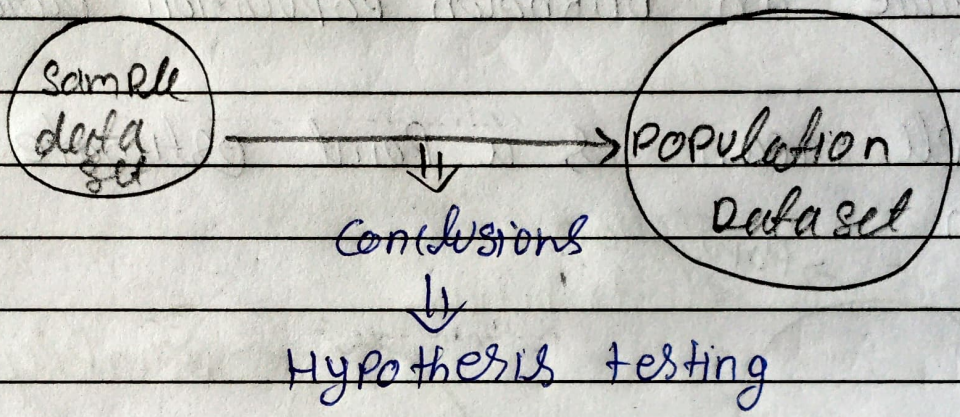
Inferential Statistics

Hypothesis Testing and Its machanism:

Inferential Stats: => Some conclusions or Infrences.



Hypothesis testing machenism:

(i) **Null Hypothesis (H_0):-**

The assumptions you are beginning with.

(ii) **Alternate Hypothesis (H_1):-**

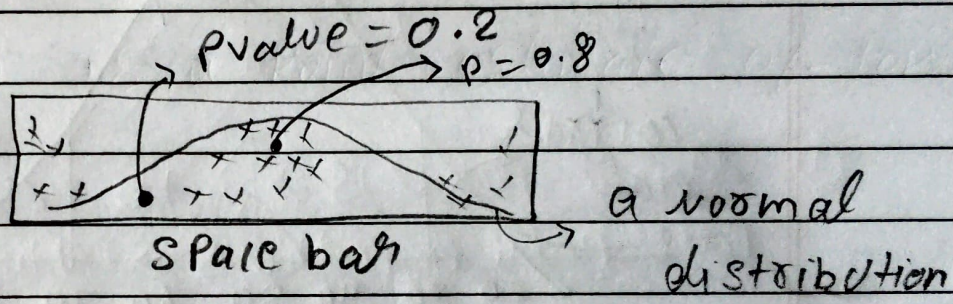
The opposite of null-Hypothesis.

(iii) **Experiments:-** → Statistical Analysis → collect the proofs.

(iv) Accept the Hypothesis or reject the hypothesis.

P-value:

The P-value is a number, calculated from a statistical test, that describes how likely you are to have hypothesis testing to help decide whether to reject the null hypothesis.



So here Pvalue = 0.2 means if we touch the SpaceBar 100 times, out of 100 touches we touch around 20 times in that region.

eg:- I am tossing a coin 100 times.

