* Probability mass fn
  - Discrete Ud.
    - Bernoulli
    - Binomial
    - Poisson
    - Geometric

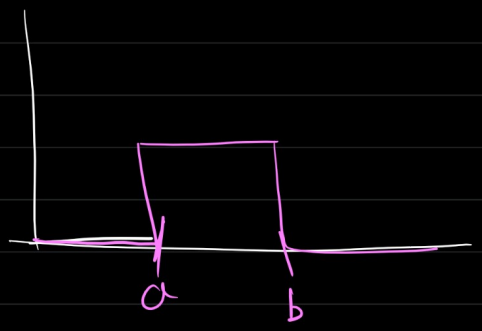
* Probdensity fn →
  - Continuous Ud.
  - Normal dist
  - Standard Normal
  - Chi-sq
  - F dist

→ Hypothesis testing

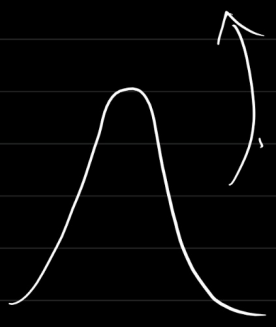
* Continuous Uniform distn

$$\text{mean} = \frac{1}{2}(a+b)$$

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



* Normal distn
  - Gaussian distn
  - Bell shaped distn.

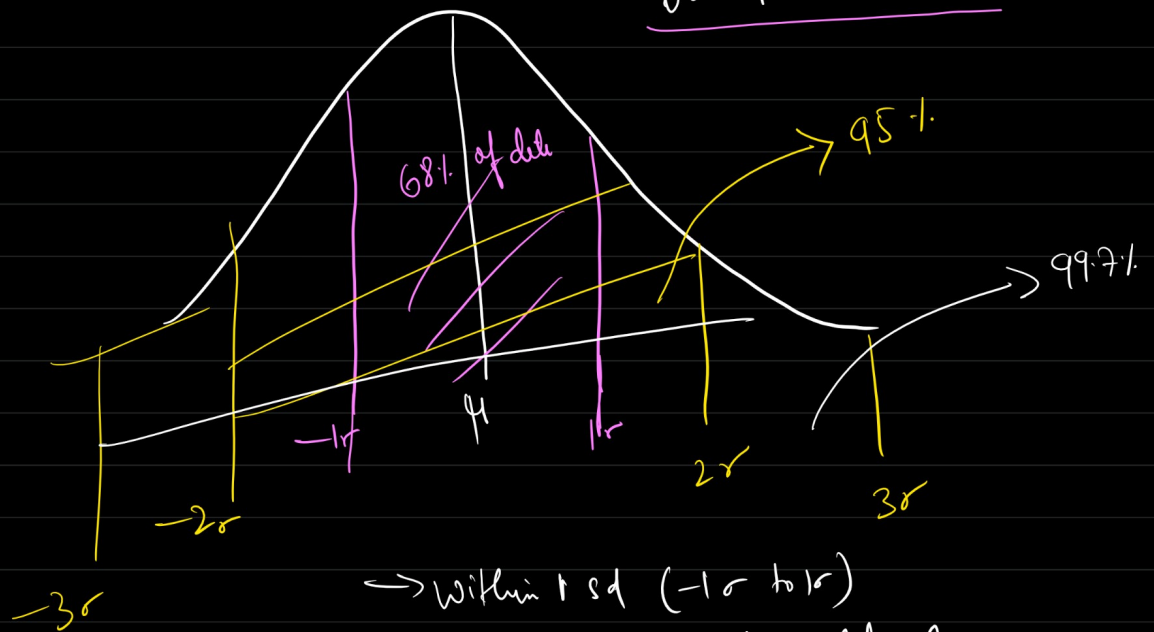


→ Most of the naturally occurring/human generated data follow a Normal distn.

Characteristics of ND

→ Symmetric ✓
→ Skewness = 0
⇒ Kurtosis = 3
→ Mean = median = mode

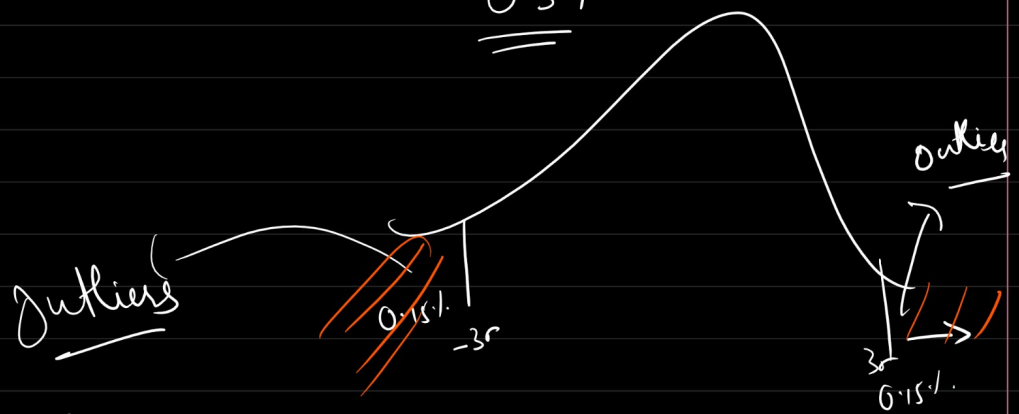
rule :- 68 – 95 – 99.7 %.



