Probability and Statistics

Introduction to Statistics

(decision making Process)

def: Statistics is the science of collecting, organising and analysing data.

Data: "facts & pieces of information"

eg: Height of students in classroom.
IQ of students in classroom.

Types of Statistics

Descriptive Statistics

It consist of organising and summary data.

(1) Measure of Central tendency.
   (i) Mean
   (ii) Median
   (iii) Mode
(2) Measure of dispersion
   (i) Variance
   (ii) Standard deviation
(3) Diff type of distribution of data.
   eg: Histogram
   Probability distribution function.
   Probability Mass function.

Inferential Statistics

It consist of using data you have measured to form conclusion.

from Sample data → Population data

concluding

(i) Z-test
(ii) t-test

} Hypothesis Testing
H_0, H_1
P value
Significa value

(iii) Chi square test

g: There are 20 Statistic Classes at your University and you have collected the heights of student in the class

Heights are recorded as

{175, 180, 140, 140, 125, 160, 135, 190}

Descriptive Question

"What is the ~~common~~ avg height of the entire classroom"

→ Sample data.

Inferential Question

"Are the height of the students in the classroom Similar to what you expect in the entire University?"

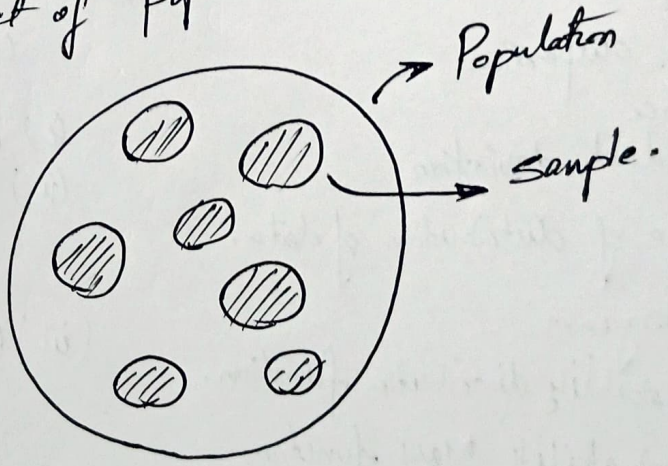
→ Population data

Population data & Sample data.

