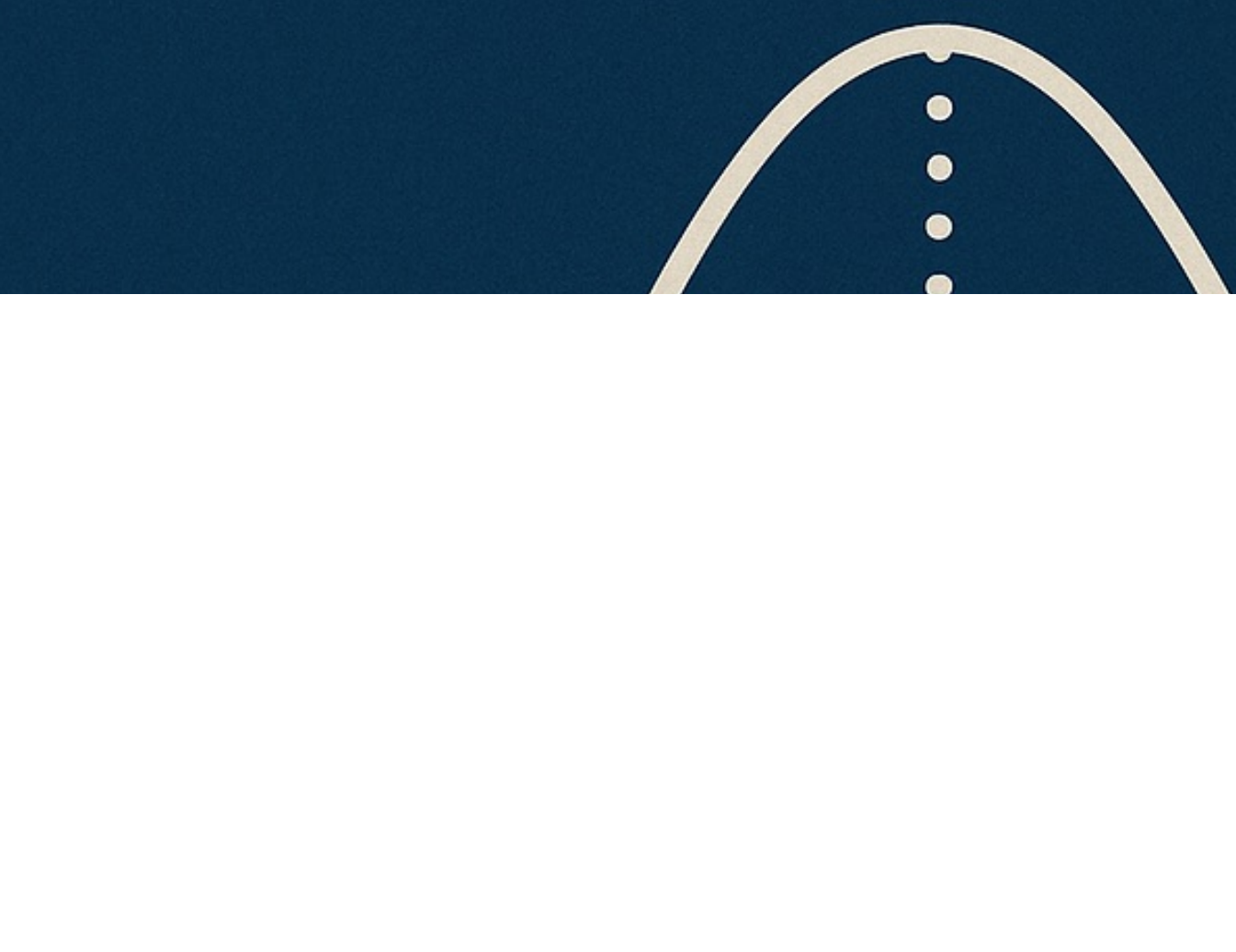
DAV-3

HYPOTHESIS TESTING

(Class starts @ 9:05 PM)



Lecture 6: ANOVA

Class starts at 9:05 PM

Agenda

1. One way ANOVA ✓
2. KW-Test (Kruskal-Wallis Test)
3. Test for Normality
   - → QQ plot ✓
   - → SW Test ✓
4. Levene's Test ✓

Types of Test

So far

✓ Numeric Vs Categorical

- 2 → (T. test) / z test
- >2 → (ANOVA)

Categorical Vs Categorical → Chi square test

Numeric Vs Numeric → Correlation

↓

in 2 days.

ANOVA

(Analysis of Variance)

Types of ANOVA

- One Way ANOVA
  - in today's class one way annova:
- Two Way ANOVA

Aerofit Example

i) Kl 781
ii) Kl 481

"C" "N"

"Product" Vs "Income"

KP781 — (i)
KP481 — (ii)
KP281 — (iii)

→ Answer

why can't we use t-test?

Boxplot
↳ 5 pt summary

Issuse wrt T-Test (Motivation for ANOVA)

T_1, T_2, T_3

3

$^3C_2 = \frac{3 \times 2}{2 \times 1}$

1. Too Many Tests

3 categories → 3 Tests

n categories → $^nC_2$ → $^{10}C_2$ → 45

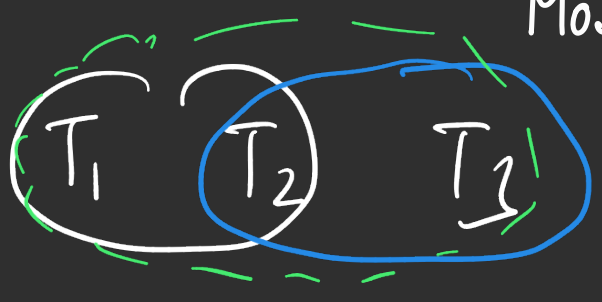
$^{10}C_2 = \frac{10 \times 9}{2 \times 1} = 45$

2. Error Compounds

α → Margin of Error (5%)

100 → 5% × 5% 5%

More Test → "more chances of Margin of Error"



$T_1\ T_2\ T_3$

$^nC_2 \Rightarrow {}^3C_2 = \frac{3 \times 2!}{2!} = 3$ Teab

3 Tests

T_1 Vs T_2
T_1 Vs T_3
T_2 Vs T_1
T_2 Vs T_3
T_3 Vs T_1
T_3 Vs T_2

permutation ⇒ comb ⇒

T_1 Vs T_2
T_2 Vs T_3
T_3 Vs T_1

2 T_1 and T_2 KPHI KPHI

2) Error rate compounds:

what is margin of error that we consider?

