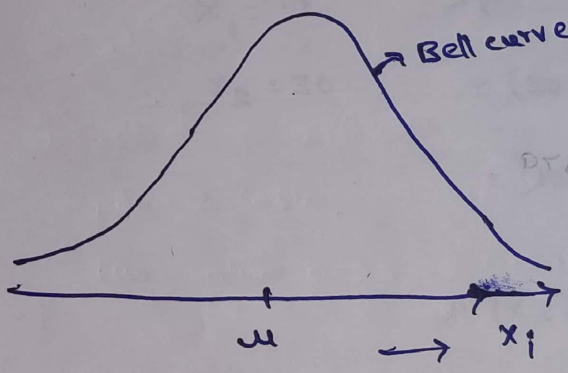
Central limit theorem:- The central limit theorem relies on the concept of a sampling distribution, which is the probability distribution of a statistic for a large number of samples taken from a population.

→ The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough. Regardless of whether the population has a normal, poisson, binomial or any other distribution, the sampling distribution of the mean will be normal.

Let's say $x \sim N(\mu, \sigma)$

n = sample size



let's take few sample from population with size n

Here, n = 1 or 2 or 3 or 100 ...

$S_1 = \{x_1, x_2, x_3 \cdots x_n\} = \bar{x}_1$

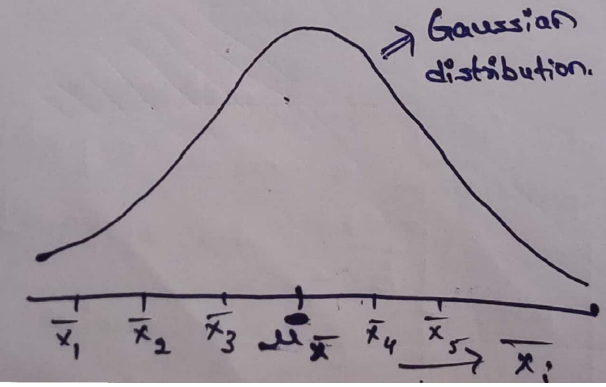
$S_2 = \{x_1, x_2, x_3 \cdots x_n\} = \bar{x}_2$

$S_3 = \{x_1, x_2, x_3, x_4 \cdots x_n\} = \bar{x}_3$

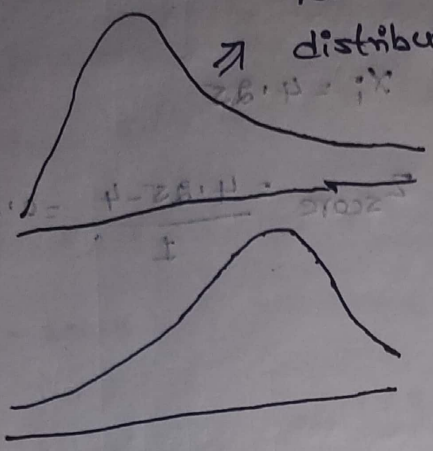
$\vdots$

$S_m = \{x_1, x_2, x_3, x_4 \cdots x_n\} = \bar{x}_m$

If we plot a graph with all the mean values of samples they also follow a Gaussian distribution.



② $X \approx N(\mu, \sigma)$

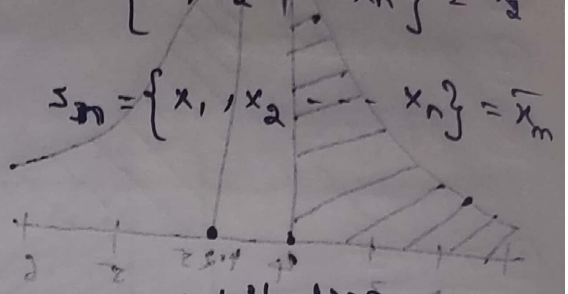


Non-Gaussian distribution.

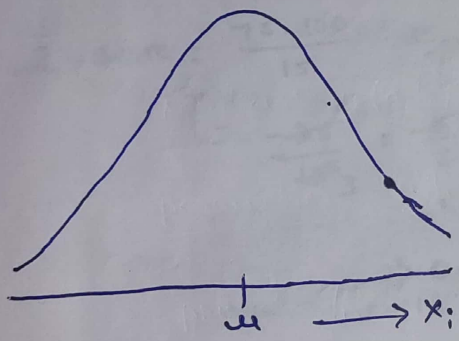
$S_1 = \{x_1, x_2 \dots x_n\} = \bar{x}_1$

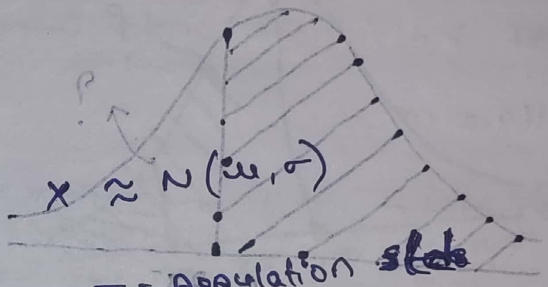
$S_2 = \{x_1, x_2 \dots x_n\} = \bar{x}_2$

$S_m = \{x_1, x_2 \dots x_n\} = \bar{x}_m$



→ If x is not following Gaussian distribution we need to consider sample size ≥ 30, in order to make sure that means of samples to follow normal distribution.



