Computer Science & DA

Probability and Statistics

SAMPLING THEORY AND DISTRIBUTION

Lecture No. 02

By- Dr. Puneet Sharma Sir

Recap of previous lecture

Topic Sampling theory (Basics)

Topics to be Covered

Topic Z - test

SAMPLING (BASICS)

Population → The group of individual under consideration (whether finite or ∞) is called pop.

Sample → A small set from population is called sample (it is always finite) & this process is called sampling.

Parameters → Numerical quantities from which we can understand population are called parameters

Statistic → " " " " " sample called statistic

for eg μ & σ are the parameters for population
while $\bar{x}$ & S " statistic " sample

with the help of statistic, we will try to understand pop also that's why sampling plays an imp role.

Proportion → The Ratio of Successful Events with Total Events known as proportion

for, eg A Coin is tossed 10 times and we are getting Head exactly 3 times then

$$\text{proportion of } H = ? = \frac{\text{Success}}{\text{Total}} = \frac{3}{10}$$

eg In a sample of 400 Children, There are exactly 210 Boys then

$$\text{Prop of Boys in Sample} = \frac{\text{Success}}{\text{Total}} = \frac{210}{400} = 0.525$$

while prop of Boys in Population $= \frac{500\,\text{cr}}{1000\,\text{cr}} = 0.500$

i.e prop in Sample $= \frac{x}{n}$ & prop in Population $= \frac{X}{N} \simeq$ Probability

i.e Sample Prop $= \frac{x}{n}$ $(\tilde{p})$ & Population Prop $= \frac{X}{N} = p_0$

(*) <u>Standard Error</u> → it is the S.D of statitic (in sample)

SE of Mean = ?

$$\boxed{SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}}$$

&

SE of proportion = ?

$$\boxed{SE(\tilde{p}) = \sqrt{\frac{\tilde{p}\tilde{q}}{n}}}$$

where $\tilde{p} = \frac{x}{n}$, $\tilde{q} = 1 - \tilde{p}$

<u>Most Probable limits for μ & p_0</u> →

$$\mu = \bar{x} \pm 3.SE(\bar{x}) \quad \& \quad p_0 = \tilde{p} \pm 3\,SE(\tilde{p})$$

$$\bar{x} - 3SE(\bar{x}) \le \mu \le \bar{x} + 3SE(\bar{x}) \quad \& \quad \tilde{p} - 3SE(\tilde{p}) \le p_0 \le \tilde{p} + 3SE(\tilde{p})$$

[NAT]

#Q. A sample of 400 members has mean 4.0. If the population is normal with standard deviation 2.6 and it's mean is unknown than find the most probable limits for population mean.

$n = 400$, $\bar{x} = 4$

$\sigma = 2.6$, $\mu = ?$

$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{2.6}{\sqrt{400}} = \frac{2.6}{20} = \frac{1.3}{10} = 0.13$

i.e $3\,SE(\bar{x}) = 0.39$

i.e Most probable limits for μ is $\bar{x} - (0.39) \le \mu \le \bar{x} + (0.39)$

$3.61 \le \mu \le 4.39$

#Q. In a town, 350 out of 600 persons were found to be vegetarian on the basis of this date can we say that majority of population in the town in vegetarian?

$n = 600$, $x = \{\text{Number of vegetarian persons}\}$ → success

proportion of Veg person (in sample) $= \dfrac{\text{No. of success}}{\text{S. size}} = \dfrac{x}{n} = \dfrac{350}{600} = 0.5833$

i.e. Sample prop $(\tilde{p}) = 0.5833$ & $\tilde{q} = 1 - 0.5833 = 0.4167$

S. Error of Sample prop $= SE(\tilde{p}) = \sqrt{\dfrac{\tilde{p}\tilde{q}}{n}} = \sqrt{\dfrac{0.5833 \times 0.4167}{600}} = 0.02$

So Most Probable limits for Population Prop $(p_0) = ?$ $\tilde{p} - 3SE(\tilde{p}) \le p_0 \le \tilde{p} + 3SE(\tilde{p})$

$$0.5833 - 0.0600 \le p_0 \le 0.5833 + 0.0600$$

$$0.5233 \le p_0 \le 0.6433$$

ie p_0 lies with in 52% to 64%

ie population proportion of Vegetarian person lies b/w 52% & 64%

Hence conclusion is "Majority of population in a town is Vegetarian"

#Q. A coin was tossed 400 times and head turned up 210 times. Discuss whether coin is unbiased or not.

