Till now:

→ Pmf.

→ Discrete Uniform dist^n — All values have equally occurrence.

→ Bernoulli distribution
  ↳ 2 possible outcome

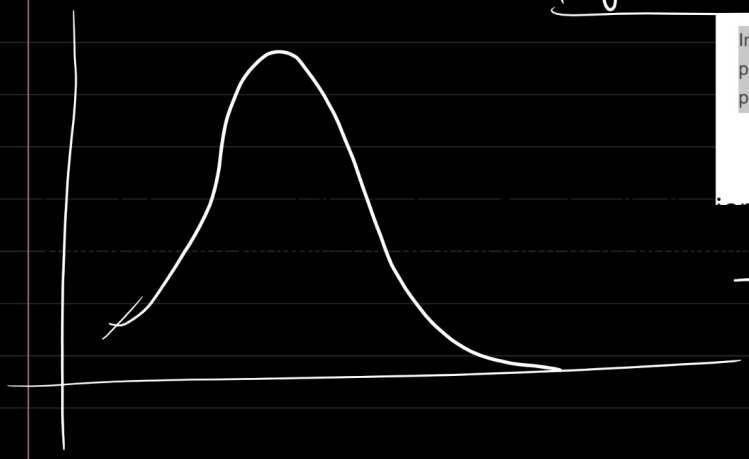
→ Binomial → n - bernoulli trial.

→ Geometric — first success after n trials

→ Poisson → Avg no of occurrence in a time interval.

* Probability density fn
  ↳ for continuous random variable.

* Normal distribution. | Gaussian distribution | Bell-shape distribution



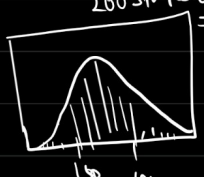
In statistics, a **normal distribution** or **Gaussian distribution** is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

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

Why ??

→



200 students.

Measure the height of students →

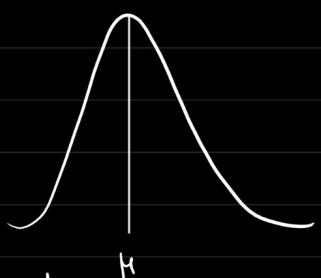
Characteristics of N.D

→ Symmetric

→ Skewness = 0

→ Kurtosis = 3

→ Mean = median = mode.



7.5th wonder of world :- figure ⇝ 68 - 95 - 99.7 %

empirical

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

→ within 2 sd (−2σ to 2σ) 95% of data lie.

→ Within 3 sd (−3σ to 3σ) 99.7% of data lie.

How ??

→ Distribution of data ✓ → data → sd → — 68% → 100 — 200

→ Outlier of data →

⇓

extreme values

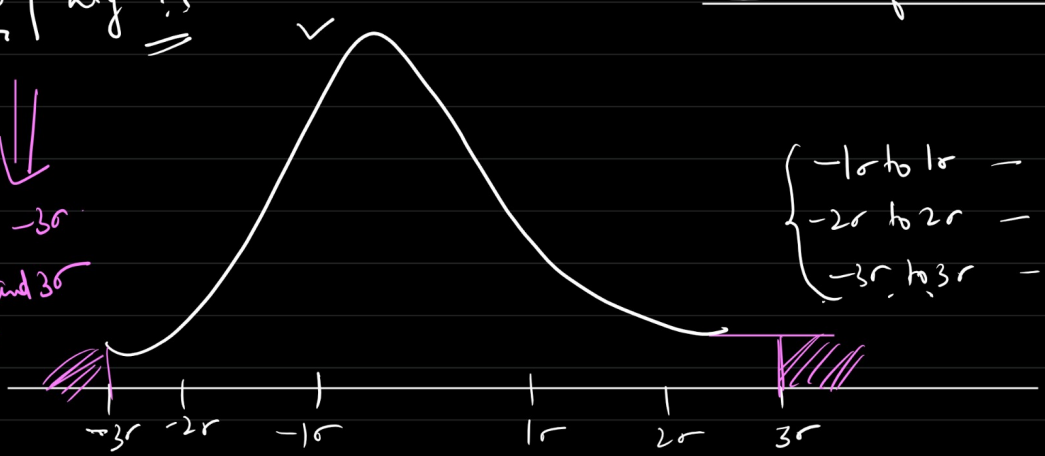
⇓

low | big ??

→ Max no of values lies b/w Q1 & Q3

−1sd − 1sd → 68%

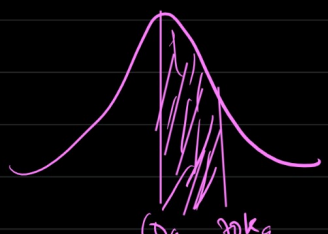
beyon −3σ and 3σ ↓ 0.3%



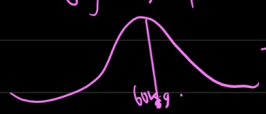
{ −1σ to 1σ — 68% ✓
−2σ to 2σ — 95% ✓
−3σ to 3σ — 99.7% ✓

* Prob densit

↳ ↳ Calculate the prob → person will fall between 60 kg & 70 kg.



→ Prob if wt is 60kg.



→ 0

Eg: Height of people

Eg: Weight

Eg: Measurement of error.

