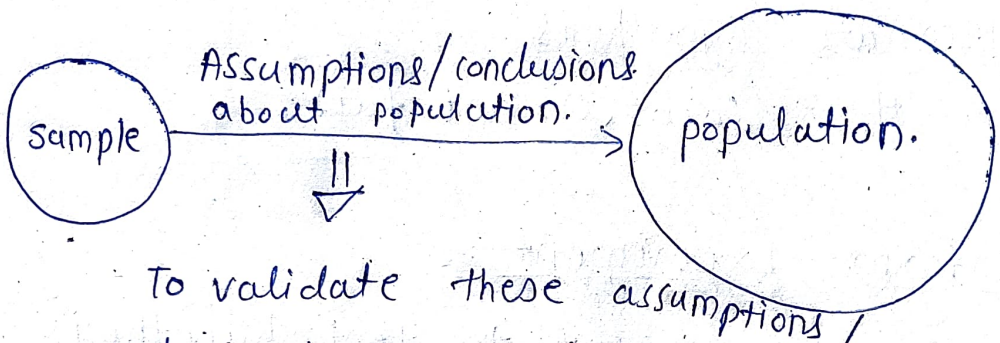
Stats Basics

* Inferential statistics

Inferential stats uses measurements from sample of subjects in the experiment to compare the samples (treatment groups) and make generalizations about larger population of subjects.



To validate these assumptions / conclusions, we use hypothesis testing

* Hypothesis testing

Hypothesis testing in stats is a way to test the results of an experiment / survey to see if we have a meaningful result or not.

* Steps to do Hypothesis testing

① Figure out your null and alternate hypothesis.

a) Null hypothesis (H_0).

Null hypothesis is always the accepted fact. Examples of null hypothesis that are generally being accepted as true are:

1) DNA is shaped like a helix.

2) There are 8 planets in solar system

b) Alternate Hypothesis (H_1)

Alternate hypothesis is accepting opposite of null hypothesis.

Eg: To determine whether a coin is fair or not?

1) Null hypothesis

H_0 = coin is fair

2) Alternate hypothesis

H_1 = coin is not fair.

② Perform Experiment

→ perform experiment to test hypothesis using below mentioned test.

1) z test 2) t-test

3) Chi-square test 4) Annova test (f test).

→ Find out confidence Interval values or Test statistic

→ Find one tail or two tail problem.

③ conclusions

① Fail to reject null hypothesis

or

② Null hypothesis rejected.

Example: A coin is tossed 100 times. Determine whether coin is fair or not.

* criterid to check if a coin is fair is: if out of 100 Head comes up in range of 80 to 20, then coin is fair or vice versa..

① H_0 (null hypothesis)

H_0 = coin is fair

H_1 (Alternate hypothesis)

H_1 = coin is not fair.

② perform experiment.

caseI when experiment is performed, it is observed that 75 times head comes up.

③ conclusion.

∵ 75 lies in range of 20 to 80.

∴ Null hypothesis cannot be rejected

⇒ coin is fair.

caseII when experiment is performed, it is observed that only 10 times head appears.

conclusion

∵ 10 doesn't lie in range 20 – 80

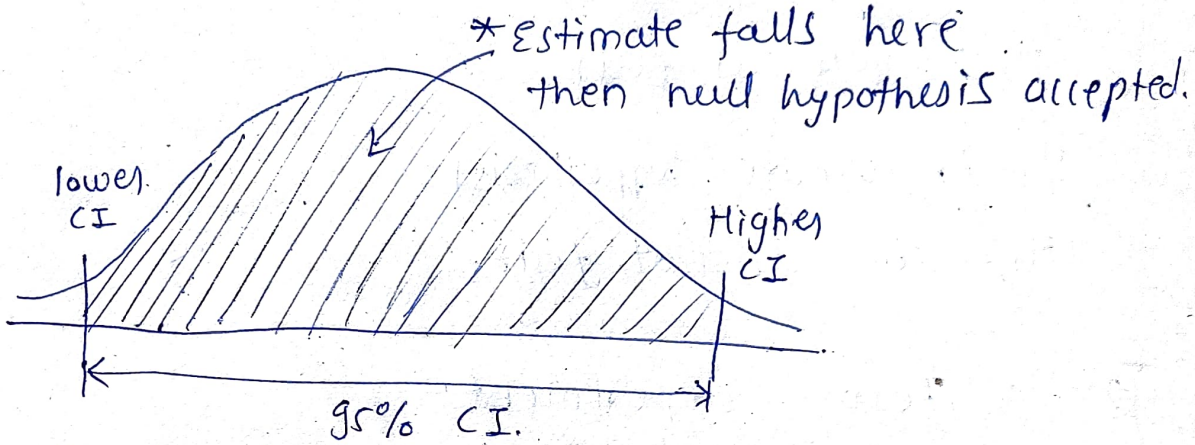
∴ Null hypothesis is rejected.