→ Within 1 sd (−1σ to 1σ)
68% of the data lie.

→ 2sd – 95%.
→ 3sd – 99.7%.

0.3 %



2nd dP 3σ के beyond
−3σ के beyond } 3 (total) outliers होते।

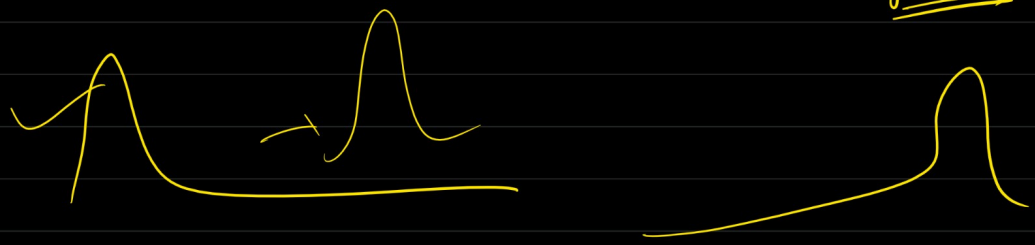
Example

- eg Ht of people
- eg Wt " "
- eg Score
- eg BP
- eg Emo
- eg IQ

* Use Case

→ Beyond 3sd → Outliers.

→ required the data to be Normally distn

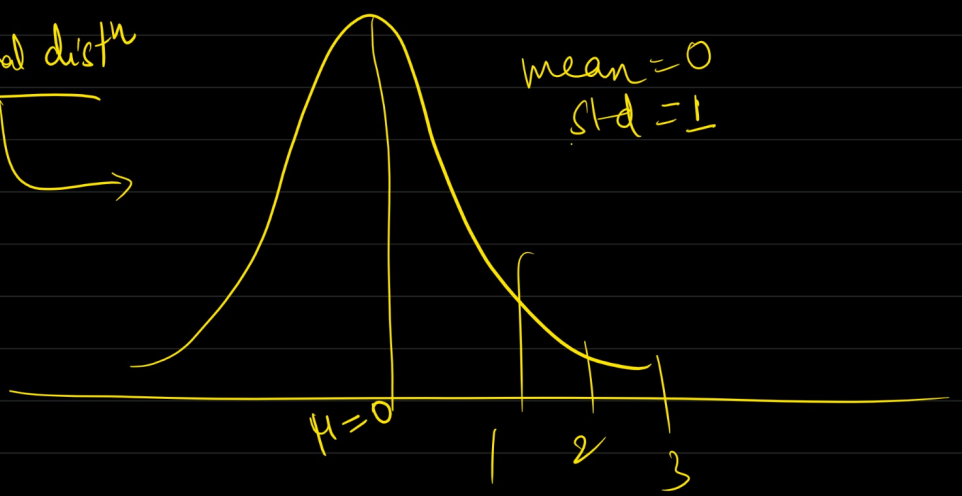


transfn → log based →
→ Square →
→ Recip →

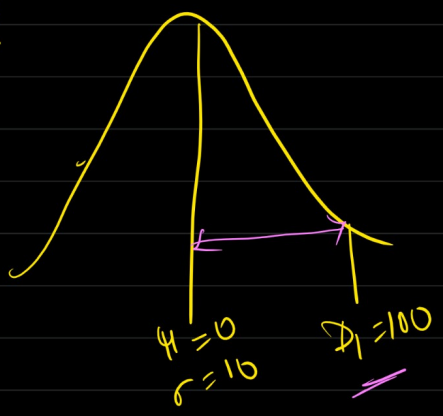
- right skewed data — log / square
- left " " — reciprocal
- -ve value — Yeo Johnson
- +ve " — Box Cox transfn
- No zero — Reciprocal trans

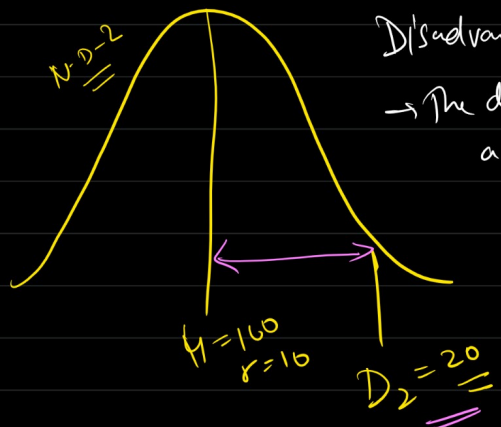
Nd — mean & Sd

* Standard Normal distn →



Why ?





