PROBABILITY DISTRIBUTION

- Random variable → A set of possible values from a random experiment.

set of random variable ← $X = \{H, T\}$

$P(X = H) = \frac{1}{2}$ in ~~a~~ Coin case

- A random variable value is unknown

Types of distributions

Probability distribution function

Probability Density function
- Normal distribution
- Standard Normal
- Log Normal
- Chi-Square
- F-distribution

Probability Mass function
- Bernoulli
- Binomial
- Poission

} continous distribution is of both types.

Probability Distributions

- eg → tossing a coin, $X = \{H, T\}$
  ↳ set of random variable
- A function that asign value to each of experiment outcome.

→ $f(x) = \frac{1}{n}$ { general fun. that tells the probability of any random value.

n → no. of outcomes

outcomes of experiment

- tossing a coin
- throwing a dice

↓

gives discrete outcomes

↓

function that capture the probability for these outcomes

↓

Probability Mass function

- calculate prob of student height b/w 160 → 170 cm

↓

Continous Outcome

↓

function for that

↓

Probability density function.

irrespective of outcome nature, if you can draw the outcomes in a form of distribution

↓

Probability distribution function

Type of experiment

① Discrete

↓

Probability mass fun.

② Continous

↓

Prob. density fun.

① Probability Mass function

↳ distribution of discrete random variable.

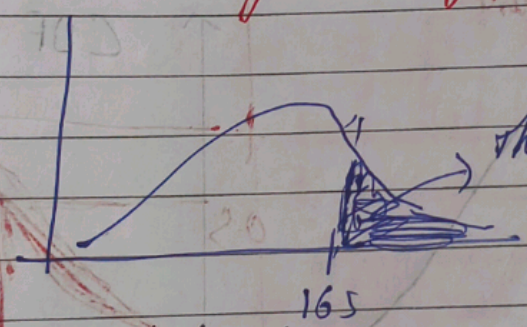
eg. $p(x \geq 3) = P(x=3) + P(x=4) + P(x=5) + P(x=6)$

$= \frac{4}{6} = 0.66$

② Probability Density function

↳ distribution of continous random variable

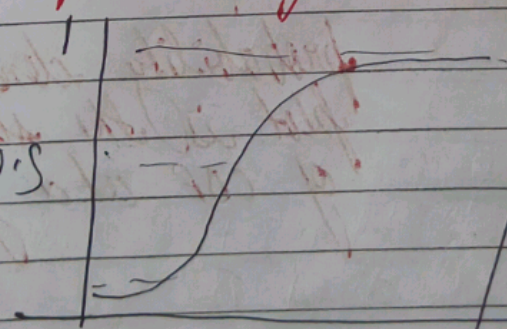
$P(x \geq 165cm) \Rightarrow$ Area under the curve for the region



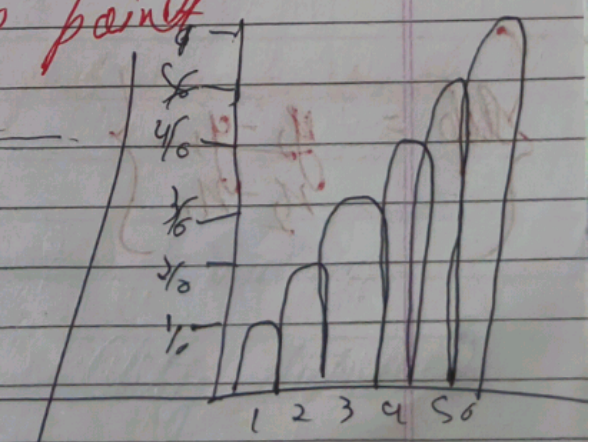
this area tells prob.

③ Cumilative dist. fun.