$\alpha = 5\%$ → this amount of error business can handle

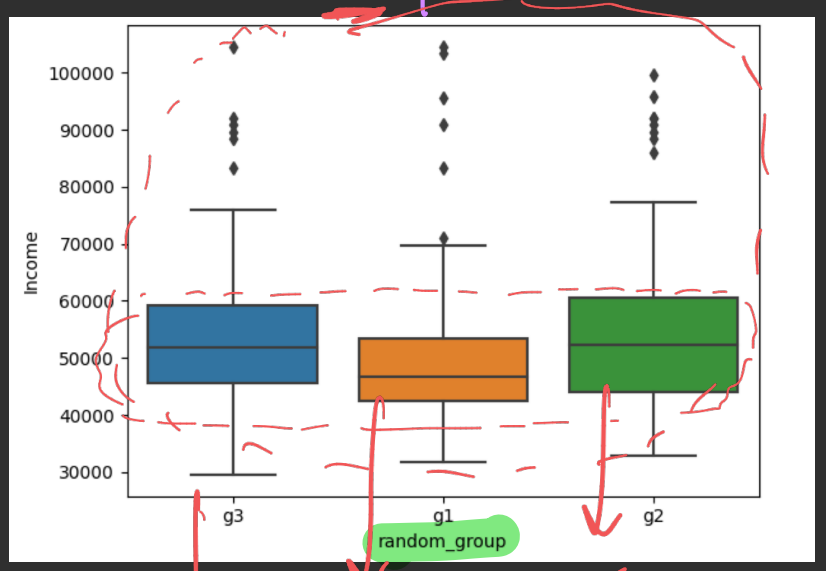
If I am doing 100 tests what will be MoE?

100 → 5% × 5% × 5% — — — 100

#Aerofit Example

(Which setup has a higher variance?)

