What is Statistics And its application?

→ Statistics is a field that deals with colection, Organization, analysis interpretation and presentation of the data.

⇓

Decision Making

e.g. Age = {24, 27, 34, 23, 28, 29, 31 32}

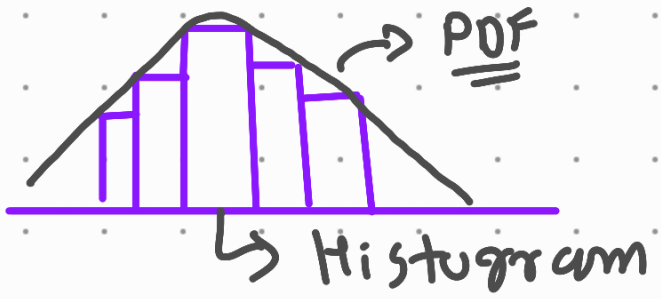
↓

{ Online Shopping }

(1) Mean
(2) Median of Age
(3) Distributions of age

→ Create Charts / Graphs

(1) Histogram
(2) PDF
(3) CDF (Cumulative Density Func)



⇒ Final Goal is to understand the data.

Application

1. Day to Day usage
2. ML / DS
3. Data Analyst
4. BA
5. Risk Analyst
6. Research

⇒ Types of Statistics

DESCRIPTIVE

→ It consists of organizing and summarizing of data.

1. Measure of central Tendency
   - i) Mean
   - ii) Median
   - iii) Mode
2. Measure of Despersion
   - i) Variance
   - ii) Standard Deviation

INFERENTIAL

→ Collect Data
- Sample Data
- Using some experiments

↓

Conclusion or Inferences

↙ other Data

{ Population Data }

Experiments:
- i) Z test
- ii) T test

Eg College A → 1000 Students

class stats

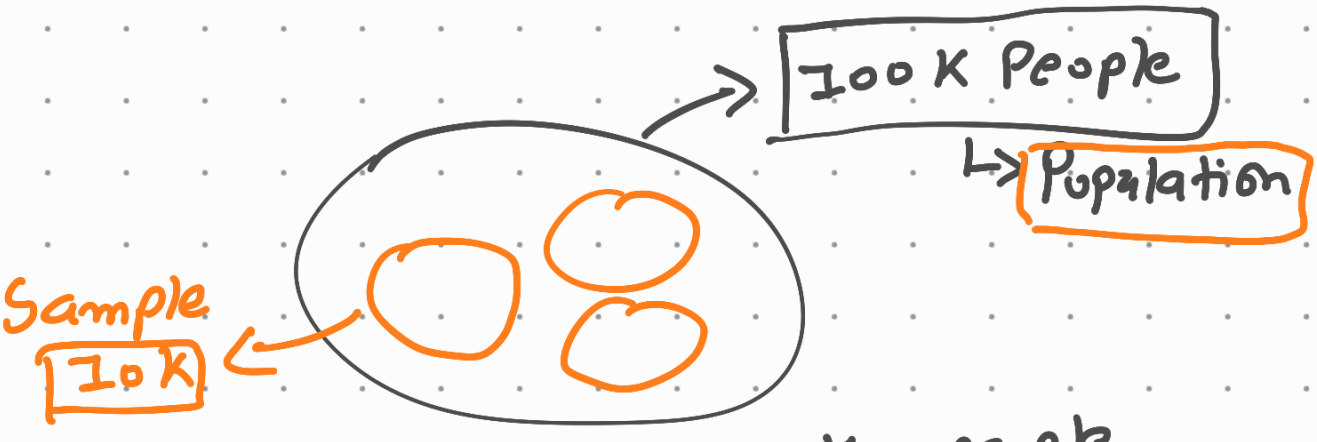
Sample → { 180 cm, 162 cm, 150 cm, 160 cm }

Descriptive → Mean, Median

Average ≈ 160

Inferential → Conclusion, Inferences

→ Population And Sample



→ Collect all the weights of all the people
(It is difficult)

→ Population → N } Notation
→ Sample → n

e.g. (1) Exit Polls

① Measure of Central Tendency

1. Mean
2. Median
3. Mod

→ Population → N

→ Sample → n

① Mean / Avg

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

↳ Respect to N

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

↳ Respect to n

eg: Age = {1, 3, 4, 5} → Distribution

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

$$\mu = 3.25$$

(2) Median

Outlier (Big Number)

Ages = {1, 3, 4, 5, 100}

$\mu = \frac{1+3+4+5+100}{5} = \frac{113}{5}$ = 22.6

→ To solve the problem of outlier we use Median.

Ages = {4, 3, 1, 5, 100}

↳ Sort the numbers

= {1, 3, 4, 5, 100} → Median = 4

① Pick the central element

if total number

① Odd → Central element

New Ages = {1, 3, 4, 5, 100, 200} → Median = 4.5

② Even → Both Central elements

$= \frac{4+5}{2} = 4.5$

③ Mod

Outlier

{4, 3, 2, 2, 2, 4, 4, 5, 2, 100}

→ Select the element with the maximum frequency.

O/P = 4 → Mod

② Measure of Dispersion

① Variance

② Standard Deviation

{Spread is Less}

Ages1 = {2, 2, 4, 4}

$\mu = \frac{2+2+4+4}{4}$

= 3

{Spread is More}

Ages2 = {1, 1, 5, 5}

$\mu = \frac{1+1+5+5}{4}$

= 3

→ Mean ←

(1) Variance

Respect to N,

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

Ages1 = {2, 2, 4, 4}

$\mu = 3$

x_i	μ	$(x_i - \mu)^2$
2	3	1
2	3	1
4	3	1
4	3	1
		$\Sigma = 4$

$\sigma^2 = \frac{4}{4} = 1$

↳ Dispersion is less.

