STATISTICS

IMPORTANT CONCEPTS

STATISTICS

Module 1 : Statistics Basics

- Introduction to statistics
- Types of statistics
- Types of sampling
- Types of data
- Level of measurement
- Measure of central tendency
- Implementation of central tendency
- Measure of spread variance
- Measure of spread variance – mean deviation
- Measure of symmetricity & skewness
- Implementation of symmetricity & skewness
- Set
- Covarience & corelation
- Covarience & corelation implementaion

Lecture 1 : Introduction to Statistics

Definition of Statistics

Statistics is a **mathematical science** that includes methods of **collecting, organizing, and analyzing data** in such a way that **meaningful conclusions** can be drawn from the data.

- **Simple Explanation:**
  - Statistics helps us **understand large amounts of data** easily.
  - It converts **raw data into useful information**.
  - It supports **fact-based decision making** instead of guessing.

Applications of Statistics

Statistics is mainly divided into **two branches** based on its usage.

1. Descriptive Statistics

- Descriptive statistics is used to **describe and summarize data**.
- It explains **what has already happened**.
- It does not predict future outcomes.

- **Topics covered under Descriptive Statistics:**
  - **Central Tendency** → Mean, Median, Mode
  - **Dispersion** → Range, Variance, Standard Deviation
  - **Symmetricity (Skewness)** → Shows direction of data spread
- **Example:**
  - Finding the **average marks** of students.
  - Finding **highest and lowest scores** in a match.

2. Inferential Statistics

- Inferential statistics is used to **draw conclusions about a population**.
- It works on **sample data**.
- It helps in **prediction and decision making**.

- **Topics covered under Inferential Statistics:**
  - **Probability Distributions**
    - PMF (Probability Mass Function)
    - PDF (Probability Density Function)
    - CDF (Cumulative Distribution Function)
  - **Central Limit Theorem (CLT)**
  - **Statistical Tests** for hypothesis testing
- **Example:**
  - Predicting **future sales** of a product.
  - Estimating **election results** using survey data.

Why Statistics?

Statistics is important because:

- It helps to **identify patterns and trends** in data.
- It improves **accuracy in decision making**.
- It reduces errors and uncertainty.
- It helps organizations **plan for the future**.

🔹 In simple words, statistics helps us **learn patterns from data**.

Data

Definition of Data

Data refers to **facts or pieces of information** that can be:

- Collected, Stored, Measured, Analyzed and Used again when needed

---

Importance of Data

- Data helps companies **increase revenue**.
- It supports **business growth and improvement**.
- Decisions based on data are **more reliable**.

Steps Involving Data

---

1. Collecting Data

Data can be collected from:

- Online portals, Surveys and questionnaires, Internet clicks and user behavior, Experiments, Customer ratings and reviews

---

2. Organizing Data

- Collected data is stored properly for easy access.
- Data is organized using:
    - SQL databases, NoSQL databases

---

3. Analyzing Data

- Data analysis is done to **find useful insights**.
- Common tools used for analysis:
    - Pandas, NumPy, Matplotlib, Seaborn, Bokeh

Examples of Data

- Scores made by cricketers in **T20 World Cup**
- Demand of products on **e-commerce websites**
- Monthly sales data of a company
- Student exam marks

Uses of Statistics

Statistics is widely used in different fields:

- **Weather Forecasting** → Rainfall and temperature prediction
- **Sports Analysis** → Highest score, strike rate, average score
- **Election Campaigns** → Opinion polls and voting trends
- **FMCG / E-Commerce** → Sales analysis and customer behavior
- **Medical / Genetics / Pharmaceutical** → Research and testing

---

📊 **Summary Table (Quick Revision)**

Topic	Key Points (Easy Revision)
Statistics	Science of collecting, organizing, analyzing data to draw conclusions
Descriptive Statistics	Summarizes data using mean, median, mode, dispersion, skewness
Inferential Statistics	Uses sample data to predict and make decisions
Probability Distribution	PMF, PDF, CDF used in inferential statistics
Why Statistics	Finds patterns, supports decisions, improves accuracy
Data	Facts or information that can be stored and analyzed
Data Collection	Surveys, portals, experiments, internet clicks, reviews
Data Organization	Stored in SQL and NoSQL databases
Data Analysis Tools	Pandas, NumPy, Matplotlib, Seaborn, Bokeh
Uses of Statistics	Weather, sports, elections, business, medical fields

---

Lecture 2 : Types of Statistics

Statistics is broadly classified into **two main types** based on how data is used and interpreted.

1. **Descriptive Statistics**
2. **Inferential Statistics**

---

1. Descriptive Statistics

- Descriptive statistics deals with **organizing, summarizing, and presenting the complete data**.
- It focuses on the **entire population**, not a sample.
- It explains **what the data shows**, without making predictions.
- It is mainly used when **exact values are required** for decision-making.

Key Features of Descriptive Statistics

- Works on **complete data / population**
- No guessing or prediction is involved
- Helps in **understanding data clearly**
- Mostly used for **business reports and analysis**

Examples of Descriptive Statistics

- Calculating the **strike rate of cricketers**
- Finding **average height or weight of students** in a class
- Measuring **average flight delay** of ABC Airlines
- Calculating **monthly average sales** of a company

Techniques Used in Descriptive Statistics

1. Measures of Central Tendency

- Mean
- Median
- Mode

  👉 Used to find the **central or average value** of data

2. Measures of Dispersion

- Variance
- Standard Deviation

  👉 Used to find **how much data varies or spreads**

3. Measures of Symmetricity

- Skewness
- Kurtosis

  👉 Used to understand the **shape of data distribution**

---

2. Inferential Statistics

- Inferential statistics uses **sample data** to draw conclusions about the **entire population**.
- It helps answer the question:
  "Can we say something about the whole population using a sample?"
- Taking a **proper sample is very important**.
- Conclusions about the population are made **based on sample results**.

Key Features of Inferential Statistics

- Works on **sample data**
- Used when population size is **very large**
- Saves **time, cost, and resources**
- Helps in **prediction and decision-making**

Why Sampling is Important

- Sometimes it is **not possible to measure the whole population**
- Population may be:
  - Very large
  - Time-consuming to measure
  - Costly to analyze
- In such cases, a **representative sample** is used

Examples of Inferential Statistics

- Estimating **average height or weight of India's population**
- Estimating the **number of trees in a national park**
- Predicting **election results using survey data**
- Estimating **future demand** of a product

Difference between Descriptive and Inferential Statistics

Basis	**Descriptive Statistics**	**Inferential Statistics**
Data Used	Complete data / population	Sample data
Purpose	Summarize and describe data	Draw conclusions and predict
Scope	Explains what data shows	Explains what data may imply
Accuracy	Exact values	Approximate values
Prediction	No prediction	Used for prediction
Cost & Time	More (if population is large)	Less (uses sample)
Example	Average marks of a class	Average marks of all students in India

---

📊 **Summary Table (Quick Revision)**

Topic	**Key Points for Quick Revision**
Types of Statistics	Descriptive and Inferential
Descriptive Statistics	Summarizes complete data or population
Used For (Descriptive)	Exact analysis and reporting
Techniques (Descriptive)	Mean, Median, Mode, Variance, SD, Skewness, Kurtosis
Inferential Statistics	Uses sample data to conclude about population
Sampling	Essential when population is large
Used For (Inferential)	Prediction and decision-making
Examples	Population height, trees in forest, election surveys

---

Lecture 3 : Types of Sampling

Sampling

Sampling is the process of **selecting a small group (sample)** from a **large population** so that conclusions about the **entire population** can be made easily.

- Sampling saves **time, cost, and effort**.
- It is mainly used in **inferential statistics**.

Types of Sampling

There are **four main types of sampling**:

1. Simple Random Sampling
2. Stratified Sampling
3. Cluster Sampling
4. Systematic Sampling

---

1. Simple Random Sampling (SRS)

Definition

Simple Random Sampling is a method in which **every member of the population (N)** has an **equal chance of being selected** in the sample.

Explanation

- Selection is done **completely at random**.
- No preference is given to any group or individual.
- It is similar to a **lottery method**.

Key Point

- Each individual has **equal probability** of selection.

Example

- Selecting **50 students randomly** from a college of 1,000 students.
- Selecting voters randomly for an opinion poll.

Disadvantage of Simple Random Sampling

- Certain groups may **not be represented properly**.
- Equal representation of all sub-groups is **not guaranteed**.

◆ **Example (India Population):**

- India has around **140 crore people**.
- While taking a sample using SRS, people from **smaller or less populated states** may not be selected.
- This causes **unequal representation**.

---

2. Stratified Sampling

Definition

Stratified Sampling is a method in which the population is divided into **different sub-groups (called strata)**, and then a **simple random sample is selected from each stratum**.

Explanation

- Population is divided based on **distinct characteristics**.
- Each group (stratum) is represented in the sample.
- Sampling is done **inside each group**.

Key Points

- Groups are **mutually exclusive and collectively exhaustive**.
- Ensures **fair representation** of all categories.

Example

- Dividing students into **male and female**, then randomly selecting students from each group.
- Dividing population by **age groups**, then sampling from each age group.

---

3. Cluster Sampling

Definition

Cluster Sampling is a method in which the population is divided into **clusters (groups)**, then **some clusters are randomly selected**, and **all individuals in the selected clusters** are included in the sample.

Explanation

- Population is divided into clusters (often based on location).
- Only **some clusters are chosen randomly**.
- Every member inside the selected cluster becomes part of the sample.

Key Points

- Sampling is done at **group level**, not individual level.
- Often used when population is **geographically spread**.

Example

- Dividing a city into **wards**, selecting a few wards randomly, and surveying **all people in those wards**.
- Selecting some schools randomly and surveying **all students** in those schools.

---

4. Systematic Sampling

Definition

Systematic Sampling is a method in which **every nth element** from the population is selected to form the sample.

Explanation

- First, a starting point is chosen.
- Then, every **nth member** is selected.
- Selection follows a **fixed pattern**.

Key Point

- Easy and quick to apply.
- Works well when population list is available.

Example

- Selecting **every 10th customer** entering a mall.
- Selecting **every 5th roll number** from a student list.

📊 **Summary Table (Quick Revision)**

Type of Sampling	Main Idea	Key Feature	Example
Simple Random Sampling	Random selection	Equal chance to all	Lottery method
Stratified Sampling	Divide into strata	Sample from each group	Male/Female groups
Cluster Sampling	Divide into clusters	Select some clusters fully	City wards
Systematic Sampling	Fixed interval selection	Every nth element	Every 10th person

---

Lecture 4 : Types of Data

Data

Data means **information or facts** that are collected for **analysis and decision making**.
Understanding data types is very important before doing any analysis.

Types of Data

Data is mainly divided into **two types**:

1. **Quantitative Data (Numerical Data)**
2. **Qualitative Data (Categorical Data)**

---

1. Quantitative Data (Numerical Data)

Definition

Quantitative data is data that is **numerical in nature** and can be **measured or counted**.

Explanation

- It is represented using **numbers**
- Mathematical operations like **addition, subtraction, average** can be performed
- Mostly used in **calculations and statistical analysis**

Examples

- Height of a person (170 cm)
- Weight of students (60 kg)
- Marks obtained in an exam (85)
- Number of products sold (200)

---

2. Qualitative Data (Categorical Data)

Definition

Qualitative data is data that **describes qualities or categories** and is **not numerical**.

Explanation

- It represents **labels, names, or categories**
- Cannot be measured using numbers directly
- Mostly used to **describe characteristics**

Examples

- Gender (Male, Female)
- Blood group (A, B, AB, O)
- Product category (Electronics, Clothing)
- Feedback (Good, Average, Bad)

Difference between Quantitative and Qualitative Data

Basis	Quantitative Data	Qualitative Data
Nature	Numerical	Non-numerical
Measurement	Can be measured	Cannot be measured
Operations	Mathematical operations possible	Mathematical operations not possible
Example	Height, weight, marks	Gender, color, feedback

Types of Quantitative Data

Quantitative data is further divided into **two types**:

1. **Discrete Quantitative Data**
2. **Continuous Quantitative Data**

---

1. Discrete Quantitative Data

Definition

Discrete quantitative data is data that **can be counted** and has **fixed values**.