↳ it is summasion of all probability possible upto a given point



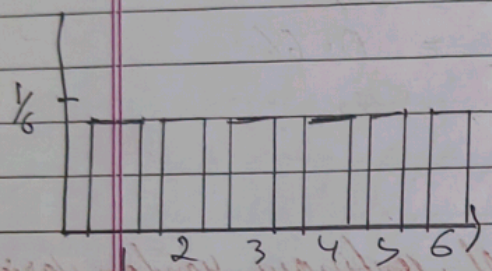
for PDF

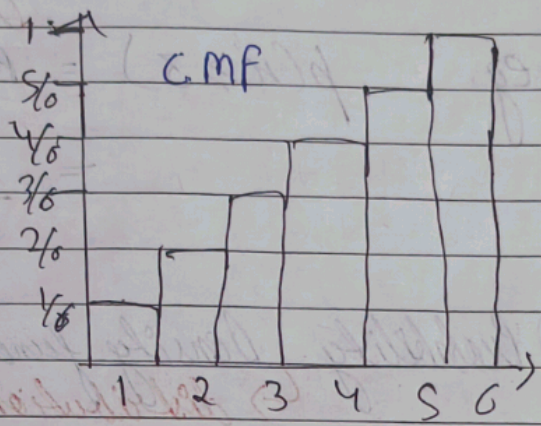


for PMF

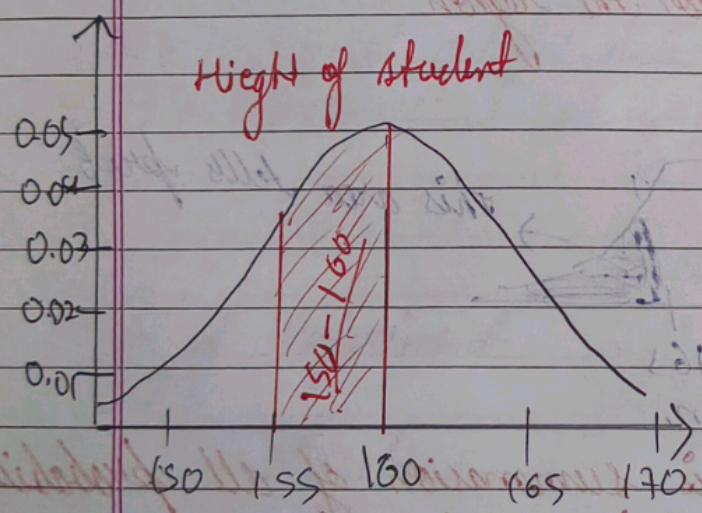
Probability distribution function & with CDF

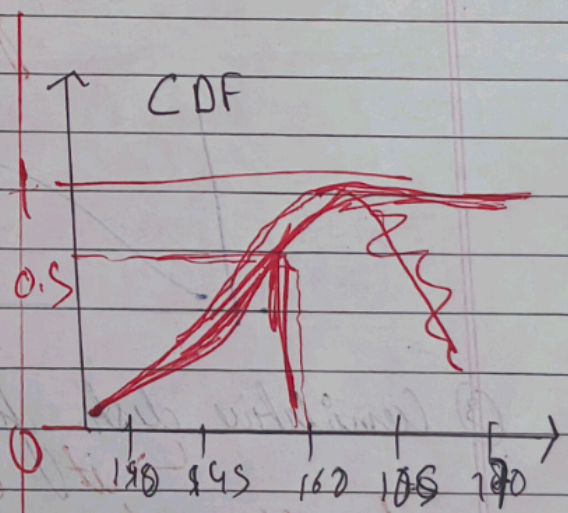
- PMF → Probability Mass function
  → Discrete random Variable





2 PDF → Random Variable is continous





$$\left\{ \text{slop} = \frac{y_2 - y_1}{x_2 - x_1} \right\}$$

probability density of a pdf plot is the slope (gradient) of CDF at a given point

Can probability density > 1 ? → yes.

- Uniform Distribution → it refers to a type of probability distribution in which outcomes are equally likely.

