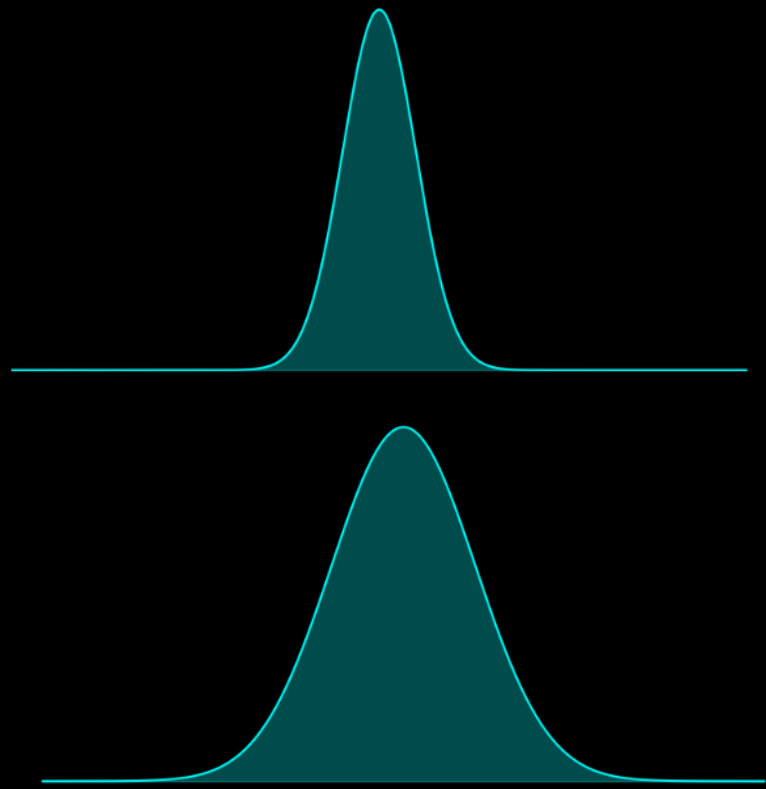
Variance and Standard deviation of population

Variance is a **measure of how much dispersed (or spread or scattered) your data is from its mean value**.

For a dataset $x_1, x_2, \ldots, x_n$ with mean μ:

$$\text{Variance } (\sigma^2) = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Variance is **difficult to interpret** because of squared units.



Standard Deviation σ is the square root of variance.

$$\text{Standard Deviation } (\sigma) = \sqrt{\text{Variance}}$$

Std. Dev. **easy to interpret** because it is in same unit as data points.

Example1: Monthly Income in a City. (Large spread around mean)

Income of population = 1, 1, 0, 0, 5, 9, 9, 10, 10, 5

Here size n = 10, and **mean** = (1 + 1 + 0 + 0 + 5 + 9 + 9 + 10 + 10 + 5)/ 10 = 5 .

$$\text{variance}(\sigma^2) = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$$

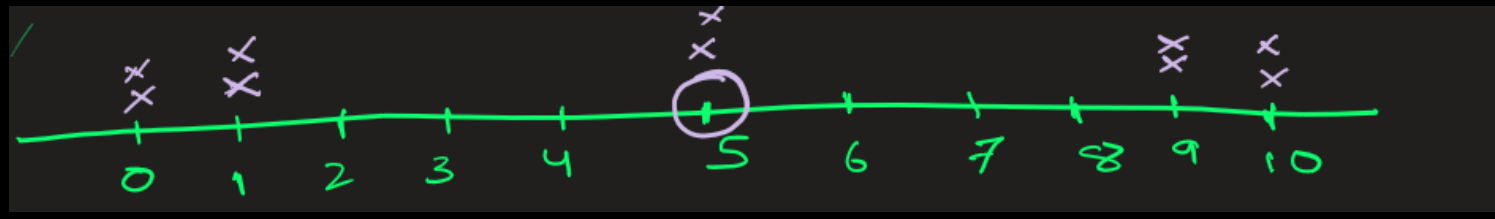
$$= \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_n-\mu)^2}{n}$$

$$= \frac{(1-5)^2 + (1-5)^2 + (0-5)^2 + (0-5)^2 + (5-5)^2 + (9-5)^2 + (9-5)^2 + (10-5)^2 + (10-5)^2 + (5-5)^2}{10}$$

$$= 16.8$$

Std. dev σ = $\sqrt{16.8}$ = 4.1

There is some dispersion / spread in data



Example2: Package Weights in a Production Line. (Few spread around mean)
Weight (Kg) of population = 4, 3, 5, 6, 7, 3, 5, 6, 4, 5

