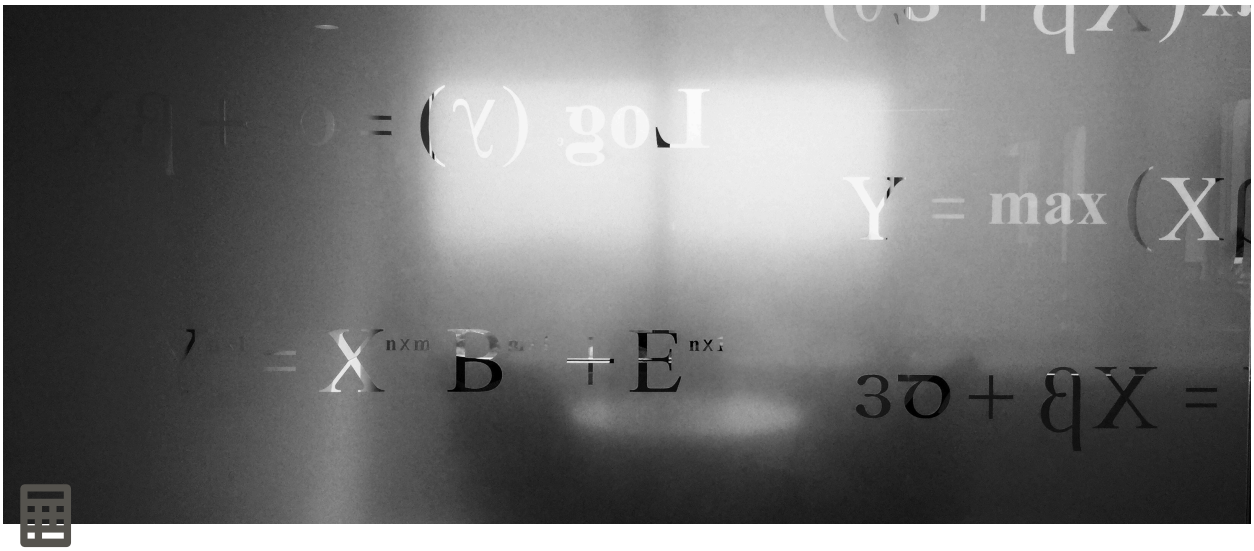

Statistics and Probability

📊 Day 1 Basics of Statistics

🌟 What is Statistics?

Definition (simple):

Statistics is the art + science of **collecting, organizing, analyzing, and interpreting data** so we can make decisions.

👉 Without statistics = guessing.

👉 With statistics = decisions based on evidence.

1️⃣ The Data Cycle (Pizza Party Example Extended)

When planning a **pizza party** 🍕, you go through these steps:

1. **Collect Data** → Ask friends their favorite pizza & slices needed.
2. **Organize Data** → Make a table of preferences.
3. **Analyze Data** → Count which flavor is most common.
4. **Interpret Data** → Decide how many of each flavor to order.
5. **Present Data** → Show in chart so all friends agree.
6. **Decide/Conclude** → Place the order! 🎯

This cycle = the **statistical process**.

2️⃣ Types of Data

a) **Categorical (Qualitative)**

- Labels, categories, groups.
- No meaningful arithmetic.

Examples:

- Blood Group (A, B, AB, O)
- Favorite color (Red, Blue, Green)
- Movie Genre (Action, Comedy, Horror)

👉 Used with **bar charts, pie charts**.

---

b) **Numerical (Quantitative)**

- Measurable, numeric values.
- Arithmetic makes sense.

Two types:

1. **Discrete (counts, no decimals):** number of students, dice rolls 🎲.
2. **Continuous (measurements, decimals allowed):** height, weight, temperature 🌡️.

👉 Used with **histograms, boxplots, scatter plots**.

3️⃣ Levels of Data Measurement

Why does this matter? Because **the math you do depends on the level**.

1. **Nominal 🏷️ (Name only)**
   - No order.
   - Example: Car brands, eye color.
2. **Ordinal 📶 (Order matters, but no exact differences)**
   - Example: Ratings (Poor < Fair < Good < Excellent).
   - Can rank, but can't say "twice as good."
3. **Interval 🌡️ (Numbers, no true zero)**
   - Example: Temperature (0°C ≠ no heat).
   - Can add/subtract, but ratios meaningless.
4. **Ratio ⚖️ (Numbers, true zero)**
   - Example: Height, weight, salary.
   - Ratios make sense (100kg is twice 50kg).

👉 Shortcut Memory Trick:

- **Nominal = Name**
- **Ordinal = Order**
- **Interval = Intervals only (no true zero)**
- **Ratio = Real zero**

4️⃣ Population vs Sample

- **Population (N):** The entire group you want to study.
- **Sample (n):** A smaller subset of the population, chosen to represent it.

📖 Example:

- Population = all students in India 🇮🇳
- Sample = 1,000 students chosen fairly from across India.