Ages2 = {1, 1, 5, 5}

$\mu = 3$

x_i	μ	$(x_i - \mu)^2$
1	3	4
1	3	4
5	3	4
5	3	4
		$\Sigma = 16$

$\sigma^2 = \frac{16}{4} = 4$

↳ Dispersion is more

Respect to n,

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

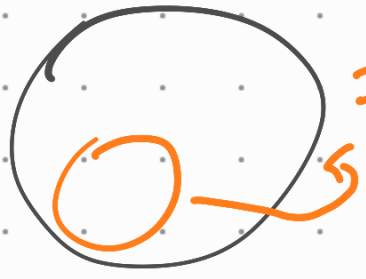
, where $\bar{x}$ = Sample mean

n = Sample Data

→ The main question is why it is divided by n-1 instead of n??

→ This is basically called as **Basal Correction.**

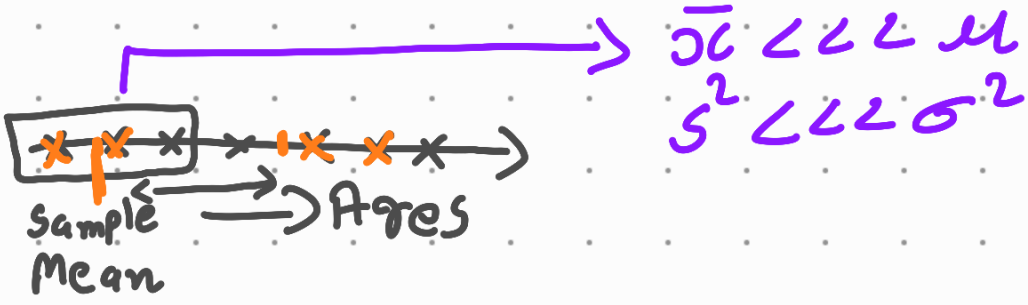
→ Let's consider that currently we use n



Area s = { ..... }

$\bar{x} \approx \mu$

$s^2 \approx \sigma^2$



$\bar{x} < < < \mu$

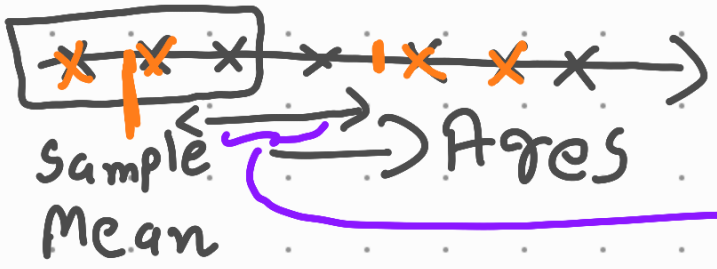
$s^2 < < < \sigma^2$

→ Lets take

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

↓

Underestimating the true population variance



→ if we use,

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

→ This is known as "Basal Correction"

↓

→ Than this value is comparitively smaller than n

→ Also not underestimating the true population variance

→ Dof = n-1

Degree of Freedom

② Standard Deviation

→ Respect to N,

$$\sigma_{(S.D.)} = \sqrt{\sigma^2}$$

→ S.D. Tells that how fars a you are from Mean.

