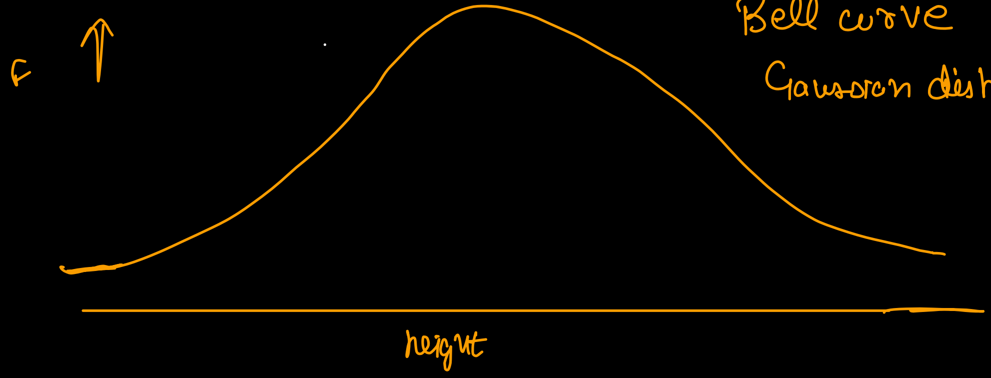
Welcome Back

Agenda for Today -

1. Graphs / Distribution in Data Science
2. Inferential Stats / Hyp. testing.

Distribution

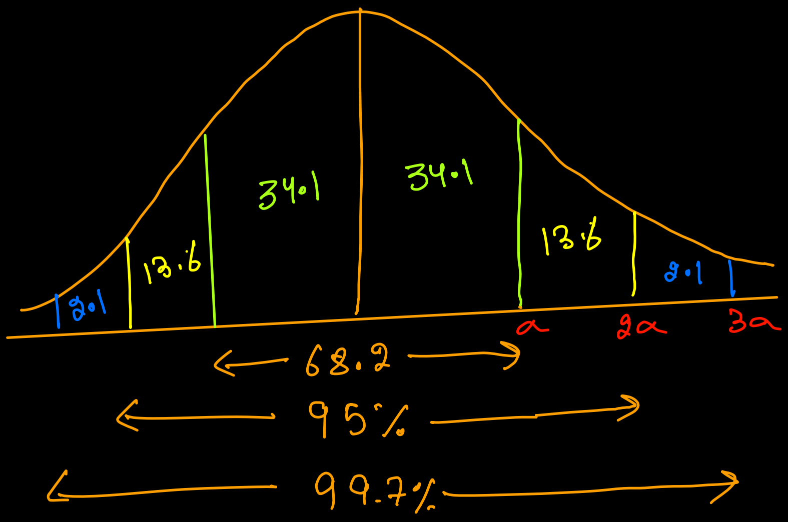
1. Normal / Gaussian Distribution



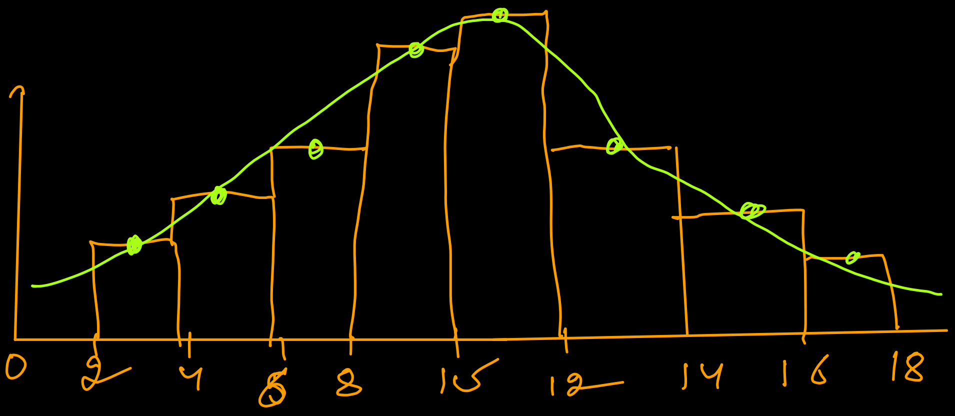
Bell curve

Gaussian distribution

⇒ Emphirecal Formula of normal distr.