$n = 400$, $x = \{$ Number of times Head occurs $\} \to$ success

$\tilde{p}$ = sample prop. of Head $= \frac{x}{n} = \frac{210}{400} = 0.525$

$\tilde{q}$ = " " of failure $= 1 - 0.525 = 0.475$

$SE(\tilde{p}) = \sqrt{\frac{\tilde{p}.\tilde{q}}{n}} = \sqrt{\frac{0.525 \times 0.475}{400}} = 0.025$

So $3\,SE(\tilde{p}) = 0.075$

$\tilde{p} - 3\,SE(\tilde{p}) = 0.450$ & $\tilde{p} + 3\,SE(\tilde{p}) = 0.600$

Most probable limits for p_0

$0.45 \le p_0 \le 0.60$

ie population Prop of Head lies in b/w 45% & 60%

$\underbrace{\text{45\% \& 60\%}}_{\text{Experimental Value}}$

while theoretical Value for $p_0 = \frac{1}{2} = 0.50$ ie 50%.

Hence Coin is Unbiassed

#Q. A die was thrown 9000 times and 1 or 6 was obtained 3120 times can we say that the die is unbiased.

$x = \{$ Number of times 1 or 6 is obtained $\}$ → success

$p_0 = \frac{X}{N} = \frac{fav}{Total} = \frac{2}{6} = \frac{1}{3} = 0.3333$

$\tilde{p} = \frac{x}{n} = \frac{3120}{9000} = 0.3466$

$n = 9000$

$$\boxed{SE(\tilde{p}) = \sqrt{\frac{\tilde{p}\tilde{q}}{n}}} = \sqrt{\frac{0.34 \times 0.66}{9000}} = 0.005$$

Hence Most Probable limits for p_0 is ?

$$\tilde{p} - 3SE(\tilde{p}) \le p_0 \le \tilde{p} + 3SE(\tilde{p})$$

$$0.3466 - 0.015 \le p_0 \le 0.3466 + 0.015$$

$$0.331 \le p_0 \le 0.361$$

∴ p_0 is almost lie in the Range of Most Probable limits so die is Certainly unbiassed

#Q. In previous question, if success is occurring 3240 times then Prove that is biased.

Sol: $n = 9000$, $x = \{$No. of times 1 or 6 is obtained$\}$ → success

$p_0 = \frac{\text{fav}}{\text{Total}} = \frac{2}{6} = \frac{1}{3} = 0.3333$

$\tilde{p} = \frac{x}{n} = \frac{3240}{9000} = 0.360$

$\tilde{q} = 1 - 0.36 = 0.640$

$$SE(\tilde{p}) = \sqrt{\frac{\tilde{p}\tilde{q}}{n}} = \sqrt{\frac{0.36 \times 0.64}{9000}} = 0.005$$

$$\tilde{p} - 3SE(\tilde{p}) \le \text{Pop. Prop of success} \le \tilde{p} + 3SE(\tilde{p})$$

$$0.360 - 0.015 \le p_0 \le 0.360 + 0.015$$

$$0.345 \le p_0 \le 0.375$$

Experimental Value of $p_0 = (0.345, 0.375)$

while Theoretical Value of $p_0 = 0.333$

∵ T. Value lies outside the E. Value so Die is certainly BIASED //

Hypothesis testing (Z-test, t-test, Chi-square test)

① Sample Value ≃ Experimental Value ≃ Exact Value ≃ Observed Value

② Population Value ≃ Theoritical Value ≃ Approx Value ≃ Expected Value

③ Hypothesis → on the Basis of Sample information, we make some assumptions for Population parameter, & these assumptions are known as Hypothesis.