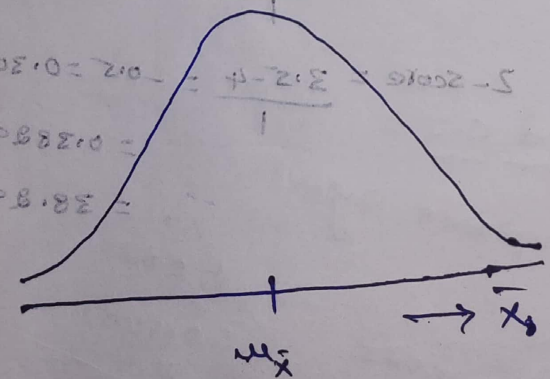


$X \approx N(\mu, \sigma)$

σ = population ~~std~~

μ = population mean (approx)

Sampling distribution of the mean

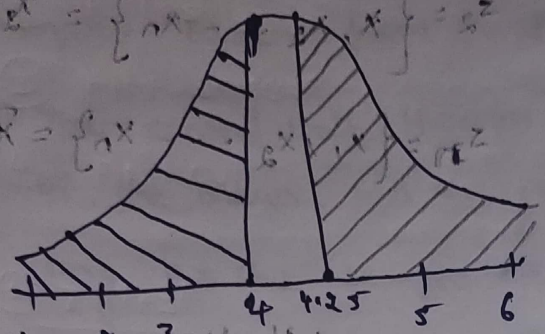


$\bar{X} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

$\frac{\sigma}{\sqrt{n}}$ → standard error.

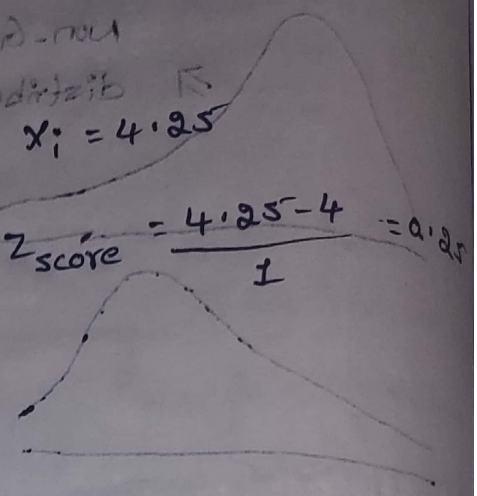
n = sample size

① $X \approx N(4,1)$

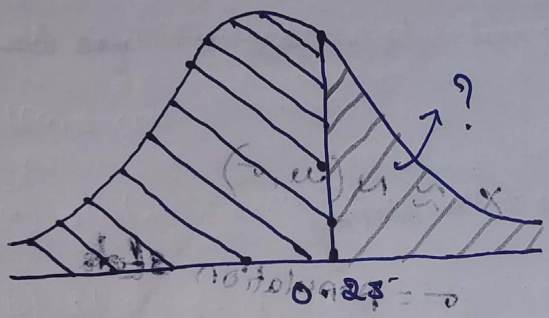


$x_i = 4.25$

$z_{score} = \frac{4.25-4}{1} = 0.25$

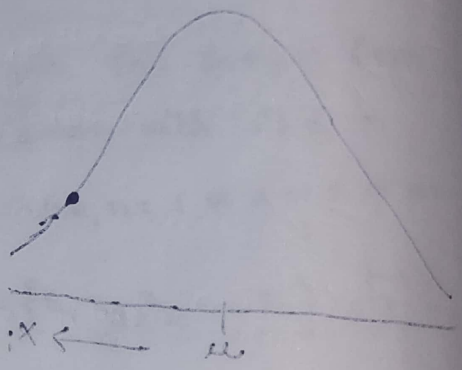


→ What percentage of scores falls above 4.25 ?



From z-table,

the value $0.25 \rightarrow 0.59871$

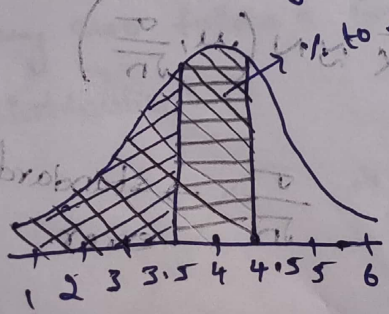


% of scores falls above 4.25 is $\Rightarrow 1 - 0.59871$

$= 0.4013$

$= 40.13\%$

→ what percentage of scores lies between 3.5 to 4.5 ?