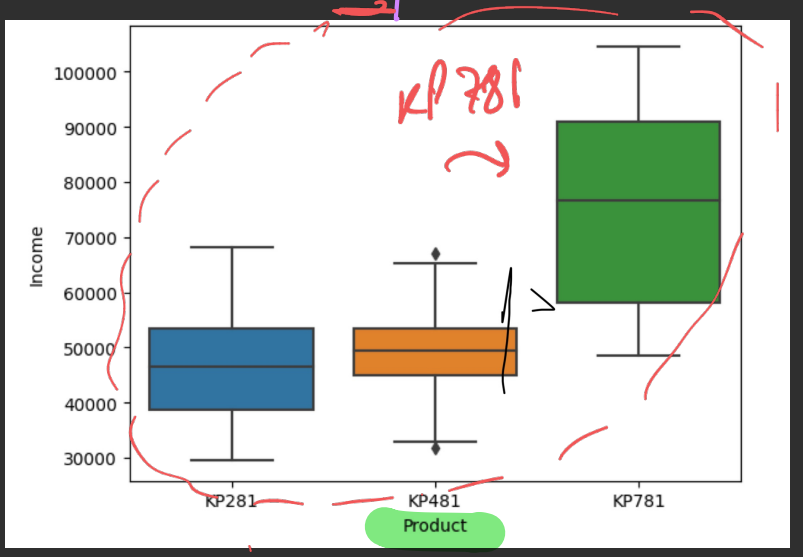
Setup I



I II III

setup 1

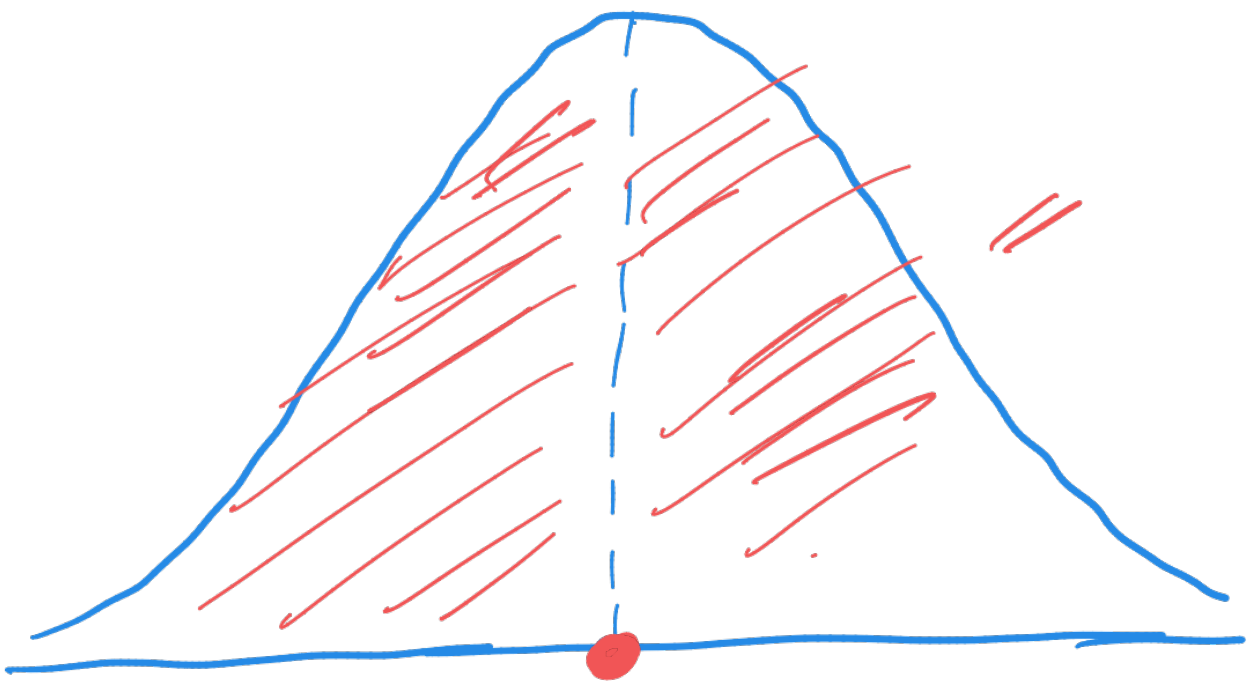
Setup II



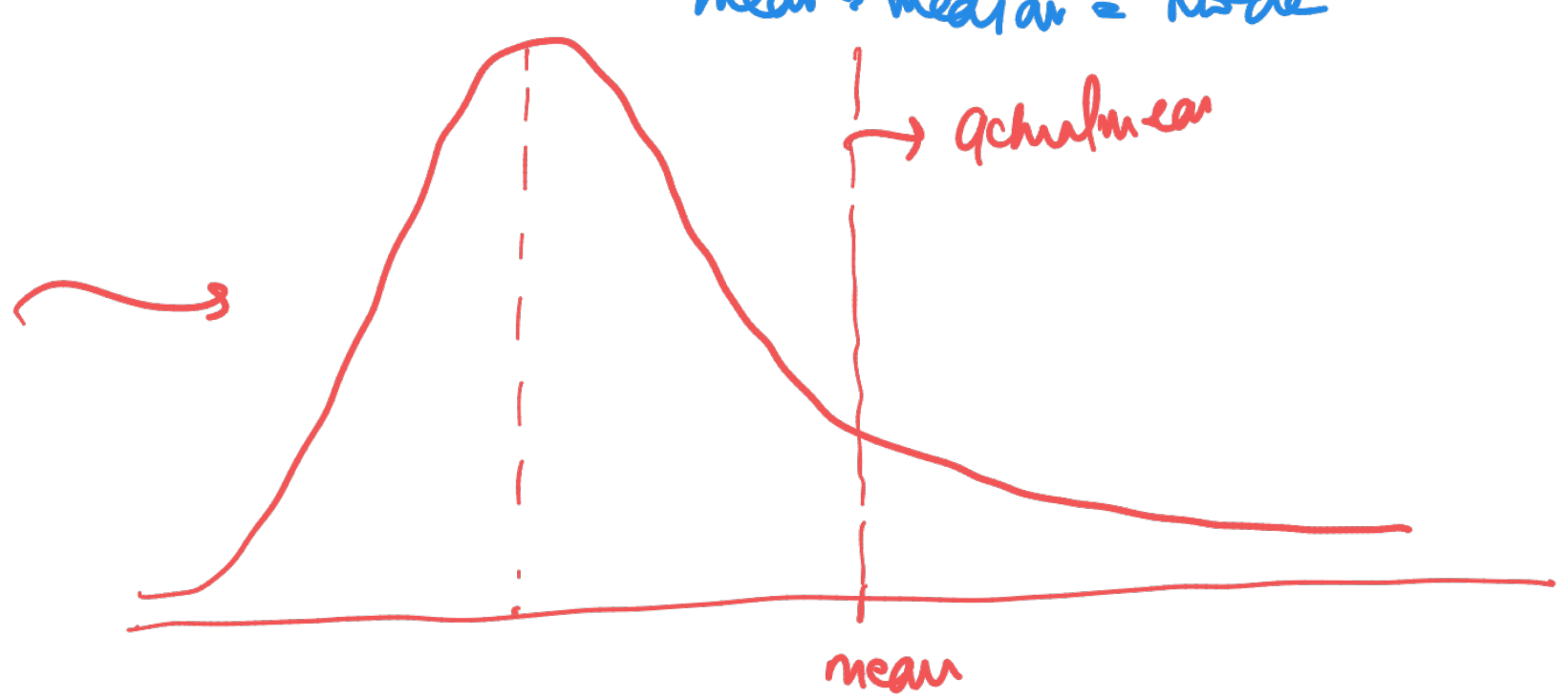
Setup 2

H_0 :- mean of all 3 groups are same

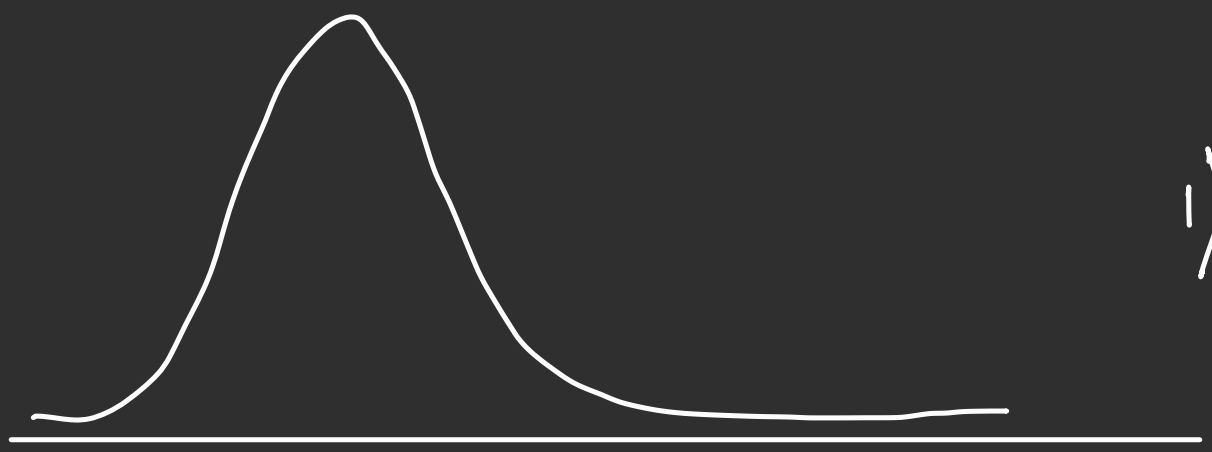
H_1 :- mean of all 3 groups are not same!



$$\frac{(x_i - \mu)^2}{N}$$

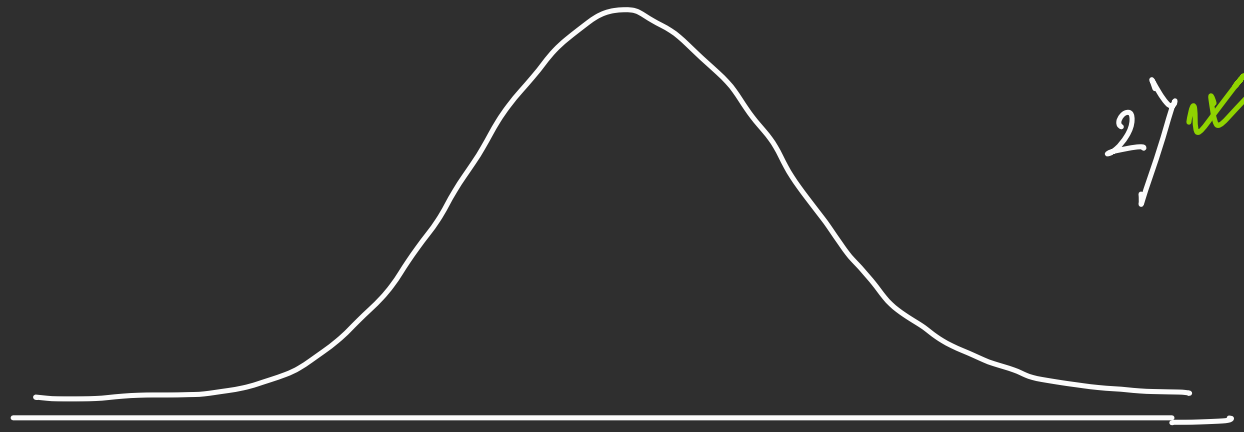


less variance here as compared to the normal distribution



1)

$$\text{var} = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$$



2) $\sqrt{\text{var}}$

$$\text{std} = \sqrt{\text{var}}$$

H_0: means of all 3 groups are same

H_a: means of all 3 groups are not same / or atleast one of the group is different

H_a:

$\mu_1 = \mu_2 \neq \mu_3$

$\mu_1 \neq \mu_2 = \mu_3$

$\mu_1 = \mu_3 \neq \mu_2$

ANOVA

↳ Analysis of variance

$H_0: g_1 = g_2 = g_3$

$H_a:$ group means are not same

$H_0:$ } Means

$H_a:$ }

New Terminologies

① Variance within each group/categories :