if P(H) = 0.5, P(T) = 0.5 ⇒ good

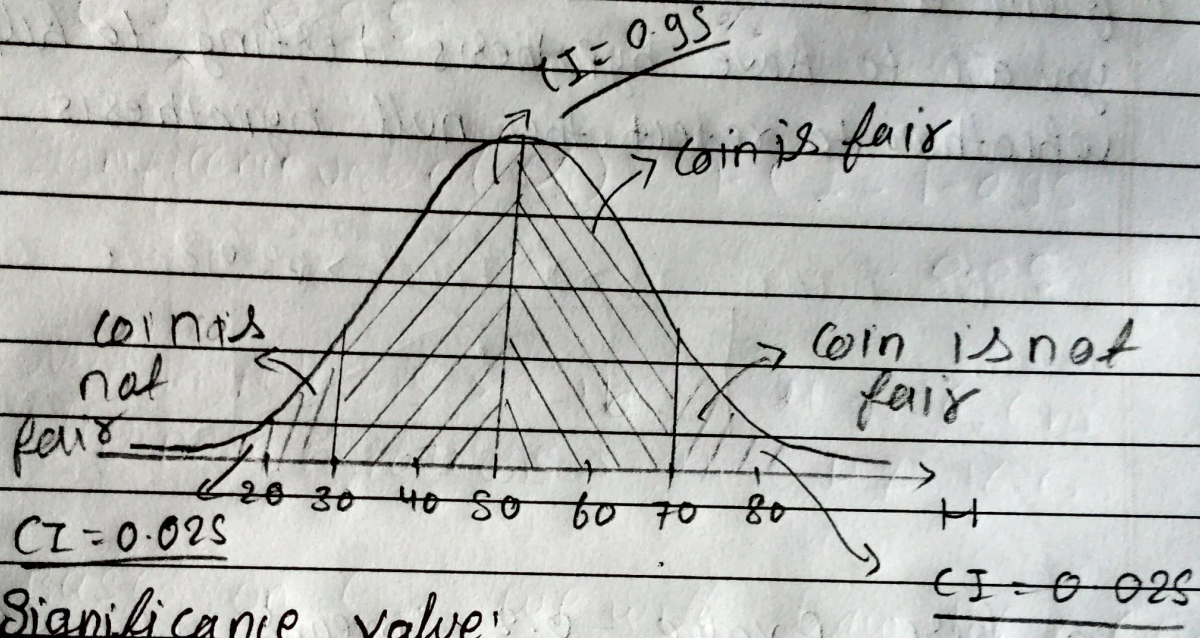
⇒ P(H) = 0.6, P(T) = 0.4 ⇒ good

⇒ P(H) = 0.7, P(T) = 0.3 ⇒ Something is wrong.

(i) Null hypothesis:- coin is fair.

(ii) Alternate hypothesis:- coin is not fair

(iii) Experiments: 100 times.



(iv) Significance value:

$\alpha = 0.05$

(area)

(CI) confidence Interval = $1 - 0.05 = 0.95$

P < Significence value => Reject the null hypothesis

Type of tests:

(i) Z-test ⇒ } ⇒ for Average data.
(ii) t-test ⇒ }

(iii) CHI-SQUARE test ⇒ for categorical data.
(iv) ANNOVA test ⇒ variance related data/analysis

Z-test:-

To apply these thing or test we have 2-conditions:

(i) Sample size > 30
(ii) we know about the Population Std.

Q. The avg heights of all residents in a city is 168cm with a (σ = 3.9) Std. A doctor believes the mean to be different. He measure the height of 36 individuals and found the avg height to be 169.5 cm

(i) State Null and Alternative hypothesis
(ii) At a 95% confidence level. Is there enough evidence to reject the null Hypothesis.

Soln

$\mu = 168 cm$

$\sigma = 3.9$

$n = 36$

$\bar{x} = 169.5 cm$

$C.I = 0.95$

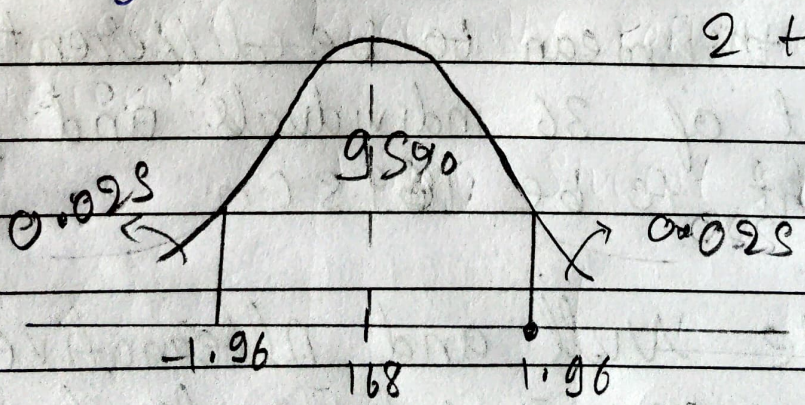
Significance value (α) $= 1 - C.I = 1 - 0.95$

$= 0.05$

① Null hypothesis $\Rightarrow$ $H_0 = \mu = 168 cm$

Alternate hypothesis $\Rightarrow$ $H_1 = \mu \neq 168 cm$

$\Rightarrow$ Based on C.I we will draw the decision boundary.



Z-score $\Rightarrow$ $1 - 0.025 = 0.975$

$\Downarrow$

area

So based on the given area we get the ± 1.96 Z-score

if Z is less than -1.96 or greather than 1.96, Reject the null hypothesis.