Eg: Score

Eg: Blood Pressure.

~~Mean is 0, Std = 1 for Normal distn.~~

* industry use case of normal distn

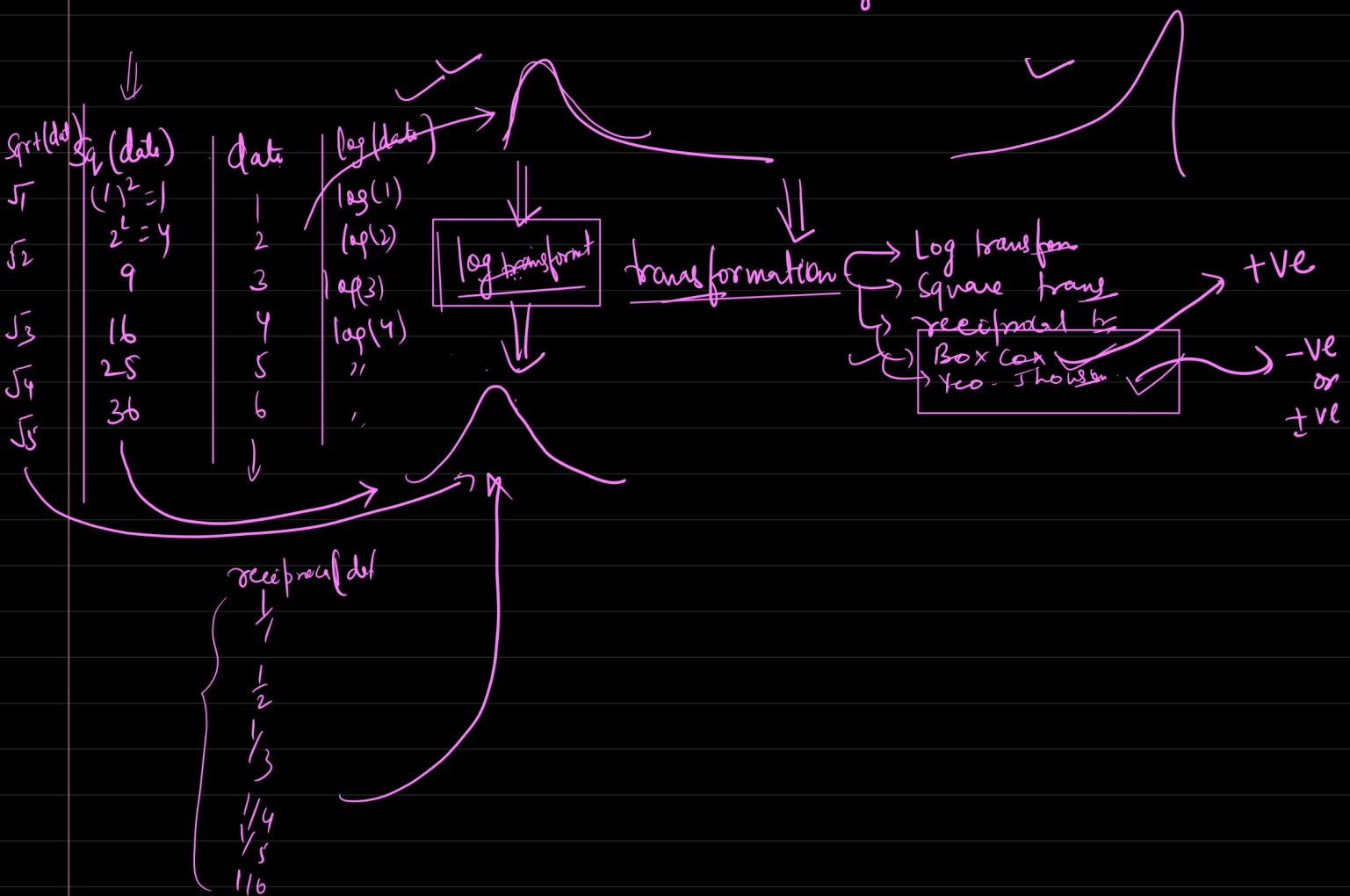
→ Beyond 3sd → outliers.

→ Some of ML | statistical models, requires the data to be Normally distributed.

① Most of data follows N.D

② Outlier identification.

③ Prediction will be wrong



→ if you are modelling statistical ML models then data should follow normal dist^n and if not do transformation.

↓

Which transformation when?

- → right skewed data — log / square.
- → left skewed data — reciprocal trans
- → −ve value → Yeo Johnson.
- → +ve non-zero — Box Cox transform
- → No zero → Reciprocal transform

Hit and trial.