→ Respect to n,

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

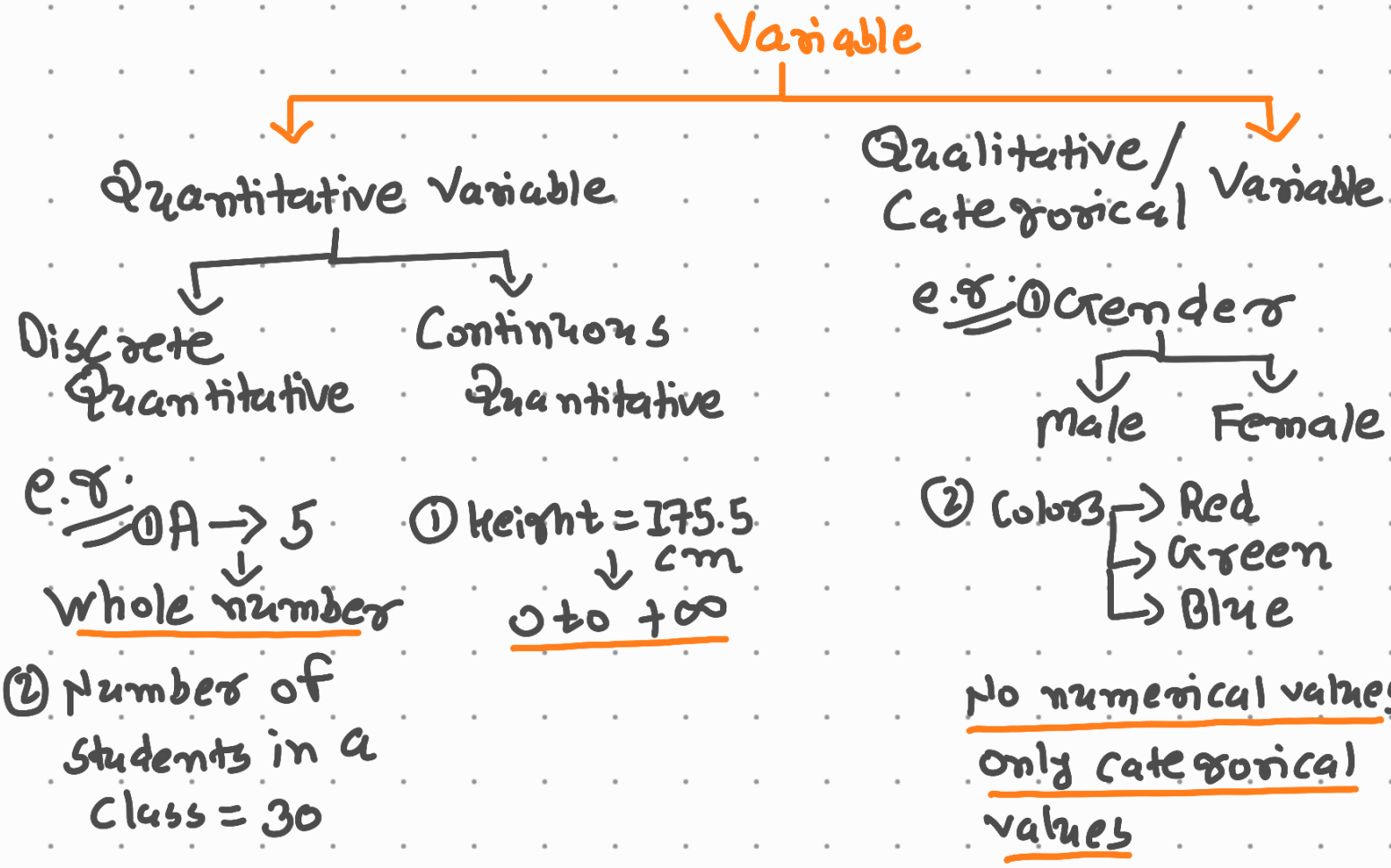
→ What is Variable?

Def:- Variable is a property that can take up any value.

e.g. Age = 25 ✓ Gender = male ✓

Ages = [12, 20, 25, 36] × Not a variable

→ Different types of Variable



→ Random Variables

Notation - X

→ Random variable is a function whose values are derived from different process or experiments.

$x = 1, x = 2, x = 3$

e.g. $y = 5x + 2$

→ Here x is random variable which can take multiple values.

e.g.

Random Process

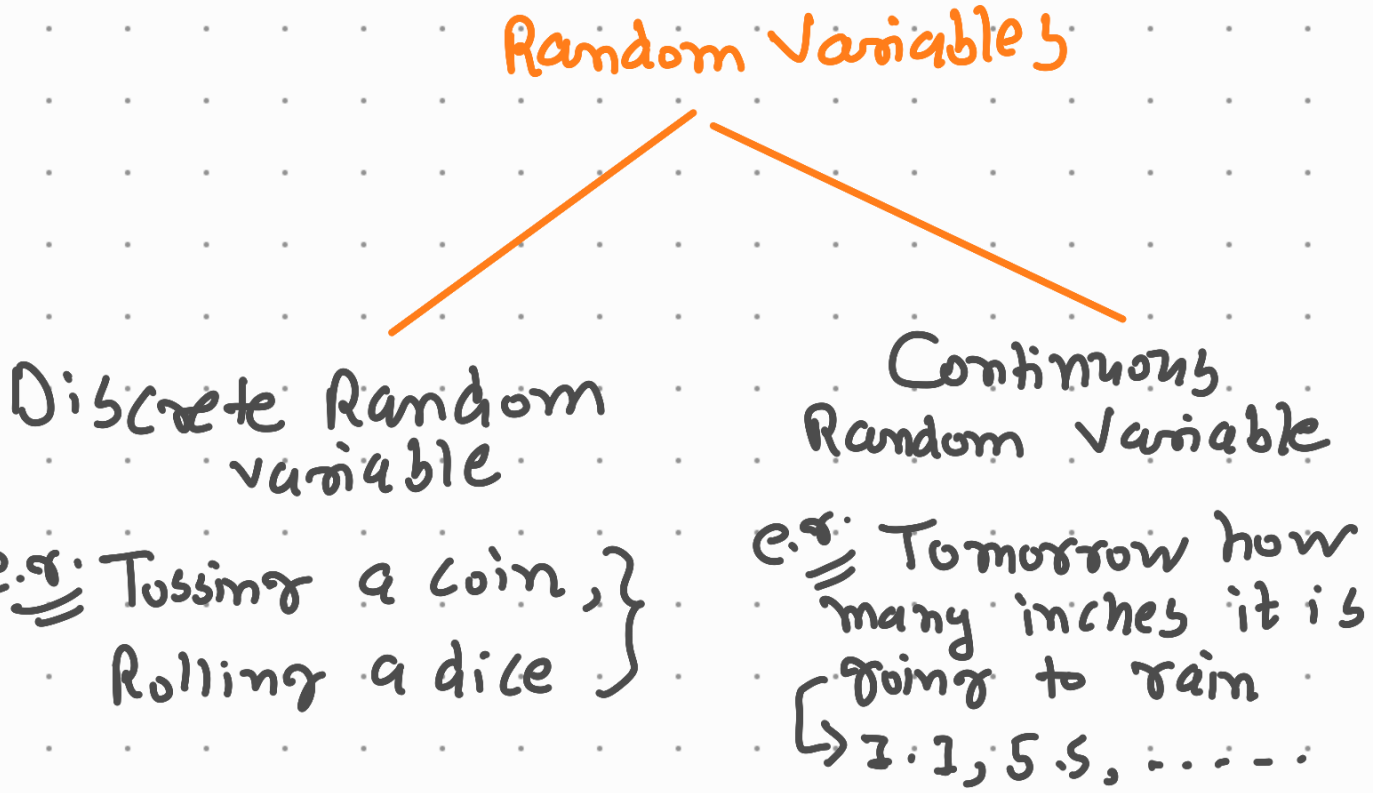
$$X = \begin{cases} 0 & H \\ 1 & T \end{cases}$$