⇒ coin is not fair

* Confidence Interval (C.I)

Confidence ~~level~~ Interval is the mean of estimate plus and minus the variation in estimate. This is the range of values you expect your estimate to fall between, if you redo the experiment, within a certain level of confidence.

Eg: 95% confidence Interval



* if estimate falls outside C.I range. then null hypothesis is rejected.

* Significance value (α).

$$\alpha = 1 - CI$$

Eg: 1) $CI = 95\%$

$$\therefore \alpha = 1 - \frac{95}{100} = 0.05$$

2) $CI = 90\%$

$$\alpha = 1 - \frac{90}{100} = 0.1$$

3) $CI = 0.8$

$$\alpha = 1 - 0.8 = 0.2.$$

* one tailed and 2 tailed tests

* one tailed Test

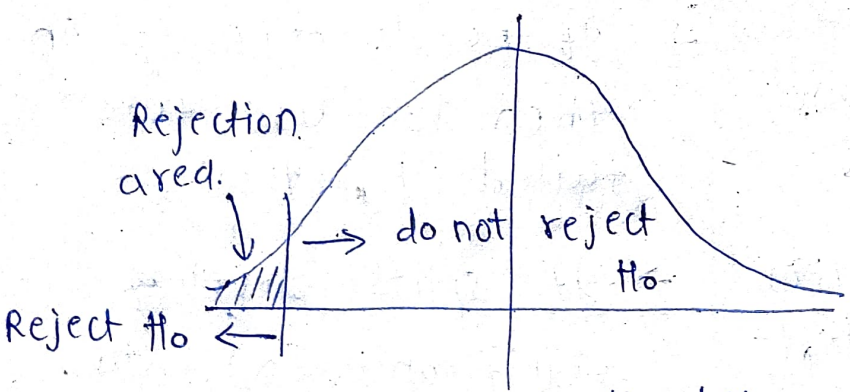
→ It is based on uni-directional hypothesis, where rejection area is on one side of sample distribution.

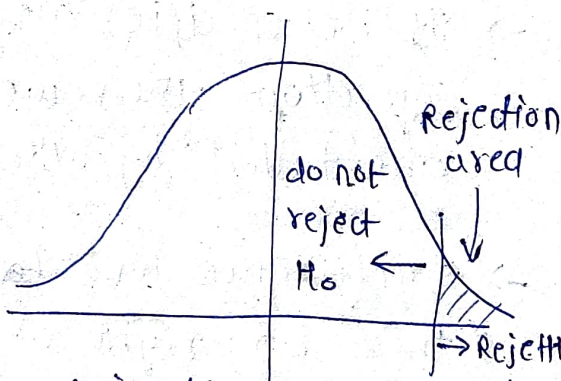
→ It determines whether a particular population parameter is larger or smaller than the predefined parameter.

Eg: A drug company manufactures a drug that is equally effective to existing drug (assumption). But your drug is cheaper.

① Since the drug is cheaper so it is best to test that the drug is less effective or not.

so this is example of one tailed Test.





