### b) Inferential stats

- From the entire population we take small samples and try to inference and conclude what the outcome will be. e.g exit poll
- Confidence intervals {**ange of values** that likely contains the **true population parameter** (like mean) with a certain level of confidence.}
- Ztest, t test, chi square test.

#### • Population Standard Deviation ( $\sigma$ )

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

- $X_i$ : data point
- $\mu$ : population mean
- $N$ : total number of points

#### • Sample Standard Deviation ( $s$ )

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

- $\bar{X}$ : sample mean
- $n$ : sample size
- Uses **(n - 1)** to reduce bias (Bessel's correction)

Variance =  $SD^2$  (how much spread from the mean)

# HEBYSHEV'S INEQUALITY

$$X \sim G.D(\mu, \sigma)$$

$$Y \not\sim G.D.$$

$$Pr(\mu - \sigma \leq x \leq \mu + \sigma) \approx 68\%$$

$$Pr(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 95\%$$

$$Pr(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 99.7\%$$

$$Pr(\mu - k\sigma \leq x \leq \mu + k\sigma) > \left[1 - \frac{1}{k^2}\right]$$

$$k=2.$$

$$Pr(\mu - 2\sigma \leq x \leq \mu + 2\sigma) > 1 - \frac{1}{2^2}$$

$$1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} = 75\% \approx 85\%$$

### • Definition

For any distribution (any shape), **at least**


$$1 - \frac{1}{k^2}$$

of the data lies within **k standard deviations** from the mean,  
for any  $k > 1$ .

### • Formula

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

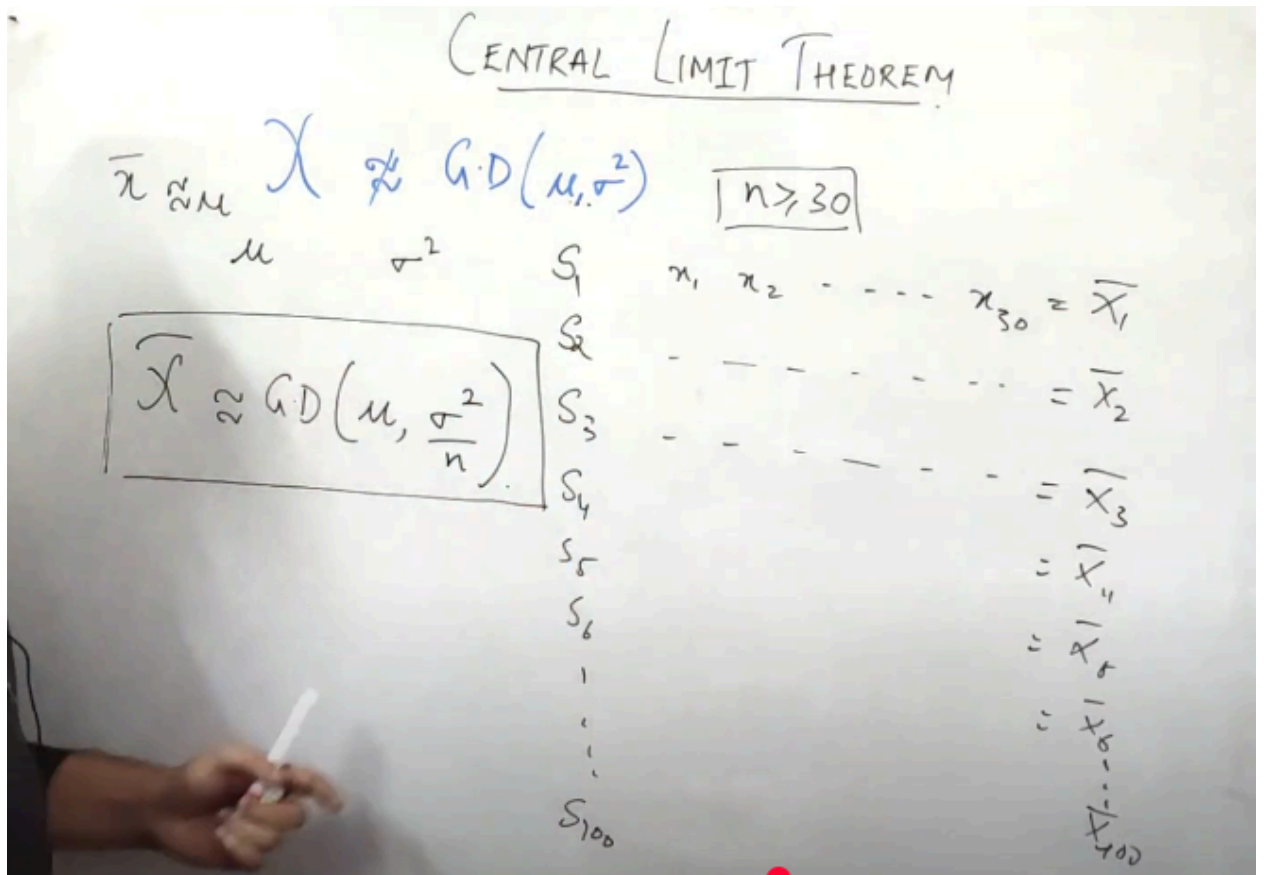
### • Example

k	At least this % within $\mu \pm k\sigma$	
2	$1 - \frac{1}{4} = 0.75 \rightarrow 75\%$	
3	$1 - \frac{1}{9} = 0.89 \rightarrow 89\%$	
4	$1 - \frac{1}{16} = 0.9375 \rightarrow 93.75\%$	

### ✓ Significance

- Works for **any distribution**, not just normal.
- Helps find **bounds** on spread without needing distribution shape.
- Useful when data is **not normal** or unknown.





If you take **many random samples** (size  $n$ ) from **any population** (not necessarily normal), the **distribution of the sample means** will **tend to follow a normal distribution** as  $n$  becomes large (usually  $n \geq 30$ ).

Makes it possible to use **normal distribution** tools (like Z-score) on **non-normal data**.

Used in **confidence intervals**, **hypothesis testing**, etc.

for `_` in range(1000):

```
sample = np.random.choice(population, size=30)
```

```
sample_means.append(np.mean(sample))
```

Population mean vs Sample mean:-

Population data may be **too large or impractical** to collect.

