📘 Day 4 – Central Tendency & Data Distribution

🔷 1. Outliers

👉 **Outlier = a data point that is far away from the rest.**

It can be extremely small or large compared to typical values.

Example

Data = [2, 3, 3, 4, 5, 100]

- Mean = ~19.5 (pulled up by 100)
- Median = 3.5 (stable)

💡 **Impact:**

- **Mean** → sensitive to outliers.
- **Median** → robust (hardly changes).
- **Mode** → unaffected unless outlier repeats.

Real-life examples

- A millionaire in a survey of middle-class salaries.
- One wrong sensor reading in temperature data.

🔷 2. Mode (in depth)

- **Definition:** Most frequent value OR the highest peak in distribution.
- Can be used for **both numerical & categorical data**.

Types of Mode

- **Unimodal** → One peak.

- **Bimodal** → Two peaks (e.g., exam scores: weak group + strong group).
- **Multimodal** → More than two peaks.

Example with Histogram

```
Interval   Count
0.5 - 1     3
1 - 1.5     0
1.5 - 2     5
2 - 2.5     0
2.5 - 3     7   ← Highest peak (Mode interval)
3 - 3.5     0
3.5 - 4     1
4 - 4.5     0
4.5 - 5     4
```

Mode = **interval 2.5 – 3**

🔷 3. Skewness

👉 Skewness = how "asymmetrical" a distribution is.

- **Right Skewed (Positive Skew):** Long tail to the right.
  - Order: **Mode < Median < Mean**
  - Example: Salaries in a company (few very rich).
- **Left Skewed (Negative Skew):** Long tail to the left.
  - Order: **Mode > Median > Mean**
  - Example: Age at death in developed countries (few early deaths).
- **Normal (No Skew):** Symmetric bell curve.
  - Order: **Mean = Median = Mode**

🔷 4. Data Transformation

👉 Why? Because many models (like regression, ML algorithms) assume **normal distribution**.

If data is skewed, we transform it.

Common Transformations

- Reciprocal: x→1xx \to \frac{1}{x}
- Log: x→log(x)x \to \log(x)
- Square Root: x→xx \to \sqrt{x}
- Exponential: x→exx \to e^x
- Box-Cox, Yeo-Johnson (advanced ML techniques).

💡 Example: Income data (right skewed) → apply log → becomes closer to normal.

🔷 5. Normal Distribution

The most important distribution in statistics 🚀

Properties

1. **Bell-shaped curve**.
2. **Symmetry** → 50% left, 50% right.
3. **Mean = Median = Mode**.
4. **Asymptotic tails** → curve never touches x-axis.
5. **Empirical Rule (68–95–99.7 Rule):**
    - 68% of data within ±1σ
    - 95% within ±2σ
    - 99.7% within ±3σ

Real-life examples

- Human heights
- IQ scores

- Measurement errors

🔷 6. Mean vs Median vs Mode – Final Comparison

Feature	Mean	Median	Mode
Definition	Arithmetic average	Middle value	Most frequent value
Best for	Symmetric data	Skewed data	Categorical data
Sensitive to outliers?	✅ Yes	❌ No	❌ No
Example use	Avg marks in exam	Typical salary	Most bought product

📝 Practice Problems

1. Dataset: [5, 6, 7, 8, 9, 100]
   - Find mean, median. Which better represents central tendency?
2. Which skewness applies?
   - (a) Salaries in India
   - (b) Ages of death in Japan
   - (c) Marks in an easy exam (most students score high).
3. True/False:
   - In a normal distribution, **mean > median**.
   - Outliers affect mean more than median.
   - A dataset can have more than one mode.

👉 That's the **Day 4 Deep Dive**. We've connected:

- Outliers 🔥
- Mode in detail
- Skewness (left/right/normal)
- Data transformations
- Normal distribution