Types of Data

Statistic For ML

Introduction

This document is part of a structured learning process focused on building a **strong statistical foundation for Machine Learning**.
Rather than memorizing formulas or relying on library functions, the emphasis is on **understanding how data behaves**, how it should be interpreted, and how incorrect assumptions can lead to misleading model results.

Statistics plays a critical role in Machine Learning, influencing data preprocessing, feature selection, model behavior, and performance evaluation. Many model failures are not due to algorithmic limitations but due to **misinterpretation of data types, scales of measurement, and statistical properties**.

Objectives

- Develop clarity in identifying different types of data
- Understand scales of measurement and their valid operations
- Prevent common statistical mistakes that lead to invalid models
- Build intuition that directly connects statistical concepts to Machine Learning decisions

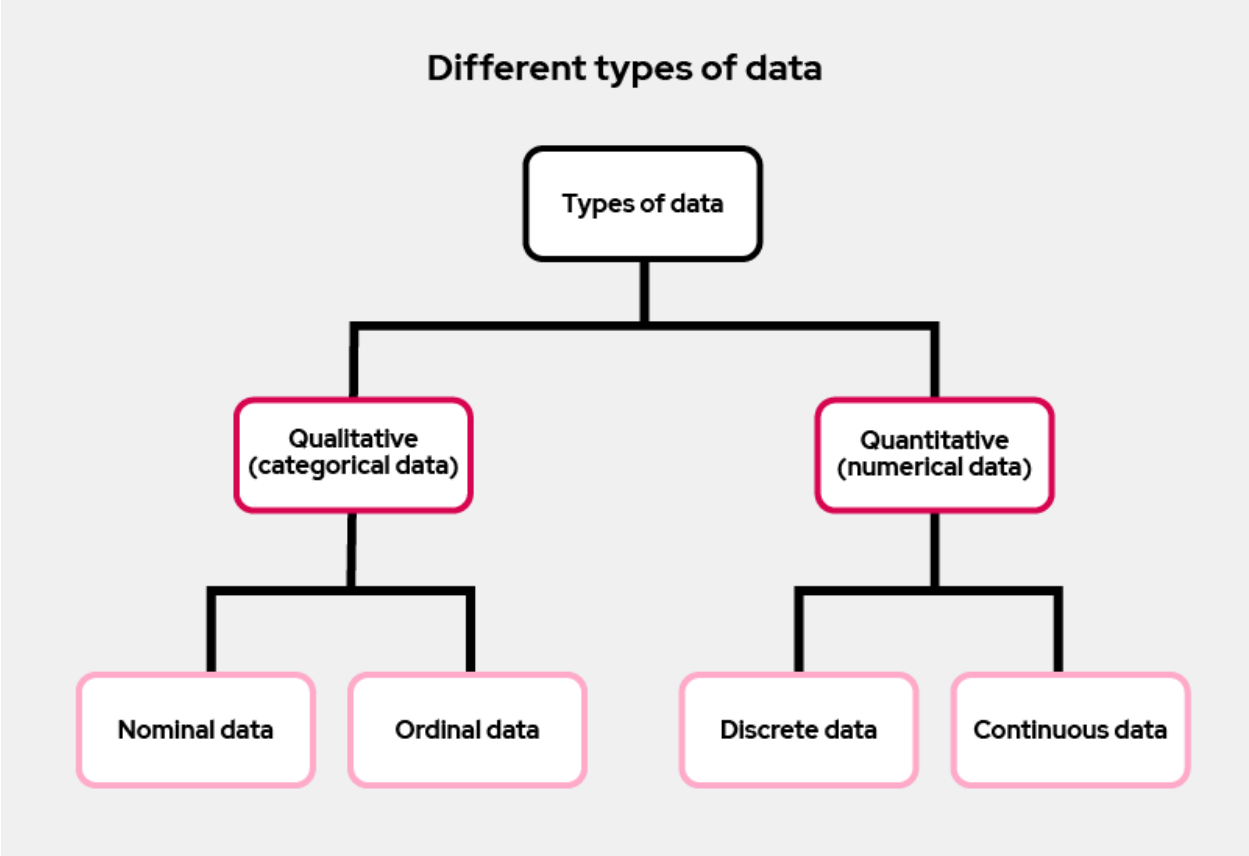
This document avoids unnecessary mathematical complexity and instead prioritizes **conceptual clarity, logical reasoning, and real-world relevance**. Each section is designed to reinforce practical understanding that can be applied directly to data analysis and machine learning workflows.

This serves as a foundational reference for subsequent topics such as data preprocessing, regression, classification, and model evaluation.

1. Types of Data

There are two Types of Data :

1. Numerical
2. Categorical



A . Numerical data :

1. **Discrete data**
2. **Continuous data**

B . Categorical Data:

1. **Nominal data**
2. **Ordinal Data**

Numerical Data :

Numerical data is data that represents **quantities** — values where **mathematical operations make sense** and the results actually mean something in the real world.

a. Discrete Data :

Discrete data is a type of **numerical data** that comes from **counting**, not measuring.It can take **only specific, separate values** — usually whole numbers.There are **gaps** between possible values.

Examples :

- Number of students in a class → 35
- Number of cars sold in a day → 12
- Number of defects in a product → 0, 1, 2, 3
- Number of goals scored → 2

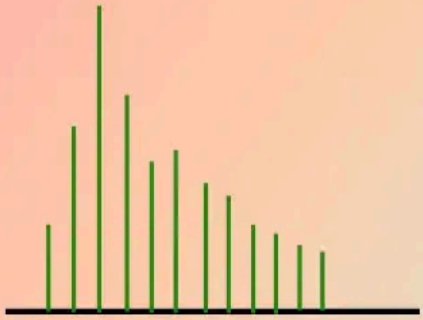
b. Continuous data:

Continuous data is a type of **numerical data** that comes from **measuring**, not counting.It can take **any value within a range**, including fractions and decimals.Between any two values, **another value is always possible**.

Examples :

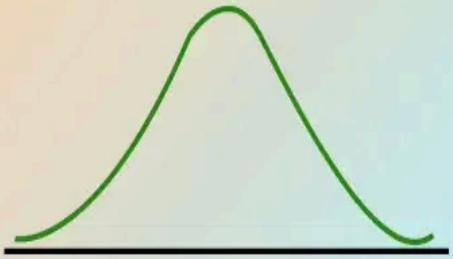
- Height = 172.63 cm
- Weight = 63.247 kg
- Time = 2.381 seconds
- Distance = 15.75 meters

Difference Between Discrete & Continuous Variable



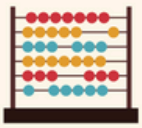
DISCRETE

- Obtained by counting values such as integers 0,1,2,3...
- Example: Your score in this upcoming mid-term exams



CONTINUOUS

- Obtained from the data that can take infinitely many values
- Example: The expected lifetime of a new light bulb

Points	Discrete Data	Continuous Data	
Meaning	Discrete data has clear spaces between values.	Continuous data falls on a continuous sequence.	
Can you count the data?	Yes, data is usually units counted in whole numbers.	Generally, NO	
Can you measure the data?	NO	YES	
Values	It has a finite number of possible values. The values cannot be divided into smaller pieces and add additional meaning.	It has an infinite number of possible values within an interval. The values can be subdivided into smaller and smaller pieces.	
Graphical Representation	Bar Chart	Histogram	
Examples	• The number of students in a class. • The number of workers in a company. • The number of parts damaged during transportation. • Shoe sizes. • Number of languages an individual speaks. • The number of home runs in a baseball game. • The number of test questions you answered correctly.	• The amount of time required to complete a project. • The height of children. • The amount of time it takes to sell shoes. • The amount of rain, in inches, that falls in a storm. • The square footage of a two-bedroom house. • The weight of a truck. • The speed of cars. • Time to wake up.	 

