03-Aug-2025

Agenda :-

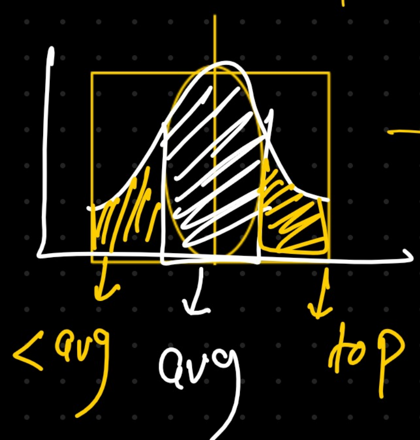
Distribution

correlation & covariance

Skewness

Skewness

measures the asymmetry of a distribution.



→ normal distributed

$n \rightarrow len(df)$

$$\text{Skewness} = \frac{\sum (x_i - \bar{x})^3 / n}{(\sigma)^3}$$

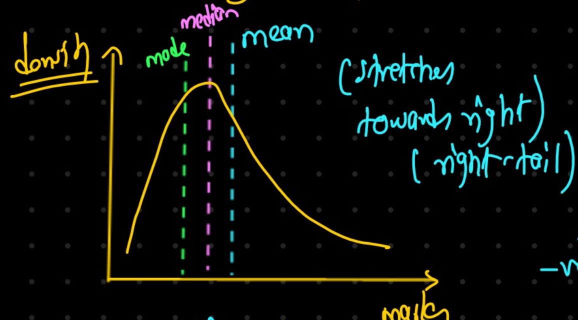
$\sigma \rightarrow$ SD

$\bar{x} \rightarrow$ mean

$x_i \rightarrow$ data point

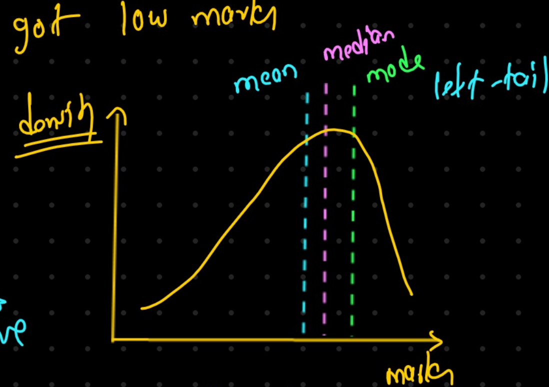
$n \rightarrow$ total size / no. of obser

Q. If most of the students scored low but few got high marks



Positive skew

Q. If most scored high but few got low marks

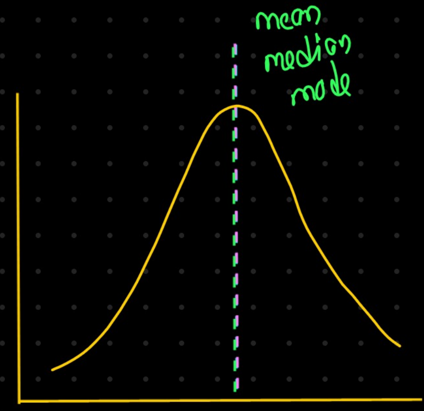


Negative skew

Note:

positive: Mean > Median > Mode

negative: Mean < Median < Mode
skew



* understand Distribution
↓
test
↓
ML (Regression model)

* Hypothesis testing

[] → 6 → $\frac{1}{6}$ → trial → huge number

6 time → 6

Correlation & Covariance

Covariance:

→ Measures how two variable vary together.
→ It tells the direction (positive/Negative relationship)
→ It does not tells you strength.