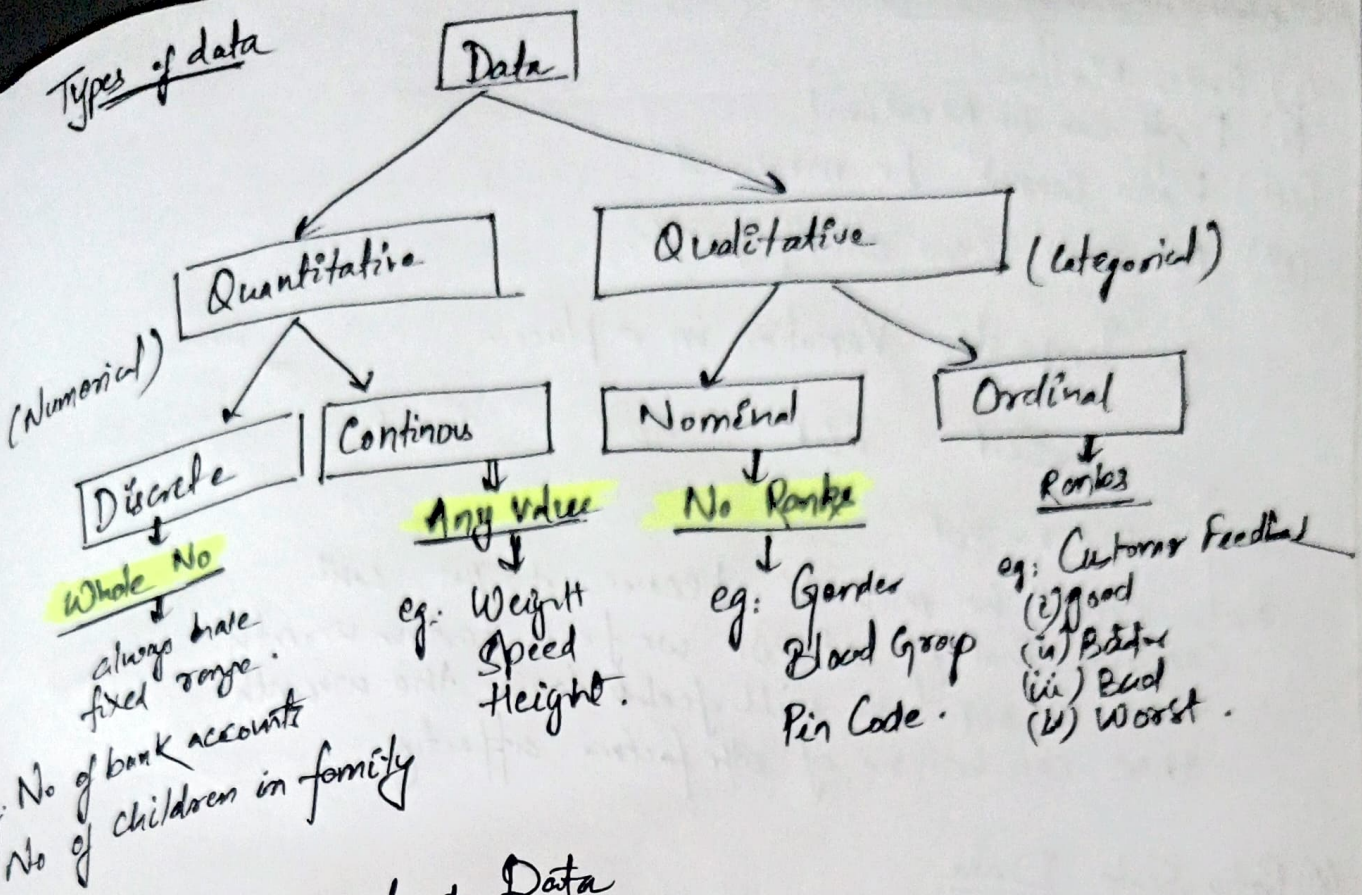
Population : The group you are interested in studying.

Sample : A subset of Population.

g: Exit Poll.



Types of data



always have fixed range.

Scale of Measurement of Data

(1) Nominal Scale Data
(2) Ordinal Scale Data
(3) Interval Scale Data
(4) Ratio Scale Data

(1) Nominal Scale Data

(i) Qualitative
(ii) Categorial Data
(iii) Order/Rank doesnot matter

eg: Survey on favorite Color.
(Here order doesnot matter on what color people like.)

(2) Ordinal Scale Data.

(i) Ranking is Important
(ii) Order Matters
(iii) Difference cannot be measured

eg: An feedback form
(i) Best
(ii) Good
(iii) Bad
(iv) Worst

↳ Just on the basis of ranks we cannot calculate difference → like in this case we donot know the reason for this ranking

(3) Integral Scale of Data

(i) Order Matters
(ii) Diff can be Measured
(iii) Ratio cannot be measured
(iv) No true Zero Starting point.

eg: Temperature Variation in a place.

30f 60f 90f 120f

Diff: 60-30 = 30f
Ratio cannot be measured because it's just cnt
Conclude that if at 30f we feel certain warmth
then at 60f we will feel double the warmth
there can be no of other factors effecting.

(4) Ratio Scale Data

(1) The Order matters
(2) Diff are measurable (including ratio)
(3) Common '0' starting point.

Grades of students
0, 90, 60, 30, 75, 40, 50

Measure of Central Tendency

(i) Mean or Average
(ii) Median
(iii) Mode

Mean

Population (N)

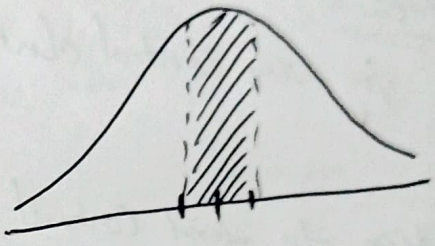
$X = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$

Population Mean (μ) $= \frac{\sum_{i=1}^{N} x_i}{N} = \frac{[1+1+2+2+3+3+4+5+5+6]}{10}$

$= 3.2$

Sample Mean (n)

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



Median

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

Step
(1) Sort the random variable X. : $\{1, 2, 3, 4, 5\}$
(2) No of elements count : 5
(3) if (count % 2 == 0)
find central element
if cont = 6 $\{1, 2, \boxed{2, 3,} 4, 5\}$

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

or if (count % 2 != 0)
find central element
$2.5 \rightarrow (3)$ Median

Why Median?

