Introduction to Statistics

Def^n:-
Statistics is the science of collecting, organizing and analyzing data.

↓

"facts / pieces of info"

eg:- Height of students in classroom.
{175cm, 180cm, 190cm, 160cm, ...}.

eg:- IQ
{100, 90, 95, 80, ...}.

→ collecting, organizing & analyzing

↓

Used for Decision Making Process.

Types of Statistics.

① Descriptive Statistics.
② Inferential Statistics.

using whole data to answer question.

Conclusion drawing from sample set

Descriptive Statistics.

↓

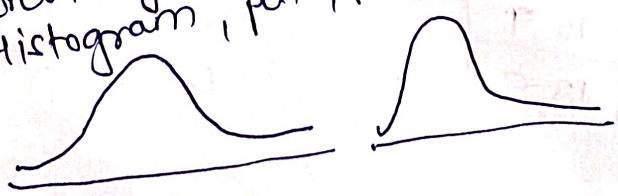
Def^n:- It consist of organizing and summarizing data.

Topics

① Measure of Central Tendency.
[Mean, Median, Mode].

② Measure of Dispersion.
[Variance & Standard deviation].

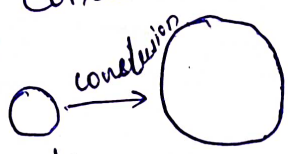
③ Different types of distribution.
→ Histogram, pdf, pmf



Inferential Stats.

↓

Def^n:- It consist of using data you have measured to form conclusion.



Sample data — Population data.

Topics

① Z-Test
② t-Test
} Hypothesis testing
Ho, H1, P-value, significance value.

③ CHI square Test

Eg:- Let's say there are 20 statistics classes at your college and you have collected the heights of student in the class.

Heights are recorded as. 175 cm, 180 cm, 140, 140, 135, 160, 135, 190

Descriptive Question

"What is the ~~common~~ ~~ao~~ average height of entire classroom".

↳ Mean

$$\frac{175+180+140+140+135+160+135+190}{8} = \text{Avg Height}$$

Inferential Question [Conclusion Drawing].

"Are the height of the Students in classroom, similar to what you expect in the entire college"

→ Sample data.

↳ Population data

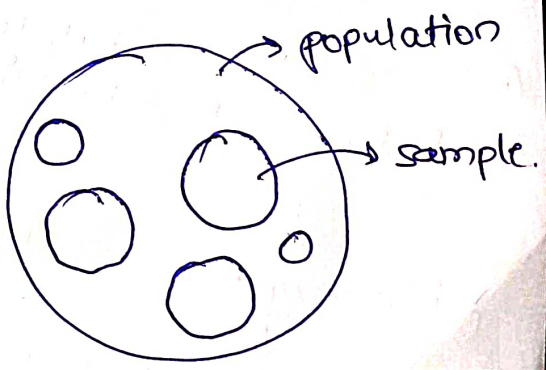
Population And Sample Data.

Population :-

The group you are interested in studying.

Sample :-

A subset of population.



Ex :- Exit poll after Election

Types of Data

DMC	DC	ISI	BUI	FWI	classes	Region
3.4	7.6	1.3	3.4	0.5	not fire	0
4.1	7.6	1	3.9	0.4	not fire	0
2.5	7.1	0.3	2.7	0.1	not fire	0
1.3	6.9	0	1.7	0	not fire	0
3	14.2	1.2	3.9	0.5	not fire	0
5.8	22.2	3.1	7	2.5	fire	0
9.9	30.5	6.4	10.9	7.2	fire	0
12.1	38.3	5.6	13.5	7.1	fire	0

DATA

① Quantitative Data
[+, -, /, *]
Numerical variable

- Discrete → Whole Number
- Continuous → Any value.

② Qualitative Data → Categorical variables

- Nominal → Not Ranked
- Ordinal → Ranked

→ In the above table :- Except Class all others are quantitative variables. As they can be measured, recorded and arithmetic operation can be performed on them.

eg:- No. of Bank Accounts of a person
No. of Children in a Family
] Discrete Quantitative Data variable.

eg:- weight, Height, Age
Temperature, Speed
] Continuous Quantitative Data variable.

eg:- Gender → M, F, Blood Group → A, B, AB, O
Pin code
] → Is categorical. → Nominal Qualitative Data. → Does not have specific Rank.

eg:- Customer Feedback → Good, Better, Bad
] → Is Ranked → Is categorical → Ordinal Qualitative Data.