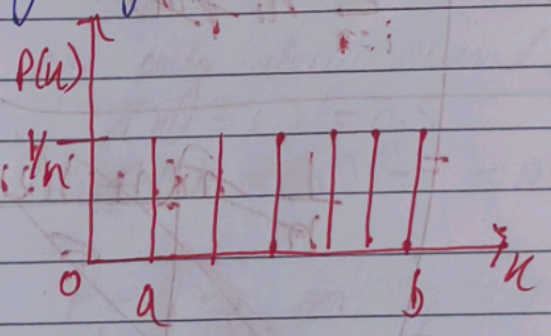
$\mu = (a+b)/2$ → mean

$\sigma^2 = \frac{n^2-1}{12}$

Types → Discrete UD
Continous UD

(1) In Discrete Uniform Dist, the outcomes are discrete and have the same probability.

Notation of uniform dist. ⇒ $U(a,b)$ // $unif(a,b)$



Expected value ⇒ in long run (EV) = mean value

$$E_p = \sum_{i=1}^{n} x_i (p(x_i))$$

x_i → random variable

$p(x_i)$ → prob. of random variable

Note → • Mean is used for frequency distribution

• Expected value is used for probability distribution

• Variance = $E(x)^2 - (E(x))^2$

$E(x) \Rightarrow \sum_{i=1}^{n} x_i \, p(x_i) \Rightarrow 1/n$

$\Rightarrow \sum_{i=1}^{n} \frac{1}{n} \Rightarrow (1+2+3 \ldots n)$

$\Rightarrow \frac{1}{N} \frac{N(N+1)}{2}$

$= \frac{N+1}{2}$

Sum of first n no $\rightarrow \frac{n(n+1)}{2}$

Sum of first n^2 no.

$= \sum_{i=1}^{n} x^2$

$= \frac{n(n+1)(2n+1)}{6}$

$E(x)^2 = \sum_{i=1}^{n} x_i^2 \, p(x_i) \Rightarrow 1/n \Rightarrow \frac{(n+1)(2n+1)}{6}$

$= \frac{1}{n} \frac{n(n+1)(2n+1)}{6}$

$= \frac{n+1}{2} - \frac{(n+1)(2n+1)}{6}$

$\Rightarrow \frac{6n+6 - 2(n+1)(2n+1)}{12}$

$\Rightarrow 2(3(n+1) - (n+1)(2n+1))$

$Var = = \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{6}\right)^2$

$= \frac{n(2n+1) + 1(2n+1)}{6} - \frac{N^2+2N+1}{4}$

$$Var = \frac{n^2-1}{12} \qquad \sigma = \sqrt{\frac{n^2-1}{12}}$$

* Bernoulli Distribution

↳ A discrete prob. distribution of a random variable which takes only two possible outcomes, typically labled as success & failure. which is fixed prob of success (1) (0) & failure

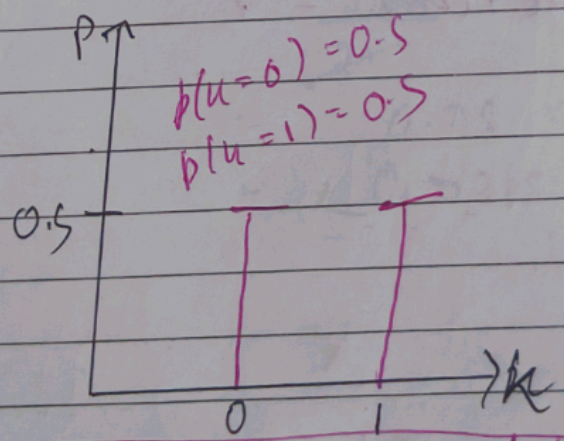
i.e → sum (f, s) = 1 of prob.