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

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

Shift

if now introduce another data.

$x = \{1, 2, 3, 4, 5, 100\}$ → (outlier) → because it does not belong to the distribution.

$\bar{x}$ (mean) $= \frac{1+2+3+4+5+100}{6} = \frac{115}{6} = 19.1$

Now Median

try the central data in (1) Col → 3

(2) Col $= \frac{3+4}{2} = 3.5$

* Median to find central tendency when outliers present.

Mode

Frequency → Maxm frequency.

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

Maxm frequency of any data: 1 so, Mode = 1

Application of Mean, Median, Mode in Feature Engineering

Age	Weight	Salary	Gender	Degree
24	70	40K	M	BE
25	80	70K	F	-
27	95	45 K	M	-
24	-	80K	-	PhD
32	-	60K	-	Masters
-	60	-	-	Bsc
-	65	55K	M	-
40	72	-	F	-

Numerical Value → We can find mean and fill null values.
↳ If outliers → Median value.

Categorical Value: We can fill it with mode values.

Measure of Dispersion
(Spread of the data)

(1) Variance
(2) Standard deviation.

Variance

Population Variance (σ^2)

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

x_i = Data Points
μ = Population mean
N = Population size

Sample Variance

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

↓
Bessel's correction

x_i
$\bar{x}$ - Sample mean
n = Sample size.

Q Why we divide sample variance by $(n-1)$?

In order to create unbiased estimators / Bessel's correction

Q What is unbiased and biased estimator?

Unbiased: Like expected value is equal to the true value of the parameter

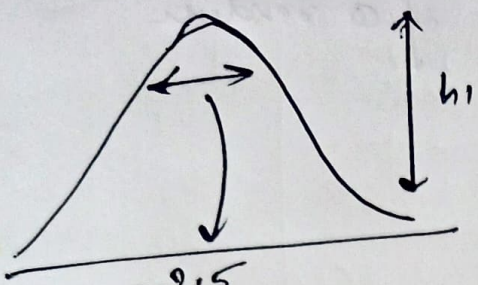
Biased: expected value is not equal to the true value due to bia

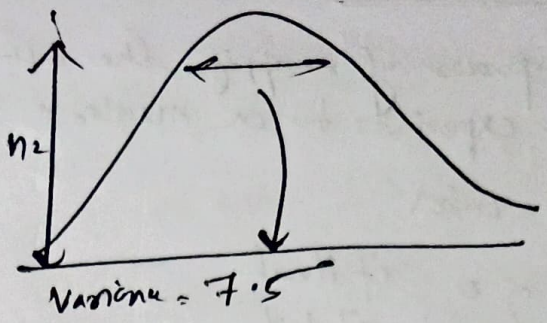
eg: If we pick 20 random students from a school and find avg

Unbiased { there is a good chance our estimator will give avg height true to the actual height

Biased { If we pick up student only from the basketball team where students are typically taller our avg value will not be equal to true value.

Dispersion or Spread (variance)