Disadvantage of SND
→ The data is on a different scale

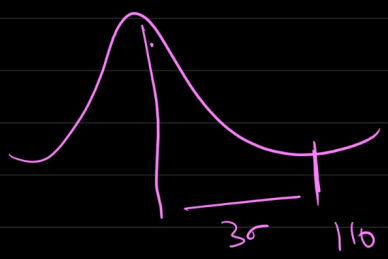
Use Case
↳ Scaling
↳ To compare two dist^n

Q: $N(\mu = 50, \sigma = 20, D_1 = 110)$.
How many s.d away this dp is from mean?

→ $Z_{score} = \frac{x - \mu}{\sigma} = \frac{110 - 50}{20} = \frac{60}{20} = 3$ sd away from mean.

Q $N(\mu = 100, \sigma = 10, D_2 = 200)$

$\frac{200 - 100}{10} = 10$ sd away.
far from μ.



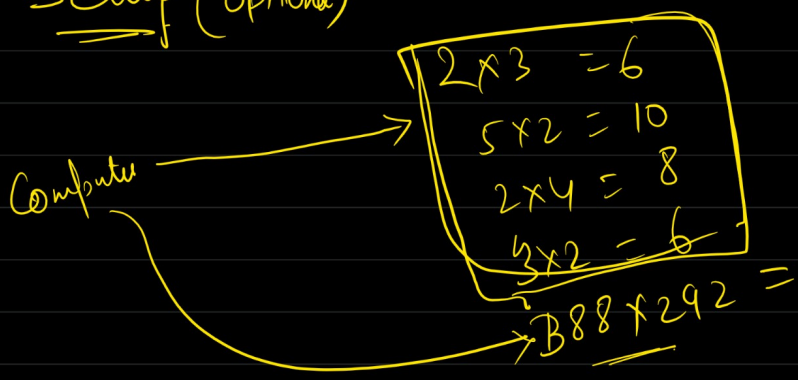
Scen-1

$Z_{score} = 3$ cm. $\left(Z_{score} = \frac{x - \mu}{\sigma}\right)$
= It means that the data point is 3 std away from mean

Scen-2

(standardize) = 3 cm.
↳ Each and every dp is on an avg is 3 cm away from mean.

Use SND → Scaling (optional)



ML

# of room	Area of house	Price of house
1	1100	5
2	900	3
1	2006	8
2	1	⋮
2		⋮
5		⋮
2		⋮
1		⋮
2		⋮

ML → To learn pattern from data

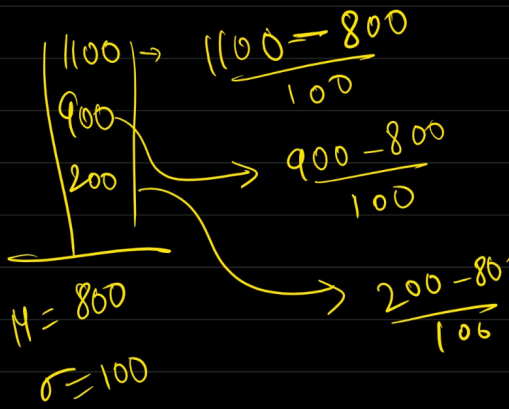
↓

Mathematical relation

(mathematical calculation)

↓

Scaling



Scaling

→ Standardisation (SND)

→ Normalisation

Q 95% of students at school are b/w 1.1m and 1.7m tall. mean & Std dev ??

