Agenda

→ covariance & Correlation
→ Probability distⁿ

→ Date

Till now

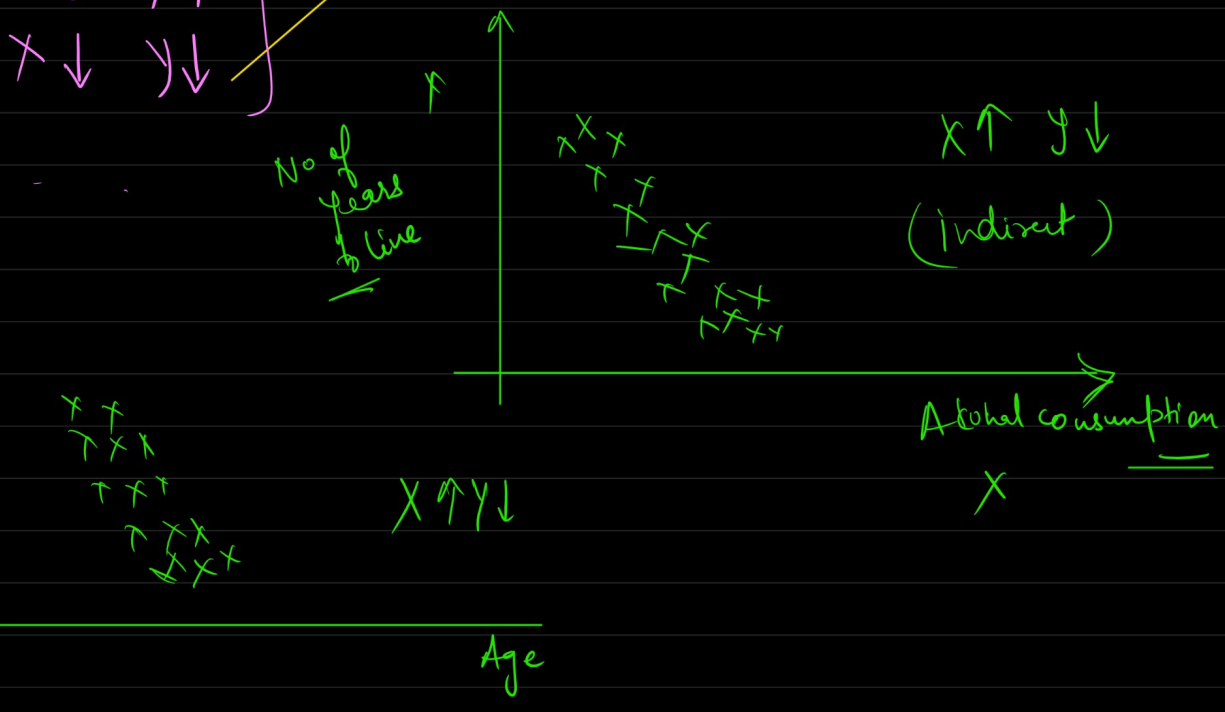
→ Descriptive Stats

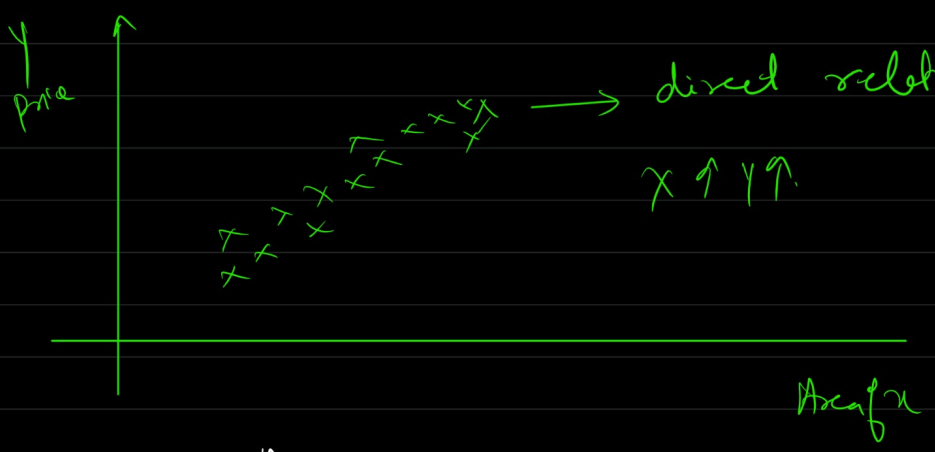
features

	Area of house	No of room	Price of house
→	1000	2	19.
→	—	—	—
→	—	—	—
→	—	—	—
→	—	—	—
→	—	—	—