2 3 5 7 8 9 10 11 12 17 18



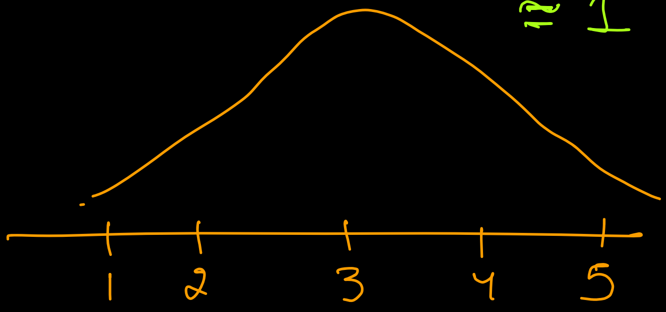
Standard Normal Distribution

$\{1, 2, 3, 4, 5\}$

$\mu = 3$

$\sigma = \sqrt{2} = 1.414$

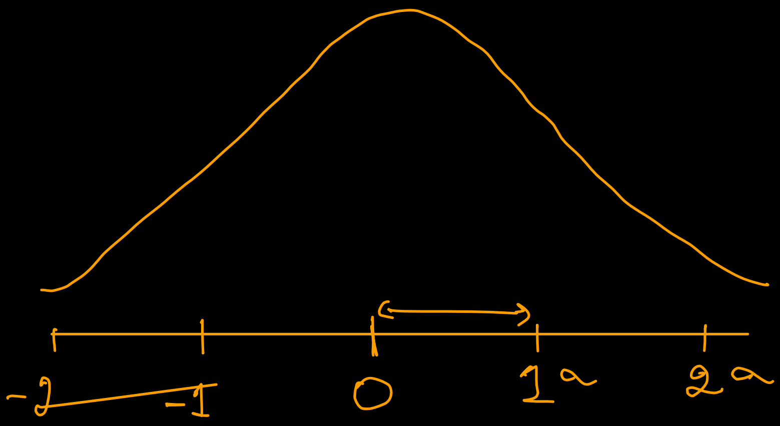
≈ 1 (assume)



$$z\text{-score} = \frac{x - \mu}{\sigma}$$

Standordization

Values (x)	μ	$x-\mu$	σ	$x-\mu/\sigma$
1	3	-2	1	-2
2	3	-1	1	-1
3	3	0	1	0
4	3	1	1	$\boxed{1}$
5	3	2	1	2

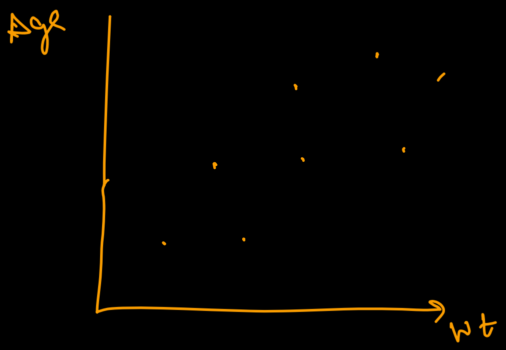


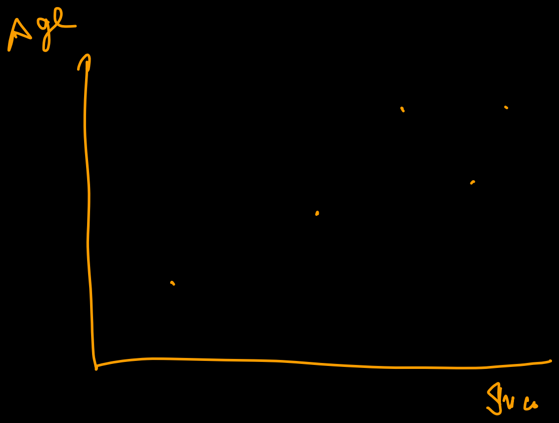
100 200 300 400 500

-200 -100 0 100 200

Age	Wt	Salary
25	60	3000
30	75	40000
32	63	500k
34	45	70k

diff unit





In order to do this we apply

Z-score

$$x_i = \frac{x-\mu}{\sigma}$$

$$\underline{3.35\sigma} \leftrightarrow 0$$

Normalization [Min-max Scale]

