⇒ Infrential Statistics:

⇒ A Hypothesis and Hypothesis Testing Mechanism:

- Hypothesis Testing Mechanism

① Null Hypothesis (H_0):
The assumption you are beginning with.
e.g: When person enter the court, considered Person is not guilty

② Alternate Hypothesis (H_1):
Opposite of null hypothesis
e.g Person is guilty (opposite of null hypothesis)

③ Experiments:
Statistical Analysis
e.g (Collect finger prints and evidences)

④ Accept/reject the null hypothesis
e.g (on the basis of experiments inferencing result)

Case Study:

Colleges at District A states it's average passed percentage of students are 85%. A new college opened in the district and it was found that a sample of student 100 have a pass percentage of 90% with a standard deviation of 4%.

Does this school have a different passed percentage?

Ans)

⇒ Hypothesis Testing

1. Null Hypothesis (h_0):

μ(Average) = 85%

2. Alternate Hypothesis (h_1):

μ(Average) $\neq$ 85%

We draw it from question

If question is does the college have greater passed percentage then:

$\mu > 85\%$ (Alternate Hypothesis)

⇒ Pvalue:

The Pvalue is a number calculated from a statistical test that describes how likely you are to have found a particular set of observations if the null hypothesis were true.

⇒ P values are used in hypothesis testing to help decide whether to reject the null hypothesis.

Example:

Coin is fair or Not

$P(H)=0.5$ $P(T)=0.5$

We find it through hypothesis testing:

Hypothesis Testing:

Null hypothesis (h_0):

① Null hypothesis

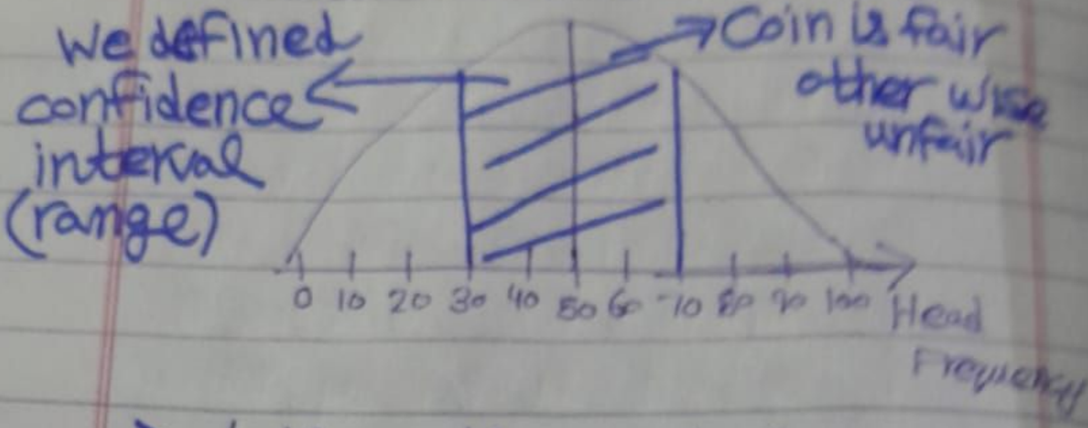
Coin is fair

② **Alternate hypothesis(h_1):**

Coin is not fair

③ Experiment:

Toss the coin 100 times.



⇒ Let's suppose in 100 tosses 50 heads come means coin is fair 60 come we can say it is fair but 70 come we think that coin may be unfair so we define

Confidence Interval

④ Significance Value (α):

Used to define confidence interval i.e $\alpha = 0.05$

Confidence Interval $(C.I) = 1 - \alpha$

$= 1 - 0.05$

$C.I = 0.95$

Now let's calculate P value by some test

Conclusion:
if P-value < significance value
 Reject the null hypothesis
else
 Fail to reject the null hypothesis

⇒ Hypothesis Testing and Statistical Analysis:

1. Z Test } ⇒ Average
2. t Test }
3. CHI SQUARE ⇒ Categorical Data
4. ANNOVA ⇒ Variance

⇒ Z test

To use Z test:
1. You should know Population Standard Deviation
2. $n >= 30$ where n is sample size

⇛ Problem Statement:

The average heights of all residents in a city is 168cm with a =3.9. The doctor believes the mean to be different. He measured the height of 36 individuals and found the average height to be 169.5cm

a) State null and alternate hypothesis

b) At a 95% confidence level is there enough evidence to reject the null hypothesis.