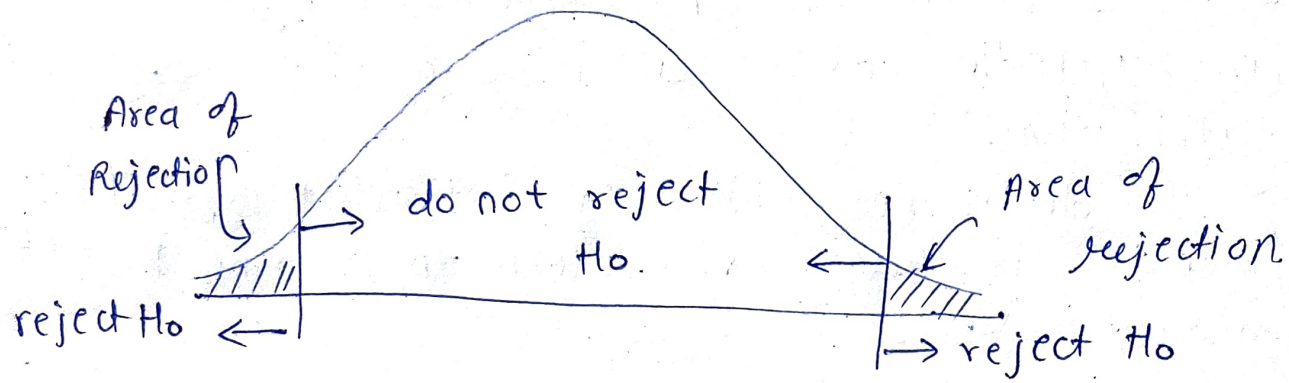
Above graphs indicates one tailed Tests graphs.

* Two tailed Tests

→ It is also called non-directional hypothesis

→ It is used for checking whether the sample is greater or less than a range of values

Eg: New loan bill for farmers is passed. To check whether new loan bill increases or decreases the loan of farmer.



one tailed	Two tailed.
→ when alternate hypothesis is one tailed, either right or left tailed then it is used.	→ when alternate hypothesis is two tailed.
→ we use either < or > sign for H_1	→ we use ≠ sign for H_1
→ If H_1 specifies any direction then we use one tailed hypothesis	→ If no direction given then we use two tailed hypothesis
→ significance level (α) lies either on right or left side of sampling distribution	→ It splits level of significance (α) into half.
→ It is used to check whether one mean is different from another mean or not	→ It is used to check whether two means different from one another or not

Ex:

In a town colleges have average placement rate of 85%. A new college is opened. To Test.

1) if new college has different placement rate:

⇒ no direction given ⇒ Two tailed Test.

$\longleftarrow$ 85% $\longrightarrow$

② a) placement rate greater than 85%

85% $\longrightarrow$ ⇒ right tailed Test.

b) placement rate less than 85%

$\longleftarrow$ 85%

⇒ left tailed Test.

* point estimate

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

Eg:

	sample	population	Inferential stats statistic	parameter.
1)	$\bar{x}$	μ	$\bar{x}$	μ
2)	s	σ	s	σ
3)	s^2	σ^2	s^2	σ^2

ie sample mean, variance, std deviation are used to estimate population mean, variance std deviation respectively

$$\text{parameter} = \left(\begin{matrix}\text{point}\\ \text{estimate}\end{matrix}\right) \pm \left(\begin{matrix}\text{margin}\\ \text{of}\\ \text{error}\end{matrix}\right)$$

Eg: for mean.

$\bar{x}$: sample mean μ: population mean

CI: confidence Interval.

***

$$\left(\begin{matrix}\text{CI range}\\ \text{for mean}\end{matrix}\right) = \left[\begin{matrix}\text{Lower CI} \\ \bar{x} - \text{MOE} \\ \mu_{Lower}\end{matrix} \quad \text{to} \quad \begin{matrix}\text{Higher CI} \\ \bar{x} + \text{MOE} \\ \mu_{upper}\end{matrix}\right]$$

MOE: margin of error

Note: margin of error (for Z test).

$$\text{MOE} = Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

where $Z_{\alpha/2}$ = critical value of Z distribution at significance level $\alpha/2$.

σ = population standard deviation

n = no. of samples.

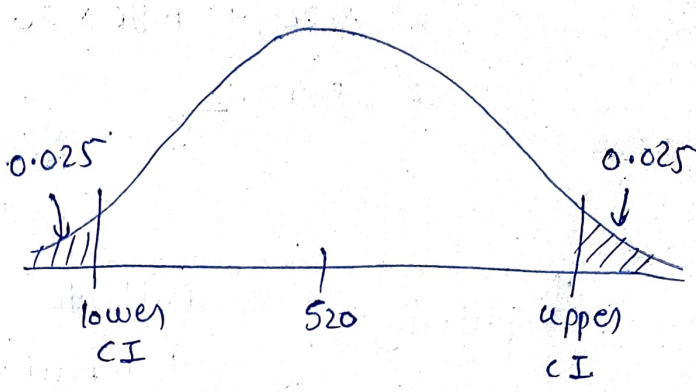
Qu: On a quant test of CAT Exam, a sample of 25 test takers has a mean of $20 with population std deviation of 100. Construct a 95% CI about the mean.