Categorical Data :

Categorical data is data that represents **groups, labels, or categories**, not quantities.If arithmetic operations **do not make sense**, the data is categorical.

a. Nominal Data :

- Nominal data is a type of **categorical data**
- Values are **names, labels, or identifiers**
- There is **no inherent order** between values
- Numbers (if used) have **no numerical meaning**
- No ranking or hierarchy
- No meaningful distance between categories
- Arithmetic operations are not allowed.

Examples :

- Gender (Male, Female)
- City (Chennai, Mumbai, Delhi)
- Blood group (A, B, AB, O)
- PIN code
- Phone number

Nominal Data — Notes

- Nominal data is a type of categorical data
- Values are names, labels, or identifiers
- There is **no inherent order** between values
- Numbers (if used) have **no** numerical meaning

Key Characteristics

- No ranking or hierarchy
- No meaningful distance between categories
- Arithmetic operations are **not allowed**

Examples

- Gender (Male, Female)
- City (Chennai, Mumbai, Delhi)
- Blood group (A, B, AB, O)
- PIN code
- Phone number
- Nationality

Not Allowed Operations

- Mean or median
- Subtraction or addition
- Distance calculations

Common Mistake

Encoding nominal data with numbers (e.g., Male = 0, Female = 1)
→ Encoding **does not** make it numerical data

ML Perspective

- Nominal data must be encoded before modeling
- Distance based models are sensitive to nominal features

> "**Nominal data is categorical data used purely** for labeling or identification, with **no order** or quantitative meaning."

b. Ordinal data:

Ordinal data is a type of **categorical data** where the values have a **meaningful order**, but the **difference between values cannot be measured**.

What we can do

- Rank values
- Compare categories
- Find the median

What we cannot do

- Calculate mean (usually misleading)
- Subtract values
- Use ratios

Examples :

- Ratings: 1, 2, 3, 4, 5
- Performance: Poor < Average < Good
- Satisfaction: Low < Medium < High
- Education level: School < UG < PG

Ordinal Data — Notes

- Ordinal data is a type of categorical data
- Categories have a defined order
- Distance between categories is **not** measurable
- Arithmetic operations are **not** valid

Examples

- Ratings (1–5)
- Poor < Average < Good
- Low < Medium < High
- Education levels

Allowed Operations

- Ranking
- Comparison
- Median

Not Allowed Operations

- Mean
- Subtraction
- Ratio calculations

Common Mistake

- Treating ordinal data as numerical because numbers are used

One-Line Definition

> "**Ordinal data is categorical data** with a meaningful order but without measurable differences between categories."

Points	Nominal Data	Ordinal Data	
Meaning	Nominal data are those items which are distinguished by a simple naming system. They are data with no numeric value, such as profession. The nominal data just name a thing without applying it to an order related to other numbered items.	Ordinal data is data which is placed into some kind of order by their position on the scale. For example, they may indicate superiority. However, you cannot do arithmetic with ordinal numbers because they only show sequence.	
Are they categorical?	Yes, nominal data are also called categorical data.	Ordinal variables are "in between" categorical and quantitative variables.	
The level of quantitative value	Without any type of quantitative value.	We can assign numbers to ordinal data but we cannot do arithmetic with ordinal numbers.	
Key Points	• Nominal data cannot be quantified. • It also cannot be assigned to any type of order. • The values are only allocated to distinct categories. • Those categories have no meaningful order.	• Ordinal data is placed into some kind of order. • Ordinal numbers only show sequence. • We can assign numbers to ordinal data. • We cannot do arithmetic with ordinal numbers. • We don't know if the differences between the values are equal.	
Examples	• Gender (Women, Men) • Religion (Muslin, Buddhist, Christian) • Hair color (Blonde, Brown, Brunette, Red, etc.) • Marital status (Married, Single, Widowed) • Ethnicity (Hispanic, Asian) • Eye color (Blue, Green, Brown).	• The first, second and third person in a competition. • Education level: the elementary school, high school, college. • Customer rating of the sales experience on a scale of 1-10. • Letter grades: A, B, C, and etc. • Economic status: low, medium and high. • Customer level of satisfaction: very satisfied, satisfied, neutral, dissatisfied, very dissatisfied.	

Scale of Measurement

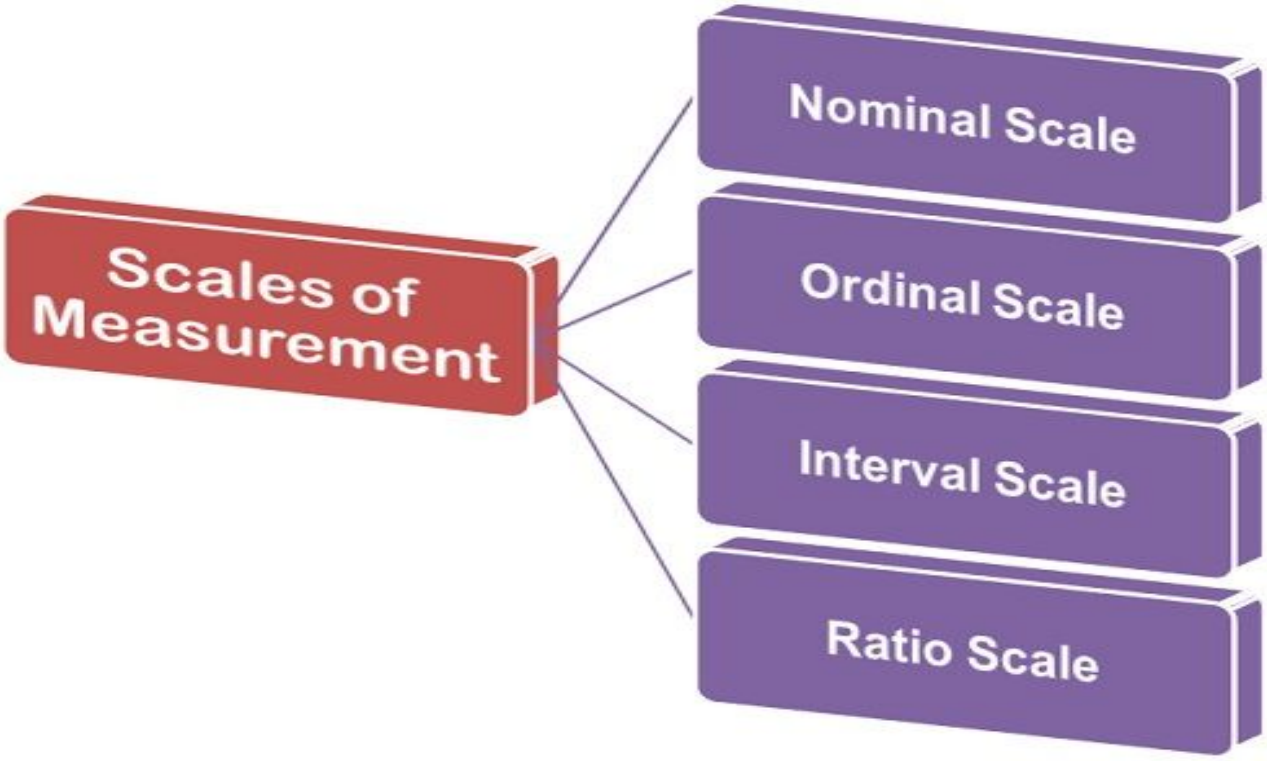
Scale Of Measurement

Scale of measurement defines **how much information a variable carries** and **what mathematical operations are valid** on it.

In simple words:

It tells **what you are allowed to do with the data**.

Same data value → different scale → different rules.



Nominal Scale

Nominal scale is the **lowest level of measurement** where data is used **only for naming or labeling**.

Core Characteristics (non-negotiable)

- ❌ **No order**
- ❌ **No ranking**
- ❌ **No magnitude**
- ❌ **No distance**
- ❌ **No arithmetic meaning**

The values are just labels.