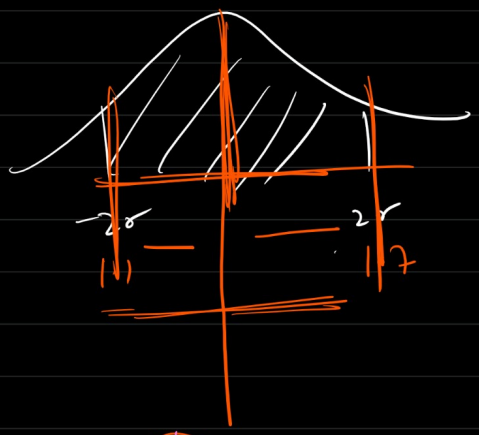
95% — 1.1m — 1.7m

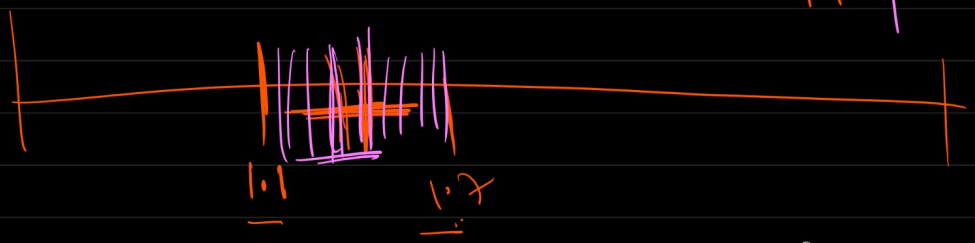
68 - 95 - 99.7

$\text{mean} = \frac{1.1+1.7}{2} = 1.4$

$4\sigma = 1.7 - 1.1$

$\sigma = \frac{1.7-1.1}{4} = 0.15$

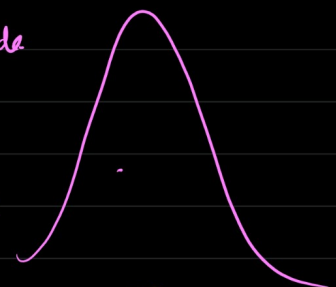


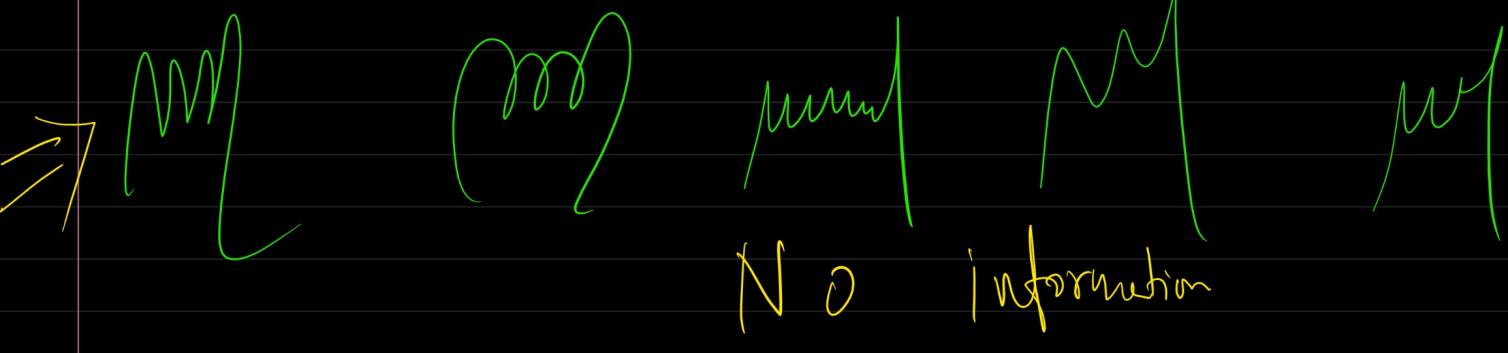


8th wonder of world → CLT (Central limit theorem)

⇓

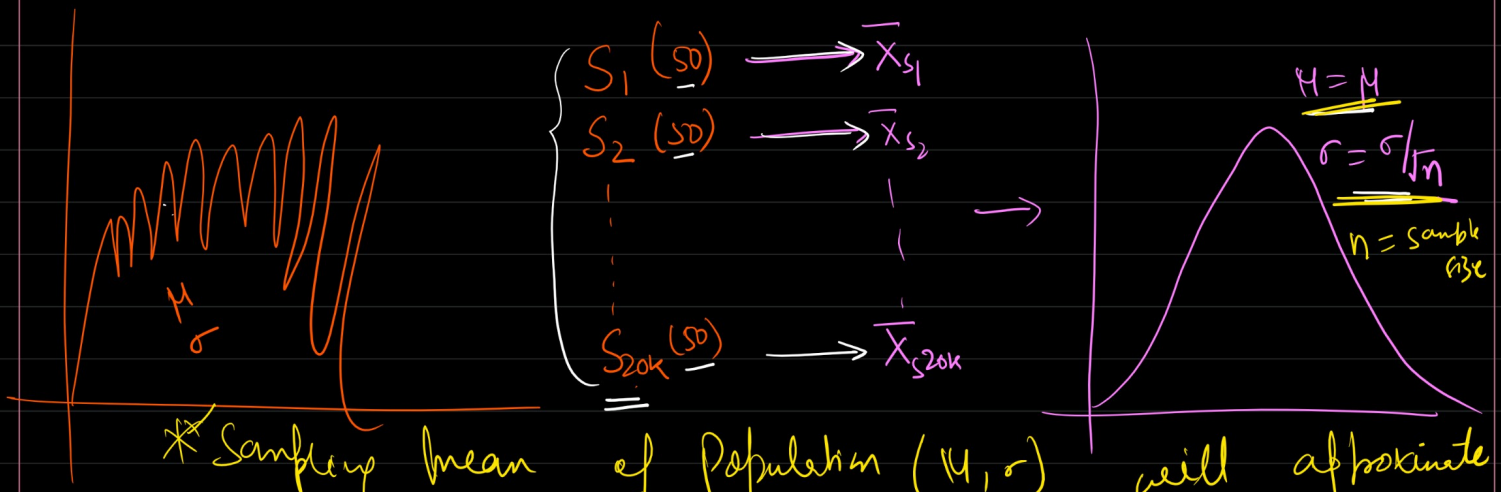
- → mean = median = mode
- → Skewness = 0
- → Symmetric
- → 68-95-99.7 rule





No information

Central limit theorem → A way to convert any form of dist^n to Normal d^n.