👉 Why use samples? Because population study is **too big, expensive, and slow**.

5️⃣ Why Sampling Matters

- Saves **time** and **cost**.
- If done well, gives results almost as good as studying the whole population.
- Used in:
  - Elections (exit polls)
  - Market surveys
  - Quality testing (check 10 items, not all 10,000)

6️⃣ Real-World Connections

- **Descriptive Statistics** → Summarize data you already have.

  👉 Example: Average marks of your class = 72.
- **Inferential Statistics** → Use sample to guess about population.

  👉 Example: From 500 voters surveyed, predict election winner.

7️⃣ Quick Recap (Sticky Notes for Revision)

- **Statistics** = turning messy data into decisions.
- **Data Types**:
  - Categorical (labels)
  - Numerical → Discrete & Continuous
- **Levels of Measurement**:
  - Nominal → Name
  - Ordinal → Order
  - Interval → No true zero
  - Ratio → True zero
- **Population vs Sample**:
  - Population = whole group
  - Sample = subset
- **Sampling** = saves time, money, effort.

8️⃣ Practice Questions

1. Identify data type (categorical/discrete/continuous/level):
   - Temperature 🌡️
   - Blood group 🩸
   - Movie rating (Flop < Hit < Blockbuster)
   - Salary 💰
2. In a survey, 200 students are in total (population). If you pick 40 at random, what's your sample size?
3. Which measure is correct for each?

- Gender distribution → Bar chart / Pie chart?
- Heights of students → Histogram or bar chart?

👉 That's the **Day 1 Deep Dive**: foundation, building blocks, and real connections.

Would you like me to also create a **mini real-world case study** (like a market survey of favorite smartphones) so you see **Day 1 applied end-to-end**?

📊 Day 2 – Data Representation

1️⃣ Why Represent Data?

- Raw data is just numbers or labels → hard to spot patterns.
- Representation = **organizing and visualizing** → patterns & trends become visible.

💡 Example: Imagine you collected marks of 50 students. If you just list them out, it's confusing. But if you group and plot → instantly clear.

2️⃣ Representing Categorical Data

$$Categorical = labels, groups, classes (gender, bloodgroup, favoritecolor).$$

a) Frequency Table

Shows how often each category occurs.

👉 Example: Favorite Fruit 🍎🍌🍇

Fruit	Frequency
Apple	10
Banana	7
Grapes	3

b) Relative Frequency Table

Converts counts into percentages (better for comparison).

Fruit	Frequency	Relative Frequency
Apple	10	50%
Banana	7	35%
Grapes	3	15%

c) Graphical Representation

- **Bar Chart** → height of bar = frequency/percentage.
- **Pie Chart** → each slice shows proportion of the whole.

👉 When to use which?

- Bar chart → better for comparing categories (clear differences).
- Pie chart → better when showing parts of a whole (market share, budget split).

3️⃣ Representing Numerical Data

$$Numerical = numbers/measurements (age, marks, salary).$$

Problem:

A raw list of marks is messy:

15, 10, 21, 35, 27, 12, 31, 24, 18, 33

Solution → Group into Intervals

Frequency Distribution Table:

Marks Range	Frequency
10–15	3
15–20	1
20–25	2
25–30	1
30–35	3

This shows how many students fall into each range.

Graphical Representation

- **Histogram** → bars touch (continuous data).
- **Frequency Polygon** → join midpoints of histogram bars with lines.

👉 Difference from Bar Chart:

- Histogram = continuous numerical data (bars touch).
- Bar Chart = categorical data (bars separated).

4️⃣ Relative Frequency

Instead of raw counts, use proportions.

Formula:

$$Relative Frequency = \frac{\text{Frequency of Class}}{\text{Total Frequency}}$$

Example: If 30 boys & 20 girls in class of 50 →

- Boys = 30/50 = 0.6 = 60%
- Girls = 20/50 = 0.4 = 40%

5️⃣ Why Group Data into Intervals?

- Large datasets → easier to read.
- Spot trends quickly.
- Helps in building graphs like histogram, polygon, ogive (cumulative frequency).

💡 Example: In our marks table, you can clearly see most students scored between **30–35 marks**.

6️⃣ Quick Comparison

Data Type	Best Representation	Example
Categorical	Bar Chart / Pie Chart	Favorite color, Gender
Numerical (Discrete)	Histogram / Frequency Table	Dice rolls, Goals scored
Numerical (Continuous)	Histogram / Polygon	Height, Marks, Salary

7️⃣ Advanced Twist

- **Cumulative Frequency (Ogive):** shows running total → good for percentiles.
- **Relative Frequency Histogram:** bar heights represent percentages, not raw counts.