we specify the min^m and max^m value

f_1	
2	0.14
5	0.57
8	1
6	0.71
1	0

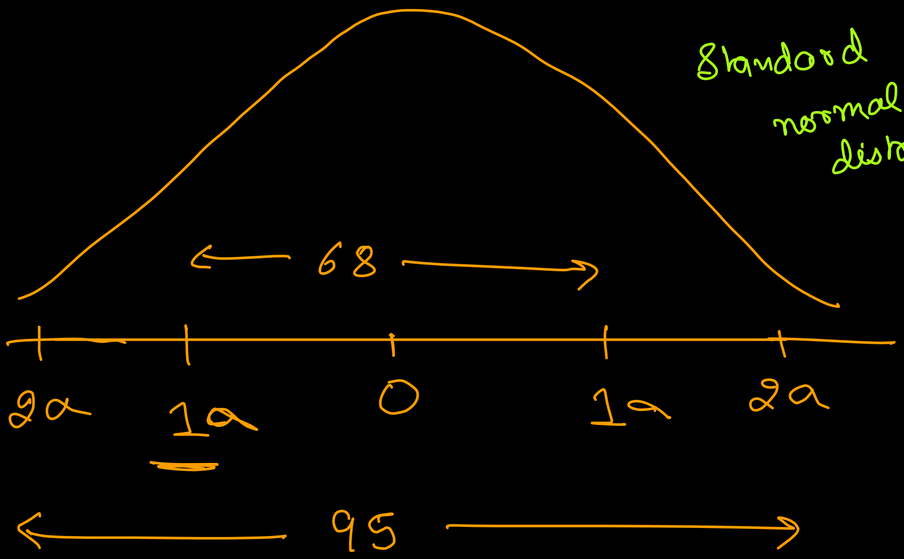
$$x_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