Ans: $\bar{x} = 520$ $\sigma = 100$ C.I = 95%, $n = 25$

$\alpha = 1 - C.I = 1 - 0.95 = 0.05$

* 95% about mean ⇒ Two Tailed Test

$\alpha/2 = \frac{0.05}{2} = 0.025$



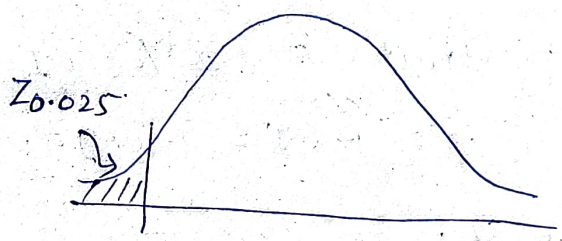
Now: $Z_{0.025}$ can be calculated using z table

① Negative z table (ztable.net)

$|Z_{0.025}| = |-(1.96 + 0.06)|$

$= |-1.96|$

$= 1.96$.



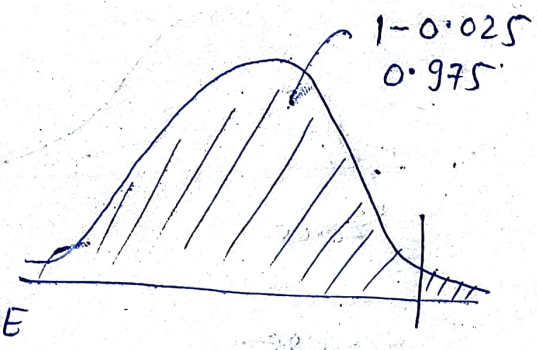
* Here taking mod value since (±) sign already included in formula

② positive z table

$Z_{0.025} = Z_{0.975}$

$= 1.9 + 0.06$

$= 1.96$.