A vs B
B vs C
C vs A

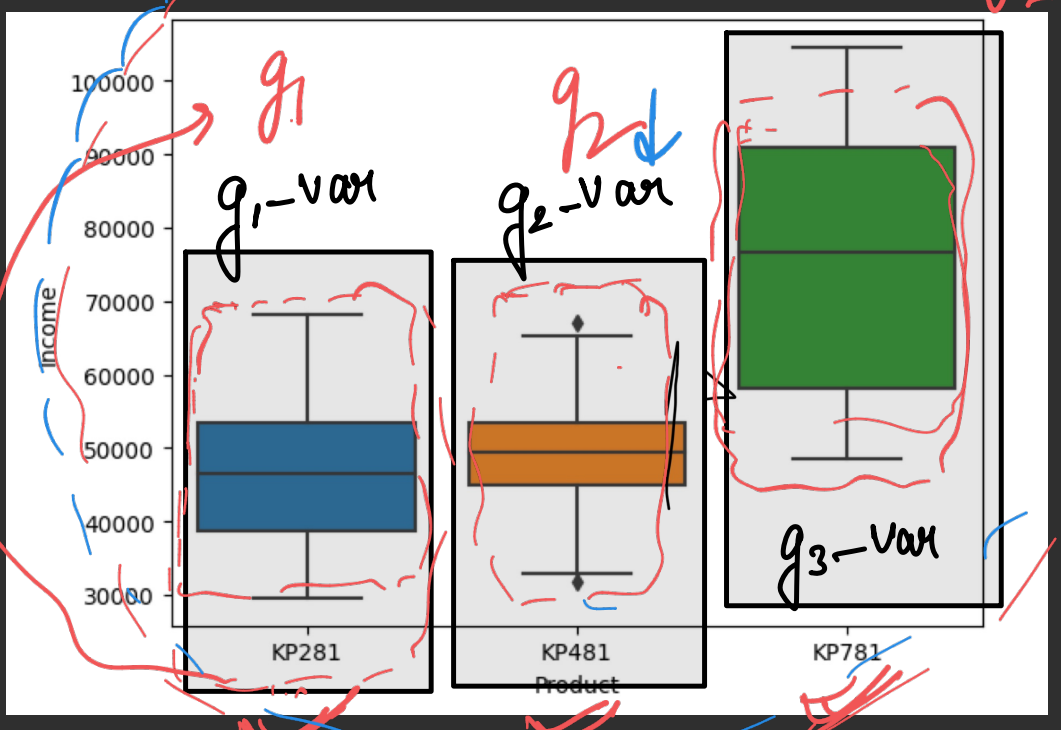
g_1, g_2, g_3

$$\text{var} \Rightarrow \frac{(x-\bar{x})^2}{n}$$

↑ 2 set of g_1

How to calculate?

$g_1, g_2, g_3 \longrightarrow$ data



g_1-avg, g_1-var

g_2-avg, g_2-var

g_3-avg, g_3-var

variance within group

within

1 2 3

we will be getting one value

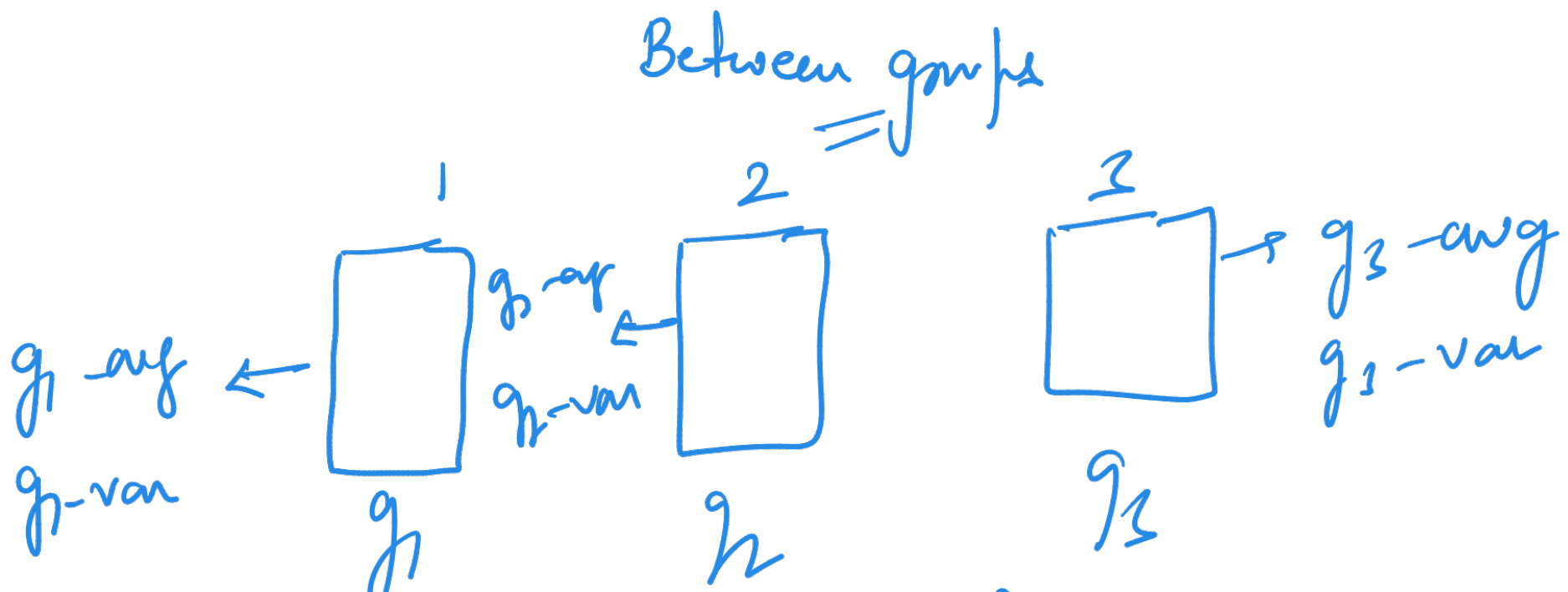
g_1-var g_2-var g_3-var

$\frac{(x_i - u)^2}{N}$ $\frac{(x_2 - u)^2}{N}$ $\frac{(x_3 - u)^2}{N}$

x y z

Total within sum of squares = $x + y + z$ → one value

Between groups



$g_1\ g_2\ g_3.\text{mean}() \rightarrow g\text{-avg}$

$$g\text{-var} = \frac{(g_1 avg - g avg)^2 + (g_2 avg - g avg)^2 + (g_3 avg - g avg)^2}{3}$$

(2) Variance between groups :

g_1, g_2, g_3 . mainly

How to calculate ?

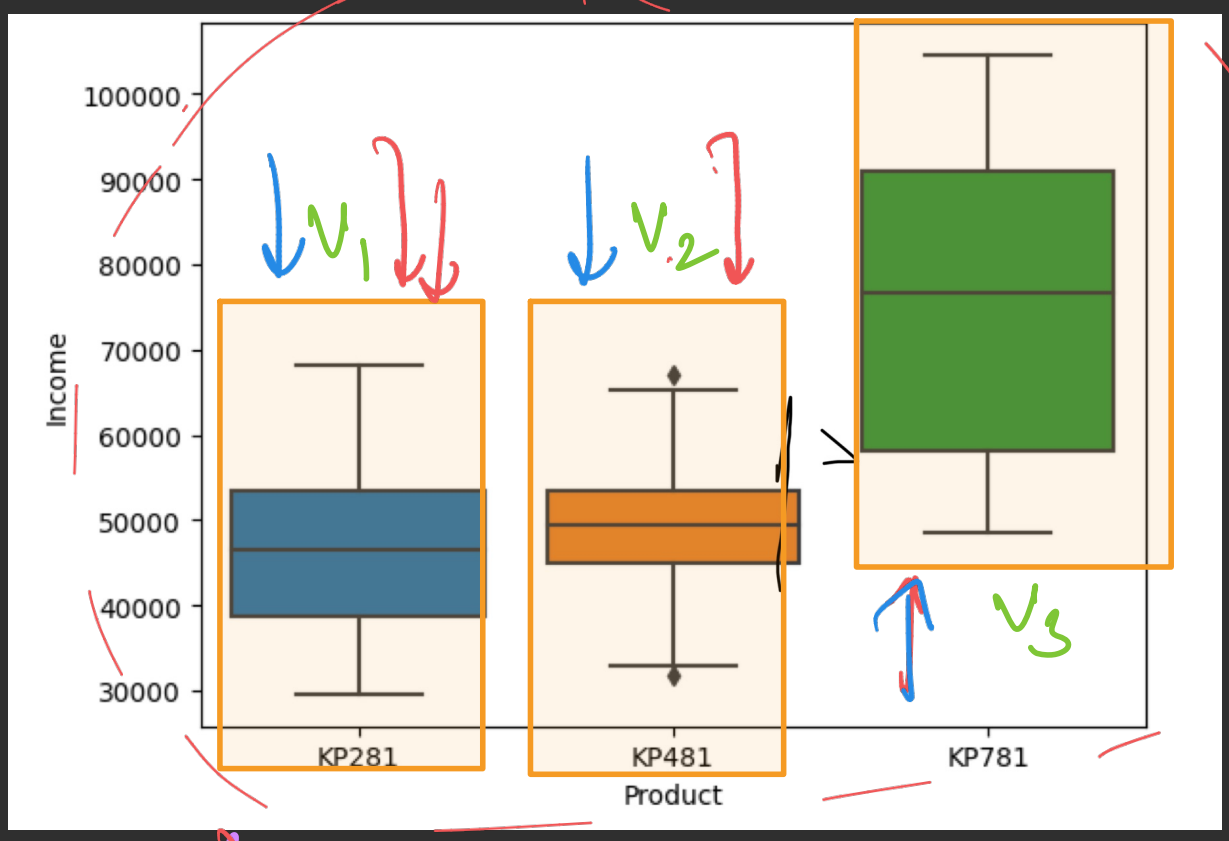
global_avg $\longrightarrow$ g_1, g_2, g_3 $\rightarrow$ g_avg