Scale of Measurement

① Nominal Scale Data.
② Ordinal Scale Data.
③ Interval Scale Data.
④ Ratio Scale Data.

① Nominal Scale Data

a] Qualitative or Categorical Data.
b] Eg:- Gender, Colour, Labels
c] Order does not matter.

Eg:- Favourite colour, survey
Red → 5 → 50%
Blue → 3 → 30%
Orange → 2 → 20%
10

② Ordinal Scale Data

a] Ranking is important.
b] Order Matters.
c] Difference cannot be measured

Eg:- App Reviews
~~① Good~~ 1 → Best
~~② Better~~ 2 → Good
3 → Bad.

Rank is assigned
Order is imp.

eg:- Working Professional Ranks days.
5 ← M
4 ← T
3 ← W
2 ← Th
1 ← F

Race Ranks.
1st 2nd 3rd
we cannot count the difference.

③ Interval Scale Data.

a] Order Matters.
b] Difference can be measured.
c] Ratio Cannot be measured.
d] No True "0" Starting point

↳ In case of temp, it can be -ve, 0 or +ve so there is no such starting point.

Eg:- Temperature Variable.
30F
60F
90F
120F
90 - 60 = 30F

60 : 30 = 2 : 1

It does not indicate that the heat felt at 60°F is twice that of at 30°F

④ Ratio Scale of Data.

a] Order Matters.
b] Differences are measurable. (including Ratio)
c] Contains a '0' starting point

Ratios : $\frac{90}{30} = \frac{3}{1} : 1$ → student scoring 90 marks get 3 times more marks.

→ Contains a fixed starting point.

eg:- Student marks in class
0, 90, 60, 30, 75, 40, 50

Ascending order
↳ 30 40 50 60 75 90

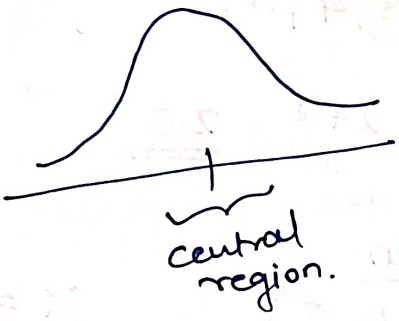
Differences
40 - 30 = 10
50 - 30 = 20
90 - 30 = 60.

Example

1. length of different rivers in the world. → Ratio Scale of Data.
2. Favourite food based on Gender. → Nominal Scale Data.
3. Marital Status → Nominal Scale Data.
4. IQ measurement → Ratio Scale Data.

Measure of Central Tendency

1. Mean or Average
2. Median
3. Mode



✓ used for EDA
✓ Also in Feature Engineering.

1. Mean

Population (N)

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

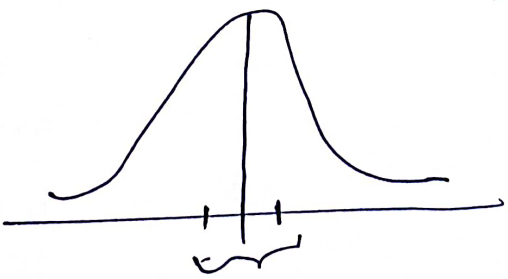
$$\text{Population Mean} = (\mu) = \sum_{i=1}^{n} \frac{x_i}{N}$$

$$= \left[\frac{1+1+2+2+3+3+4+5+5+6}{10}\right]$$

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

Sample (n)

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



② Median

Steps:- $X = \{4, 5, 2, 3, 2, 1\}$

① Sorting the elements of Random variable X

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

② Check the No. of elements.

Count = 6

③ If Count == Even

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

$\Downarrow$

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

④ if Count == Odd.

$\{1, 2, 2, 3, 4, 5, 6\}$

$\Downarrow$

Median $= \underline{\underline{3}}$

Why Median

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

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

$= \frac{15}{5} = 3$

$\boxed{\bar{x} = 3}$

$X = \{1, 2, 3, 4, 5, 100\}$ ← Outlier → does not belong to distribution

$\bar{x} = \frac{1+2+3+4+5+100}{6}$

$= \frac{115}{6}$

$= \underline{\underline{19.1}}$ → Because of outlier Huge shift in the Sample mean

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

Count = Even

$\therefore$ Median $= \frac{3+4}{2} = \underline{\underline{3.5}}$

→ Median is used to find the central tendencies when outliers is present

③ **Mode** :- Frequency Maximum.

