Statistical Tests - 3

- ✓ χ^2-test
- ANOVA

↓

we already know the framework

Next-class

↓

- aov ML
- DL

Test of independence: χ^2-test (pearson)

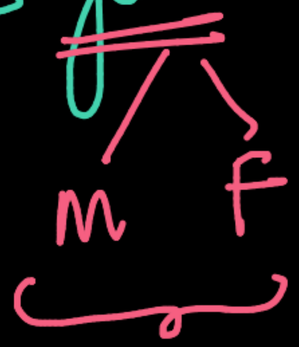
Framework

Biz:

Youtube / Spotify / Amazon

premium-model

Q Does gender matter in becoming a premium sub?



Q Is premium-sub dependent on gender

Y/N

M/F

Framework:

1

Ho: Gender has no effect on premium sub

Ha: " has effect on " "

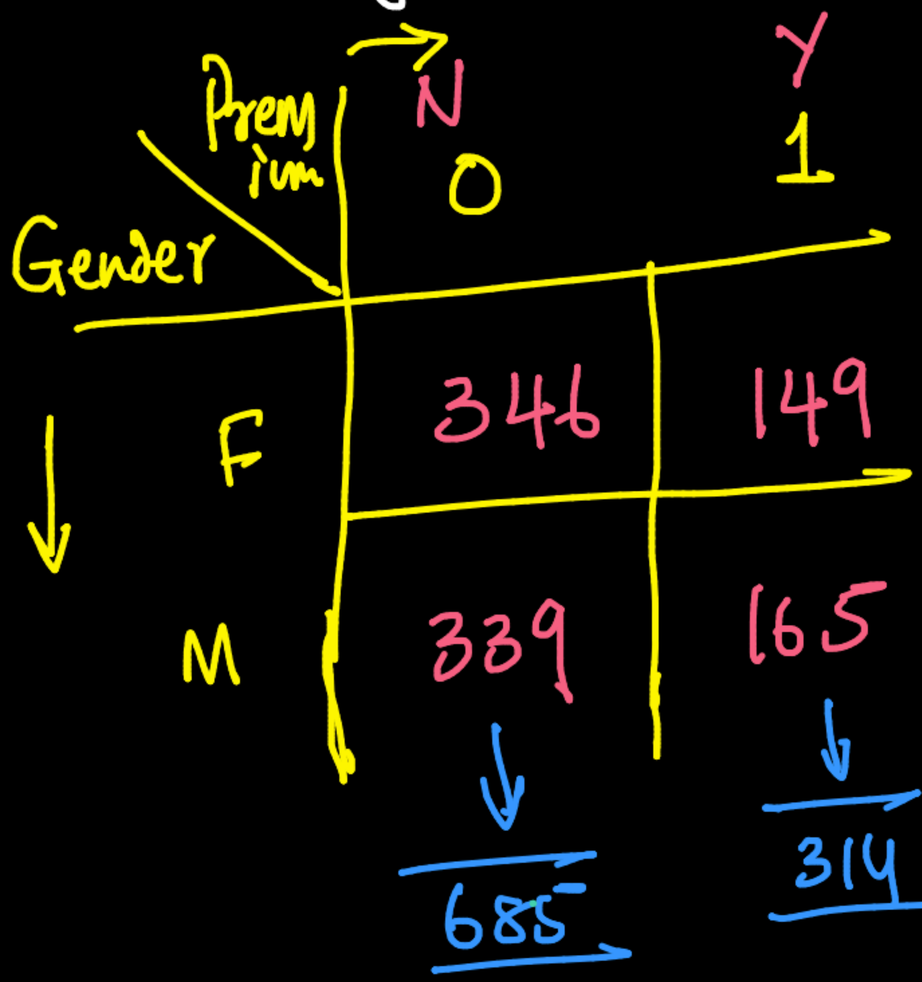
Safari File Edit View History Bookmarks Develop Window Help

Data: Contigency table:

F : M = 495 : 504

0 : 1 = 685 : 314

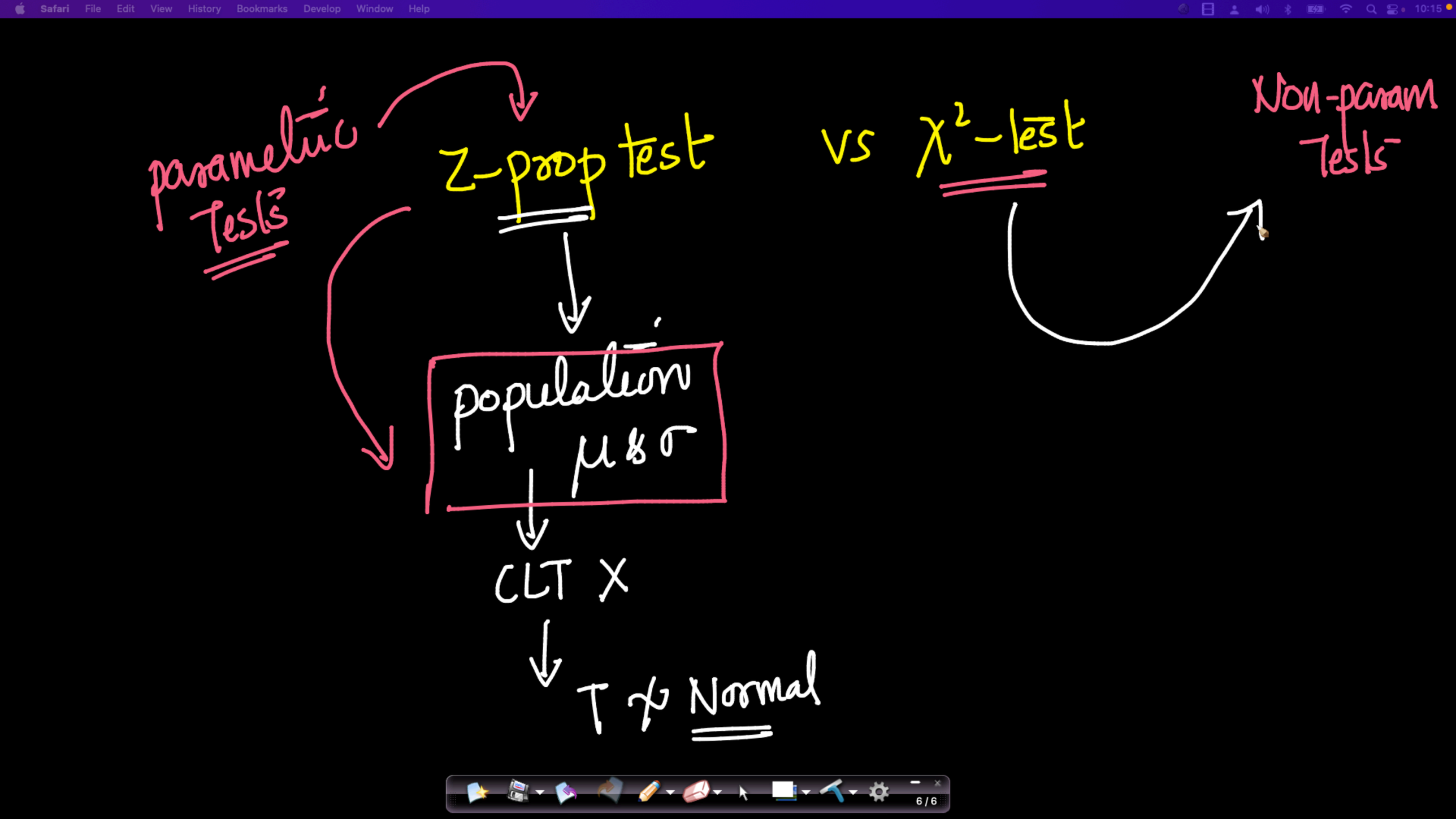
observed



→ 495

→ 504

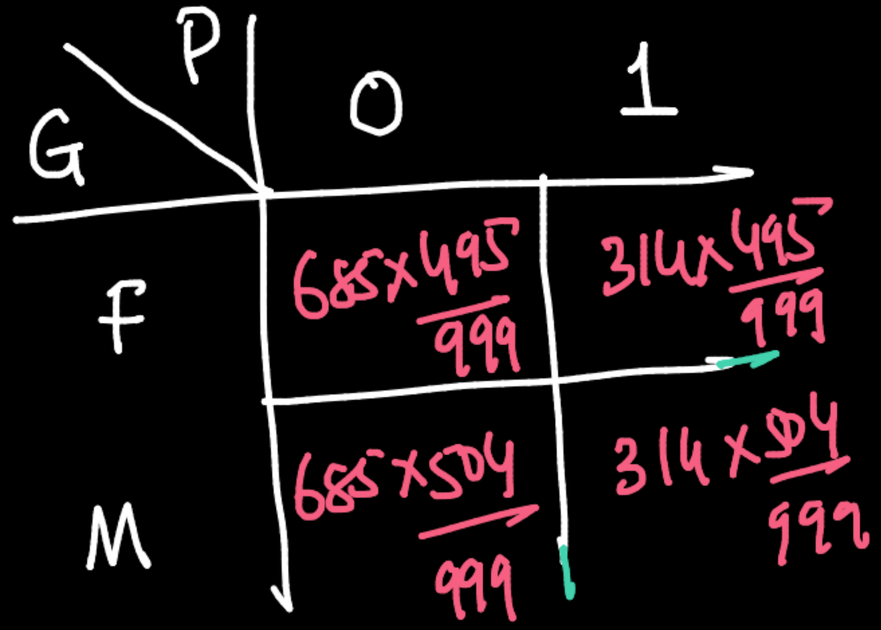
999


parametric tests
Z-prop test
vs χ^2-test
Non-param tests
Population
μ & σ