Z-test ⇒

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

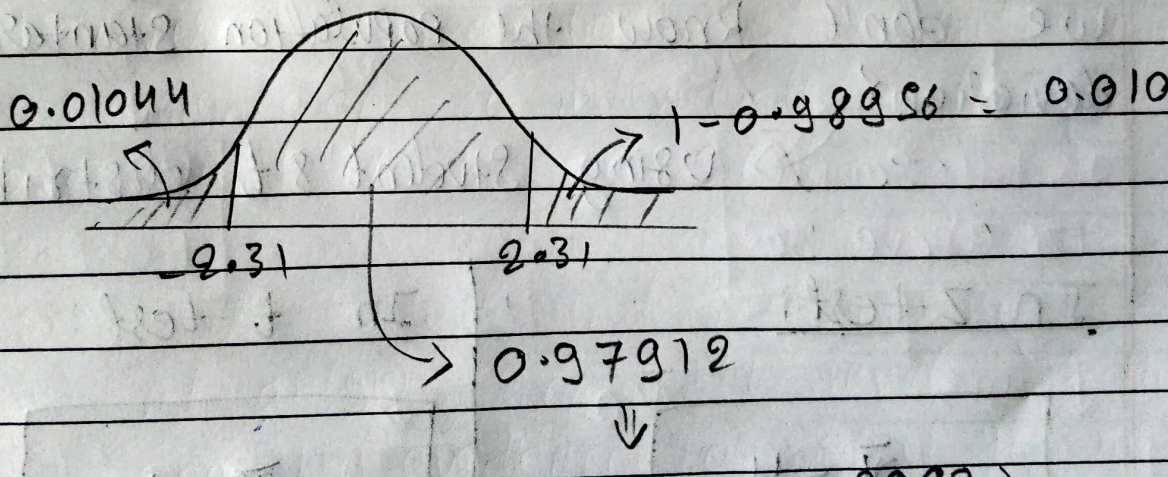
$$Z\text{-score} = \frac{\bar{x} - \mu}{\sigma}$$

apply central limit theorem $\sigma \Rightarrow \frac{\sigma}{\sqrt{n}}$

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

$$= \frac{1.5}{0.65} = 2.31$$

2.31 > 1.96 Reject the null hypothesis.



Pvalue = 0.01044 + 0.01044

⇒ 0.02088

P < 0.05

0.02088 < 0.05 => Reject the Null Hypothesis

=> Final Conclusion the Avg Heights ≠ 168cm

The avg height seems to increasing based on Sample height.

Student t-test:

In z-test when we perform any analysis using z-score we require σ (Population Standared deviation) → is already known.

=> How do we perform any analysis when we don't know the population standard deviation?

→ using Student's t distribution.

In Z-test:

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

σ → Population Std

In t-test:

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

s → sample Std

In t-test we use a parameter Degree of freedom.

$$\{Dof = n-1\}$$

n = no of data(sample)

Q- CI = 95%

$\alpha = 0.05$

$\mu = 100$, $S = 20$

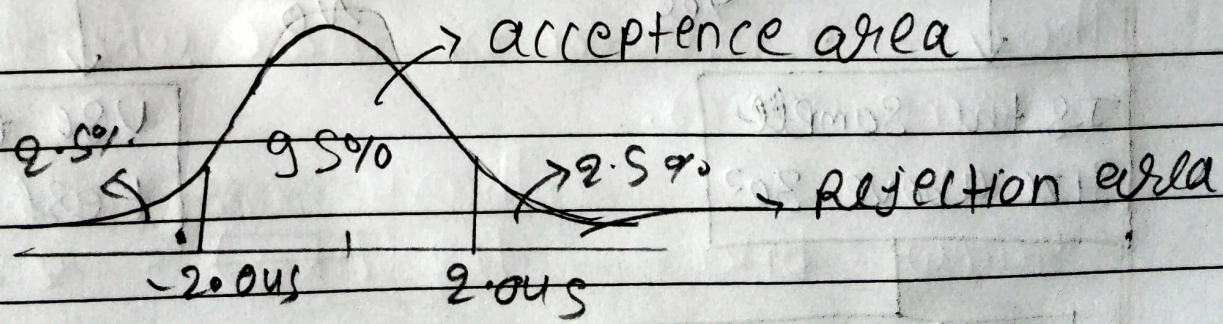
$n = 30$, $\bar{x} = 140$,

Null hypothesis :- $H_0 = \mu = 100$

Alternate hypothesis :- $H_1 = \mu \neq 100$

$dof = n-1 = 30-1$

$= 29$



If my t-test is less than -2.045 or greather than 2.045, we have to reject the null hypothesis.