8️⃣ Practice Mini-Tasks

1. Survey 10 people: "Favorite Social Media App (Instagram, YouTube, WhatsApp, X)". Make:
   - Frequency Table
   - Relative Frequency (%)
   - Bar Chart (choose app with highest share).
2. Collect ages of 8 family members. Group into intervals (e.g., 0–10, 11–20...) and make a histogram.
3. From the histogram, answer:
   - Which age group is most common?
   - Is data spread evenly or skewed?

✅ **Summary (Sticky Notes for Revision):**

- Categorical → Frequency Table, Bar, Pie.
- Numerical → Frequency Distribution, Histogram, Polygon.
- Histogram vs Bar Chart → touching vs separated bars.
- Relative Frequency = % comparison.
- Grouping → makes patterns visible.

📘 Day 3 — Central Tendency

🌟 Why Central Tendency?

When we have a dataset, we often want to answer:

👉 "What is a typical value here?"

For example, if I ask **"How much do students score in SSC?"**, you won't read me 100 marks one by one — you'll give me one **representative number**.

That's what central tendency gives us.

There are **3 tools**:

1. **Mean (average)**
2. **Median (middle)**
3. **Mode (most frequent)**

1️⃣ Mean (Average)

📌 **Formula**

$$Mean \bar{x} = \frac{\text{Sum of observations}}{\text{Number of observations}}$$

📖 **Example**:

Marks = 91, 81, 92, 89, 90, 94

$x^- = 91 + 81 + 92 + 89 + 90 + 946 = 5376 = 89.5\bar{x} = \frac{91+81+92+89+90+94}{6} = \frac{537}{6} = 89.5$

👉 **Interpretation**: The average student scored ~90.

💡 **Strength**: Uses all values.

⚠️ **Weakness**: Very sensitive to outliers.

2️⃣ Median (Middle Value)

📌 **Steps**

1. Sort the data.
2. If odd count → middle element.
3. If even count → average of 2 middle elements.

📖 **Example 1 (Odd count)**

Data: 1, 5, 20, 21, 16, 17, 3

Sorted → 1, 3, 5, 16, 17, 20, 21

Median = **16**

📖 **Example 2 (Even count)**

Data: 1, 5, 20, 21, 16, 17, 3, 7

Sorted → 1, 3, 5, 7, 16, 17, 20, 21

Median = (7+16)/2 = **11.5**

👉 **Interpretation**: Half of values are below, half above.

💡 **Strength**: Not affected by outliers.

3️⃣ Mode (Most Frequent)

📌 **Definition**

The value that occurs most often.

📖 **Example**

Data = 10, 15, 20, 20, 25, 30, 20

Mode = **20** (appears 3 times).

👉 **Best for** categorical data:

- "Most common blood group?"
- "Most sold pizza flavor?"

⚖️ Mean vs Median (Impact of Outliers)

📖 **Case 1 – Balanced Salaries**

50K, 75K, 1L, 2L

- Mean = 1.06L
- Median = 87.5K

  👉 Both give fair idea.

📖 **Case 2 – With Outlier**

0.5 Paise, 50K, 1L, 1000 Cr

- Mean = ~200 Cr 🤯
- Median = 1L ✅

  👉 Median is more reliable when extreme values exist.

🔑 Summary Table

Measure	Formula/Logic	Good For	Weakness
Mean	Sum ÷ Count	Balanced, normal data	Affected by outliers
Median	Middle value	Skewed data, income, property prices	Ignores exact magnitudes
Mode	Most frequent	Categories, popularity	Can be multiple / not exist

---

💡 Extra Insights (for deeper understanding)

1. **Multiple Modes**:
   - 1 peak → **Unimodal**
   - 2 peaks → **Bimodal**
   - Many peaks → **Multimodal**
2. **Link with Shape**:
   - In **normal distribution** → Mean = Median = Mode.
   - In **skewed data** → they spread apart (we'll explore in Day 4).
3. **Why we need all three**:
   - **Mean** tells "mathematical average."
   - **Median** tells "middle typical person."
   - **Mode** tells "most common occurrence."

     👉 Together, they give a **full picture** of the data.

🏋️ Practice Problems

1. Data: {5, 7, 8, 10, 10, 15, 20}
   - Find Mean, Median, Mode.
2. Salaries (in ₹000): {20, 22, 25, 28, 30, 90}
   - Compute Mean & Median. Which represents the data better?
3. Survey results: {Poor, Good, Excellent, Good, Fair, Good, Excellent}
   - Which central tendency measure is appropriate?

📘 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

$Reciprocal: x \rightarrow 1xx \rightarrow \frac{1}{x}$

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

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

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

- 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

📘 Day 5 – Deviation & Measures of Spread

1️⃣ Deviation – Why does total deviation = 0?

👉 **Deviation** = how far a data point is from the mean.

Formula for each value:

$$di = xi - x^{-}d_i = x_i - \bar{x}$$

- If xix_i > mean → deviation is **positive**.
- If xix_i < mean → deviation is **negative**.

Example: Data = {1, 2, 3, 4, 5}, Mean = 3

Deviations = (-2, -1, 0, +1, +2).