Here size n = 10, and **mean** = (4 + 3 + 5 + 6 + 7 + 3 + 5 + 6 + 4 + 5)/ 10 = 4.8

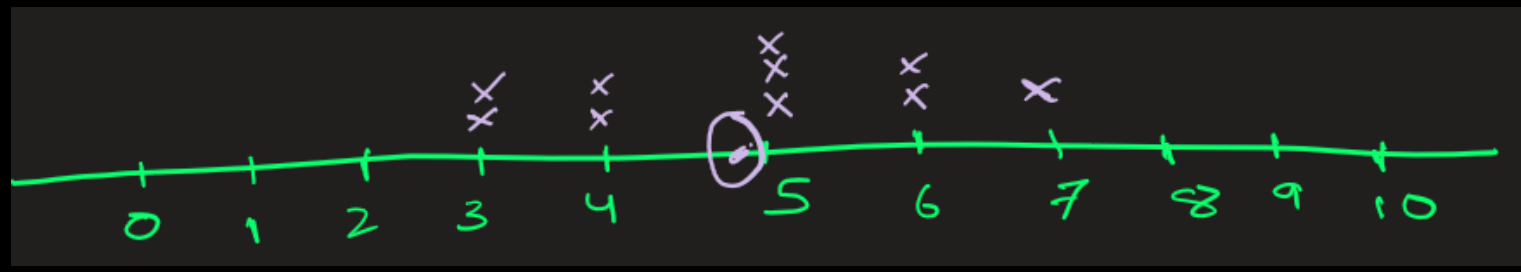
$$\text{variance}(\sigma^2) = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$$

$$= \frac{(4-4.8)^2 + (3-4.8)^2 + (5-4.8)^2 + \cdots + (5-4.8)^2}{10}$$

$$= 1.56$$

Std. dev σ = $\sqrt{1.56}$ = 1.3 **Kg**

Here data is less dispersed.
Variance in data went down



Example3: Time intervals (in seconds) between each operation in an assembly line (high precision)

Time interval of population i.e. all 10 machines = 5, 5, 5, 5, 5, 5, 5, 5, 5, 5

Here size n = 10, and **mean** = (5 + 5 + 5 + … + 5)/ 10 = 5

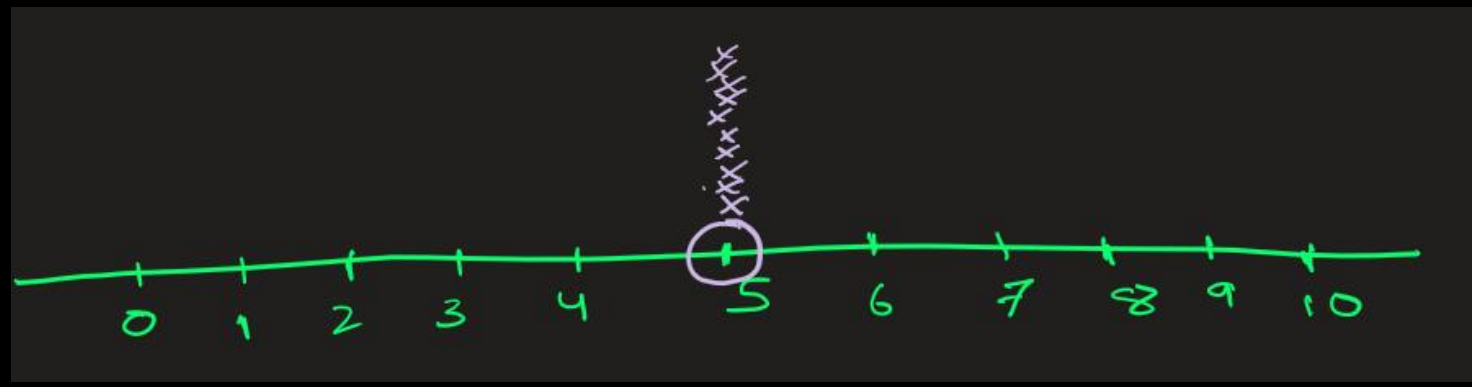
$$\text{variance}(\sigma^2) = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$$

$$= \frac{(5-5)^2 + (5-5)^2 + (5-5)^2 + \cdots + (5-5)^2}{10}$$

$$= 0$$

Std. dev σ = $\sqrt{0} = 0$

The data is not dispersed at all.
Zero variance in data.



