# ⊛ Random Variables:

Take an equation → $x + y = 4$
$x - y = 10$.

Here, x & y are normal variables.
Random variables are different from these.

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

eg → • Tossing a coin. This is a random process as we get (H) or (T)

• Rolling a dice {1, 2, 3, 4, 5, 6}
• measure temp for next day.

These all are random processes.

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

Hence, it quantifies a random variable

e.g. $Y = \{\text{sum of rolling of dice 4 times}\}$

Hence, random variable can take any value depending upon outcome of the random process whereas, Normal variables have fix value.

eg. $P(Y \geq 15)$

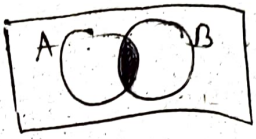
NOTE: Random Variable should be denoted by CAPITAL LETTER.

# Sets

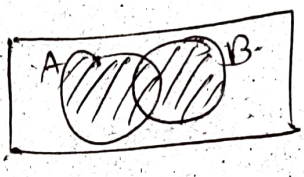
$A = \{1, 2, 3, 4, 5, 6, 7, 8\}$

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

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

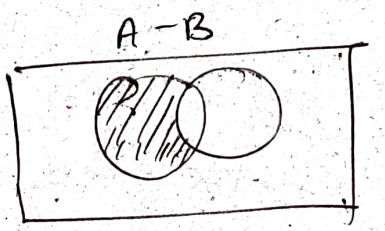



**Union**: $A \cup B = \{1, 2, 3, 4, 5, 6, 7, 8\}$




**Difference**:

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




**Subset**:

$A \rightarrow B \Rightarrow$ False

$B \rightarrow A \Rightarrow$ True

**Superset**:

$A \rightarrow B \Rightarrow$ True

$B \rightarrow A \Rightarrow$ False

## (4) Covariance and Correlation:

| X | Y |
|---|---|
| 2 | 3 |
| 4 | 5 |
| 6 | 7 |
| 8 | 9 |

{Relation between X and Y}

| | |
|---|---|
| X ↑ | Y ↑ |
| X ↑ | Y ↓ |
| X ↓ | Y ↑ |
| X ↓ | Y ↓ |

Correlation is nothing but relationship bet^n two variables.

⇒

| | |
|---|---|
| X ↑ | Y ↑ |
| X ↓ | Y ↓ |

⇒

| | |
|---|---|
| X ↓ | Y ↑ |
| X ↑ | Y ↓ |

Covariance (X, Y).

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

Hence, var(x) is nothing but covariance (X, X)

Note:

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

For | X↑ Y↑ / X↓ Y↓ | covariance will always be (+ve)

For | X↑ Y↓ / X↓ Y↑ | covariance will be (−ve).

| X | Y |
|---|---|
| 2 | 3 |
| 4 | 5 |
| 6 | 7 |
| $\bar{x}=4$ | $\bar{y}=5$ |

$$cov(X,Y) = \sum_{i=1}^{3} \frac{(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

$$= \frac{(2-4)(3-5)+(4-4)(5-5)+(6-4)(7-5)}{2}$$

$$= \frac{4+0+4}{2} = 4 \rightarrow +ve$$

Hence (+ve) covariance

**Adv. of covariance:**

1) Relatn. betn X & Y.

**Disadv. of covariance.**

1) Does not have specific Limit value.

(2) **Pearson Correlation Coefficient:** [−1 to +1]

It has specific limit value

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y} \rightarrow [-1 \text{ to } +1]$$

(1) more the value towards +1, the more +ve correlation it is. → (x,y)

(2) more the value towards −1, the more −ve correlation it is → (x,y)

What is use case?

→ Suppose you have 1000 features and you can't use all for creating ML models.

Finding correlation between features & target can help you to decide which features & more ~~dependent~~ related to target. Hence, you can drop feature whose correlation with target is close to zero.

(4) Spearman Rank Correlation:

$$r_s = \frac{cov(R(x), R(y))}{\sqrt{R(x)} \cdot \sqrt{R(y)}}$$

| X | Y | R(x) | R(Y) |
|---|---|---|---|
| 1 | 2 | 5 | 5 |
| 3 | 4 | 4 | 4 |
| 5 | 6 | 3 | 3 |
| 7 | 8 | 2 | 1 |
| 0 | 7 | 6 | 2 |
| 8 | 1 | 1 | 6 |