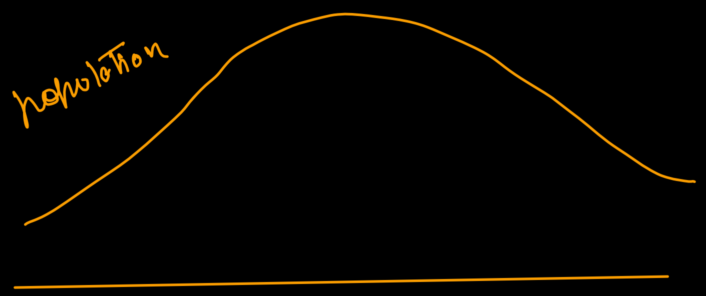
Inferential Stats

1. Central limit theoram
2. Type 1 - Type 2 error
3. One tail vs Two tail
4. z-test , t-test , pvalue
5. z-test for proportional population
6. Chi-square
7. Annova

Central Limit theoram

If you have a distribution (Gaussian or non gaussian)



If we now take multiple samples from this distribution.

$S1 = \{x_1 \cdots x_n\} \rightarrow \bar{x}_1$

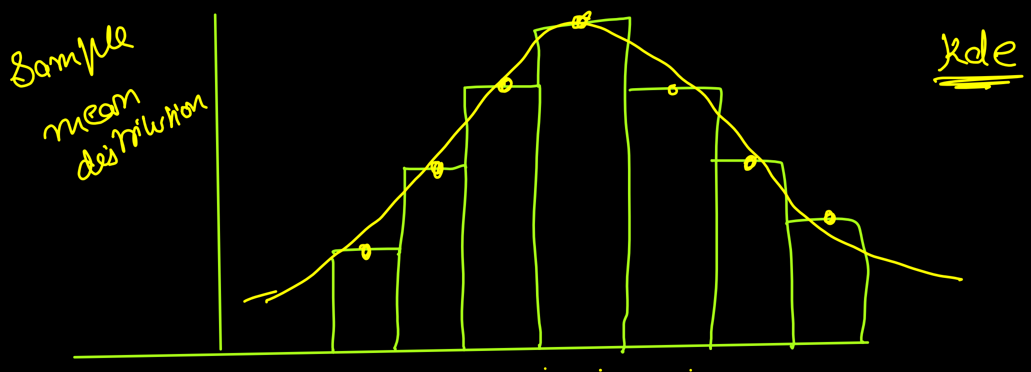
$S2 = \{x_2 \cdots x_n\} \rightarrow \bar{x}_2$

$S3 \rightarrow \bar{x}_3$

$\vdots$

$S_i \rightarrow \bar{x}_n$

then, if we now plot all sample means on a graph. that will actually be a normal / gaussian distribution



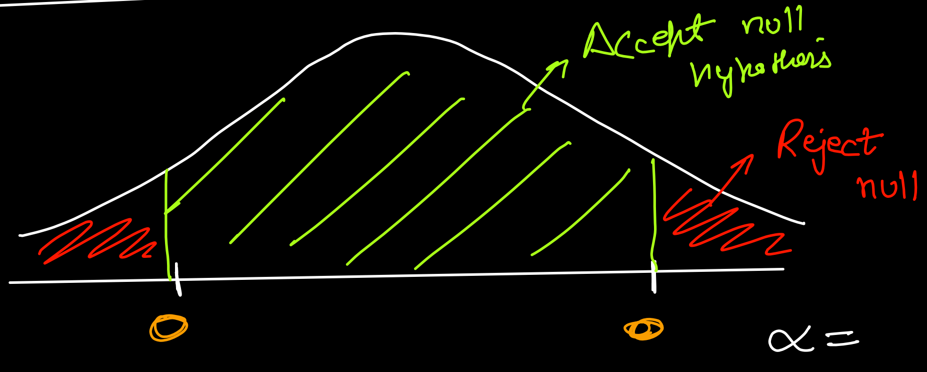
$n \geq 30$ usually applies for a non gaussian distribution

n is the sample size i.e. elements in each sample

Works all the time.