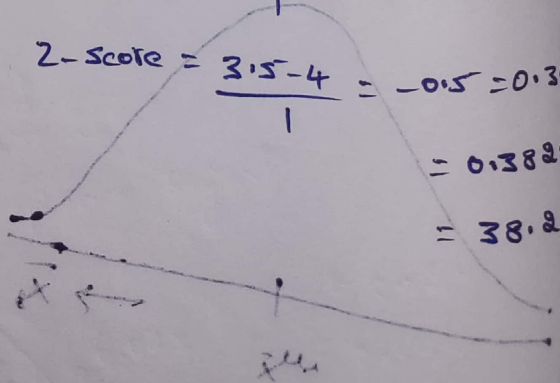


$z\text{-score} = \frac{4.5-4}{1} = +0.5 = 0.69146$

$z\text{-score} = \frac{3.5-4}{1} = -0.5 = 0.30853$

$= 0.3829$

$= 38.29\%$

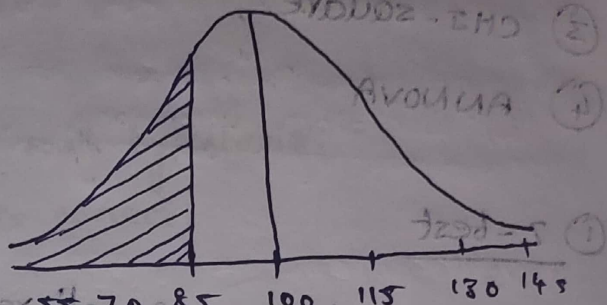


* In India the average IQ is 100, with a standard deviation of 15. What is the percentage of the population would you expect to have an IQ lower than 85?

→ $\sigma = 15$, $\mu = 100$

$$Z\text{-score} = \frac{85-100}{15} = -1$$

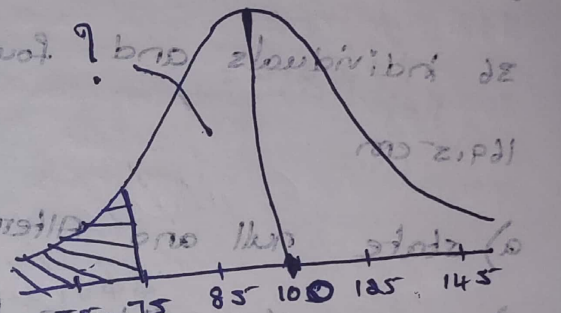
$\Rightarrow 0.15866$



$IQ \geq 85 \Rightarrow 1 - 0.1586$

$= 84.13\%$

ii) $75 \geq IQ \leq 100$?



$$Z\text{-score} = \frac{75-100}{15}$$

$$= \frac{-25}{15_3} = -\frac{5}{3}$$

$= -1.633$

$\Rightarrow 0.516$

Area $= 1 - 0.516$

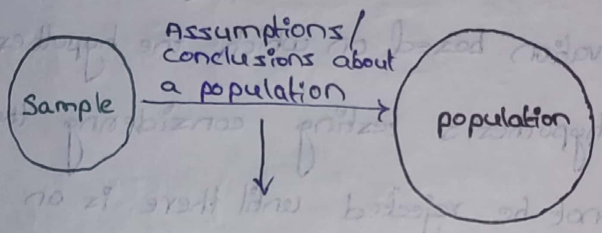
$$Z\text{-score} = \frac{100-100}{15} = 0$$

Inferential Statistics

Statistical Inference:

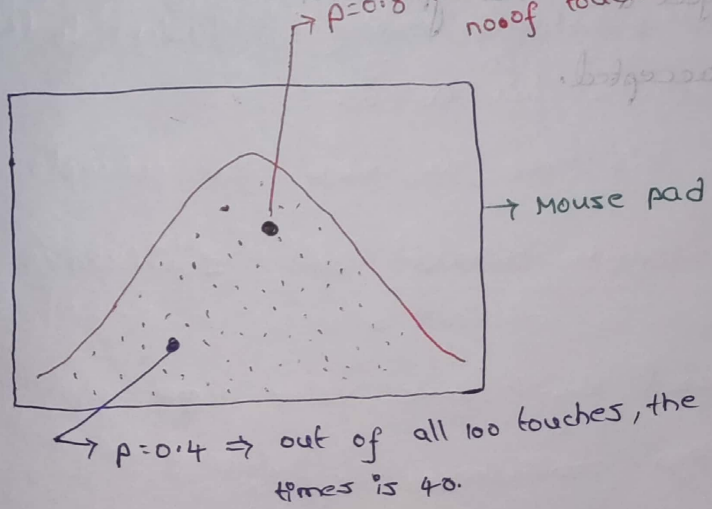
→ In statistical inference various sample statistics are analyzed to determine the population parametres.