$$x_2 = \frac{2-1}{8-1} = \frac{1}{7}$$

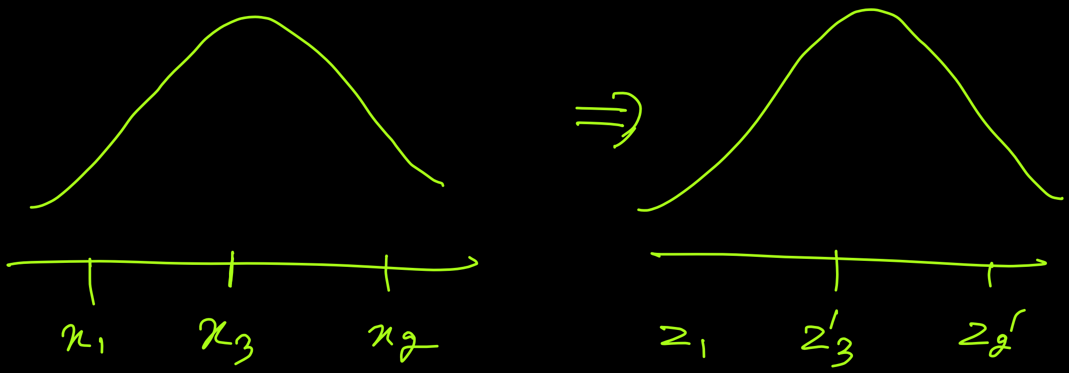
In most ML cases, standardization

In DL, CNN, we have pixels

we normalize to 0-1

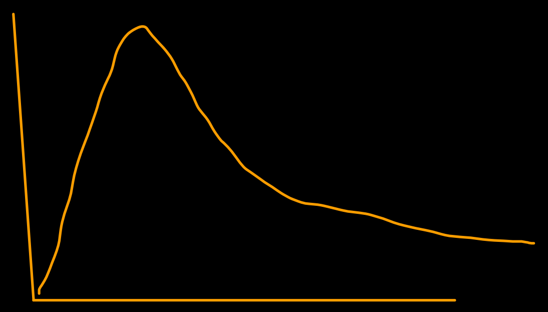


Standard normal distribution



Log Normal Distribution

One of the type of skewed data is log normal distribution.



wealth of people

length of comments on Youtube

Can we convert this into Gaussian distribution?

If we take natural log of each and every value, I will get normal / Gaussian distr.

$$\cancel{x_i'} = \ln(x_i)$$

$\log_{10}$ normal log

$\log_e$ natural log

$\underline{2}^3 = 8$

$3 = \log_{\textcircled{2}} 8$

$\log_a b = x$ | $\log_{10} 1000 = 3$

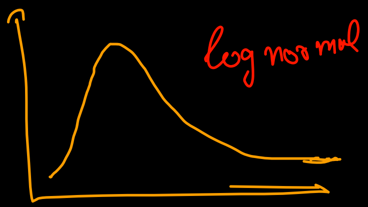
$a^x = b$

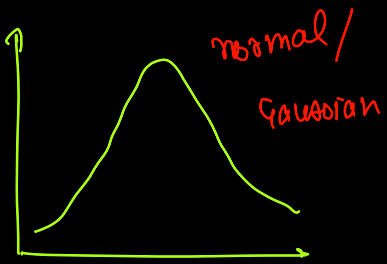
$\underline{e} = 2.73 \ldots$ π

$\underline{\log_e}$

If a variable x is following a log-normal distribution then $y = \ln(x)$ will follow normal distribution

x	$\ln(x) = y$
2	0.693
5	1.609
7	1.95
8	
9	
10	



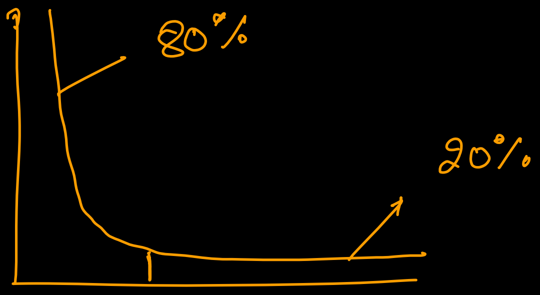


Power Law

80:20 rule

pareto principle

80% of your outcome comes from 20% effort



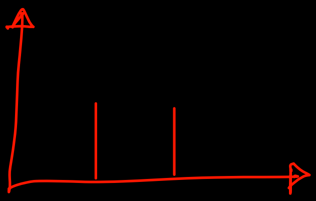
Bernaulis Distribution → discrete values