Tossing a coin

$$X = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{cases}$$

Rolling a fair dice

→ Different Types of Random Variables



Random Variable

$X \longrightarrow$ Function $\longrightarrow$ Values

$\Downarrow$

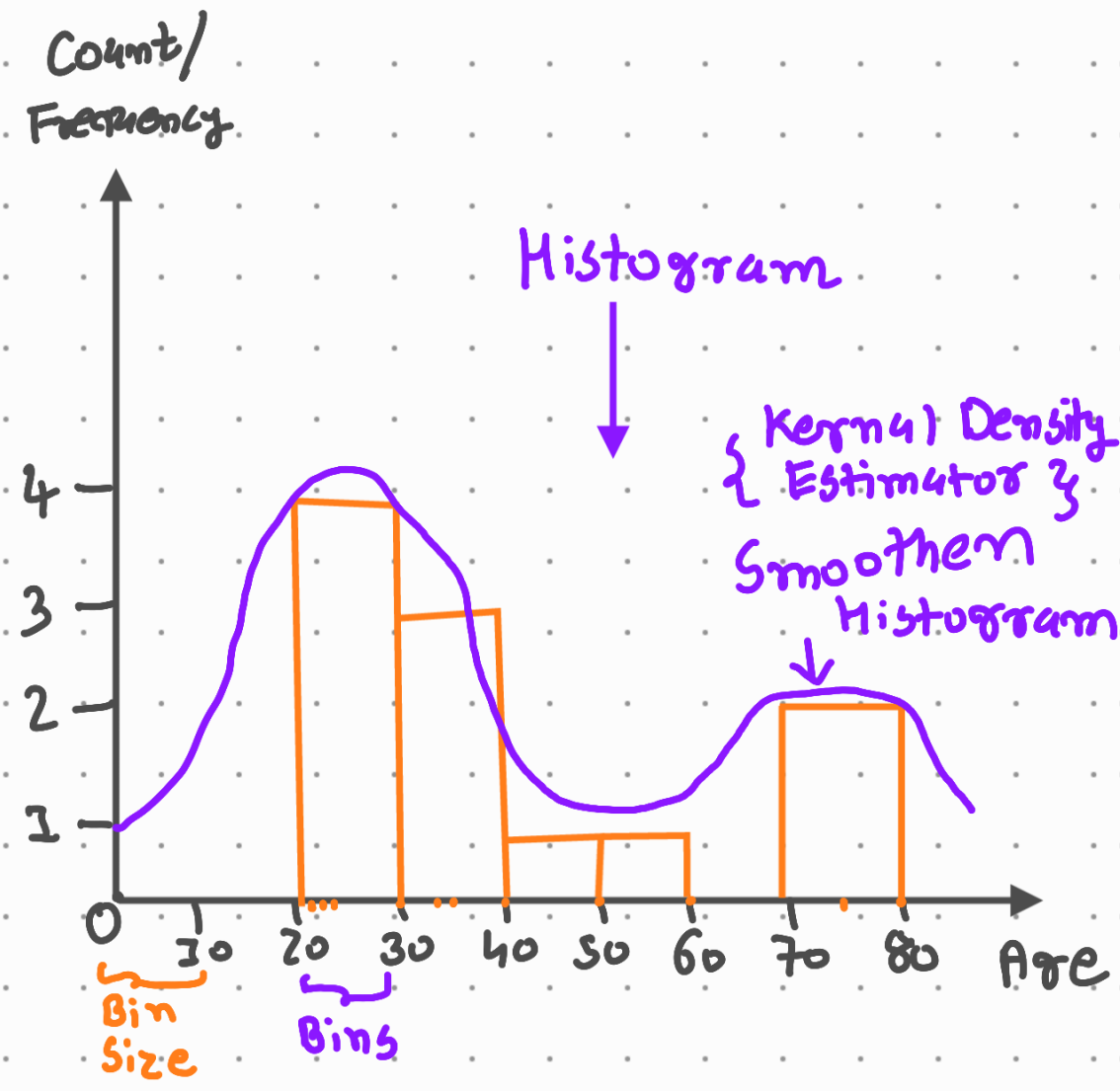
Process or Experiments

Histograms

Ages

$X = \{23, 24, 25, 30, 34, 36, 40, 50, 60, 75, 80\}$

Bins
- 20-30 = 4
- 31-40 = 3
- 41-50 = 2
- 51-60 = 1
- 61-70 = 0
- 71-80 = 2



Percentile And Quartiles

Percentage => {1, 2, 3, 4, 5, 6}

No of ODD Numbers = 3

Percentage of ODD Numbers in this group = $\frac{3}{6} \times 100$

$= 50\%$

Percentiles: A percentile is a value below which a certain percentage of observations lie.