CLT ✗
T ✗ Normal

Correlation - coeff → numerical variables

(PCC)

Expected under Ho : no effect of Gender on premium subs ..



F : M = 495 : 504

685 314 999

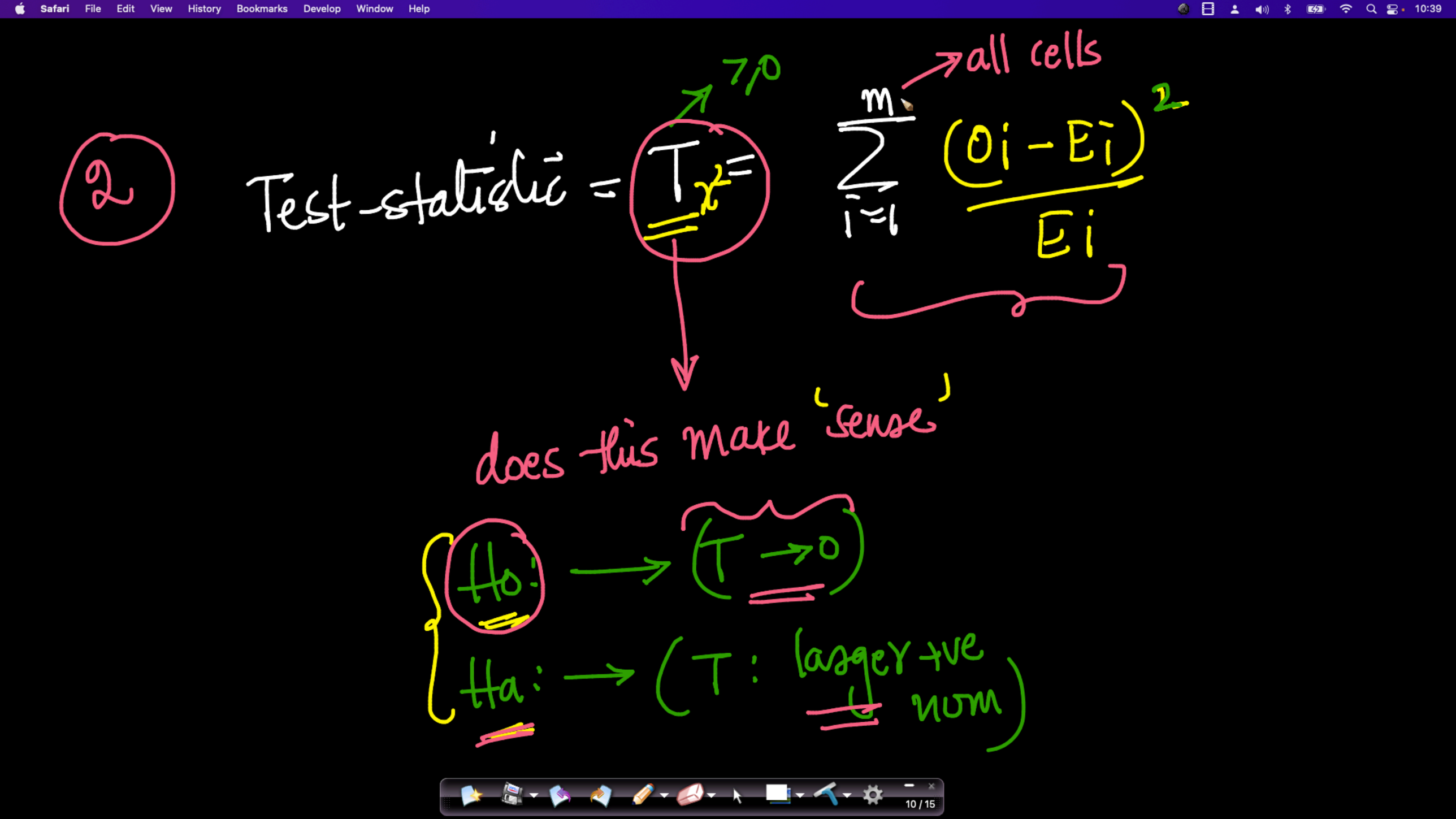
Expected (under Ho)