Term	Denotes	Example
→ Statistic	a quantity/property of a sample	sample mean, sample variance s^2
parameter	a quantity/property of a population.	population mean, population variance



To validate these assumptions, we use hypothesis testing.

P-value: It is the probability for null hypothesis to be true



→ This p-value can determine probability.

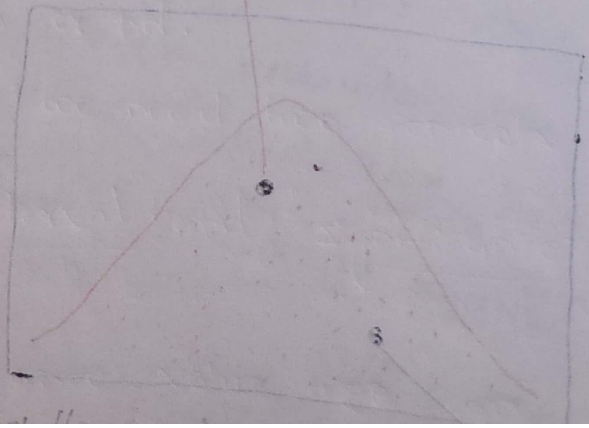
Hypothesis Testing:-

In statistics, hypothesis is a statement about the population which is represented by some numerical values. Hypothesis testing deals with collecting enough evidence about the hypothesis. Then, based on evidence collected, the test either accepts or reject null hypothesis.

→ each hypothesis test usually includes two competing hypothesis about the population. They are null hypothesis and alternate hypothesis.

Null Hypothesis (H_0): It is a statement/hypothesis that disapproves/rejects the observation based on which the hypothesis is made. you can start the hypothesis testing considering the null hypothesis to be true. It cannot be rejected until there is an evidence which suggests otherwise.

Alternate Hypothesis (H_1):- It is a statement/hypothesis which is contradictory to the null hypothesis. If you find enough evidence to reject the null hypothesis, then the alternate hypothesis is accepted.



Steps involved in performing Hypothesis Testing:-

Step 1:-

Figure out null hypothesis and Alternative hypothesis.

Null hypothesis (H_0)

Alternate hypothesis (H_1)

Significance level ∝

confidence interval

step 2:- Decision boundary.

calculate the probability of getting the observed data (p value) assuming null hypothesis to be true.

→ calculate the test statistic.

1) z-test 2) t-test 3) chi-square 4) ANOVA (F-test)

→ Find out true problem is one tail or two tail.

→ Find out confidence interval values

→ calculate true p-value.

step 3:- conclusions.

(I) Fail to reject null hypothesis (when p value >= significance level)

(II) Reject null hypothesis (when p value < significance level)

Q) Test whether the coin is fair coin or not by performing 100 tosses.

① Null hypothesis - coin is fair → (H_0)

② Alternate hypothesis - coin is not fair → (H_1)

③ Experiment

④ conclusions

→ In the experiment performed we got following outcomes.

100 tosses → 50 times head, 50 times tails } coin is fair.

60 times head, 40 times Tail } coin is fair

30 times head, 70 times tail } ?

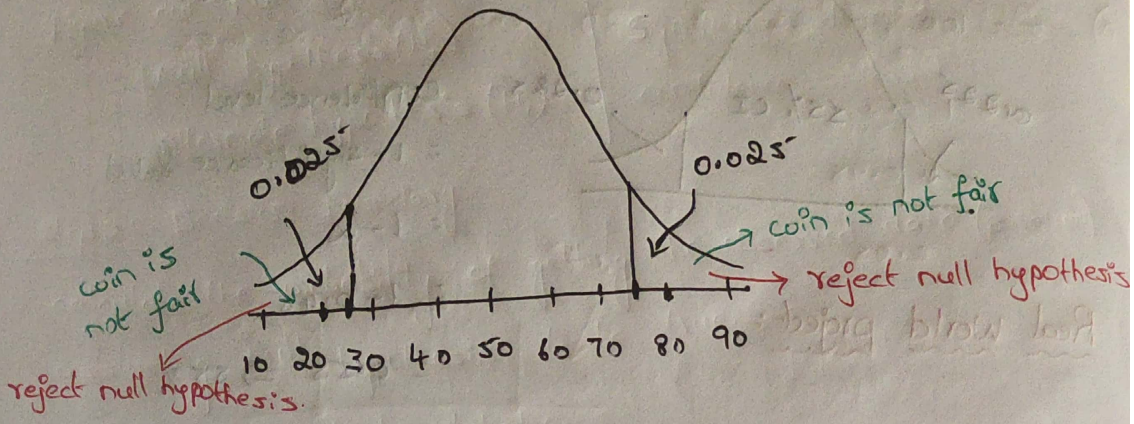
In this case we may get confusion to tell whether the coin is fair or not.

