PW SKILLS

① Measure of Central Tendency

① Mean
② Median
③ Mode

[EDA and Feature Engineering]

① Mean

Population (N)

Sample (n)

$X : \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\} \rightarrow$

$n = 10$

Population mean $(\mu) = \frac{\sum_{i=1}^{N} x_i}{N}$

Sample mean $(s) = \frac{\sum_{i=1}^{n} x_i}{n}$

$$\mu = \frac{1+1+2+2+3+3+4+5+5+6}{10}$$

$$= \frac{32}{10} = 3.2$$

② Median

4, 5, 2, 3, 2, 1

Sort ⟶ 1, 2, 2, 3, 4, 5

Median

Even count

1, 2, [2, 3] 4, 5

Odd count

{1, 2, 2, [3] 4, 5, 7}

⇓

$\frac{2+3}{2} = 2.5$

Median = 3

Median = 2.5

Why Median ?

$\{1, 2, 3, 4, 5\}$

$\{1, 2, \boxed{3}, 4, 5\}$

$S = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3\}$

Median = 3

$\{1, 2, 3, 4, 5, \boxed{100}\}$

$\{1, 2, 3, 4, 5, \boxed{100}\}$

$S = \frac{1+2+3+4+5+100}{6} = \frac{115}{6} \approx 18$

Median = 3.5

③ **Mode** = Frequency Maximum

$\{2, 1, 1, 1, 4, 5, 7, 8, 9, 10\}$

Mode = 1

EDA And Feature Engineering

Type of Flower	Age
Lilly	10
Rose	3
→ Rose	5
Sunflower	Mean or Median
Rose	8

outliers

② Measure of Dispersion

① Variance

② Standard deviation

① Variance

Population Variance

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

x_i = Data points

μ = Population Mean

N = Population size

Sample Variance

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

Bessel's Correction

WHY DOES THE SAMPLE VARIANCE HAVE N-1 IN THE DENOMINATOR? The reason we use n-1 rather than n is so that the sample variance will be what is called an unbiased estimator of the population variance

x_i – Data points

$\bar{x}$ – Sample mean

n → Sample size

Eg: $\{1,2,3,4,5\}$ ⇒ Sample

$$\bar{x} = \frac{1+2+3+4+5}{5} = 3$$

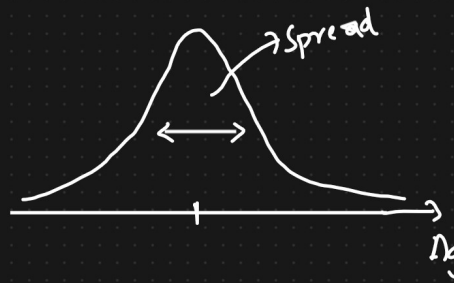
$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

x_i	$\bar{x}$	$(x_i - \bar{x})^2$
1	3	4
2	3	1
3	3	0
4	3	1
5	3	4

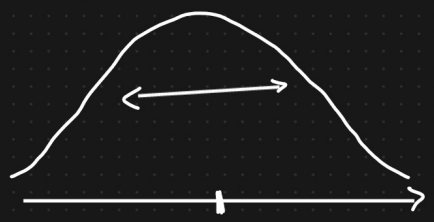
$$s^2 = \frac{10}{4} = 2.5$$

Variance : Spread of the data.

$s^2 = 2.5$



$s^2 = 6.5$ ↑↑



② Standard deviation

Population std

$\sigma = \sqrt{\text{Variance}}$

Sample std

$s = \sqrt{\text{Sample variance}}$

$s^2 = 2.5$

$\sqrt{s^2}$ = Sample Std

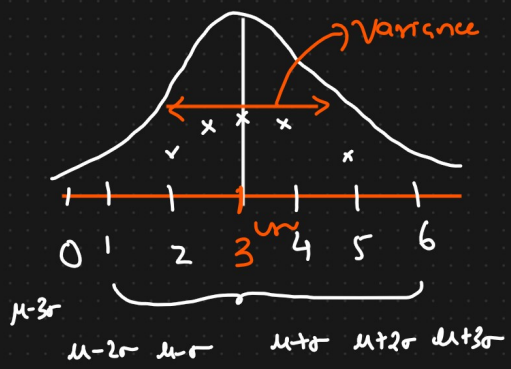
Consider

$\{1, 2, 3, 4, 5\}$

↳ $\mu = 3$ ✓

↳ $\sigma = 1$ ✓

5



③ Random Variables

$$\left\{ \begin{array}{l} x+5=7 \\ y+x=10 \end{array} \right\} \quad x=2$$

$y+2=10$

$y=8$ //.

Random Variable is a process of mapping the output of a random process or experiment to a number

Eg : Tossing a Coin

Rolling a dice

Measure the Temperature for the next day

$$X = \begin{cases} 0 & \text{if } H \\ 1 & \text{if } T \end{cases}$$

Quantifying a Random Process

$$Y = \begin{cases} \text{Sum of the rolling of dice 7 times} \\ \{4,5,6,1,2,2\} = 20 \end{cases}$$

$P(Y \geq 15)$ $\quad$ $P(H)$

④ Sets

$A = \{1, 2, 3, 4, 5, 6, 7, 8\}$

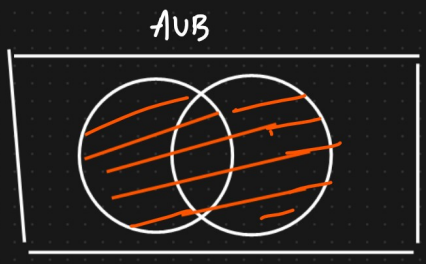
$B = \{3, 4, 5, 6, 7\}$

① Intersection

$A \cap B = \{3, 4, 5, 6, 7\}$

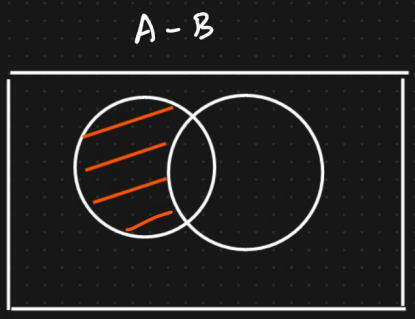
② Union

$A \cup B = \{1, 2, 3, 4, 5, 6, 7, 8\}$



③ Difference

$A - B = \{1, 2, 8\}$



④ Subset

$A \longrightarrow B \Rightarrow$ FALSE

$B \longrightarrow A \Rightarrow$ TRUE

⑤ Superset

$A \longrightarrow B \Rightarrow$ True

$B \longrightarrow A \Rightarrow$ False.