Ans):

Given Data:

μ (Population Mean) = 168cm

σ (Population Standard Deviation) = 3.9

n (Sample Size) = 36

$\bar{X}$ (Sample Mean) = 169.5

Confidence Interval (CI) = 95%

CI = 0.95

using Confidence Interval (C.I)
calculate Significance Value (α)
α(Significance Value) = 1 − C.I
= 1 − 0.95

α(Significance Value) = 0.05

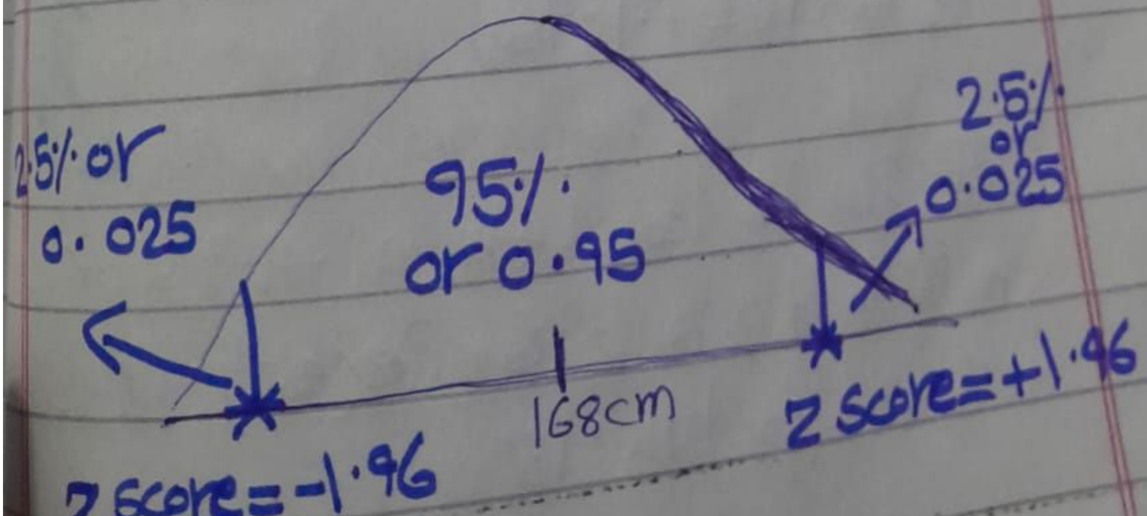
As we know population standard deviation and n>=30 so we will use **z test**

① **Null Hypothesis(h_0):**
$\mu = 168$ cm

② **Alternate Hypothesis (h_1):**
$\mu \neq 168$ cm (opposite to null hypothesis)

③ **Based on C.I we will draw Decision Boundary**

As C.I (Confidence Interval) = 95%.



⇒ It is two tail test as value can be less than or greater, than to reject null hypothesis

Now we will calculate the z-score for both marked points using area under the curve

So, for right most

Area under the curve = $\underbrace{1}_{\text{total area}} - \underbrace{0.025}_{\text{that region}}$

Area under the curve = 0.9750

To find Z-Score see it in Z-table

$\boxed{\text{Z-Score} = +1.96}$

and for left point as distribution is symmetric

$\boxed{\text{Z-Score} = -1.96}$

If **Z** is less than -1.96 or greater than 1.96, Reject the null hypothesis

4) Z-test:

To calculate Z we will use Z-test

$$Z_d\ (\text{sample}) = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

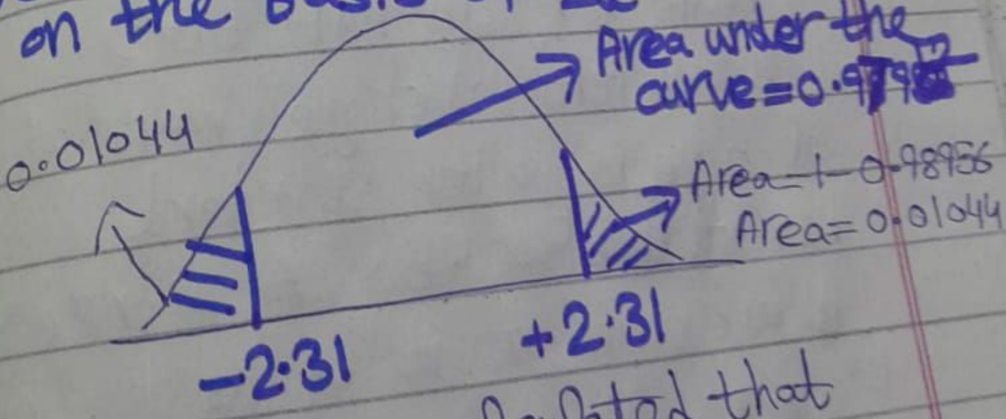
$$Z_d = \frac{169.5 - 168}{3.9/\sqrt{36}} = \frac{1.5}{0.65}$$

$$\boxed{Z_d = 2.31}$$

5) Conclusion:

As $Z_d > 1.96$ so we reject the null hypothesis

⇒ Constructing Decision Boundary on the basis of Z_d



* Again we calculated that Area under the curve using Z-table by +2.31 which is 0.98956

• Using P-value:

P-value = 0.01044 + 0.01044

P-value = 0.02088

As P-value (0.2088) < Significance value (0.05)

So, Null hypothesis is rejected

Final Conclusion:

Average ≠ 168 cm

The average height seems to be increasing based on sample heights.