In inferential statistics, with any sample data, we will thus be focusing on normal / Gaussian distribution.

Confidence Interval



A confidence interval from sample data is a range of values which are likely to have pop. parameter at some confidence level.

Point estimation

The value of any statistic that estimate the value of a parameter is called point estimate.

$\bar{x}$ (sample) $\quad\quad$ μ (pop)

2.5, ~~2.8~~, 3.1, ~~3.3~~ $\quad\quad$ 3

$\bar{x}$ is point estimate of μ

Inferential stats :

We have some sample data from which we get value of pop.

This point estimate can help us get the value of pop. parameter.

* estimate of pop. = Point estimate ± margin of error.

Q: On a Data science test of a company the σ pop is known to be 100. [80]

A sample of 25 test takers has a mean of 520.

Construct a 95% CI about the mean?

488.64

551.36

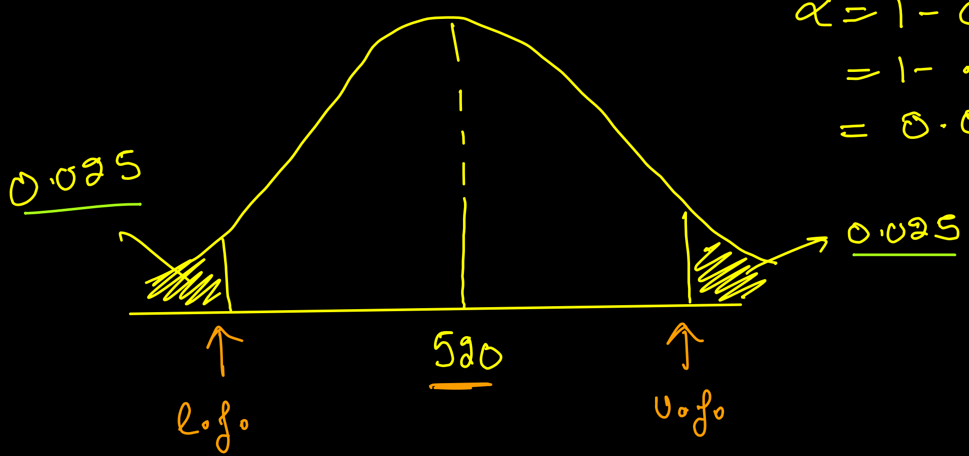
$\sigma = 100$ $\quad \bar{x}$ (sample) $= 520$

$n = 25$ $\quad CI = 95\%$ $\quad \alpha = 0.05$

$\alpha = 1 - CI$

$= 1 - .95$

$= 0.05$



Population σ is given $\Rightarrow$ {z-test}

$\bar{x} \pm$ margin of error

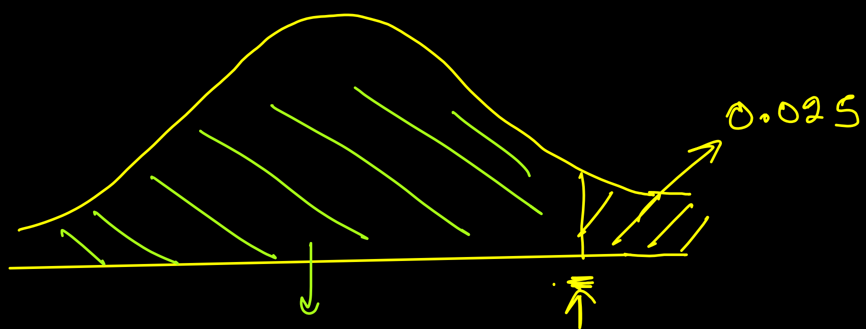
margin of error = $Z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$

$\sigma \rightarrow$ stand. deviation

$\sqrt{n} \rightarrow$ sample size

Z-test is a test which is used to compare your sample mean to your population mean when population σ is known & sample size is large enough