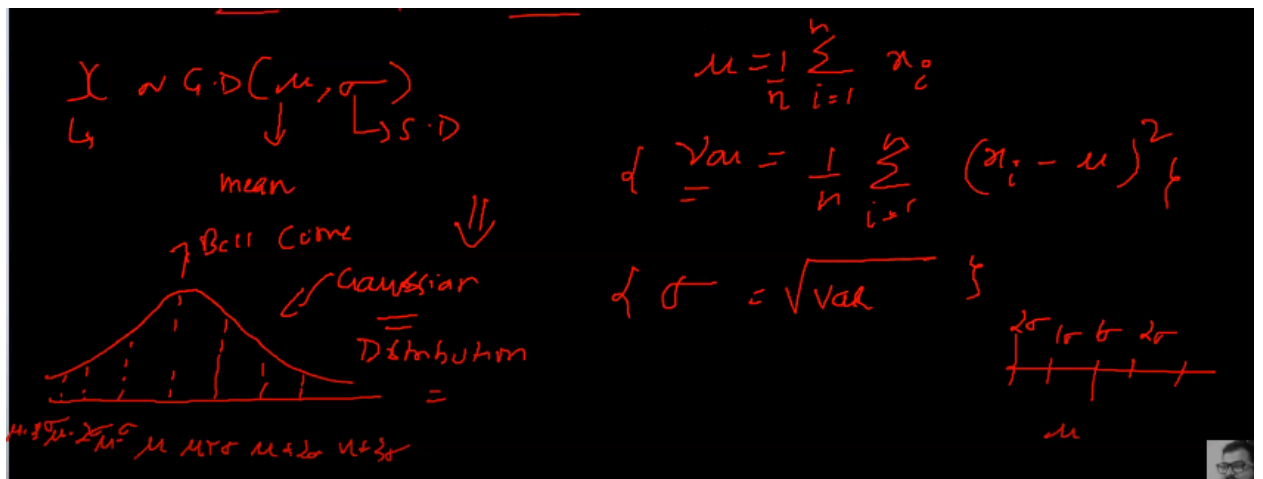
Sample mean helps to **estimate** the population mean. (exit poll)

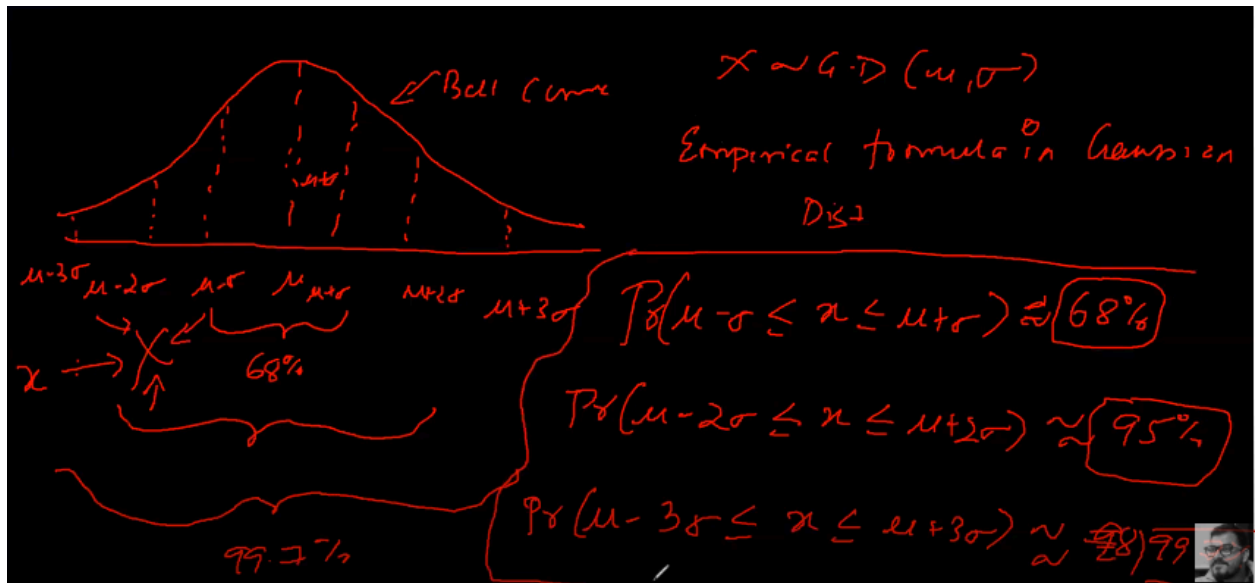
---

Random variable

- Discrete - int
- Continuous - within a range any value

Gaussian / Normal Distribution (Data is symmetrically distributed e.g height)





Empirical formula

---

log Normal Distribution (skewed distribution e.g Salary, feedback)

Handwritten text defining a Log Normal distribution:

$X \sim \text{Log Normal}$   
 if  $\ln(x) \sim \mathcal{N}(\mu, \sigma)$

- To make the distribution normal or normally distributed using log fn.
- **Scaling and standardscaler is used to normalise the data.**

- Use Scaling (MinMaxScaler) when you need bounded values, e.g., in neural networks.

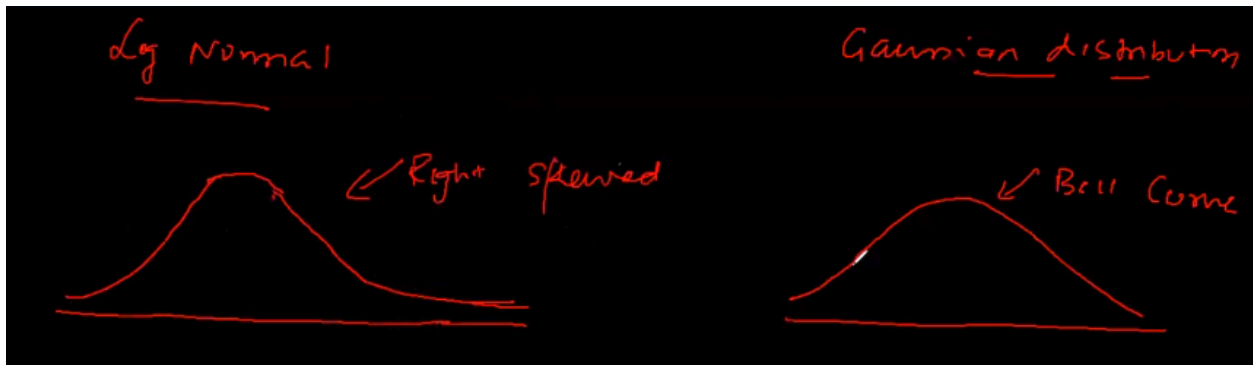
$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

```
scaler = MinMaxScaler()
scaled_data = scaler.fit_transform(data)
```

- Use StandardScaler when data is Gaussian or for PCA, SVM, Logistic Regression.