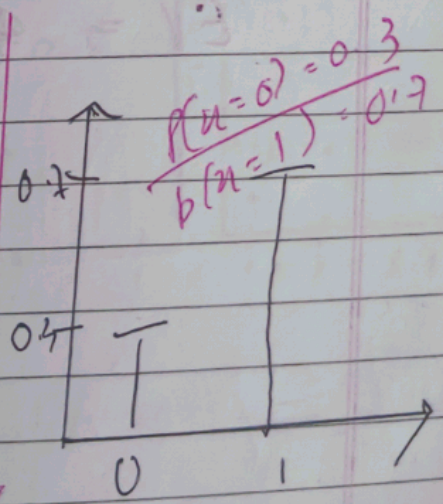
Since there are only two event (p, q)

So if $P(x = 1) = 0.5$

then $P(x = 0) = 1 - 0.5 = 0.5$

$$P = 1 - q \text{ or } q = 1 - P$$





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

* condition of bernouli distribution

1 -> No of trial should be finite

2 -> each trial should be independent

3 -> there must be only two possible outcomes.

4 -> prob of each output should be same in every trial.

ex -> • tossing a coin.
• Someone has commited a froud or not.
• pass / fail

• Mean of Bernouli => P

• Variance => $P(1-P)$

$EV = \sum x P(x) = P$

Binomial Distribution

Bi → Two outcomes

Binomial distribution is the 'n' Bernoulli trial

Pmf of bernoulli ⇒ $Pmf = p^k (1-p)^{1-k}$ because it is 1 trial $n=1$

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

{ mean = np , variance = npq

Eg. prob. of getting exactly 2 heads in 3 trial

$n = 3$, $k = 2$

$P(k=2) = {}^nC_k \, p^k (1-p)^{n-k}$

$= {}^3C_2 \, p^2 (1-s)^{3-2}$

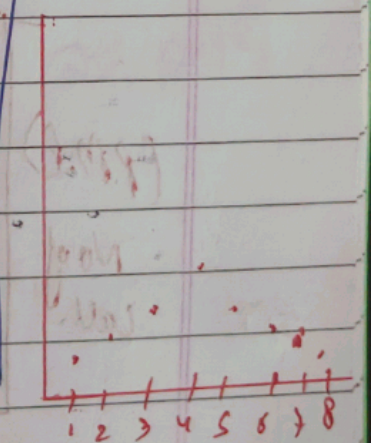
$= 3C_2 (0.5)^2 \times (0.5)^1$

$= \frac{3 \times 2!}{2!} \times 0.25 \times 0.5$

$= 0.75 \times 0.5$

$= 0.375$

if $n = 10$
$k = 3$



Possion Distribution

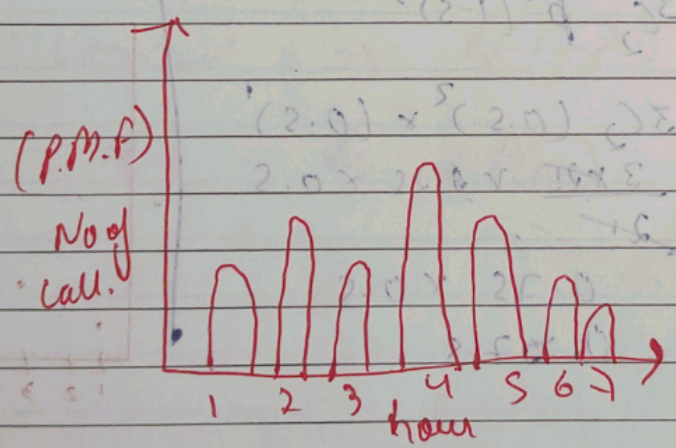
The ~~pos~~ poisson distribution is a discrete probability distribution that describe the no. of event that occur within a fixed interval of time or space given a known average rate of occurance

* No of ~~to~~ ~~that~~ events occuring within a fixed time interval

eg. → Customer Care

λ → expected no. of call (event) to occur every time interval

$\lambda = 100$



$$PMF = p(x = x) = \frac{e^{-\lambda} \cdot \lambda^{x}}{x!}$$

λ = avg. rate
$e = 2.71828$

Q => Person call at 5th hour.

$$P(x=5) = \frac{e^{-\lambda} \cdot \lambda^{x}}{x!}$$

$$= \frac{(2.718)^{-10} \cdot 10^{5}}{5!}$$

$$= \boxed{\frac{10^{5}}{5! \times (2.718)^{10}}}$$

Q = Avg no of customer entering a store is 5. What is the prob. of exactly 3 customer will enter the store next hour.

