What is Variance?

Variance is a measure of how much the data points **differ from the average (mean)** value.

> It shows how **spread out** the values are in a dataset.

Example Student Marks

Let's compare marks of two classes:

Class A (Low Variance)

$$\text{Variance} = \frac{1}{N}\sum(x_i - \mu)^2$$

- x_i : each data value
- μ: mean
- N : number of data points

It's the **average of the squared differences** from the mean.

Student	Marks
A1	69
A2	70
A3	71

- Mean = (69 + 70 + 71) / 3 = **70**
- All marks are **close to the mean**
- **Low Variance**

Class B (High Variance)

Student	Marks
B1	50
B2	70
B3	90

- Mean = (50 + 70 + 90) / 3 = **70**
- Marks are **spread far from the mean**
- **High Variance**

What is Standard Deviation?

Standard Deviation (SD) is the **square root of variance**.

It tells you **how much the data varies from the mean** in the **original units** (like marks, rupees, etc.).

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

Example: score={10,20,30,40,50}

◆ Step 1: Calculate the Mean (μ)

$$\text{Mean} = \frac{10+20+30+40+50}{5} = \frac{150}{5} = 30$$

✅ So, the average score is **30.**

◆ Step 2: Find Deviations from the Mean

Score (X)	Deviation (X − μ)	Squared Deviation ((X - μ)^2)
10	10 − 30 = -20	$(-20)^2$ = 400
20	20 − 30 = -10	$(-10)^2$ = 100
30	30 − 30 = 0	0^2 = 0
40	40 − 30 = 10	10^2 = 100
50	50 − 30 = 20	20^2 = 400

- **Mean** is the average score: 30
- **Variance** shows the average of the squared differences from the mean: 200
- **Standard Deviation** is the square root of variance: ≈ 14.14
- A **low standard deviation** means scores are tightly clustered around the mean.
- A **high standard deviation** means the scores are more spread out.

◆ Step 3: Calculate Variance (σ^2)

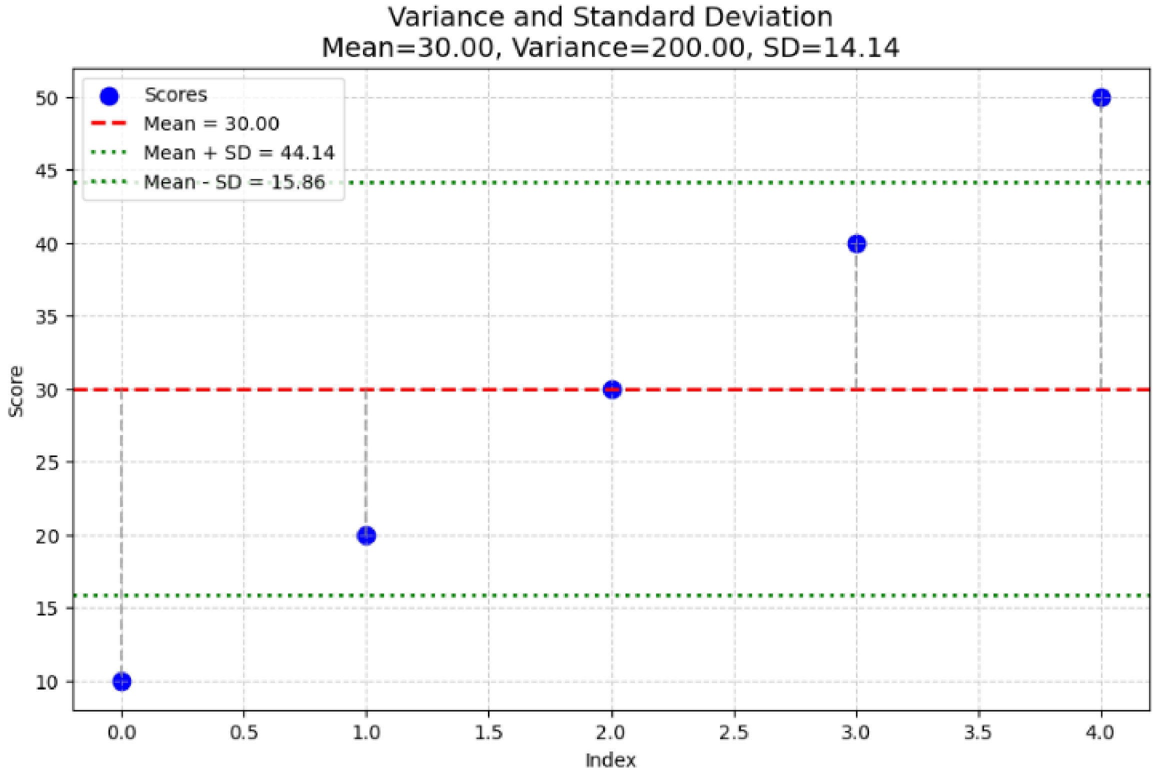
$$\text{Variance} = \frac{\sum(X-\mu)^2}{N} = \frac{400+100+0+100+400}{5} = \frac{1000}{5} = 200$$