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

```
standard_scaler = StandardScaler()
standardized_data = standard_scaler.fit_transform(data)
```



E.g,

R&D , Marketing, Campaign & Profit are parameters provided.

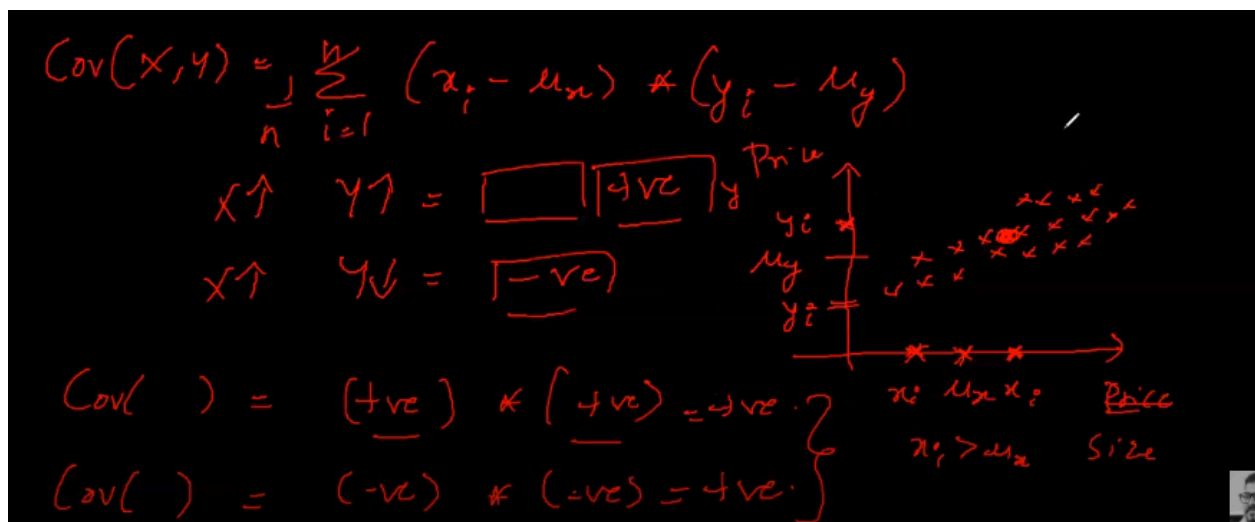
- We need to find out what distribution each parameter follows.
- There will be huge gap between the values of each parameter
- Suppose R&D follows gaussian dist.
- Convert the G.D to **Standard normal deviation**( $\mu = 0$  &  $\sigma = 1$ )
- **Standard scaler corresponds to Zscore.**

Feature	Scaling (Normalization)	StandardScaler (Standardization)
Goal	Rescale data to a fixed range	Transform data to mean = 0, std = 1
Formula	$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$	$X' = \frac{X - \mu}{\sigma}$
Resulting Range	[0, 1] or [-1, 1]	No fixed range (depends on data)
Sensitive to	Outliers	Less sensitive than scaling
Use When	Feature ranges vary a lot	Data follows normal/Gaussian distribution
Tool in sklearn	<code>MinMaxScaler()</code>	<code>StandardScaler()</code>

- If marketing follows log normal dist. Then we can directly do  $\ln(x)$  to normalise the value.
- It will make it a gaussian dist. Then find std.

Covariance,

How **two variables change together**.