$Z_{\alpha/2} = Z_{0.5/2} = Z_{0.025}$



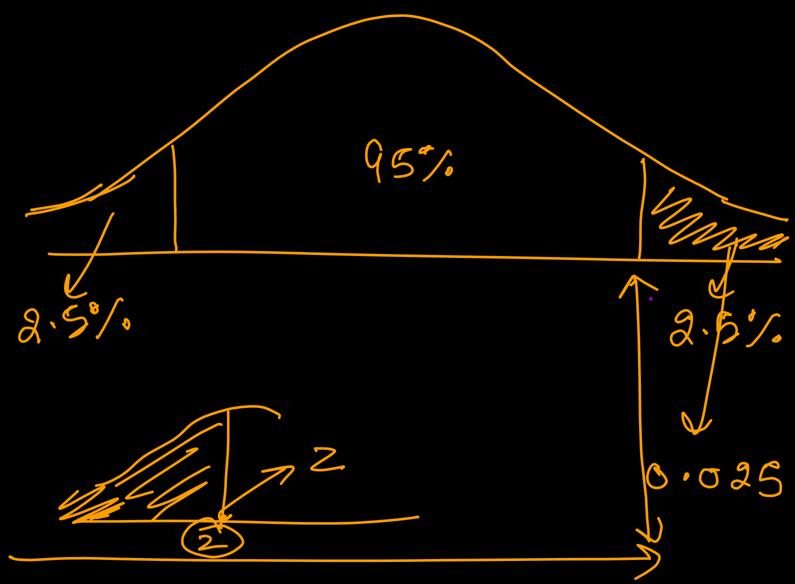
$1 - 0.025 = 0.975$

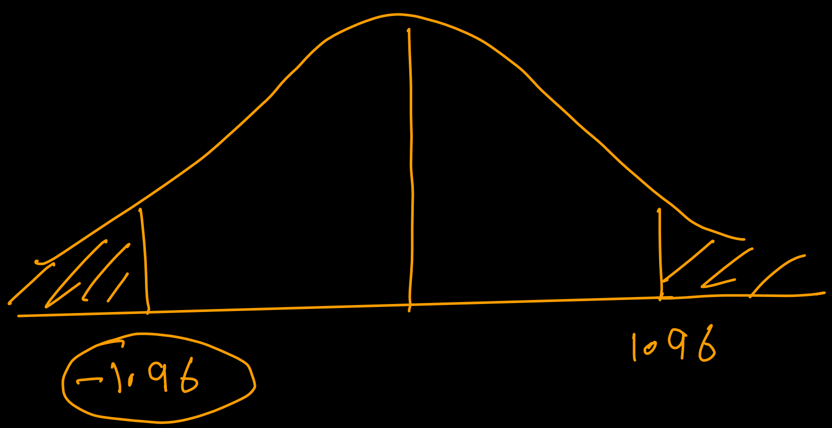
body area = 0.975

tail area = 0.025

C.I = 95%

$\alpha = 5\%$





margin of error

$$= 1.96 \times \frac{100}{\sqrt{25}} = = z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

$$= 39.2$$

population estimate $= 520 \pm 39.2$

$L.F = 520 - 39.2 = 480.8$

$U.F. = 520 + 39.2 = 559.2$

$\Rightarrow$ we are 95% confident that the population mean is b/w 480.8 & 559.2.

σ standard deviation = 80

T - test

t-test is used to compare mean of 2 groups. normally when sample size is less and standard deviation.

50gm

$\Rightarrow$ 25 sample

48gms

$\sigma_s =$

Drug A

50%

Drug B

55%

z-test	t-test
① $n \geq 30$	① sample σ
② population or sample σ is given	② $n < 30$

Q- In a data science test, a sample of 25 test takers have a mean of 520 and a $\underline{\sigma}$ of 80

(s) $\Downarrow$ sample

Construct a 95% CI about mean?