Examples (lock these in)

- **Gender (Male, Female)**
- **Blood Group (A, B, AB, O)**
- **City (Chennai, Mumbai, Delhi)**
- **Country**
- **Customer ID**
- **Phone number**
- **Email ID**

Even if numbers are used → still nominal.

Ordinal Scale :

Ordinal scale is a level of measurement where data values **can be ordered or ranked**, but the **difference between values is not measurable**.

Core Characteristics (remember these)

- ✅ **Order exists**
- ❌ **Distance is unknown**
- ❌ **No true zero**
- ❌ **Ratios are meaningless**

Order ≠ quantity..

Common Examples

- **Satisfaction: Low < Medium < High**
- **Performance: Poor < Average < Good**
- **Education level: School < UG < PG**
- **Pain level: Mild < Moderate < Severe**
- **Ratings: 1, 2, 3, 4, 5**

Even though numbers are used, they represent rank only.

Interval Scale:

Interval scale is a level of measurement where:

- Data values have a **clear order**
- The **difference between values is meaningful and equal**
- BUT **zero does NOT mean absence**

Core Characteristics (lock these)

- ✅ Order exists
- ✅ Equal intervals (gaps are the same)
- ❌ No true zero
- ❌ Ratios are meaningless

Common Examples

- Temperature in Fahrenheit (°F)
- Calendar years (2010, 2020, 2025)
- Time of day (hours on a clock)

NOTE :

Interval scale measures ordered data with equal intervals between values, but without a meaningful zero point.

Ratio Scale:

Ratio scale is the highest level of measurement where:

- Data has a **clear order**
- **Equal intervals** between values
- A **true, meaningful zero** exists

Core Characteristics (lock these forever)

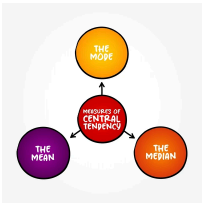
- ✅ Order exists
- ✅ Equal gaps
- ✅ **True zero exists**
- ✅ Ratios are meaningful

This is the **only scale** where multiplication and division make sense.

Examples

- **Age (years)**
- **Height (cm, m)**
- **Weight (kg)**
- **Distance (km)**
- **Blood Pressure (mmHg)**
- **Cholesterol level**
- **Time duration (seconds, minutes)**
- **Salary / Income**

Measure of Central Tendency

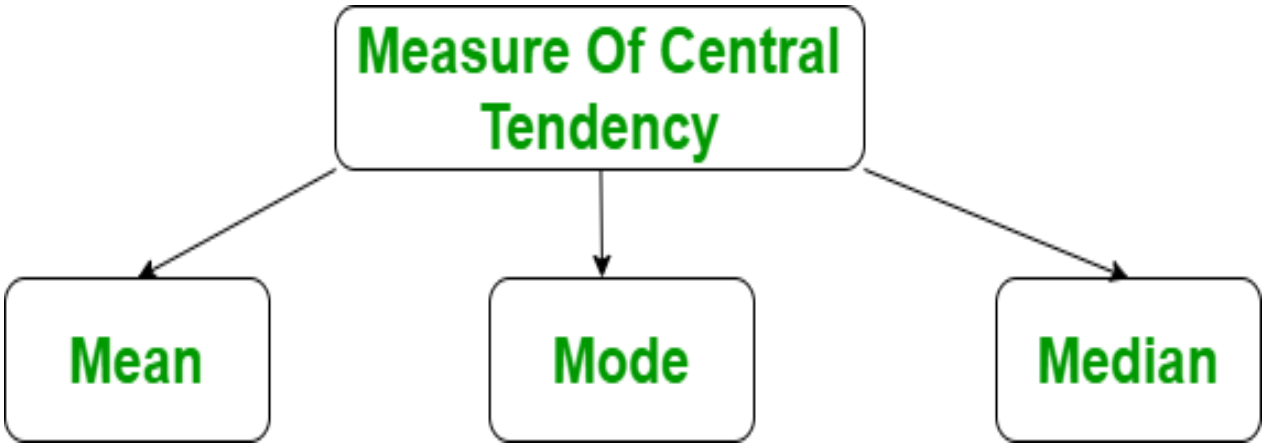


Measure of Central Tendency

Central tendency answers one question only:

"What single value best represents this entire data?"

That's it.
Mean, Median, Mode are three different answers to that same question — used in different situations.



Mean

Mean (commonly called *average*) is the value obtained by **adding all observations and dividing by the number of observations**.

Formula

Mean = Number of values / Sum of all values

Example

Data:

`10, 20, 30, 40, 50`

Sum = 150

Count = 5

Mean = 150 / 5 = **30**

Here, mean makes sense because the data is **balanced**.

When TO Use Mean

Use the mean **only if ALL of these are true**:

1. Data is numerical

Examples:

- Age
- Height
- Weight

2. Data is on interval or ratio scale

- Temperature (°C) → Interval
- Age, height, weight → Ratio

If subtraction and averaging make sense → mean is allowed.

3. Data is roughly Symmetry

Mean works best when:

- No heavy skew
- Values spread evenly

4. No extreme outliers

Mean assumes:

- No single value dominates the data

  Example:

  Exam scores in a well-designed test.

Purpose of Mean

The **mean** is used to **summarize numerical data with a single representative value** that captures the **overall level** of the data.

Purpose 1: Find a Central Reference Point

Examples:

- Age
- Height
- Weight

Purpose 2: Compare Groups Fairly

Mean allows comparison across groups.

Example:

- Average marks of Class A vs Class B
- Average temperature of two cities

Purpose 3: Minimize Squared Error (ML & Math)

The mean is the value that:

> **Minimizes the total squared deviation from all data points**

That's why:

- Regression predicts values around the mean
- MSE is built around the mean
- Mean is the "optimal center" mathematically

Purpose 4: Act as a Baseline

In regression:

- Predicting the mean is the **simplest possible model**
- Any ML model must beat this baseline

If your model performs worse than mean prediction → model is useless.

Median

Median is the **middle value** of a dataset **after arranging the data in order**

In simple words:

Median answers: "What is the central value when data is ordered?"

How to Find the Median

Step 1: Sort the data

Either Ascending or Descending

Step 2: Check the number of observation

Case 1 : Odd number of values

Let consider a data : 10, 20, 30, 40, 50

Middle value = 30

Median = 30

Case 2: Even number of values

Let consider the data : 10, 20, 30, 40

Two middle values = 20 and 30

Median = (20 + 30) / 2 = 25

When to Use Median

- Data is **skewed**
- Extreme values exist

Purpose of Median

The **median** is used to represent the typical central value of data when order matters more than magnitude.

Purpose 1: Represent the “Typical” Value

Examples:

- Values are not evenly spread
- Extremes exist

Purpose 2: Handle Outliers Safely

Median ignores how far values are,

So,

- Extreme values do not pull it
- It stays stable

Purpose 3: Work with Ordinal Data

Median depends only on order, not distance.

So it is Valid For :

- **Ratings**
- **Ranks**
- **Satisfaction levels**

Purpose 4: Split Data into Two Equal Halves

Median ensures:

- 50% of values are below
- 50% of values are above

Median-based models don't explode like mean-based ones.

Purpose 5: Provide Robust Baselines

In Machine Learning:

- Minimizing **absolute error (MAE)** leads to median
- Used in **robust regression**

Median-based models don't explode like mean-based ones.

Purpose 6: Summarize skewed Distribution

When data is:

- Right-skewed (income, house prices)
- Left-skewed (easy exams)

Median gives an **honest summary**.

Median Baseline VS Mean Baseline

Mean baseline vs Median baseline

Mean baseline

- Predicts the **mean** of the target
- Optimal when minimizing **squared error (MSE)**
- BUT extremely sensitive to outliers

Median baseline

- Predicts the **median** of the target
- Optimal when minimizing **absolute error (MAE)**
- **Robust to outliers**

This is the key difference.

Mode :

Mode is the value that occurs most frequently in a dataset.