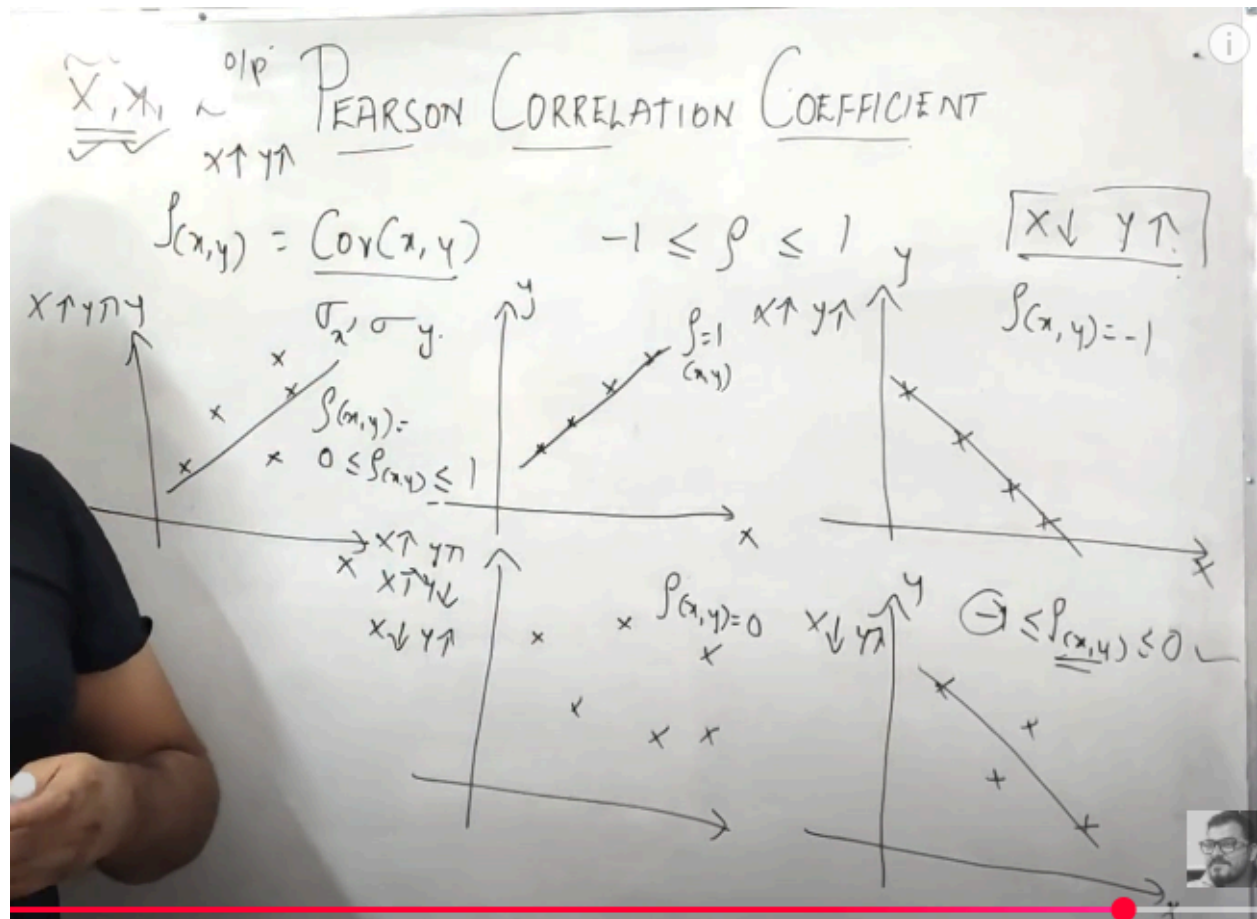
Meu = mean

`cov_manual = np.sum((np.array(X) - mean_X) * (np.array(Y) - mean_Y)) / len(X)`



# 2. Using NumPy's cov()

cov\_matrix = np.cov(X, Y, bias=True) # bias=True for population covariance



$$\sigma_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}$$

- Helps to understand the power and direction of correlation
- If the  $r = \pm 1$  then we can know both are same and can omit any one.

## Spearman Correlation

It measures the **strength and direction of the monotonic relationship** between two variables using their **ranks**, not raw values.

### • Formula:

If no tied ranks:

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

- $d_i$  = difference between ranks of each pair
- $n$  = number of pairs

### ✓ When to Use:

- Data is **not normally distributed**
- Relationship is **monotonic** (always increases or decreases, not necessarily linear)
- **Ordinal data** or ranks

```
rho, p_value = stats.spearmanr(X, Y)
```

Step 1: Rank the values			
X	Rank(X)	Y	Rank(Y)
10	1	30	3
20	2	20	2
30	3	10	1

Step 2: Find the difference in ranks (d)			
Rank(X)	Rank(Y)	$d = R_X - R_Y$	$d^2$
1	3	-2	4
2	2	0	0
3	1	2	4

## Outlier

- It is a data point in a dataset that is distant from all other observations i.e data point that lies outside the overall distribution. (e.g salary of one ceo too high when compared to mean salary)
- Any data outside the third sd is considered an outlier i.e if  $Z \leq 3$  then it will not be considered outlier.

### What is the reason for an outlier to exists in a dataset?

1. Variability in the data
2. An experimental measurement error

### What are the impacts of having outliers in a dataset?

1. It causes various problems during our statistical analysis
2. It may cause a significant impact on the mean and the standard deviation

## Various ways of finding the outlier.

1. Using scatter plots
2. Box plot
3. using z score
4. using the IQR Interquartile range

```
outliers=[]
def detect_outliers(data):

    threshold=3
    mean = np.mean(data)
    std =np.std(data)

    for i in data:
        z_score= (i - mean)/std
        if np.abs(z_score) > threshold:
            outliers.append(y)
    return outliers
```

```
outlier_pt=detect_outliers(dataset)
```

```
outlier_pt
```

```
[102, 107, 108]
```

## InterQuantile Range

75%- 25% values in a dataset

### Steps

1. Arrange the data in increasing order
2. Calculate first(q1) and third quartile(q3)
3. Find interquartile range (q3-q1)
4. Find lower bound  $q1*1.5$
5. Find upper bound  $q3*1.5$

Anything that lies outside of lower and upper bound is an outlier

Step 1 to 2

```
quantile1, quantile3= np.percentile(dataset,[25,75])  
  
print(quantile1,quantile3)  
  
12.0 15.0
```

---

P value

- p-value is the **chance of seeing your result just by luck, assuming the null hypothesis is true.**
- If a feature has a **high p-value**, we assume:  
"This feature **may not affect** the target significantly." (**random chance.**)
- So omit the feature.

✓ **Step-by-step:**

**1. Compute t-statistic:**

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Where:

- $\bar{X}$ : sample mean
- $\mu_0$ : hypothesized population mean
- $s$ : sample standard deviation
- $n$ : sample size

**2. Use the t-distribution to find p-value:**

$$\text{p-value} = 2 \times P(T > |t|)$$

(using degrees of freedom  $df = n - 1$ )

## ♦ 1. Feature Selection

p-value helps you decide if a **feature (input variable)** actually influences the **target**.

p-value	Interpretation
Low ( $\leq 0.05$ )	Feature is <b>significant</b> → keep it
High ( $> 0.05$ )	Feature likely <b>random</b> → drop it

✓ Especially useful in **Linear Regression, Logistic Regression, and Statistical ML models**.

## ♦ 2. Model Interpretability


- p-value shows which features are **actually contributing**
- Helps explain the model in **interpretable ML**, useful in fields like:
  - Healthcare
  - Finance
  - Policy

## ♦ 3. Avoiding Overfitting

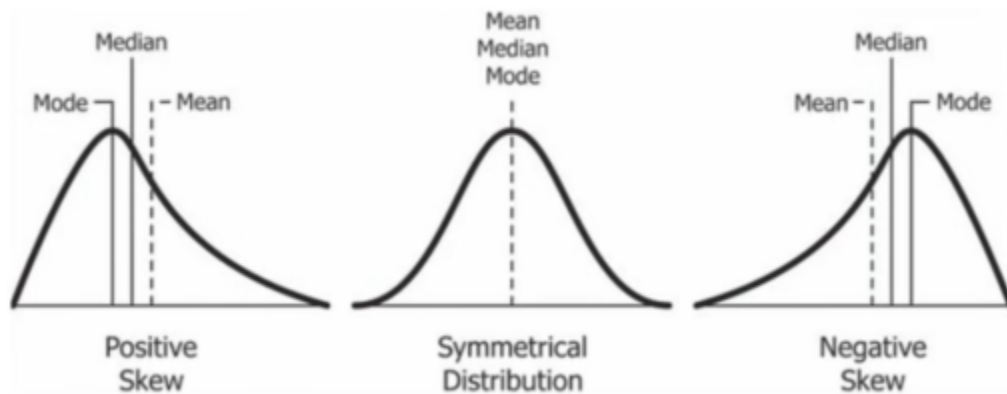
By removing high-p-value (irrelevant) features, your model becomes **simpler** and more **generalized**.