Sum = -2 -1 + 0 + 1 + 2 = **0** ✅

💡 **Reason:** The mean is the balancing point of data, so deviations cancel.

👉 That's why we **don't use plain deviation** as a measure of spread.

2️⃣ Absolute Mean Deviation (AMD or MAD)

👉 To avoid negative + positive canceling, we take **absolute value**:

$$MAD = 1n\sum \mid xi - x^{-} \mid MAD = \frac{1}{n}\sum |x_i - \bar{x}|$$

Example: Data = {1, 2, 3, 4, 5}, Mean = 3

- |1-3| = 2

- |2-3| = 1
- |3-3| = 0
- |4-3| = 1
- |5-3| = 2

Sum = 6 → MAD = 6/5 = **1.2**

👉 Interpretation: On average, data points are **1.2 units away from mean**.

⚠️ Math note: Absolute value is hard in calculus (nondifferentiable at 0), so we prefer squaring → leads us to variance.

3️⃣ Variance

👉 To avoid cancellation and make math smooth → **square deviations**.

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

(for population variance).

For sample variance (statistical estimation):

$s2 = 1n - 1 \sum (xi - x^-)2s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$

Example: Data = {1, 2, 3, 4, 5}, Mean = 3

Deviations = (-2, -1, 0, 1, 2)

Squares = (4, 1, 0, 1, 4)

Sum = 10

- Population variance = 10/5 = **2**
- Sample variance = 10/4 = **2.5**

👉 Variance measures **average squared spread** of data.

⚠️ But it's in **squared units** (marks², km², rupees²). Hard to interpret directly.

4️⃣ Standard Deviation (SD)

👉 To bring back original units → take square root of variance.

$$\sigma = \sigma 2, s = s2\sigma = \sqrt{\sigma^2}, \quad s = \sqrt{s^2}$$

From above example:

- Population SD = √2 ≈ **1.41**
- Sample SD = √2.5 ≈ **1.58**

👉 Interpretation: On average, data points are ~1.4 units away from mean.

📌 SD is the **most widely used measure of spread**.

5️⃣ Scaling Effect (Important Concept)

If you multiply data by a constant cc:

- Mean → ×c
- Variance → ×c2c^2
- SD → ×c

Example:

Data = {2, 3, 4, 5, 6} km

Multiply by 10 → {20, 30, 40, 50, 60} rupees

- Variance (km) = 2
- Variance (rs) = 200 = 100 × bigger
- SD (km) = 1.41
- SD (rs) = 14.1 = 10 × bigger

👉 This is why variance has **squared units**.

6️⃣ Coefficient of Variation (CV)

👉 To compare spread across different units/scales, use **CV** (unitless).

$$CV = SDMean \times 100CV = \frac{SD}{Mean} \times 100$$

Example:

- For km → CV ≈ 35.35%
- For rupees → CV ≈ 35.35% (same, scale doesn't matter).

👉 CV helps compare consistency across different measurements.

7️⃣ Practical Summary

- **Deviation**: cancels to 0 (not useful).
- **MAD**: uses absolute deviation, easy to understand but less used in advanced math.
- **Variance**: squared average deviation, math-friendly but hard units.
- **SD**: square root of variance, best real-world measure of spread.
- **CV**: compares spread across different scales, unitless.

📝 Practice Problems

1. Data = {2, 4, 6, 8, 10}
   - Find Mean, MAD, Variance, SD.
2. A company's monthly profits (in lakh ₹): {20, 22, 25, 23, 100}
   - Find mean and SD. Which is more reliable, mean or median, and why?
3. Two batsmen scored:
   - Player A: {40, 50, 60, 70, 80}
   - Player B: {10, 30, 50, 70, 90}
   - Who is more consistent? (Use SD).

📘 Day 6 – Variance, Covariance & Correlation

1️⃣ Variance – Spread of a Single Variable

👉 **Definition:** How much a variable varies around its mean.

$$Var(X) = 1N \sum (xi - x^-)2 Var(X) = \frac{1}{N} \sum (x_i - \bar{x})^2$$

Example: Ages = {20, 21, 22, 23, 60}

- Mean ≈ 29.2
- Variance is large because of the outlier (60).

💡 **Intuition:** If values are close to mean → low variance.

If values are spread out → high variance.

2️⃣ Covariance – Relationship Between Two Variables

👉 **Definition:** How two variables vary together.

$$Cov(X,Y) = 1N \sum (xi - x^-)(yi - y^-) Cov(X,Y) = \frac{1}{N} \sum (x_i - \bar{x})(y_i - \bar{y})$$

Interpretation:

- **Positive covariance** → X ↑, Y ↑ (move together).
- **Negative covariance** → X ↑, Y ↓ (move opposite).
- **Zero covariance** → no relationship.

