📘 Day 6 – Variance, Covariance & Correlation

1️⃣ Variance – Spread of a Single Variable

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

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

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

Do you want me to **solve these practice problems step by step** right now, or prepare the **Day 7 Deep Dive (Probability Basics)** next?