⇒ Problem Statement: 02

A factory manufactures bulbs with average Warranty of 5 years with standard deviation of 0.50. A worker believes that it is less than 5 years. He tests a sample of 40 bulbs and find the average time to be 4.8 years

a) State null and alternate hypothesis

b)

) At a 2% Significance level is there enough evidence to support idea that warranty should be revised

Ans:

Data Given:

μ (population average) = 5 years

σ (Standard Deviation) = 0.50

$\bar{X}$ (Sample Mean) = 4.8 years

n (Sample Size) = 40 bulbs

α (Significance Level) = 2% = 0.02

Confidence Interval (C.I) = 1 − 0.02

C.I = 0.98 or 98%

① Null Hypothesis (h_0):

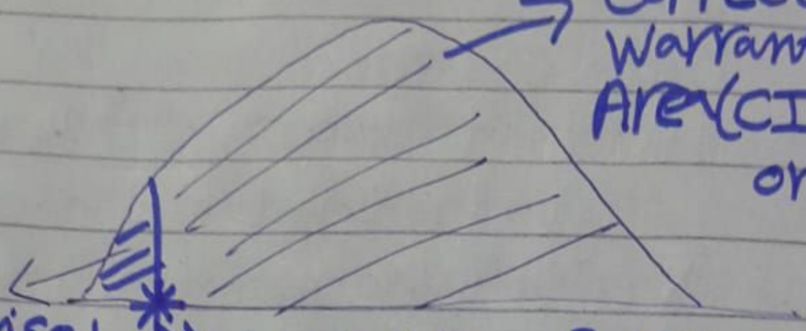
$\mu = 5$ years

② Alternate Hypothesis (h_1):

$\mu < 5$ years

③ Draw decision boundary based on C.I

It is a one tail test as testing for only less than



Correct Warranty Area (CI = 0.98 or 98%)

Revised Area (0.02 or 2%)

z-score : -2.05

We have to find z-score for this point

Area under the curve for point = 1 - 0.98

As point is on left = 0.02

so z-score will be negative. Now let's see z-score from z-table

z-score of point = $\boxed{-2.05}$

Now if $z < -2.05$ then Null Hypothesis reject

Z-test:

$$Z_d = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{4.8 - 5}{0.50/\sqrt{98}}$$

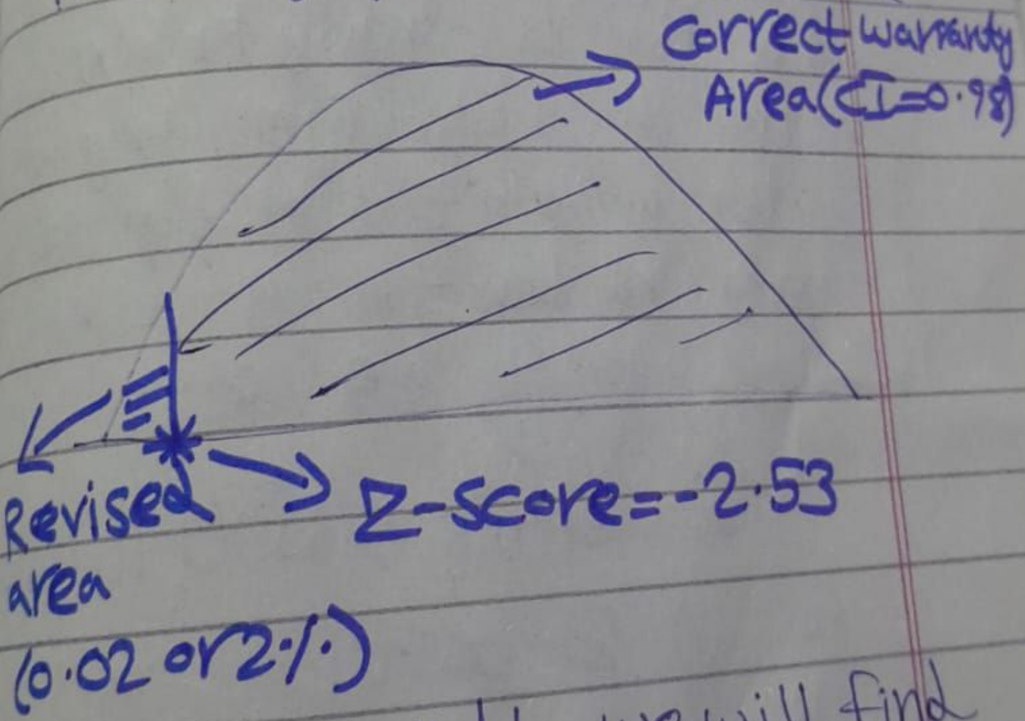
$$Z_d = -2.53$$

Conclusion:

As $Z_d < -2.05$ so null hypothesis rejected.