$(h_2 < h_1)$

Standard Deviation

Population S.D

$$\sigma = \sqrt{\text{variance}} \quad (\sigma^2)$$

Sample S.D

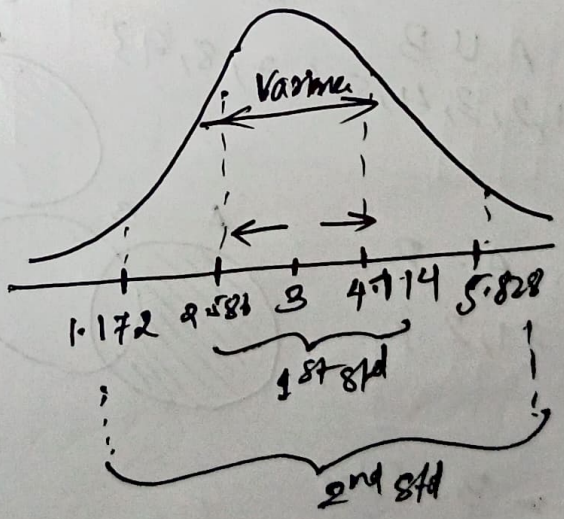
$$\text{Std} = \sqrt{S^2}$$

S^2 → Sample variance

ex: $x = \{1, 2, 3, 4, 5\}$

$\bar{x} = 3$

$\sigma = 1.414$



Random Variable

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

eg:- Tossing a coin

$$x \begin{cases} 0 & \text{if Head} \\ 1 & \text{if tail} \end{cases}$$

(2) Rolling a dice

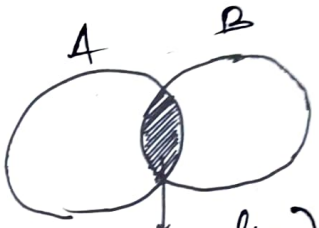
Sets

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

$B = \{3, 4, 5, 6, 7, 8, 9\}$

(1) Intersection: $A \cap B$

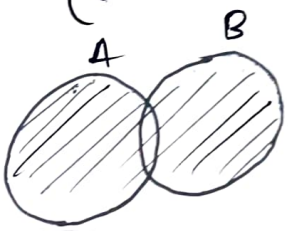
$\{3, 4, 5, 6, 7, 8\}$



(Intersection)

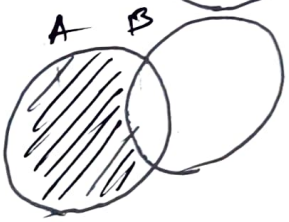
(2) Union: $A \cup B$

$\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$



(3) difference: $A - B$

$\{1, 2\}$



(4) Subset:

$A \rightarrow B$ = False

$B \rightarrow A$ = False

Histogram & Skewness → (Frequency)

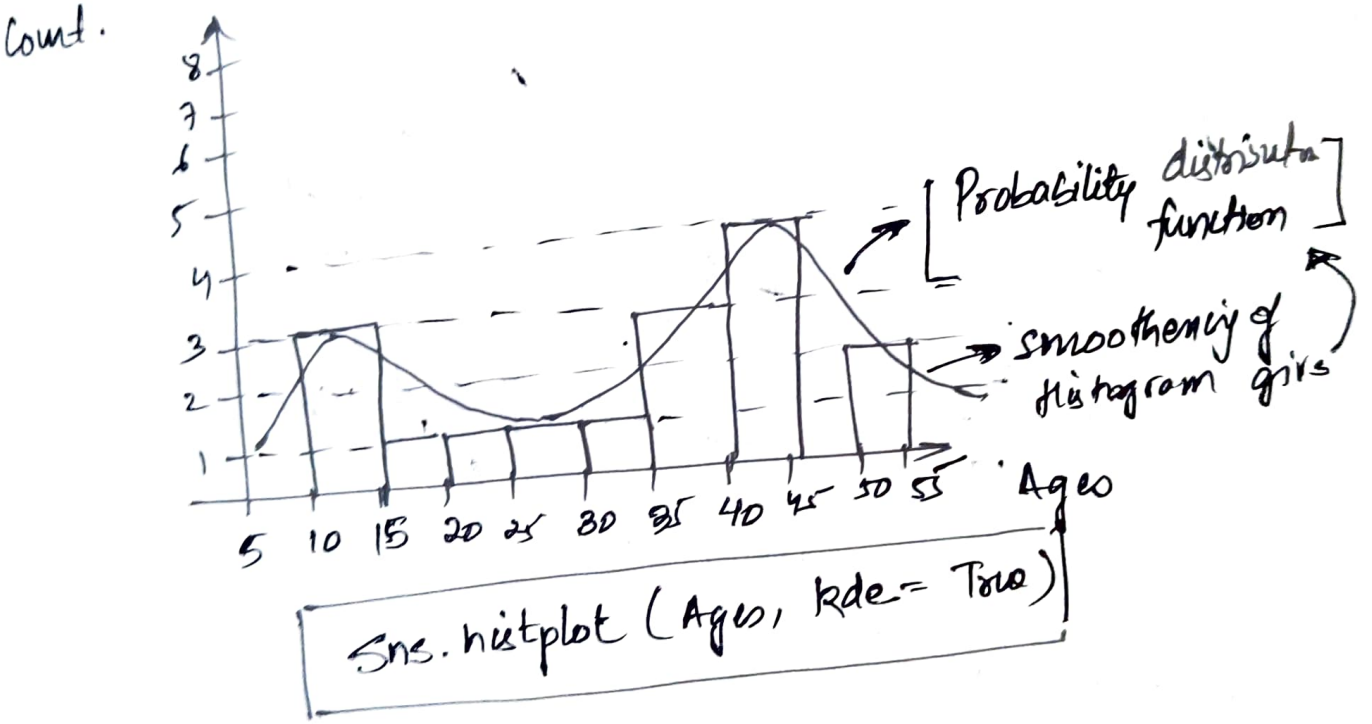
Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51}