Area of House ↑ Price of house ↑

X ↑	Y ↑	→ direct relation	
X ↑	Y ↓	→ indirect relation (inverse)	
X ↓	Y ↑		
X ↓	Y ↓		





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

↓

Variance $\Rightarrow \dfrac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1} = \dfrac{\sum_{i=1}^{n}(x-\bar{x})(x-\bar{x})}{n-1}$

Variance was spread of data $\Rightarrow$ relationship of feature with itself

Cov → wrt other feats

Sample

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

* Cov → relationship b/w two variable

$\begin{cases} x\uparrow\ y\uparrow \\ x\downarrow\ y\downarrow \end{cases}$ (+ve Cov)

$x\uparrow\ y\downarrow$
$x\downarrow\ y\uparrow$
(-ve Cov)

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

Sample

x	y
5	6
3	2
2	3
1	5

$\bar{x} = 2.75$ $\quad \bar{y} = \frac{16}{4} = 4$

$$\Rightarrow \frac{(5-2.75)(6-4)+(3-2.75)(2-4)+(2-2.75)(3-4)+(1-2.75)(5-4)}{4-1}$$

$\Rightarrow$ -46.33

↳ X and Y $-ve$

If $+ve$ value $\rightarrow$ $+ve$ relationship

Advantage

→ Relation b/w x & y → $+ve$, $-ve$

* Disadvantage

O X Y

$Cov(X,Y) = 50$

$-\infty$ to ∞

A − B

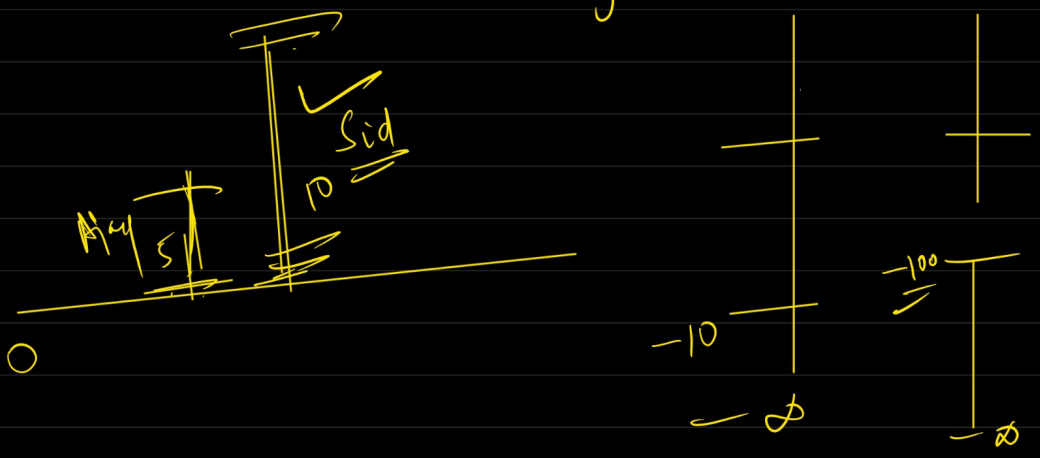
$Cov(A,B) = 100.$

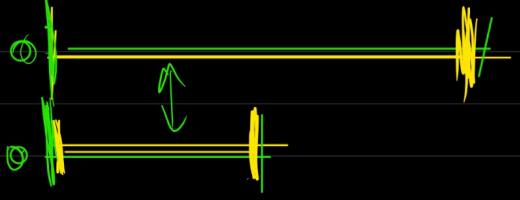
A−B is twice the relationship as compared to X−Y.

→ Covariance only measures the direction.

→ Never talks about strength of relation

→ No any standardise scale to interpret the strength.





Cov → direction of relation

② Covariance has dimension

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

x → height → ft

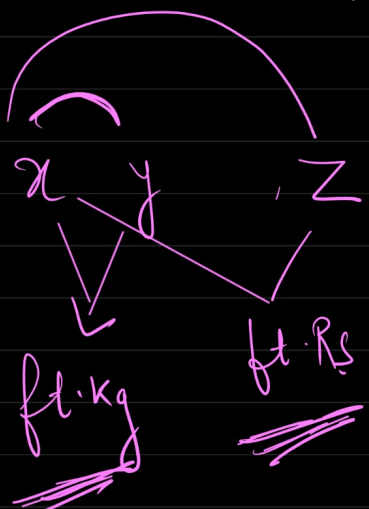
y → wt → kg

$Cov(x,y) \Rightarrow$ ft·kg.