(i) Null Hypothesis (H_0) → it is a kind of statement in which we assume that there is No difference b/w Sample statistic & population parameters

(ii) Alternative Hypothesis (H_1) → Any Hypothesis which is Complementary to Null Hyp is called A-Hyp

for eg H_0: Population Mean is μ_0 then H_1: $\mu \neq \mu_0$ or $\mu > \mu_0$ or $\mu < \mu_0$

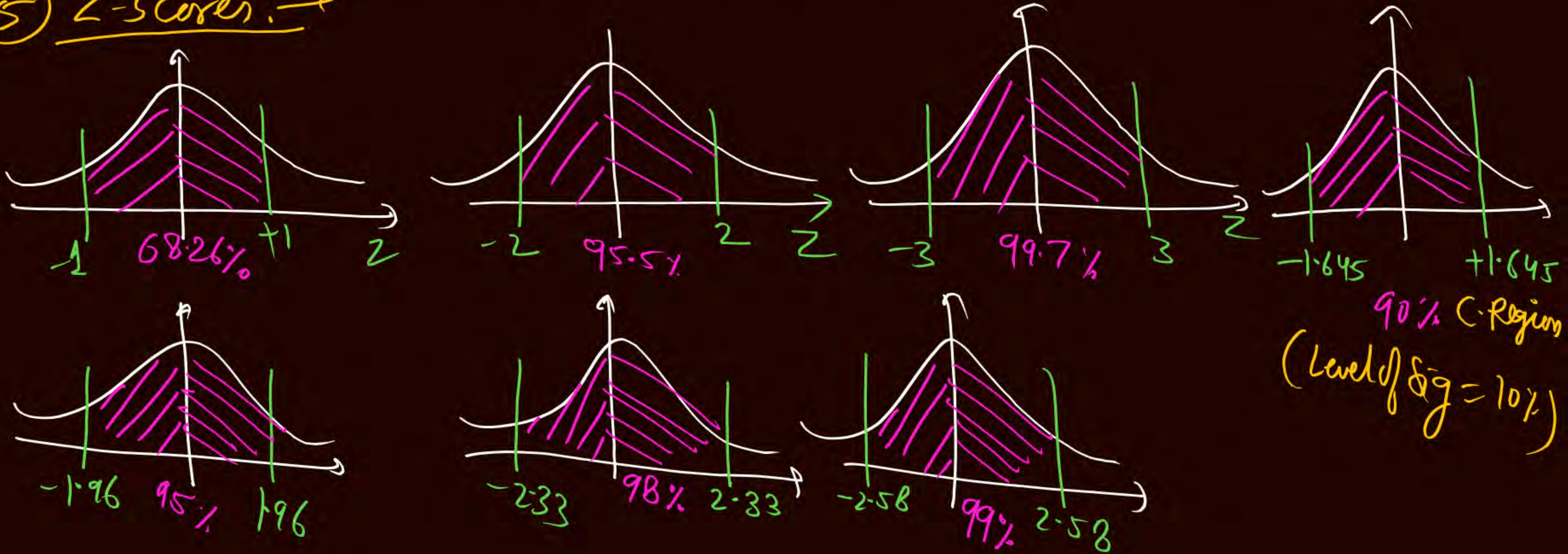
ie $\mu = \mu_0$

(4) Error in Sampling → While Sampling we may commit following types of Errors;

(i) Type 1 Error → H_0 is rejected when it is True (Producer's Risk)

(ii) Type 2 " → H_0 is accepted when it is false (Consumer's Risk)

