📘 Day 7 – Outliers, Mode in Distributions, Skewness & Data Transformation

---

1️⃣ Outliers

👉 **Definition:**

An **outlier** is an observation that is **unusually high or unusually low** compared to the rest of the data.

Key effects:

- **Mean** → heavily affected (pulled toward the outlier).
- **Median** → not affected (depends only on position).
- **Mode** → usually not affected unless outlier repeats.

🔍 How to Detect Outliers

1. **Rule of Thumb:**
   - Any value **< Q1 – 1.5×IQR** or **> Q3 + 1.5×IQR** is an outlier.

     (Q1 = 25th percentile, Q3 = 75th percentile, IQR = Q3 − Q1).
2. **Standard Deviation Rule (Z-Score):**
   - If $|z| > 3$, value is an outlier.
   - $z = \frac{x - \bar{x}}{\sigma}$
3. **Boxplot:**
   - Outliers appear as dots beyond whiskers.

⚡ How to Deal with Outliers

- **Investigate** → Sometimes an outlier is a real, meaningful event (e.g., stock market crash).
- **Remove** → If clearly a mistake (e.g., typing 10000 instead of 100).
- **Transform** → Apply $log(x)$ or $\sqrt{x}$ transformation to reduce their effect.
- **Use robust statistics** → Median instead of mean.

2️⃣ Mode from Distribution

👉 **Mode in a raw dataset:** Most frequent value.

👉 **Mode in a distribution plot:** The **highest peak (modal class/interval)**.

Example Histogram Data

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

- Mode = **interval 2.5 – 3**
- Frequency = **7 values** fall inside.

👉 **Important:** For distributions, you can't say "exact mode = 2.7". You say:

"The modal class is 2.5 – 3, with frequency 7."

Types of Modes

- **Unimodal** → One peak.
- **Bimodal** → Two peaks.

- **Multimodal** → More than two peaks.

3️⃣ Skewness

👉 Skewness = asymmetry of data due to outliers.

- **Left Skew (Negative Skew):**
  - Caused by low outliers.
  - Order: Mode > Median > Mean
- **Right Skew (Positive Skew):**
  - Caused by high outliers.
  - Order: Mode < Median < Mean
- **No Skew (Normal Distribution):**
  - Balanced, symmetric.
  - Order: Mean = Median = Mode

4️⃣ Data Transformations

👉 Why transform?

- Many math models assume **normal distribution**.
- Real-world data often skewed.
- Transformations make data closer to normal.

Common Transformations

$Reciprocal : x \rightarrow 1/xx \rightarrow 1/x$

$Log : x \rightarrow log(x)x \rightarrow \log(x)$

$SquareRoot : x \rightarrow xx \rightarrow \sqrt{x}$

$Exponential : x \rightarrow exx \rightarrow e^x$

- Power methods (Box-Cox, Yeo-Johnson).

👉 Example: Income data (very skewed) → log transform → distribution becomes closer to normal.

5️⃣ Normal Distribution (Revisited)

Properties:

1. Bell-shaped curve.
2. Symmetry → 50% data left, 50% right.
3. Used in exams (CAT, GMAT, GATE), natural phenomena.
4. **Asymptotes:** curve never touches x-axis.
5. **Mean = Median = Mode.**

✅ Summary

- **Outliers** → extreme values, affect mean but not median.
- **Mode** in distributions → modal class, highest peak.
- **Skewness:**
  - Left skew: Mode > Median > Mean.
  - Right skew: Mode < Median < Mean.
  - Normal: Mode = Median = Mean.
- **Transformations** help convert skewed data into normal.
- **Normal distribution** is the gold standard assumption in many math models.

📝 Practice Questions

1. Data = {2, 3, 4, 5, 100}
   - Mean, Median, Outlier detection.
   - Which central tendency measure is better here?
2. In a histogram, the tallest bar is in interval 40–50 with frequency 12.

- What is the mode?

3. A dataset of exam marks shows Mode < Median < Mean.
    - Is the data skewed left, right, or normal?