$n = 25$ $\quad \bar{x} = 520$ $\quad \sigma_{sam} = 80$
(s)

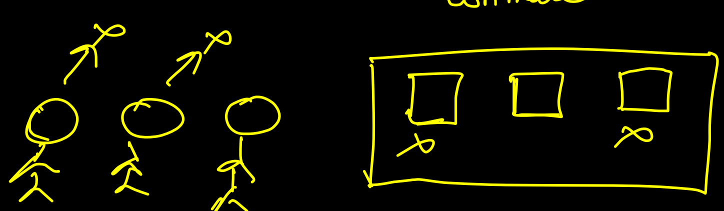
population estimate $= \bar{x} \pm$ error

$\Downarrow$

$$\boxed{t\text{-test} = t_{\alpha/2} \frac{s}{\sqrt{n}}}$$

We need to find degree of freedom in t-test which is $\underline{n-1}$. $\Rightarrow$ dof

dof = how many choices we have to estimate



1
1

$t_{\alpha/2} = t_{0.025}$

$\underline{dof} = n-1 = 24 = 25-1$

$t_{0.025} = 2.064$

$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$

$520 - 2.064 * \frac{80}{5}$

$\Rightarrow 486.97$

We calculate dof. <u>for checking</u> value in t-table

but error is $\boxed{t_{\alpha/2} \frac{s}{\sqrt{n}}}$

Q:- estimate the avg. wt of sharks in the ocean?

Type 1 error vs Type 2 error

H_0 : coin is fair (null)

H_1 : Coin is not fair (alternate)

experiment will be done

null true or null false

Real

Outcome

O1 : we reject the null hypo. when in reality it is false

O2 : we reject null hyp. when in reality is is true → type 1 error

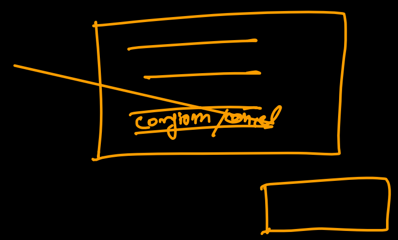
O3 : we accept the null hyp. when in reality it is false. → type 2 error.

O4 :- we accept the null hyp. when in reality it is true

Ho

real / predict	Cancer	no Cancer
Cancer	✓	✗
No Cancer	$\boxed{\times}$	✓

real / Pred	Confirm	Cancel
Confirm		
Cancel		



Pred \ Real	+ve	-ve
+ve		Type 1 error
-ve	Type 2 error	

Ho ⇒ null hypothesis, it assumes that there is no statistical diff b/w groups

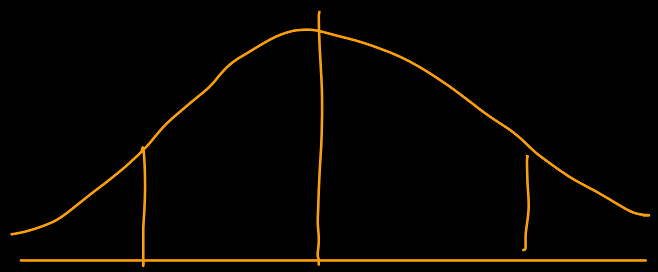
	null hyp T	null hyp F
Reject null hyp.	Type 1 error	✓
Fail to reject null	✓	Type 2 error

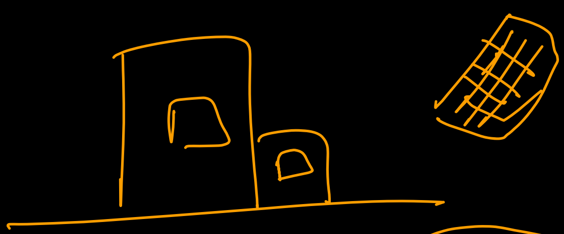
One tail and 2 tail test

Q ⇒ A factory is making chocolate bars which have a mean wt of 50gm.