∴ Lower CI = point estimate − MOE

$= \text{mean} - \text{MOE}$

$= 520 - Z_{0.025} \frac{\sigma}{\sqrt{n}}$

$= 520 - (1.96) \times \frac{100}{\sqrt{25}}$

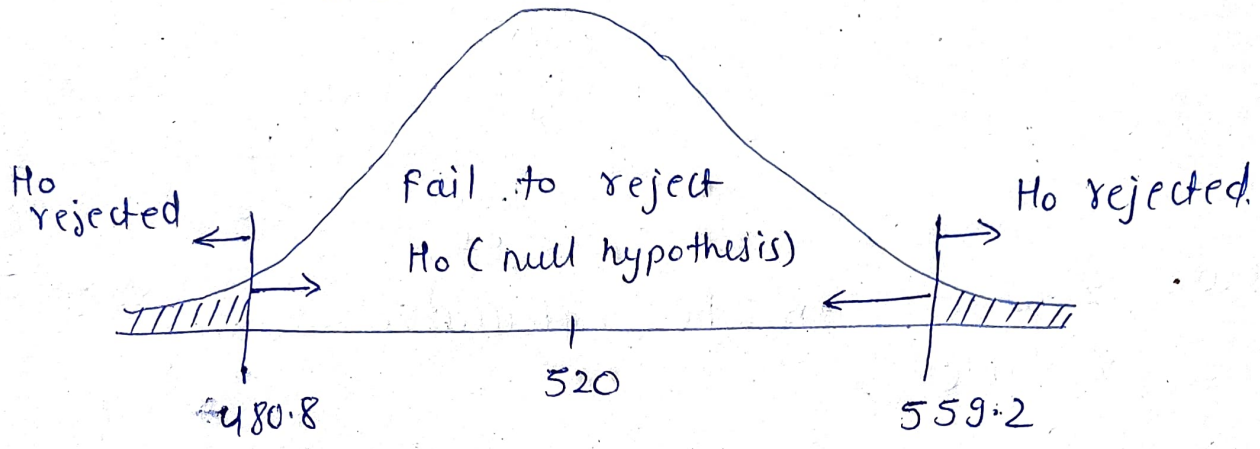
lower CI = $520 - 1.96 \times 20$

$= 480.8$

upper CI = point estimate + MOE

$= 520 + 1.96 \times 20$

$= 559.2$



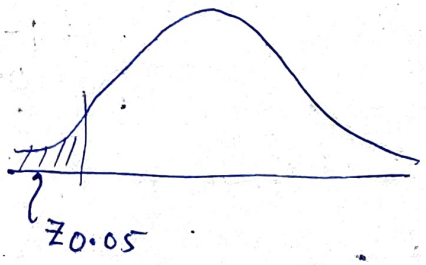
Que: $\bar{x} = 480$ $\sigma = 85$ $n = 25$, CI = 0.90

construct a 90% CI about mean.

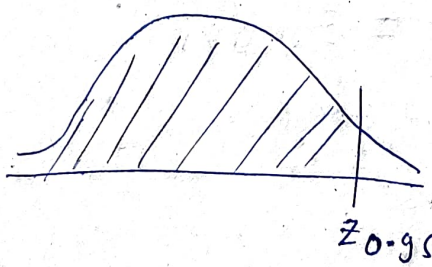
$\alpha = 1 - CI = 1 - 0.9 = 0.1$

$MOE = Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} = Z_{0.1/2} \times \frac{85}{\sqrt{25}}$

$= Z_{0.05} \times 17$



or



~~$Z_{0.05}$~~ =

Now 0.95 lies b/w

0.94950 and 0.95053

ie ⇒ 1.604 and 1.605

	①	variable	②
x	0.94950	0.95	0.95053
y	1.604	$Z_{0.05}$	1.605

$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1)$$ — Interpolation formula (linear)

$$Z_{0.05} = 1.604 + \left(\frac{1.605 - 1.604}{0.95053 - 0.94950}\right) \times (0.95 - 0.94950)$$

$$Z_{0.05} = 1.60448$$

$$MOE = 1.60448 \times 17$$

$$= 27.28$$

$$\text{Lower CI} = \bar{x} - MOE$$

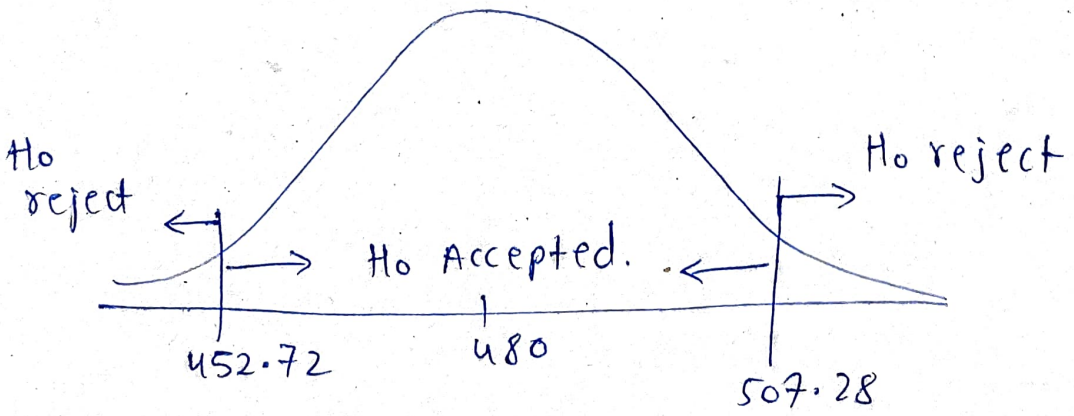
$$= 480 - 27.28$$

$$= 452.72$$

$$\text{upper CI} = \bar{x} + MOE$$

$$= 480 + 27.28$$

$$= 507.28$$



* t-Test

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

where, $\bar{x}$ = mean of sample

$t_{\alpha/2}$ = critical value of t-table at significance value $\alpha/2$.

s = sample standard deviation

n = no. of samples

Also, degree of freedom (dof).

$$dof = n-1$$

Suppose 3 seats are available

P1 P2 P3

[S1] [S2] [S3]

3 person
- 1st person (P_1) → 3 choices to sit
- 2nd person (P_2) → 2 choices to sit
- 3rd person (P_3) → no choice must sit on last seat.

$\therefore dof = n-1$ ⇒ 3-1 = (2) only 2 people can choose.