Consider, we have 60 values and we count bin of 10 bins.

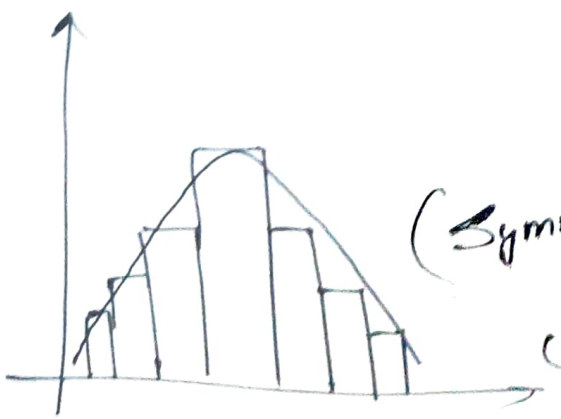
$\frac{60}{10} = \boxed{6}$ → bin size (No of bins = 10)

$\frac{60}{20} = \boxed{3}$ bin size (No of bins = 20)

Histogram



Symmetrical



(Symmetrical Distribution)

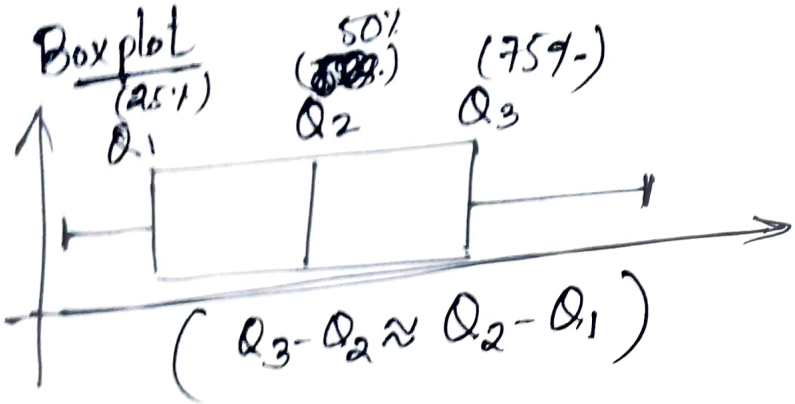
↓

(No skewness)

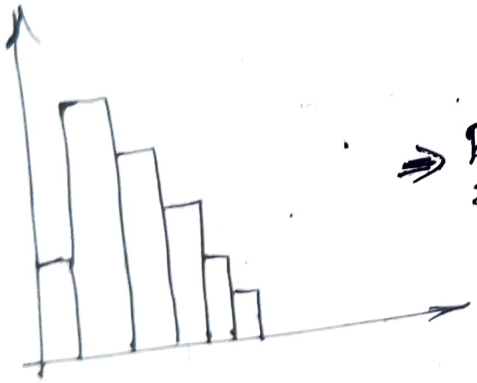
No skewness

① The mean, median, mode all are at perfect centre
→ (Mean = Median = Mode)