(5) Z-Scores :→



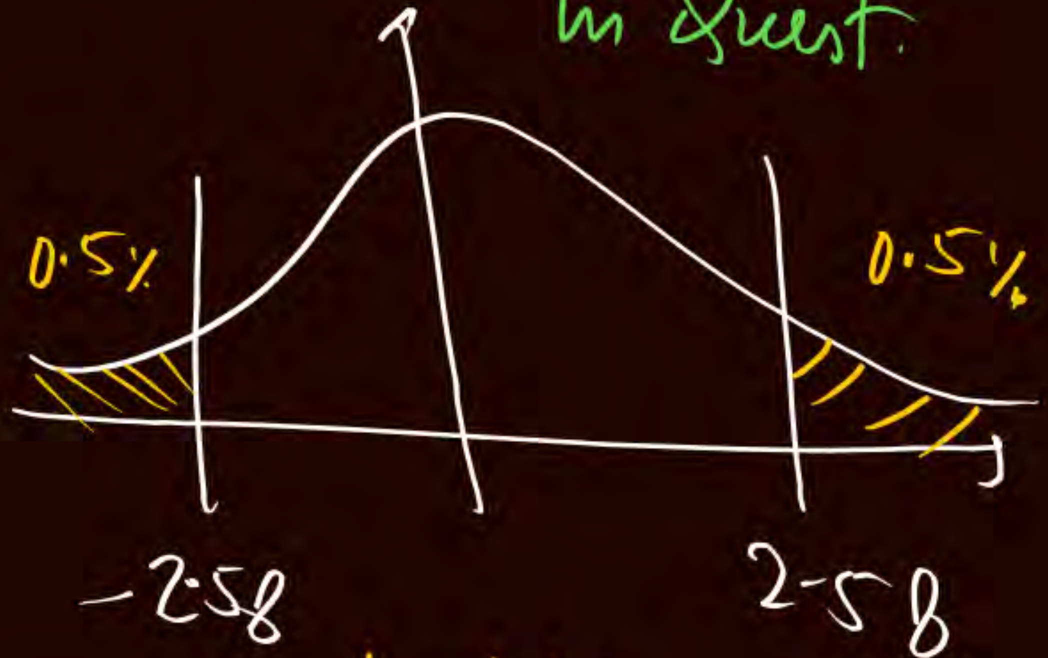
(*) Acceptance Region → (Confidence Region) → The Region in which H_0 is accepted known as A-Region.

(*) Rejection Region (Critical Region) → " " " H_0 is Rejected " " R Region

(*) Level of Significance (α) → The probability of statistic falls in the Rejection Region is called α. Generally it is represented in terms of % and it is predecided in Quest.

L.of S = 10%

L of Sig = 5%



level of Sig = 1%

Z-TEST (Large sample test is $n \geq 30$)

with the help of z test we can solve following types of Questions;

Type I Testing the significance of Pop. Mean

$$Z = \frac{\bar{x} - \mu_0}{SE(\bar{x})} = \boxed{\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}}$$

Here $H_0 : \mu = \mu_0$ & $H_1 : \mu \neq \mu_0$

Type II Testing the significance of Difference b/w two Means — $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$