→ In this case we need two concepts one is "confidence interval" and "significance value".

→ Based significance value we can able to define confidence interval.

→ The confidence interval significance value is given by domain expertise.

when I toss a coin 100 times, I may have chance to get heads 10 times head or 90 times head

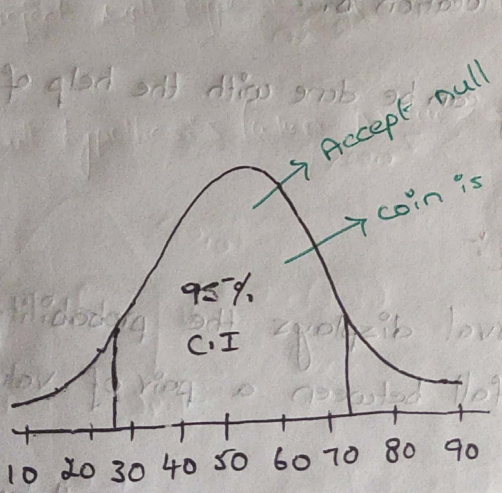


Conclusion:-

Here 10, or 90 does not lies in range 20-80, so that null hypothesis is rejected.

⇒ coin is not fair.

If i perform a experiment I may have a chance to get 60 times head.



Conclusion:

Here 60 lies in range 20-80

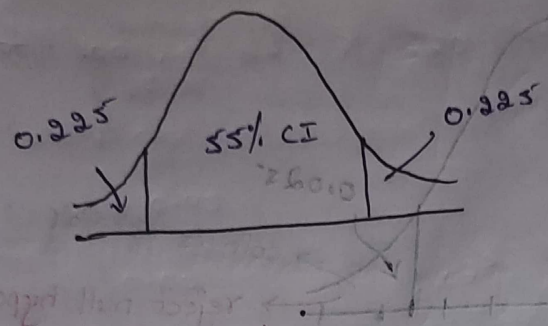
→ In this case we will accept null hypothesis

→ coin is fair.

For medical problem, $\alpha = 0.45$

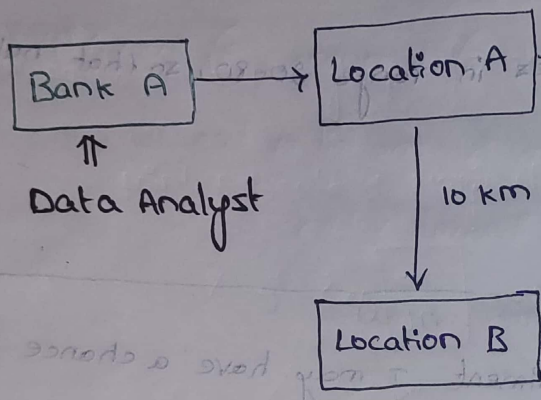
if $\alpha = 0.45$, C.I = ?

$\frac{0.45}{2} = 0.225$



Confidence level
= 1 - significance level
= $1 - \alpha$
= $1 - 0.45$
= $0.55 \Rightarrow 55\%$

Real World project:

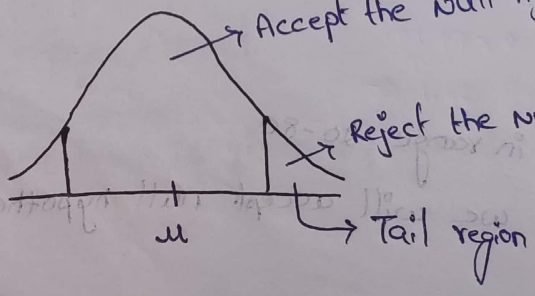


→ Now the task is based on the data that we have at Location 'A'. we need to find whether should we open a atm or not at location B.

→ This entire thing can be done with the help of hypothesis testing.

Confidence Interval :-

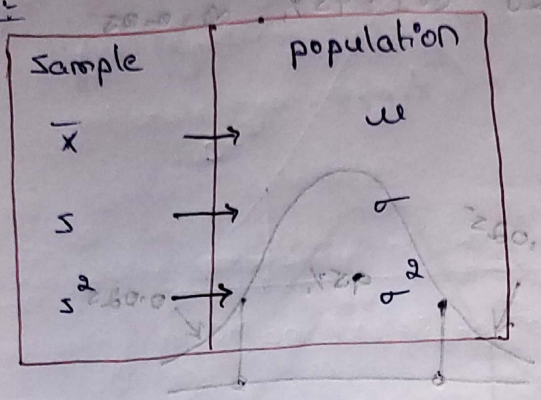
A confidence interval displays the probability that a parameter will fall between a pair of values around mean.



point estimate:

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

Eg:



statistic parameter.