g_1, g_2, g_3, g_4

global_var $\Rightarrow$

$$\frac{(g_1 avg - g avg)^2 + (g_2 avg - g avg)^2 + (g_3 avg - g avg)^2 + (g_4 avg - g avg)^2}{N\text{-}s}$$

Here N $\longrightarrow$ 3 (3 categories/group)



- V → Variance within group

V → Variance b/w groups

Variance b/w groups

Test Statistics

F ratio = Variance b/w groups / Variance within group

F statistic

F ratio = $\frac{(x_i - \mu)^2}{N}$ / $\frac{(x_j - \mu_j)^2}{N}$ = $\frac{(O-E)^2}{E}$ / $\frac{(O-E)^2}{E}$

F ratio = B/W

F ratio ↑ ⟶ low pvalue ⟶ Reject H_0

- W F ratio ↑ ∝ Variance b/w group ↑
- W F ratio ↑ ∝ 1/Variance within group ↓

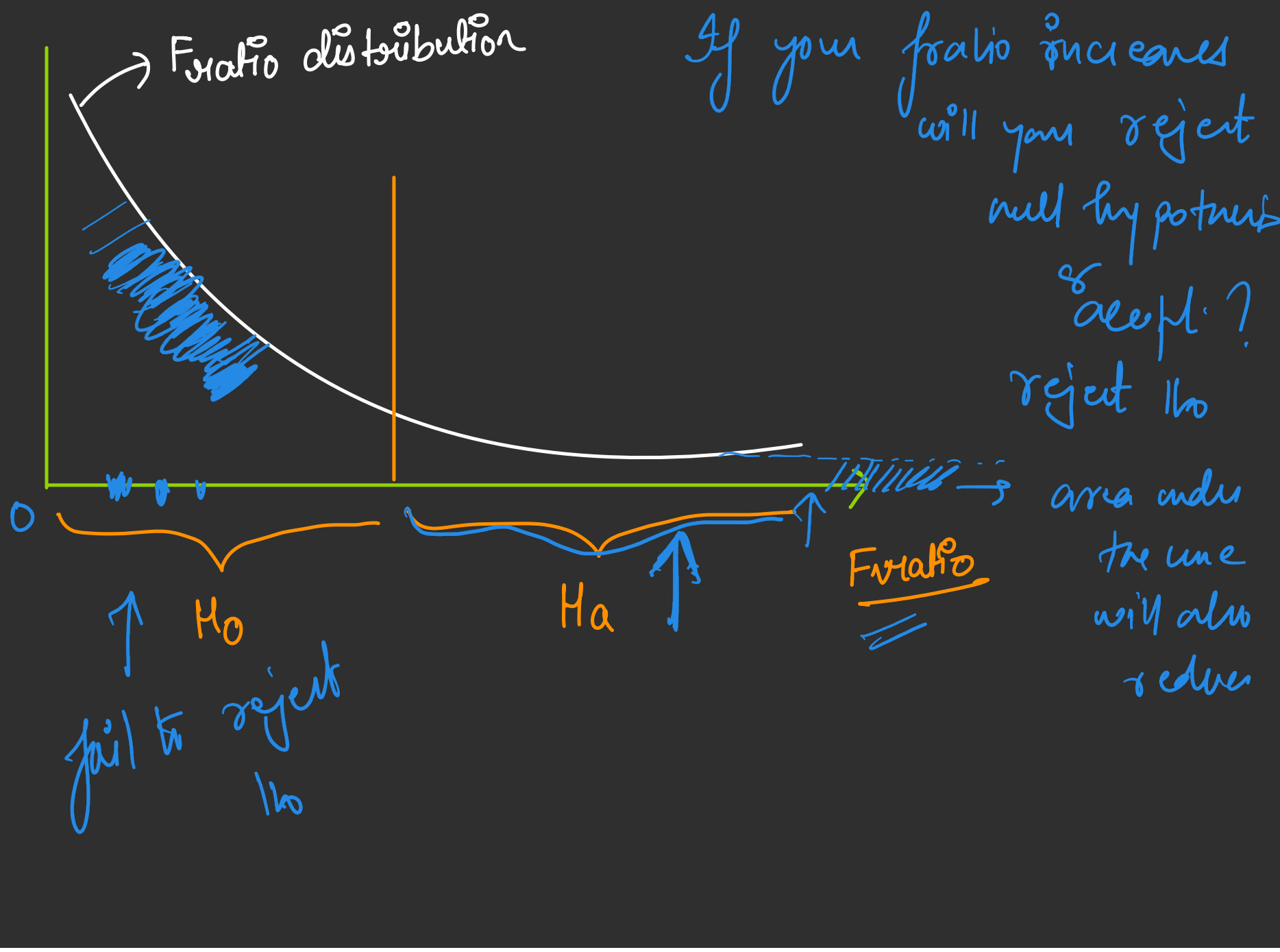
$$F_{ratio} = \frac{\text{Var b/w groups}}{\text{Var within groups}} = \frac{\frac{(x-\mu)^2}{N}}{\frac{(x-\mu)^2}{N}}$$

$df_1 = v$

$df_2 = v$

$Chi^2 \rightarrow$ Always pos

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


F ratio distribution
If your f ratio increases will you reject null hypothesis or accept?
reject Ho
area under the curve will also reduce
0
Ho
Ha
F ratio
fail to reject Ho