Explanation

- Values are usually **whole numbers**
- No intermediate values exist between two numbers
- Comes from **counting**

Examples

- Number of students in a class (40)
- Number of cars in a parking lot
- Number of mobile phones sold in a day

---

2. Continuous Quantitative Data

Definition

Continuous quantitative data is data that **can be measured** and can take **any value within a range**.

Explanation

- Values can include **decimals**
- Infinite values are possible between two points
- Comes from **measurement**

Examples

- Height of a person (165.5 cm)
- Weight of fruits (2.75 kg)
- Temperature (36.6°C)

Difference between Discrete and Continuous Quantitative Data

Basis	Discrete Data	Continuous Data
Nature	Countable	Measurable
Values	Whole numbers	Decimal values possible
Range	Fixed values	Infinite values in a range
Example	Number of students	Height, weight

Types of Qualitative Data

Qualitative data is further divided into **two types**:

1. **Nominal Qualitative Data**
2. **Ordinal Qualitative Data**

---

1. Nominal Qualitative Data

Definition

Nominal data is qualitative data where **categories have no specific order or ranking**.

Explanation

- Categories are just **names or labels**
- No comparison like greater or smaller is possible

Examples

- Gender (Male, Female)
- Blood group (A, B, AB, O)
- Nationality (Indian, American)

---

2. Ordinal Qualitative Data

Definition

Ordinal data is qualitative data where **categories have a meaningful order or ranking**.

Explanation

- Order matters
- Exact difference between categories is **not known**

Examples

- Feedback (Good, Better, Best)
- Education level (School, College, University)
- Customer rating (Low, Medium, High)

Difference between Nominal and Ordinal Qualitative Data

Basis	Nominal Data	Ordinal Data
Order	No order	Order exists
Ranking	Not possible	Possible
Comparison	Cannot compare	Can compare
Example	Gender, blood group	Good, Better, Best

Use Case of Data Types (Very Important)

- To **analyze data** or **build machine learning models**, knowing data types is very important.
- Computers understand **numbers**, not categories.
- So, **categorical (qualitative) data must be converted into numbers**.

Example

Category	Converted Value
Good	1
Better	2
Best	3

This process helps in:

- Data analysis
- Statistical modeling
- Machine learning algorithms

📊 **Summary Table (Quick Revision)**

Topic	**Key Points (Easy Revision)**
Types of Data	Quantitative and Qualitative
Quantitative Data	Numerical, measurable or countable
Qualitative Data	Categorical, descriptive
Discrete Data	Countable, whole numbers
Continuous Data	Measurable, decimal values
Nominal Data	No order, only labels
Ordinal Data	Ordered categories
Use Case	Convert categories to numbers for analysis

Lecture 5 : Level of Measurement

Level of Measurement

Level of measurement explains **how data is measured, classified, and compared**.

It tells us **what kind of mathematical operations and analysis** can be done on data.

Data is measured using **four types of scales**.

Types of Measurement Scales

For Quantitative Measures (Numeric Data)

1. **Interval Scale Data**
2. **Ratio Scale Data**

For Qualitative Measures (Non-Numeric Data)

3. **Nominal Scale Data**
4. **Ordinal Scale Data**

---

1. Interval Scale Data

Definition

Interval scale data is numeric data where:

- The **difference between values is meaningful**
- But there is **no true zero starting point**

Explanation

- Data values are equally spaced
- Zero does not mean "nothing"
- Ratios are **not meaningful**

Examples

- Temperature in Celsius or Fahrenheit
  (0°C does not mean no temperature)
- Calendar years (2020, 2021, 2022)
- Time on a clock (10 AM, 11 AM)

Charts Used

- Histogram, Scatter Plot, Line Chart

---

2. Ratio Scale Data

Definition

Ratio scale data is numeric data where:

- Difference between values is meaningful
- It has a **true zero starting point**

Explanation

- Zero means **absence of quantity**
- All mathematical operations are possible
- Ratios like **twice, half** are meaningful

Examples

- Height (0 cm means no height)
- Weight (0 kg means no weight)
- Age (0 years)
- Income (₹0 means no income)

Charts Used

- Histogram, Scatter Plot, Line graphs and bar charts

---

3. Nominal Scale Data

Definition

Nominal scale data is qualitative data where:

- Data is **classified into categories**
- There is **no order or ranking**

Explanation

- Categories are just **names or labels**
- No comparison like greater or smaller

Examples

- Gender (Male, Female), Blood group (A, B, AB, O), Country name, Product type

---

Charts Used

- Pie Chart, Bar Plot

---

4. Ordinal Scale Data

Definition

Ordinal scale data is qualitative data where:

- Categories have a **meaningful order**
- Exact difference between categories is **not measurable**

Explanation

- Ranking is possible
- Difference between ranks is not fixed

Examples

- Feedback (Poor, Average, Good, Excellent)
- Education level (School, College, University)
- Customer satisfaction (Low, Medium, High)

Charts Used

- Pie Chart, Bar Plot

Difference between Levels of Measurement

Basis	Nominal	Ordinal	Interval	Ratio
Data Type	Qualitative	Qualitative	Quantitative	Quantitative
Labelled Categories	✓ Yes	✓ Yes	✗ No	✗ No
Meaningful Order	✗ No	✓ Yes	✓ Yes	✓ Yes
Measurable Difference	✗ No	✗ No	✓ Yes	✓ Yes
True Zero Point	✗ No	✗ No	✗ No	✓ Yes
Mathematical Operations	✗ No	Limited	+, −	+, −, ×, ÷
Ratio Meaningful	✗ No	✗ No	✗ No	✓ Yes
Example	Gender	Feedback	Temperature	Weight

📊 **Summary Table (Quick Revision)**

Scale Type	Nature	Key Feature	Example
Nominal	Qualitative	Categories only	Gender
Ordinal	Qualitative	Ordered categories	Feedback
Interval	Quantitative	No true zero	Temperature
Ratio	Quantitative	True zero exists	Weight

Lecture 6 : Measure of Central Tendency

Descriptive Statistics

Descriptive statistics is used for **summarizing data** without changing it at a specific time or instance.

Techniques of Descriptive Statistics

- **Measure of Central Tendency, Measure of Dispersion, Measure of Symmetricity**

Measure of Central Tendency

Definition

Measure of central tendency is a statistical method used to **find the central or typical value** of a dataset.

Explanation

- It shows **where the center of data lies**
- It helps in **quick understanding of data**
- One value represents the **whole dataset**

Example (Simple Idea)

If marks of students are:

60, 65, 70, 75, 80

→ The central value is around **70**

Central Tendency Represents the Center Point of a Dataset

The three main measures of central tendency are:

1. Mean, 2. Median, 3. Mode

---

1. Mean

Definition

Mean is the **average value** of a dataset.

It is calculated by **adding all values and dividing by the total number of values**.

Formula

Mean = (Sum of all values) / (Total number of values)

Example with Solution

Data: 10, 20, 30, 40, 50

Step 1: Add all values

= 10 + 20 + 30 + 40 + 50 = **150**

Step 2: Count total values = 5

Step 3: Divide

Mean = 150 / 5 = **30**

---

2. Median

Median is the **middle value** of a dataset when the data is arranged in **ascending or descending order**.

Explanation

- Median divides the data into **two equal parts**
- It is **not affected by extreme values (outliers)**

Example with Solution (Odd number of values)

Data: 5, 10, 15, 20, 25

Middle value = **15**

So, Median = **15**

Example with Solution (Even number of values)

Data: 10, 20, 30, 40

Middle values = 20 and 30

Median = (20 + 30) / 2 = **25**

---

3. Mode

Definition

Mode is the value that **appears most frequently** in a dataset.

Explanation

- Dataset can have **one mode, more than one mode, or no mode**
- Mode is the **only central tendency measure used for categorical data**

Example with Solution

Data: 2, 4, 4, 6, 8

- 4 appears **most frequently**

So, Mode = **4**

When to Choose Mean or Median

- If data has **outliers (very large or very small values)**:
  - **Mean is affected**
  - **Median is not affected**

Example

Data: 10, 12, 14, 16, 100

- Mean = (10+12+14+16+100)/5 = **30.4**
- Median = **14**

👉 Median gives a **better central value** here.

Data Types Used in Central Tendency

- **Mean** → Numerical data
- **Median** → Numerical data
- **Mode** → Numerical and Categorical data

Use Cases of Central Tendency

1. Handling Missing Values (Imputation)

- If data has **no outliers** → Replace missing value with **Mean**
- If data has **outliers** → Replace missing value with **Median**

Example

Data: 10, 12, ?, 14, 16

- No outliers → Missing value can be replaced with **Mean**

Data: 10, 12, ?, 14, 100

- Outlier present → Missing value should be replaced with **Median**

📊 **Summary Table (Quick Revision)**

Measure	**Meaning**	**Used For**	**Affected by Outliers**	**Example**
Mean	Average value	Numerical data	Yes	Marks average
Median	Middle value	Numerical data	No	Income data
Mode	Most frequent value	Numerical & categorical	No	Most sold product
Central Tendency	Center of data	Data summarization	—	Single representative value
Missing Value Handling	Data cleaning	Mean / Median	Depends on outliers	Replace null values

---

Lecture 7 : Implementation of Central Tendency

Measure of Central Tendency (Practical Implementation)

In this lecture, we **implement mean, median, and mode using Python**.

These measures help us **summarize numerical data** using a single value.

We will use:

- Basic Python, NumPy, SciPy, Statistics module

---

1. Mean (Average)

Definition

Mean is the **average value** of a dataset.

It is calculated by **adding all values and dividing by total count**.

Example 1: Mean using basic Python

```
age = [28, 56, 35, 28, 76, 89]

(28 + 56 + 35 + 28 + 76 + 89) / 6
```

Output:

52.0

Example 2: Mean using NumPy

```
import numpy as np

age = [28, 56, 35, 28, 76, 89]
np.mean(age)
```

Output:

52.0

Example 3: Mean of height data

```
height = [170, 180, 100, 120, 122]
np.mean(height)
```

Output:

138.4

---

2. Median

Definition

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

It is **not affected by outliers**.

Example: Median using NumPy

```
nos = [4, 2, 3, 7, 8]
np.median(nos)
```

Output:

4.0

---

3. Mode

Definition

Mode is the value that **appears most frequently** in the dataset.

It can be used for **numerical and categorical data**.

Example Dataset

nos = [4, 2, 2, 3, 7, 8]

Mode using SciPy

```
from scipy import stats

stats.mode(nos)
```

Output:

ModeResult(mode=array([2]), count=array([2]))

Mode using statistics module

```
import statistics

statistics.mode(nos)
```

Output:

2

Mean is Affected by Outliers

An **outlier** is a value that is **very large or very small** compared to other values.

Example with Outlier

```
height = [170, 1800, 100, 120, 122]
np.mean(height)
```

Output:

462.4

👉 Mean becomes **misleading** due to outlier (1800).

Example without Outlier

```
height = [170, 180, 100, 120, 122]
np.mean(height)
```

Output:

138.4

Median with Outlier

```
height = [170, 1800, 100, 120, 122]
np.median(height)
```

Output:

122.0

👉 Median gives a **better central value** when outliers exist.

Another Simple Median Example

height = [160, 120, 125, 34]

np.median(height)

Output:

122.5

Key Observations

- **Mean** changes when outliers are present
- **Median** remains stable even with outliers
- **Mode** shows most frequent value

📊 **Summary Table (Quick Revision)**

Measure	Python Function	Module Used	Affected by Outliers	Used For
Mean	np.mean()	NumPy	Yes	Average calculation
Median	np.median()	NumPy	No	Central value
Mode	stats.mode()	SciPy	No	Most frequent value
Mode	statistics.mode()	statistics	No	Categorical & numeric
Outlier Handling	Prefer median	—	—	Robust analysis

Lecture 8 : Measure of Spread (Variance / Dispersion)

Measure of Dispersion

Definition

Measure of dispersion is a statistical method used to **show how data values are spread or scattered** around the central value.

Explanation

- Central tendency tells the **center of data**
- Dispersion tells **how far data points are from each other**
- It helps to understand **data variability**

Example (Simple Idea)

Two classes have the same average marks,
but one class has **very different marks**.

👉 Measure of dispersion helps identify this difference.

Measure of Dispersion Tells How the Data is Spread

The main measures of dispersion are:

- Range, Percentage, Percentile, Quartiles (Box Plot), Variance, Standard Deviation

---

1. Range

Definition

Range is the **difference between the maximum and minimum value** in a dataset.