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

2.8, 2.9, 2.5, 3.5 3

→ Sample (mean, variance, std deviation) are used to estimate population (mean, variance, std deviation).

Confidence Interval

point estimate ± Margin of Error.

→ when population standard deviation is given:

$$C.I = \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$M.O.E = Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

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

σ = population std deviation, n = no. of samples

C.I range for mean = Lower C.I to Higher fence C.I

PE − MOE PE + MOE

C.I range for mean : $\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ to $\bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

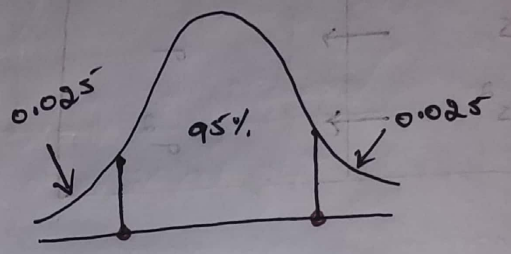
Q) On a quant test of CAT exam, the standard deviation is known to be 100. A sample of 25 test takers has a mean of 520. Construct a 95% C.I about the mean?

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

$\alpha = 1 - C.I$

$= 1 - 0.95$

$= 0.05$



when population std is given {Z-score} → Z table

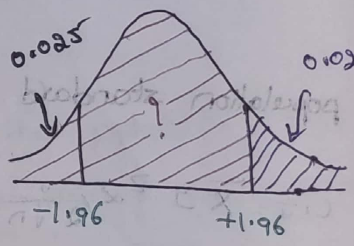
Point Estimate ± Margin of Error.

$\bar{X} \pm Z_{\alpha/2} \left[\frac{\sigma}{\sqrt{n}}\right]$ → standard Error.

Lower fence C.I $= \bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Higher fence C.I $= \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

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



$= 1 - 0.025$

$= 0.9750$

⇒ 1.96 through Z-table

B

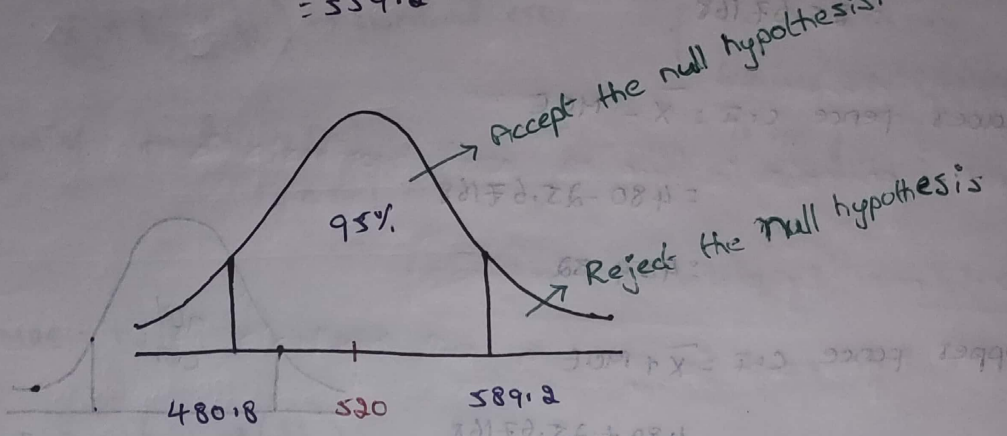
lower fence $= 520 - 1.96 * \frac{100}{\sqrt{25}}$

$= 520 - 1.96 \times 20$

$= 480.8$

Higher fence $= 520 + 1.96 \times 20$

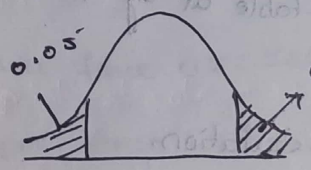
$= 559.2$



b) $\sigma = 80, n = 25, CI = 0.90, \bar{x} = 480$. construct a 90% C.I about mean. $\alpha = 1 - CI = 1 - 0.9 = 0.1$

$$MOE = Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} = Z_{\frac{0.1}{2}} \times \frac{80}{\sqrt{25}}$$

$$= Z_{0.05} \times 16$$



$= 1 - 0.05$

$= 0.95$

0.95 lies between 0.9450 and 0.95053

ie $\Rightarrow$ 1.604 and 1.605

x	0.9450	0.95	0.95053
y	1.604	$Z_{0.05}$	1.605

$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1)$ $\rightarrow$ interpolation