{2, 2, 3, 4, 5, 5, 6, 7, 8, 8, 8, 9, 9, 10}

Percentile of value x = $\frac{\text{no of values below } x}{n} \times 100$

$x = 9 \Rightarrow$ Percentile $= \frac{11}{14} \times 100$

Percentile Ranking = 78.57 %ile of value 9

Value for given Percentile = $\frac{\text{Percentile}}{100} \times (n+1)$

Value $= \frac{25}{100} \times (15)$ $= 3.75$

Now take $\frac{3+4}{2}$ $= 3.5$

Quartile

25.1.ile = 1st Quartile
50.1.ile = 2nd Quartile
75.1.ile = 3rd Quartile

5 Numbers Summary

1) Minimum
2) 1st Quartile (25.1.ile)
3) Median
4) 3rd Quartile (75.1.ile)
5) Maximum

e.g. 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27 (27 ← Outlier)

[Lower fence ⟷ Higher fence]

Lower fence = $Q_1 - 1.5(IQR)$
Higher fence = $Q_3 + 1.5(IQR)$

$IQR = Q_3 - Q_1$ (Inter Quartile Range)

Q_1 (25.1.) $= \frac{\text{Percentile}}{100} \times (n+1) = \frac{25}{100}(19+1) = 5^{th}$ Position $= 3$

Q_3 (75.1.) $= \frac{\text{Percentile}}{100} \times (n+1) = \frac{75}{100}(19+1) = 15^{th}$ Position $= 7$

$IQR = Q_3 - Q_1 = 7 - 3 = 4$

Lower Fence = $3 - 1.5(4) = -3$

Higher Fence = $7 + 1.5(4) = 13$

$[-3 \longleftrightarrow 13]$ → Anything out of this range is considered as an Outlier.

Outlier ↓

1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, ~~27~~

↳ Removed

→ Minimum = 1

→ 1st Quartile = 3

→ Median = 5

→ 3rd Quartile = 7

→ Maximum = 9

5-Number Summary ⇓ → Box Plot (Use to see the Outlier)

Covariance And Correlation

→ These are two statistical measures used to determine the relationship between two variables.

→ Both are used to understand how changes in one variable are associated with changes in another variable.

(1) Covariance

Covariance is a measure of the relationship between two random variables. The metric evaluates how much – to what extent – the variables change together. In other words, it is essentially a measure of the variance between two variables. However, the metric does not assess the dependency between variables.

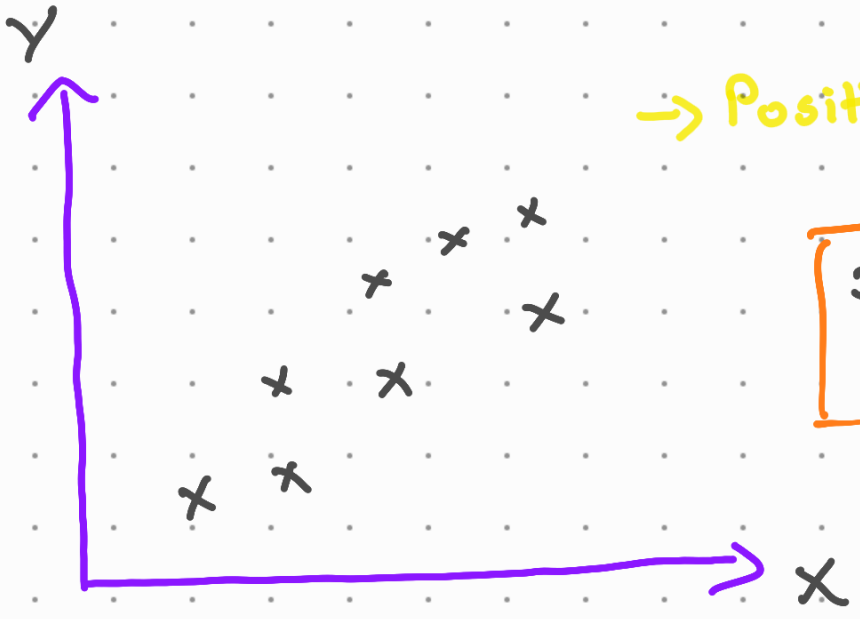
- **Positive covariance**: Indicates that two variables tend to move in the same direction.
- **Negative covariance**: Reveals that two variables tend to move in inverse directions.

e.g.

x	y
2	3
4	5
6	7
8	9

[Quantify the relationship between x and y]