## ♦ 4. A/B Testing & Experiments

In ML pipelines or MLOps:

- Use p-value to check if a new model version performs **statistically better**
- Common in **product experiments, UI**  **ing, feature launches**

## Skewed Distribution and its relation with Mean, Median and Mode



Mean>Median>Mode

same

Mode>Median>Mean

E.g wealth, comment

height. Weight

Lifespan

---

## Confidence Intervals About Mean

- **Definition:**

A **confidence interval** gives a **range of values** that likely contains the **true population mean** with a certain level of confidence (e.g. 95%).

- **Formula (when population std is unknown):**

$$CI = \bar{X} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Where:

- $\bar{X}$  = sample mean
- $s$  = sample standard deviation
- $n$  = sample size
- $t_{\alpha/2}$  = t-score for desired confidence (from t-table)

- **Example:**

Sample: [65, 68, 72, 66, 70]

→  $\bar{X} = 68.2, s = 2.86, n = 5$

At **95% confidence**,  $t_{0.025, df=4} \approx 2.776$

$$CI = 68.2 \pm 2.776 \cdot \frac{2.86}{\sqrt{5}} \approx 68.2 \pm 3.56$$

$$CI \approx [64.64, 71.76]$$

- **✓ Interpretation:**

"We are 95% confident that the **true mean** lies between **64.64 and 71.76**."





## Bernoulli Distribution

As the number of independent trials increases, the relative frequency of success approaches the actual probability of success.

### ✓ Formula (Law of Large Numbers form):

If a Bernoulli trial is repeated  $n$  times, and the probability of success is  $p$ , then:

$$\frac{\text{Number of Successes}}{n} \rightarrow p \text{ as } n \rightarrow \infty$$

#### • Example:

- Toss a fair coin ( $p = 0.5$  for heads)
- If you toss it 10 times → you may get 7 heads
- Toss it 10,000 times → heads will be ~50% of the time

✓ As trials increase, observed frequency  $\approx$  true probability

Use	Why It Matters
Model evaluation	Accuracy stabilizes after enough samples
Data sampling	Justifies training on large datasets
Confidence estimation	Probability estimates improve with more data

**PMF** gives the **probability** of a **discrete random variable** taking a specific value.

### 🎯 Scenario:

A **biased coin** has:

- $P(\text{Heads}) = 0.7$
- $P(\text{Tails}) = 0.3$

Let random variable  $X$ :

- $X = 1$  for Heads
- $X = 0$  for Tails

### ✅ PMF Table:

$X = x$	Meaning	$P(X = x)$
1	Heads	0.7
0	Tails	0.3

### ✅ PMF Formula:

$$P(X = x) = \begin{cases} 0.7 & \text{if } x = 1 \\ 0.3 & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

## ✓ 1. Simple Random Sampling

| Every item has an **equal chance** of being selected.

● **How:** Use random number generator or lottery method.

- **Example:** Picking 10 students randomly from a class of 100.

## ✓ 2. Stratified Sampling

| Population is divided into **strata (groups)**, then random samples are taken from each group.

● **How:** Divide by gender, age, income group, etc., and sample from each.

- **Example:** Selecting 5 boys and 5 girls from a class with 50 boys & 50 girls.

## ✓ 3. Systematic Sampling

| Select every **k-th item** from a list after a random starting point.

● **How:** Choose every 5th name in a list after randomly starting at, say, 3rd name.

- **Example:** From a list of 1000 people, select every 10th person.

## ✓ 4. Convenience Sampling

| Select the **easiest available** members of the population.

● **How:** Ask whoever is nearby or available.

- **Example:** Surveying people at the entrance of a mall.

## 5 Number summary and how to handle outliers using IQR (Interquartile Range)

- Minimum
- 25% (Q1)
- Median
- 75% (Q3)
- Maximum

Percentile formula =  $x/100 * (n+1)$

X = 25, 50, 75 etc.

- The result denoted the index that value corresponds to the index denotes that 25% of values are less than that index. (n should be in sorted manner).

E.g, 1,2,3,4,5,5,6,7,8,9,9,10

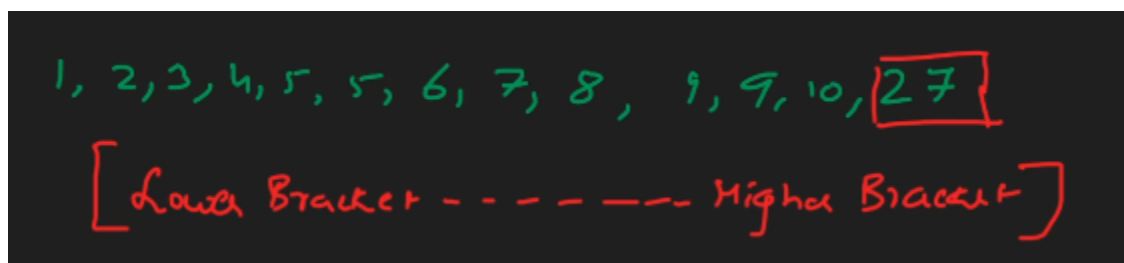
For 25, index = 3.25 i.e 3th index which is value 3

For 75, index = 10 i.e 9

Therefore,

- Minimum = 1
- 25% = 3
- Median = 6
- 75% = 9
- Maximum = 10

IQR = Q3 - Q1 = 6 (spread of 50%)



To remove the outlier we need to find the values which will lie between higher and lower brackets. Anything beyond the range can be omitted.