Conclusion:

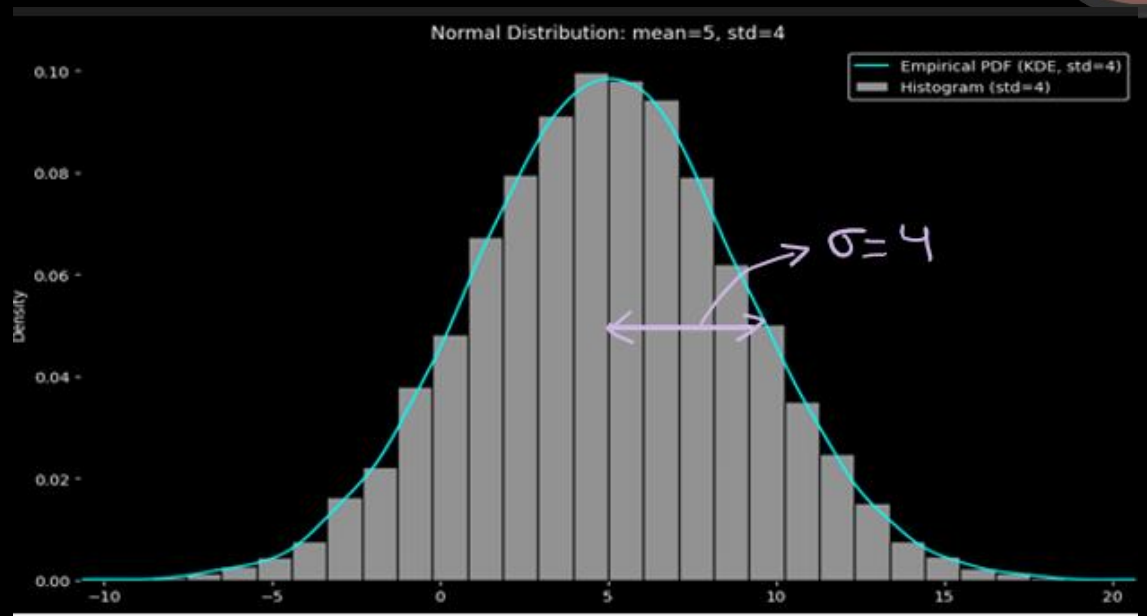
- If all data points are **close to the mean** -> Variance is **small**. Data is less dispersed or less spread
- If data points are **far from the mean** -> Variance is **large**. Data is more dispersed or more spread

How standard deviation controls dispersion/spread of distribution curve:

1) Following is 100 samples taken from population with 1050 data points with high std dev. (This could be temperature.)

6.987, 4.447, 7.591, 11.092, 4.063, 4.063, 11.317, 8.070, 3.122, 7.170, 3.146, 3.137, 5.968, -2.653, -1.900, 2.751, 0.949, 6.257, 1.368, -0.649, 10.863, 4.097, 5.270, -0.699, 2.822, 5.444, 0.396, 6.503, 2.597, 3.833, 2.593, 12.409, 4.946, 0.769, 8.290, 0.117, 5.835, -2.839, -0.313, 5.787, 7.954, 5.685, 4.537, 3.796, -0.914, 2.121, 3.157, 9.228, 6.374, -2.052, 6.296, 3.460, 2.292, 7.447, 9.124, 8.725, 1.643, 3.763, 6.325, 8.902, 3.083, 4.257, 0.575, 0.215, 8.250, 10.425, 4.712, 9.014, 6.447, 2.420, 6.446, 11.152, 4.857, 11.259, -5.479, 8.288, 5.348, 3.804, 5.367, -2.950, 4.121, 6.428, 10.912, 2.927, 1.766, 2.993, 8.662, 6.315, 2.881, 7.053, 5.388, 8.875, 2.192, 3.689, 3.432, -0.854, 6.184, 6.044, 5.020, 4.062