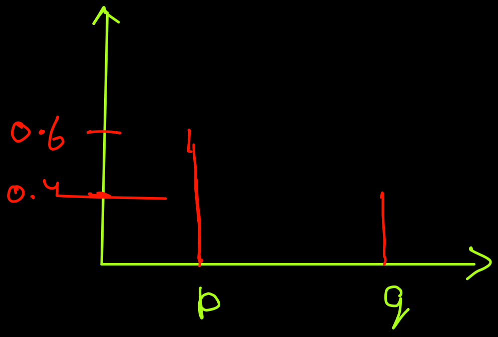
Flip a coin

6 heads
10 total

$$p = \frac{6}{10}$$

$$q = 1 - p = \frac{4}{10}$$





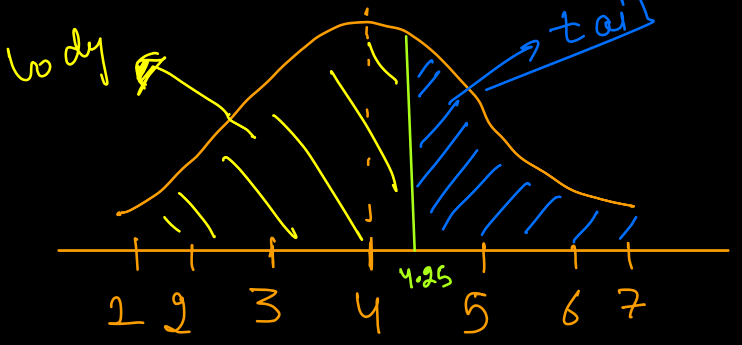
Injerential Statistics

$$z = \frac{x_i - \bar{x}}{\sigma}$$

Stats Interview Q

We take a standard normal distr.

with $\mu = 4$ $\sigma = 1$



Where does 4.25 fall?

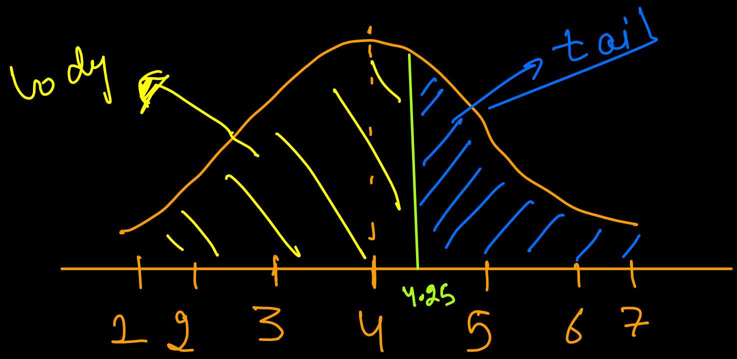
⇓

How many S.D. does our 4.25 falls from mean?

$$z = \frac{4.25 - 4}{1}$$

$$= .25$$

What % of Scores/values falls above 4.25 ?



I need to calculate the area of value

Can I assume full area to be 1 ?

For are, let us calculate & see z-score.

If body area is .5987

area of tail.

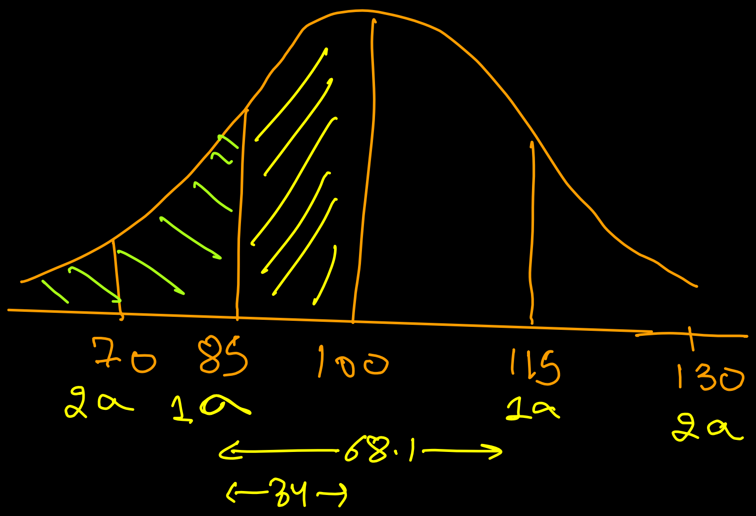
$\Rightarrow 1 - .5987$

$\Rightarrow 40.13\%$

In Mars, avg IQ is 100 with a S.D. of 15.

What % of population would you expect to have IQ less than 85?

$\mu = 100$
$\sigma = 15$



$$Z = \frac{85 - 100}{15} \Rightarrow$$

$$= \frac{-15}{15} = -1$$

Value from z-table for $z = -1$

$= 0.1587$

$\Rightarrow$ 15.87 %

% of pop^n having IQ b/w

85 – 100

$0.5 - 0.1587$

Hypothesis testing , Confidence Interval ,

Significance value.