In simple words:

Mode answers: “Which value appears the most?”

No math. No averages. Just frequency.

How to Find the Mode

Step 1: Count how many times each value appears

Step 2: The value with the highest count is the mode

When To Use Mode

- Use mode to find the most common value
- Use mode when data is categorical
- Use mode when data is in names or labels
- Use mode when mean and median don’t make sense

Easy Example :

- Most common gender → Mode
- Most common blood group → Mode

Purpose of Mode

- To find the **most common value**
- To know **what appears the most**
- To summarize **categorical data**
- To identify **popular / frequent choices**
- To handle data where **mean and median are not useful**
- To understand **dominant category or class**

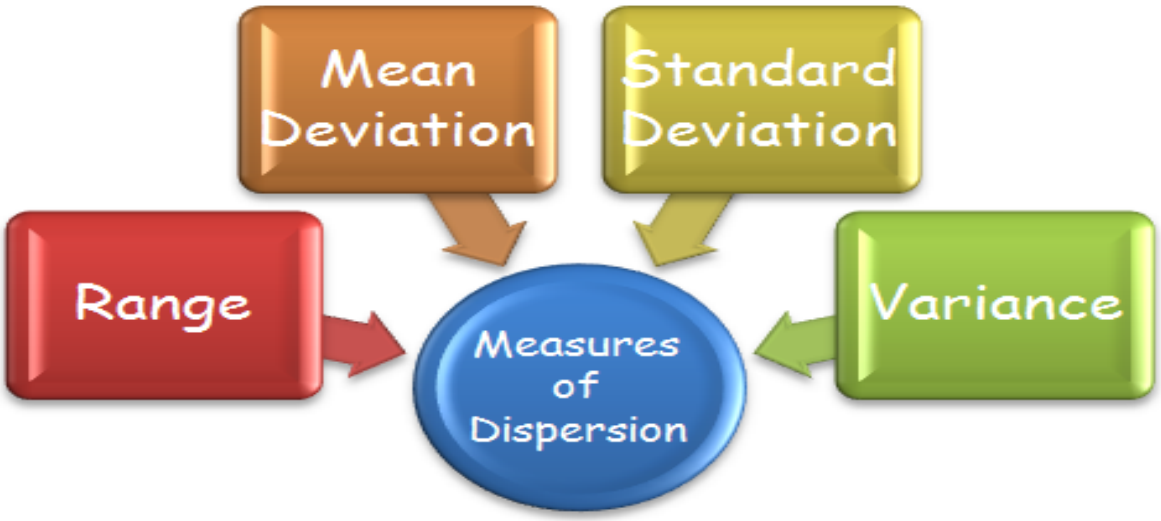
Measure of Dispersion

Measure of Dispersion

Dispersion tells:

How much the data values spread away from the center

Two datasets can have the same mean, but very different dispersion.



Range

How far the data spreads from the smallest value to the largest value.
It gives a quick idea of spread.

Formula

Range = Maximum value − Minimum value

Example

Data:

`10, 20, 90, 120, 150, 130`

Max = 150

Min = 10

Range = 150 - 10 = **140**

What Range Tells Us

- Small range → data is **closely packed**
- Large range → data is **widely spread**

Where Range is Useful

- Quick comparison of datasets
- Rough idea of variability
- Small datasets
- Initial data exploration

Range In Machine Learning

- Used only for **initial understanding**
- Rarely used alone
- Helps spot extreme values quickly

Limitation In Range

- ❌ Uses only **two values**
- ❌ Ignores all middle values
- ❌ Very sensitive to **outliers**

Example

Data = [10, 24, 34, 16, 160]

Range = 160 - 10 = 150

But Values are too close, so the range can mislead.

Standard Deviation

On average, how far the data values are from the mean

In simple words, It answer :

How much do values usually differ from the average?

Why is Standard Deviation Needed ?

Two datasets can have the **same mean**, but:

- one can be **tightly packed**
- one can be **widely spread**

Standard deviation tells how consistent or inconsistent the data is.

Small vs Large Standard Deviation :

Small SD → values are close to the mean (stable data)

Large SD → values are far from the mean (spread out data)

Standard Deviation - Step-by-Step Calculation

Data: 10, 20, 30, 40, 50

Mean = **30**

1 Step 1: Find the Mean

$$\text{Mean} = \frac{(10+20+30+40+50)}{5} = 150/5 = 30$$

Mean = **30**

2 Step 2: Find Deviation from Mean

10 – 30 = –20	20 – 30 = –10	40 – 30 = –20
20 – 30 = –10	30 – 30 = 0	50 – 30 = +10
40 – 30 = +10	40 – 30 = +10	50 – 30 = +20

3 Step 3: Square the Deviations

$(-20)^2 = 400$
$(-10)^2 = 100$
$0^2 = 0$

$(-20)^2 = -100$
$(+10)^2 = 100$
$(+20)^2 = 100$

$-100 = 0^2 = 100$
$0^2 = 0^2 = 100$
$(10)^2 = 100^2 = 300$

$(+10) = 400$
$(+20)^2 = 400$

4 Step 4: Find Average of Squared Deviations

$$\text{Variance} = \frac{(400+100+0+100+400)}{5} = 1000/5 = 200$$

5 Step 5: Take Square Root

$$\text{Standard Deviation} = \sqrt{200} = \sqrt{20} \approx 14.14$$

Standard Deviation ≈ 14

Why do we take square roots?

Because,

Variance is in **squared units** (cm^2, kg^2 ❌)

SD brings it back to **original units** (cm, kg ✅)

This makes SD **easy to understand**.

When to Use Standard Deviation

- Data is **numerical**
- You want to measure **consistency**
- Comparing variability between datasets
- Used in **ML preprocessing** (standardization)

Variance

On average, how far the data values are from the mean

In simple words, It answer :

Variance is the average of squared deviations from the mean.

Why is Variance Needed ?

Variance tells:

- **how scattered the data is around that centre**

Two data may have same mean but doesn't have same variance

Distribution

Distribution

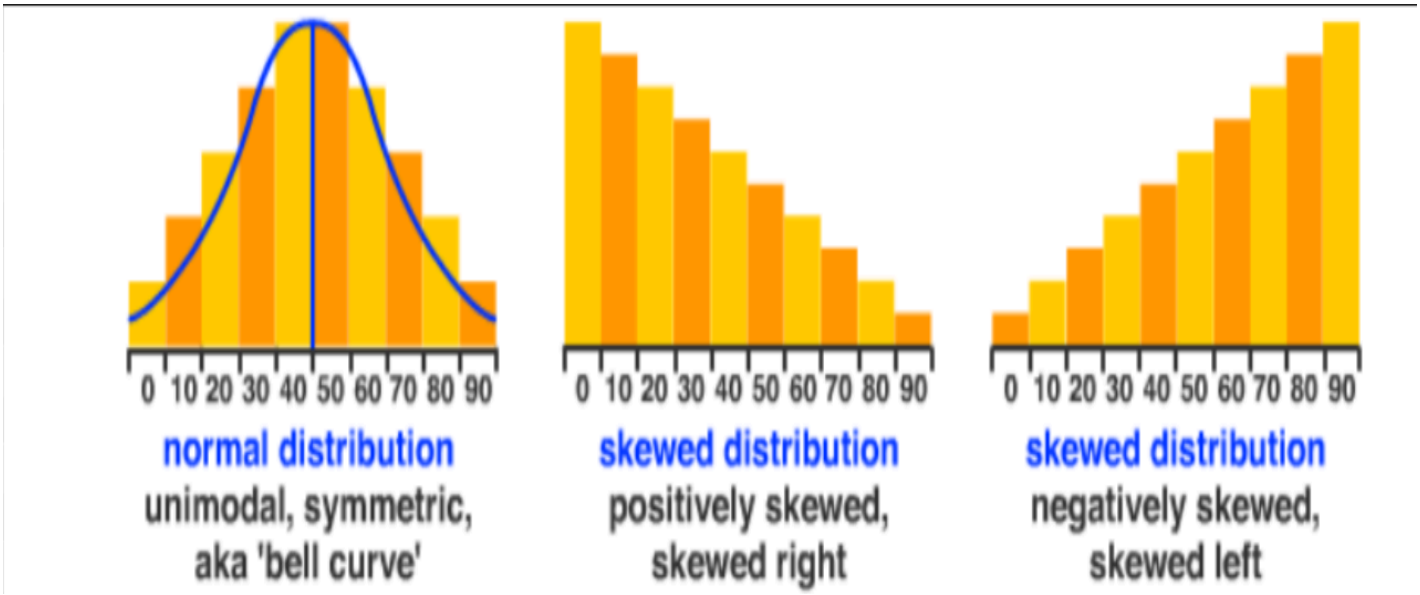
Distribution describes:

-How data values are spread across different values and how frequently they occur

In plain words:

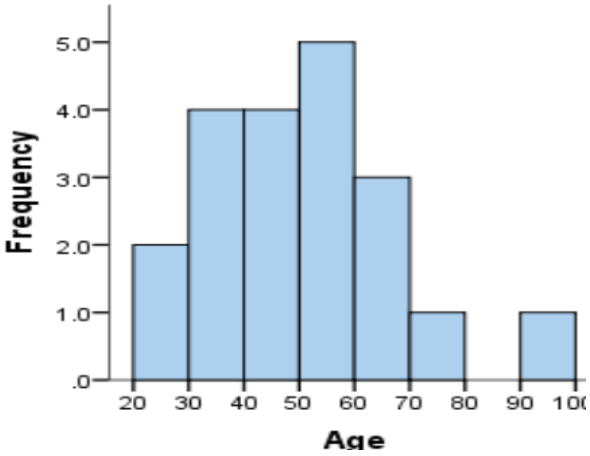
1. **What values appear?**
2. **How often do they appear?**
3. **How are they spread?**

Shape of Distribution



Histogram

Histogram is used to represent the distribution.



Histogram Basics :

X - axis = Data values

Y - axis = Frequency

Shape = Shape of the Distribution

Example

Mark of the student :

```
52, 49, 90, 50, 54
```

Distribution help us to understand ;

1. Are most students around 50 ?

   Ans : Since most of the student got near to the mark 50, they are around 50.

2. Is outlier present ?

   Ans : 90 is outlier here.

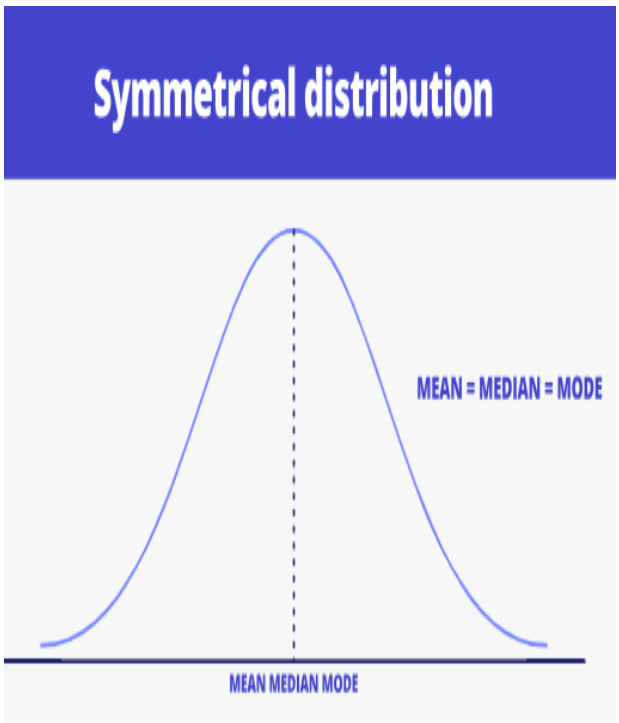
3. Is Data Balanced or not ?

   Ans : No this data follows right skew

Symmetric Distribution

Data is balanced around the center

I.e : Left side ≈ Right side



Properties :

- Mean ≈ Median ≈ Mode
- No strong Outlier
- Speard is even

Examples :

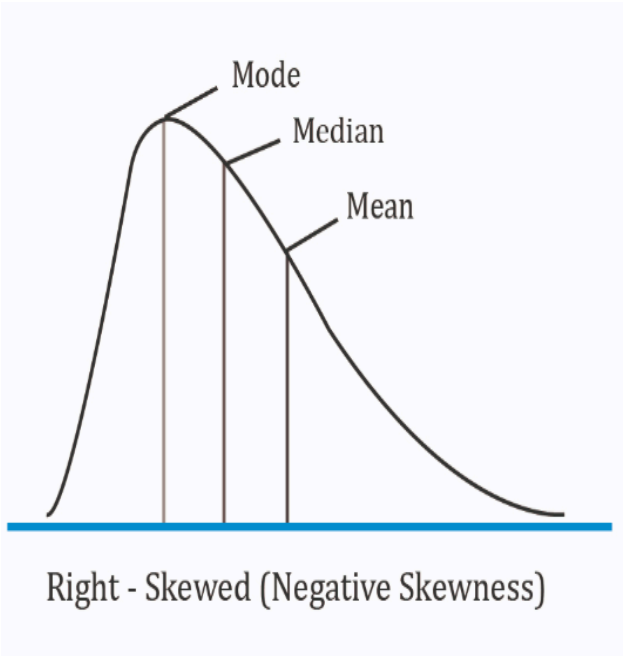
1. Height of adult
2. Measurement error
3. Sensor Value

For ML ;

- Mean is reliable
- Standard deviation makes sense.
- MSE works well.

Right Skewed Distribution (Positive Skew) :

Most of the values in data are small , few values are large



Properties :

- Mode > Median > Mean
- Mean will be in right
- Outliers will be in right.

Examples :

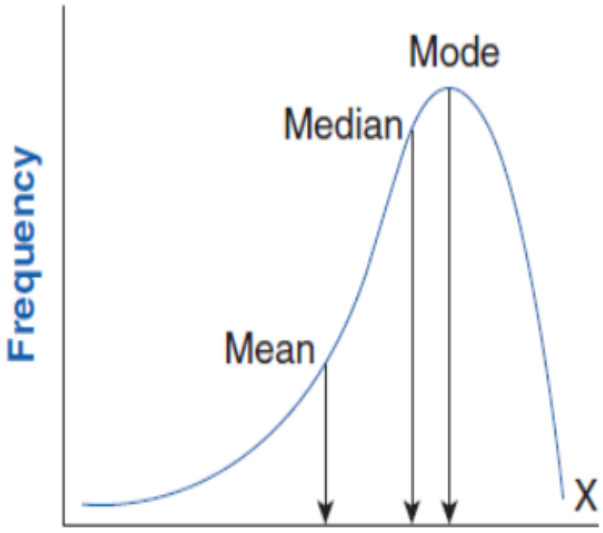
- Income
- House Price
- Waiting time

For ML :

- ❌ Mean is misleading
- ✅ Median is better
- ✅ MAE preferred over MSE

Left Skewed Distribution (Negative Distribution) :

Most of the values are large , some are small



Properties :

- Mean < Median < Mode
- Outliers are in Left side

Examples :

- Easy exam scores
- High passing exam

For ML :

- ❌ **Mean may lie**
- ✅ **Median is safer**

Probability

Probability

Probability measures how likely an event is to happen.

Values range from:

0 → impossible
1 → certain

Example:

Probability of rain = 0.8
→ High chance of rain

Probability Scale

Probability is always between 0 and 1.

a) 0 → Impossible
b) 0.5 → may or may not happen
c) 1 → definitely will happen

Examples:

1. Probability of sun rising tomorrow → 1
2. Probability of rain today → 0.6
3. Probability of humans flying naturally → 0

Mental lock 🔒

High probability ≠ guarantee
Low probability ≠ impossible

Sample Space

Sample space is all possible outcomes of an experiment.