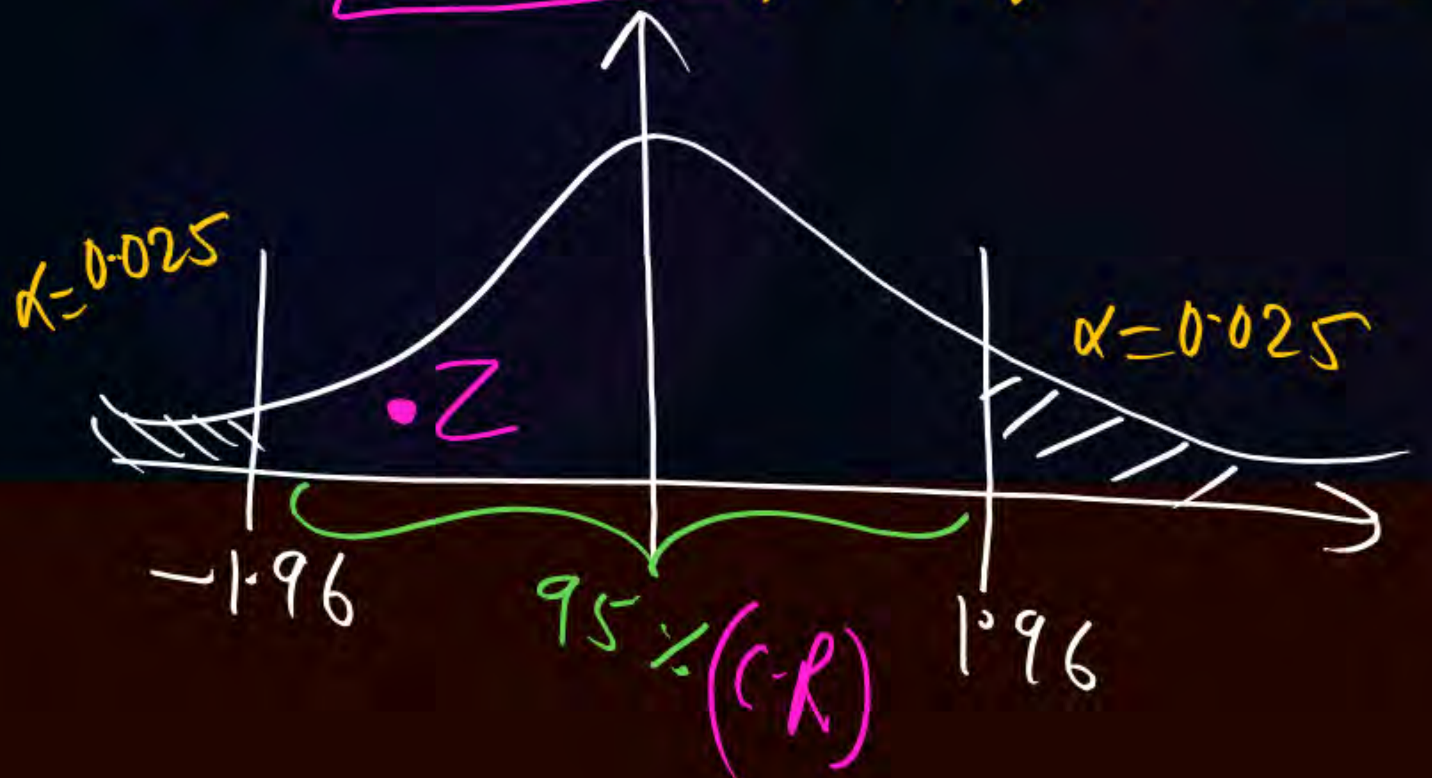
$$Z = \frac{\bar{x} - \bar{y}}{SE(\bar{x} - \bar{y})} = \boxed{\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}$$

#Q. A sample of 400 members has a mean = 4 where sample in taken from normal population with unknown mean and standard deviation 2.6 can we say that population mean is 4.2 with 5% level of significance. It is given that for two tailed test, $Z_\alpha = 1.96$ for $\alpha = 0.025$.

Sol: $n = 400$, $\bar{x} = 4$

$\mu_0 = ?$, $\sigma = 2.6$

$H_0 : \boxed{\mu = 4.2}$, $H_1 : \mu \neq 4.2$



Now $Z = ? = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{4 - 4.2}{2.6/\sqrt{400}} = \dfrac{-0.2 \times 20}{2.6}$

$Z = -1.538$

∵ Z lies in the acceptance Region So H_0 is accepted

i.e. Pop. Mean can be taken as $\mu_0 = 4.2$ Ans

Note: ∵ $|Z| < Z_\alpha (5\%)$ So H_0 is Accepted

i.e $(1.538 < 1.96)$

#Q. While calculating the average monthly income of a family in a town a sample of 81 families was taken. The mean income and standard deviations of these 81 families were found to be 4108 Rs and ~~24 Rs~~ respectively shown that the assumption "average income of family in a town in 4100 Rs" is not reasonable for 1% level of significant if $Z_\alpha = 2.58$.