{2, 1, 1, 1, 4, 5, 7, 8, 9, 9, 10}.

Mode = 1 (Max^m occuring frequency)

EDA and Feature Engineering.

Age	Weight	Salary	Gender	Degree.
24	75	40K	M	BE
25	80	70K	F	–
27	95	45K	~~M~~	–
24	–	50K	M	Phd
32	–	60k	–	BE
Missing Data –	60	–	–	Masters
–	65	55K	–	BSC
40	72	–	M	BE
↓	Missing data. ↓	↓	↓	↓
Mean	Mean	Mean	Mode	Mode

Implementation of Measures of Central Tendency

Mean (Average).

→ age = [12, 21, 23, 45, 65, 43, 56, 45, 32, 67, 54, 34]

import numpy as np

⇒ np.mean(age)

→ weights = [54, 45, 34, 54, 34, 67, 54, 56, 75]

np.mean(weights)

```
import seaborn as sns.
df = sns.load_dataset('tips')
df.head()
```

⇒ np.mean(df['total_bill'])

Median

→ np.median(age)

→ age = [12, 21, 23, 45, 65, 43, 56, 45, 32, 67, 54, 34, 200]

outlier.

→ np.mean(age)

→ np.median(age)

Mode

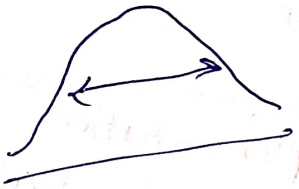
from scipy import stats.

stats.mode(age).

np.median(df['total_bill']

Measure of Dispersion

① Measure of Dispersion [Spread of the data].



A] Variance.

B] Standard deviation.

A] Variance

Population Variance

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

x_i = data points

μ = Popln mean

N = Populn size.

Sample Variance

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

x_i → data points.

$\bar{x}$ → sample mean

n → sample size.

Why we divide sample variance by n-1 ?

Ans) The sample variance is divided by n-1 so that we can create an unbiased estimator of the population variance

⇓

Bessele Correction.

Eg:- $\{1, 2, 3, 4, 5\}$

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

x	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	3	-2	4
2	3	-1	1
3	3	0	0
4	3	1	1
5	3	2	4
$\Sigma = 15$			$\Sigma = 10$

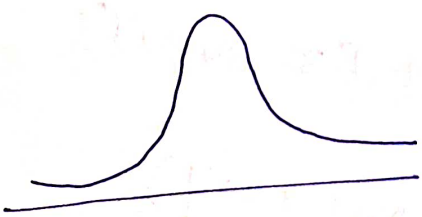
$$S^2 = \frac{10}{5-1} = \frac{10}{4} = 2.5$$

$X = \{ \quad \}$

$S^2 = 2.5$

⇓

Dispersion or spread



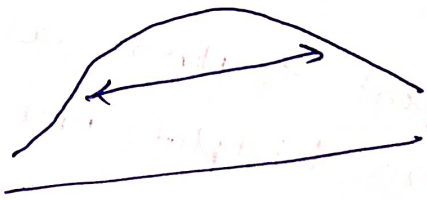
Spread is less.

$Y = \{ \quad \}$

$S^2 = 7.5$

⇓

Dispersion or spread



Spread increases.