#Quiz

F ratio = Variance b/w groups / Variance within group

B/W

F statistic

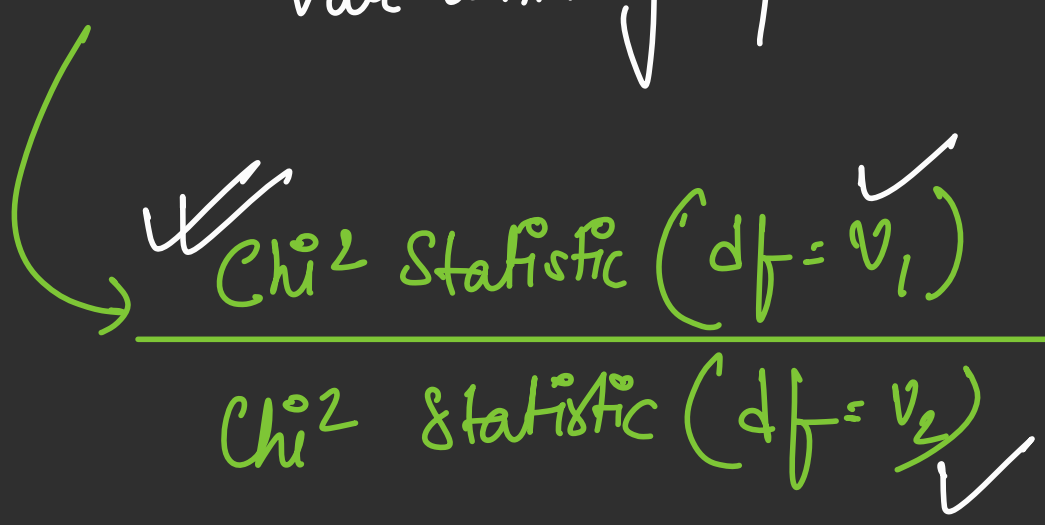
(Test Statistic)

H_0: means are same b/w groups (There is no significant d/f b/w groups)

H_a: Opp

F-ratio $\Rightarrow$ $\dfrac{\text{Var b/w groups}}{\text{Var within groups}}$

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



Assumptions of ANOVA

1. Data should be Normal / Gaussian (Income Example) (Normality Test)
2. Data is independent of each other (Income of one person is not dependent on each other)
3. Equal variance in different group

Gaussian → QQ plot

Equal variance → Levene's Test

Equal sample size

ANOVA assumption NOT follow

↓ Alternative

non-par

(Kruskal-Wallis Test) (KW Test)

(Alternative of ANOVA)

Kruskal - Wallis Test

Test to compare the median of two or more independent group.

↳ Alternative to one-way ANOVA (useful when data is non-linear)

Normality

H_0: Population of all groups have same median

H_a: At least one of the group has different median

Non parametric Test → Does not care what the underlying dist of data

Advantages of KW Test

1. No assumption of normality ✗
2. Robust to outliers ?

Limitations of KW Test

1. Doesn't really tells you which group is different (Also same for ANOVA) ← anova also does not tell that
2. Less powerful than ANOVA.