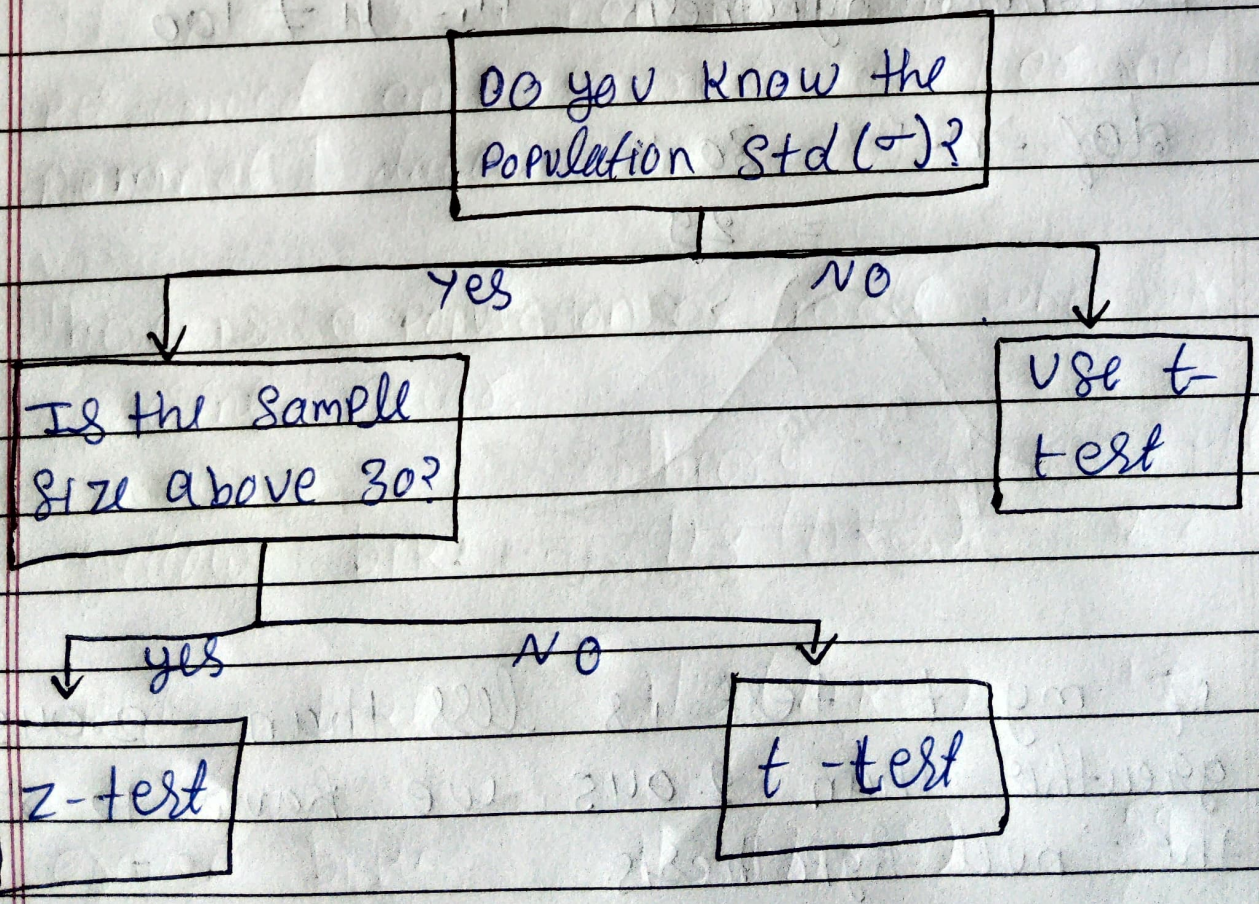
t-test

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

$$t = \frac{140 - 100}{20/\sqrt{30}} = \frac{40}{3.65} = 10.96$$

Since $t = 10.96 > 2.045$ we have to reject the null hypothesis.

When to use z-test and t-test



Bayes theorem:

Bayesian statistics is an approach to data analysis and the parameter estimation based on bayes' theorem.

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

$$\Rightarrow P(A) * P(B/A) = P(B) * P(A/B)$$

$$\Rightarrow P(B/A) = \frac{P(B) * P(A/B)}{P(A)}$$

$$\Rightarrow P(A/B) = \frac{P(A) * P(B/A)}{P(B)}$$

Bayes theorem

Dataset

Size of house	No of rooms	location	price
			O/P
x_1	x_2	x_3	y

Independent

dependent

$$P(y/x_1,x_2,x_3) = \frac{P(y) * P(x_1,x_2,x_3/y)}{P(x_1,x_2,x_3)}$$

→ Bayes theorem on your given dataset

CHI SQUARE TEST :-

The Chi square test for goodness of fit test claims about population proportion.