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

$$Z_{0.05} = 1.60448$$

$$MOE = 1.60448 \times 16$$

$$= 25.67168$$

Lower fence C.I $= \bar{x} - MOE$

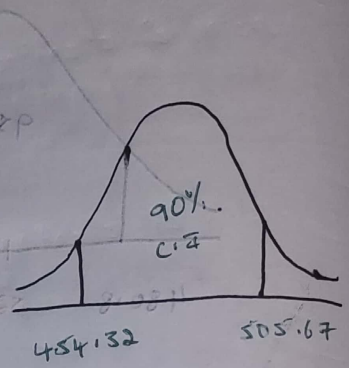
$= 480 - 25.67168$

$= 454.32$

Upper fence C.I $= \bar{x} + MOE$

$= 480 + 25.67168$

$= 505.67$

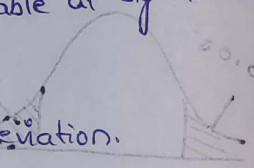


t-test

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

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

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

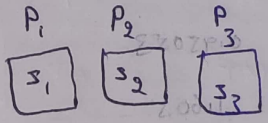


s = sample standard deviation.

n = no. of samples

Degree of freedom (d.o.f) $= n - 1$

suppose 3 seats are available



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

d.o.f $= n - 1 = 3 - 1 = 2$ ⇒ only two people have choice to choose seat

→ On the Quant test of cat exam, a sample of 25 test takers has a mean of 520 with a sample standard deviation of 80. construct 95% C.I about the mean?

Ans:) $\bar{x} = 520$, $s = 80$, $\alpha = 0.05$, $n = 25$, C.I $= 0.95$

when sample standard deviation is given we have to use t-test ⇒ t-table

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \rightarrow \text{standard error.}$$

{ t-test we are using because population std is not given }

Degree of freedom $= n-1$

$= 25-1 \Rightarrow 24$

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

$$= t_{0.025} \times \left(\frac{80}{\sqrt{25}}\right)$$

From t-table $t_{0.025}$ and dof $= 24$

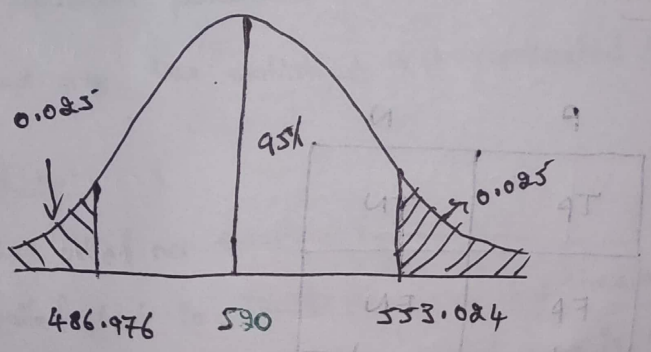
$t_{0.025} = 2.064$

$MOE = 2.064 \times 16$

$= 33.024$

lower fence C.I $= 520 - 33.024 = 486.976$

upper fence C.I $= 520 + 33.024 = 553.024$



Type-I and Type-2 Error :-

Null hypothesis (H_0) = coin is fair

Alternate hypothesis (H_1) = coin is not fair.

Reality check :-

Null hypothesis is true or Null hypothesis is false.

Decision :- {By performing Experiments}

Null hypothesis is true or Null hypothesis is false.

outcome : 1

we reject the null hypothesis, when in reality it is false → yes (Good decision)

outcome : 2

we reject the null hypothesis when in reality is ~~false~~ true → Type-I Error

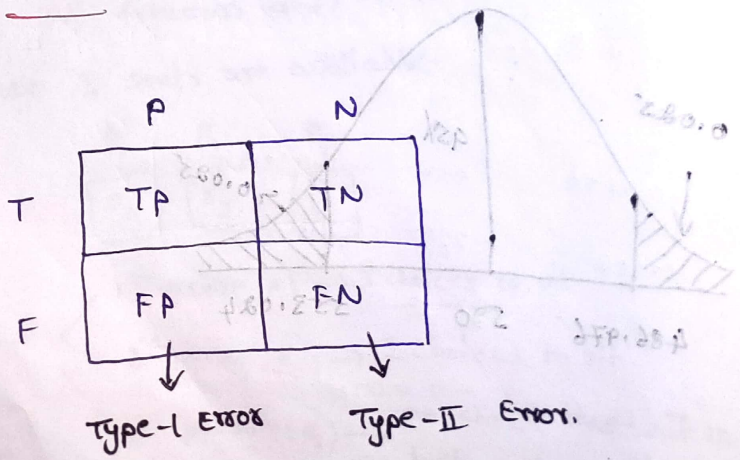
outcome : 3

we accept the null hypothesis when in reality it is false → Type-II Error

outcome : 4

we accept the Null hypothesis when in reality it is true → yes (Good decision)

Confusion Matrix :-

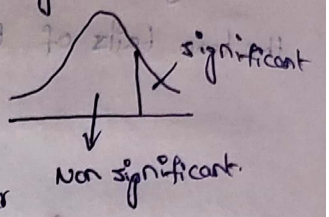


One-tailed test

→ It is based on uni-directional hypothesis.