Lower =  $Q1 - 1.5 * (IQR)$     ||    Higher =  $Q3 + 1.5 * (IQR)$

From our example,

$L.b = -6$

$U.B = 18$ , thus we can remove 27

---

### ✓ Compare:

Term	Meaning
False Positive (FP)	Predict <b>Yes</b> , but actually <b>No</b>
False Negative (FN)	Predict <b>No</b> , but actually <b>Yes</b>

### ✓ Which to reduce? Depends on scenario:

Use Case	Reduce	Why?
Cancer Detection	False Negative	Missing a real cancer is dangerous
Spam Filter	False Positive	Don't mark real emails as spam
Loan Approval	False Positive	Don't approve bad customers
Fraud Detection	False Negative	Catch all fraud, even if a few false alarms happen

---

Z score & its application (**standardization**, **compare scores b/w diff distribution**, )

**SD**: Describes **spread** of all data

**Z-score**: Describes **position** of a single data point within that spread

Formula same as **standard scaler**

E.g, compare

The image shows handwritten calculations for Z-scores on a dark background. On the left, under the heading 'India', are the statistics for 2020: Avg = 181,  $\sigma = 12$ , and Final Score = 187. The Z-score calculation is  $Z_{2020} = \frac{187 - 181}{12} = \frac{6}{12} = 0.5$ . On the right, for 2021: Avg = 182,  $\sigma = 5$ , and Final Score = 185. The Z-score calculation is  $Z_{2021} = \frac{185 - 182}{5} = \frac{3}{5} = 0.6$ , with the final result underlined.

<u>India</u>	2020 (Cricket)	2021 (Cricket)
	Avg = 181	Avg = 182
	$\sigma = 12$	$\sigma = 5$
	Final Score = 187	Final Score = 185
	$Z_{2020} = \frac{187 - 181}{12} = \frac{6}{12} = 0.5$	$Z_{2021} = \frac{185 - 182}{5} = \frac{3}{5} = 0.6$

0.6 > 0.5

---

Power Law Distribution

A relative change in one quantity revokes in a proportional change in other quantity (80-20 rule).

$$P(x) \propto x^{-\alpha}$$

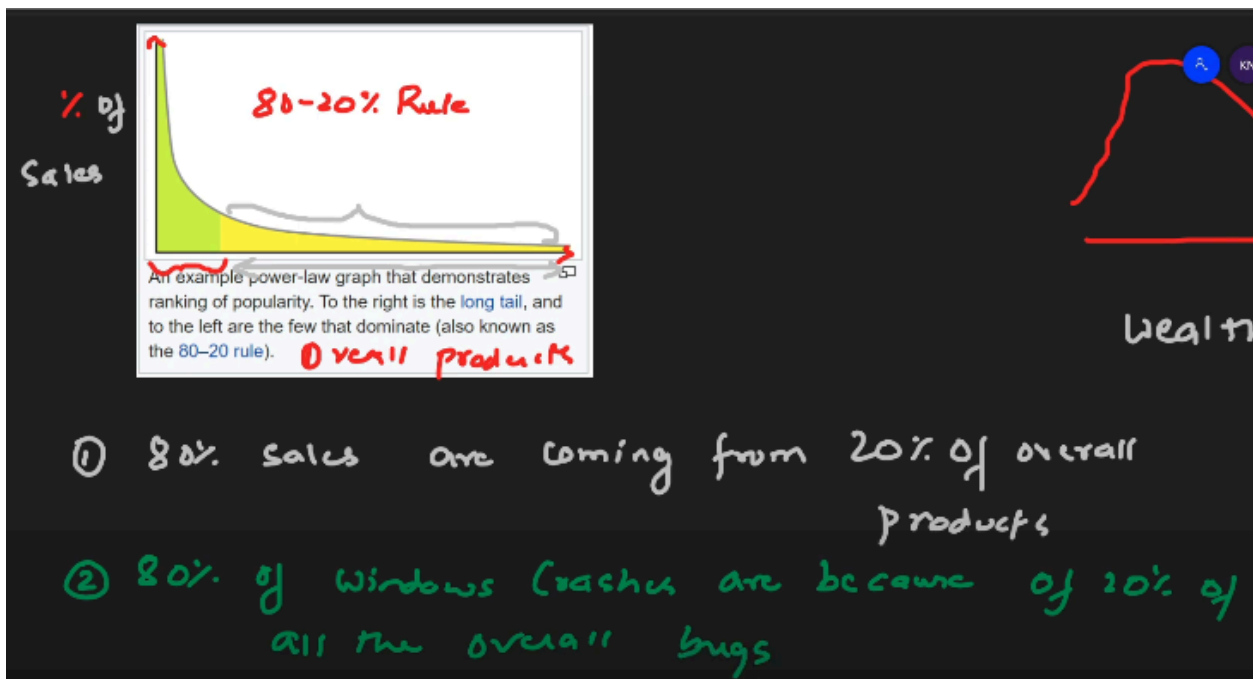
Where:

- $x$  = value
- $\alpha > 1$  = exponent (controls the "steepness")
- $P(x)$  = probability of  $x$

### • Key Properties:

Feature	Meaning
Heavy tail	Large values have <b>non-zero probability</b>
Scale-free	Pattern looks same across scales
No fixed mean/variance	If $\alpha \leq 2$ , mean $\rightarrow \infty$

E.g



Pareto Dist.

### ✓ Relation:

Term	Description
Power law	General rule: $P(x) \propto x^{-\alpha}$
Pareto distribution	A specific case where the distribution follows:

$$P(X > x) = \left(\frac{x_m}{x}\right)^\alpha \quad \text{for } x \geq x_m$$

### ✓ Difference in View:

Feature	Power Law	Pareto Distribution
General or specific	General formula	Specific statistical distribution
Domain	Applies to various real-world data	Often used in economics, wealth, etc.
Function form	$x^{-\alpha}$	$\left(\frac{x_m}{x}\right)^\alpha$

---

Important topics of Probability

- Mutually exclusive event - Two events that **cannot happen at the same time**. (tossing coin)
- Not mutually exclusive event - Two events that **can occur together**. (taking card)

Handwritten formulas on a chalkboard:

ME  $Pr(A \text{ or } B) = Pr(A) + Pr(B)$

NME  $Pr(A \text{ or } B) = Pr(A) + Pr(B) - Pr(A \cap B)$



Multiplicative Rule

① Independent Event

⊕ Tossing a Coin

2 Outcomes  
H, T, H  
 $\{1, 2, 3\}$

$Pr(H) = \frac{1}{2}$     $Pr(H) = \frac{1}{2}$   
 $Pr(T) = \frac{1}{2}$

No. of Outcomes will not reduce

Eg: Rolling a dice  $\{Pr(1 \text{ and } 3)\}$

Independent Event  
 $\rightarrow Pr(A \text{ and } B) = Pr(A) \times Pr(B)$   
 $Pr(1 \text{ and } 3) = Pr(1) \times Pr(3)$