② Standard deviation

Population SD

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

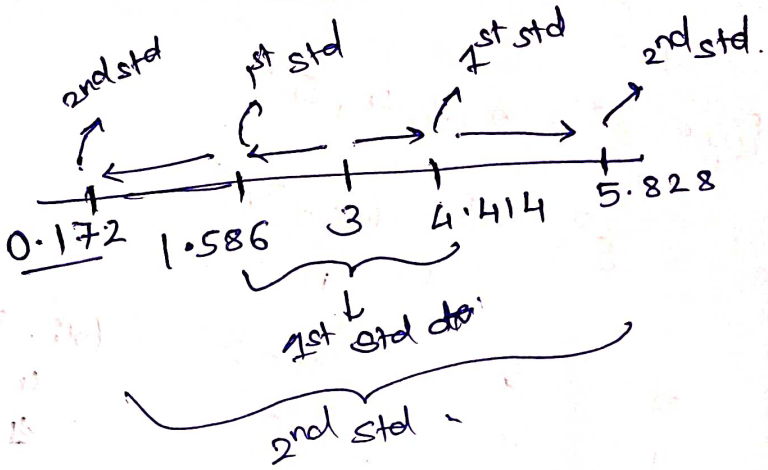
Sample Standard Deviation

$\text{Std} = \sqrt{s^2} \rightarrow$ Sample Variance

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

$\bar{X} = 3$

$\sigma = 1.414$



Measure of Dispersion → Python

```
## Measure of Dispersion.
ages_lst = [23, 43, 23, 56, 74, 32, 68, 98, 45, 32]
import numpy as np
np.mean(ages_lst).
np.var(ages_lst)
np.std(ages_lst)
import seaborn as sns
sns.histplot(ages_list, kde= True)
import pandas as pd
data = [[10, 12, 13], [34, 23, 45], [32, 34, 21]]
data
pd.DataFrame(data, columns = ["A", "B", "C"])
df.head()
```

df.var() # to get variance column wise
df.var(axis=1) # to get variance row wise.

Random Variables

Random Variable

Def^n :- Is a process of mapping the o/p of a random process or experiments to a number.

Eg :- Tossing a Coin.

$$x = \begin{cases} 0 & \text{if Head} \\ 1 & \text{if Tails.} \end{cases}$$

Eg :- Rolling a dice

Eg :- Measure the Temperature of the next day

$$Y = \{\text{sum of Rolling of dice 7 times}\}$$

$\Downarrow$

$P_r(Y \geq 15)$. $P(Y < 10)$

$\downarrow$

Probability

Sets

① Sets :- Organized collection of Elements.

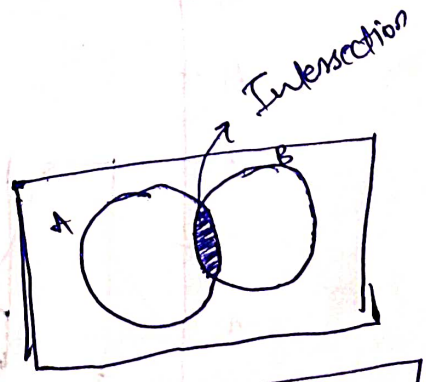
$A = \{1, 2, 3, 4, 5, 6, 7, 8\}$

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

Operations on Sets.

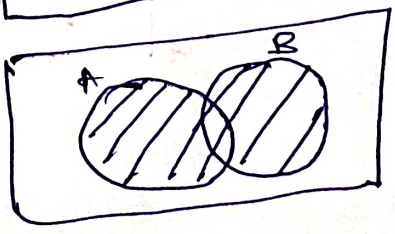
① Intersection

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



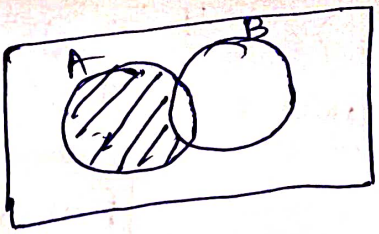
② Union

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



③ Difference

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



④ Subset

$A \to B \Rightarrow$ False.

$B \to A \Rightarrow$ True. [B/c all elements of B are present in A]

⑤ Superset

$A \to B \Rightarrow$ True

$B \to A \Rightarrow$ False.

Histograms and Skewness

Histogram → [Frequency]

Ages = { 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51 }.