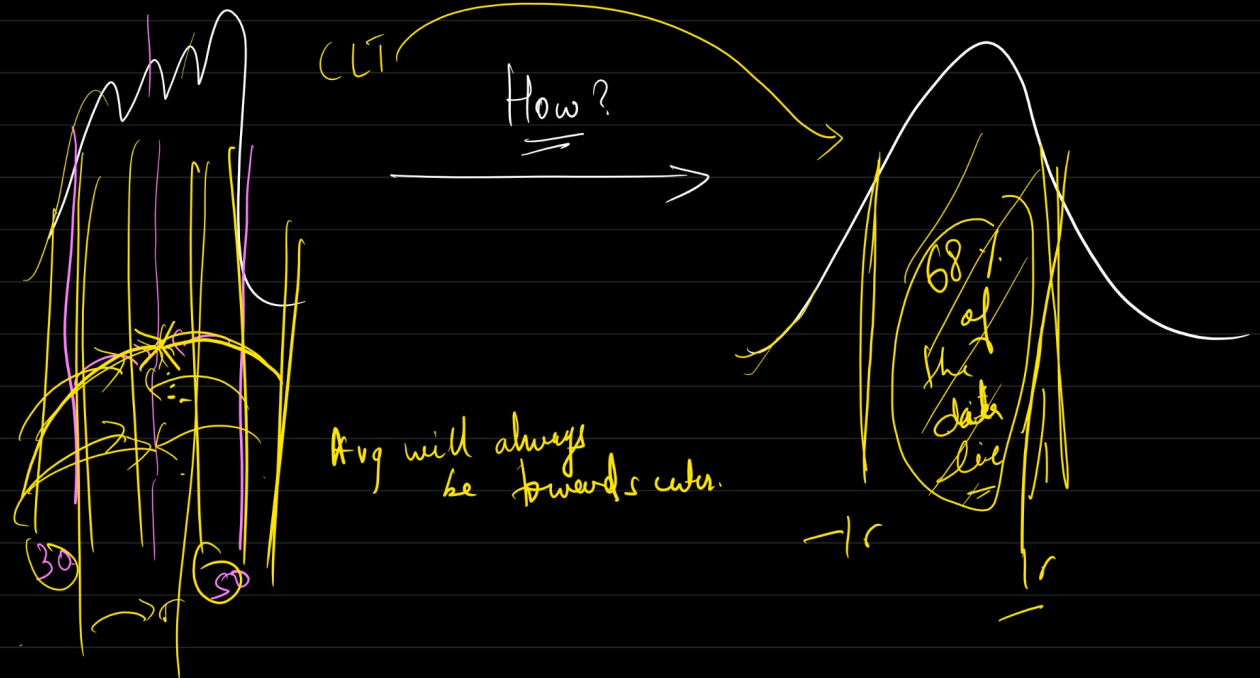


* Sampling mean of Population (μ, σ) will approximate to a normal dist^n

* Two condition of CLT

→ The no of Samples should be large enough

→ The Sample Size ≥ 30



SND

$$Z_{SND} = \frac{\bar{x} - \mu}{\sigma}$$

CLT

$$Z_{SND} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

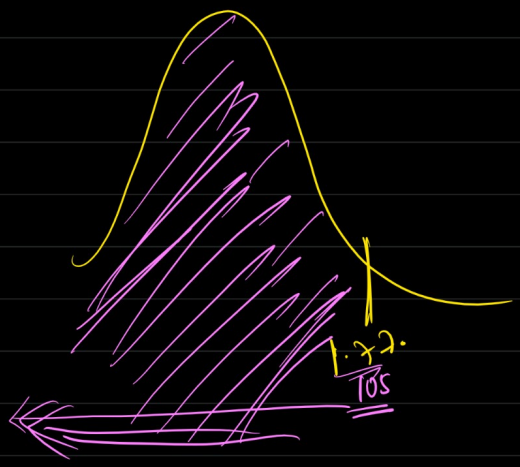
S.S < 30 → t distn

If you have a population with $\mu = 100$ & $\sigma = 20$, if you have SS of 50 from this pop. what is prob that sample mean will be less than 105?

→

$\mu = 100$ $\sigma = 20$, $n = 50$, $\bar{x} = 105$

$$Z_{sor} = \frac{x - \mu}{\sigma} = CLT = \frac{105 - 100}{20/\sqrt{50}} = \frac{5\sqrt{2}}{4} = 1.765 = 1.77$$



Z dist ≠ Z table

= 1.77

- Normal dist
- Standard Normal dist
- Log Normal distn
- t-distn
- Pareto
- Exponential
- Chi-square
- F-distn

① t distn

→ Shape is determined by d.o.f $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ || σ of Population is given & S.S ≥ 30