② Dependent Event

Eg- Taking a card from a deck

1 Experiment  $\rightarrow$   $\boxed{K}$     $\rightarrow$  2nd Expt  $\rightarrow$   $\boxed{Q}$  & red  
3rd Expt  $\rightarrow$   $\boxed{J}$

$Pr(K) = \frac{1}{52}$     $Pr(Q) = \frac{1}{51}$     $Pr(J) = \frac{1}{50}$

$\nearrow$  It will reduce

Eg:  
 $Pr(K \text{ and } Q)$    Conditional Events  
 $Pr(A \text{ and } B) = Pr(A) \times Pr(B/A)$

## Permutation and Combination

✓ In ML – Where It Is Used	
Area	Use of Perm/Comb
Feature selection	Choosing k features from n (Combinations)
Hyperparameter tuning	Trying all settings (Permutations)
Data augmentation	Rearranging data or combinations of input
Cross-validation	Choosing different subsets

## • 1. Permutation (Order matters)

Number of ways to **arrange**  $r$  items out of  $n$ .

$${}^nP_r = \frac{n!}{(n-r)!}$$

### • Example:

Arrange 3 books from 5 →

$${}^5P_3 = \frac{5!}{(5-3)!} = 60$$

## • 2. Combination (Order doesn't matter)

Number of ways to **choose**  $r$  items out of  $n$ .

$${}^nC_r = \frac{n!}{r!(n-r)!}$$

### • Example:

Choose 3 students from 5 →

$${}^5C_3 = \frac{5!}{3! \cdot 2!} = 10$$

---

Why Sample Variance is Divided by  $n-1$

Use  $n-1$  when working with **samples**

Use  $n$  only when working with **entire population**

We divide by **(n - 1)** to correct for **bias** — this is called **Bessel's correction**.

- **Explanation:**

- In sample variance, we use **sample mean** ( $\bar{X}$ ) instead of population mean.
- Since  $\bar{X}$  is calculated from the same data, it **underestimates variability**.
- Dividing by **n - 1** gives an **unbiased estimate** of true population variance.

---

- **1. Covariance**

- **What it tells:** Direction of relationship between two variables
- **Formula:**

$$\text{Cov}(X, Y) = \frac{1}{n} \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

- **✗ Disadvantages:**

Point	Why it's a problem
No fixed scale	Values are hard to interpret directly
Units-dependent	Changes with unit (e.g., cm vs m)
Can't compare across datasets	Due to scaling issue

- **2. Pearson Correlation Coefficient (r)**

- **What it tells:** **Linear** relationship (-1 to +1 scale)
- **Formula:**

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

**✗ Disadvantages:**

Point	Why it's a problem
Only measures <b>linear</b> relation	Fails with non-linear patterns
Sensitive to <b>outliers</b>	Outliers can distort the value
Assumes <b>normal distribution</b>	Not robust for skewed data

- **3. Spearman Rank Correlation (ρ)**

- **What it tells:** **Monotonic** relationship using **ranks**
- **Formula:**

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

**✗ Disadvantages:**

Point	Why it's a problem
Less precise for linear patterns	Loses actual magnitude (uses ranks)
Ties in ranks can affect accuracy	Especially in small datasets
Slower to compute for large datasets	Due to sorting and ranking