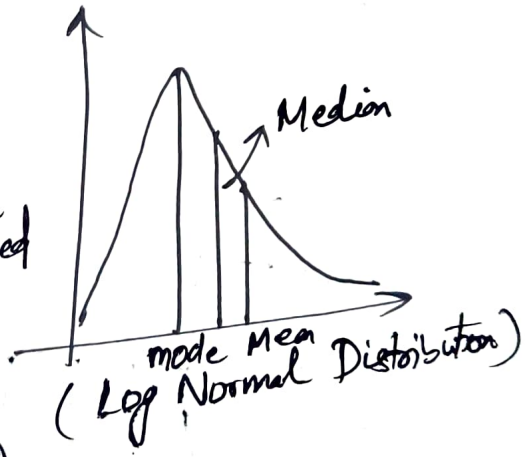
Boxplot



② Right Skewed data

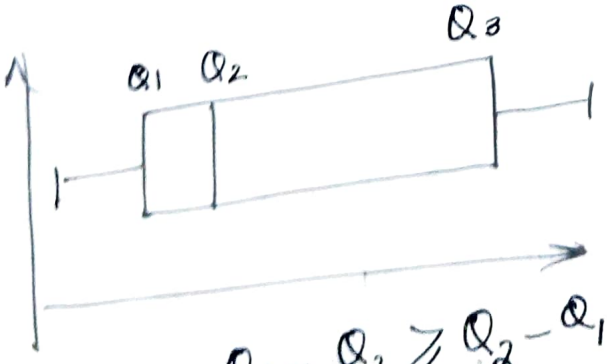


⇒ Positive Skewed

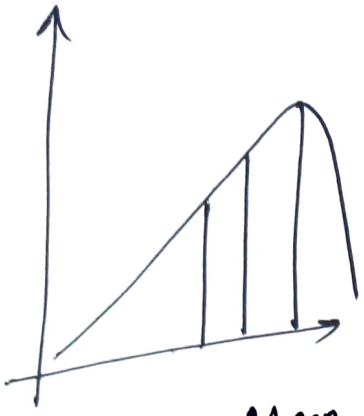


(Relationship b/w mean, median, mode)

Mean > Median > Mode

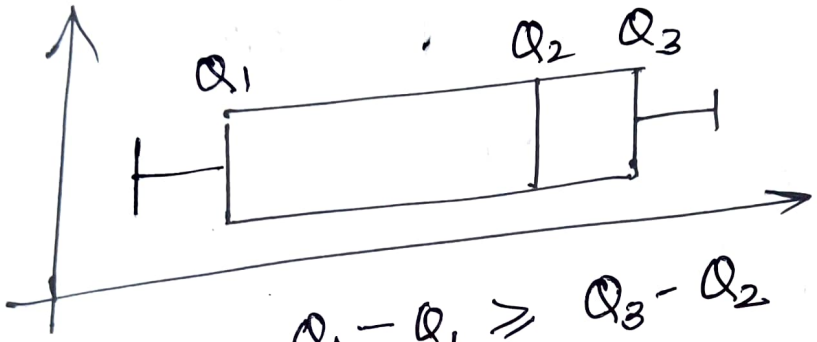


Left skewed distribution



⇒ Negative skewed

Mean ≤ Median ≤ Mode



$Q_2 - Q_1 \geq Q_3 - Q_2$

Covariance & Correlation

[Relationship b/w X and Y]