Formula

Range = Maximum value − Minimum value

Example with Solution

Data: 10, 20, 30, 40, 50

- Maximum value = 50
- Minimum value = 10

Range = 50 − 10 = **40**

Important Points about Range

- Very **easy to calculate**
- Gives a **basic idea of spread**
- **Outliers strongly affect the range**

Outlier Effect on Range

Data: 10, 20, 30, 40, 500

- Range = 500 − 10 = **490**

👉 Range becomes very large due to **outlier (500)**.

---

2. Percentage

Definition

Percentage shows a value **out of 100**.

It represents a part of the whole.

Formula

Percentage = (Part / Total) × 100

Example with Solution

A student scored **45 marks out of 60**.

Percentage = (45 / 60) × 100 = **75%**

Use of Percentage

- Exam results, Success rate, Accuracy measurement

---

3. Percentile

Definition

A percentile is a value **below which a certain percentage of observations lie**.

Explanation

- It helps compare an individual value with the group
- Widely used in **exams and rankings**

Example with Solution

If a student is in the **80th percentile**,

it means **80% of students scored less than or equal to that student**.

Simple Data Example

Data (sorted): 10, 20, 30, 40, 50

- 50th percentile = **Median = 30**
- 100th percentile = **50**

---

4. Quartiles

Definition

Quartiles are values that **divide a dataset into four equal parts**.

Explanation

- Data is first arranged in **ascending order**
- Then divided into **4 equal parts**
- Quartiles are present at the **cut points**

Types of Quartiles

- **Q1 (First Quartile)** → 25% of data
- **Q2 (Second Quartile)** → 50% of data (Median)
- **Q3 (Third Quartile)** → 75% of data

Example (Odd Number of Values)

Data: 10, 20, 30, 40, 50

- Q2 (Median) = 30
- Q1 = Median of 10, 20 = **15**
- Q3 = Median of 40, 50 = **45**

Example (Even Number of Values)

Data: 10, 20, 30, 40, 50, 60

- Q2 = (30 + 40) / 2 = **35**
- Q1 = Median of 10, 20, 30 = **20**
- Q3 = Median of 40, 50, 60 = **50**

---

5. Inter Quartile Range (IQR)

Definition

Inter Quartile Range (IQR) is the **difference between the third quartile and first quartile**.

Formula

IQR = Q3 − Q1

Example with Solution

From previous example:

- Q3 = 50 and Q1 = 20

IQR = 50 − 20 = **30**

Use Case of IQR

- IQR is mainly used to **detect outliers**
- It is **not affected by extreme values**

Which Plot to Use if Outlier is Present and Why?

Best Plot: Box Plot

Reason

- Box plot clearly shows:
  - Minimum value, Maximum value, Quartiles, **Outliers as separate points**
- It is based on **quartiles and IQR**
- Best visualization to **detect outliers**

📊 **Summary Table (Quick Revision)**

Measure	What it Shows	Affected by Outliers	Example
Range	Max – Min	Yes	Age difference
Percentage	Value out of 100	No	Exam score
Percentile	Position in data	No	Rank
Quartiles	Divide data into 4 parts	Less	Box plot
IQR	Q3 − Q1	No	Outlier detection

Lecture 9 : Measure of Spread – Mean Deviation, Variance & Standard Deviation

---

Measure of Spread / Dispersion

Definition (Simple & Easy)

Measure of spread (or dispersion) tells us **how much the data values are spread out** or **how far they are from the center (mean)**.

Explanation

- Central tendency tells **where the center of data is**
- Dispersion tells **how scattered the data is**
- If data values are far apart → spread is **high** and If data values are close → spread is **low**

Main Measures of Dispersion

- **Mean Deviation, Variance, Standard Deviation**

---

1. Mean Deviation

Definition (Simple & Easy)

Mean deviation is the **average distance of all data values from the mean**.

Explanation

- It shows how much data values **deviate from the mean**
- Absolute values are taken so that **negative values do not cancel positives**

Important Formulas for Mean Deviation

Mean Deviation about Mean

$$\text{Mean Deviation} = \frac{\sum |x - \bar{x}|}{n}$$

Where:

- x = data value
- $\bar{x}$ = mean
- n = total number of observations

Example with Solution

Data: `2, 4, 6, 8`

Step 1: Calculate Mean

$$\bar{x} = \frac{2+4+6+8}{4} = 5$$

Step 2: Find deviation from mean

Value(x)	x – Mean	|x – Mean|
2	-3	3
4	-1	1
6	1	1
8	3	3

Step 3: Add absolute deviations

3 + 1 + 1 + 3 = 8

Step 4: Divide by total values

Mean Deviation = 8/4 = 2

---

2. Variance

Definition (Simple & Easy)

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

Explanation

- It shows spread at an **overall level**
- If spread increases → **variance increases**
- Squaring removes negative values but changes units

Key Concept

Variance tells us **how far data is spread from the mean**, on average.

Important Formulas for Variance

Population Variance

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

Sample Variance

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

Where:

- μ = population mean
- $\bar{x}$ = sample mean
- N = population size
- n = sample size

How to Calculate Variance (Step-by-Step)

1. Calculate the **mean**
2. Subtract mean from each value
3. Square each difference
4. Find the **average of squared differences**

Example with Solution

Data: 2, 4, 6, 8

Step 1: Mean

$$\bar{x} = 5$$

Step 2: Calculate squared deviations

x	x − Mean	$(x - \text{Mean})^2$
2	−3	9
4	−1	1
6	1	1
8	3	9

Step 3: Add squared deviations

9 + 1 + 1 + 9 = 20

Step 4: Divide

- Population variance:

$$\sigma^2 = \frac{20}{4} = \mathbf{5}$$

- Sample variance:

$$s^2 = \frac{20}{3} = \mathbf{6.67}$$

---

3. Standard Deviation

Definition (Simple & Easy)

Standard deviation is a measure of **how spread out numbers are from the mean**.

Explanation

- It is the **square root of variance**
- Unit is the **same as original data**
- Easier to understand than variance

Important Formulas for Standard Deviation

Population Standard Deviation

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

Sample Standard Deviation

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

Why Standard Deviation is Used

- Variance uses squared units, which are **hard to interpret**
- Standard deviation converts it back to **original unit**
- It clearly shows **where data values lie**

Example with Solution

From previous variance example:

- Population variance = 5

$$\sigma = \sqrt{5} \approx \mathbf{2.24}$$

- Sample variance = 6.67

$$s = \sqrt{6.67} \approx \mathbf{2.58}$$

Some Important Points

- In real life, we usually **do not have access to the full population**
- So, we calculate statistics using a **sample**
- Sample variance uses **(n – 1)** instead of **n**
- This makes sample variance an **unbiased estimator**
- We are **estimating population variance using sample data**

📊 **Summary Table (Quick Revision)**

Measure	Meaning	Formula Idea	Unit
Mean Deviation	Avg. distance from mean	$\text{Mean Deviation} = \frac{\sum \lvert x - \bar{x} \rvert}{n}$	$x - \bar{x}$
Variance	Avg. squared spread	$\sum (x - \bar{x})^2 / (n-1)$	Squared unit
Standard Deviation	Spread from mean	√Variance	Same as data
Population Variance	Full data spread	Divide by N	Squared
Sample Variance	Estimated spread	Divide by n−1	Squared

Lecture 10 : Measure of Symmetricity & Skewness

Measure of Symmetricity

Definition

Measure of symmetricity tells us **whether the data is evenly balanced around the center** or **tilted to one side**.

Explanation

- If data is evenly spread on both sides of the center → **Symmetric data**
- If data is more spread on one side → **Asymmetric (Skewed) data**
- Symmetricity helps us understand the **shape of data distribution**

Example

- Heights of students in a class are mostly balanced → **Symmetric**
- Income data where few people earn very high income → **Asymmetric**

Skewness (Measure of Symmetricity)

Definition

Skewness is a statistical measure that tells **how much and in which direction** the data distribution is **tilted**.

Important Meaning

- **No skewness** → Data is **perfectly symmetric**
- **Skewness = 0** → No inclination on either side

Important Formulas of Skewness

1. Karl Pearson's Coefficient of Skewness

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

If mode is not available:

$$\text{Skewness} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

---

2. Bowley's Coefficient of Skewness (Quartile Based)

$$\text{Skewness} = \frac{(Q_3 + Q_1 - 2Q_2)}{(Q_3 - Q_1)}$$

Where:

- Q_1 = First quartile
- Q_2 = Median
- Q_3 = Third quartile

Types of Skewness

---

1. Right Skewed Distribution (Positive Skewness)

Explanation

- Tail of the distribution is on the **right side**
- Most of the data lies on the **left side**
- Few very large values pull the distribution to the right

Important Points

- Tail is on the **right side**
- **Mean ≥ Median ≥ Mode**
- Skewness value is **positive**

Example

- Income distribution
- Waiting time data
- **Log-normal distribution**

---

2. Left Skewed Distribution (Negative Skewness)

Explanation

- Tail of the distribution is on the **left side**
- Most of the data lies on the **right side**
- Few very small values pull the distribution to the left

Important Points

- Tail is on the **left side**
- **Mode ≥ Median ≥ Mean**
- Skewness value is **negative**

Example

- Exam marks where most students score high
- Age at retirement in a company

Use Cases of Skewness

1. To Check Whether Data is Skewed or Not

- Using **visualization** (distribution plot / histogram)
- Using **skewness value**
- Helps understand **data behavior**

---

2. In Machine Learning Models

- Some ML algorithms assume **symmetric (normally distributed) data**
- Skewed data can reduce **model performance**
- Example: Linear Regression, Logistic Regression

If Data is Not Symmetric (Skewed Data)

We apply **data transformations** to reduce skewness.

Common Transformations

- **Log transformation**
- **Exponential transformation**
- **Reciprocal transformation**
- **Square transformation**
- **Square root transformation**
- **Box-Cox transformation**
- **Yeo-Johnson transformation**
- **Outlier treatment**

Why Data is Skewed

1. Nature of Data

- Some data is naturally skewed
  (e.g., income, sales, population)

---

2. Presence of Outliers

- Very large or very small values
- To **accommodate outliers**, distribution becomes skewed

Summary Table (Quick Revision)

Concept	Key Points (Easy Revision)
Symmetricity	Balance of data around center
Skewness	Direction and amount of tilt
No Skewness	Perfectly symmetric data
Right Skewed	Tail right, Mean ≥ Median ≥ Mode
Left Skewed	Tail left, Mode ≥ Median ≥ Mean
Skewness Formulas	Pearson, Bowley
Detection	Histogram, distribution plot, skewness
Use in ML	Some models need symmetric data
Transformations	Log, sqrt, Box-Cox, Yeo-Johnson
Reason for Skew	Nature of data, outliers

Lecture 11 : Implementation of Symmetricity & Skewness

Dataset Used

data = [2, 3, 4, 5, 8, 9, 12, 40, 41, 15]

data

Output:

[2, 3, 4, 5, 8, 9, 12, 40, 41, 15]

This data has **small values and some very large values**, so it is useful to understand **spread and skewness**.

Measure of Dispersion / Spread (Implementation)

Measure of dispersion tells us **how much the data is spread**.

---

1. Range

Definition (Simple)

Range is the **difference between maximum and minimum value**.

Python Code

max(data) - min(data)

Output:

39

Explanation

- Maximum value = 41
- Minimum value = 2
- Range = 41 − 2 = **39**
- Large range indicates **high spread and possible outliers**

---

2. Percentile

Definition (Simple)

Percentile tells us **below which a certain percentage of data lies**.

Python Code

import numpy as np

np.percentile(data, [0, 25, 50, 75, 100])

Output:

[2. 4.25 8.5 14.25 41.]

Explanation

- 0th percentile → Minimum = 2
- 25th percentile (Q1) → 4.25
- 50th percentile (Median) → 8.5
- 75th percentile (Q3) → 14.25
- 100th percentile → Maximum = 41

---

3. Inter Quartile Range (IQR)

Definition (Simple)

IQR shows the **spread of the middle 50% of data**.