$\lambda = 5$

$$p(x=3) = \frac{e^{-\lambda} \lambda^{x}}{x!} = \frac{(2.718)^{-5} \, 125}{3!}$$

$$= \frac{0.00674 \times 125}{6} \approx 0.14$$

Continous Uniform Distribution

Uniform distribution is a type of distribution in which outcomes are equally likely.

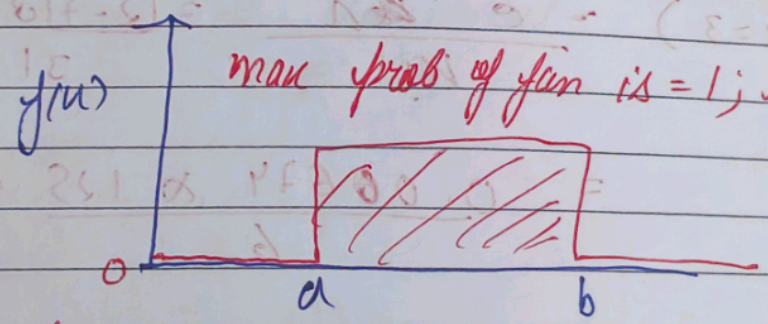
A continous uniform probability distribution is a distribution in which it has infinite no. of values defined in specified range/bound.

- random variable is continous.
- rectangular distribution.

Notation = $U(a,b)$

Parameter = $-\infty < a < b > \infty$

$b >> a$; a is min & b is max



max prob of fun is = 1; the area = 1

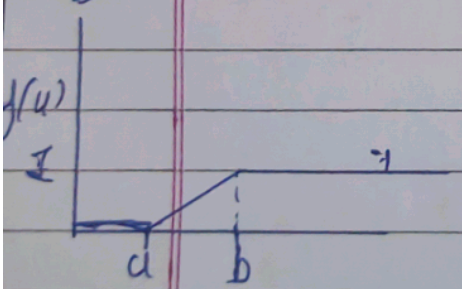
Prob = Area of rectangle = 1

$= d \times b = 1$

$= (b-a) \times f(x) = 1$

$f(x) = \frac{1}{b-a}$

CDF



$$cdf = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & x \in [a,b] \\ 1 & \text{for } x > b \end{cases}$$

mean / median of Uniform Continous dist $= \frac{1}{2}(a+b)$

Variance $= \frac{1}{12}(b-a)^2$

Q1-> The no. of items sold at a shop daily is uniformly distributed with max & min items sold. 50 & 20 resp.

(1) Prob of daily sales to fall 25 to 40?

$\cancel{f(u) = \frac{1}{40-25} = \frac{1}{15} = 6.66}$

$\cancel{= 15 \times \frac{1}{30} = 0.5}$

$= b \times h$

$= b-a \times f(u)$

$= b-a \times \frac{1}{50-20} = 15 \times \frac{1}{30}$

$= 0.5$

(2) prob. of sales ≥ 30

$P(x \geq 30) = 50-30 \times \frac{1}{50-20}$

$= 20 \times \frac{1}{30} = 66.66$

Q-) The amount of time for pizza delivery is uniformly distributed b/w 15 & 60 min.

What is std deviation of amount of time it take for pizza to be delivered

$$V = \frac{1}{12}(b-a)^2 \quad \frac{1}{12}(45)^2$$

$$V^2 = \frac{(45)^2}{12} \Big| \sigma = \frac{45}{2\sqrt{3}}$$

$$\sigma = \frac{22.5}{\sqrt{3}} = 12.99 \approx 13$$

Proof of mean/expected value of a variance

$$E(x) = \int_a^b x \cdot f(x)\, dx$$

$$= \int_a^b x \frac{1}{b-a}\, dx = \frac{1}{b-a} \int_a^b x\, dx$$

$$= \frac{1}{b-a} \times \left[\frac{x^2}{2}\right]_a^b$$

$$= \frac{1}{b-a} \frac{b^2-a^2}{2} = \frac{1}{2} \frac{(a+b)\cancel{(a-b)}}{\cancel{(b-a)}}$$