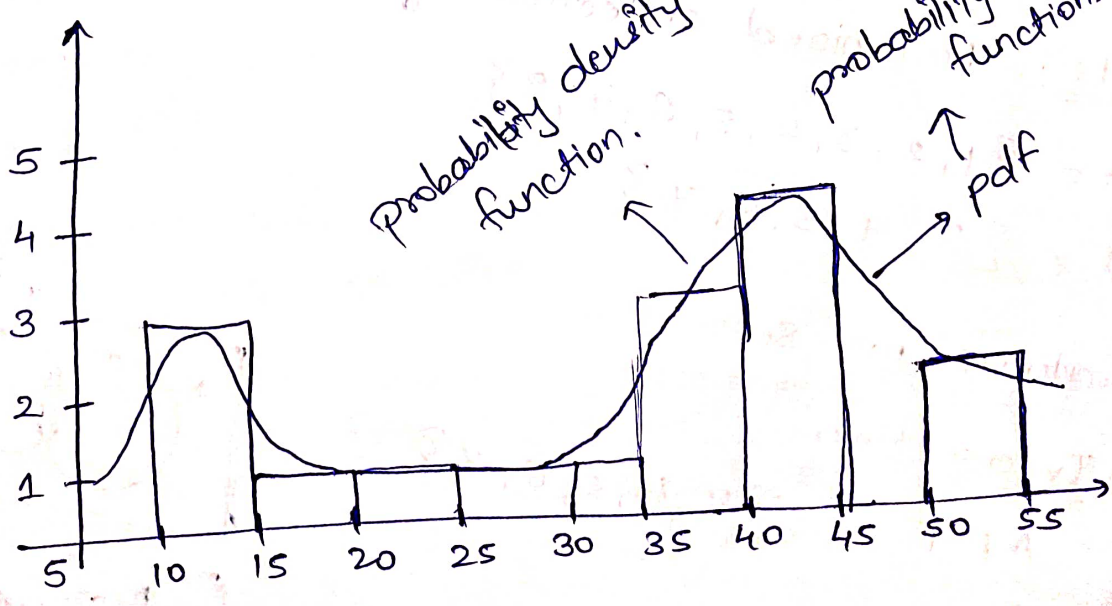
3 1 1 3 4 2

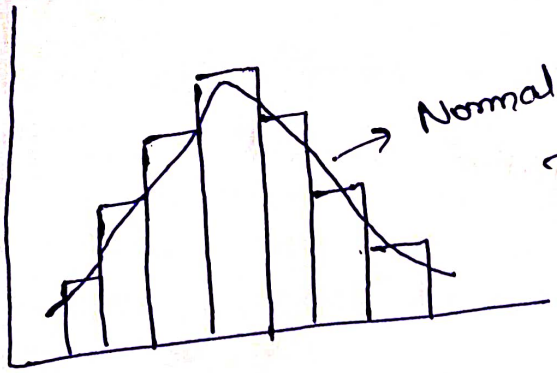
{No. of bins | buckets = 10}.

$\frac{50}{10} = \boxed{5}$ → bin size

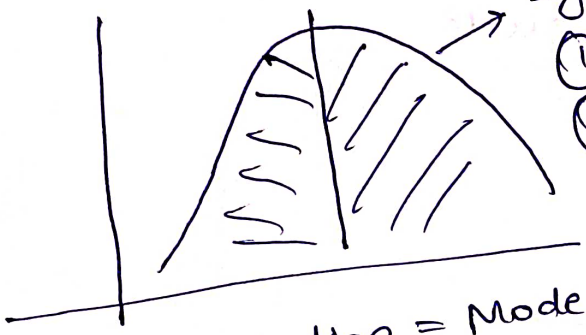
{No. of bins = 20}

$\frac{50}{20} = 2.5$ → bin size





→ Normal | Gaussian Distribution.



→ Symmetrical Distribution

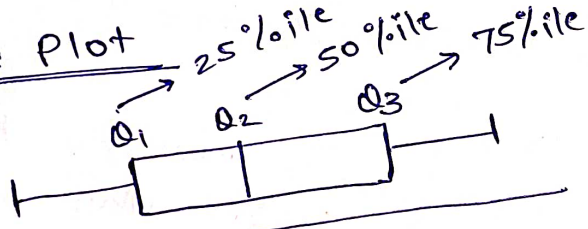
① No skewness

② Median, Mean & Mode

↳ All are perfectly at center

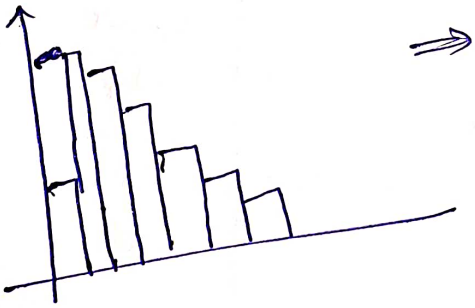
→ Mean = Median = Mode.

Box Plot

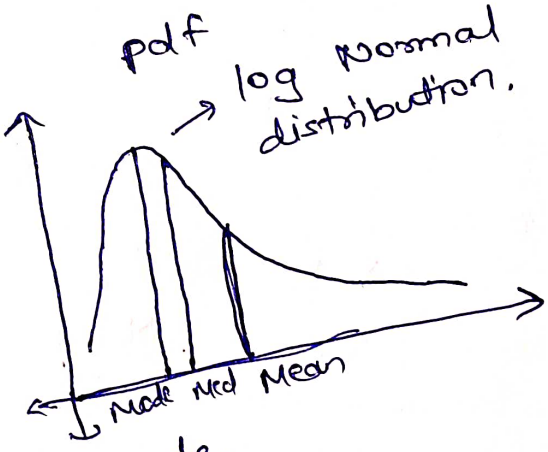


$Q_3 - Q_2 \cong Q_2 - Q_1$

② Right Skewed Data.