G \ P	0	1
F	339 (E_1)	155 (E_2)
M	345 (E_3)	158 (E_4)

Observed

G \ P	0	1
F	346 (O_1)	149 (O_2)
M	339 (O_3)	165 (O_4)

999


2
Test-statistic = T
χ²
≥0
m
i=1
all cells
(Oi − Ei)²
Ei
does this make 'sense'
Ho:
T → 0
Ha: → (T: larger +ve num)

Pearson

under H_0

$$T_{\chi^2} \sim \chi^2(k)$$

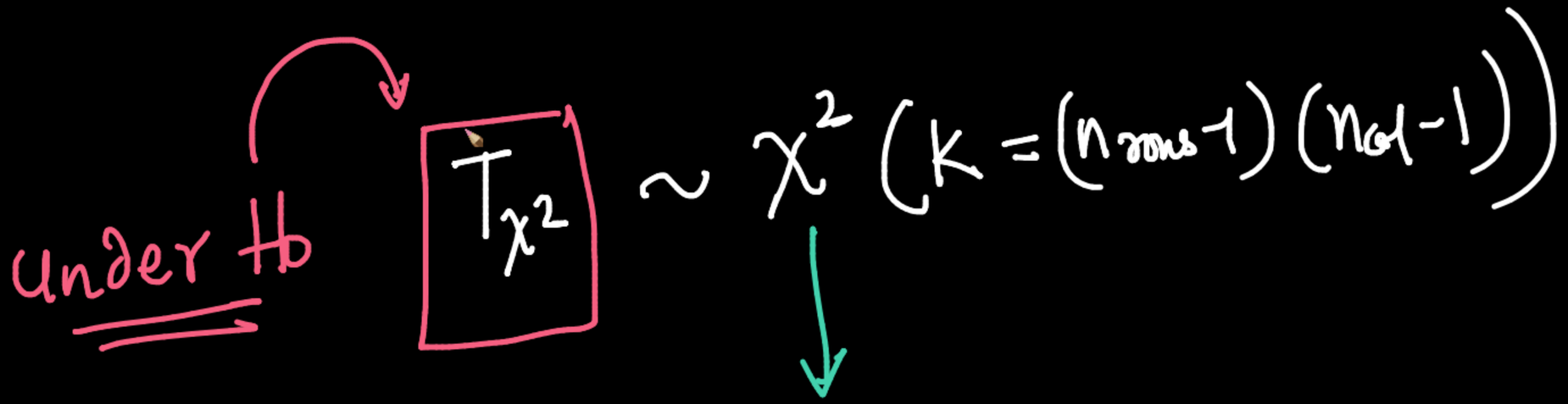