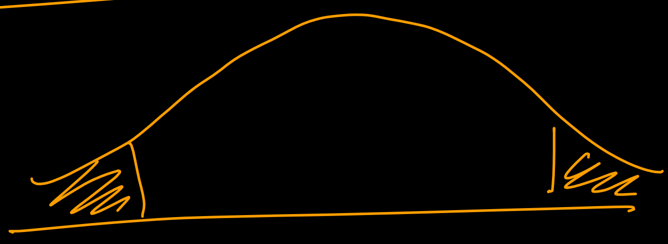
You took 25 samples & observed mean to be 49 gms. with sd. 25.

Is the avg. wt actually 50 gm

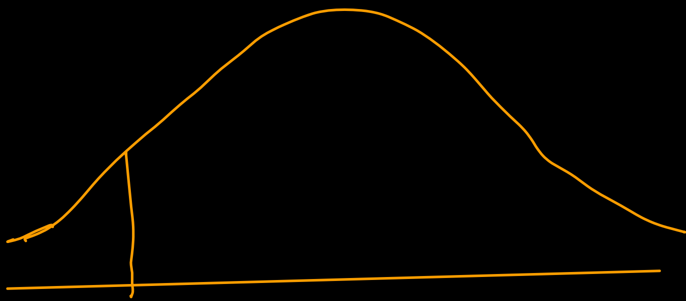




50gm



$\mu \neq 50$



$\mu < 50$

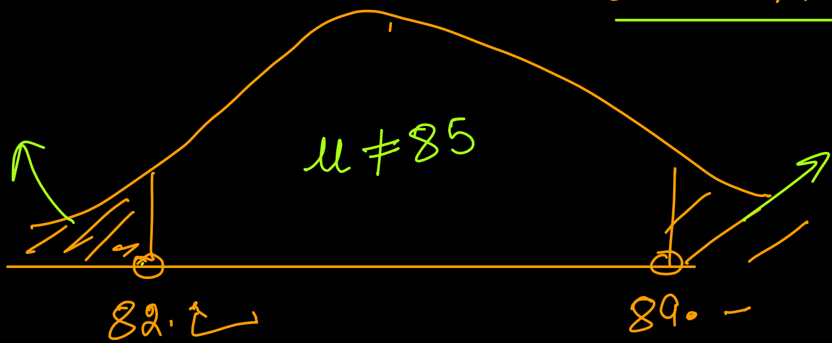
NSIT

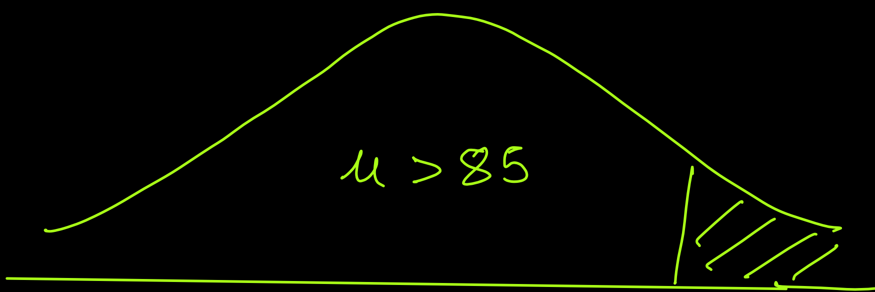
$\boxed{p.r = 85\%}$

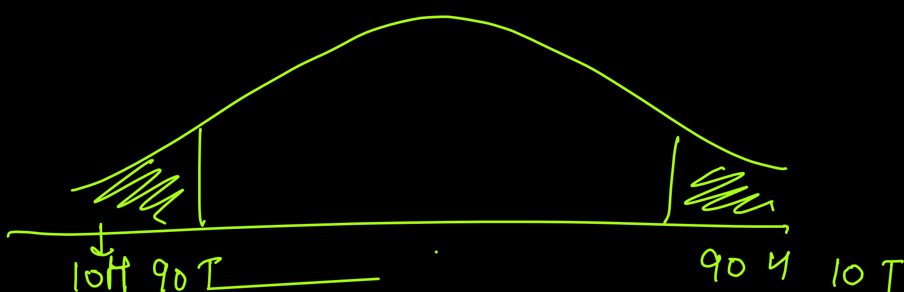
200 student sample

p.r (sample) = 88%

$\alpha = 4\%$







Q ⇒ A factory has a m/c that fills 80 ml of medicine in a bottle

Using 40 samples, we measure the avg quantity dispense / filled by m/c to be 78 ml with a s.d of 2.5.