⇒ Using P-value

Redraw graph with new z-score



Using Z-table we will find the area under the curve which is 0.0057 So,

P-value = 0.0057

As

P-value (0.0057) < Significance value (0.02)

So,

· **Conclusion:**

We reject the null hypothesis

⇒ Student t Distribution:

In Z test when we perform any analysis using Z-score we require σ (population standard deviation) is already known.

* How do we perform any analysis when we don't know population standard deviation

⇓ we use

Student's t distribution

As in Z-test

$$Z = \frac{x-\mu}{\sigma/\sqrt{n}}$$

In Student t distribution

$$\boxed{t = \frac{x-\mu}{s/\sqrt{n}}}$$

sample standard Deviation

As we used Z table in Z test we use t-table in Student t Distribution.

Degree of Freedom (dof)

$\boxed{dof = n-1}$

For example, we are in a room having 3 chairs 3 people enter the room the 1st and 2nd person can choose chair but the 3rd has no option he has to sit on the only chair empty so Degree of Freedom (dof) in this case is:

$dof = 3-1$

$\boxed{dof = 2}$

⇛ T-stats and T-test:
↘ One-Sample T-test

Case Study:

In the population the average IQ is 100. A team of researcher want to test a new medication to see if it has either positive or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140 with a standard deviation of 20. Did the medication affect intelligence. and C.I = 95%.

• Ans:

Given data:

μ (Population Mean) = 100

n (Sample size) = 30

$\bar{X}$ (Sample Mean) = 140

s (Sample Standard deviation) = 20

C.I (Confidence Interval) = 95% = 0.95

Significance Value (α) = 0.05 = 5%

① **Null Hypothesis (h_0):**

$\mu = 100$

② **Alternate Hypothesis (h_1):**

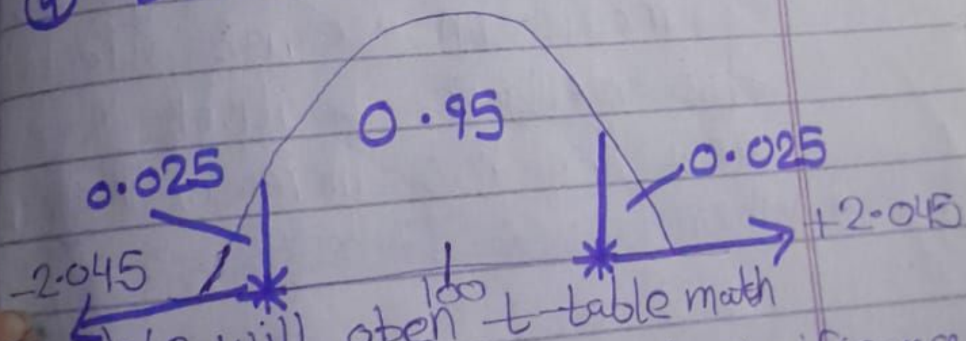
$\mu \neq 100$ (2 tail test can be greater than or less than)

③ **Degree of Freedom (dof):**

dof = n − 1 = 30 − 1

$\boxed{dof = 29}$

④ **Decision Rule:**



We will open t-table math from Y-axis 2 tail and significance value ($\alpha = 0.05$) and from x-axis dof = 29 and note value

⇒ If t-test is less than −2.045 or greater than 2.045, reject the null hypothesis

⑤ Calculate test statistics

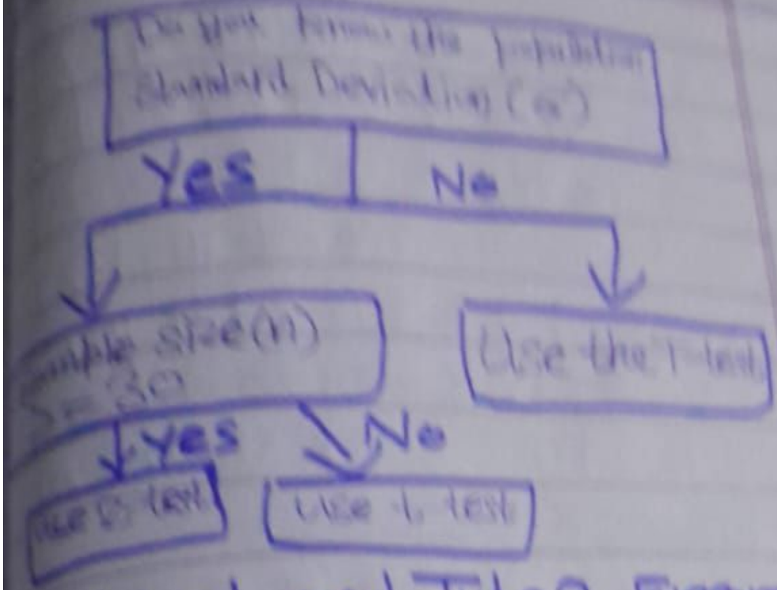
$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{140-100}{20/\sqrt{30}} = \frac{40}{3.65}$$

$$\boxed{t = 10.96}$$

⑥ Conclusion:

As $t(10.96) > 2.045$ so null hypothesis is rejected. So, medication use has affected the intelligence positively. Means medication has increased intelligence.