* ~~For ND mean 0, σ=1~~

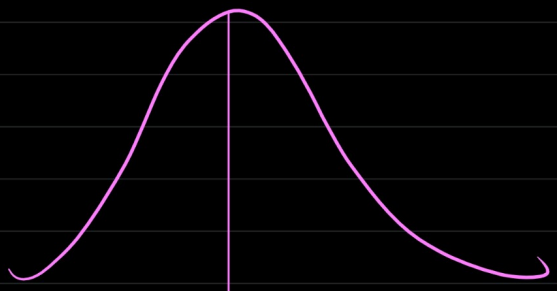
⇓

true or not

↳ Mean & std dev^n — for Normal distribution

Can be anything

mean = 0

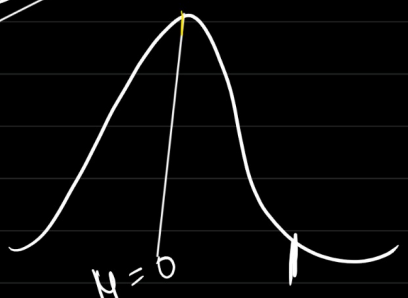


μ = 1010
σ = 20

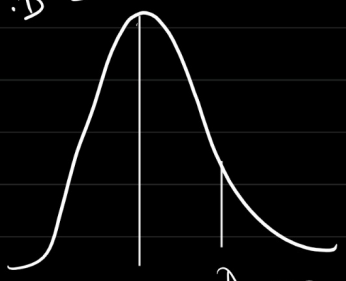
→ mean of the data is 1010

→ On an avg the data point is 20 units away mean

N.D-1



N.D-2



→ which dp among D_1 & D_2 is closer to μ ?

⇓ =

We can't say

⇓

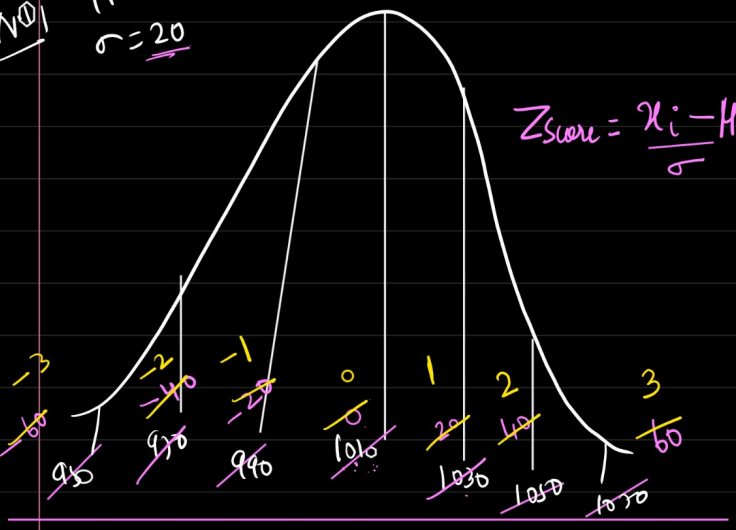
Standard Normal distn — Scales of both the distn is not same

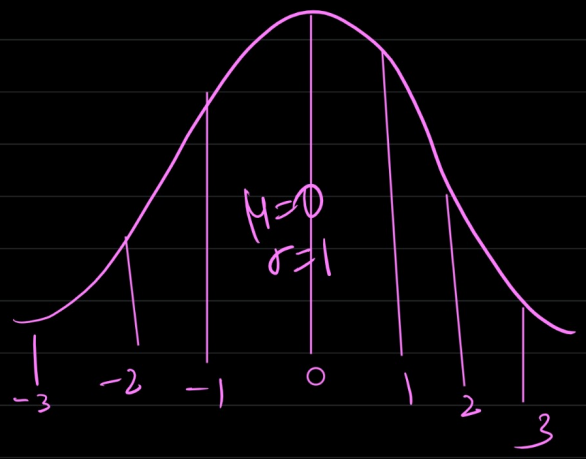
mean, variance

two classes

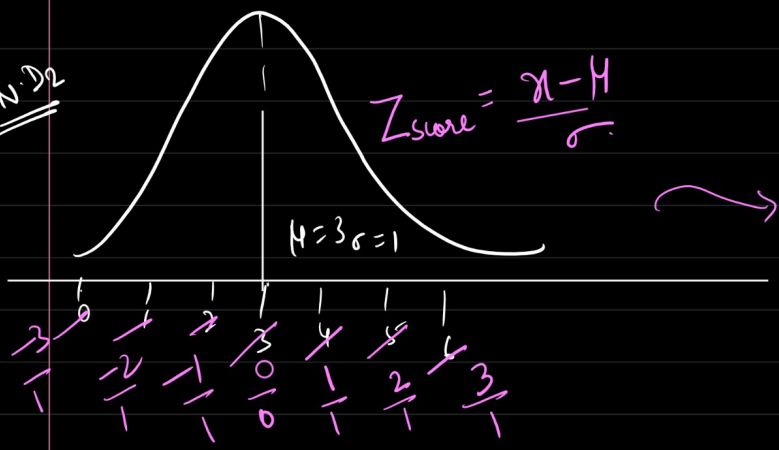
SC ↙ ↘ Aprts.