Formula: —

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

H ↑ | W ↑ ↓

+ve

X → Y

Engine CC ↑ | Mileage ↑ ↓

-ve covariance

Engine CC | Height

no → 0

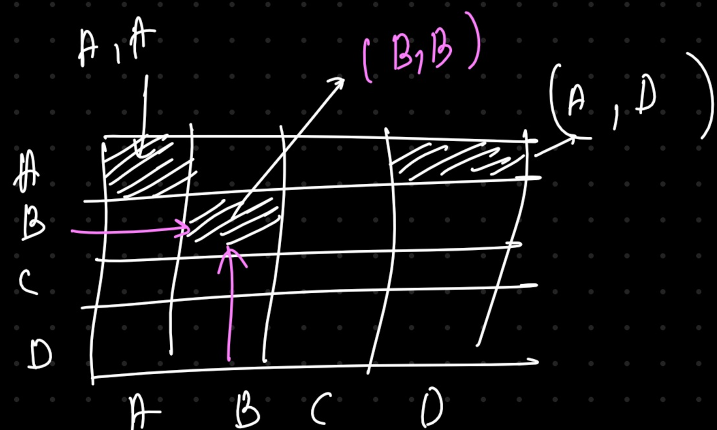
0.001

Correlation:—

- A scaled version of covariance that also tells you the strength & direction of the relationship between variables.
- The value lies between -1 & +1

$$\text{correlation formula}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

A w.r.t A → 1



positive (A, B) ⟶ +ve → ↑ study, ↑ marks

A ↓ study hours

B ↘ higher more

negative (A, B) → -ve ⟶ ↑ time, ↓ decreasing

A ↓ more time on social media

B ↘ lower marks

(Survived, Pclass) → -ve → $\underline{\underline{\text{Survived } \uparrow}}$, $\boxed{\text{Pclass } \downarrow}$

Survived ↓ A

Pclass ↓ B

Pclass → 1, 2, 3

Pclass ↓ (higher the class of the person)

<u>Finding</u> :—

Survided & pclass → -0.34 (Higher class (1st class) passenger were likely to survive.

SibSp & Parch → 0.41 (people with more family aboard (sibling, parent, etc) often had both types of releative together)

SibSp (No. of sibling or spouse abord with the passenger)

Parch (No. of parents or children abord with the passenger)

Survided & Fare → 0.26 (passengers who paid more had "better chance" of survival. (probably in 1st class))

$-\infty$ 187 198 ∞

-1 $\boxed{1}$ $\boxed{1}$

cov → +ve

→ covariance 90 −160 −1000

160

200

→ correlation → -1 to 1

	Covariance	Correlation
Measure	Direction of linear Relationship	Direction & Strength
Scale	Raw unit	Standardized (−1 to +1)
Interpret	positive / Negative / zero	+1 (strong +ve), 0 (none), −1 (strong −ve)
Use	step in calculating correlation	use for feature relationship.

Probability Distributions

(1) Bernoulli Distribution:

coin → H
→ T

email → spam
→ not

sales → purchase
→ Not

match → lose
→ win

Models a single experiment with exactly 2 outcomes:

success (1)

Failure (0)

prob(H) = 0.5 → success of getting Head

p(1−H) = 0.5

success of not getting Hea

coin → $\boxed{H, T}$

A

$P(A) = 0.7$ → success

$p(1-A) = (1-0.7) = 0.3$

0.7

70% prob

Ad →

100 →

P.)

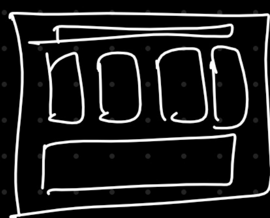
SU → $\boxed{100}$

$\boxed{0.3}$

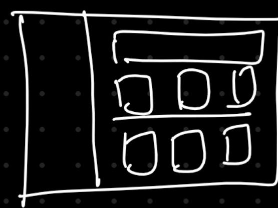
Hypo Testing

A | B Testing

Company A



Type A



Type B

different layout

A, B
→ purchase
→ full working website

→ https://company-amzn.com

Simulating Test distribution

$\boxed{100}$ → data

1000

A → 500 → 1

Checked more product

how many added more item to the ca

1000

B → 500 → 3

10 only

A

B

Binomial distribution:—

Models the number of success (k) in n independent Bernoulli trials.

$H = 0.6$

Bern → coin → $f = 0.4$, succe $= 0.6$

Bin → 10 → success rate

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Q: salesperson has 20% chance of making a sales per call. They make 15 calls.

: what's the prob of exactly 3 sales?

$$P(3) = \binom{15}{3} (0.2)^3 (0.8)^{12} \approx 0.25$$

Assignment :—

Task 1 :

Create 1000 student data with numpy.

Colum A → Student id
Colum B → Height
Colum C → weight

} synthetic →

Task 1.1 → numpy or pandas, manully write formula for covariance & correlation.

Task 1.2 → calculate it.

Task 1.3 → skewness plot for any one column.
→ skewness value.