Que: on a quant test of CAT exam, a sample of 25 Tests takers has a mean of 520 with a sample standard deviation of 80. construct 95% CI about mean.

Ans: 1) since about mean ⇒ Two tailed Test.

2) Sample std dev given ⇒ t-test and $n < 30$

$\bar{x} = 520$, $s = 80$, $n = 25$ $CI = 0.95$

$\alpha = 1 - 0.95 = 0.05$

$MOE = t_{\alpha/2} \left(\frac{s}{\sqrt{n}}\right)$

$= t_{\left(\frac{0.05}{2}\right)} \left(\frac{80}{\sqrt{25}}\right)$

$= t_{0.025} \times 16.$

$dof = n - 1$

$= 25 - 1$

$\boxed{dof = 24.}$

from t table (ttable.org)

$t_{0.025}$ or $t_{0.05}$ and $dof = 24$

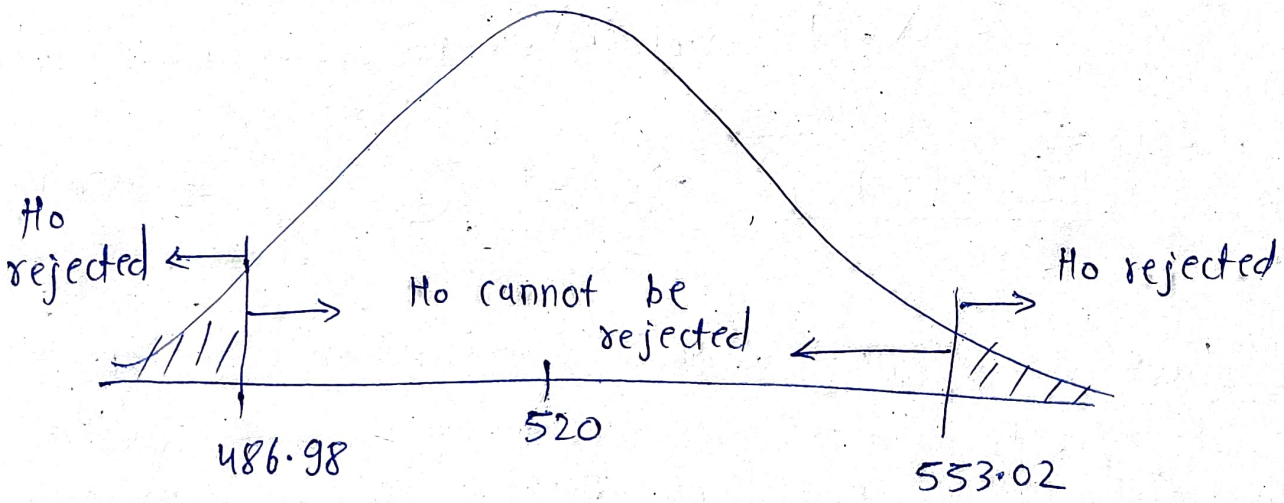
$t_{0.025} \Rightarrow 2.064$

$MOE = 2.064 \times 16$

$= 33.024$

lower CI $= 520 - 33.024 = 486.976$

upper CI $= 520 + 33.024 = 553.024$



$CI = [486.98 \text{ to } 553.02]$

Note:

1) when to use Z test

a) we know population std deviation (σ) OR

b) we do not know population std deviation but our sample is large ($n \geq 30$).

2) when to use t test

a) we do not know population std. deviation (σ)

b) our sample size is small ($n \leq 30$)

c) we know sample std deviation (s).

Questions on Hypothesis testing

Que: 1 A factory has a m/c that fills 80ml of Baby medicine in a bottle. An employee belives that average amount of baby medicine is not 80 ml. using 40 samples he measures the avg. amount of medicine dispersed by machine is 78 ml. with a std deviation of 2.5

1) state null (H_0) and alternate (H_1) hypothesis

2) At 95% CI, is there enough evidence to support machine is working properly or not.

Given: $\mu = 80$, $n = 40$, $\bar{x} = 78$, $s = 2.5$

$CI = 0.95$

$\because$ $n > 30 \Rightarrow$ Z test

1) $H_0 \Rightarrow \mu = 80$ (m/c is working properly)

$H_1 \Rightarrow \mu \neq 80$ (m/c is not working properly).