Formula

IQR = Q3 - Q1

Python Calculation

14.25 - 4.25

Output:

10.0

Explanation

- IQR = **10**
- IQR is **not affected by outliers**
- Mainly used for **outlier detection**

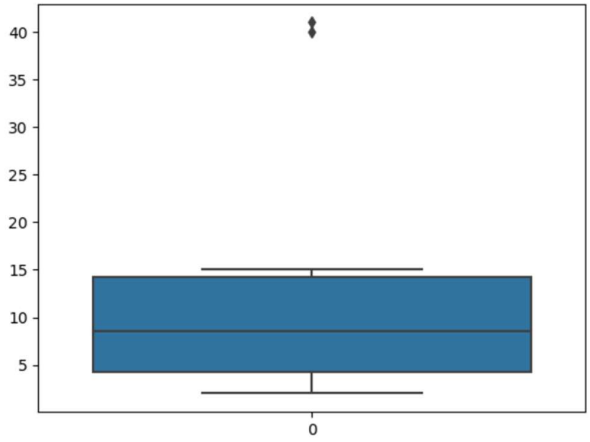
---

4. Box Plot (Outlier Detection)

Python Code

```
import seaborn as sns
sns.boxplot(data)
```

Output (Visualization):



Important Concept

- Upper fence = Q3 + 1.5 × IQR
- Lower fence = Q1 − 1.5 × IQR

Why Box Plot?

- Best plot when **outliers are present**
- Clearly shows **median, quartiles, and outliers**

---

5. Variance

Definition (Simple)

Variance tells the **average squared spread from the mean**.

Python Code (Population Variance)

```
np.var(data)
```

Output:

198.64

Explanation

- Uses denominator **N**
- Assumes data is the **entire population**
- High variance → data is **widely spread**

---

6. Standard Deviation

Definition (Simple)

Standard deviation tells **how far data values are from the mean on average**.

Python Code

np.std(data)

Output:

14.094

Explanation

- Square root of variance
- Same unit as original data
- Easier to understand than variance

Using statistics Module

Sample Variance (n – 1)

import statistics

statistics.variance(data)

Output:

220.71

Explanation

- Uses **n – 1**
- Gives an **unbiased estimate** of population variance
- Used when population data is not fully available

Population Variance

statistics.pvariance(data)

Output:

198.64

Measure of Symmetricity – Skewness

Definition (Simple)

Skewness tells **whether data is symmetric or tilted to one side**.

Python Code for Skewness

from scipy.stats import skew

skew(data)

Output:

1.24

Explanation

- Skewness > 0 → **Right skewed (positive skew)**
- Large values like **40 and 41** pull data to the right
- Mean > Median (generally)

Final Interpretation of Data

- Data has **outliers**
- Spread is **high**
- Distribution is **right skewed**
- **Median** is better than mean for central value
- **Box plot** is the best visualization

📊 **Summary Table (Quick Revision)**

Measure	**Python Function**	**Output Meaning**
Range	max() - min()	Overall spread
Percentile	np.percentile()	Data position
IQR	Q3 - Q1	Middle 50% spread
Box Plot	sns.boxplot()	Outliers detection
Variance	np.var()	Squared spread
Std Deviation	np.std()	Average spread
Sample Variance	statistics.variance()	Unbiased estimate
Skewness	skew()	Data symmetry

Lecture 12 : Set

Set

Definition (Simple & Easy)

A set is a **collection of well-defined and distinct objects**.

The objects inside a set are called **elements**.

Explanation

- Elements in a set are **unique** (no repetition)
- Order of elements **does not matter**
- Sets are usually written using **curly brackets { }**

Example

Set of numbers: A = {1, 2, 3, 4}

Set of vowels: V = {a, e, i, o, u}

Properties of a Set

1. Elements are **distinct** (no duplicates)
2. Order of elements is **not important**
3. Sets can contain **numbers, letters, or objects**
4. Sets are written using **{ } brackets**

Union of Sets

Definition

The union of two sets contains **all elements that are in either set or in both sets**.

Symbol

$A \cup B$

Example

A = {1, 2, 3}

B = {3, 4, 5}

Union:

$A \cup B$ = {1, 2, 3, 4, 5}

Explanation

- Common elements are written **only once**
- Union combines both sets

Intersection of Sets

Definition

The intersection of two sets contains **only the common elements** present in both sets.

Symbol

$A \cap B$

Example

$A = \{1, 2, 3\}$

$B = \{3, 4, 5\}$

Intersection:

$A \cap B = \{3\}$

Explanation

- Only elements present in **both sets**
- If no common element → empty set

Difference of Sets

Definition

The difference of two sets contains elements that are **present in the first set but not in the second set**.

Symbol

$A - B$

Example

$A = \{1, 2, 3\}$

$B = \{3, 4, 5\}$

Difference:

$A - B = \{1, 2\}$

Explanation

- Removes common elements
- Order matters ($A - B \neq B - A$)

Subset

Definition

A set A is a subset of set B if **all elements of A are present in B**.

Symbol

$A \subseteq B$

Example

$A = \{1, 2\}$

$B = \{1, 2, 3, 4\}$

Here:

$A \subseteq B$

Explanation

- Every element of A is inside B
- A can be **equal to B** or smaller

Superset

A set A is a superset of set B if **A contains all elements of B**.

Symbol

$A \supseteq B$

Example

A = {1, 2, 3, 4}

B = {1, 2}

Here:

$A \supseteq B$

Explanation

- Superset is the **larger set**
- It means A contains B

Symmetric Difference of Sets

Symmetric difference contains elements that are **in either of the sets but not in both**.

Symbol

$A \triangle B$

Example

A = {1, 2, 3}

B = {3, 4, 5}

Symmetric Difference:

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

Explanation

- Common elements are **removed**
- Remaining unique elements are included

📊 **Summary Table (Quick Revision)**

Operation	Meaning	Symbol	Example Result
Union	All elements of both sets	$A \cup B$	{1,2,3,4,5}
Intersection	Common elements	$A \cap B$	{3}
Difference	A elements not in B	$A - B$	{1,2}
Subset	A inside B	$A \subseteq B$	True
Superset	A contains B	$A \supseteq B$	True
Symmetric Difference	Non-common elements	$A \triangle B$	{1,2,4,5}

Lecture 13 : Covariance & Correlation

Covariance and Correlation

Covariance and correlation are statistical measures used to **understand and quantify the relationship between two variables**.

- They tell us **how two variables move together**
- Used widely in **data analysis, statistics, and machine learning**

---

1. Covariance

Definition

Covariance measures **how two variables change together**.

Explanation (Important Concepts)

- If both variables **increase together** → covariance is **positive**
- If one increases and the other decreases → covariance is **negative**
- If there is **no relationship** → covariance is **zero**
- Covariance tells only the **direction**, not the strength

Formula of Covariance

Population Covariance

$$\mathrm{Cov}(X,Y) = \frac{\sum(X_i - \mu_X)(Y_i - \mu_Y)}{N}$$

Sample Covariance

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

Example with Solution

Let

X (Study Hours): 2, 4, 6

Y (Marks): 40, 60, 80

Step 1: Mean

- Mean of X = 4
- Mean of Y = 60

Step 2: Deviations and products

X	Y	X−X̄	Y−Ȳ	Product
2	40	−2	−20	40
4	60	0	0	0
6	80	2	20	40

Sum of products = 80

Step 3: Sample Covariance

{Cov}(X,Y) = 80/2 = 40

👉 Positive covariance → **direct relationship**

Advantages of Covariance

- Shows **direction of relationship**
- Easy to calculate
- Useful in **portfolio analysis**

Disadvantages of Covariance

- Value depends on **units of data**
- Cannot compare relationships easily
- Does not show **strength of relationship**

---

2. Correlation

Definition (Simple & Easy)

Correlation measures **both direction and strength** of the relationship between two variables.

Explanation (Important Concepts)

- Correlation value lies between **−1 and +1**
- Independent of units
- More informative than covariance

Pearson Correlation Coefficient

Definition

Pearson correlation coefficient measures the **linear relationship** between two variables.

Formula

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

Where:

- r = correlation coefficient
- σ_X, σ_Y = standard deviations

Range of Pearson Correlation

Value of r	Meaning
+1	Perfect positive correlation
0	No correlation
−1	Perfect negative correlation

Example

If correlation between **study hours and marks** is:

- r = 0.85 → Strong positive relationship

Use Cases of Pearson Correlation

- Feature selection in ML
- Checking relationship between variables
- Financial market analysis
- Data preprocessing

Understanding the Relationship

1. Direct Relationship (Positive Relationship)

- Both variables move in the **same direction**
- Increase in one → increase in other

Example:

- Study hours ↑ → Marks ↑
- Height ↑ → Weight ↑

👉 Covariance > 0, Correlation > 0

---

2. Indirect Relationship (Negative Relationship)

- Variables move in **opposite directions**
- Increase in one → decrease in other

Example:

- Price ↑ → Demand ↓
- Speed ↑ → Time ↓

👉 Covariance < 0, Correlation < 0

To Quantify / Measure the Relationship

- **Covariance** → tells **direction**
- **Correlation** → tells **direction + strength**

Key Difference: Covariance vs Correlation

Basis	Covariance	Correlation
Meaning	Direction of relationship	Direction + strength
Unit dependency	Depends on units	Unit free
Range	$-\infty$ to $+\infty$	−1 to +1
Interpretability	Difficult	Easy
Use	Direction check	Relationship measurement

📊 **Summary Table (Quick Revision)**

Concept	Key Points
Covariance	Direction of relationship
Positive Covariance	Variables move together
Negative Covariance	Variables move opposite
Correlation	Strength + direction
Pearson r	Linear relationship
Direct Relationship	Both increase/decrease
Indirect Relationship	One up, other down
ML Use	Feature selection

Lecture 14 : Covariance & Correlation – Implementation

Libraries Used

import seaborn as sns

Seaborn is a **data visualization library** built on top of Matplotlib.

It also provides **built-in datasets** for practice.

Checking Available Datasets in Seaborn

sns.get_dataset_names()

Output (partial list):

['tips', 'iris', 'titanic', 'flights', 'diamonds', ...]

👉 We will use the **tips dataset**.

Loading the Dataset

df = sns.load_dataset('tips')

df

Output (sample rows):

total_bill	tip	sex	smoker	day	time	size
16.99	1.01	Female	No	Sun	Dinner	2
10.34	1.66	Male	No	Sun	Dinner	3
21.01	3.50	Male	No	Sun	Dinner	3
23.68	3.31	Male	No	Sun	Dinner	2
24.59	3.61	Female	No	Sun	Dinner	4

👉 This dataset contains **both numerical and categorical colum**

Checking First Few Rows

df.head()

Output:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

Checking Data Types

df.dtypes

Output:

total_bill	float64
tip	float64
sex	category
smoker	category
day	category
time	Category
size	Int64

dtype: object

👉 **Important Point**

- Covariance and correlation are calculated **only for numerical columns**
- Numerical columns here are:
  - total_bill
  - tip
  - size

Covariance Calculation

Important Concept

- Covariance shows **direction of relationship**
- Range can be **any value** (−∞ to +∞)
- Depends on **units of data**
- Calculated only for **numerical features**

Python Code

df.cov(numeric_only=True)

Output:

	total_bill	**tip**	**size**
total_bill	79.252939	8.323502	5.065983
tip	8.323502	1.914455	0.644875
size	5.065983	0.644875	1.489274

Explanation

- Positive values → **direct relationship**
- Example:
  - total_bill & tip covariance is positive
  - As bill increases, tip also increases

Correlation Calculation

Important Concept

- Correlation measures **direction + strength**
- Range is always between **−1 and +1**
- **Dimensionless** (unit free)
- Easier to interpret than covariance

Python Code

df.corr(numeric_only=True)

Output:

	total_bill	**tip**	**size**
total_bill	1.000000	0.675734	0.598315
tip	0.675734	1.000000	0.598315
size	0.675734	1.000000	0.489299

Explanation

- Correlation of a variable with itself = **1**
- total_bill & tip ≈ **0.67**
  - Strong **positive relationship**
- tip & size ≈ **0.49**
  - Moderate positive relationship

Heatmap for Correlation (Visualization)

Why Heatmap?