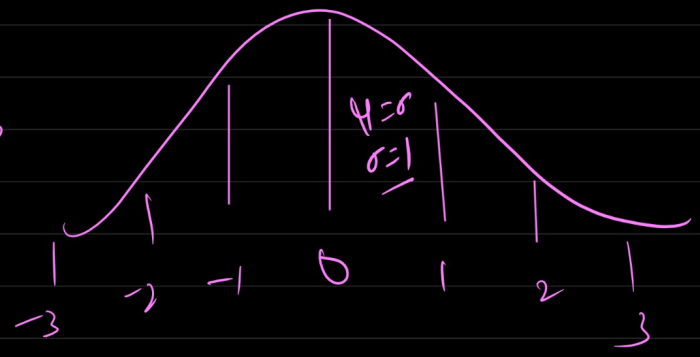
N(0, 1) $\mu = 1010$, $\sigma = 20$





N.D2



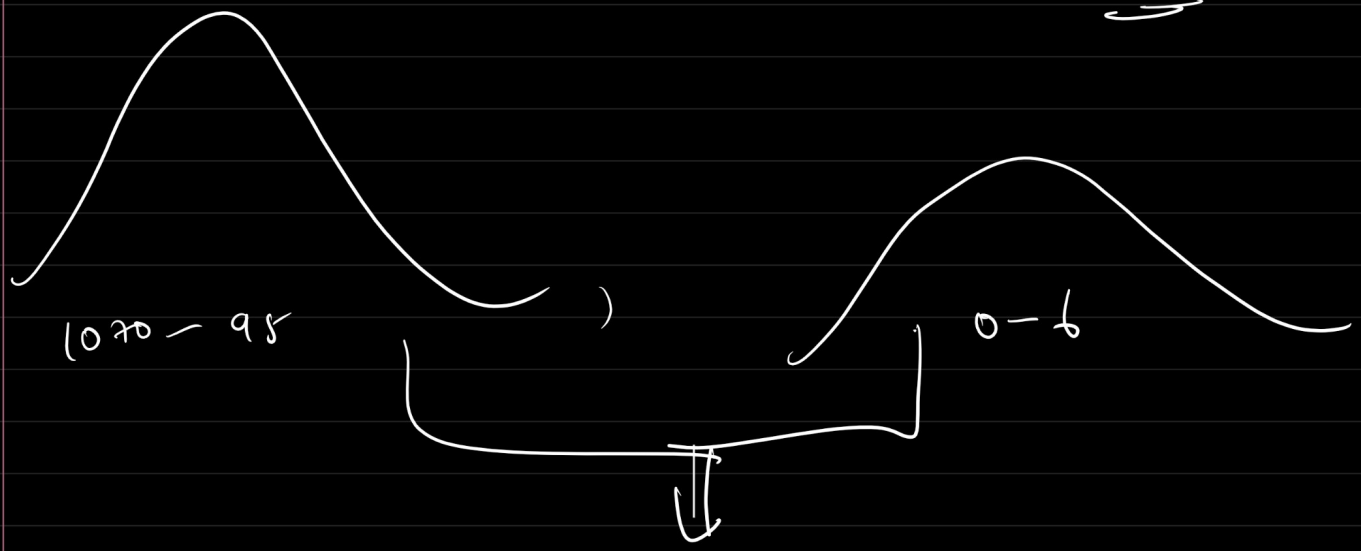


Std N.D

↳ Standardisation of data

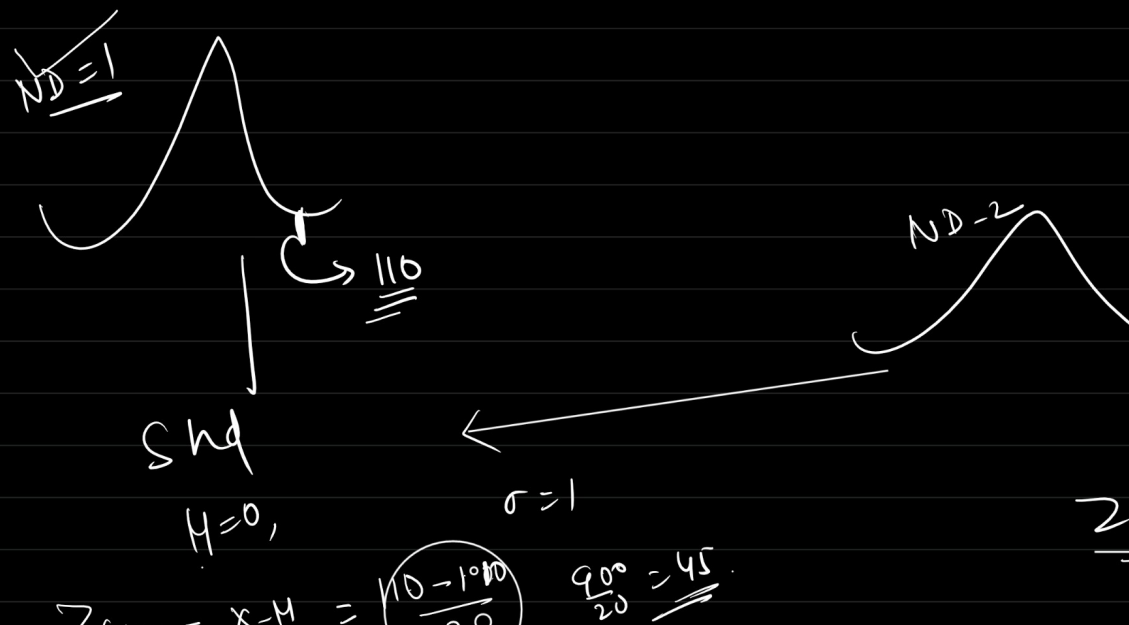
$\mu = 0, \sigma = 1$

* The normal distn which has $\mu = 0$, $\sigma = 1$ is Std normal distn



Standardising

$\mu = 0, \sigma = 1$



Whichever Z score is highest, the dp will be far from mean.