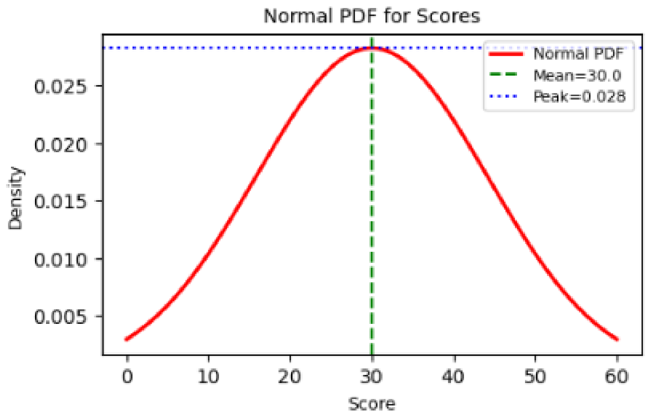
📌 Variance tells us how **spread out** the data is (in squared units).

◆ Step 4: Calculate Standard Deviation (σ)

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{200} \approx 14.14$$

📌 This means, on average, each score is about **14.14 points away** from the mean.


Variance and Standard Deviation
Mean=30.00, Variance=200.00, SD=14.14
Scores
Mean = 30.00
Mean + SD = 44.14
Mean - SD = 15.86
Score
Index
50
45
40
35
30
25
20
15
10
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Normal PDF for Scores
Normal PDF
Mean=30.0
Peak=0.028
Density
Score
0.025
0.020
0.015
0.010
0.005
0
10
20
30
40
50
60

Percentile

- A percentile is a number that tells us what percentage of the data lies below that value.
- It helps us understand where a particular data point stands in the dataset.

Dataset:[12, 18, 25, 26, 30, 34, 40, 45, 50]

Number of elements, **n** = **9**

$$position = (n - 1) * q$$

Where:

- `n` is the number of elements
- `q` is the desired quantile (e.g., 0.25, 0.5, 0.75)

Step-by-Step Example

Q1 (25th percentile):

- $Position = (9 - 1) * 0.25 = 2.0 \rightarrow index2$
- Value at index 2 = **25.0**

Q2 (50th percentile / Median):

- $Position = (9 - 1) * 0.50 = 4.0 \rightarrow index4$
- Value at index 4 = **30.0**

Q3 (75th percentile):

- $Position = (9 - 1) * 0.75 = 6.0 \rightarrow index6$
- Value at index 6 = **40.0**