- Easy to **visualize relationships**
- Color shows **strength of correlation**
- Darker color → stronger relationship

Python Code

```
corr = df.corr(numeric_only=True)
corr
```

Output:

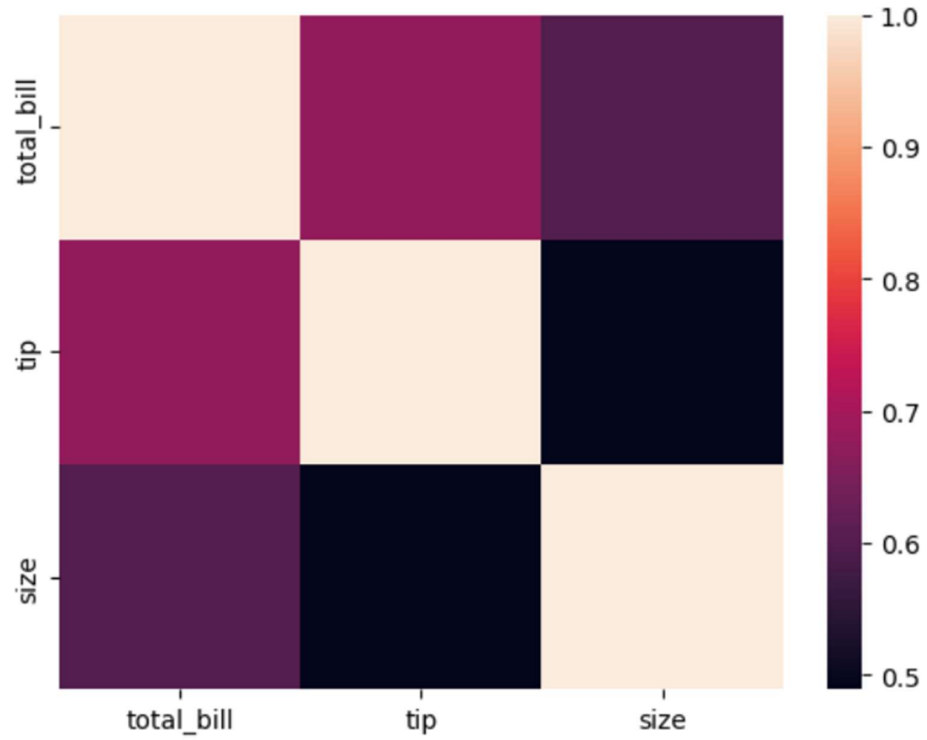
	total_bill	tip	size
total_bill	1.000000	0.675734	0.598315
tip	0.675734	1.000000	0. 489299
size	0.598315	0.489299	1.000000

Heatmap Code

```
sns.heatmap(corr)
```

Output (Visualization):

```
<AxesSubplot: >
```



Key Observations from Heatmap

- total_bill and tip → **strong positive correlation**
- size also positively related to both
- No negative correlation in this dataset

Important Points to Remember

- Covariance & correlation are calculated **only for numerical data**
- Covariance → direction only
- Correlation → direction + strength
- Heatmap helps in **quick relationship analysis**
- Very useful in **EDA (Exploratory Data Analysis)**

📊 **Summary Table (Quick Revision)**

Concept	Code Used	What It Shows
Dataset load	sns.load_dataset()	Sample data
Data types	df.dtypes	Numeric vs categorical
Covariance	df.cov()	Direction of relationship
Correlation	df.corr()	Strength + direction
Heatmap	sns.heatmap()	Visual relationship
Range (Correlation)	−1 to +1	Easy interpretation

Statistics

Module 2 : statistics advance – 1

- Random variable
- ..

Lecture 1 : Random Variable

Random Variable

A random variable is a **variable that takes numerical values** based on the **outcome of a random experiment**.

Explanation

- A random experiment is an experiment whose outcome is **not fixed**
- Random variable converts **outcomes into numbers**
- It helps in **mathematical and statistical analysis**

👉 In short: A random variable **assigns numbers to outcomes**.

Why Random Variable is Needed

- To represent random outcomes in **numerical form**
- To apply **probability and statistics formulas**
- Used in **probability distributions**

Example of Random Variable

Example 1: Tossing a Coin

Random Experiment : Tossing a coin

Possible Outcomes : Head (H), Tail (T)

Define a Random Variable

Let **X** be a random variable such that:

- X = 1 if Head occurs
- X = 0 if Tail occurs

Explanation of Example

Outcome	Random Variable Value (X)
Head	1
Tail	0

- The outcome is random
- X converts outcomes into **numbers**
- Now we can easily perform **probability calculations**

Example 2: Rolling a Dice

Random Experiment

- Rolling a dice

Possible Outcomes

- 1, 2, 3, 4, 5, 6

Define a Random Variable

Let **Y** be a random variable representing the **number obtained on dice**.

Outcome	Y
1	1
2	2
3	3
4	4
5	5
6	6

Explanation

- Each outcome has a **numerical value**
- Y is a random variable
- Values depend on chance

Example 3: Number of Heads in Two Coin Tosses

Random Experiment : Tossing two coins

Possible Outcomes : HH, HT, TH, TT

Define Random Variable

Let **Z** be the **number of heads**.

Outcome	Z
HH	2
HT	1
TH	1
TT	0

Explanation

- Z changes based on outcome
- Z takes values **0, 1, or 2**
- This is a **random variable**

Important Points about Random Variable

- Takes **numerical values**
- Value depends on **chance**
- Used in **probability distributions**
- Can be **discrete or continuous** (explained later)

📊 **Summary Table (Quick Revision)**

Concept	Key Points (Easy Revision)
Random Variable	Assigns numbers to random outcomes
Purpose	Helps in probability calculation
Coin Toss Example	Head → 1, Tail → 0
Dice Example	Dice number itself
Two Coin Toss	Counts number of heads
Nature	Depends on chance

Lecture 2 : Types of Distribution

Distribution

- A distribution shows **how data values are spread or arranged** over possible values.
- It tells us **which values occur more and which occur less**.

Explanation

- Distribution helps understand the **pattern of data**
- It shows the **behavior and shape** of data
- Used widely in **statistics, probability, and data analysis**

Different Types of Distribution

Data can follow different types of distributions depending on:

- Nature of data
- How data is generated
- Type of random variable

Main types include:

- Probability distributions
- Uniform distributions

Probability Distribution Function

A probability distribution function describes **how probability is assigned to different values of a random variable**.

Explanation

- It connects **random variables with probability**
- Total probability is always **equal to 1**
- Used to understand **likelihood of values**

Types of Probability Distribution Function

There are **two main types**:

1. Probability Density Distribution
2. Probability Mass Distribution

I. Probability Density Distribution (Continuous Data)

Meaning

Probability density distribution is used for **continuous random variables**.

Explanation

- Data can take **any value within a range**
- Probability is defined over an **interval**
- Area under the curve represents probability

Types of Probability Density Distribution

1. Normal Distribution (Bell Shaped / Gaussian Distribution)

- Data is **symmetrically distributed**
- Most values lie around the **mean**
- Mean, median, and mode are equal

---

2. Standard Normal Distribution

- Special case of normal distribution
- Mean is **0**
- Standard deviation is **1**
- Used for standardization

---

3. Log-Normal Distribution

- Data is **positively skewed**
- Log of data follows a normal distribution
- Used when values grow exponentially

---

4. Chi-Square Distribution

- Data is **positively skewed**
- Used in **hypothesis testing**
- Depends on degrees of freedom

---

5. F-Distribution

- Positively skewed distribution
- Used to **compare variances**
- Common in ANOVA tests

II. Probability Mass Distribution (Discrete Data)

Probability mass distribution is used for **discrete random variables**.

Explanation

- Data takes **countable values**
- Probability is assigned to **individual values**
- Sum of probabilities equals **1**

Types of Probability Mass Distribution

---

1. Bernoulli Distribution

- Represents **single trial** experiment
- Only two possible outcomes
- Used for yes/no type situations

---

2. Binomial Distribution

- Used for **multiple independent trials**
- Fixed number of trials
- Each trial has same probabilit

3. Poisson Distribution

- Used for **counting events**
- Events occur independently
- Used for rare events in fixed interval

Uniform Distribution

Uniform distribution is a distribution where **all values have equal probability**.

Explanation

- No value is more important than another
- Probability is **evenly spread**
- Data has no preference

Types of Uniform Distribution

---

1. Discrete Uniform Distribution

- Used for **discrete values**
- Each possible value has **same probability**
- Values are countable

---

2. Continuous Uniform Distribution

- Used for **continuous values**
- Probability is equal over an interval
- Data lies within a fixed range

📊 **Summary Table (Quick Revision)**

Topic	Key Idea (Easy Words)
Distribution	Shows how data is spread
Probability Distribution	Assigns probability to values
PDF	For continuous data
PMF	For discrete data
Normal Distribution	Symmetric, bell-shaped
Standard Normal	Mean 0, SD 1
Log-Normal	Positively skewed
Chi-Square	Hypothesis testing
F-Distribution	Variance comparison
Bernoulli	Single trial
Binomial	Multiple trials
Poisson	Event counting
Uniform Distribution	Equal probability
Discrete Uniform	Equal probability for countable values
Continuous Uniform	Equal probability over interval

Lecture 3 : Probability Distribution Function

Probability Distribution

A probability distribution describes **all possible outcomes of an experiment** and tells **how likely each outcome is**.

Example

Consider a coin toss.

Possible outcomes: Head, Tail

Each outcome has a probability:

- P(Head) = 0.5
- P(Tail) = 0.5

Explanation

- Outcomes are known
- Probability is assigned to each outcome
- Total probability is always **1**

Outcome of an Experiment

Outcomes of an experiment can be of **two types**:

1. Discrete
2. Continuous

---

1. Discrete Outcome : A discrete outcome has **countable values**.

Example

Rolling a dice.

Possible outcomes: 1, 2, 3, 4, 5, 6

Explanation

- Outcomes are **fixed and countable**
- No values exist between 1 and 2
- Used in **PMF**

---

2. Continuous Outcome : A continuous outcome can take **any value within a range**.

Example

Height of students.

Possible values: 150.2 cm, 150.25 cm, 150.256 cm, etc.

Explanation

- Infinite values are possible
- Values are **measured**, not counted
- Used in **PDF**

Probability Distribution Function (PDF – General)

A probability distribution function shows **how probabilities are distributed** over values of a random variable.

Example

If X represents number of heads in one coin toss:

- X = 0 or 1
- Each value has a probability

Explanation

- Links **random variable** to probability
- Total probability = **1**
- Can be discrete or continuous

Probability Mass Function (PMF)

- Probability Mass Function is used for **discrete random variables**.
- It assigns probability to **each individual value**.

Example

Random variable X = outcome of a coin toss

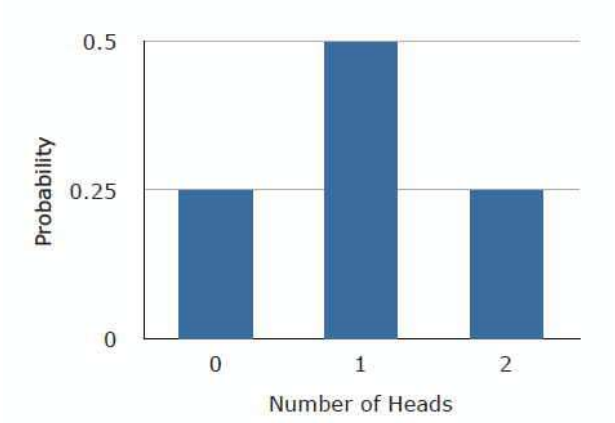
- X = 1 (Head)
- X = 0 (Tail)

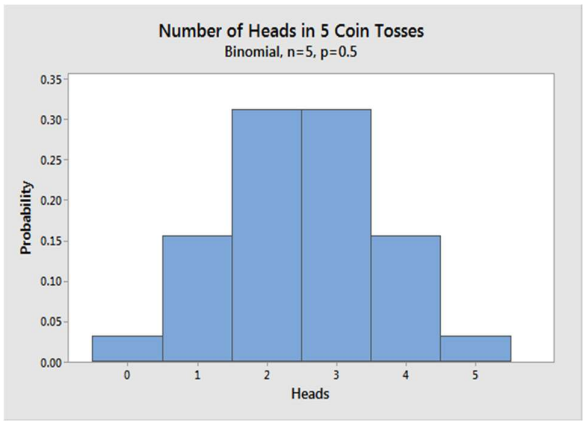
X	Probability
0	0.5
1	0.5

Explanation

- Each value has a fixed probability
- Probabilities add up to **1**
- Used for **discrete data**

PMF Graph (Conceptual)





Probability Density Function (PDF)

- Probability Density Function is used for **continuous random variables**.
- It shows probability over a **range of values**.