Q $N(\mu = 50, \sigma = 20, D_1 = 110)$

Can you tell how many standard deviation away this data point is lying away from mean value?

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

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

Can you tell how many std dev away D_2 is lying away from mean?

$\Rightarrow Z_{score} = \frac{200-100}{10} = 10$ S.d away from mean.

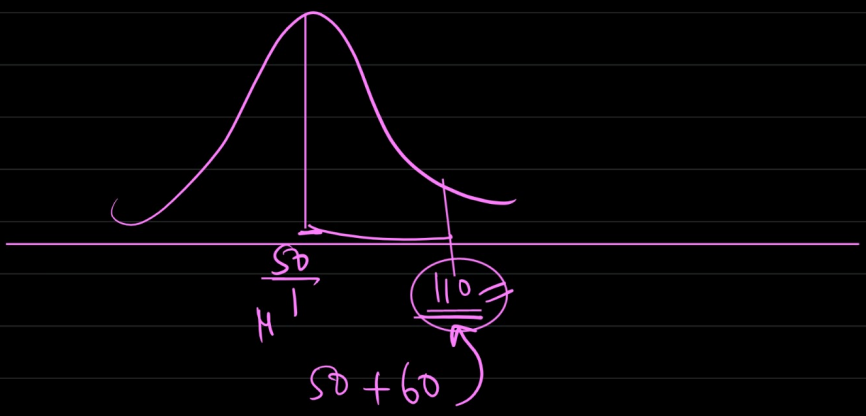
Q Which dp from D_1 & D_2 is more away from mean value? → D_2

Scenario-1

$Z_{score} = 3$ cms

It means that the datapoint is 3 std dev away from mean.

$D_1 \rightarrow 20 \times 3 = 60$ units away from mean



Scenario-2

$\sigma = 3$ cm

→ each and every datapoint on an arg is 3cm away from mean

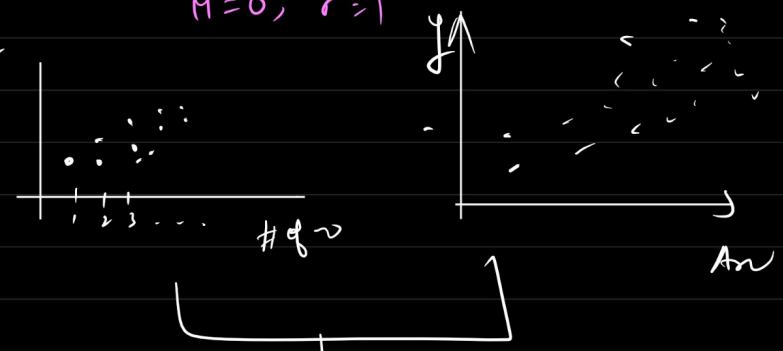
* industry use case of Std N.D

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

# of room	Area of house (sqft)	Price of house (y) (in Lr)
1	1100 sq ft	5
2	900 sqft	3
2	2000 sq	8
1	900 sq	
2		
5	950 sq	
3		
3	11006	
1		
2		
9		
10		
11		

$\mu = 5$
$\sigma = 1$

$\mu = 0, \sigma = 1$



Both data have different scales.

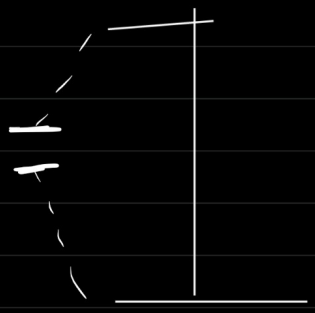
⇓

→ learning parameter will also vary much in terms of scale.

⇓

Scaling (optional)

Robuscaler → removes effect of outlier



Scaling

↳ Standardization ($\mu = 0, \sigma = 1$) (outlier)

↳ Normalization (0 and 1) (min-max Scaler)

$$x_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{mi}}$$

$x_{min} = 1$
$x_{max} = 5$

1	$\frac{1-1}{5-1} = 0$
2	$\frac{2-1}{4} = \frac{1}{4} = 0.25$
3	$\frac{3-1}{4} = \frac{2}{4} = 0.5$
1	0
2	0.25
3	0.5
4	$\frac{4-1}{4} = \frac{3}{4} =$
5	$\frac{5-1}{4} = 1$

Why Scaling?

→ Better interpretation

→ faster optimisation (gradient descent)

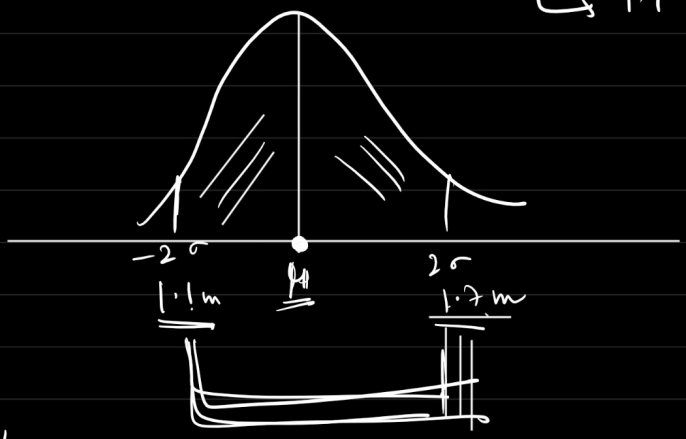
Q. 95% of students at school are between 1.1 m and 1.7 m tall. Can you calculate mean and std dev.??