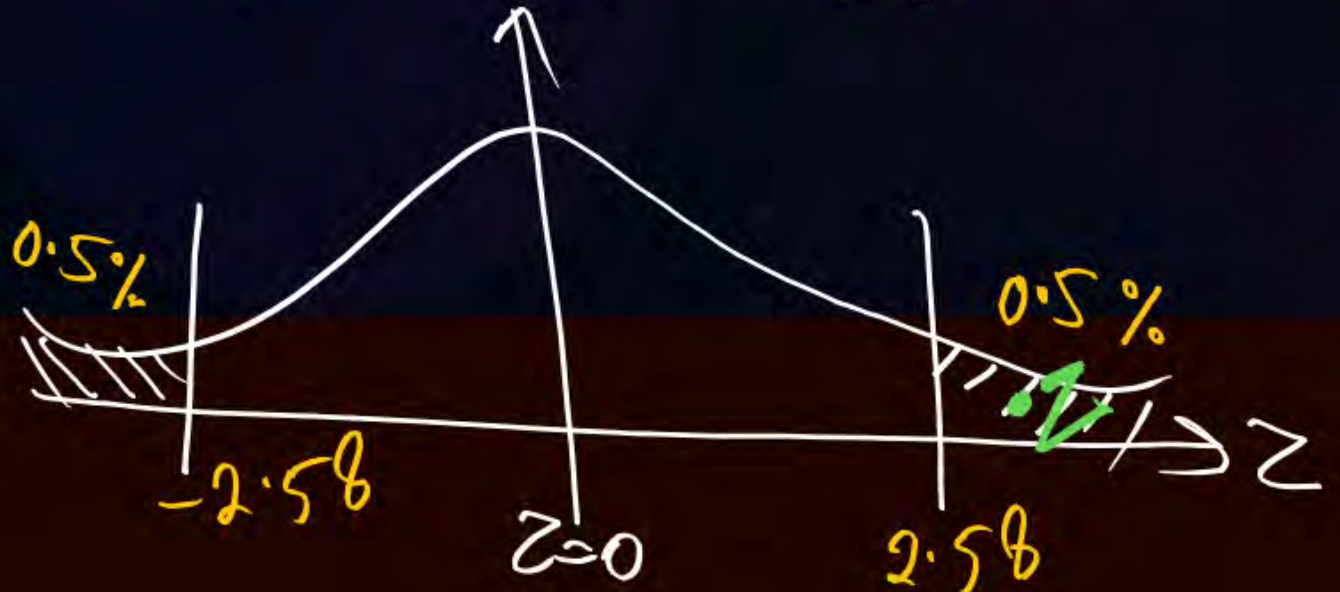
SD of pop in 24 Rs.

H_0

(ii) Also find the most probable limits for average income.

$n = 81, \ \bar{x} = 4108, \ \sigma = 24$

$H_0: \mu_0 = 4100, \ H_1: \mu_0 \neq 4100$



$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{4108 - 4100}{24/\sqrt{81}} = 3$$

$\because |Z| > Z_\alpha (1\%)$ i.e. Z lies in Rejection Region

So H_0 is Rejected & H_1 is accepted.

i.e. Av monthly income of family $\neq 4100$

(ii) $\mu_0 = ?$, $\bar{x} = 4108$, $n = 81$, $\sigma = 24$

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{24}{\sqrt{81}} = \frac{24}{9} = \frac{8}{3} \Longrightarrow 3 \cdot SE(\bar{x}) = 8$$

$$\text{Sample Mean} - 3SE(\bar{x}) \le \text{Pop Mean} \le \text{Sample Mean} + 3SE(\bar{x})$$

