Analysis of Variance and its Types

Is a statistical analysis or method used to compare "means" of two or more groups

ANOVA
1. factors (Variables)
2. Levels

Factors :- independent variables or categorical variables that divide data into groups can be qualitative eg - treatment types, gender, region. OR quantitative eg - age group if categorized (10-20, 30-40, 60-80 age groups)

Levels :- these are distinct categories or values within each factor. Each level represents treatment, condition or group being compared in a factor.

For Ex - Medicine is a Factor then doses 10mg, 20mg 30mg can be levels

Factor → $\boxed{\text{Mode of Payment}}$

GPAY phonepe NEFT EMIs ⇒ Levels.

Assumptions in ANNOVA

we perform ANNOVA test only when these

assumption are satisfied.

① Normality of Sample Distribution Mean
- means that the "Means" avg of groups being compared should follow a Normal (bell-curve) distribution.

② Absence of Outliers
outliers need to be removed from dataset

③ Homogenity of Variance $[\sigma_1^2 = \sigma_2^2 = \sigma_3^2]$
Variability (σ^2) of data points should be roughly equal across all groups being compared

④ Samples are Independent and Random

Types of Annova

① One way Annova → one Factor with atleast two levels and, these levels are independent of each other

Eg → Doctor want to test a new "medication" to decrease headache. They split participant in 3 conditions [10mg, 20mg, 30mg] Doctor ask participants to rate headache

⇒ Medication → Factor   10mg 20mg 30mg → levels.

(2) Repeated measure Anova :- one factor, atleast two levels, levels are dependent.

Factor → Running   levels - Day 1, Day 2, Day 3

User - Day 1 - 8km   Day 2 - 5km   Day 3 - 3km

cause exhausted   distance effected due to tiredness.

(3) factorial ANOVA :- two or more factors (each of which with atleast 2 levels), levels can be independent or Dependent

Gender	Day 1	Running Day 2	Day 3
Male	8	6	7
Female	8	5	2

→ Male & Female have different energy levels.

Note - ANOVA test is done in F distribution

Hypothesis Testing in ANOVA

comparing mean

Null Hypothesis : $H_0 : \mu_1 = \mu_2 = \mu_3 \dots \mu_k$

Alternative " : $H_1 = \mu_1 \neq \mu_2 \neq \mu_3 \dots \mu_k$

or

Atleast one the mean is not equal

Test statistics :-

$$F\text{-Test} = \frac{\text{Variance } \boxed{\text{btw}} \text{ Sample}}{\text{Variance } \boxed{\text{within}} \text{ Sample}}$$

Eg $\rightarrow$

X_1	X_2	X_3
1	6	5
2	7	6
4	3	3
5	2	2
3	1	4
$\sum x_1 = 15$	$\sum x_2 = 19$	$\sum x_3 = 20$
$\bar{x}_1 = 3$	$\bar{x}_2 = 19/5$	$\bar{x}_3 = 4$

Choosing a sample & calculating Variance within it is Variance within sample

If we compare the 3 groups variance $\rightarrow$ Variance between sample

F Test $\quad H_0 : \bar{x}_1 = \bar{x}_2 = \bar{x}_3$

H_1 : Atleast one sample is not equal.

- Lets solve a problem

One factor with atleast 2 levels, levels are independent.

(1) Doctors want to test a new medication which reduces headache. They split participant into 3 condition [15, 30, 45]. Later the doctor ask patients to rate medicat[ion] [1-10]. Are there any difference btw 3 condition using alpha = 0.05?

Solution $\Rightarrow$

15mg	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

Headache rating for medicat

① Define Null and Alternate Hypothesis

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

H_1 : Atleast one μ is not equal.