x ↑ y ↑
x ↓ y ↑
x ↑ y ↓
x ↓ y ↓



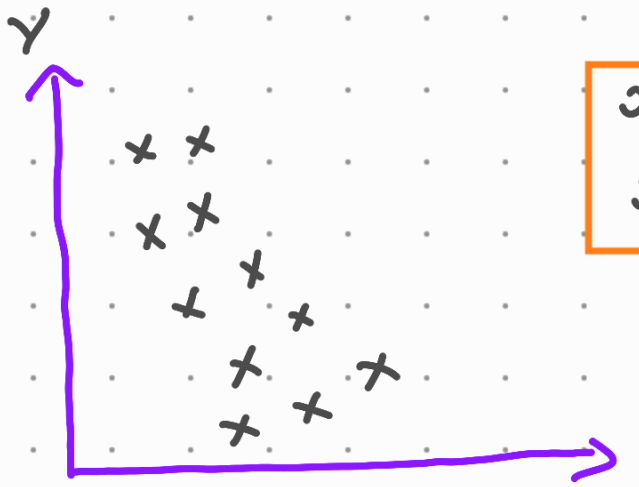
→ Positive(+ve) Covariance

x ↑ y ↑
x ↓ y ↓

→ Dataset

Size of House	Price (lakhs)
1200	45
1300	50
1500	75

→ Size of House ↑ Price ↑
→ Size of House ↓ Price ↓



x	y
7	10
6	12
5	14
4	16

$x \downarrow \; y \uparrow$
$x \uparrow \; y \downarrow$

→ Negative (-ve) covariance

→ Formula

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

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

$$Cov(x,x) = Var(x)$$

→ x_i → Datapoints of Random Variable x
→ $\bar{x}$ → Sample mean of x
→ y_i → Datapoints of Random Variable y
→ $\bar{y}$ → Sample mean of y

e.g. Students Data

Hour Studied (x)	Exam Score (y)
2	50
3	60
4	70
5	80
6	90

→ $x\uparrow$ $y\uparrow$
→ $x\downarrow$ $y\downarrow$
} +ve Covariance

① $\bar{x} = \frac{2+3+4+5+6}{5} = 4$

② $\bar{y} = \frac{50+60+70+80+90}{5} = 70$

$$Cov(x,y) = \frac{(2-4)+(50-70)+(3-4)+(60-70)+(4-4)+(70-70)+(5-4)+(80-70)+(6-4)+(90-70)}{5-1}$$

$$\boxed{Cov(x,y) = 20} \Rightarrow \text{+ve Covariance}$$

→ The number of hours studied increases the exam score.

e.g.

x	y
7	10
6	12
5	14

→ $x\downarrow$ $y\uparrow$ } -ve covariance
→ $x\uparrow$ $y\downarrow$ }

(1) $\bar{x} = \frac{7+6+5}{3} = 6$

(2) $\bar{y} = \frac{10+12+14}{3} = 12$

$$Cov(x,y) = \frac{(7-6)+(10-12)+(6-6)+(5-6)+(14-12)+(12-12)}{3-1}$$

$$= \frac{(1)+(-1)+(0)+(0)+(-1)}{2}$$

$$\boxed{Cov(x,y) = 0}$$

Advantages

① Quantify the Relationship between X and Y

Disadvantages

① Covariance doesn't have a specific limit value

$Cov(x,y) \Rightarrow -\infty$ to $+\infty$

② Correlation

- → Pearson Correlation Coefficient
- → Spearman Rank Correlation

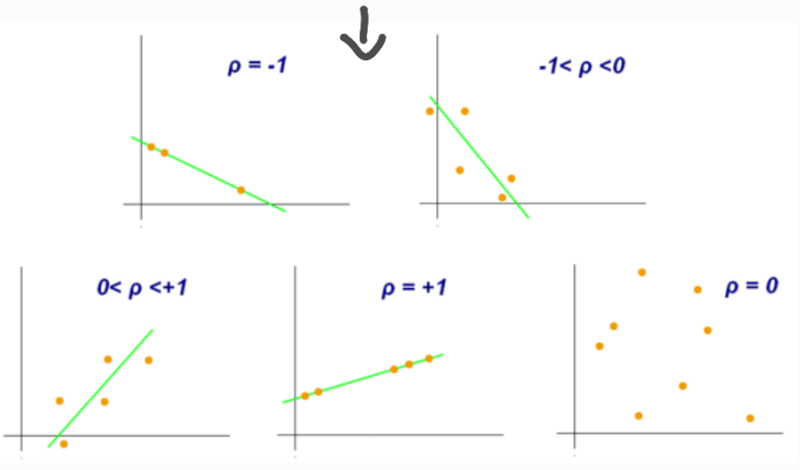
① Pearson Correlation Coefficient

→ It limits the value $\Rightarrow [-1 \text{ to } 1]$

All points in straight Line ↓

→ Formula

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$$



→ The more the value towards +1, the more +ve correlated x & y

→ The more the value towards -1, the more -ve correlated x & y

e.g: Students Data

Hour Studied (x)	Exam Score (y)
2	50
3	60
4	70
5	80
6	90

→ x↑ y↑ } +ve Covariance
→ x↓ y↓ }

① $\bar{x} = \frac{2+3+4+5+6}{5} = 4$

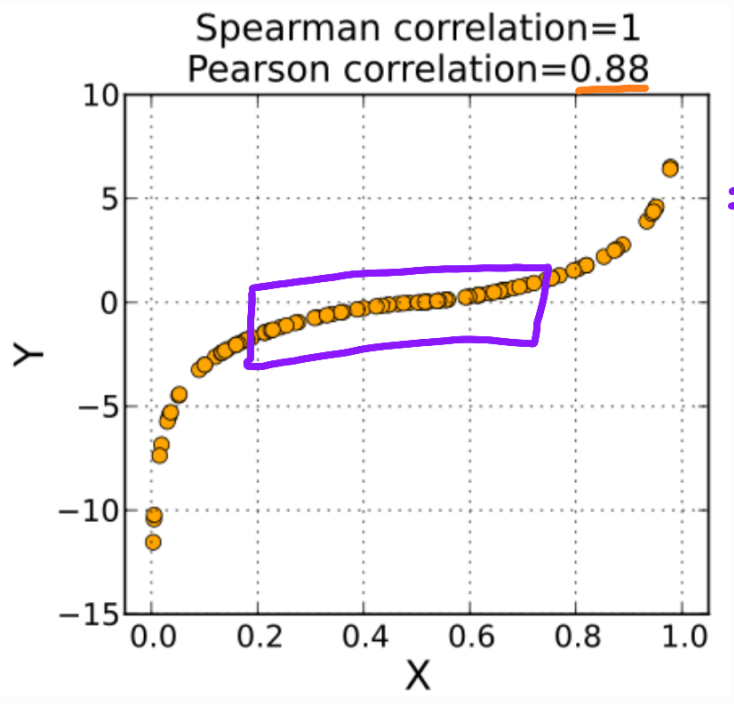
② $\bar{y} = \frac{50+60+70+80+90}{5} = 70$