→ 95% ⟶ 1.1 m and 1.7 m tall

-2σ to 2σ

-2σ to 2σ ↳ 1.1 to 1.7

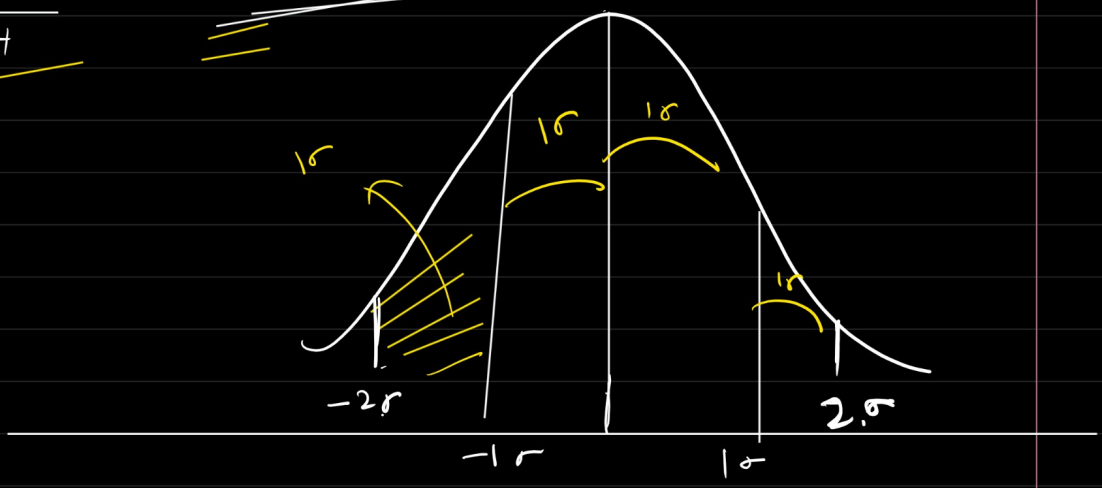
$$\frac{1.1+1.7}{2} = 1.4\text{m}$$



95% is 2 s.d b/w

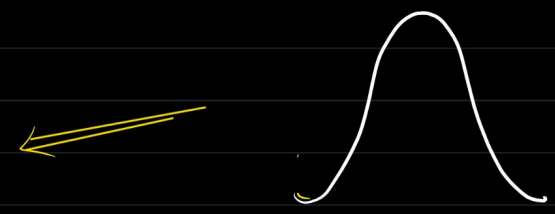
$$4 \text{ s.d} = 1.7 - 1.1$$

$$1 \text{ s.d} = \frac{1.7-1.1}{4} = 0.15 \text{ m}$$



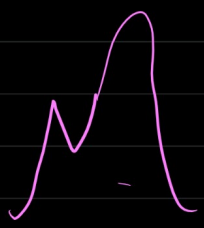
* Central limit theorem (8th wonder of world)

* For a N.D
  - → Symmetric
  - → Skewness = 0
  - → Kurtosis = 3
  - → mean = median = mode
  - → 68-95-99.7% rule ✓



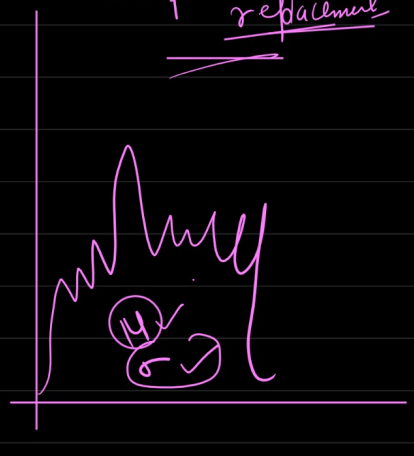
Std N.D — $\mu = 0$, $\sigma = 1$

Haphazard | irregular



Central limit theorem → A way to convert any form of distribution to Normal distribution.

Sample with replacement.

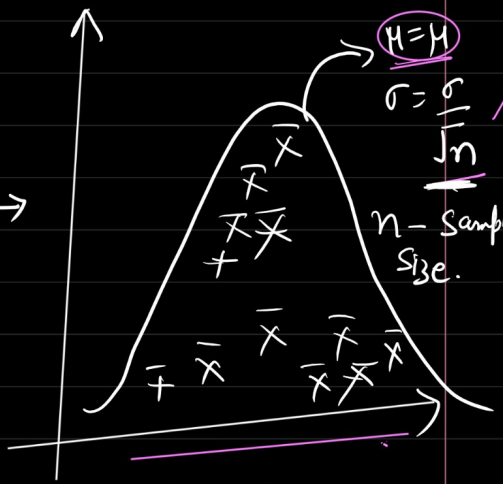


S_1 (50) → $\bar{x}_{s_1}$

S_2 (50) → $\bar{x}_{s_2}$

S_3 (50) → $\bar{x}_{s_3}$

S_{20K} (50) → $\bar{x}_{s_{20K}}$



→ The CLT states that if you have any pop with μ, σ and you take sufficiently large random sample from population with replacement, then the distribution of sampling mean will be approximately

normally distributed.