② State Significance Value $\alpha = 0.05$

$CI = 0.95$

③ Degree of Freedom

N = Total No of Sample $= 21$

a = No of Categories $= 3$

n = No. sample of in a category $= 7$

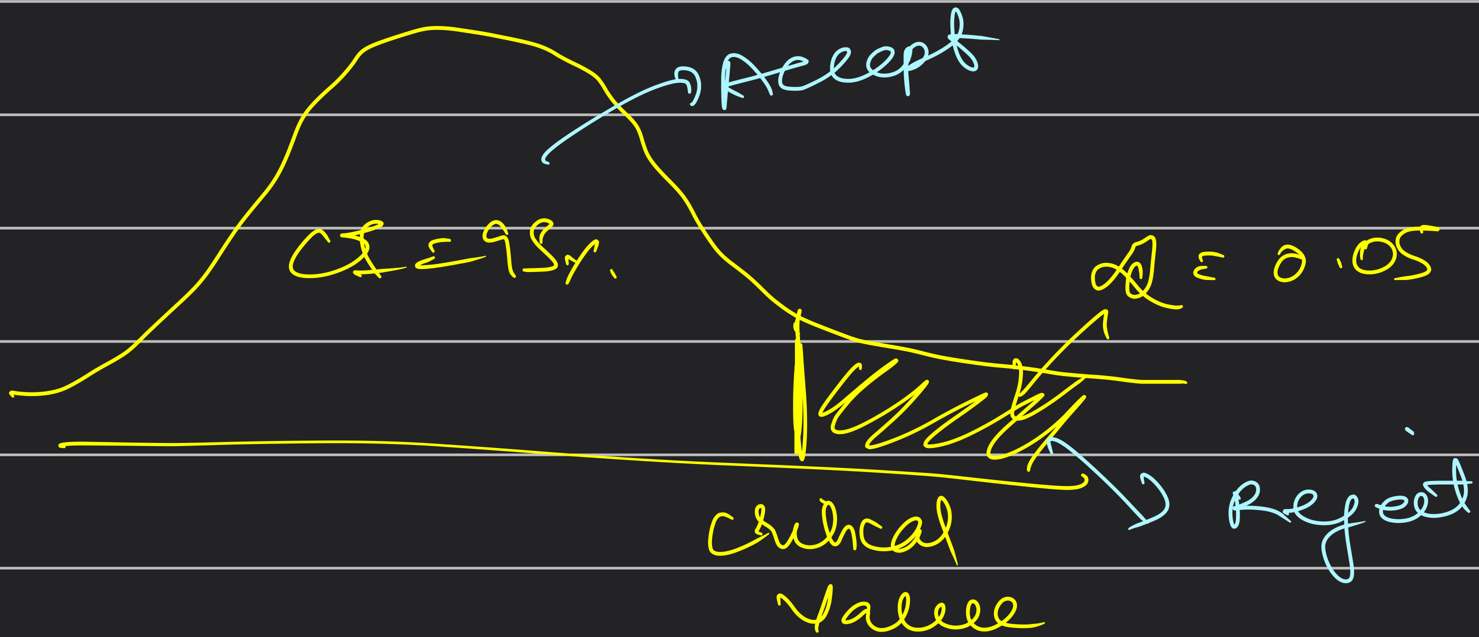
$df_{btw} = a - 1 = 3 - 1 = 2$

$df_{within} = N - a = 21 - 3 = 18$

$df_{total} = N - 1 = 20 = df_{btw} + df_{within}$
$= 2 + 18 = 20.$

$[2, 18]$ will help us determine value in F table

for F table we use F distribution a right-skewed graph



Now In F distribution
take $\alpha = 0.05$ table go to 2 column & 18th row where you get value
~~3.556~~ 3.5546

④ Decision Rule

If F test is > than Critical Value (3.5546) then we reject Null Hypothesis Ho.

(5) Calculate F test

Sum of Square | df | MS | F

SS between $= \boxed{\dfrac{\Sigma(\Sigma a_i)^2}{n} - \dfrac{T^2}{N}}$

15mg $= 9 + 8 + \ldots + 9 + 8 = 57$

20 mg $= 7 \ldots + 6 = 47$

45mg $= 3 + 4 \ldots + 3 = 21$

$T^2 = 57 + 47 + 21$

$SS_{btw} = \dfrac{57^2 + 47^2 + 21^2}{7} - \dfrac{[57+47+21]}{21}$

$\to n$

$SS_{btw} = 98.67$

$SS_{within} = \boxed{\Sigma y^2 - \dfrac{\Sigma(\Sigma a_i)^2}{n}}$

$= \Sigma y^2 - \dfrac{57^2 + 47^2 + 21^2}{7}$

$\Sigma y^2 = 9^2 + 8^2 + 7^2 \ldots + 4^2 + 3^2$
$+ 3^2 + 2^2 \to$ summing square of 21 values

$= 853 - \left[\dfrac{57^2 + 47^2 + 21^2}{7}\right] = 10.29$

	SS	df	MS (SS/df)	F
btw	98.67	2	49.34	
within	10.29	18	0.54	
Total	108.95	20		

$$F = \frac{MS_{btw}}{MS_{within}} = \frac{49.34}{0.54} = 86.56$$

Conclusion ⟹ 86.56 is greater than 3.55 we reject Ho ∴ Yes there is a difference and we are 95% confident abt it.