$$Cov(x,y) = \frac{(2-4)+(50-70)+(3-4)+(60-70)+(4-4)+(70-70)+(5-4)+(80-70)+(6-4)+(90-70)}{5-1}$$

$$Cov(x,y) = 20$$

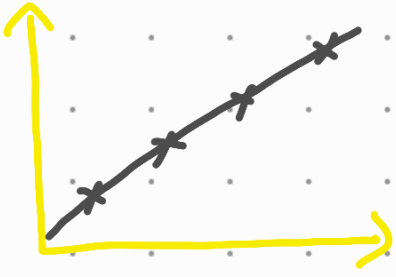
$$\rho_{x,y} = \frac{2\sigma}{\sigma_x \cdot \sigma_y} \Rightarrow 0 \text{ to } 1$$

(2) Spearman Rank Correlation



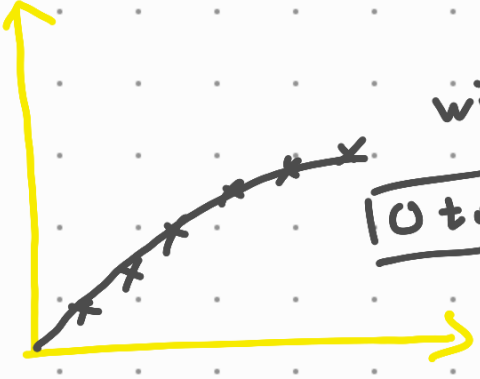
$\Rightarrow$ x↑ y↑ }
x↓ y↓ }

→ Pearson will give -1 if



For linear Data

if



will not give 1

0 to 1

→ Pearson correlation can't capture for non-linear data.

→ Formula

$$\gamma_s = \frac{Cov(R(x), R(y))}{\sigma(R(x)) * \sigma(R(y))}$$

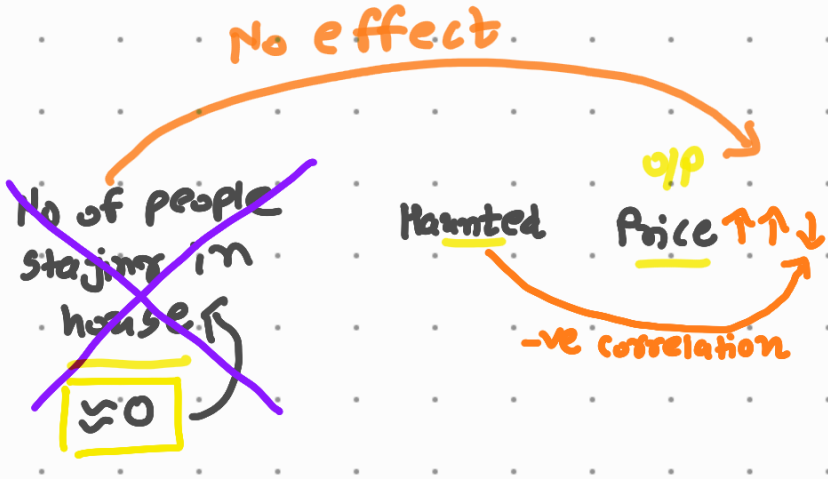
R = Rank

e.g.

x	y	R(x)	R(y)
1	2	2	1
3	4	3	2
5	6	4	3
7	8	5	5
0	7	1	4

Feature Selection

Size of ↑ | No of Room ↑ | Location ↑



Probability

→ **Probability** - It is about determining the likelihood of an event.

e.g. ① Toss a coin → {H, T}

$P_s(H) = 1/2 = 50\%$.

$P_s(T) = 1 - 1/2 = 1/2 = 50\%$.

② Rolling a dice → {1, 2, 3, 4, 5, 6}

$P_s(x=1) = \frac{1}{6}$

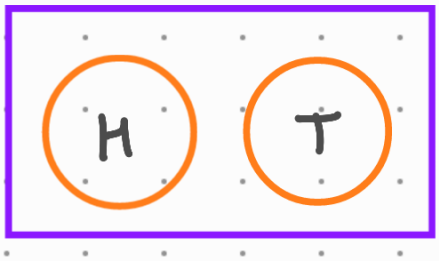
→ Formula

$$P(A) = \frac{\text{Number of Favourable Outcome}}{\text{Total Number of Favourable Outcomes}}$$

→ **Mutual Exclusion Events**

→ Two events are Mutual Exclusive if they can not occur at the same time.