$$\boxed{E(x) = \frac{1}{2}(a+b)}$$

$$Var = E(x)^2 - (E(x))^2$$

$$= \int_a^b x^2 \cdot f(x)\, dx - \frac{1}{(2)^2}(a+b)^2$$

$$= \frac{1}{b-a} \left[\frac{x^3}{3}\right]_a^b \qquad (b^3-a^3) = (b-a)(b^2+ab+a^2)$$

$$= \frac{1}{\cancel{b-a}} \frac{\cancel{(b-a)}(b^2+ab-a^2)}{3}$$

$$= \frac{b^2+ab+a^2}{3} - \frac{a^2+b^2+2ab}{4}$$

$$= \frac{b^2+ab+a^2}{3} - \frac{a^2+b^2+2b}{4}$$

$$= \frac{1}{12}\left(4(b^2+ab+a^2) - 3(a^2+b^2+2ab)\right)$$

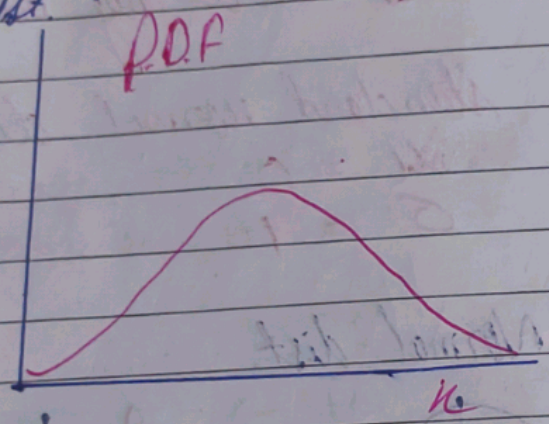
$$= \frac{1}{12}\left(\overset{3^2}{\cancel{4b^2}} + \cancel{4ab} + \overset{a^2}{\cancel{4a^2}} - \cancel{3a^2} - 3b^2 - \overset{2ab}{\cancel{6ab}}\right)$$

$$= \frac{1}{12}\left(b^2 + a^2 - 2ab\right)$$

$$= \boxed{\frac{1}{12}(b-a)^2}$$

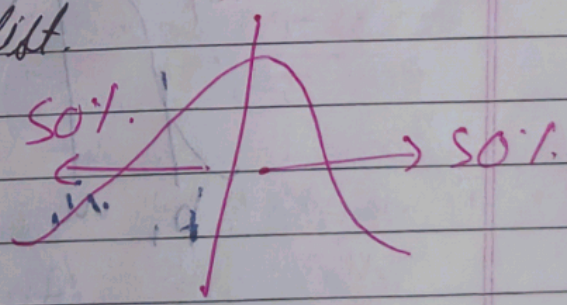
Normal / gaussian Distribution

- type of ~~contit~~ continous prob. dist
- also called as Bell-shaped data.
- Most of data in real world follows normal dist.



Eg → ① mark of student
② height of peoples
③ blood pressure
etc.

characterstic of Normal dist.



- Symmetrial around mean
- mean = mode = median
- Skewness = 0
- follows emperial rule → it say that