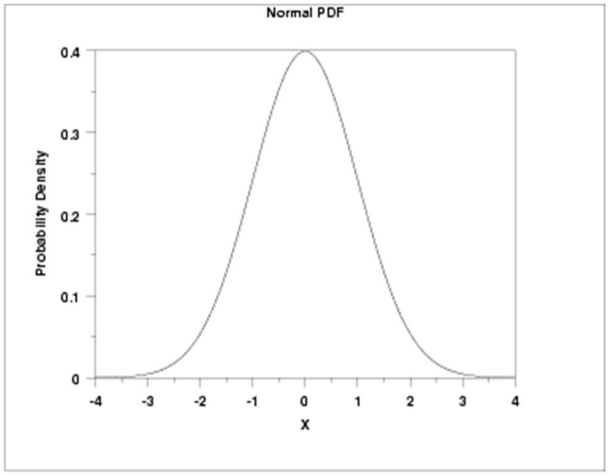
Example

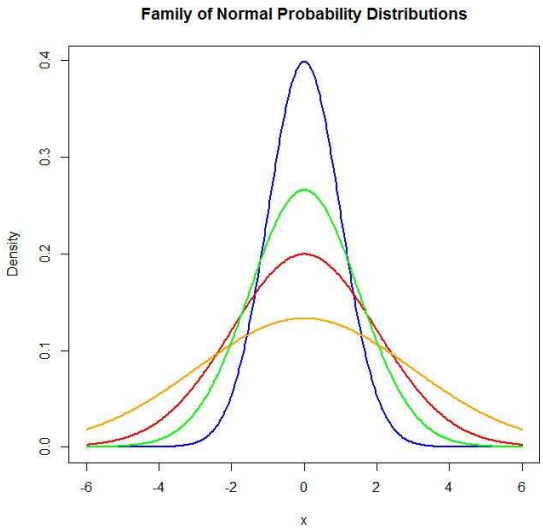
Random variable X = height of students.

Explanation

- Probability at a single point is **0**
- Probability is calculated over an **interval**
- Area under the curve = **probability**

PDF Graph (Conceptual)





Cumulative Distribution Function (CDF)

Cumulative Distribution Function gives the **probability that a random variable is less than or equal to a value**.

Example

Let X be marks of a student.

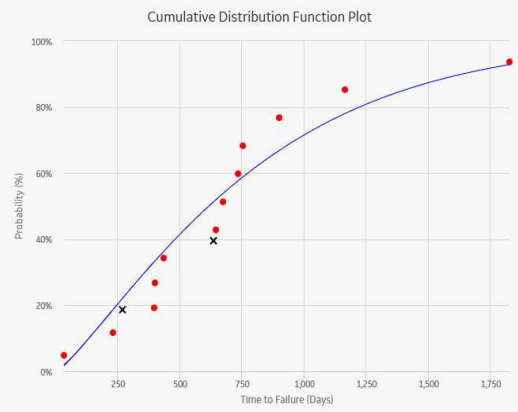
CDF tells:

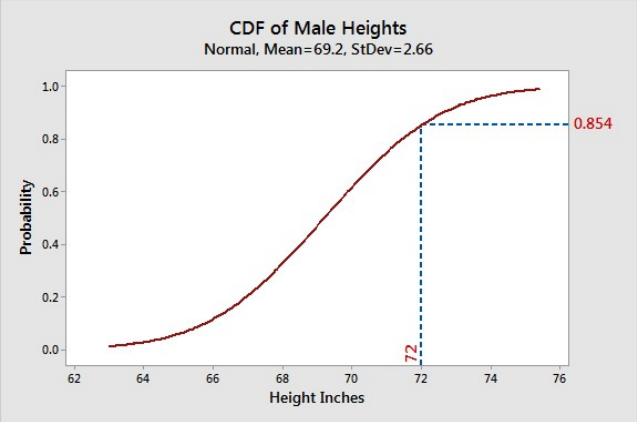
- Probability that marks ≤ 60

Explanation

- CDF is **increasing**
- Final value of CDF is **1**
- Used to find probabilities **up to a value**

CDF Graph (Conceptual)





Important Differences (Quick Understanding)

Function	Used For	Data Type	Probability Meaning
PMF	Discrete values	Countable	Exact value
PDF	Continuous values	Measurable	Range of values
CDF	Both	Both	≤ given value

Summary Table (Quick Revision)

Topic	Key Idea (Easy Words)
Probability Distribution	Shows outcomes with probabilities
Discrete Outcome	Countable values
Continuous Outcome	Measurable values
PMF	Probability of exact discrete values
PDF	Probability over a range
CDF	Probability up to a value
Total Probability	Always equals 1

Lecture 4 : Probability Distribution Function with CDF

Probability Mass Function (PMF)

Definition (Simple & Easy)

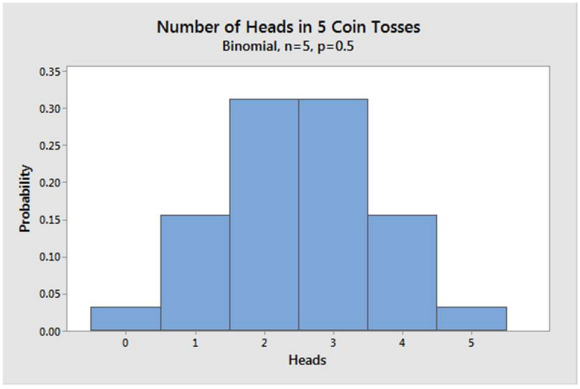
Probability Mass Function (PMF) is used for **discrete random variables**.

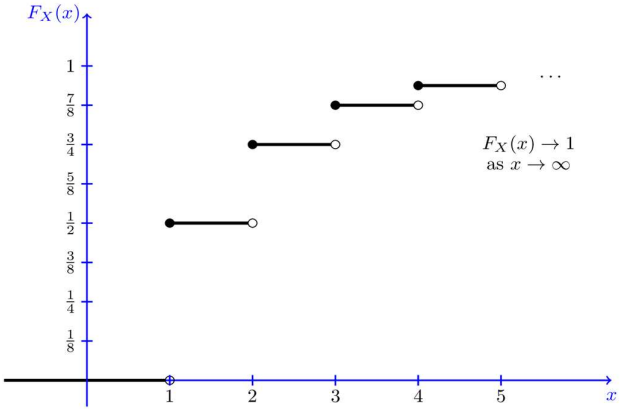
It gives the **probability of each exact value** of a random variable.

Important Points of PMF

- Used for **discrete data**
- Probability is assigned to **individual values**
- Probability values are **between 0 and 1**
- Sum of all PMF values is **1**
- Represented using **bar graph**

PMF and CDF for the Same Example (Conceptual)





Explanation

- PMF graph shows **separate bars** for each value
- CDF graph is a **step-wise increasing graph**
- CDF at a point = **sum of PMF values up to that point**

Probability Density Function (PDF)

Definition (Simple & Easy)

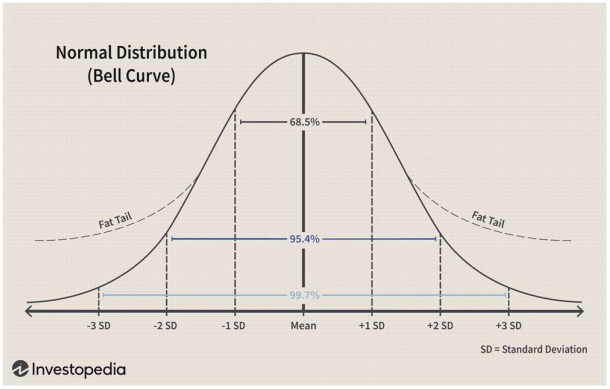
Probability Density Function (PDF) is used for **continuous random variables**.

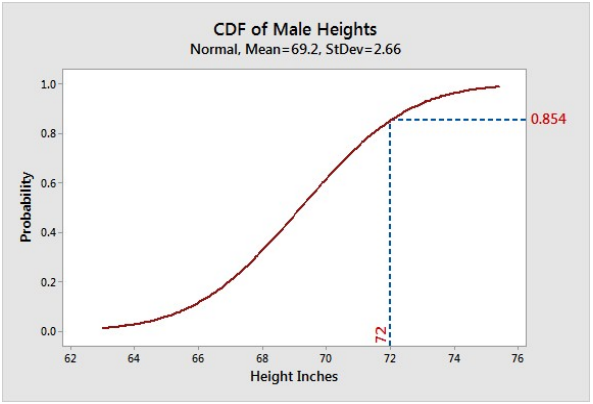
It shows how **density of probability** is spread over values.

Important Points of PDF

- Used for **continuous data**
- Probability at a single point is **0**
- Probability is found over a **range**
- Area under PDF curve = **probability**
- Represented using a **smooth curve**

PDF and CDF for the Same Example (Conceptual)





Explanation

- PDF curve shows **density of data**
- CDF curve is **smooth and increasing**
- CDF reaches **1 at the end**

Relationship between PDF and CDF

Observation 1

👉 **Probability density of PDF is the slope (gradient) of the CDF at a given point**

Simple Meaning:

- PDF tells **how fast CDF is increasing**
- Higher PDF → **steeper CDF slope**
- Lower PDF → **flatter CDF slope**

Observation 2

👉 **Initially in PDF, changes are small → slope of CDF is also small**

👉 **Near the center (around 0.5 probability), PDF peaks → CDF slope is highest**

👉 **At the end, PDF decreases → CDF slope decreases again**

⚠️ **Important Note:**

We are talking about **slope of CDF**, **not the CDF value itself**.

Can Probability Density be Greater than 1?

Answer: YES, it can be greater than 1

Why? (Simple Explanation)

- PDF value is **not probability**
- PDF represents **density**, not exact probability
- Actual probability is calculated using **area under the curve**
- As long as total area = **1**, PDF value can exceed 1

Key Point to Remember

- **Probability ≤ 1**
- **Probability Density can be > 1**
- Probability = **area under PDF curve**, not height

Quick Difference (PMF vs PDF with CDF)

Feature	**PMF**	**PDF**
Data type	Discrete	Continuous
Graph	Bar graph	Smooth curve
CDF type	Step function	Smooth curve
Probability at point	Possible	Always 0
Relation with CDF	Sum	Area / slope

📊 **Summary Table (Quick Revision)**

Concept	**Key Idea (Easy Words)**
PMF	Probability of exact discrete values
PMF Graph	Bar graph
PMF CDF	Step-wise increasing
PDF	Probability density for continuous data
PDF Graph	Smooth curve
PDF CDF	Smooth increasing curve
PDF–CDF Relation	PDF = slope of CDF
Density > 1	Possible (area matters)

Lecture 5 : Discrete Uniform Distribution

Uniform Distribution

Definition (Simple & Easy)

A uniform distribution is a distribution in which **all possible values have equal probability**.

Explanation

- Every outcome has the **same chance**
- No value is more important than another
- Probability is **evenly spread**

Example

- Outcomes of a **fair dice**
- Each number has equal probability

Types of Uniform Distribution

Uniform distribution is divided into **two types**:

1. **Discrete Uniform Distribution**
2. **Continuous Uniform Distribution**

Discrete Uniform Distribution

Definition (Simple & Easy)

Discrete uniform distribution is a distribution where:

- Values are **discrete (countable)**
- **Each value has the same probability**

Explanation

- Outcomes are fixed and countable
- Probability of each outcome is equal
- Used when outcomes are **equally likely**

Example

Consider a fair dice.