⇒ When to use T-test vs Z-test



⇒ Type1 and Type2 Errors

Reality: Null Hypothesis is True or Null Hypothesis is false

Decision: Null Hypothesis is True or Null Hypothesis is False.

Outcome1: We reject the Null hypothesis when in reality it is false → Good

Outcome2: We reject the null hypothesis when in reality it is True → Type1 Error

Outcome 3: We retain the null hypothesis when in reality it is False → Type 2 error

Outcome 4: We retain the null hypothesis when in reality it is true → Good

⇒ Bayes Statistics:

Bayes Statistics is an approach to data analysis and parameter estimation based on Bayes theorem

• Bayes Theorem:

Probability → Independent events
→ Dependent events

① Independent events:

e.g Rolling a dice

outcomes {1, 2, 3, 4, 5, 6}

$Pr(1) = \frac{1}{6}$, $Pr(2) = \frac{1}{6}$, ...

Means one event is not impacting other event

e.g Tossing a coin
$Pr(H) = 0.5$ $Pr(T) = 0.5$
Event are not affecting each other if head comes first even then the probability of getting tail is 0.5

② **Dependent Events:**
When one event is affecting other

e.g I have 5 marbles in a box out of which 2 are Red and 3 are Yellow Now if I pick a marble (red)

$$Pr(red) = \frac{2}{5}, \quad Pr(Yellow) = \frac{3}{5}$$

~~Now I want to pick another~~ Yellow marble after picking red so probability will change

$$Pr(red) = \frac{1}{4}, \quad Pr(Yellow) = \frac{3}{4}$$

This is dependent event

Conditional Probability

Now

$$\boxed{Pr(R \text{ and } Y) = Pr(R) \times \boxed{Pr\left(\frac{Y}{R}\right)}}$$

Means 1st Red then Yellow

Probability of Yellow if red has taken

Now we know

$Pr(A \text{ and } B) = Pr(B \text{ and } A)$

Let's expand

$$Pr(A) \times Pr\left(\frac{B}{A}\right) = Pr(B) \times Pr\left(\frac{A}{B}\right) \quad \text{—(i)}$$

$$\boxed{Pr\left(\frac{B}{A}\right) = \frac{Pr(B) \times Pr\left(\frac{A}{B}\right)}{Pr(A)}}$$

$\Rightarrow$ Bayes Theorem

OR

From equ(i)

$$\boxed{Pr\left(\frac{A}{B}\right) = \frac{Pr(A) \times Pr\left(\frac{B}{A}\right)}{Pr(B)}}$$

A, B = Events

$Pr\left(\frac{A}{B}\right)$ = Probability of A given B is true

$Pr\left(\frac{B}{A}\right)$ = Probability of B given A is true

$Pr(A), Pr(B)$ = Independent probabilities of A and B

Example Dataset:

Let's suppose we have 4 features

Size of house	No. of rooms	Location	Price
x_1	x_2	x_3	y

$\underbrace{x_1 \quad x_2 \quad x_3}$ → **Independent Features**

y → **Output/ Dependent Feature**

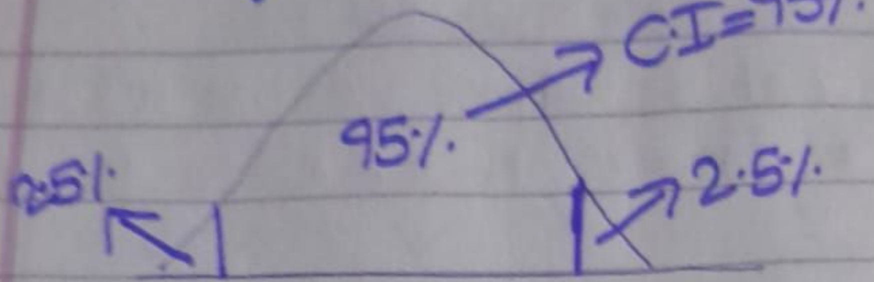
Now use Bayes Theorem Equation

$$\boxed{Pr\left(\frac{y}{x_1, x_2, x_3}\right) = \frac{Pr(y) \times Pr(x_1, x_2, x_3/y)}{Pr(x_1, x_2, x_3)}}$$

⇓

Baye's Theorem used in Naive Baye's Machine Learning Algorithm

⇒ Confidence Interval and Margin of Error:



Point Estimate:

$\bar{x}$ (Sample Mean) is point estimate to μ (population Mean)
Let $\bar{x} = 2.5$, $\mu = 3$ so it will many times not exactly same so we will define a range that my point estimate lies within this range (confidence Interval)

Confidence Interval,

$C.I = \text{Point Estimate} \pm \text{Margin of Error}$

Now Margin of Error (In Z-test) $= Z\left(\frac{\alpha}{2}\right) \frac{\sigma}{\sqrt{n}}$