📍 Example:

- Age & Income → usually **positive covariance**.
- Exercise hours & Weight → usually **negative covariance**.
- Shoe size & Exam marks → **zero covariance**.

3️⃣ Covariance Matrix

👉 For multiple variables, we arrange variances & covariances into a matrix.

Example: Variables = Age, Salary

	Age	Salary
Age	Var(Age)	Cov(Age, Salary)
Salary	Cov(Salary,Age)	Var(Salary)

💡 Diagonal = variances, Off-diagonal = covariances.

💡 Always symmetric: Cov(X,Y) = Cov(Y,X).

4️⃣ Scatter Plot (Visual Tool)

A **scatter plot** helps see relationships:

- **Positive slope** → positive relation.
- **Negative slope** → negative relation.
- **Cloudy / random** → no relation.

👉 Example: Age vs Salary plotted = upward sloping scatter.

5️⃣ Correlation Coefficient (r)

👉 Problem: Covariance values are unbounded (can be −∞ to +∞).

👉 Solution: Normalize covariance → correlation.

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

- Always between −1 and +1.

Interpretation:

- **r = +1** → perfect positive relation.
- **r = −1** → perfect negative relation.
- **r = 0** → no relation.
- **|r| close to 1** = strong relation, **|r| close to 0** = weak relation.

📍 Example:

- Age vs Income, r = 0.8 → strong positive.
- Age vs Income, r = −0.5 → moderate negative.
- Age vs Income, r = 0.05 → almost no relation.

✅ Quick Summary

- **Variance** → Spread of one variable.
- **Covariance** → Direction of relationship (positive/negative/none).
- **Covariance Matrix** → Table of variance + covariance for multiple variables.
- **Scatter Plot** → Visualize relation.
- **Correlation (r)** → Strength & direction of relationship (−1 ≤ r ≤ +1).

📝 Practice Problems

1. For dataset:

   X = {2, 4, 6}, Y = {1, 2, 3}

   - Find covariance. Is it positive or negative?
2. Suppose the correlation between **study hours & marks** is r = 0.9.
   - Interpret this result in plain words.
3. Which pair likely has:
   - Positive correlation?
   - Negative correlation?
   - Near-zero correlation?

     👉 (a) Height & Weight

     👉 (b) Hours of Sleep & Stress

     👉 (c) Shoe size & Salary

📘 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.

   $z = x - x^{-}\sigma z = \frac{x-\bar{x}}{\sigma}$
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/√ 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

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

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

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

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

- 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?

📘 Day 8 – Outliers & Standardization

🔷 1. Outliers & IQR Rule

We already learned outliers are **extreme values**.

Day 8 formalizes the method to detect them using **quartiles (Q1, Q3)**.

- **Q1** = 25th percentile
- **Q2 (Median)** = 50th percentile
- **Q3** = 75th percentile
- **IQR (Interquartile Range)** = Q3 − Q1

👉 Outlier thresholds:

- **Mild outliers:**

$$Lowerbound = Q1 - 1.5 \times IQR, Upperbound = Q3 + 1.5 \times IQR \text{Lower bound} = Q1 - 1.5 \times IQR, \quad \text{Upper b}$$

- **Strong outliers:**

$$Lowerbound = Q1 - 3 \times IQR, Upperbound = Q3 + 3 \times IQR \text{Lower bound} = Q1 - 3 \times IQR, \quad \text{Upper bound}$$

Example:

- Q1 = 10k, Q2 = 1 lakh, Q3 = 5 lakh
- IQR = 5L − 10k = 4.9L
- Upper bound = 5L + 1.5×4.9L ≈ 12.35L

👉 Anyone earning **≥12.35L per month** = outlier.

📌 In **Python boxplots**, the default threshold = **1.5×IQR** (mild outliers).

🔷 2. How to Deal with Outliers

Outliers impact **mean** but not **median**.

Options:

1. **Drop them (not recommended)**
   - If only 1–2% of data are outliers → dropping might be fine.
   - BUT: you also lose related info (other columns).
2. **Replace with Median**
   - Since median is robust, we can replace outliers with the **50th percentile value**.
3. **Cap with Q1 & Q3 (Winsorization)**
   - If value > Q3 → set = Q3.
   - If value < Q1 → set = Q1.
   - This reduces outlier effect but keeps all rows.

👉 Example:

If income = 100L, Q3 = 5L → cap it at 5L.

🔷 3. Why Scale Data?

In datasets, features have different ranges.

- Age ≈ 20–80
- Income ≈ 10k – 1 crore

⚠️ Problem:

Some ML models use **distance-based calculations** (like k-NN, SVM, clustering).

If features are not scaled → large-value features dominate.

Example:

Distance between (20, 50,000) and (30, 20,000):

$(30 - 20)2 + (20000 - 50000)2\sqrt{(30 - 20)^2 + (20000 - 50000)^2}$

→ Income dominates, age ignored.