Possible values:

1, 2, 3, 4, 5, 6

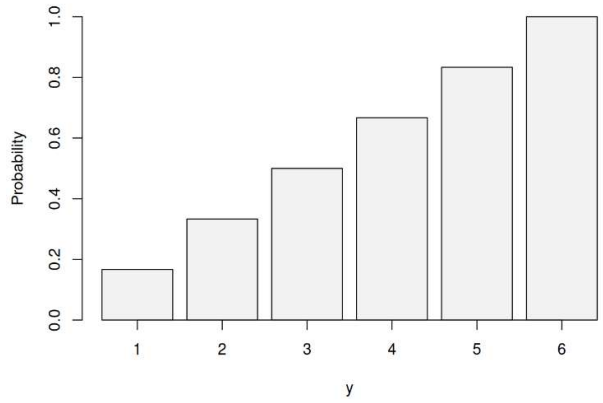
Probability of each value:

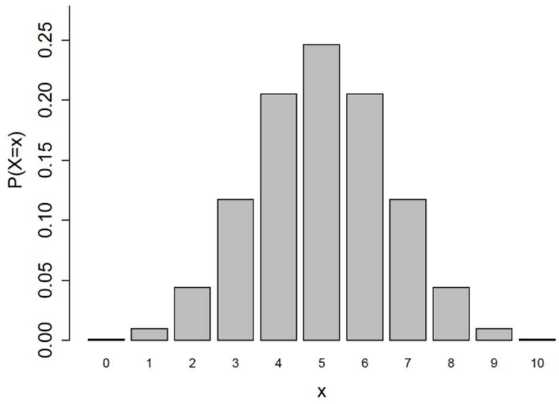
$P(X = x) = 1/6$

PMF of Discrete Uniform Distribution

Explanation

- PMF shows probability of **each exact value**
- Since it is uniform, **all bars are of equal height**
- PMF is represented using a **bar graph**



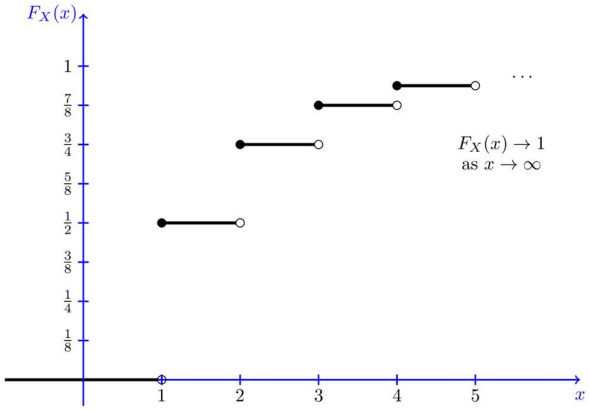


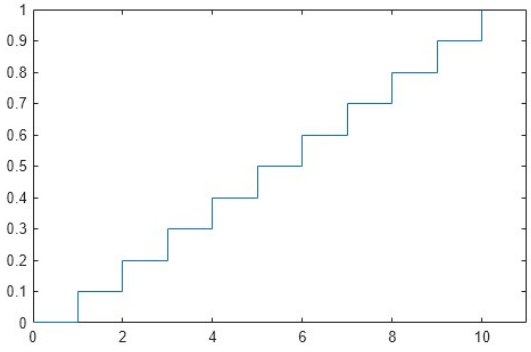
PMF Graph

CDF of Discrete Uniform Distribution

Explanation

- CDF is the **cumulative sum of PMF values**
- Graph increases in **steps**
- Final value of CDF is always **1**





CDF Graph

Mean of Discrete Uniform Distribution

Formula

$$\text{Mean} = \frac{a+b}{2}$$

Where:

- **a** = first (smallest) value
- **b** = last (largest) value

Example (Mean) with Solution

Values: 1, 2, 3, 4, 5, 6
Here: a = 1, b = 6

Mean = (1 + 6)/2 = 7/2 = 3.5

Variance of Discrete Uniform Distribution

Formula

$$\text{Variance} = \frac{(b - a + 1)^2 - 1}{12}$$

Where:

- **a** = first (smallest) value
- **b** = last (largest) value
- **(b − a + 1)** = total number of possible values

Example (Variance) with Solution

Values: 1, 2, 3, 4, 5, 6
Here: a = 1, b = 6

$$\text{Variance} = \frac{(6 - 1 + 1)^2 - 1}{12}$$

$$= \frac{6^2 - 1}{12} = \frac{35}{12} \approx \mathbf{2.92}$$

Important Points to Remember

- All values have **equal probability**
- PMF bars are of **equal height**
- CDF is **step-wise increasing**
- Mean lies at the **center**
- Variance depends on the **range of values**

📊 **Summary Table (Quick Revision)**

Topic	Key Points (Easy Words)
Uniform Distribution	Equal probability for all values
Discrete Uniform	Countable values
PMF	Equal height bars
CDF	Step-wise increasing
Mean	(a + b) / 2
Variance	((b − a + 1)^2 − 1) / 12
A	First value
B	Last value

Lecture 6 : Bernoulli Distribution

Bernoulli Distribution

Definition (Simple & Easy)

Bernoulli distribution is a probability distribution that represents an experiment with **only two possible outcomes**.

Explanation

- There are **only two outcomes**
- One outcome is called **success**
- The other outcome is called **failure**
- Probability of success is **p**
- Probability of failure is **1 – p**
- Total probability = **1**

Example of Bernoulli Distribution

Example

Tossing a coin.

- Head → Success
- Tail → Failure

Let:

- P(Head) = p
- P(Tail) = 1 − p

Explanation of Example

- Only two outcomes are possible
- If probability of one outcome is **p**
- Then probability of the other outcome must be **1 – p**
- Because total probability cannot exceed **1**

Probability Calculation

If:

- P(X = 1) = p (success)
- P(X = 0) = 1 − p (failure)

This is because:

p + (1 - p) = 1

PMF Formula of Bernoulli Distribution

Formula

$$P(X = k) = p^k \times (1 - p)^{1-k}$$

Where:

- **X** = random variable
- **k** = outcome value (0 or 1)
- **p** = probability of success
- **1 – p** = probability of failure

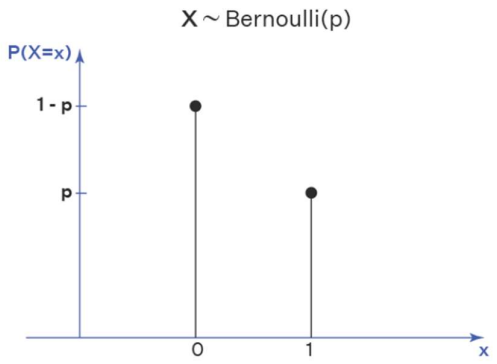
Meaning of Formula

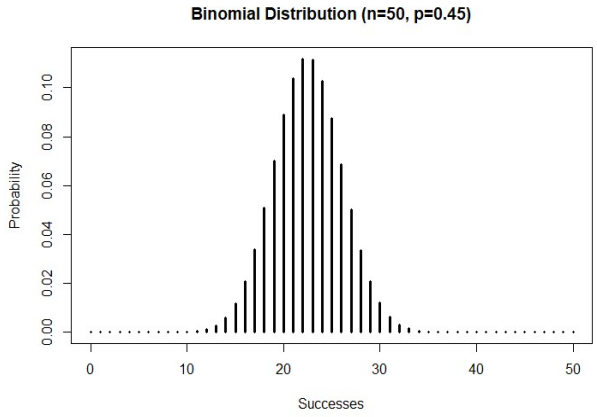
- If **k = 1**, probability = p
- If **k = 0**, probability = 1 − p

Graph of Bernoulli Distribution

Explanation

- PMF is shown using a **bar graph**
- One bar for **success (1)**
- One bar for **failure (0)**
- Heights depend on **p** and **1 − p**





PMF Graph (Bernoulli Distribution)

Conditions for Bernoulli Distribution

Bernoulli distribution is applicable when:

1. Only **two possible outcomes**
2. Outcomes are **mutually exclusive**
3. Probability of success is **constant**
4. One trial is performed
5. Outcomes are independent

Examples of Bernoulli Distribution

- Tossing a coin (Head / Tail)
- Exam result (Pass / Fail)
- Match result (Win / Loss)
- Answer (Yes / No)
- Light bulb (Working / Not working)

Solved Question

Question

Bumrah bowls 6 balls at the wicket.

The probability of **hitting the stump** on each ball is **0.6**.

What is the probability of **not hitting the wicket**?

Solution

Given:

- Probability of hitting the wicket = p = 0.6

Since only two outcomes are possible:

P(Not hitting) = 1 - p

= 1 - 0.6

= 0.4

👉 Probability of **not hitting the wicket** = **0.4**

Mean of Bernoulli Distribution

Formula

Mean = p

Where:

- **p** = probability of success

Explanation

- Average value of Bernoulli trial equals probability of success

---

Variance of Bernoulli Distribution

Formula

Variance} = p(1 - p)

Where:

- **p** = probability of success
- **1 – p** = probability of failure

Explanation

- Variance is maximum when **p = 0.5**
- Variance shows **uncertainty in outcome**

📊 **Summary Table (Quick Revision)**

Concept	**Key Points (Easy Words)**
Bernoulli Distribution	Two outcomes only
Outcomes	Success (1), Failure (0)
Probability	p and 1 − p
PMF Formula	p^k × (1−p)^(1−k)
Mean	P
Variance	p(1−p)
Graph	Bar graph
Examples	Coin toss, pass/fail

Lecture 7 : Binomial Distribution

Binomial Distribution

Definition (Simple & Easy)

Binomial distribution is a probability distribution that shows the **number of successes** in a **fixed number of independent trials**, where **each trial has only two possible outcomes**.

Explanation / Important Points

To understand binomial distribution clearly, remember:

- There are **multiple trials** (more than one)
- Each trial has **two outcomes**: success or failure
- Probability of success (**p**) remains **same** in every trial
- Trials are **independent**
- We count **number of successes**

How Binomial Distribution is Different from Bernoulli Distribution

- **Bernoulli distribution** → only **one trial**
- **Binomial distribution** → **multiple Bernoulli trials together**

👉 Binomial distribution is a **collection of Bernoulli trials**.

Example of Binomial Distribution

- Tossing a coin **10 times** and counting number of heads
- Exam with **10 MCQs**, counting number of correct answers
- Factory produces bulbs, counting **defective bulbs** in a batch

---

PMF Formula of Binomial Distribution

Formula :

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Where:

- **X** = random variable (number of successes)
- **n** = total number of trials
- **k** = number of successes
- **p** = probability of success in one trial
- **1 – p** = probability of failure
- **(\binom{n}{k})** = number of ways to choose k successes from n trials

---

Question 1

Question

A coin is tossed **3 times**.

What is the probability of getting **exactly 2 heads**?

Solution

Given:

- Number of trials = **n = 3**
- Number of heads = **k = 2**
- Probability of head = **p = 0.5**
- Probability of tail = **1 − p = 0.5**

Step-by-Step Calculation

$$P(X = 2) = \binom{3}{2}(0.5)^2(0.5)^1$$

$$= 3 \times 0.25 \times 0.5$$

$$= \mathbf{0.375}$$

👉 Probability of getting exactly **2 heads = 0.375**

Question 2

Question

A coin is tossed **10 times**.

What is the probability of getting **exactly 3 tails**?