$Z\left(\frac{\alpha}{2}\right)$ → Z-score of $\frac{\alpha}{2}$

For t-test
$C.I = \bar{x} \pm t \frac{s}{\sqrt{n}}$

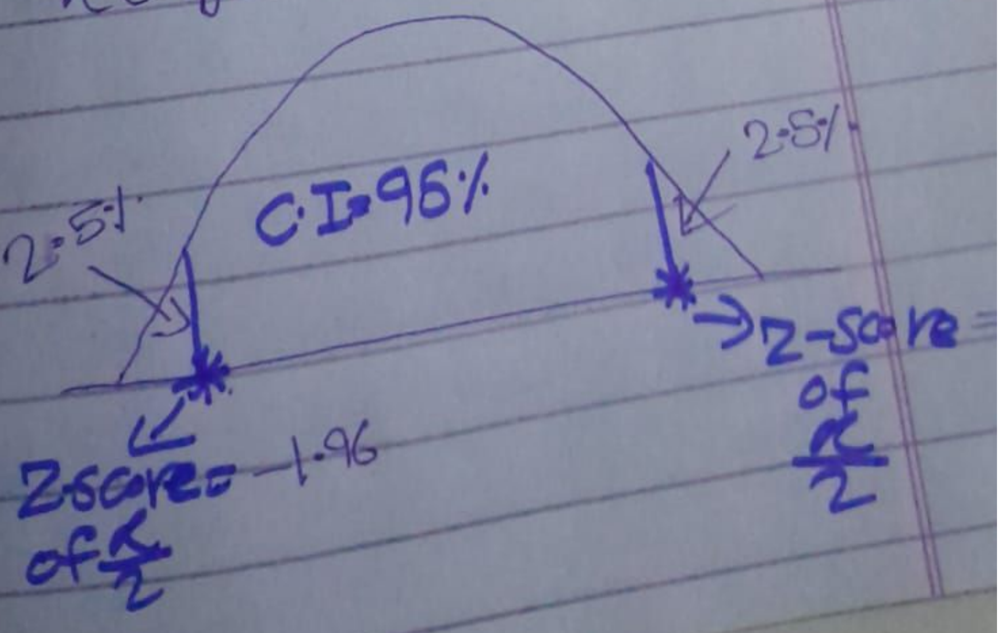
Z-score of $\frac{\alpha}{2}$

For Z-test
$C.I = \bar{x} \pm \boxed{Z\left(\frac{\alpha}{2}\right)} \frac{\sigma}{\sqrt{n}}$

E.g. On the verbal section of cat Exam the standard deviation is known to be 100. A sample of 30 test taken is a mean of 520. Construct 95% of confidence interval about the mean

Given:

σ (Population Standard Deviation) = 100
n (Sample Size) = 30
$\bar{x}$ (Sample Mean) = 520
C.I = 95% = 0.95
α (Significance Value) = 1 − 0.95 = 0.05 or 5%



$Lower\ C.I = \text{Point estimate} - Z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}$

$Lower\ C.I = \bar{x} - Z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}$

$Lower\ C.I = 520 - (1.96)\frac{100}{\sqrt{25}}$

$\boxed{Lower\ C.I = 480.8}$

$Higher\ C.I = \bar{x} + Z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}$

$Higher\ C.I = 520 + (1.96)\frac{100}{\sqrt{25}}$

$\boxed{Higher\ C.I = 559.2}$

- Conclusion:

I am 95% confident about the mean of CAT score is between 480.8 and 559.2

CHI SQUARE TEST:

Also called "The chi square test for goodness of fit" claims about population proportions

- It is a non-parameteric test that is performed on **categorical data (ordinal and nominal) data**

- **Example:**

There is a population of Male who likes different colour bikes

Bikes are	Theory categorical Distribution Theory (this %)	observed category and Distribution (like this) Sample
1. Yellow Bike	1/3	22
2. Red Bike	1/3	17
3. Orange Bike	1/3	59

We are checking the correctness of theory according to sample called goodness of fit

Goodness of fit test:

In a science class of 75 students, 11 are left handed. Does this class fit the theory that 12% of people are left handed?

Ans):

observed categorical Distribution → O

Theory categorical Distribution → E

Categories	O	E
Left Handed	11	$12\% = \frac{12}{100} \times 75 = \boxed{9}$
Right Handed	64	66
Total:	75	75

Now we will check the Goodness of fit through CHI SQUARE Test.

CHI SQUARE for Goodness of Fit:

In 2010 census of city, the weight of individuals in small city were found to be following

<50 Kg	50–75	>75
20%	30%	50%

In 2020, weight of n=500 individuals were sampled. Below are results

<50	50–75	>75
140	160	200

Usig $\alpha = 0.05$, Would you conclude the population difference of weights has changed in last 10 years?