🔷 4. Standardization (Z-Score Scaling)

👉 Formula:

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

- Mean becomes 0
- Standard deviation becomes 1
- Result = **standard normal distribution**

Properties:

- After standardization:

  $\mu = 0, \sigma = 1\mu = 0, \quad \sigma = 1$

- Variance = 1
- Works well when data is normally distributed.

👉 Example:

If student marks = 70, mean = 60, SD = 5:

$z = 70 - 605 = 2z = \frac{70-60}{5} = 2$

= "score is 2 SDs above average".

🔷 5. Normalization (Min–Max Scaling)

👉 Formula:

$$x\prime = x - min(x)max(x) - min(x)x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Scales values to [0,1] range.
- Useful when we want all features between 0–1 (e.g., neural networks).

Example:

Data = {20, 30, 50}, min = 20, max = 50

- For 30:

$$30 - 2050 - 20 = 1030 = 0.33\frac{30 - 20}{50 - 20} = \frac{10}{30} = 0.33$$

🔷 6. Empirical Rule (for normal data)

- 68% of data → within 1 SD of mean
- 95% → within 2 SD
- 99.7% → within 3 SD

👉 Outliers beyond ±3 SD = strong outliers.

✅ Summary

- **Outliers**: detected using IQR rule (1.5×IQR = mild, 3×IQR = strong).
- **Handling outliers**: Drop, Replace with median, or Winsorize.
- **Scaling**: ensures features are comparable in models.
- **Standardization**: (z-score) → mean=0, SD=1.
- **Normalization**: (min-max) → scale between 0–1.
- **Empirical Rule**: 68-95-99.7 rule for normal distribution.

📝 Practice Problems

1. Dataset incomes: {10k, 20k, 30k, 40k, 12L}.
   - Find Q1, Q3, IQR.
   - Detect if 12L is an outlier (mild or strong).
2. Marks: {20, 30, 40, 50, 60}, mean = 40, SD = 10.

- Standardize each mark (z-score).

3. Heights: {150, 160, 170, 180}, min=150, max=180.
   - Normalize values into [0,1].

📘 Day 8 – Outliers & Standardization

🔷 1. Outliers & IQR Rule

We already learned outliers are **extreme values**.

Day 8 formalizes the method to detect them using **quartiles (Q1, Q3)**.

- **Q1** = 25th percentile
- **Q2 (Median)** = 50th percentile
- **Q3** = 75th percentile
- **IQR (Interquartile Range)** = Q3 − Q1

👉 Outlier thresholds:

- **Mild outliers:**

$$Lower\ bound = Q1 - 1.5 \times IQR, Upper\ bound = Q3 + 1.5 \times IQR\text{Lower bound} = Q1 - 1.5 \times IQR, \quad \text{Upper}$$

- **Strong outliers:**

$$Lower\ bound = Q1 - 3 \times IQR, Upper\ bound = Q3 + 3 \times IQR\text{Lower bound} = Q1 - 3 \times IQR, \quad \text{Upper boun}$$

Example:

- Q1 = 10k, Q2 = 1 lakh, Q3 = 5 lakh
- IQR = 5L − 10k = 4.9L
- Upper bound = 5L + 1.5×4.9L ≈ 12.35L

👉 Anyone earning **≥12.35L per month** = outlier.

📌 In **Python boxplots**, the default threshold = **1.5×IQR** (mild outliers).

🔷 2. How to Deal with Outliers

Outliers impact **mean** but not **median**.

Options:

1. **Drop them (not recommended)**
   - If only 1–2% of data are outliers → dropping might be fine.
   - BUT: you also lose related info (other columns).
2. **Replace with Median**
   - Since median is robust, we can replace outliers with the **50th percentile value**.
3. **Cap with Q1 & Q3 (Winsorization)**
   - If value > Q3 → set = Q3.
   - If value < Q1 → set = Q1.
   - This reduces outlier effect but keeps all rows.

👉 Example:

If income = 100L, Q3 = 5L → cap it at 5L.

🔷 3. Why Scale Data?

In datasets, features have different ranges.

- Age ≈ 20–80
- Income ≈ 10k – 1 crore

⚠️ Problem:

Some ML models use **distance-based calculations** (like k-NN, SVM, clustering).

If features are not scaled → large-value features dominate.

Example:

Distance between (20, 50,000) and (30, 20,000):

$(30 - 20)2 + (20000 - 50000)2\sqrt{(30 - 20)^2 + (20000 - 50000)^2}$

→ Income dominates, age ignored.

🔷 4. Standardization (Z-Score Scaling)

👉 Formula:

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

- Mean becomes 0
- Standard deviation becomes 1
- Result = **standard normal distribution**

Properties:

- After standardization:

  $\mu = 0, \sigma = 1\mu = 0, \quad \sigma = 1$
- Variance = 1
- Works well when data is normally distributed.

👉 Example:

If student marks = 70, mean = 60, SD = 5:

$z = 70 - 605 = 2z = \frac{70-60}{5} = 2$

= "score is 2 SDs above average".