* Sampling mean of population (μ, σ) will approximate to a normal distribution.

Population (μ, σ) → large no of Sample → Sampling → ND

$\mu = \mu$

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

Two conditions of CLT

→ The no of samples should be large enough.

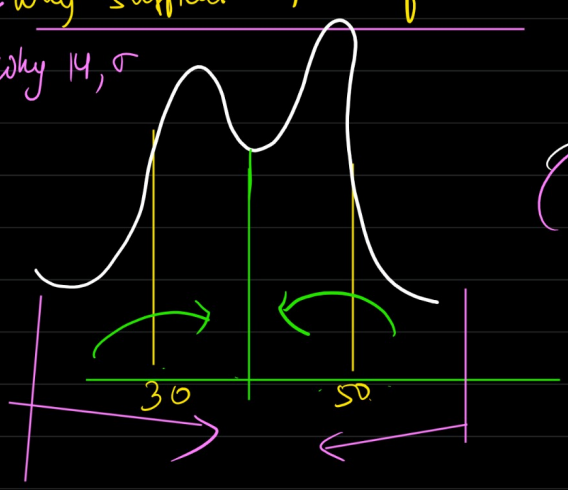
→ The Sample size should be equal to and greater than 30 (Except Pop is ND)

Why $n \geq 30$?? → Sampling mean follow 2 distribution → $z_{sur} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

CLT ??

Why sufficient large no of sample?

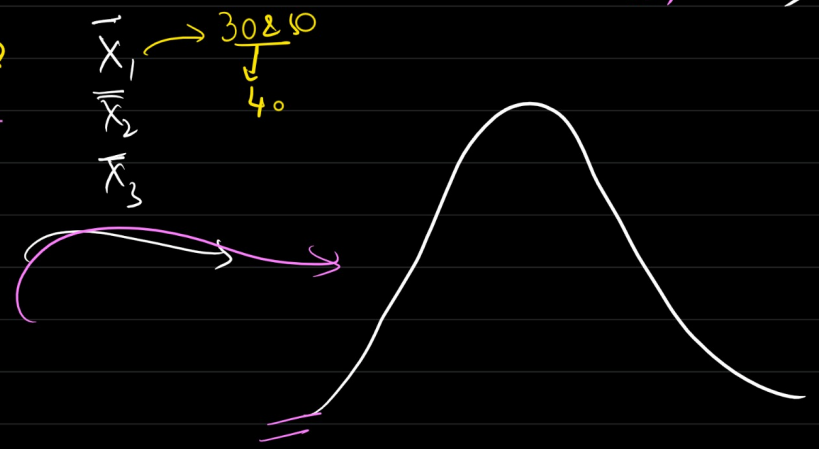
Why μ, σ



$\bar{x}_1$ → 30 & 50

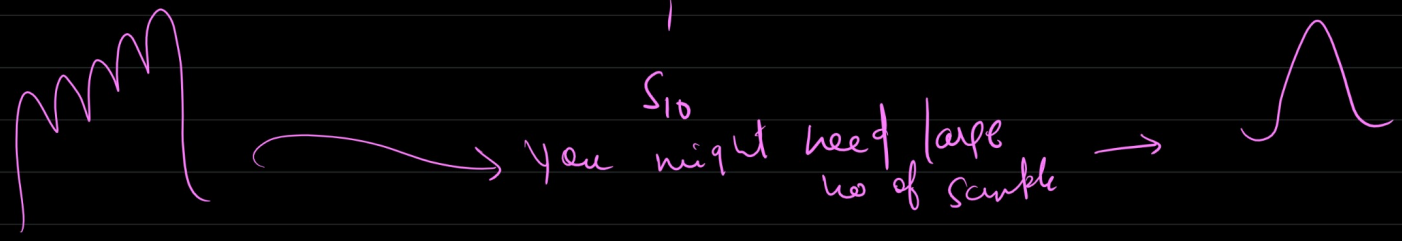
$\bar{x}_2$

$\bar{x}_3$



→ N.D

Most (68%) → will lie in the centre → 1SD to 1SD



Population $\Rightarrow$ $Z_{score} = \frac{x-\mu}{\sigma}$

$\Rightarrow$ S.S < 30

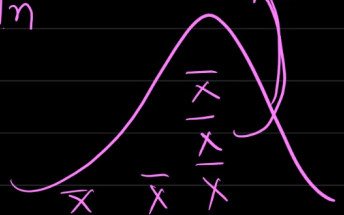
$\rightarrow$ S.S ≥ 30

$\Downarrow$

CLT

$\rightarrow \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \rightarrow$ Z-distribution