$Cov(\text{transmit}, \text{ht}) = \text{Rs} \cdot \text{ft}$.

transmit → Rs

ht → ft



f1 (Area)	Yield	Price

Not comparable

↳ different dimensions

Soln

→ −1 to 1

→ dimensionless

f1 f2 f3 f4

② Pearson correlation coeff [−1 to 1]

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

−1 0 1

← −ve

+ve →

the more −ve correlation is more on neg side

the more +ve relationship is higher on +ve direction

$\rho_{x,y} = 0.4$, $\rho_{A,B} = 0.8$.

⇒ feature A, B is highly correlated as compared to x, y.

* → Pearson correlation coefficient always measures the linear relationship.

What to do in Non-linear relationship - ??

* Spearman Rank correlation.

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

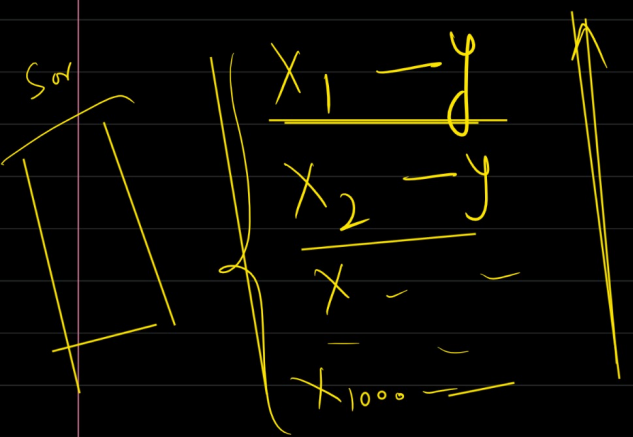
$R(x)$ → Rank of x

$R(y)$ → Rank of y

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

$x \rightarrow 5, 7, 8, 1, 2 \rightarrow 1, 2, 5, 7, 8$ (5th, 4th, 3rd, 2nd, 1st)

x_1 x_2 x_3 Temp — Castle — Delhi — — x_{1000} → Price



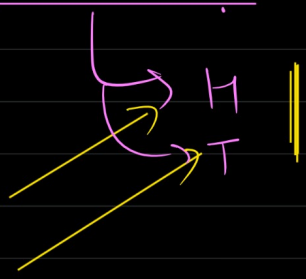
+ve — −0.8

−ve — +0.8

* Random Variables

↳ A set of Possible values from a random experiment.

→ Tossing a coin



Experiment is random.

↓

Outcomes is random

$$X = \begin{cases} 0 :- \text{tail} \\ 1 :- \text{Head} \end{cases}$$

↑ random Variable

↳ quantifies the experiment

$X = \{1, 0\}$

$P(X) = \frac{1}{n}$ (where n is total possible outcomes)

* dice → 1, 2, 3, 4, 5, 6

$\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ . . . $\frac{1}{6}$

$\boxed{P(X) = \frac{1}{n}}$ (where $n = 6$)

$P(X=1) = \frac{1}{6}$

$P(X=2) = 1/6$ ...

Dice — 1, 2, 3, 4, 5, 6.

Outcome

Outcome

Discrete

→ tossing coin/dice

↓

discrete outcomes

↓

Probability mass function (pmf)

Continuous

↓

Ht of students in class

160 – 165 cm

⇓

Continuous outcome

⇓

(pdf) probability density function

Discrete

Dice → Discrete outcomes.