🔷 5. Normalization (Min–Max Scaling)

👉 Formula:

$$x\prime = x - min(x)max(x) - min(x)x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Scales values to [0,1] range.
- Useful when we want all features between 0–1 (e.g., neural networks).

Example:

Data = {20, 30, 50}, min = 20, max = 50

- For 30:

$30 - 2050 - 20 = 1030 = 0.33 \frac{30-20}{50-20} = \frac{10}{30} = 0.33$

🔷 6. Empirical Rule (for normal data)

- 68% of data → within 1 SD of mean
- 95% → within 2 SD
- 99.7% → within 3 SD

👉 Outliers beyond ±3 SD = strong outliers.

✅ Summary

- **Outliers**: detected using IQR rule (1.5×IQR = mild, 3×IQR = strong).
- **Handling outliers**: Drop, Replace with median, or Winsorize.
- **Scaling**: ensures features are comparable in models.
- **Standardization**: (z-score) → mean=0, SD=1.
- **Normalization**: (min-max) → scale between 0–1.
- **Empirical Rule**: 68-95-99.7 rule for normal distribution.

📝 Practice Problems

1. Dataset incomes: {10k, 20k, 30k, 40k, 12L}.
   - Find Q1, Q3, IQR.
   - Detect if 12L is an outlier (mild or strong).
2. Marks: {20, 30, 40, 50, 60}, mean = 40, SD = 10.
   - Standardize each mark (z-score).
3. Heights: {150, 160, 170, 180}, min=150, max=180.
   - Normalize values into [0,1].

📘 Day 9 – Hypothesis Testing + Probability Foundations

🔷 1. Probability Basics

Probability = likelihood of an event happening.

- Formula:

$$P(E)) = \frac{\text{Number of favorable outcomes}}{\text{Total possible outcomes}}$$

📍 Example: Toss a coin

- Sample space S = {H, T}
- P(H) = 1/2 = 0.5

📍 Example: Roll a die

- S = {1,2,3,4,5,6}
- P(odd) = 3/6 = 0.5

🔷 2. Types of Events

- **Independent**: One event does not affect the other.

  Example: Tossing 2 coins.

  P(H on 1st AND H on 2nd) = 0.5 × 0.5 = 0.25
- **Mutually exclusive**: Both cannot happen at the same time.

  Example: Drawing a King and Queen at the same time from 1 card = 0.

🔷 3. Conditional Probability

$$P(A \mid B) = P(A \cap B)P(B)P(A|B) = \frac{P(A \cap B)}{P(B)}$$

📍 Example: Deck of 52 cards

- P(King | Face card) = King face cards / total face cards = 4/12 = 1/3

🔷 4. Law of Total Probability

If events B1,B2,...BnB_1, B_2, ... B_n partition the sample space:

$$P(A) = \sum i = 1nP(A \mid Bi) \cdot P(Bi)P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)$$

📍 Example:

Factory has 3 machines: A(40%), B(35%), C(25%).

Defective probability: A(2%), B(3%), C(4%).

Overall defect rate = (0.4×0.02) + (0.35×0.03) + (0.25×0.04) = **0.0295 = 2.95%**

🔷 5. Bayes' Theorem

$$P(A \mid B) = P(B \mid A)P(A)P(B)P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

📍 Example: Medical Testing

- Disease prevalence = 1%
- Test detects correctly = 99%
- False positive = 5%

If test is positive, what is P(person has disease)?

$$P(Disease \mid Positive) = 0.99 \times 0.010.99 \times 0.01 + 0.05 \times 0.99 \approx 0.167P(Disease|Positive) = \frac{0.99 \times 0.0}{0.99 \times 0.01 + 0.0}$$

👉 Even if test says "Positive", probability is only **16.7%** (because disease is rare).

🔷 6. Hypothesis Testing

- Hypothesis = claim about a population.
- **Null Hypothesis (H_o)** → No effect, no difference.
- **Alternative Hypothesis (H_1)** → There is effect, difference.

📍 Example: Average student height = 170 cm.

- H_o: μ = 170
- H_1: μ ≠ 170

🔷 7. Errors in Testing

- **Type I error (α)** → Rejecting H_o when true (false alarm).
- **Type II error (β)** → Not rejecting H_o when false (missed detection).

👉 α = significance level (usually 0.05 = 5%).

🔷 8. Test Statistics

- **Z-test** → population variance known, large n (>30).

$$Z = x^- - \mu\sigma/nZ = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- **t-test** → population variance unknown, small n (<30).

$$t = x^- - \mu s/nt = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

📍 Example: Company claims avg. salary = ₹50,000.

Sample of 25 employees → mean = ₹48,000, s = ₹4,000.

$$t = 48000 - 500004000/25 = -2000800 = -2.5t = \frac{48000 - 50000}{4000/\sqrt{25}} = \frac{-2000}{800} = -2.5$$

Compare with t-table (df=24). If |t| > critical value → reject H_o.