→ one tailed test is a hypothesis test in which the region of rejection appears on one end of sampling distribution.

→ Critical values are boundary between non significant and significant results in a hypothesis test.

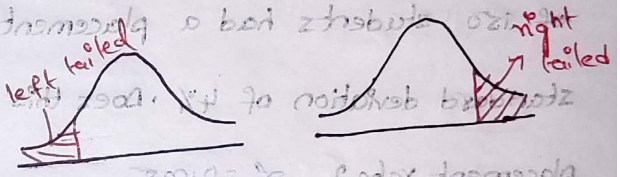


→ one-tailed test represents estimated parameter is greater or less than the critical value.

→ when the sample tested falls in the region of rejection either left or right side, at this time it leads to acceptance of alternate hypothesis rather than null hypothesis

→ In this statistical hypothesis test, all the critical region related to α, is placed any one of the two tails.

one-tailed test can be :



Left-tailed test :-

→ when the population parameter is believed to be lower than the assumed one, the hypothesis test carried out is the left tailed test.

Right-tailed test :-

→ when the population parameter is supposed to be greater than the assumed one, the statistical test conducted is a right tailed test.

Two tailed test :-

→ It is also called non-directional hypothesis.

→ The two tailed test is described as a hypothesis test, in which the region of rejection or say the critical area is on both ends of the normal distribution.

→ It determines whether the sample tested falls within or outside of a certain range of values.

→ Therefore an alternate hypothesis is accepted in place of the null hypothesis, if the calculated values falls in either of the two tails of the probability distribution.

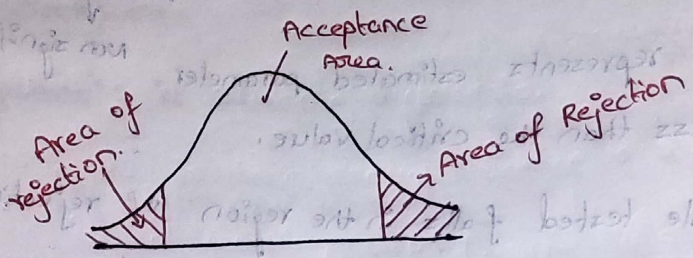


Fig: Two-tailed test

(2) Colleges in karnataka has an 85% placement rate. A new college was recently opened and it was found that a sample of 150 students had a placement rate of 88%. with a standard deviation of 4%. Does this college has different placement rate? $\alpha = 0.005$

→ Her the placement rate is may be less than or greater than 85%.

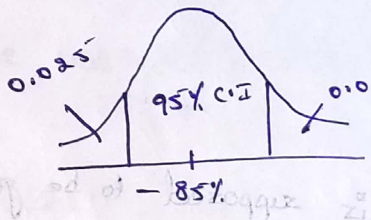
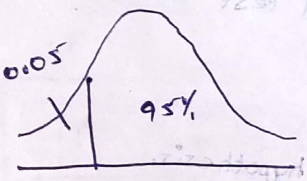


Fig:- Two tailed test

→ If placement is less than 85%.



Difference between one tailed and two tailed test.

one-tailed test	Two-tailed test.
→ used to test a directional hypothesis.	→ used to test a non-directional hypothesis.
→ The null hypothesis is rejected if the test static falls entirely on one side of sampling distribution.	→ The null hypothesis is rejected if the test statistic falls in either tail of the sampling distribution.
→ The critical value or p-value is only calculated for one tail of the distribution.	→ The critical value or p-value is calculated for both tails of the distribution.
→ we use either "<" or ">" sign for (H_1).	→ we use $\neq$ 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.

When to use z-test?

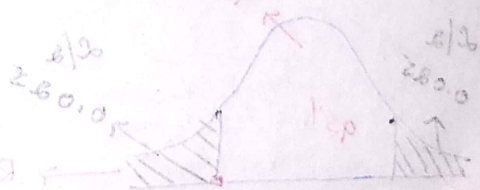
① when sample size ($n \geq 30$)

② population standard deviation or sample standard deviation either of them should given.

when to use t-test?

① when sample size $n < 30$

② when sample standard deviation is given.



when $n = 1 \Rightarrow z = \frac{\bar{x} - \mu}{\sigma}$

$n = 30 \Rightarrow z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$