$p(1) = 1/6$

$p(2) = 1/6$

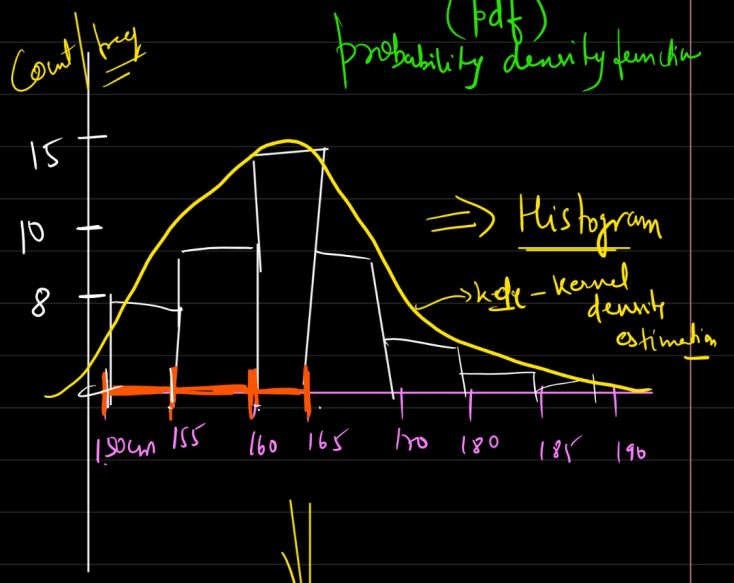
$P(3) = 1/6$

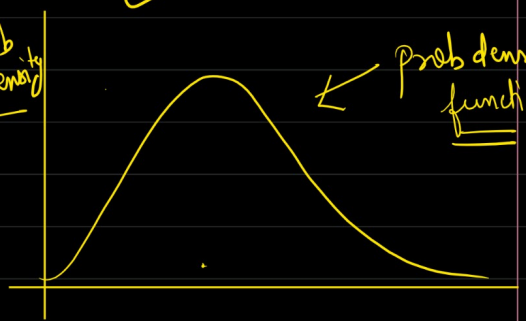
$P(4) = 1/6$

$P(5) = 1/6$

$P(6) = 1/6$

Pmf





Categorical data — Bar

Continuous — Historical

Discrete — 1, 2, 3, 4 (which takes finite value)

Continuous → ht of stu — 1, 2
150 cm,
150.001 cm
100 Kg
100 Kg 2 parts

Two types of Experiment

- discrete → Prob mass function
- Continuous → Prob density function

Cumulative dist^n function

X	P(X)	cdf
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6 = 1

cdf → 0 to 1

Prob dist fn

Prob density fn

→ Normal / Gaussian
→ Standard Normal dist
→ log Normal dist
→ Chi-Square dist
→ F-distn.

Prob mass fn.

→ Bernoulli
→ Binomial
→ Poisson

Uniform dist → discrete
Uniform dist → Cont. Prob

* Discrete Uniform distn

→ Outcomes will be discrete
→ Outcomes are equally likely.

Uniform dist

Discrete (pmf)

Continuous (pdf)

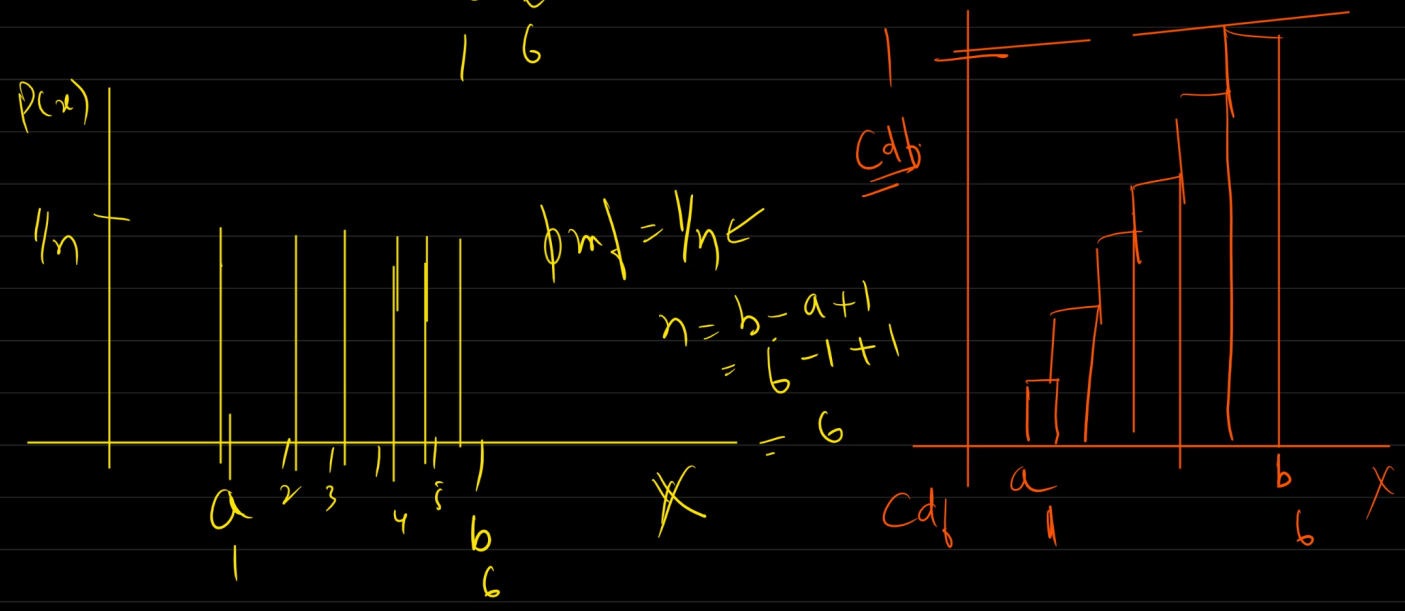
→ Outcomes are discrete & equally likely

equally likely ⇒ Same/Equally probability

eg rolling a dice

$P(1) = 1/6$
$P(2) = 1/6$
$\vdots$
$P(6) = 1/6$

$U(a,b)$ | Unif $\{a, b\}$



Q what is prob of getting 3 when you throw a dice?