Q ⇒ a) State null and alternate hypothesis

b) At 95% C.I – is there enough evidence that m/c is working properly?

Step 1 $\quad H_0 \Rightarrow \mu = 80$

$H_1 \Rightarrow \mu \neq 80$

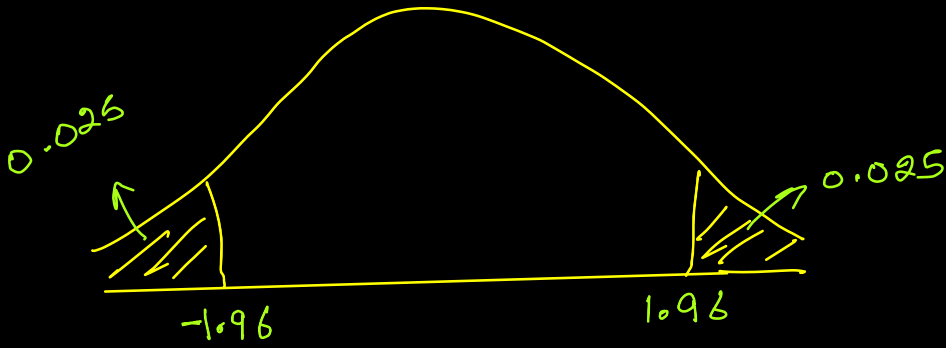
$n = 40 \quad \bar{x} = 78 \quad \sigma = 2.5 \quad \mu = 80$

$C.I. = 95\% \quad \alpha = 0.05$

⇒ z-test

⇒ 2-tail

Step 3 :- Decision boundary



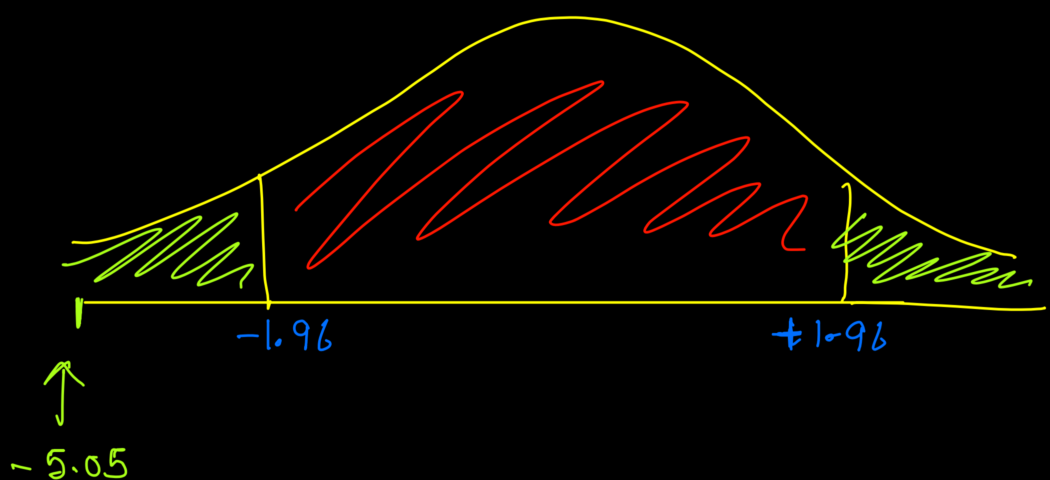
④ Calculate test statistics.

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

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

$$Z = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$ → for sample

$$= \frac{78-80}{2.5/\sqrt{40}} = \boxed{-5.05}$$



5. Conclusion / State the result:

If $z = -5.05$ is less than -1.96 or greater than $+1.96$ then we reject null hyp. with 95% confidence

$\Rightarrow$ There is a fault in the m/c.