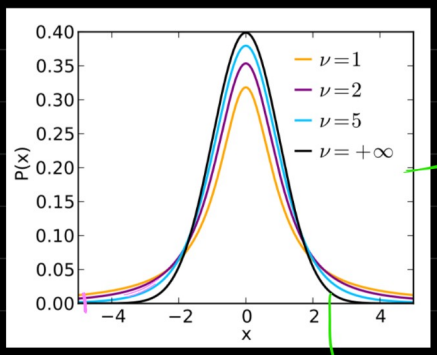
What if σ pop not given or S.S < 30

↓

t-distribution

$$t_{score} = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

S → Sample standard



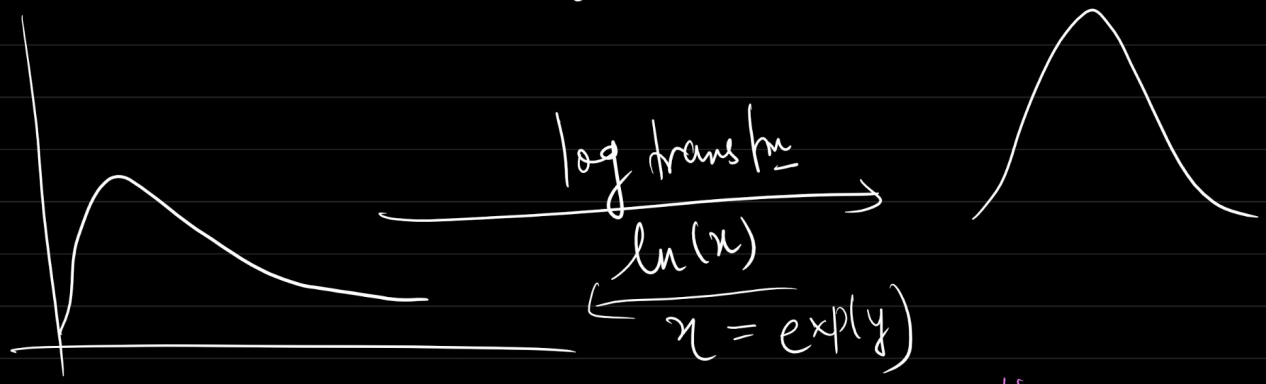
dof = 5-5-1

t dist – tails becomes thicker
Shape – becomes thinner.

→ Normal / SND

② log Normal distn

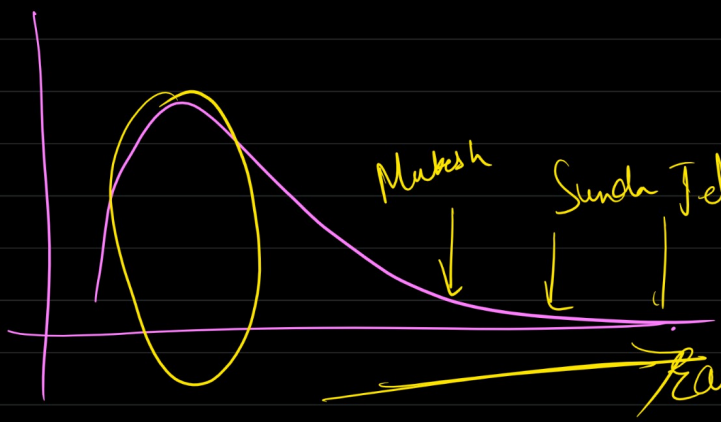
↳ A contn prob distn whose logarithm is normally distributed.