⇒ positive skewed

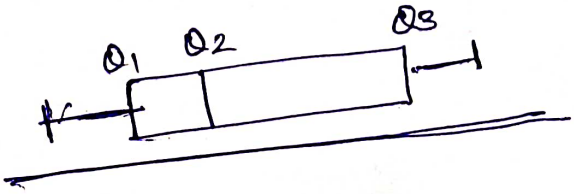


→ Relationship between Mean, Median, Mode.

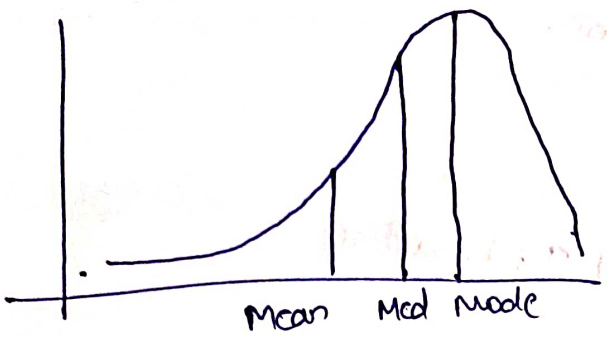
$\boxed{\text{Mean} \geq \text{Median} \geq \text{Mode}}$

$\boxed{Q_3 - Q_2 \geq Q_2 - Q_1}$

→ Box Plot

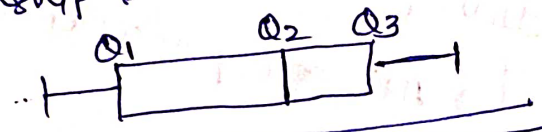


③ Left Skewed Distribution



⇒ Negative Skewed

Relationship :- Mean ≤ Median ≤ Mode.



$$\boxed{Q_2 - Q_3 \leq Q_2 - Q_1}$$

Covariance and Correlation.

Random variables

X	Y
2	3
4	5
6	7
8	9

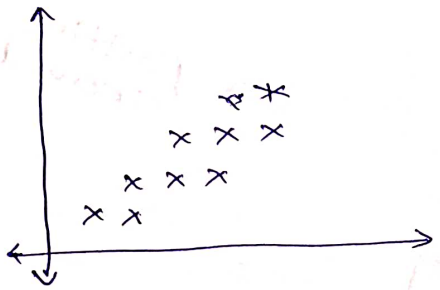
[Relationship between X and Y]

X↑	Y↑
X↓	Y↑
X↓	Y↓
X↑	Y↓

→ To quantify the relationship we use covariance & correlation.

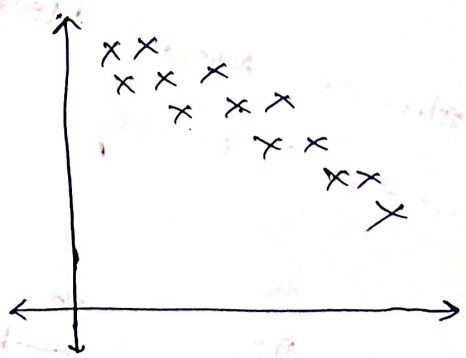
X↑	Y↑
X↓	Y↓

→



X↑	Y↓
X↓	Y↑

→



↑ Size of House → ↑ Price of House

↓ → ↓

Covariance

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

$x_i, y_i \rightarrow$ Data points of x & y

$\bar{x}, \bar{y} \rightarrow$ sample mean of $x | y$

$n \rightarrow$ sample size.

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

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

$= Cov(X, X) \Rightarrow$ Spread of Data.

$Cov(X, Y)$

$X\uparrow$	$Y\uparrow$
$X\downarrow$	$Y\downarrow$

+ve Covariance.

$X\uparrow$	$Y\downarrow$
$X\downarrow$	$Y\uparrow$

-ve Covariance

Eg

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

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

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

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

$$Cov(X,Y) = \frac{4+4}{2} = \frac{8}{2} = 4 \text{ +ve}$$

$\Downarrow$

Positive Co-variance

X & Y are having Positive Co-variance.