parametric Test

Z-test

T-test

ANOVA

Non-parametric

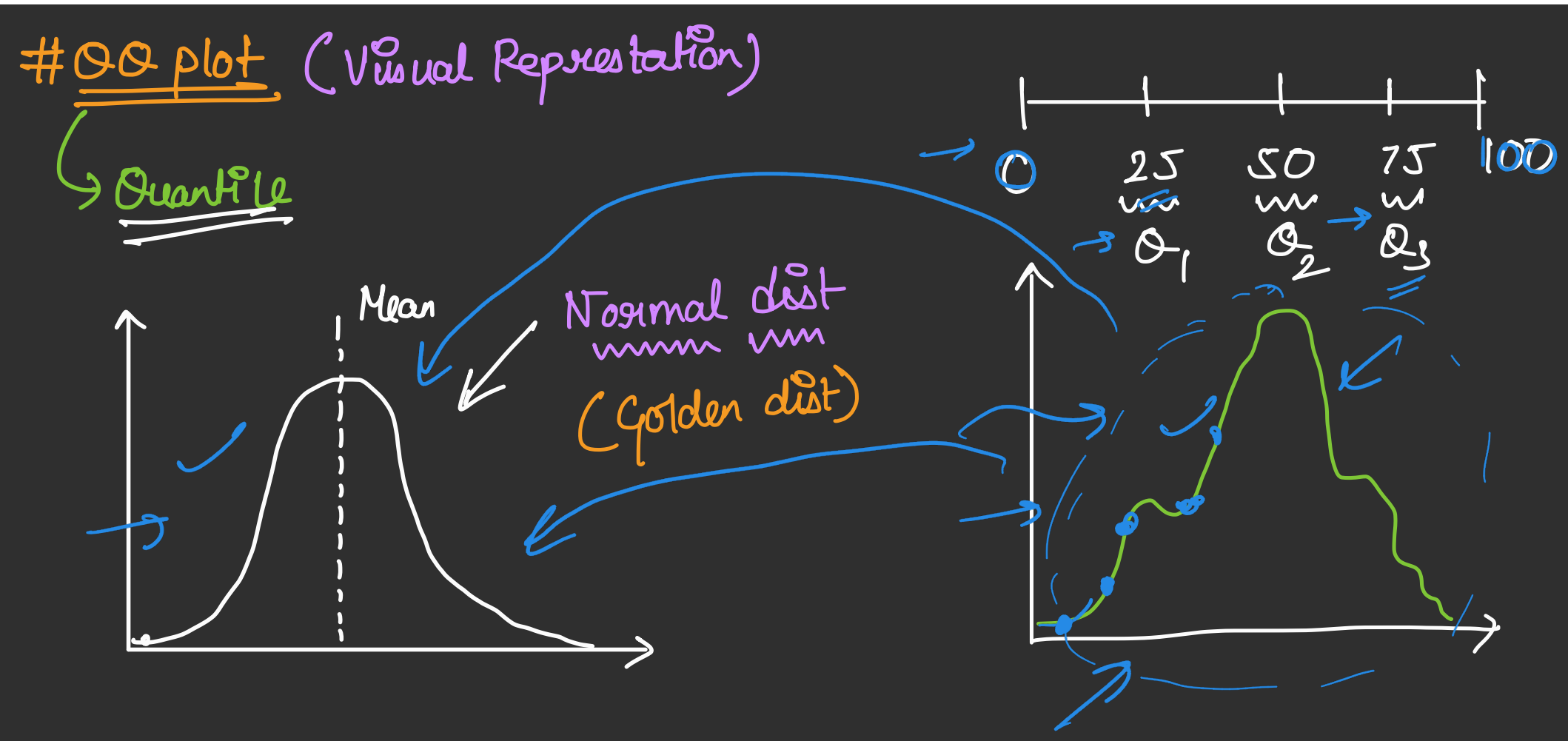
KW Test

Normality Test

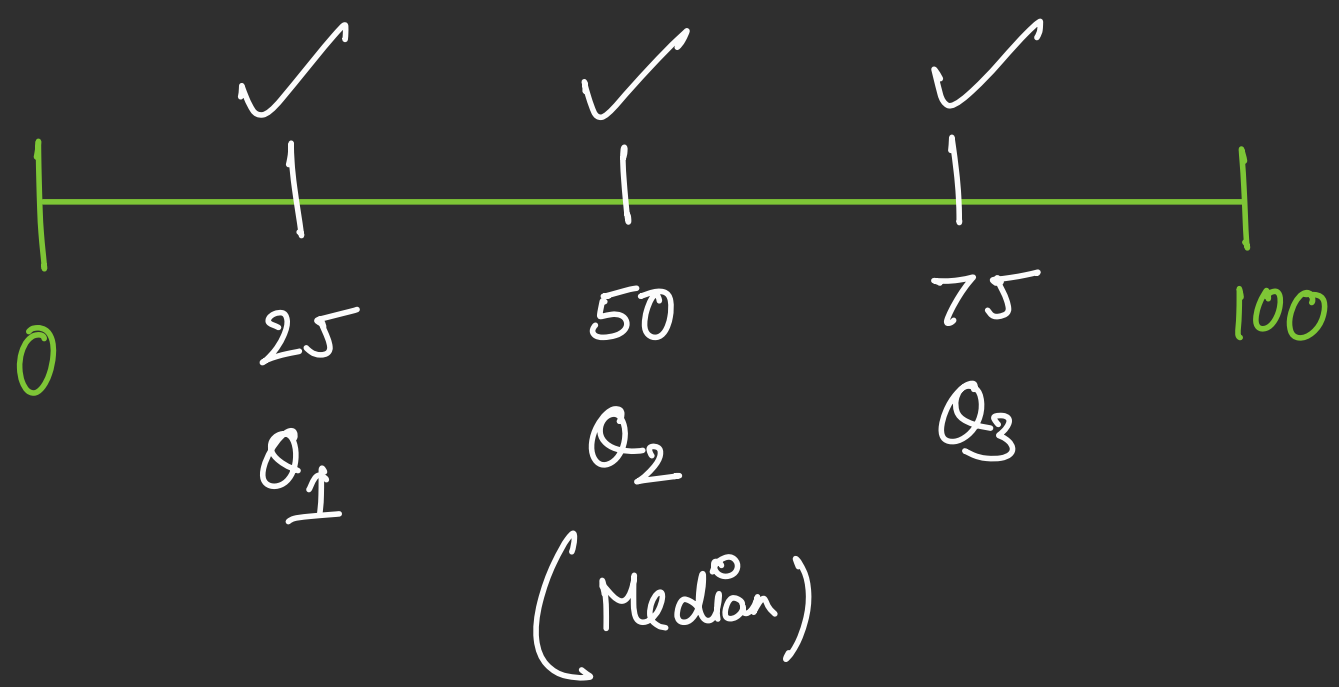
1. QQ plot [Visual Approach]
2. Shapiro-Wilk Test [HT approach]

QQ plot (Visual Representation)

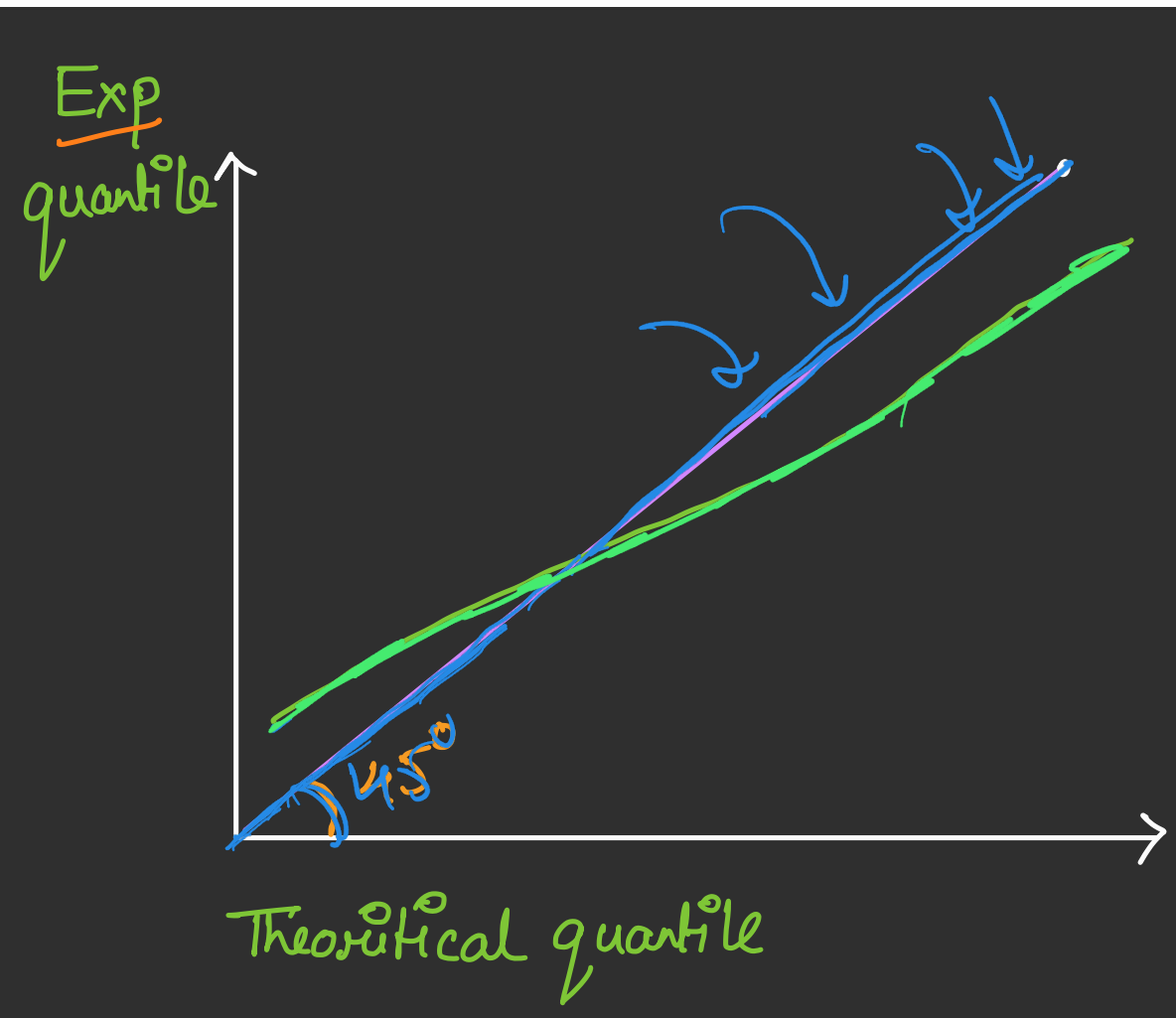
→ Quantile



"superimposition"