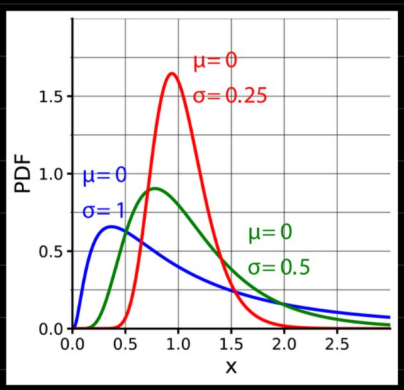


logarithm distn to normal distn

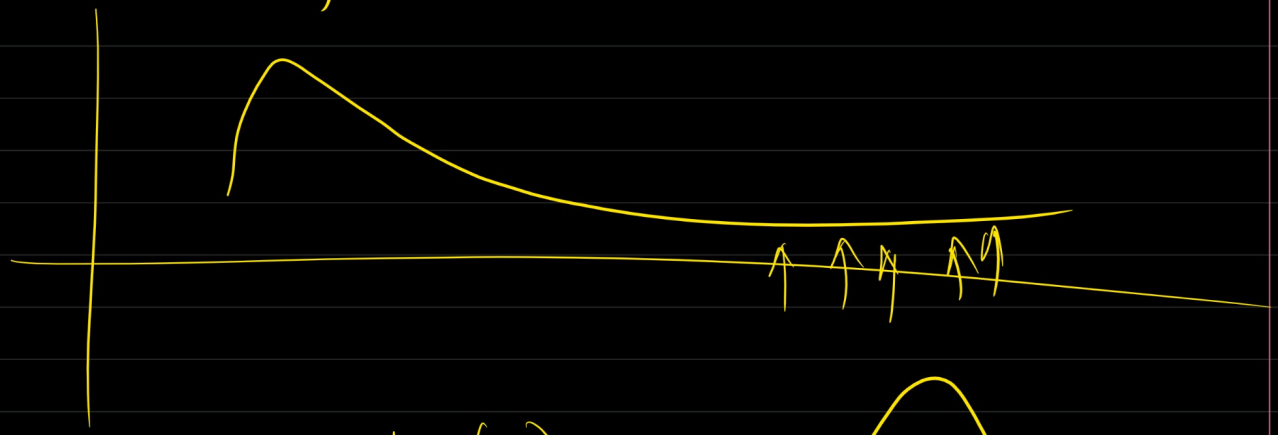
Example

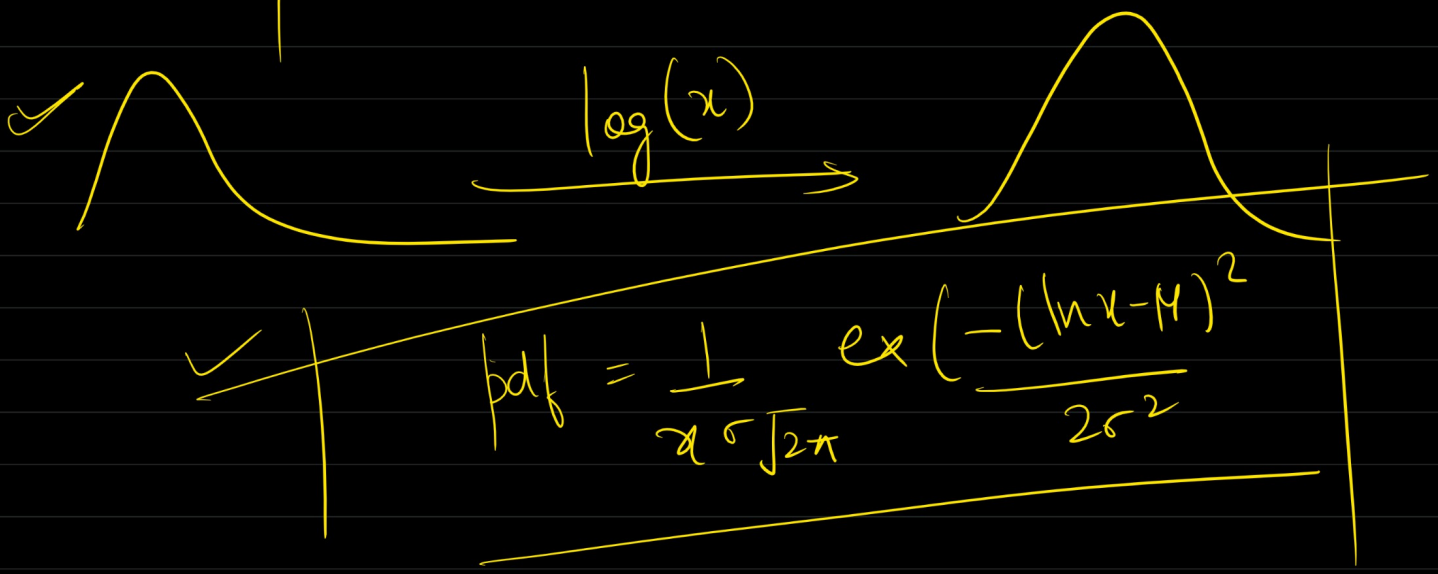
Wealth distn of world





* People virality comments in Post





$$pdf = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$

* Power law distn (Pareto principle)