68% → 1st std. dev away from mean
95 → 2nd " "
99.7 → 3rd std dev

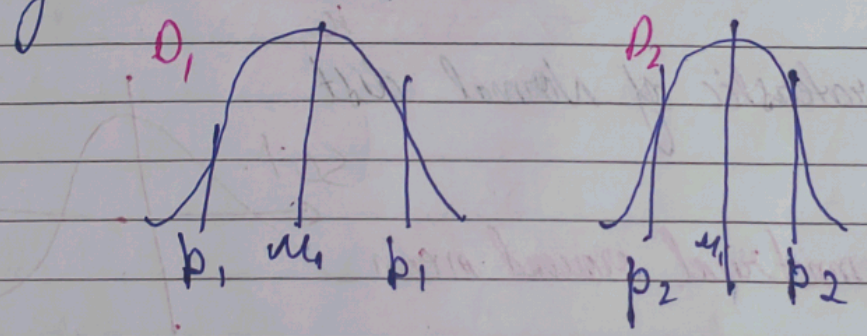
Standard normal distribution

→ it is also continous prob. dist
→ it is a special case of normal dis

Where in normal dist
μ & σ can be any value

but in standard normal dist
$\mu = 0$
$\sigma = 1$

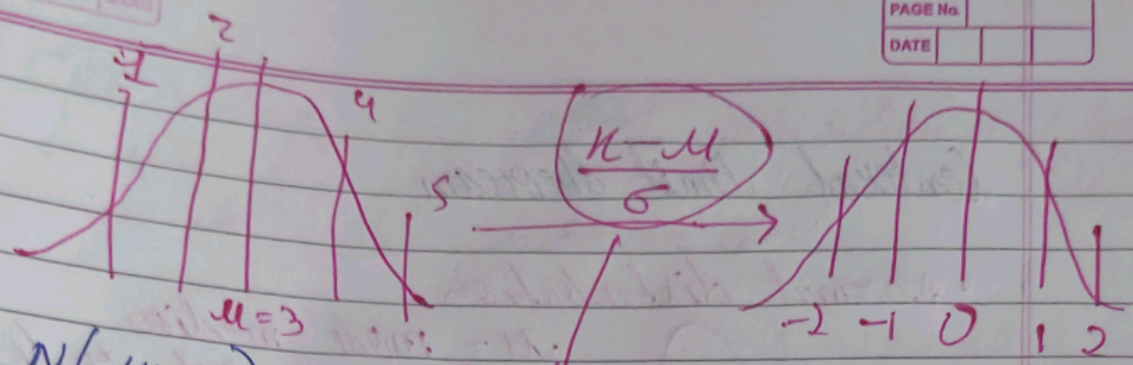
Why Std. Normal dist.



from this can i say p_1 is closer to μ_1 in D_1 compared to p_2 of $D_2(\mu_2)$

* you can not compare science student steam topper to Arts stream topper.

so we changes ND ⟶ SND



$N(\mu, \sigma)$

it is also called a Z score

$SN(\mu=0, \sigma=1)$

- Use Case
  - many ML algo like (linear regression, clustering, Logistic) requires scaling
    ⇓
    SND ⇒ faster scaling

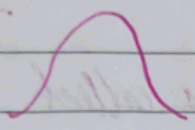
Standardization = Z score = $\frac{x-\mu}{\sigma}$

No. of food	price	distance from city	price in loc
1	10000	20	70
2	13000	15	80
3	14000	30	85
2	11000	30	75
1	⋮	⋮	⋮
4	⋮	⋮	⋮

in above data all value are in different standard. so using <u>SND</u> (zscore) it becomes easy.

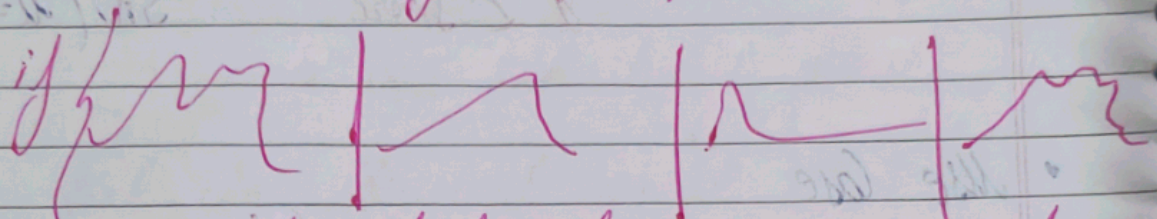
Centeral limit theoream

- Normal distribution.