0
25
Q_1
50
Q_2
(Median)
75
Q_3
100

normal dist

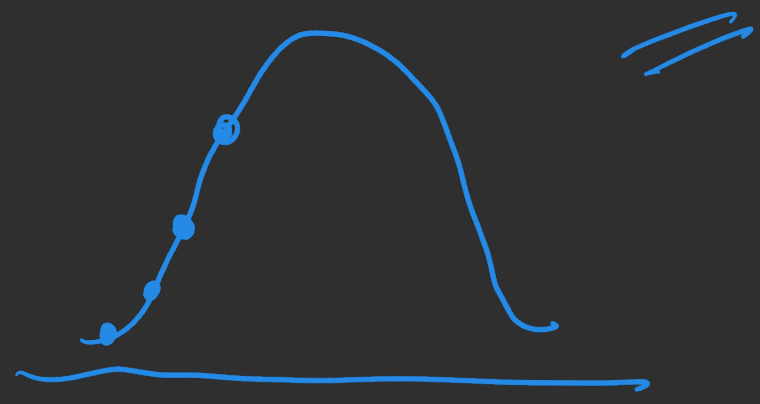


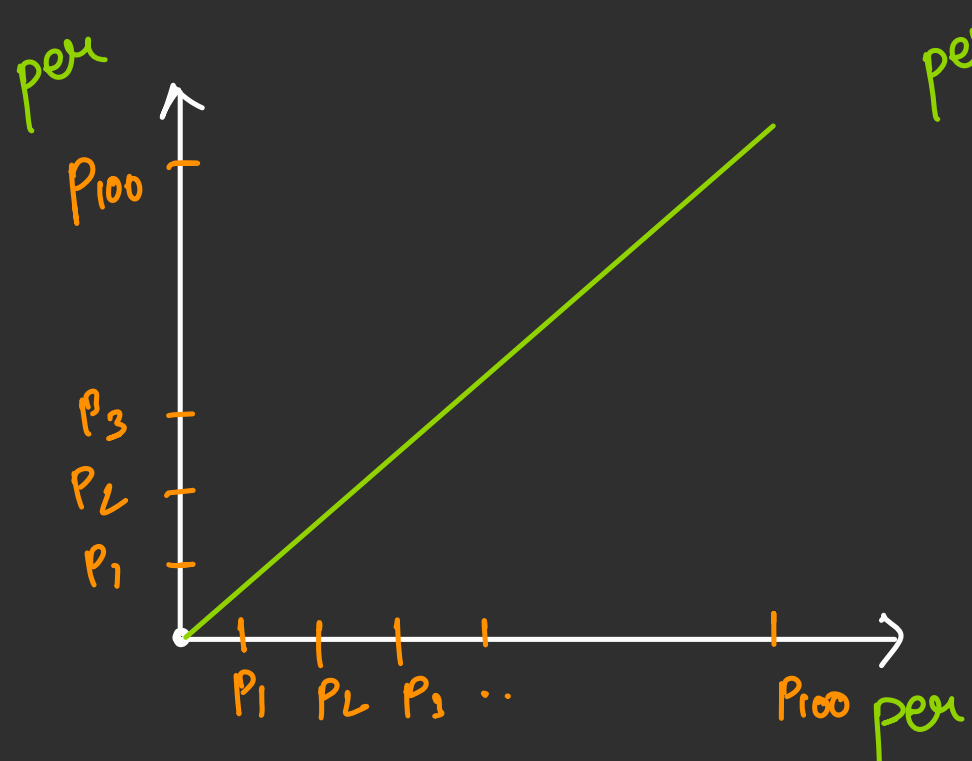
Theoritical

$P_1 \rightarrow$
$P_2 \rightarrow$
$P_3 \rightarrow$
P_4
⋮
P_{100}

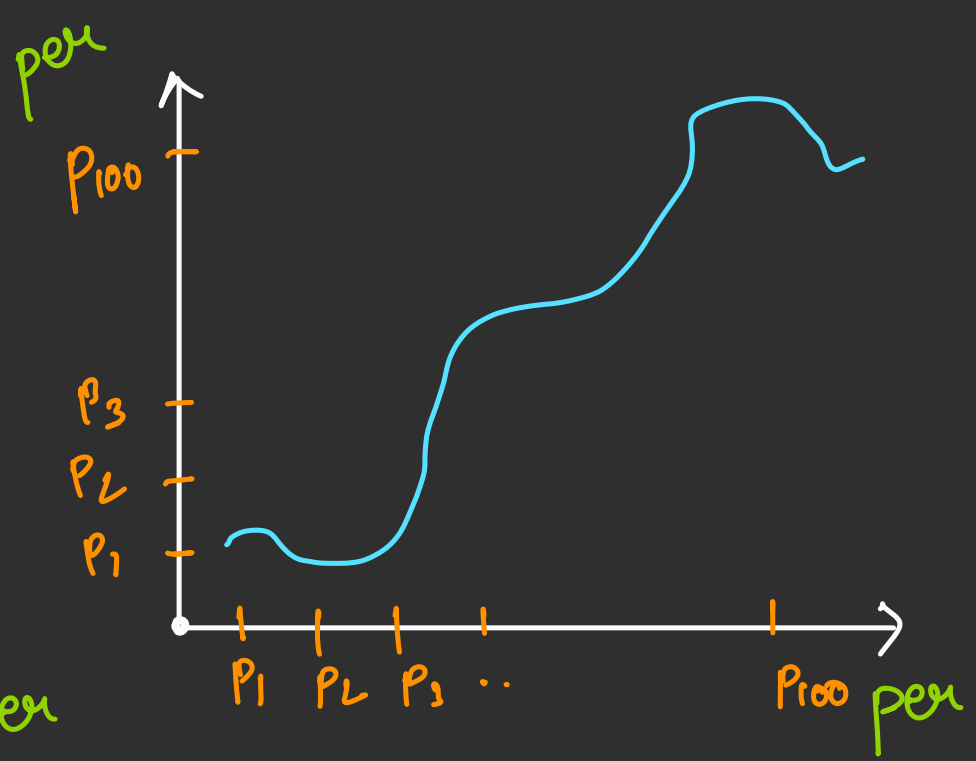
Exp

P_1
P_2
⋮
P_{100}





Emperical Dist
(Normal)



Your dist

Shapiro-Wilk Test (Statistical Test)

$\begin{cases} H_0: \text{Data is Gaussian} \\ H_a: \text{Data is not Gaussian} \end{cases}$

[Note:
This test may not work well when data is too large]

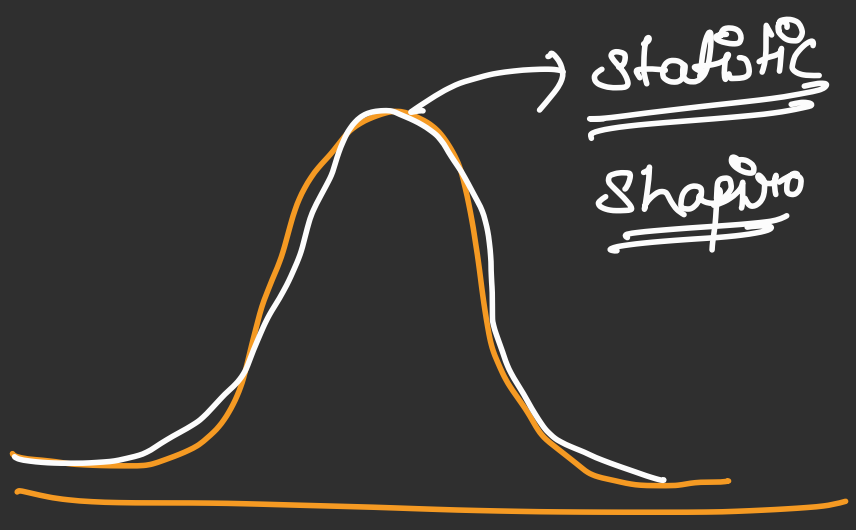
$\alpha = 5\%$

if pvalue $< \alpha$:

Reject H_0

else:

fail to Reject H_0



Levene's Test (Aka Test of Variability)

Assumption of ANOVA: "There should be equal variance in diff groups/categories in our data".

$$f\text{ratio} = \frac{(\text{Var across group})}{\text{Var within group}}$$