e.g. ① Tossing a coin



$P_s(H \text{ or } T) = P_s(H) + P_s(T)$

↳ {Additive Rule for mutual Exclusive Event}

$P_s(H) = 1/2 \quad P_s(T) = 1/2$

$P_s(H \text{ or } T) = 1$

② Rolling a dice → {1, 2, 3, 4, 5, 6}

$P_s(1 \text{ or } 5) = P_s(1) + P_s(5)$

$= 1/6 + 1/6$

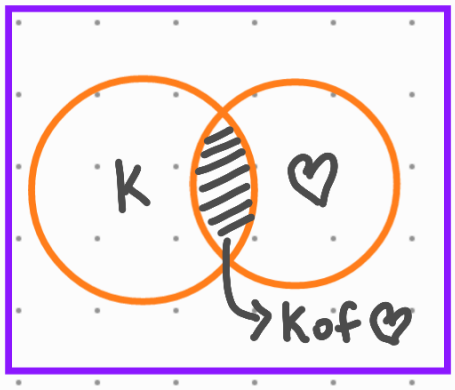
$P_s(1 \text{ or } 5) = 1/3$

→ Non-Mutual Exclusive Events

e.g. Taking a Card from the deck → Total-52 cards

$P_s(K \text{ or } ♡)$

↳ Here a card can be a K or K of ♡



$= P_s(K) + P_s(♡) - P(K \text{ and } ♡)$

$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$ ↳ {Additive Rule for Non-Mutual Exclusive events}

$= \frac{16}{52} = \frac{4}{13}$

→ Multiplication Rule → {Independent And Dependent Events}

① Independent Events

→ 2 events are independent if they do not effect one another.

e.g: ① Tossing a coin → {H and then T}

$P_s(H) = 1/2 \quad P_s(T) = 1/2$

→ Getting Tail first time does not effect getting Head then.

② Rolling a dice → {1, 2, 3, 4, 5, 6}

$P_s(1) = 1/6 \quad P_s(2) = 1/6$

→ Both events does not effect each other same as tossing a coin.

② Dependent Events

→ 2 events are dependent if and only if they effect each other.

e.g. ① Take a king card from the deck and then queen card from the deck. (without putting cards back)

$$P_s(K) = \frac{4}{52} \qquad P_s(Q) = \frac{4}{51}$$

↳ These type of events are Dependent events.

Multiplication Rule Formula

① Independent events

$$P_s(A \text{ and } B) = P_s(A) * P_s(B)$$

② Dependent events

$$P_s(A \text{ and } B) = P_s(A) * P_s(B/A)$$

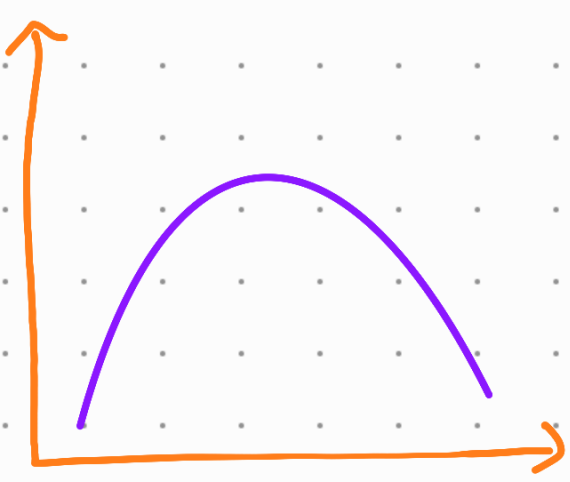
↳ Conditional Probability

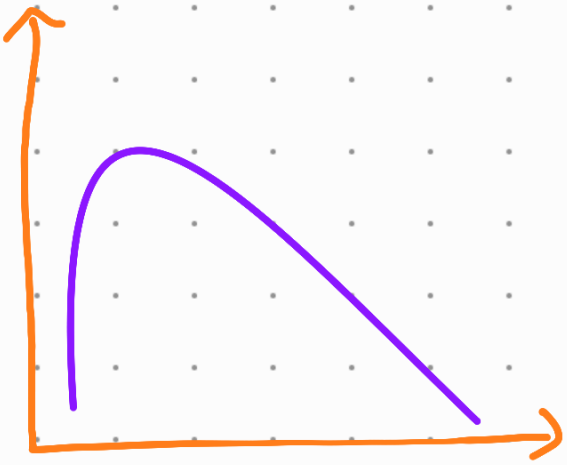
Probability Distribution Functions

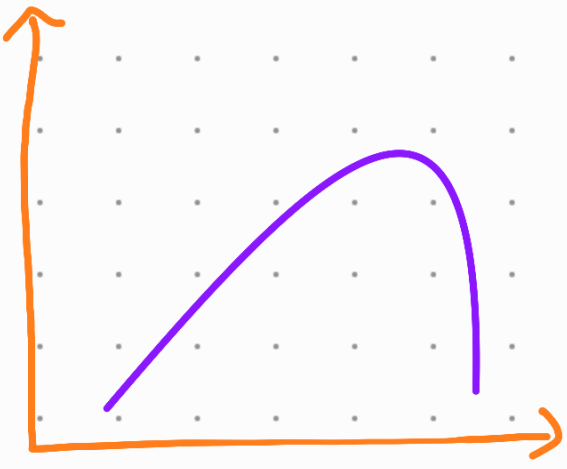
→ PDF describe how the probabilities are distributed over the values of a random variable.

e.g. Ages = { ..... } ⇒ Continuous Random Variables

→ Various PDFs







2 Main types of Probability Distribution Function

(1) Probability Mass Functions (PMF)

↳ Used for discrete random variables

(2) Probability Density Function (PDF)

↳ Used for Continuous random variables