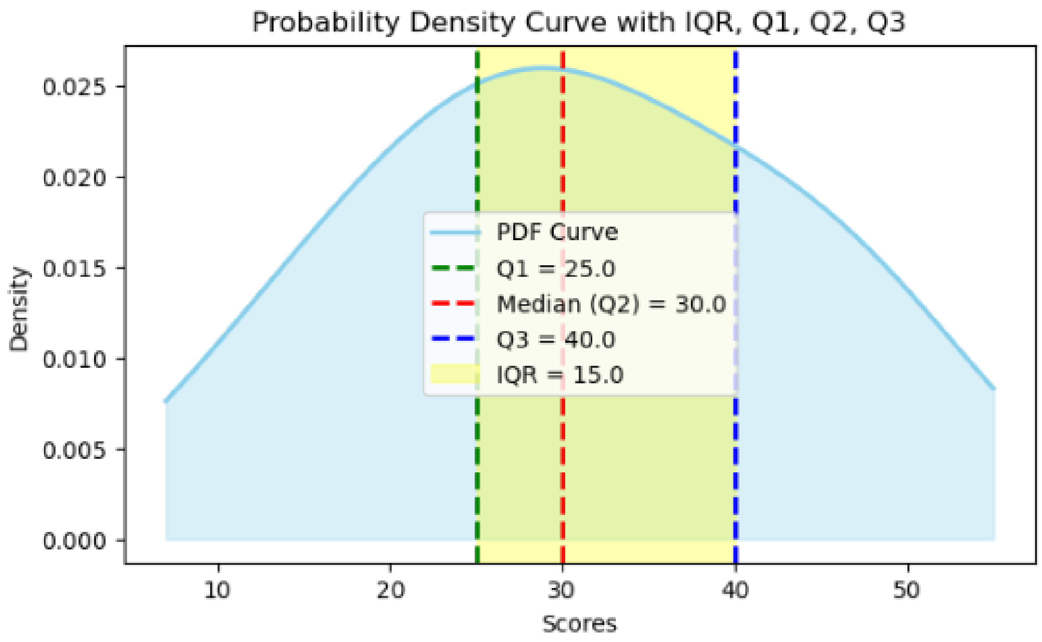
1

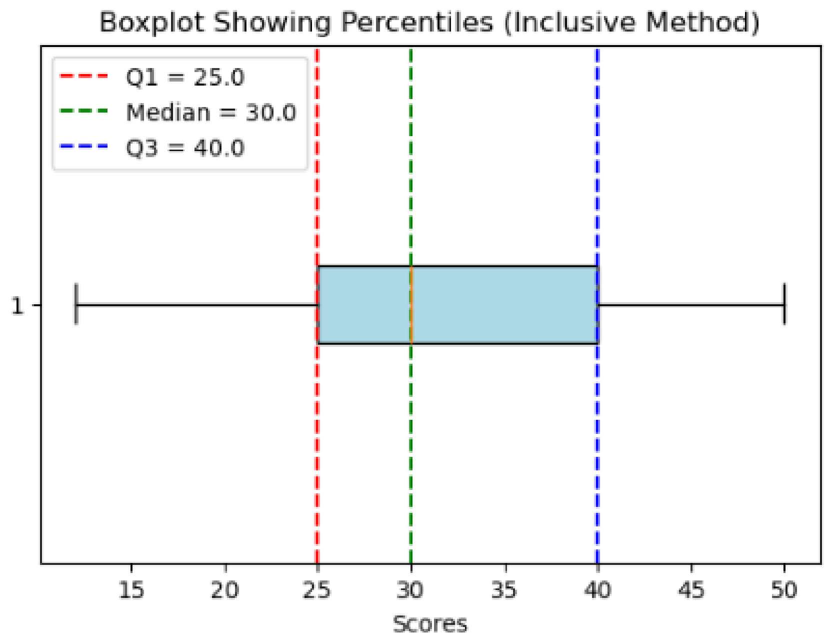
$$IQR = Q3 - Q1 = 40.0 - 25.0 = 15.0$$

$$Lower Bound = Q1 - 1.5 \times IQR = 25.0 - 1.5 \times 15.0 = 25.0 - 22.5 = **2.5**$$

$$Upper Bound = Q3 + 1.5 \times IQR = 40.0 + 1.5 \times 15.0 = 40.0 + 22.5 = **62.5**$$

Metric	Value
Q1	25.0
Q2 (Median)	30.0
Q3	40.0
IQR	15.0
Lower Bound	2.5
Upper Bound	62.5


Probability Density Curve with IQR, Q1, Q2, Q3
PDF Curve
Q1 = 25.0
Median (Q2) = 30.0
Q3 = 40.0
IQR = 15.0
Density
0.000
0.005
0.010
0.015
0.020
0.025
10
20
30
40
50
Scores


Boxplot Showing Percentiles (Inclusive Method)
Q1 = 25.0
Median = 30.0
Q3 = 40.0
1
15
20
25
30
35
40
45
50
Scores

Scaling

Scaling is the process of transforming features (variables) so they fit into a specific range or distribution.

1. **Standardization (Z-score Scaling)**

- **Definition:** Rescales data so it has mean = 0 and standard deviation = 1.
- **Formula:**

$$Zscore = \frac{X - \bar{x}}{s}$$

- **Example:**
  Data: [10, 20, 30]
  - Mean (x') = 20, Std (s) ≈ 8.16
  - Transformed: [-1.22, 0, +1.22]

Z-Score Standardization

- **Formula:** (X - μ) / σ
- **Range:** Not fixed (can be negative, >1, or < -1).
- Typically: most values lie between **-3 and +3** (for normal data).

🎓 Z-Score and Probability Examples

We'll explore how to calculate:

- Z-scores
- Probabilities from Z-scores
- Number of values above or below a certain score in a dataset

---

📘 Dataset Used:

Student	Score
A	60
B	70
C	80
D	90
E	100

Example 1: How many values are **below 80**?

Step 1: Z-score for X = 80

$$Z = \frac{X-\mu}{\sigma} = \frac{80-80}{14.14} = 0$$

📊 Step 2: Z-table Probability

$$P(Z =< 0) = 0.5$$

👉 This means **50%** of values are less than 80.

🧮 Step 3: Apply to Dataset

$$0.5 \times 5 = 2.5 \approx 2 \text{ or } 3\ values$$

🎓 Z-Score and Probability Examples

We'll explore how to calculate:

- Z-scores
- Probabilities from Z-scores
- Number of values above or below a certain score in a dataset

---

📘 Dataset Used:

Student	Score
A	60
B	70
C	80
D	90
E	100

✅ Example 2: How many values are **greater than 90**?

🔢 Step 1: Z-score for X = 90

$Z = \frac{90-80}{14.14} \approx 0.71$

📊 Step 2: Z-table Probability

$P(Z < 0.71) \approx 0.7611$

$P(Z > 0.71) = 1 - 0.7611 = 0.2389$

So **23.89%** of values are greater than 90.

Step 3: Apply to Dataset

$$0.2389 \times 5 = 1.1945 \approx 1 \text{ value}$$