Coin toss experiment : Check whether a coin is fair or not by doing 100 tosses

$p(h) = 0.5$

$p(t) = 0.5$

Cov 1 → ↓ 90%

Cov 2 ⇒ ↓ 95%

50H	60H	80H	90H
50T	40T	20T	10T

80H 20T ⇒ fair

90H 10T ⇒ ~~not fair~~

90H 10T ⇓ fair

Hypothesis testing

H_0 = null hypothesis = default assumpn

= oin is fair

H_1 = alternate hypothesis ⇒ opp. of default assumpn

= Coin is not fair

Perform experiment

Reject / Accept our null Hypothesis

50 H	80 H	90 H
50 T	20 T	10 T

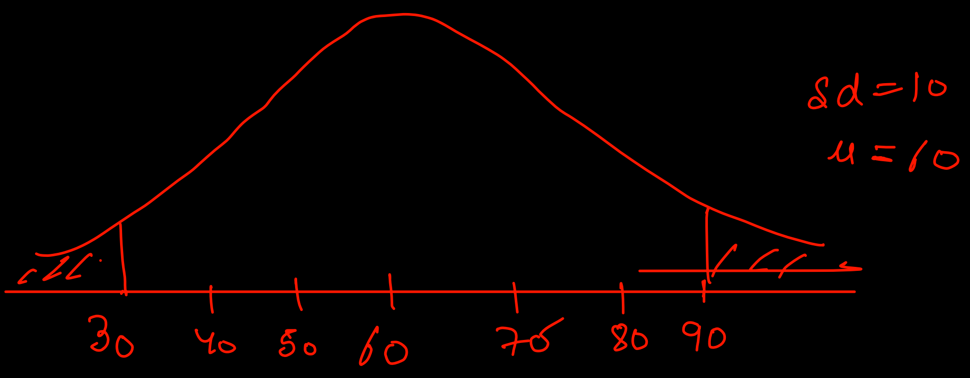
⇒ Whenever we have a condition /situation like above we define few more param.

2 params.

Confidence Interval , significant value
(α)

⇓

Using α, we calculate confidence interval



below 30, we have a doubt

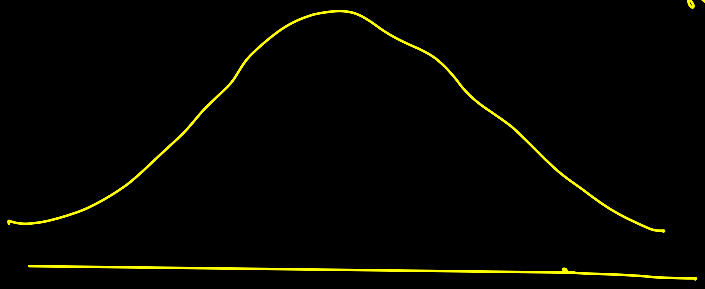
So, confidence interval basically says that we will be defining some range & if our value falls b/w that range then our coin is fair (H_0 ✓)

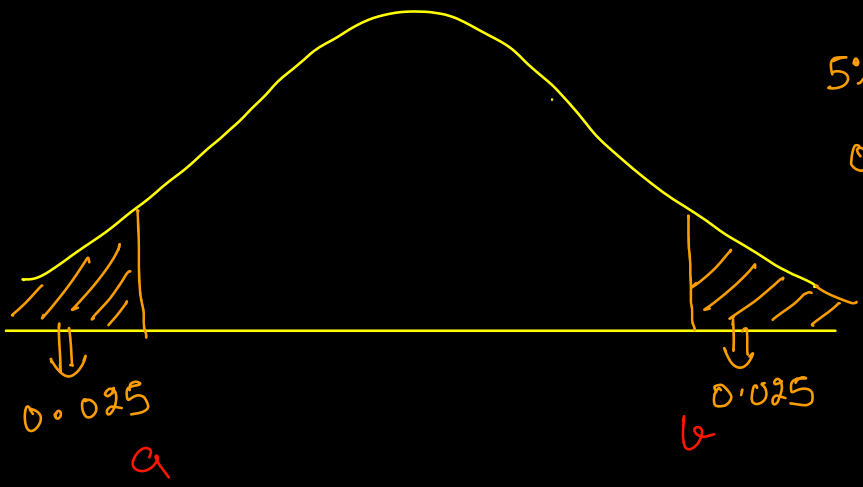
who define/ this confidence Interval ?
decides ?