ML intuition :

Sample Space -> All possible Class

Example :

Spam / not spam

Disease / not Disease

Example :

Coin Toss

All possible Outcomes ;

Sample Space = {HEAD , TAIL}

Dice Roll

All possible Outcomes ;

Sample Space = {1, 2, 3, 4, 5, 6}

Event

An Event is ;

"A specific outcome or a group of outcomes we care about"

ML intuition :

- Event = ' Predicted Class '

Example :

- Event = 'Spam'
- Event = ' Not Disease'

Examples :

1.Coin Toss

- Event = 'Head'
- Event = 'Tail'

Relationship between Space and Event :

- Space -> All Possible Outcome
- Event -> Particular Outcome

Note : Types of Event —-----> **Independent** and **Dependent** .

Independent Event :

The two Events are independent if and only if one event doesn't affect the others

In other words :

If one event has occurred then it should not affect the chance of another event

Examples :

Toss a Coin :

- Toss 1 —> 'Head'
- Toss 2 —> 'Tail' or 'Head'

Roll a Dice :

- Roll 1 —> '6'
- Roll 2 —> ' Anything from 1 to 6 , 6 may also come again '

For ML :

- Independent features mean one feature gives **no information** about another
- Naive Bayes **assumes features are independent**

Dependent Event :

If two events are dependent then

One event DOES affect the other event.

Example

1. Taking a Card from a Deck

- First card → Ace
- Second card → chances change

2. Medical Testing

- First card → Ace
- Second card → chances change

For ML :

In real data:

Features are often dependent

Example:

High BP and High Cholesterol are related

But some models simplify by assuming independence.

Conditional Probability :

Conditional probability means finding the chance of something AFTER you already know some information.

In simple Words ;

Conditional probability is the likelihood of an event occurring after considering additional known information.

Let consider same Events and two situations

Event : ‘Student passes exam ‘

Situation 1 : Before knowing anything

For the question :

‘Will student pass the exam ‘

The reply will be :

‘May be , chance of 50% ‘

Situation 2: After knowing something

The known Condition :

‘The student studied 8hrs for the exam’

Now for this question :

‘Will student pass the exam ‘

The reply will be :

‘May be , chance of 90%

Here, the condition changed the probability , This change is known as conditional probability .

Another Examples :

Now the sky is so clear and sunny

So the probability of rain is **‘LOW**’

If the sky is dark and cloudy

Then the probability of rain is **‘High’**

Term ‘given’ in Conditional Probability :

Sometimes the statement will be :

Probability of A is given by B

It means that :

What is the chance of ‘A’ after I already know ‘B’ is true?

Example : ‘Probability of disease **given** test is positive’

It doesn’t know :

- Two probabilities
- Two events happening together

It means:

- **Second info changes the first probability**

Why conditional probability is NOT optional :

Because in real life :

- We never decide with zero information
- Decisions are made after seeing evidence

Conditional probability = **decision with evidence**

ML connection :

Example:

- Probability of spam given words in email
- Probability of disease given BP & cholesterol

ML = conditional probability machine.

Summary:

Example:

- After info → conditional probability
- Before info -> Normal Probability

Random Variables

Random Variables

A random variable is:

A rule that assigns numerical values to the outcomes of an uncertain process.

"Random" refers to the outcome, not the variable

The variable itself follows a defined rule

Examples

Rolling a Dice :

Random variable: X

X = number obtained when rolling a dice

Possible values of X:

$X \in \{1, 2, 3, 4, 5, 6\}$

Probabilities:

P(X) = { P(X=1), P(X=2), ..., P(X=6) }

Coin toss :

Random variable: Y

Y = 1 → Head

Y = 0 → Tail

Possible values:

$Y \in \{0, 1\}$

Probabilities:

P(Y) = { P(Y=1), P(Y=0) }

Exam result

Random variable: Z

Z = 1 → Pass
Z = 0 → Fail

Possible values:

$Z \in \{0, 1\}$

Probabilities:

P(Z=1), P(Z=0)

Number of emails received today

Random variable: E

E = number of emails received in one day

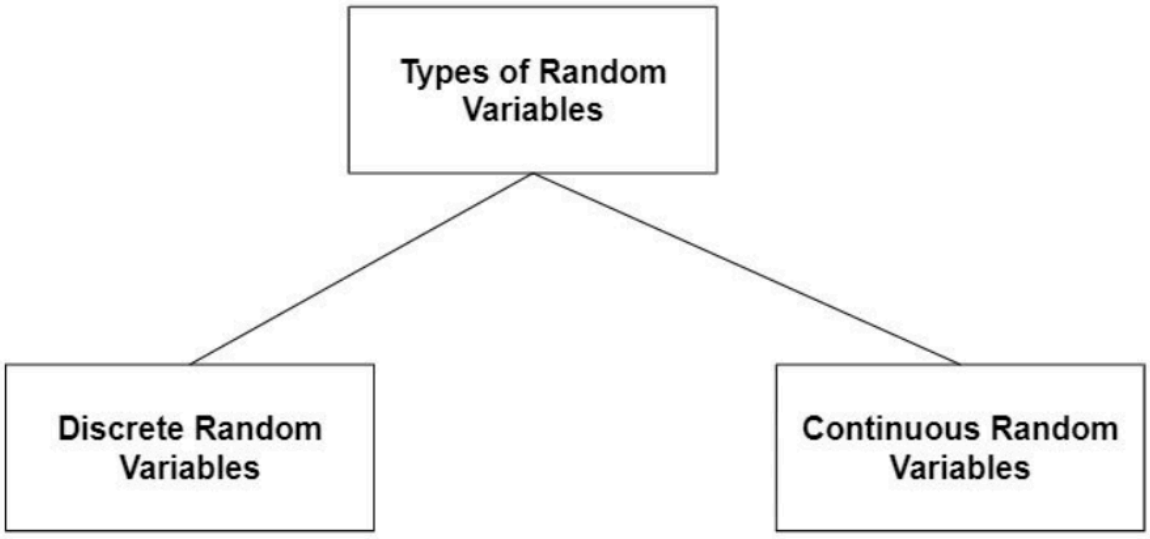
Possible values:

$E \in \{0, 1, 2, 3, 4, \ldots\}$

Probabilities:

P(E=0), P(E=1), P(E=2), ...

Types of Random Variables



Discrete Random Variables

A discrete random variable is a random variable that can take a finite or countably infinite set of values.

Or

Countable values

Characteristic of discrete random variable:

- Comes from counting
- Usually integers (0, 1, 2, 3, ...)
- Has gaps between values

Connection with Probability :

- Each exact value can have a probability
- That probability is given by PMF

ML connection :

Discrete random variables appear in ML as:

- Classification labels
- Binary outputs
- Count targets

Examples:

- Spam / Not spam
- Pass / Fail
- Number of clicks

Best Example :

Number of heads in coin tosses

Why this is the best:

- Possible values:
  `0, 1, 2, 3, ...`
- Values are **countable**
- No in-between values (❌ 1.5 heads)
- Comes from **counting**

Continuous Random Variable

Definition :

"A continuous random variable is a random variable that can take any value within a given range."