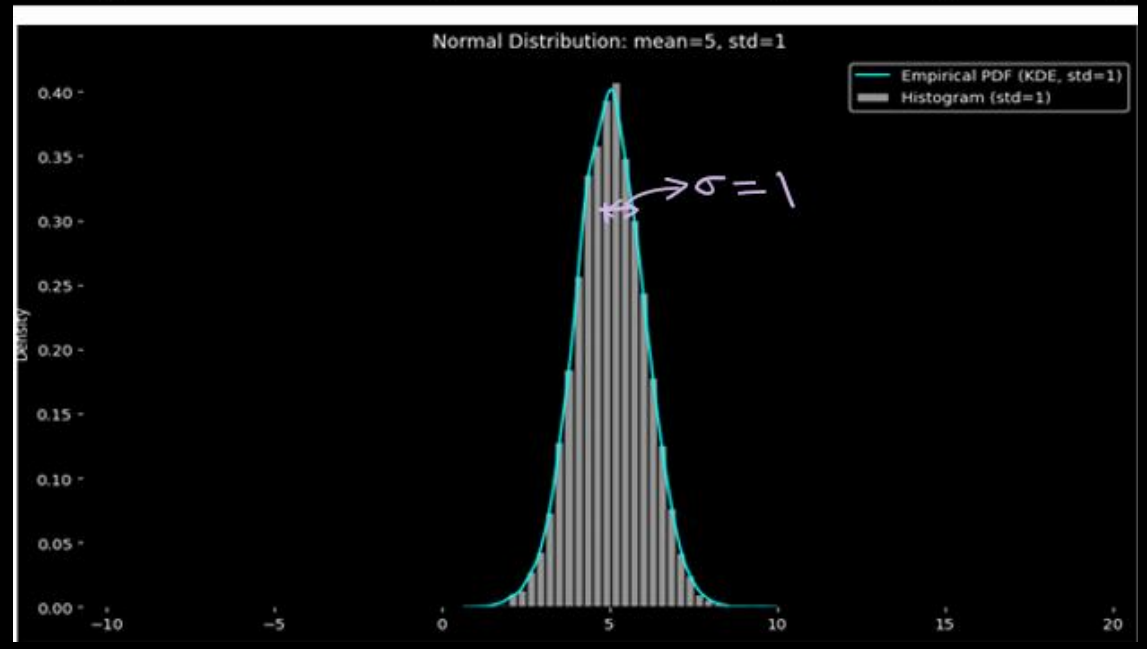
Sample mean: 5, Sample std dev: **4**



2) Following is 100 samples taken from population with 1050 data points with low std dev

4.322, 4.695, 4.403, 5.110, 6.197, 4.229, 6.001, 4.218, 4.152, 5.819, 5.922, 5.851, 3.684, 4.534, 5.823, 5.042, 3.926, 5.458, 4.285, 6.795, 6.545, 5.604, 6.361, 5.065, 5.765, 6.478, 5.245, 4.745, 3.295, 4.917, 5.823, 5.946, 5.504, 4.459, 3.023, 4.505, 4.696, 4.687, 5.619, 6.986, 5.124, 4.783, 4.741, 5.124, 4.172, 5.120, 5.451, 5.210, 5.458, 5.434, 3.228, 5.637, 4.329, 3.904, 3.896, 5.434, 4.777, 3.318, 5.478, 3.560, 5.140, 5.238, 5.886, 6.782, 3.635, 4.948, 4.724, 5.434, 5.216, 4.647, 4.559, 3.884, 5.988, 5.459, 5.918, 4.690, 4.343, 3.922, 5.359, 3.960, 6.849, 3.692, 5.274, 6.567, 5.599, 5.176, 4.032, 3.618, 5.770, 6.506, 4.561, 3.861, 2.542, 4.019, 4.886, 6.599, 5.673, 5.305, 4.600, 4.417

Sample mean: 5, Sample std dev: **1**



Why the 2 measure: variance and std. dev.?

Short Answer: Variance is for math; standard deviation is for humans!

The unit's matter

- Variance is the average squared deviation from the mean
- So, **if your data is in, say, cm, then variance is in square cm (cm^2).**
- **Squared units aren't directly interpretable: nobody says "the spread of your height is 100 cm^2".**
- By taking the square root, you bring the units back to the original scale

If the mean height is 170 cm and the standard deviation is 10 cm, you can immediately say: **"Most data is within 10 cm above or below the mean."**

That's intuitive. But "the variance is 100 cm^2" is not easily interpretable by our brain — squared differences are abstract

Population variance versus Sample variance

If you have a **population** of size N with values $x_1, x_2, \ldots, x_N$, the **population variance** is:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

x_i = each individual value

μ = population mean = $\frac{1}{N}\sum_{i=1}^{N} x_i$

N = size of the population

If you have a **sample** of size n drawn from a population, the **sample variance** is:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

x_i = each sample value

$\bar{x}$ = sample mean = $\frac{1}{n}\sum_{i=1}^{n} x_i$

n = size of the sample

$n-1$ is used instead of n (Bessel's correction) to get an **unbiased estimate** of the population variance

- For **population** variance, divide by n
- For **sample** variance, divide by n-1

For large n, the result is almost same

How to decide which variance to use: population or sample?

- If you are analyzing **all the data from the group** you're interested in, then use **population**.
- If you are analyzing a **subset (a sample) of the full group** and trying to generalize, then use **sample**.

Situation	Population or Sample?	Formula to Use
You recorded heights of **all 100 students** in a school	Population	Divide by n
You took a **random sample of 10 students** to estimate the average height in the school	Sample	Divide by n - 1
You analyzed **every** product made in a factory on one day	Population	Divide by n
You tested **20 random products** to estimate quality	Sample	Divide by n - 1

STOP