Domain Expert person defines this interval.

eg :- health care

Covid test → false -ve report → symptoms.





5% = α

0.05

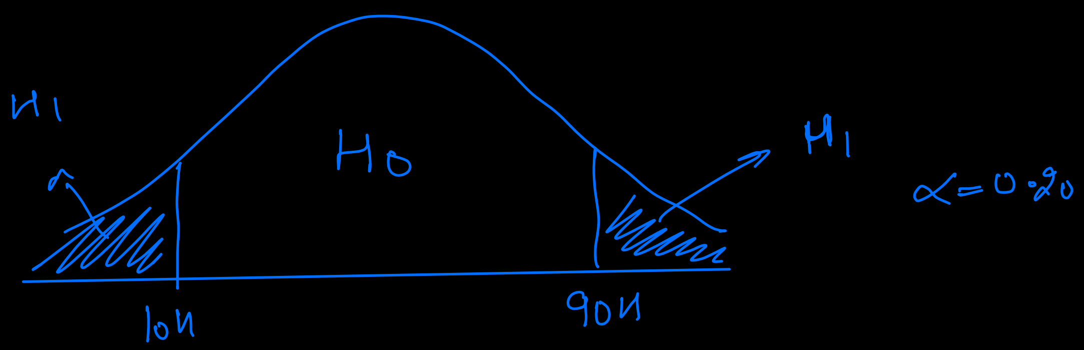
significance value = 0.5%

How to find confidence Interval (C.I.) from α ?

$$C.I. = 1 - 0.025 - 0.025$$
$$= 95\%$$

If my experiment yields that value fall b/w $\underline{a - b}$

I will accept H_0



Who define this α (significant value)