Characteristics :

- Values are **not countable**
- There are **no gaps**
- Between any two values, another value exists

Nature of continuous random variables :

- Comes from **measurement**
- Can take **decimal values**
- Can be infinitely precise

ML connection :

- Regression targets
- Sensor readings
- Prices, scores, durations

Discrete Random Variable

Values	Countable (0,1,2,3...)
Source	Counting
Decimals	No Decimals (0, 1, 2)
ML Task:	Classification

Continuous Random Variable

Values	Uncountable (Any Value)
Source	5.3, 7,28, 8.95...
Decimals	
ML Task:	Includes Decimals (5.7, 8.95)

Discrete vs Continuous

Countable vs Uncountable

PMF vs PDF

Probability Mass function (PMF):

Used for ***Discrete Random Variables***

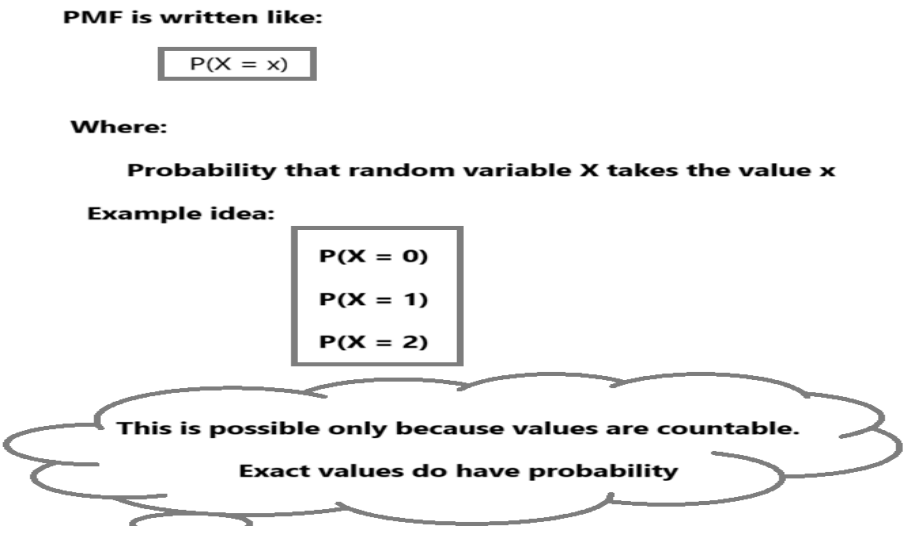
Meaning:

PMF gives the **probability of an exact value.**

Characteristic of PMF :

- Works only with **countable values**
- Probability can be assigned to **each exact value**
- Probabilities add up to **1**

Important Intuition :



Probability Density Function:

Used for Continuous Random Variables .

Meaning :

PDF describes how dense the values are around a point.

Note : PDF doesn't provide exact values of the probability

Characteristic of PDF :

- Works with **uncountable values**
- Probability of an **exact value = 0**
- Probability is found over a **range (interval)**

PDF intuition :

For continuous variables:

- Exact value has no width
- So it has zero probability

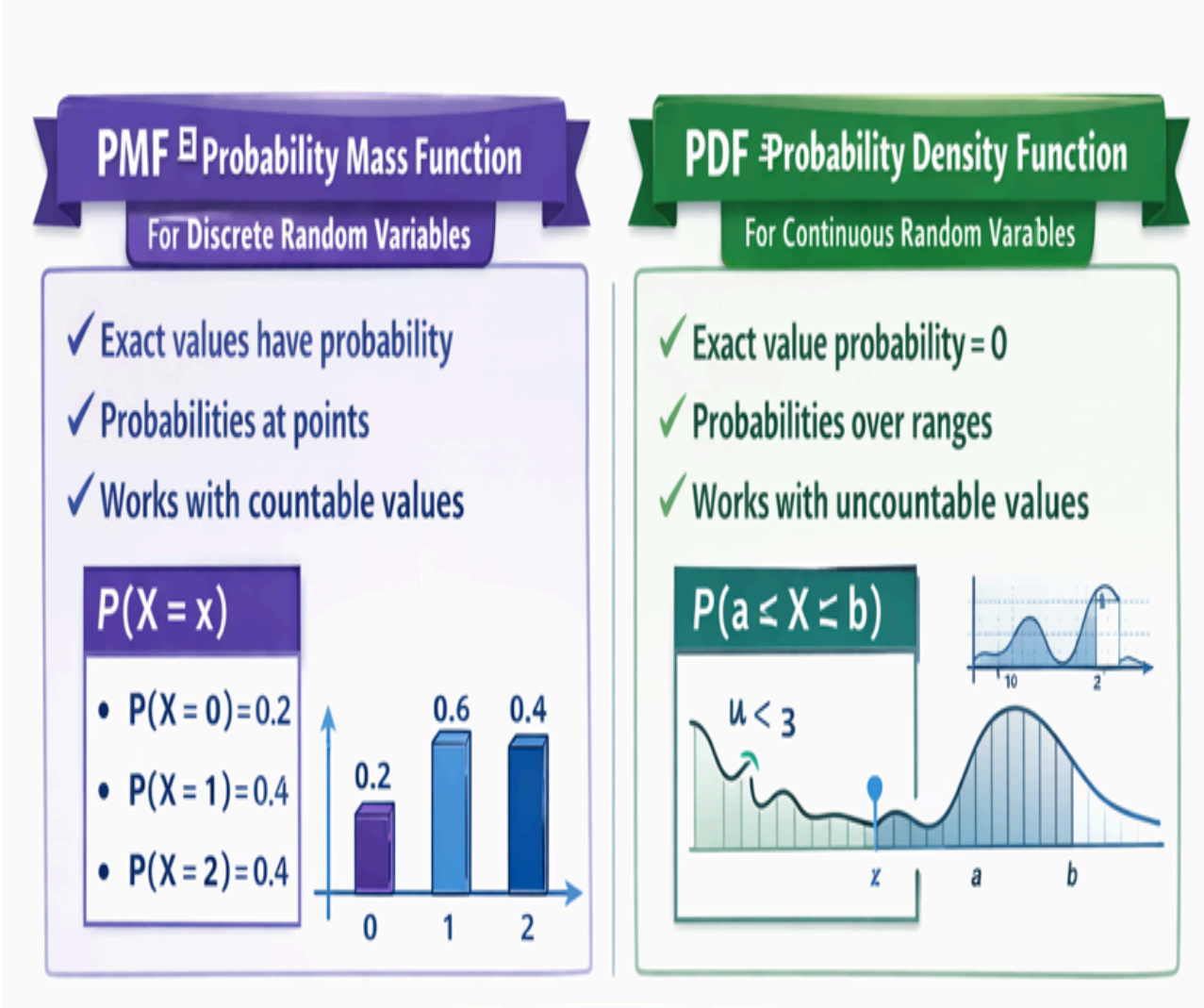
Instead:

Probability comes from an interval, not a point

Represented as :

- ❌ P(X = 45)
- ✅ P(40 ≤ X ≤ 45)

  Probability = area under the curve in that interval.


PMF Probability Mass Function
For Discrete Random Variables
✓ Exact values have probability
✓ Probabilities at points
✓ Works with countable values
P(X = x)
• P(X = 0) = 0.2
• P(X = 1) = 0.4
• P(X = 2) = 0.4
0.2
0.6
0.4
0
1
2
PDF Probability Density Function
For Continuous Random Varables
✓ Exact value probability = 0
✓ Probabilities over ranges
✓ Works with uncountable values
P(a ≤ X ≤ b)
u < 3
x
a
b
10
2

Important Distribution

Bernoulli Distribution

A random variable is:

A rule that assigns numerical values to the outcomes of an uncertain process.

"Random" refers to the outcome, not the variable

The variable itself follows a defined rule

Examples

Rolling a Dice :

Random variable: X

X = number obtained when rolling a dice

Possible values of X:

$X \in \{1, 2, 3, 4, 5, 6\}$

Probabilities:

P(X) = { P(X=1), P(X=2), ..., P(X=6) }

Coin toss :

Random variable: Y

Y = 1 → Head

Y = 0 → Tail

Possible values:

$Y \in \{0, 1\}$

Probabilities:

P(Y) = { P(Y=1), P(Y=0) }

Exam result

Random variable: Z

Z = 1 → Pass
Z = 0 → Fail

Possible values:

Z ∈ {0, 1}

Probabilities:

P(Z=1), P(Z=0)

Number of emails received today

Random variable: E

E = number of emails received in one day

Possible values:

E ∈ {0, 1, 2, 3, 4, ...}

Probabilities:

P(E=0), P(E=1), P(E=2), ...