It is a non-Parametric test that is performed on categorical data [ordinal and nominal] data.

eg:- There is a population of mall who likes different - color.

		Theory	Sample
yellow	bike	1/3 (Likes)	22
Red	bike	1/3 (Likes)	17
orange	bike	1/3 (Likes)	99

↓ theory categorical distribution

↓ observed categorical distribution

classmate
Date
Page

⇒ Goodness of fit test we are try to implement

#

$$\chi^2 = \frac{(\text{observed} - \text{Expected})^2}{\text{Expected}}$$

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

Analysis of variance (ANOVA):-

⇒ Anova is a statistical method used to compare the means of 2 or more groups.

(i) Factors (variable)
(ii) Levels

Eg: medicine (factor)

Dosage { 10 mg, 5 mg, 15 mg } ⇒ Levels

mode of Payment (factor)

Levels <- [GPay PhonePay IMPS].

ASSUMPTIONS in ANOVA:-

① Normality of sampling distribution of mean → The distribution of sample mean is normally distributed

② Absence of outliers → outlying score need to be removed from the data set.

③ Homoginity of variance → population variance in different levels of each independent variables are equal.

$$[\sigma_1^2 = \sigma_2^2 = \sigma_3^2]$$

④ Samples are random and Independent.

Hypothesis testing in ANOVA:-

① Null Hypothesis :- $H_0 = \mu_1 = \mu_2 = \mu_3 \cdots \mu_k$

② Alternate hypothis:- H_1 = Atleast one of the sample mean is not equal

F-test :-

$$\left\{ F = \frac{\text{variance between samples}}{\text{variance within samples.}} \right\}$$

[variance between samples]

[variance within sample]

x_1	x_2	x_3
1	6	5
2	7	6
4	3	3
5	2	2
3	1	4
$\bar{x}_1 = 3$	$\bar{x}_2 = \frac{19}{5}$	$\bar{x}_3 = 4$

like the data is spread with respect to mean.

Q.

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

① Define Null and alternate hypothesis.

$H_0 = \mu_{15mg} = \mu_{30mg} = \mu_{45mg}$

H_1 = not all μ are equal

② Significance score: $\alpha = 0.05$

C.I = 0.95

③ Calculate the Degree of freedom

$N = 21$, $a = 3$, $n = 7$

$df\ (between) = a - 1 = 3 - 1 = 2$

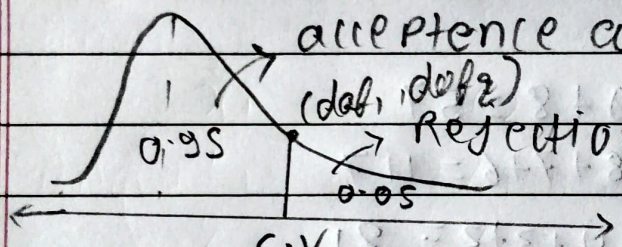
dof (within) = N − a = 21 − 3 = 18

dof (total) = N − 1 = 21 − 1 = 20

(2, 18) → dof 2, dof 1

{α = 0.05} → (C.V) it is a critical value.

(4) Decision boundry:



↳ based on the F-table

Value of C.V at (2, 18) = 3.5546

Rule:

If F is greater than 3.5546, reject the Null hypotheris.

⑤ Calculate the F-test.

	sum of square	df	mean square	F
→ between	98.67	2	49.34	
→ within	10.29	18	0.54	
→ total	108.96	20	5.44	

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

15mg = 9+8+7+8+9+8 = 57

30mg = 7+6+6+7+8+7+6 = 47

45mg = 4+3+2+3+4+3+2 = 21

$\Rightarrow \frac{57^2 + 47^2 + 21^2}{7} - \frac{[57^2 + 47^2 + 21^2]}{21}$

$\Rightarrow$ 98.67

② **SS within** $\Rightarrow \sum y^2 - \frac{\sum (\sum a_i)^2}{n}$

$\sum y^2$ = sum of square of all the value.

$$\Sigma y^2 = 9^2 + 8^2 + 7^2 + \dots$$

$$= 853$$

$$\Rightarrow 853 - \frac{[57^2 + 47^2 + 21^2]}{7}$$

$$\Rightarrow 10.29$$

$$\Rightarrow \boxed{\text{mean square} = \frac{SS}{dof}}$$

$$F\text{-test} = \frac{ms\ (between)}{ms\ (within)}$$

$$\Rightarrow \frac{49.34}{0.54}$$

$$\Rightarrow \underline{86.56}$$

$\Rightarrow 86.56 > 3.5946$ so we have to reject the Null hypothesis.