Vaccine

100 people take it

⇓

30 don't get covid

60 don't get covid

75 don't get covid

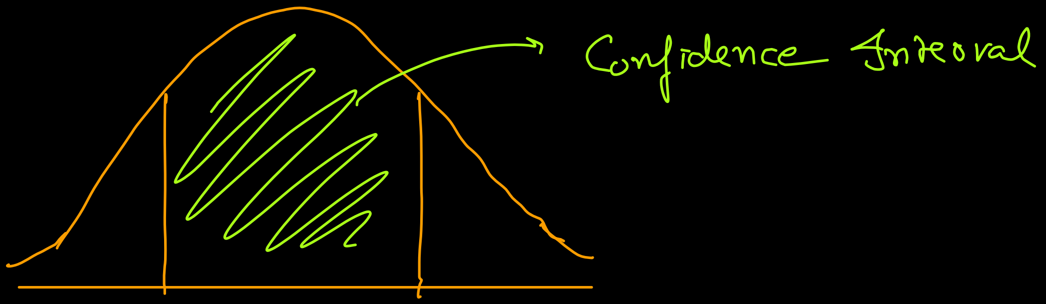
domain expert

medical cases we normally have a very high α

Confidence Interval

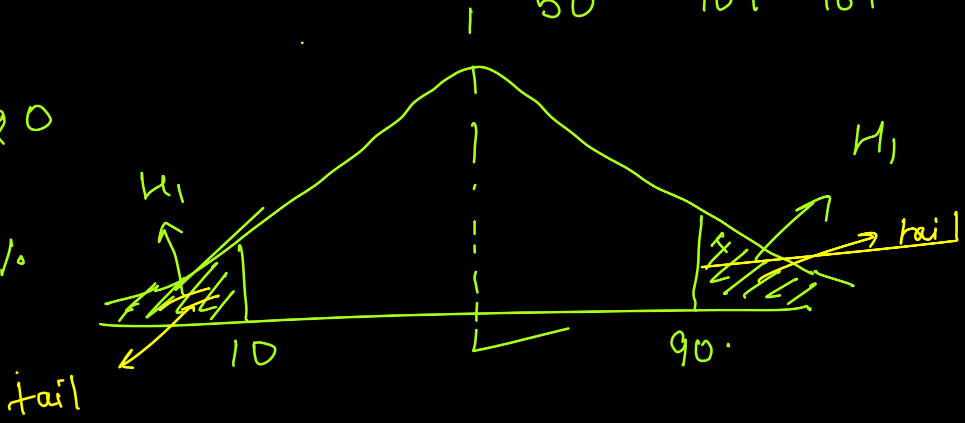
⇓

some ~~arrange~~ range, within which if our value falls then we accept null hypotheses



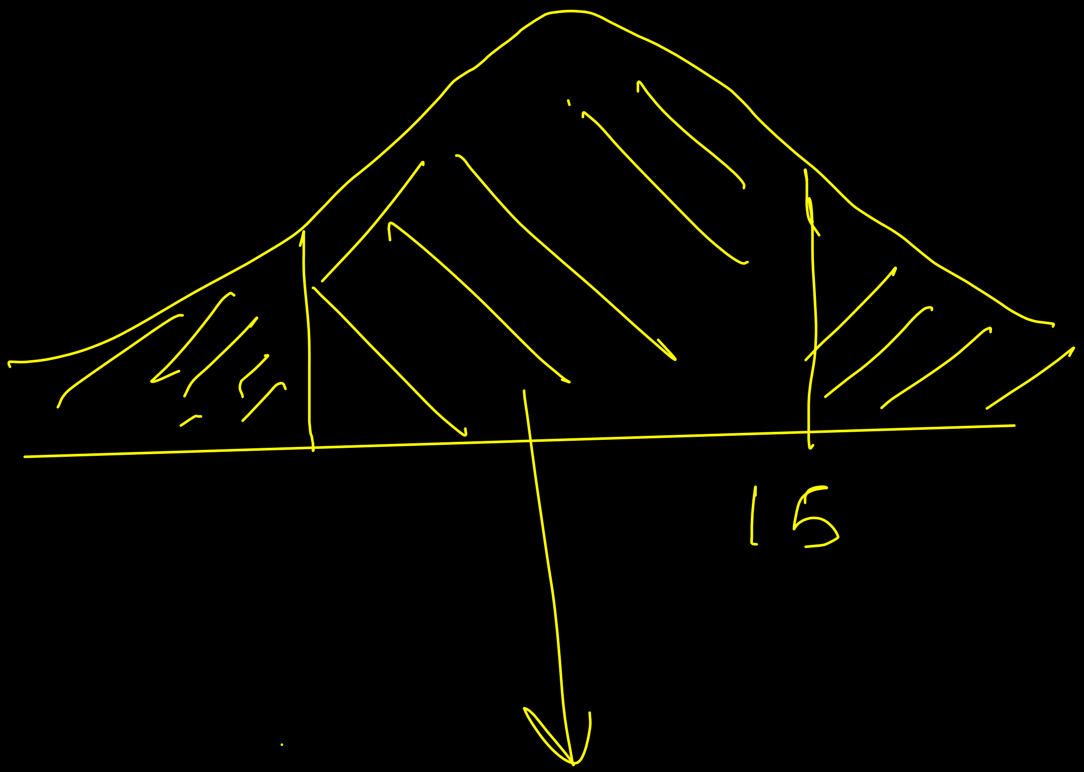
100 → 50 50   90H 10T   10H 90T

$\alpha = .20$ or 20%



$$C I. = 1 - \underbrace{0.1 - \underline{0.1}}$$

$$= 80\%$$



$$\left.\begin{matrix} 854H \\ 15T \end{matrix}\right\} \Rightarrow$$ unfair coin

$\alpha = .3$

$604H\ 40T \Rightarrow$ unfair