$$4108 - 8 \le \mu_0 \le 4108 + 8$$

$$\boxed{4100 \le \mu_0 \le 4116}$$

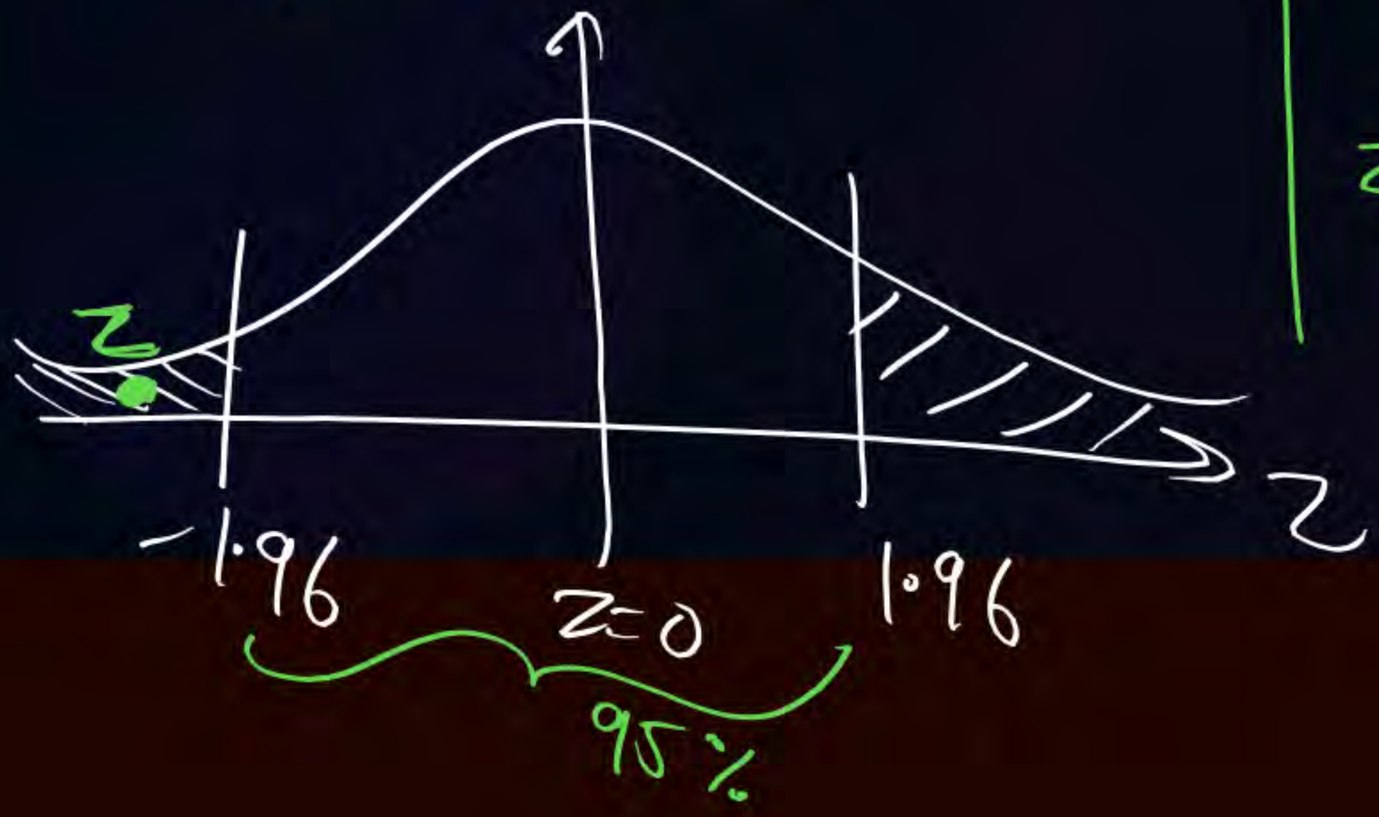
#Q. The mean of two samples of 1000 and 2000 members are 67.5 and 68.0 inches respectively can the sample be regarded as drawn from the same population of standard deviation 2.5 inches take α = ~~0.5~~ 0·05

i.e. $Z_\alpha\ (0.05) = 1.96$

Sol: $n_1 = 1000$, $n_2 = 2000$

$\bar{x} = 67.5$, $\bar{y} = 68.0$

$\sigma_1 = \sigma_2 = 2.5$



$H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{67.5 - 68.0}{\sqrt{\frac{6.25}{1000} + \frac{6.25}{2000}}} = -5.16$$