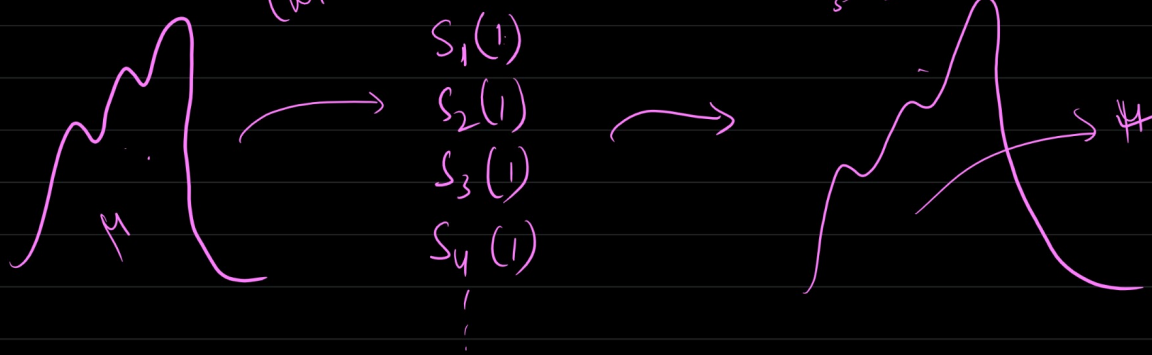
if S.S < 30 $\rightarrow$ t-dist



Why μ is μ after CLT?

Why σ is $\sigma/\sqrt{n}$ after CLT??

(without)



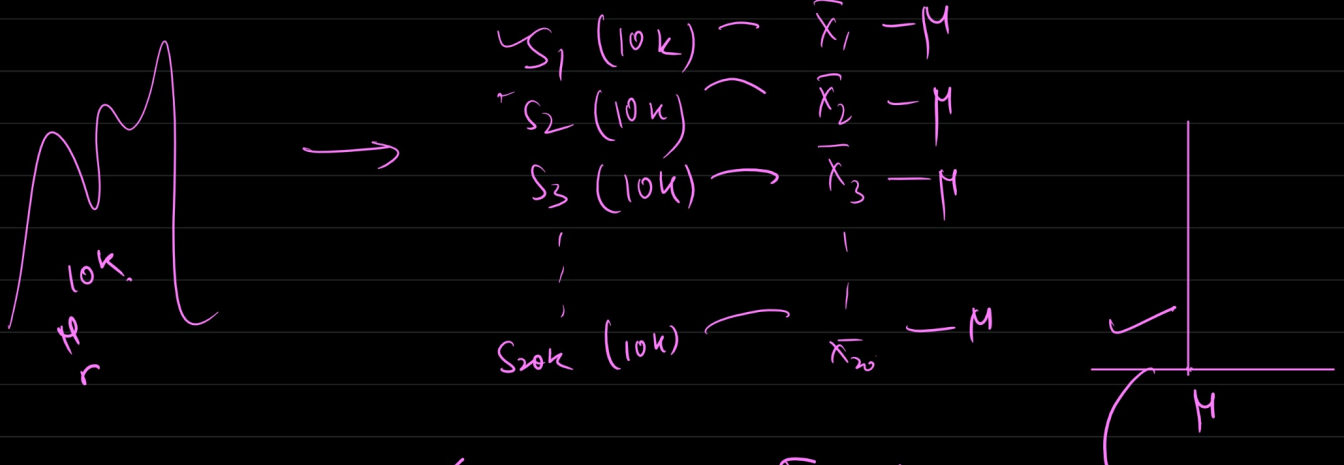
$2 \rightarrow 2$

$3 \rightarrow 3$

$5 \rightarrow 4$

$n = \infty$

$S(\infty) \Rightarrow$ Whatever no of elements are there in population, you take all element in Sample.



$\checkmark \sigma \to \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\infty} = 0$

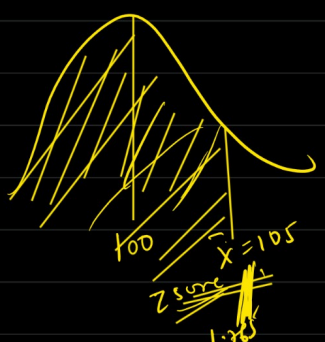
Q You have a population with $\mu = 100$ and std dev $\sigma = 20$. If you have sample size of 50 from this population. What is the prob that sample mean will be less than 105.

$\mu, \bar{x}$

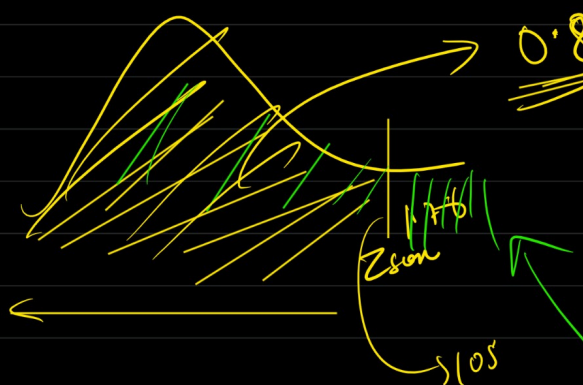
$\Rightarrow \mu = 100, \sigma = 20, n = 50 \quad \bar{x} = 105$

$$Z_{cur} = \frac{x - \mu}{\sigma} \xrightarrow{(CLT)} \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{105 - 100}{\frac{20}{\sqrt{50}}} = \frac{5\sqrt{2}}{4} = 1.767$$

area upto



Using Z table → Area = 0.89



Z disc m

↓

Z table

1 — 0.89

= 0.11

Zsim = 1.76

↓

0.89