Solution

Given:

- Number of trials = **n = 10**
- Number of tails = **k = 3**
- Probability of tail = **p = 0.5**
- Probability of head = **1 − p = 0.5**

Step-by-Step Calculation

$$P(X = 3) = \binom{10}{3}(0.5)^3(0.5)^7$$

$$= \binom{10}{3}(0.5)^{10}$$

$$= 120 \times \frac{1}{1024}$$

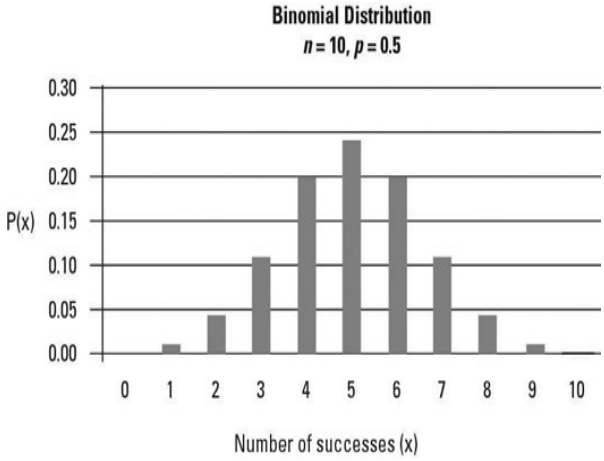
$$= \mathbf{0.117}$$

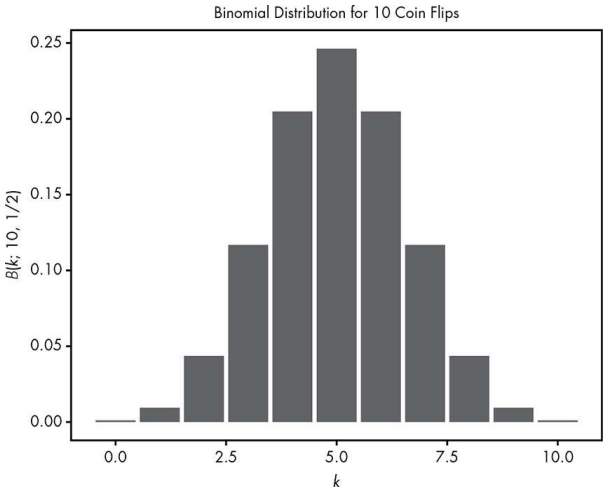
👉 Probability of getting exactly **3 tails = 0.117**

Graph Representation (for Question 2) : PMF Graph (Binomial Distribution)

Explanation

- X-axis → number of tails
- Y-axis → probability
- Bar graph shows probability for each possible value
- Highest bar appears near the center





Mean of Binomial Distribution

Formula

Mean = n * p

Where:

- **n** = number of trials
- **p** = probability of success

Variance of Binomial Distribution

Formula

Variance = n * p * (1 - p)

Where:

- **n** = number of trials
- **p** = probability of success
- **1 – p** = probability of failure

Important Points to Remember

- Binomial distribution = **multiple Bernoulli trials**
- Probability **p** must be constant
- Outcomes must be **independent**
- Used widely in **probability, statistics, and ML**

📊 Summary Table (Quick Revision)

Concept	Key Points (Easy Words)
Binomial Distribution	Multiple trials, two outcomes
Difference from Bernoulli	Bernoulli = 1 trial
PMF Formula	nCk · p^k · (1−p)^(n−k)
Mean	n × p
Variance	n × p × (1−p)
Example	Coin toss multiple times
Graph	Bar graph (PMF)

Lecture 8 : Poisson Distribution

Poisson Distribution

Definition (Simple & Easy)

Poisson distribution is a probability distribution that shows the **number of times an event occurs in a fixed interval of time or space**, when events happen **randomly and independently**.

Explanation / Important Points

- Used to **count events**
- Events occur **randomly**
- Events are **independent**
- Average rate of occurrence is **constant**
- Time or space interval is **fixed**
- Probability of more than one event at the same instant is **very small**

Where Poisson Distribution is Used

- Number of customers entering a shop
- Number of calls in a call center
- Number of accidents in a day
- Number of defects in a product

PMF Formula of Poisson Distribution

Formula

$$P(X = k) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

Where:

- **X** = random variable (number of events)
- **k** = number of occurrences (0, 1, 2, ...)
- **λ (lambda)** = average number of occurrences in a fixed interval
- **e** = Euler's number (approximately 2.718)
- **k!** = factorial of k

Solved Question

Question

The average number of customers entering a store in **one hour** is **5**.

What is the probability that **exactly 3 customers** will enter the store in the next hour?

Solution

Given:

- Average rate = λ = **5**
- Number of customers = k = **3**

Step-by-Step Calculation

$$P(X = 3) = \frac{e^{-5} \times 5^3}{3!}$$

$$= \frac{e^{-5} \times 125}{6}$$

$$\approx \frac{0.0067 \times 125}{6}$$

$$\approx \frac{0.8375}{6}$$

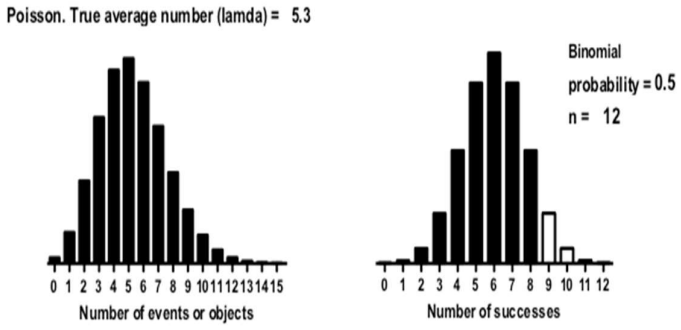
$$\approx \mathbf{0.1396}$$

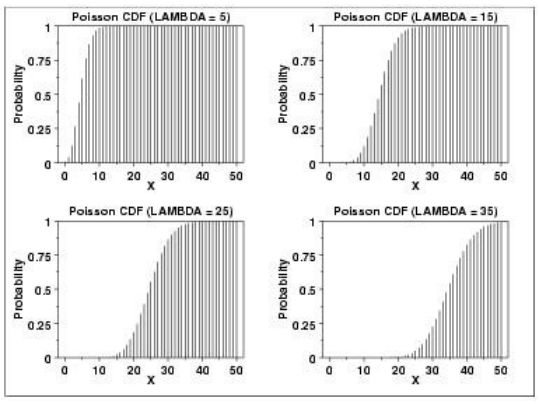
👉 Probability that **exactly 3 customers** enter the store = **0.1396**

Graph Representation (Poisson Distribution) : PMF Graph (Poisson Distribution)

Explanation

- X-axis → number of customers
- Y-axis → probability
- Bar graph shows probability for each value
- Highest bar appears near the **average value (λ)**





Mean of Poisson Distribution

Formula

Mean = λ

Where:

- **λ (lambda)** = average number of events

Explanation

- Mean equals the **average rate of occurrence**
- Center of the distribution

Variance of Poisson Distribution

Formula

Variance = λ

Where:

- **λ (lambda)** = average number of events

Explanation

- Variance is equal to mean
- Spread depends only on λ

Important Points to Remember

- Poisson distribution is used for **rare or count events**
- Mean = Variance = λ
- Used when events occur **randomly**
- No fixed number of trials is required
- Works best for **large population and small probability**

📊 **Summary Table (Quick Revision)**

Concept	**Key Points (Easy Words)**
Poisson Distribution	Counts events in fixed interval
λ (Lambda)	Average rate
PMF Formula	(e^−λ · λ^k) / k!
Mean	Λ
Variance	Λ
Graph	Bar graph
Example	Customers entering store

Lecture 9 : Continuous Uniform Distribution

Uniform Distribution

Definition (Simple & Easy)

A uniform distribution is a distribution in which **all values within a given range have equal probability**.

Explanation

- Every value in the range is **equally likely**
- There is **no preference** for any value
- Probability is spread **evenly over the interval**

Important Points

- Used for **continuous data**
- Probability depends on a **range**, not a single value
- Total probability is always **1**

Examples

- **OTP guessing** between 0000 and 9999
- **Guessing exact time** between 2:00 PM and 3:00 PM
- **Waiting time** for a bus between 5 and 10 minutes
- **Temperature** between 20°C and 25°C during a day

Notation of Continuous Uniform Distribution

Notation

$$X \sim U(a, b)$$

Where:

- **a** = minimum value
- **b** = maximum value
- **U** = uniform distribution

PDF of Continuous Uniform Distribution

Formula

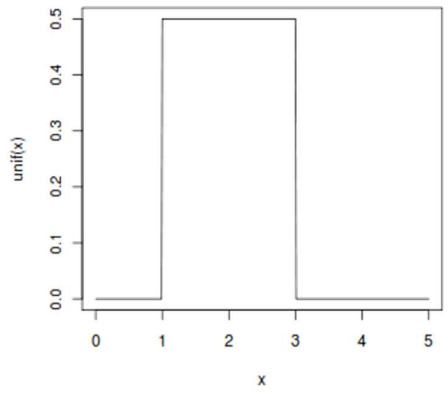
$$f(x) = \frac{1}{b - a}$$

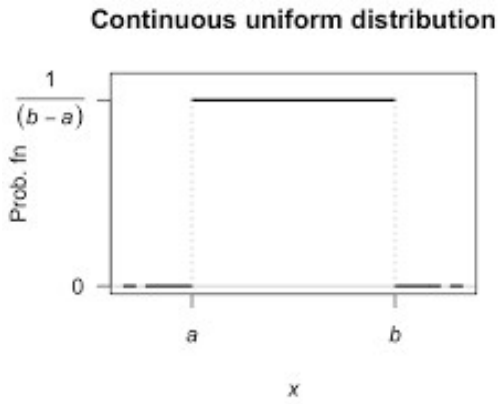
Where:

- **a** = minimum value
- **b** = maximum value
- **x** = any value between a and b

Explanation

- PDF is a **constant horizontal line**
- Height is same for all values
- Area under the curve = **1**





PDF Graph (Continuous Uniform Distribution)

CDF of Continuous Uniform Distribution

Definition

CDF gives the **probability that the random variable is less than or equal to x**.

CDF Formula

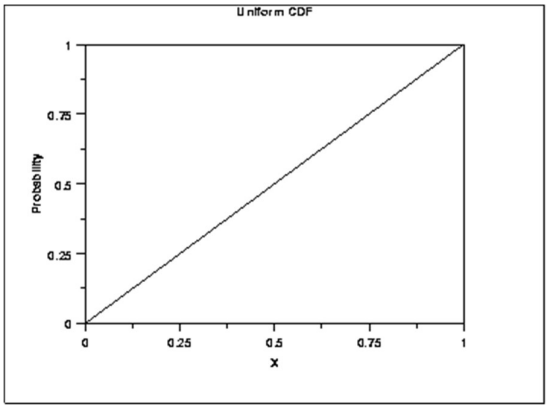
$$F(x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & a \le x \le b \\ 1, & x > b \end{cases}$$

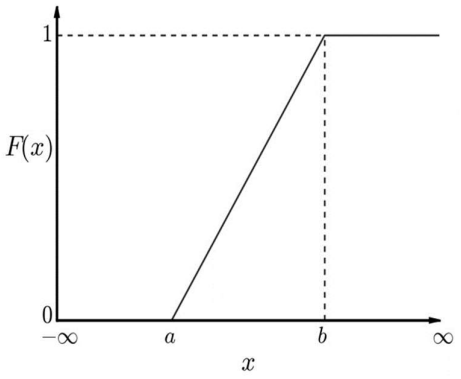
Where:

- **a** = minimum value
- **b** = maximum value
- **x** = value of random variable

Explanation

- Before **a**, probability is 0
- Between **a and b**, probability increases linearly
- After **b**, probability becomes 1





CDF Graph (Continuous Uniform Distribution)

Mean / Median of Continuous Uniform Distribution
Formula

$$\text{Mean} = \text{Median} = \frac{a+b}{2}$$

Where:

- **a** = minimum value
- **b** = maximum value

Explanation

- Center of the distribution
- Mean and median are **same**

Variance of Continuous Uniform Distribution
Formula

$$\text{Variance} = \frac{(b-a)^2}{12}$$

Where:

- **a** = minimum value
- **b** = maximum value

Question 1
Question
The number of items sold daily in a shop is **uniformly distributed**.
Minimum sales = **20**, Maximum sales = **50**.

Given

- a = 20
- b = 50

I. Probability that sales fall between 25 and 40

Formula

$$P(25 \leq X \leq 40) = \frac{40 - 25}{50 - 20}$$

Solution

$$= \frac{15}{30} = \mathbf{0.5}$$

Final Answer : Probability = **0.5**

II. Probability that sales are more than 30 and up to maximum

$$P(30 \leq X \leq 50) = \frac{50 - 30}{50 - 20}$$

Solution

$$= \frac{20}{30} = \mathbf{0.6667}$$

Final Answer : Probability ≈ **0.67**

Question 2

Question

The delivery time of pizza is **uniformly distributed** between **15 and 30 minutes**.

Find the **standard deviation** of delivery time.

Given

- a = 15
- b = 30

Step 1: Variance

$$\text{Variance} = \frac{(30 - 15)^2}{12}$$

$$= \frac{225}{12} = 18.75$$

Step 2: Standard Deviation

$$\text{Standard Deviation} = \sqrt{18.75} \approx \mathbf{4.33}$$

Final Answer : Standard deviation ≈ **4.33 minutes**

📊 Summary Table (Quick Revision)

Topic	Key Points (Easy Words)
Uniform Distribution	Equal probability
Continuous Uniform	Values in a range
Notation	U(a, b)
PDF	$1 / (b - a)$
CDF	Linear increase
Mean	$(a + b) / 2$
Variance	$(b - a)^2 / 12$
Use Case	Time, distance, guessing