RCB

KGF

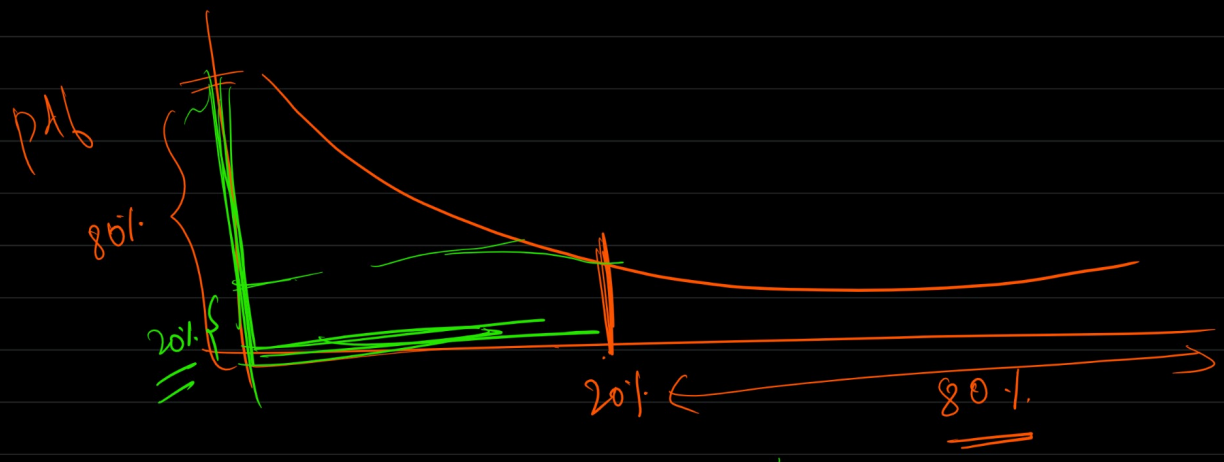
20%

20% of the people does the 80% of thing

→ 20% of employee in a team does 80% of work

→ 20%

80 | 20 forms

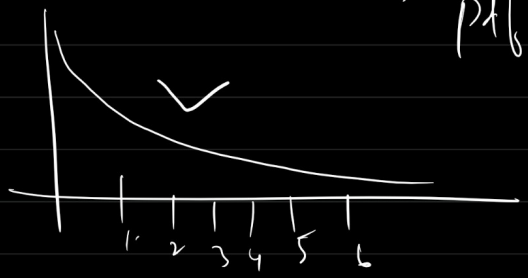


$$pdf = \frac{\alpha x_m^{\alpha}}{x_m^{\alpha+1}}$$

* Exponential dist^n

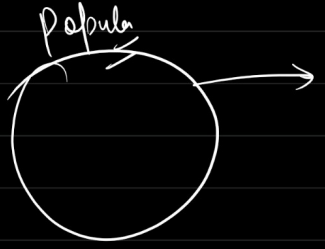
$$pdf = \lambda e^{-\lambda x} \quad x \geq 0$$

→ Radio active decay
→ life span of battery



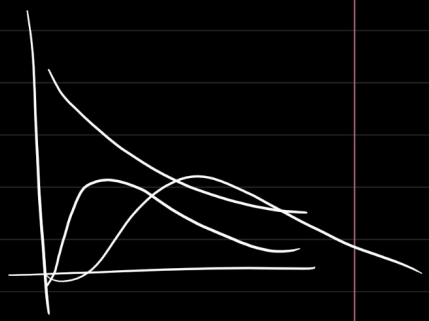
* Chi-square dist^n

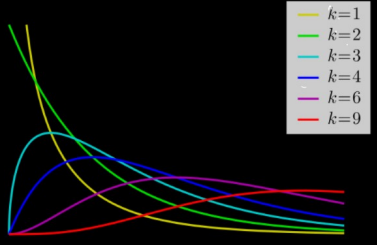
→ describe the dist^n of Sum of Square of k random variable.



$S_1 (1,2)$
$S_2 (3,4)$
$S_3 (2,4)$

Square the sum & add
1^2+2^2
3^2+4^2 →
2^2+4^2



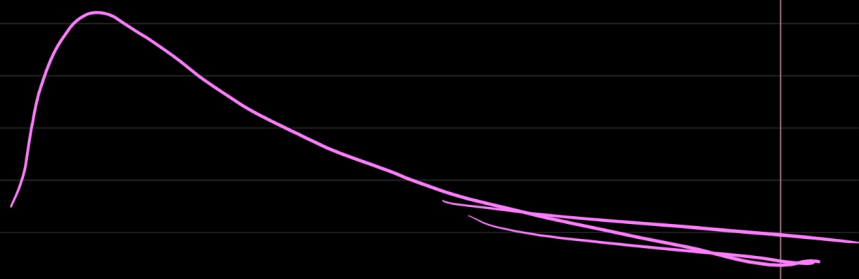


→ determined by k
→ Non-neg dist^n
→ right skewed data
for small k — Pareto pa

$(0.2)^2 = 0.04$

2^2 — 4

3^2 — 9

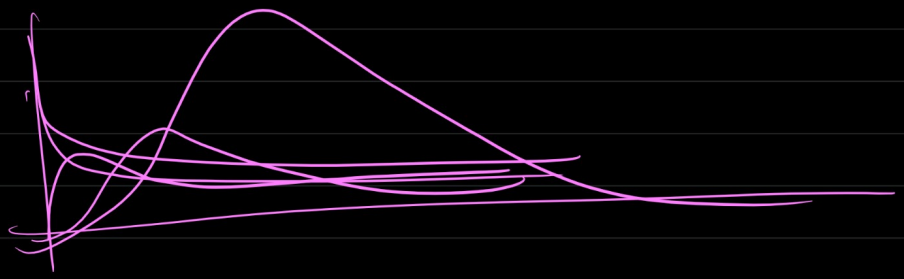


* F distn

→ Compares variances

$d_1\ d_2$, $s_1\ s_2$ Fisher Snedecor

$$X = \frac{s_1 | d_1}{s_2 | d_2}$$



dof = SampleSize − 1

t dist → With change in No of Sample, the shape changes

Bernoulli ← Pass / fail: $p^k(1-p)^{1-k}$

⇓

n no. of Bernoulli — Binomial dist^n

⇒ $\binom{n}{k} p^k (1-p)^{n-k}$