Ans)

2010 Theory/ Expected

<50 kg	50–75	>75
20%	30%	50%

2020 n = 500 observed

<50	50–75	>75
140	160	200

Expected from sample according to 2010

<50	70–75	>75
0·2×500 = 100	0·3×160 = 150	0·5×500 = 250

① Null Hypothesis (h_0)
The data meets the expectation

② Alternate Hypothesis (h_1)
The data does not ~~meets~~ the expectation

③

α (Significance Value) = 0·05 or 5%
C·I (Confidence Interval) = 1 − 0·05
= 0·95 or 95%

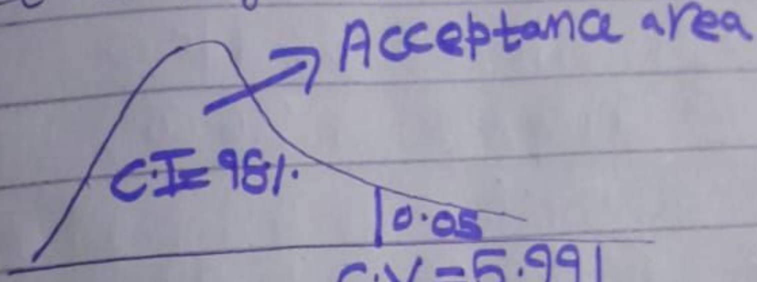
(4) Degree of Freedom(dof) = $K-1$ ← No. of categories

dof $= 3-1$

$\boxed{dof = 2}$

(5) Decision Boundary:

In CHI SQUARE distribution is not symmetric It is right skewed



C.V=5.991
↓
(critical Value)

Now we will find critical value from CHI SQUARE Table using significance value (α) and dof which is 5.991 in this case.

⇒ If X^2 (Chi square) is greater than 5.991 reject the null hypothesis else fail to reject the null hypothesis

Assumptions in ANOVA:

① Normality of Sampling Distribution of Mean (The distribution of sample mean is normally distributed)
② Absence of outliers (outliers need to be removed from data set)
③ Homogenity of Variance (Population Variance in different levels of each independent variable are equal) i.e
$\sigma_1^2 = \sigma_2^2 = \sigma_3^2$
④ Samples are independent and randomly selected.

⇒ Types of ANOVA (3 Types)

① One Way ANOVA:
one factor with atleast 2 levels and these levels are independent

(6) CHI SQUARE Test Statistics:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Observed, Expected

Now put values from data

$$\chi^2 = \frac{(140-100)^2}{100} + \frac{(160-150)^2}{150} + \frac{(200-250)^2}{250}$$

$$\chi^2 = 16 + 0.66 + 100$$

$$\boxed{\chi^2 \text{(Chi square)} = 26.66}$$

- **Conclusion:**

As $\chi^2 = 26.66 > 5.99$ so reject the null hypothesis

The weights of 2020 population are different than those expected in the 2010 population.

Analysis of Variance (ANOVA):

Definition:

ANOVA is a statistical method used to compare the means of 2 or more groups

ANOVA has:

① Factors (variable)
② Levels

e.g

Medicine (Factor)

Dosage 5mg 10mg 15mg → Levels

Mode of payment (Factor)

GPAY JAZZCASH CASH ON → Levels

Eg: Doctor wants to test a new medication to decrease headache

They split the participants in 3 conditions [10mg, 20mg, 30mg]

Doctor ask the participant to rate the headache (1-10)

Medication → Factor

$\underbrace{\text{10mg} \quad \text{20mg} \quad \text{30mg}}_{}$ Levels (Independent)

② Repeated Measures ANOVA:

one factor with atleast two levels and these levels are dependent.

E.g:

Running → Factor

Levels (Dependent) (As same person is running the days)

Day 1 Day 2 Day 3

(3) Factorial ANOVA:

Two or more factors each of which with different levels, levels can be dependent or independent.

E.g.

Running → Factor

Levels

Gender → factor / levels		Day 1	Day 2	Day 3
Male		6	8	10
		7	9	11
Female		12	14	16
		13	15	17

Hypothesis Testing in ANOVA (Partitioning of variance in ANOVA)

Null Hypothesis (h_0):

$\mu_1 = \mu_2 = \mu_3 \dots \mu_k$

(As we compare mean of different groups)

Alternate Hypothesis (h_1):

Atleast one of the sample mean is not equal

F Test Statistics:

$$F = \frac{\text{variance between samples}}{\text{Variance within samples}}$$

Let 3 samples: → Variance b/w samples

ance thin nple

x_1	x_2	x_3
1	6	5
2	7	6
4	3	3
5	2	2
3	1	4

$\bar{x}_1 = 3$ $\quad \bar{x}_2 = \frac{19}{5}$ $\quad \bar{x}_3 = 4$

Null hypothesis (h_0):

$\overline{x_1} = \overline{x_2} = \overline{x_3}$

Alternate hypothesis (h_1):

Atleast one sample mean is different

OneWay ANOVA - Case Study:

Doctor want to test a new medication which reduces headache. They split the participants into 3 conditions [15mg, 30mg, 45mg]. Later on the doctor ask the patient to rate the headache between [1-10]. Are there any difference b/w the 3 conditions using $\alpha = 0.05$?