🔷 9. Confidence Intervals

CI gives range of plausible values for mean.

$$CI = x^- \pm Z\alpha/2 \cdot \sigma nCI = \bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

📍 Example: Mean = 100, σ=15, n=36, 95% CI.

$$CI = 100 \pm 1.96 \cdot 156 = 100 \pm 4.9CI = 100 \pm 1.96 \cdot \frac{15}{6} = 100 \pm 4.9$$

CI = (95.1, 104.9)

✅ Summary of Day 9

- Probability basics → events, independence, conditional, total probability, Bayes.
- Hypothesis testing → Null vs Alternative, errors, p-values.
- Z-test, t-test.

- Confidence Intervals.

📘 Day 10 – Distributions + Advanced Hypothesis Testing

🔷 1. Probability Distributions

A **distribution** describes how data values are spread.

Two main types:

- **Discrete** → specific values (0,1,2,...)
- **Continuous** → any value in a range

1.1 Discrete Distributions

- **Bernoulli** → Single trial, success (1) or failure (0).

  $$P(X = 1) = p,\ P(X = 0) = 1 - pP(X = 1) = p,\ P(X = 0) = 1 - p$$

  Example: Coin toss (H=1, T=0).

- **Binomial** → Repeated Bernoulli trials (n trials).

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

  Example: Probability of 4 heads in 10 tosses.

- **Poisson** → Number of events in fixed interval (rare events).

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

  Example: Calls at a call center per hour.

- **Geometric** → Number of trials until first success.
- **Negative Binomial** → Trials until r successes.
- **Uniform (discrete)** → All outcomes equally likely.

1.2 Continuous Distributions

- **Normal (Gaussian)**

  Bell curve, symmetric, mean=median=mode.

  $$f(x) = 1\sigma 2\pi e - (x - \mu)22\sigma 2f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  Example: Heights, test scores.

- **Exponential**

  Time between events (waiting time).

  $$f(x) = \lambda e - \lambda x,\ x \geq 0f(x) = \lambda e^{-\lambda x},\ x \geq 0$$

  Example: Time between earthquakes.

- **Uniform (continuous)**

Equal probability in a range [a, b].

- **t-distribution**

  Used for small sample hypothesis testing. Tails fatter than normal.

- **χ^2 distribution**

  Sum of squared normal variables. Used in Chi-square test.

- **F-distribution**

  Ratio of two variances. Used in ANOVA.

1.3 Central Limit Theorem (CLT)

No matter population distribution, as sample size grows (n > 30):

- Sampling distribution of mean → **Normal**.
- Mean = μ, SD = σ/√n.

👉 This is why Z/t tests work!

🔷 2. Advanced Hypothesis Testing

- **Chi-Square Test (χ^2)**

  Tests association between categorical variables.

  $\chi 2 = \sum (O - E) 2E \chi^2 = \sum \frac{(O-E)^2}{E}$

  O = observed, E = expected.

📍 Example: Is gender independent of voting preference?

- **ANOVA (Analysis of Variance)**

  Compares means of 3+ groups.

  - H_o: All group means equal.
  - H_1: At least one mean differs.

    Test statistic → F-distribution.

📍 Example: Do 3 diets give same average weight loss?

- **Non-parametric tests** (no normality assumption):
  - Mann-Whitney U test (2 groups).
  - Kruskal-Wallis test (3+ groups).
  - Wilcoxon signed-rank test (paired data).

🔷 3. Correlation (Beyond Pearson)

- **Pearson's r** → linear correlation.
- **Spearman's rank correlation (ρ)** → based on rank (good for monotonic relationships).
- **Kendall's Tau (τ)** → rank concordance measure.

🔷 4. Regression Assumptions

When using Linear Regression:

1. Linearity → relation is linear.

2. Independence → errors not correlated.
3. Homoscedasticity → equal variance of errors.
4. Normality of errors.
5. No multicollinearity (independent variables not highly correlated).

📍 If assumptions fail → use transformations, regularization, or non-parametric models.

🔷 5. Effect Size & Power Analysis

- **Effect size** → strength of a relationship.
  - Cohen's d (difference between two means in SD units).
  - η^2 (eta squared) in ANOVA.
- **Statistical Power** → probability of detecting a true effect.

  $Power = 1 - \beta Power = 1 - \beta$

  Higher power → lower risk of Type II error.

  Desired power = 0.8 (80%).

✅ Summary of Day 10

- Distributions → discrete (Bernoulli, Binomial, Poisson) & continuous (Normal, Exponential, t, χ^2, F).
- Central Limit Theorem.
- Hypothesis testing advanced → Chi-Square, ANOVA, Non-parametric.
- Correlation extensions → Spearman, Kendall.
- Regression assumptions.
- Effect size & power analysis.