Advantages

① Helps to find relationship b/w X and Y
-ve or +ve value.

Disadvantages

① Covariance does not have a specific limit value.

② **Pearson Correlation Coefficient** O/P ranges b/w [-1 to 1]

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

A] The more the value towards +1 the more +ve correlated it is. (X, Y)

B] The more the value towards -1, the more -ve correlated it is (X, Y).

③ **Spearman Rank Correlation** [-1 to 1]

$$\gamma_s = \frac{Cov(R(X), R(Y))}{\sigma(R(x)) \times \sigma(R(x))}$$

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

Feature Selection

Size of House ↑ +ve

No. of Rooms ↑ +ve

Locat^n ↑ +ve

No. of People staying ≈ 0 → This variable is not important to determine price.

Price ↑ +ve

Haunted ↓ -ve → Correlation =

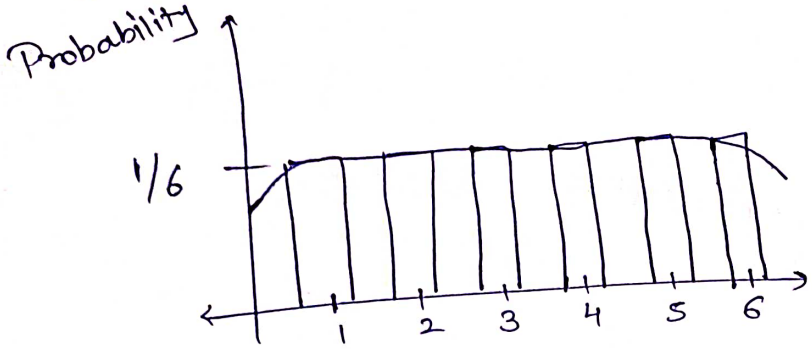
② Probability Mass Function (pmf)

Variable ⟶ Discrete Random Variable.

Eg :- Rolling a Dice. $\{1, 2, 3, 4, 5, 6\}$.

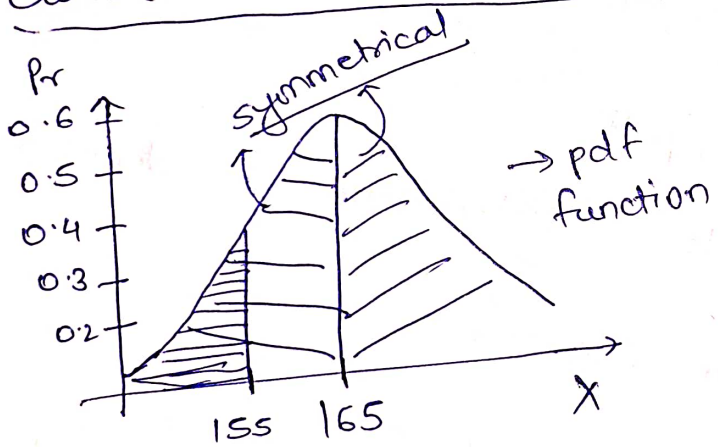
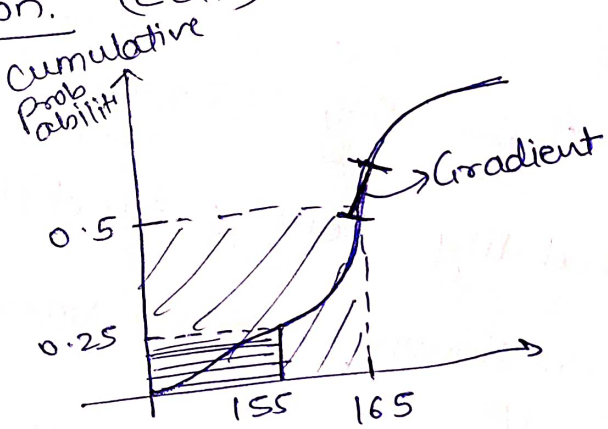
$Pr(1) = 1/6$

$Pr(2) = 1/6$



$$Pr(X \leq 4) = Pr(X=1) + Pr(X=2) + Pr(X=3) + Pr(X=4)$$

$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$$

$$= \frac{4}{6} = 2/3$$

③ Cumulative distribution function. (cdf)