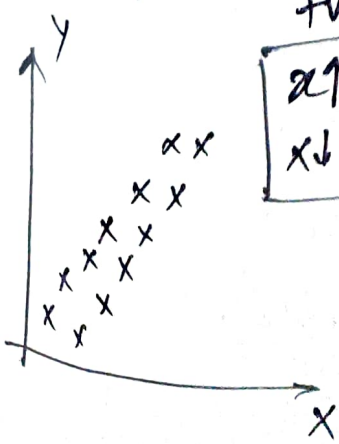
X	Y
2	3
4	5
6	7
8	9

can be

X↑ Y↑
X↓ Y↑
X↓ Y↓
X↑ Y↓

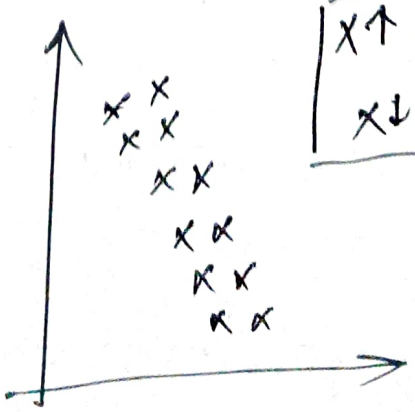
+ve (Cov)

X↑	Y↑
X↓	Y↓



-ve (Cov)

X↑	Y↓
X↓	Y↑



Covariance

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

x_i = data point
$\bar{x}$ = mean (sample)
y_i = data point
$\bar{y}$ = mean (sample)

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

$\Downarrow$

$Cov(X,X)$

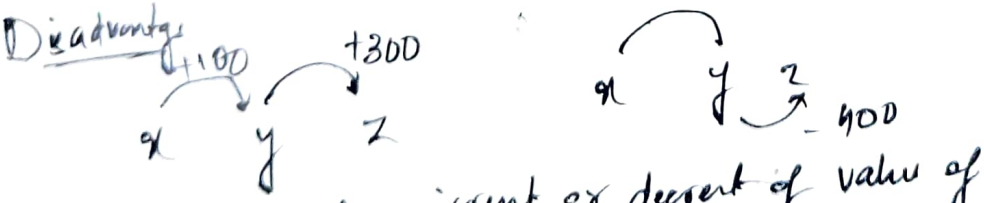
** ($Cov(X,X)$ is basically $var(X)$ which is the spread of data on X.)

Advantages of Covariance

(1) find out the relationship b/w x and y.

(2)

Disadvantage



* No limit on the increase or decrease of value of random variable, so corelation is difficult.
* Covariance does not have limit value.

To fix this we use: Pearson Correlation Coefficient.

Pearson Correlation Coeffi [-1 to 1]

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

σ_x : Std of x
σ_y : Std of y

(for linear relationships preffered) or appox (Normal distribution) → because of this Cov is limited

① The more the value towards +1, the more +vely correlated it is.

② The more the value toward -1, the more -vely correlated it is.

Spearman Rank Correlation [-1 to 1]

$$r_s = \frac{Cov(R(x), R(y))}{\sigma_{R(x)} * \sigma_{R(y)}}$$

R(x) = Rank of x
R(y) = Rank of y.

(preffered)

eg

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

for data Not
① Normally distbtd
② has Outliers
③ Not strictly linear

Implementation (Choose imp features)

① Feature Selection

Size of flat (+ve) ↑ | No of Rooms (+ve) ↑ | Location (+ve) ↑ | No of people stay ≈ 0 (No relation) | Leakage (-ve) | Price ↑

(we can delete feature which do not affect the output)

★ Correlation and Covariance Relationship

Covariance : tells you how two random variables are connected to each other. but does not give you a measure of strength.

Correlation : makes the relationship a standardised relationship which is a simple scale from -1 to 1.
It tells you the strength of a relationship.

Kendall Correlation Coefficient

It also tells the strength and direction, but it maps the value on basis of its agreement or disagreement.

C - Concordant : if most of all pairs agree with each other.

D - Discordant, if most of all the pairs don't agree with each other.

formula

$$T = \frac{C - D}{\frac{1}{2} n(n-1)}$$

If most pairs are in same order = 1
If most pairs are in opposite order = -1
If most pairs are in random = 0

* Use "Pearson" = if you care about the actual values and assume a straight line relationship.
* Use "Spearman" if you care about the ranks of the values
* Use "Kendall" if you want to measure the agreement or disagreement of the order of data points

Syntax

Covariance

df.cov()

Corr

① df.corr(method="pearson", numeric_only=True)
② df.corr(method="spearman", numeric_only=True)
③ df.corr(method="kendall", numeric_only=True)