μ = mode = median
- follows emperial rule

- if



these are the distribution. How we get info about these distribution

⇓

CLT comes → it say that if we have a population with (μ) mean, (σ) std dev and take sufficiently large no. of random sample from population with replacement, then distribution of sample mean will be normaly distributed

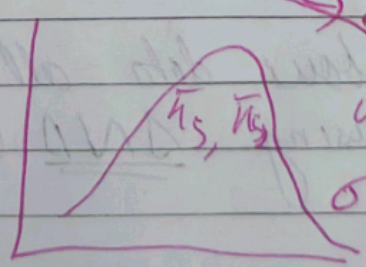
population's

→ The sampling mean (μ, σ) of a population will approximately be normal distribution (μ, $\frac{\sigma}{\sqrt{n}}$)

sample's

μ
σ

→ $S_1(SO) \rightarrow \bar{x}_{s_1}$
$S_2(SO) \rightarrow \bar{x}_{s_2}$
⋮
$S_n(SO)$ $\bar{x}_{s_3}$



$\mu = \mu$

$\sigma = \frac{\sigma}{\sqrt{n}}$

sample size

CLT → irrespective of the distribution of population the distribution of sampling mean will always be normal $\begin{cases} \mu = \mu \\ \sigma = \frac{\sigma}{\sqrt{n}} \end{cases}$ → sample size

- Condition of CLT
  - no of samples should be sufficiently large
  - sample size ($n > 30$) except the population which is already normal.

std error $= \frac{\sigma}{\sqrt{n}}$ $\{\uparrow n, \downarrow SE\}$

Q → you have pop. with $\mu = 100$ & $\sigma = 20$. if you have sample size $n = 50$ from this population what is prob. that sample mean will be less than 105

$\mu = 100$, $\sigma = 20$, $n = 50$

$\bar{x} = 105$

$$Z\text{score} = \frac{105 - 100}{\frac{20}{\sqrt{50}}} = \frac{5}{\frac{20}{4}} = 0.25 \times \sqrt{50}$$

$$= 1.76$$

$$= \cancel{1.76} \approx 0.89$$

- Application Z test.

$$Z = \frac{\bar{x} - \mu}{\sigma}$$

Q → $N(\mu = 50, \sigma = 20, D = 110)$

How many std div D is away from mean

$$Z = \frac{x_i - \mu}{\sigma} = \frac{110 - 50}{20} = \frac{60}{20}$$

$$\boxed{= 3\sigma}$$

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

$\mu = 4, \quad \sigma = 1$

what is score of getting no. greater that 4.5

$$Z_{score} = \frac{4.5 - 4}{1} = 0.5 \simeq 0.6915$$

4.5 is 0.5σ away from mean

for each Z_{score} → we have prob. value (area under curve)

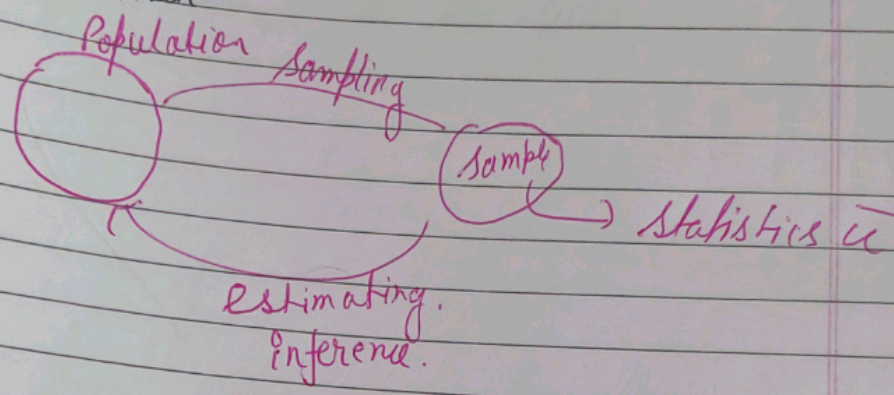
↓

Z Table Pdf

for more than 4.5 $= 1 - 0.6915$

$= 0.31$

• Point estimate



estimate → A specified observed value used to estimate an unknown population parameter using sample

1 → point estimate → A single that is used to estimate true value of population parameter.

2 → Interval estimate → A range of values used to estimate unknown population parameter.