Ans) Medication → Factor

Levels

15 mg	30 mg	45 mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

This is dataset

Steps:

1. Define null and alternate hypothesis

h_0 (null) : $\mu_{15} = \mu_{30} = \mu_{45}$

h_1 (alternate) : Not all μ are equal

2. Significance and C.I

Significance (α) = 0.05

Confidence Interval (C.I) = 1 - 0.05 = 0.95

3. Degree of Freedom

N (Total Sample Size) = No. of Rows × No. of Columns

$N = 7 \times 3 \Rightarrow \boxed{N = 21}$

n (sample size) = 7

a (levels) = 3

$dof_{between} = a - 1 = 3 - 1$

$\boxed{dof_{between} = 2}$

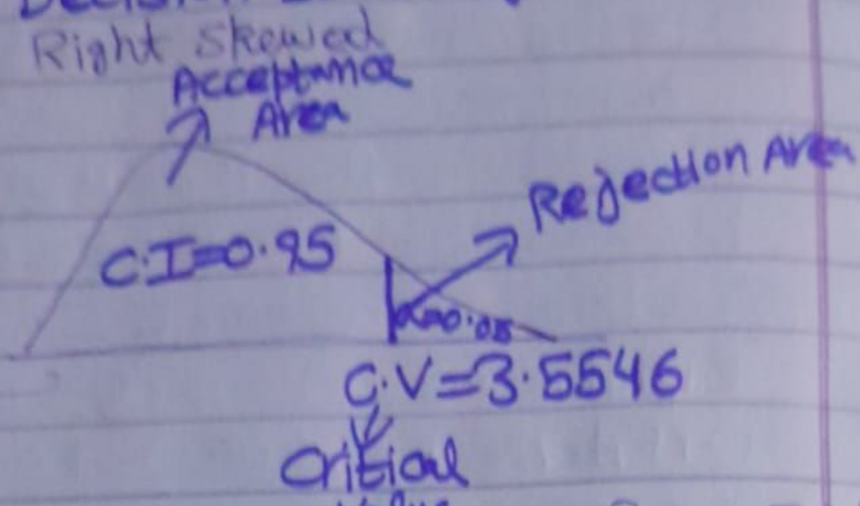
$dof_{within} = N - a = 21 - 3$

$\boxed{dof_{within} = 18}$

$dof_{Total} = N - 1 = 21 - 1$

$\boxed{dof_{total} = 20}$

④ Decision Boundary:



ow get critical value from F table
sing $\alpha=0.05$ and $df_1=2$ and $df_2=18$
hich is 3.5546

ecision Rule: If F is greater
han 3.5546 we reject the null
pothesis else fail to reject h_0.

alculate Ftest statistics:

$$F = \frac{\text{variance b/w samples}}{\text{variance within samples}}$$

(Further calculation on
next Register)

⇒ AI/ML Course:
- Inferential Statistics
- ANOVA Test

⑤ Calculate F Test Statistics Continued:

$$F = \frac{\text{variance b/w samples}}{\text{variance within samples}}$$

data:

15 mg	30 mg	45 mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

SS → Sum of square
dof → Degree of Freedom
MS → Mean square $= \frac{SS}{dof}$

	SS	dof	MS	F
Between	98.67	2	49.34	86.56
Within	10.29	18	0.54	
Total (just sum the values)	108.96	20	5.45	

① SS(Sum of Square) between:

① $SS\ between = \frac{\sum(\sum a_i)^2}{n} - \frac{T^2}{N}$

$$\left.\begin{aligned} &15\,mg: 9+8+7+8+8+9+8 = 57 \\ &30\,mg: 7+6+6+7+8+7+6 = 47 \\ &45\,mg: 4+3+2+3+4+3+2 = 21 \end{aligned}\right\} \sum a_i$$

$$SS\ between = \frac{(57)^2+(47)^2+(21)^2}{7} - \frac{(57)^2+(47)^2+(21)^2}{21}$$

$$\boxed{SS\ between = 98.67}$$

② $SS\ within = \sum Y^2 - \frac{\sum(\sum a_i)^2}{n}$

$$\sum Y^2 = 9^2+8^2+7^2+8^2+8^2+9^2+8^2+7^2+6^2+6^2+7^2+8^2+7^2+6^2+4^2+3^2+2^2+3^2+4^2+3^2+2^2$$

$$\sum Y^2 = 853$$

$$SS\ within = 853 - \left(\frac{57^2+47^2+21^2}{7}\right)$$

$$SS\ within = 10.21$$

→ $$F = \frac{MS_{between}}{MS_{within}} = \frac{49.34}{0.54}$$

$\boxed{F = 86.56}$: **Ans**

• Conclusion:

As $F(96.86) > 3.3546$

Reject the null Hypothesis (H_0)

So, Yes there is differences between 3 conditions not the mean of all are same.

model biased

balance the

both almost

• Schemes dataset:

Down