Conditions :

Bernoulli works ONLY when:

✅ Single trial

✅ Only two outcomes

✅ Fixed probability

✅ Outcome coded as 0 or 1

If any condition breaks → Not Bernoulli.

Binomial Distribution

Binomial distribution models the number of successes in multiple independent Bernoulli trials.

Simple Understand:

- **Bernoulli →** One yes/no experiment
- **Binomial →** Many yes/no experiments → Count successes

Examples :

1. 10 Coin Tosses

Each toss:

Head → Success
Tail → Failure

Question:
👉 How many heads will appear in 10 tosses?

This is Binomial, not Bernoulli.

Conditions of Binomial Distribution :

✅ Fixed Number of Trials (n is fixed)

✅ Each Trial Has Only Two Outcomes

✅ Probability of Success is Constant (p is same)

✅ Trials Must Be Independent

✅ Counting Number of Successes

❌ Example That Looks Binomial But Is NOT

Drawing 5 Cards Without Replacement

For this Example the observation are

Trials fixed ✔

Binary maybe ✔

Probability same ❌

Independent ❌

Reason Not to be Binomial :

1.The probability are not same for all outcomes, if red card is drawn from first outcome, the probability will be 26/52 , then again a red card is drawn then the probability will be 25/52.

2. Each trial are Dependent

👉 So it is Not Binomial.

Difference Between Bernoulli and Binomial Distribution

	Bernoulli Distribution	Binomial Distribution
Number of Trials	1 Trial Only	Multiple Trials (n Fixed)
Outcomes	Two Outcomes: 0 or 1 (Success / Failure)	Count of Surceeses in *n* Trials (0 to n)
Probability	Probability: *p* (Fixed)	Probability: *p* (Same for Each Trial)
Example	Single Coin Toss (Head or Tail)	Number of Heads in 10 Coin Tosses
Question Asked	Did it Happen? Yes or No?	How Many Times Did it Happen?

Normal Distribution

Definition :

Normal distribution is a continuous probability distribution where data clusters symmetrically around a mean value, forming a bell-shaped curve due to many small random effects combining.

Example: Class Height

Suppose class average height = 170 cm.

Reality:

Many students → 165–175 cm
Some → 155 or 185
Very few → 145 or 195

Characteristics :

- ✔ Symmetric
- ✔ Single center point (mean = median = mode)
- ✔ Most values near average.

Covariance

Covariance

It measures:

1. Do X and Y go above their averages together?
2. Do X and Y go below their averages together?

Formula

$$(X - \text{Mean of } X) \times (Y - \text{Mean of } Y)$$

Why does subtract mean?

Because we want to know:

👉 Above average or below average?

Why multiply?

Because sign tells direction that is positive or negative

Limitation 👀:

Example:

Covariance = 500 → Good? Bad? Strong? Weak?

I can't tell.

Why?

Because covariance depends on:

👉 Units

👉 Scale

👉 Magnitude of data

Another Example:

Height in cm vs height in meters → covariance changes.

Correlation

Correlation

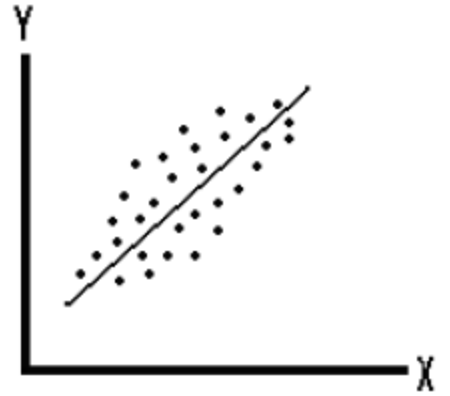
Definition

Correlation measures how strongly two variables are related AND in which direction.

Covariance → Direction only

Correlation → Direction + Strength

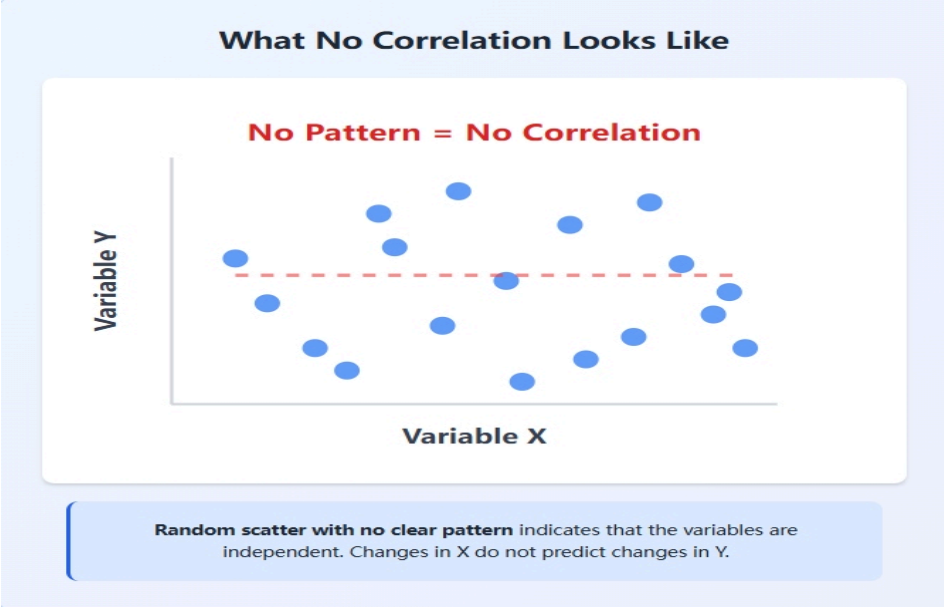
How the correlation looks in the graph of a data



Positive Correlation

Negative Correlation

If no correlation :



Range of the Correlation value :

Strength Meaning

Correlation Value	Strength
0.7 to 1	Strong
0.3 to 0.7	Moderate
0 to 0.3	Weak

Mini Analysis

Mini Statistical Analysis Report

Dataset : “Happiness Factors Dataset”

Dataset Overview :

The dataset contains **158 observations** and **5 numerical features**:

- Economy
- Family (Fam)
- Health
- Freedom
- Happiness Score (H_Score)

Description :

	Economy	Fam	Health	Freedom	H_Score
count	158.000000	158.000000	158.000000	158.000000	158.000000
mean	0.846137	0.991046	0.630259	0.428615	5.375734
std	0.403121	0.272369	0.247078	0.150693	1.145010
min	0.000000	0.000000	0.000000	0.000000	2.839000
25%	0.545808	0.856823	0.439185	0.328330	4.526000
50%	0.910245	1.029510	0.696705	0.435515	5.232500
75%	1.158448	1.214405	0.811013	0.549092	6.243750
max	1.690420	1.402230	1.025250	0.669730	7.587000

All columns are numerical and there are **no missing values**, making the dataset clean and suitable for statistical analysis and machine learning preprocessing.

Correlation Analysis of a data

```
Correlation :
           Economy       Fam    Health   Freedom   H_Score
Economy   1.000000  0.645299  0.816478  0.370300  0.780966
Fam       0.645299  1.000000  0.531104  0.441518  0.740605
Health    0.816478  0.531104  1.000000  0.360477  0.724200
Freedom   0.370300  0.441518  0.360477  1.000000  0.568211
H_Score   0.780966  0.740605  0.724200  0.568211  1.000000
```

Strength Meaning

Correlation Value	Strength
0.7 to 1	Strong
0.3 to 0.7	Moderate
0 to 0.3	Weak

High correlation observed between:

Economy ↔ Health ≈ 0.81

This suggests:

- Economically strong regions often have better healthcare systems
- Possible multicollinearity risk in ML models
- May require feature selection or dimensionality reduction later

3