$\Rightarrow$ since no specific direction is given

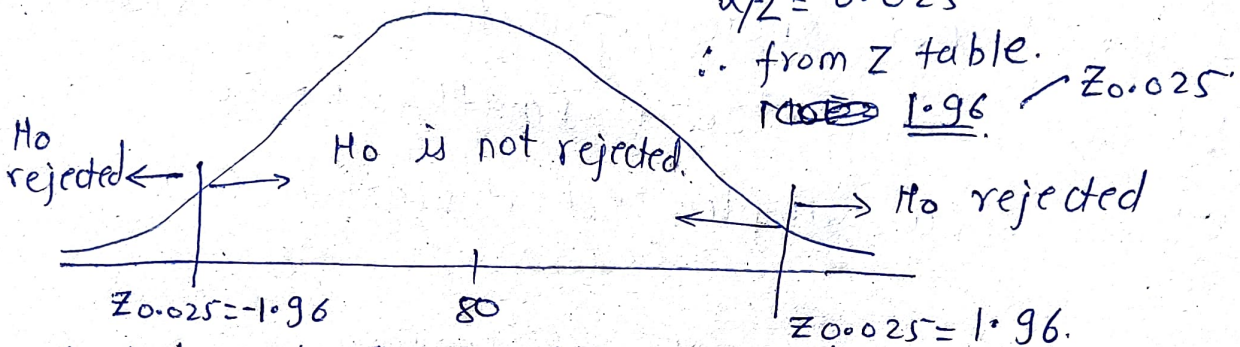
$\leftarrow 80 \rightarrow$ $\therefore$ Two tailed Test.

2) $CI = 0.95 \Rightarrow \alpha = 1 - 0.95 \Rightarrow 0.05$

$\alpha/2 = 0.025$

$\therefore$ from Z table.

1.96 $\checkmark Z_{0.025}$



* calculate test statistic

Z test.

$$Z_{score} = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{78 - 80}{\frac{2.5}{\sqrt{40}}}$$

Standard error $\leftarrow \left(\frac{s}{\sqrt{n}}\right)$

$$Z_{score} = -5.0596$$

Now Z doesn't lie in range -1.96 to 1.96.

3) <u>conclusion</u> (Decision rule).

if $Z = -5.05$ is less than -1.96 or greater than $+1.96$, Reject the null hypothesis with 95% CI values.

ie: There is some fault in machine

Ans: $\mu = 100$, $\sigma = 15$, $n = 30$, $\bar{x} = 140$, CI = 95%.

$\because$ σ given and $n \geq 30$

$\Rightarrow$ z test

-ve ← No effect 100 → +ve

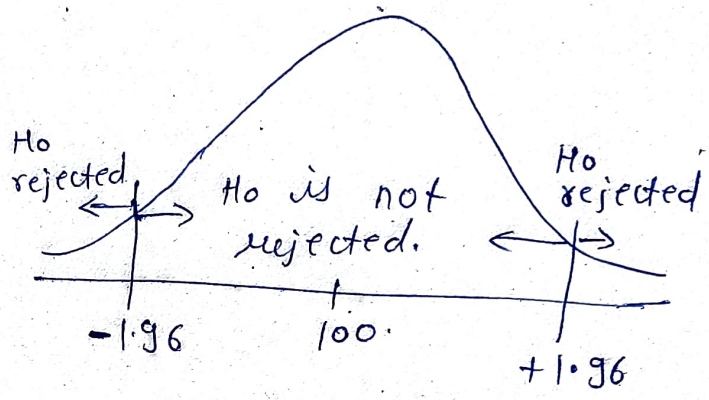
① $H_0 \Rightarrow \mu = 100$ (No effect) Two tailed Test

$H_1 \Rightarrow \mu \neq 100$ (increase or decrease in IQ)

② $\alpha = 1 - 0.95 = 0.05$

$\alpha/2 = 0.025$

$Z_{0.025} \Rightarrow \pm 1.96$.



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

$$= \frac{140 - 100}{15/\sqrt{30}} \Rightarrow Z_{score} = 14.60$$

③ conclusion

$\because$ 14.60 doesn't lie in blw -1.96 to $+1.96$

$\therefore$ H_0 is rejected.

$\therefore$ medication may increase or decrease the IQ.