→ 1/6

* mean of discrete Uniform dist^n $\Rightarrow \frac{a+b}{2}$ ✓

Variance " " $\Rightarrow \frac{n^2-1}{12}$ ✓

② Bernoulli dist^n

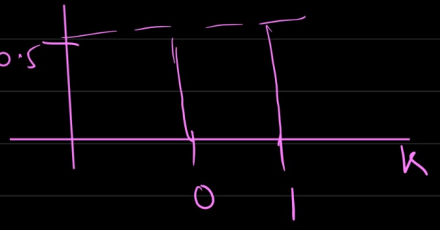
→ discrete

→ Outcomes — 2 → Pass | fail

→ Tossing a coin.

$$P(X=k) = \begin{cases} p & \text{if } k=1 \\ 1-p & \text{if } k=0 \end{cases}$$

$$\boxed{P(X=k) = p^k (1-p)^{1-k}}$$



$$P_{success} + P_{failure} = 1$$

$$P_{Head} + P_{tail} = 1$$

$$0.5 + 0.5 = 1$$

$$P(x=k) = p^k (1-p)^{1-k}$$

$k = 1$

$$p^1 (1-p)^{1-1}$$

$$p^1 (1-p)^0$$

$$p \cdot 1$$

$$p$$

$k = 0$

$$p^0 (1-p)^{1-0}$$

$$1 \cdot (1-p)$$

$$1-p$$

$$p + (1-p) = 1$$

* Conditions of Bernoulli dist
  - → No of trials should be finite
  - → Each trial should be Independent.
  - → only two possible outcomes
  - → Prob of each outcome should be same in every trial.

Example:

tossing a coin

fraud / not

Six or no

Pass / fail

Q. Bumrah bowls 6 balls at wicket with prob of 0.6 of hitting the stump with each ball. What is prob not hitting wicket

→ $p(H) = 0.6$

$p(\text{Not } H) = 1 - 0.6 = 0.4$

* mean $= p$
  variance $= p(1-p)$

* Bi-nomial dist

bi – two → bernoulli

n – bernoulli → Binomial

* Binomial dist^n is n Bernoulli trial

pmf of Bernoulli $= p^k (1-p)^{1-k}$

pmf of Binomial dist $= {}^nC_k \, p^k (1-p)^{n-k}$

1st toss 2nd toss 3rd toss 4th .. 10th

* What is prob of getting k heads out of n trial

$\binom{n}{k} = \frac{\lfloor n}{\lfloor n-k \, \lfloor k}$

n → total no. of trail
k – no of events

$\lfloor 4 = 4 \times 3 \times 2 \times 1$

$\lfloor 10 = 10 \times 9 \times 8 - - - - \times 1$

Q With 3 tosses what is prob of getting exactly 2 heads?

$$n=3, \quad k=2$$

$$b(x=2) = {}^{n}C_{k}\, p^{k} (1-p)^{n-k}$$

$$= {}^{3}C_{2} (0.5)^{2} (0.5)^{3-2}$$

$$= \frac{\lfloor 3}{\lfloor 3-2 \cdot \lfloor 2} \cdot (0.5)^{2} \cdot 0.5$$

$$= \frac{3 \times 2}{1 \times 2} (0.5)^{3} = 0.375$$

mean $= np$

variance $= np(1-p)$

* Poisson distribution

→ discrete Uniforms.

→ describes the no. of events that occur within a fixed interval of time, given average rate of occurence.

→ No of events occuring in a fixed time interval.

$$pmf \Rightarrow P(X=x) = \frac{e^{-\lambda} \cdot \lambda^{x}}{x!}$$

$e = 2.718$

λ – avg rate of events every interval.

- ex No of calls received by a customer care every hour
- ex No of People visiting hospital / temple
- ex No of accidents
- ex No of emails received

Q The avg no of customers entering a store in an hour is 5. What is the prob of exactly 3 customers will enter the store next hour?

$\lambda = 5$

$P(X=3)$

$$\frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{e^{-5}\cdot 5^{3}}{3!}$$

$$= \frac{(2.718)^{-5} \times 125}{6}$$

$$\simeq 0.14$$

Mean $= \lambda t$

Varian $= \lambda t$