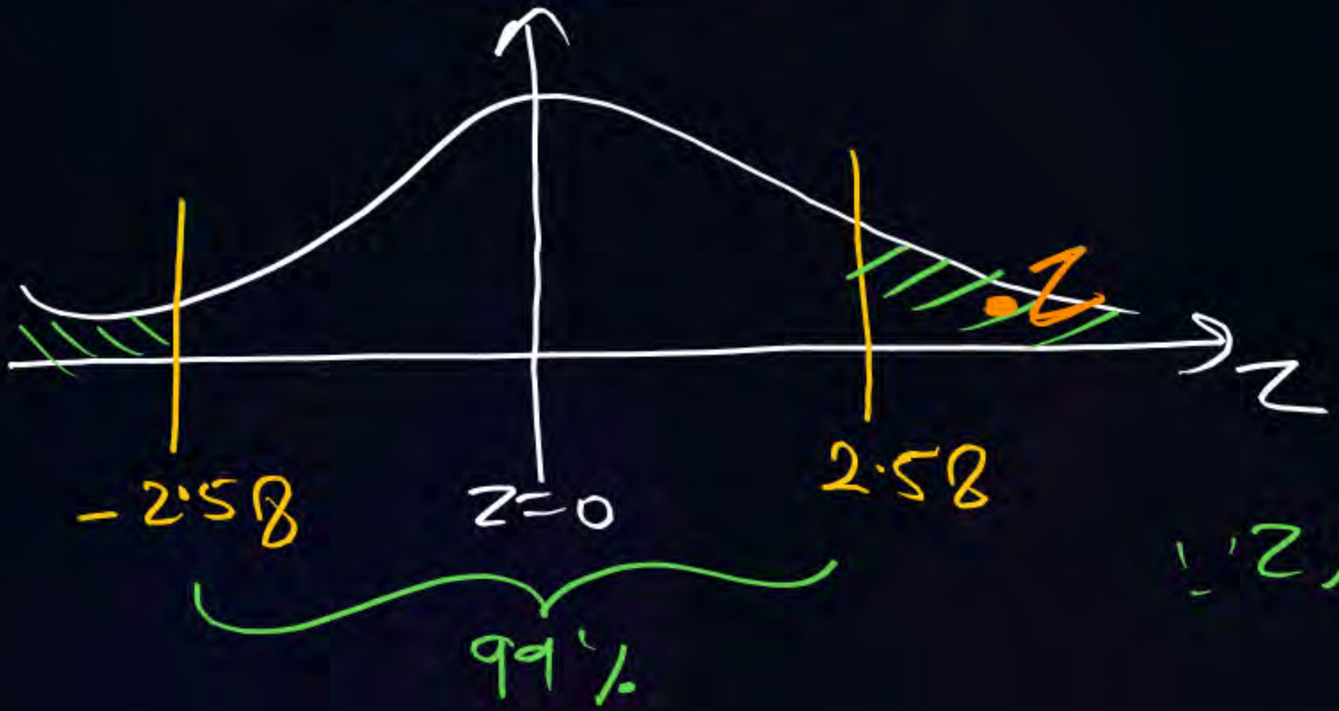
z lies in Rejection Region so H_0 is Rejected & H_1 is accepted i.e Samples are not drawn from same population.

#Q. If means of two samples (of same size 400) are 250 and 220 with their standard deviation 40 and 55 respectively then test $\boxed{H_0: \mu_1 = \mu_2}$ against $H_1: \mu_1 \neq \mu_2$ at 1% level of significance

i.e. $Z_\alpha\ (0.01) = 2.58$

$n_1 = n_2 = 400$, $\bar{x} = 250$, $\sigma_1 = 40$

$\bar{y} = 220$, $\sigma_2 = 55$



$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{250 - 220}{\sqrt{\frac{(40)^2}{400} + \frac{(55)^2}{400}}} = 8.82$$

∵ Z lies in the Rejection Region so H_0 is Rejected & H_1 is accepted.

THANK - YOU