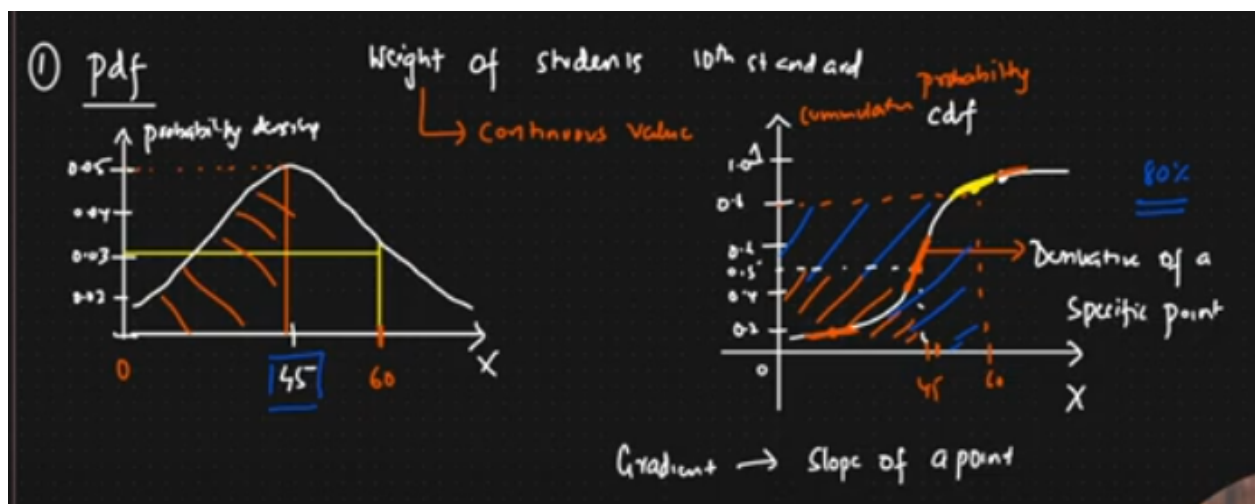
---

## Probability distribution Func

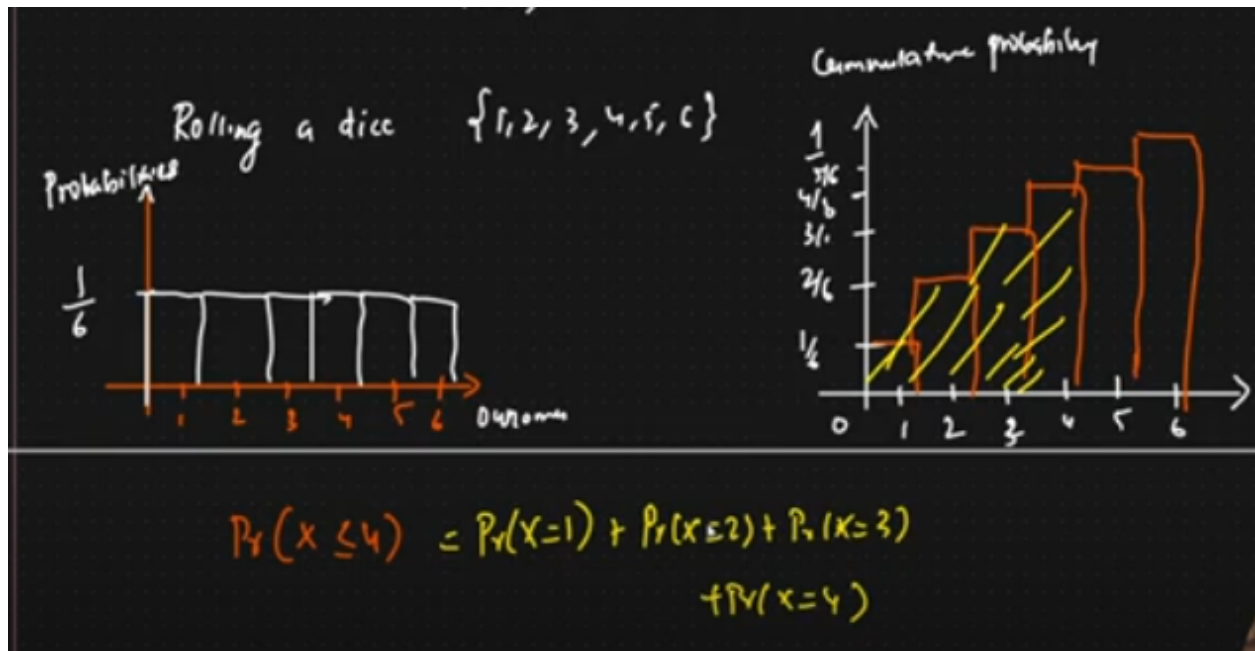
- 1) Probability Density Function -> Continuous random variable
- 2) Probability Mass Function -> Discrete random variable
- 3) Cumulative Distribution Function -> area

PDF -> is the derivative of CDF at any point

## Bell Curve



PMF ->



## 1. PMF (Probability Mass Function)

**Significance:** Used for **discrete random variables** (e.g., number of heads in coin toss).

**Problem Statements it solves:**

- What is the probability of getting exactly 3 heads in 5 tosses?
  - What is the probability of rolling a 6 on a die?
  - In a dataset of defects per product, what's the chance of getting exactly 2 defects?
- 

## 2. PDF (Probability Density Function)

**Significance:** Used for **continuous random variables** (e.g., height, weight, time).

**Problem Statements it solves:**

- What's the probability density of a person being 170 cm tall?
- Modeling time taken to complete a task.
- What's the likelihood that a signal amplitude lies between 1.5V and 2.0V?

Note: PDF gives density, not exact probability. Probability is found over an **interval** using PDF.

---

## 3. CDF (Cumulative Distribution Function)

**Significance:** Gives **cumulative probability** —  $P(X \leq x)$

**Problem Statements it solves:**

- What is the probability of getting **at most** 2 defects?
  - What is the chance a person's weight is less than 65 kg?
  - Find percentile ranking of a test score.
-

✓ **Entropy** is a measure of **uncertainty** or **randomness** in data.

### 🔗 Why it's important in MI?

Mutual Information (MI) is **based on entropy**. It tells how much **uncertainty about one variable** (e.g., target) is **reduced by knowing another** (e.g., a feature).

### 📄 Entropy Formula:

For a variable  $X$  with possible values  $x_1, x_2, \dots, x_n$ :

$$H(X) = - \sum P(x_i) \cdot \log_2 P(x_i)$$

- If all values are equally likely → high entropy (max uncertainty)
- If one value dominates → low entropy

### 🔗 MI Formula:

$$MI(X, Y) = H(X) + H(Y) - H(X, Y)$$

→ How much knowing  $X$  reduces entropy (uncertainty) in  $Y$

### 🧠 In Feature Selection:

- Compute **MI between each feature and the target**
- Feature with **higher MI = more informative**

### ✓ Example:

- Target: `Rain` (Yes/No)
- Feature: `Humidity`
  - If `Humidity` = High always when `Rain` = Yes → MI is high
  - If `Humidity` has no pattern with `Rain` → MI is near 0

Entropy ?

---



**EDA (Exploratory Data Analysis)** in ML means:

- **Definition**

EDA is the process of **analyzing and visualizing** data to:

- Understand its structure
- Identify patterns
- Detect anomalies
- Test assumptions
- Prepare for modeling

- **Significance**

- Helps decide which ML model to use
- Improves model accuracy
- Saves time during debugging
- Reveals missing or corrupt data
- Highlights data distribution and outliers

- **Why Required?**


- Garbage in → garbage out.
- Without understanding data, model training is **blind** and often wrong.

---

- **Categorical Features vs Target**

Test	When to Use	Purpose / Significance
Chi-Square Test	Categorical feature + Categorical target	Tests independence between variables.
ANOVA (F-test)	Categorical target + Continuous feature	Tests variance between groups (for >2 classes).
Mutual Information (MI)	Any feature type	Measures information gain (non-linear too).
Cramér's V	Categorical + Categorical	Strength of association between two categorical vars.

### • Numerical Features vs Target

Test	When to Use	Purpose / Significance	
T-Test	Binary target + Continuous feature	Compare means between two groups.	
ANOVA (F-test)	Multi-class target + Continuous feature	Compare variance between >2 groups.	
Pearson Correlation	Continuous target + Continuous feature	Measures linear relationship strength.	
Spearman Rank Correlation	Continuous + Continuous (non-linear)	Measures monotonic relationships.	
Mutual Information (MI)	Any combination	Captures both linear & non-linear dependencies.	

### • Wrapper & Embedded Methods (Not statistical, but popular)

Method	Type	Why Useful
RFE (Recursive Feature Elimination)	Wrapper	Selects features by recursively removing least important ones.
Lasso (L1 Regularization)	Embedded	Penalizes irrelevant features to shrink them to 0.
Tree-based feature importance	Embedded	Uses decision tree splits to rank feature importance.