k = dof = param

$$= [(n_{rows} - 1) + (n_{col} - 1)]$$

in the Contingency Table

(+) χ^2 → no assumptions about the population

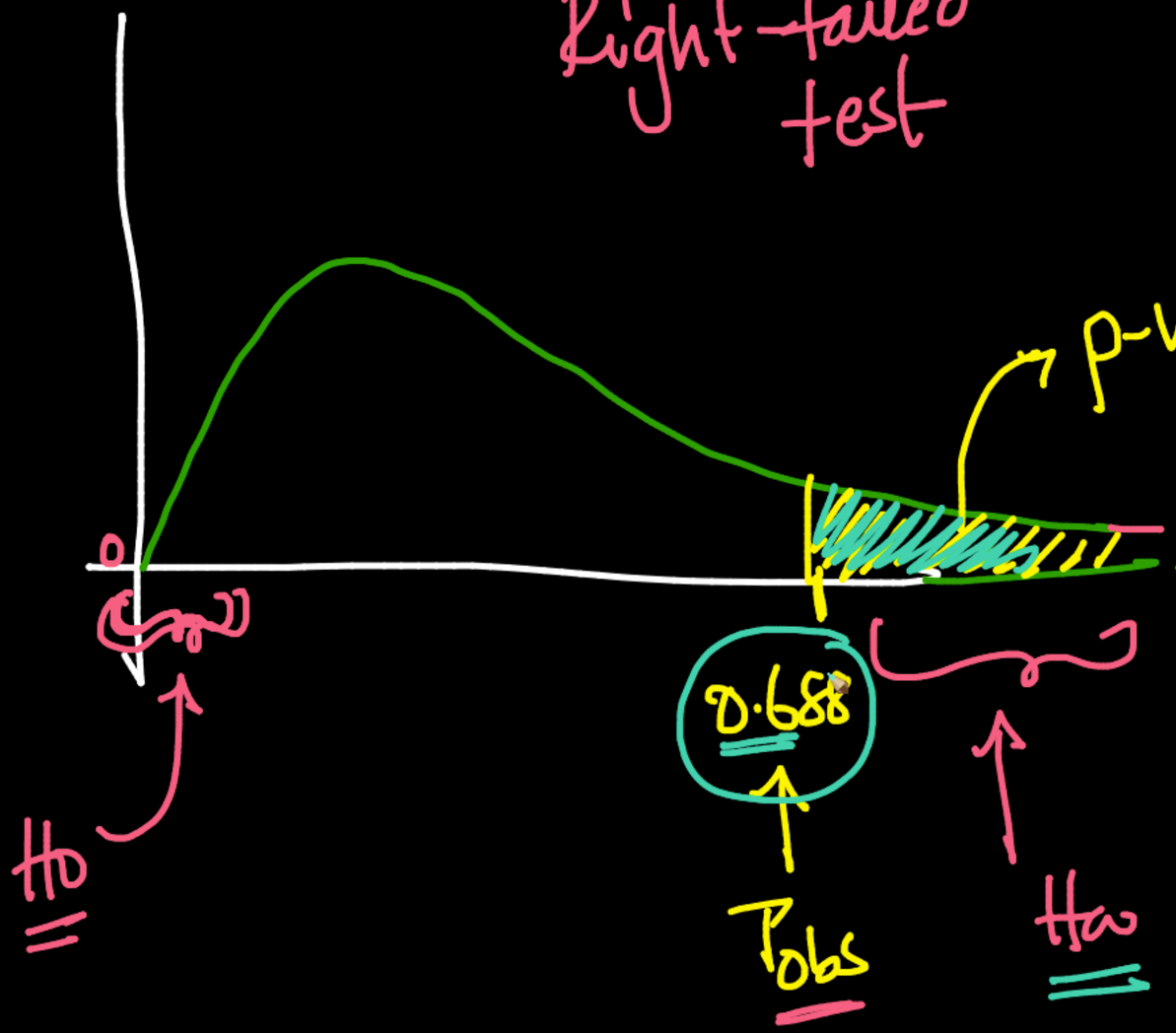
↳ non-param test


under Ho
T_{χ^2}
$\sim \chi^2 \left(k = (n_{rows}-1)(n_{col}-1)\right)$

(3) PDF

Right-tailed test

$T_{obs} = 0.688$ (let)



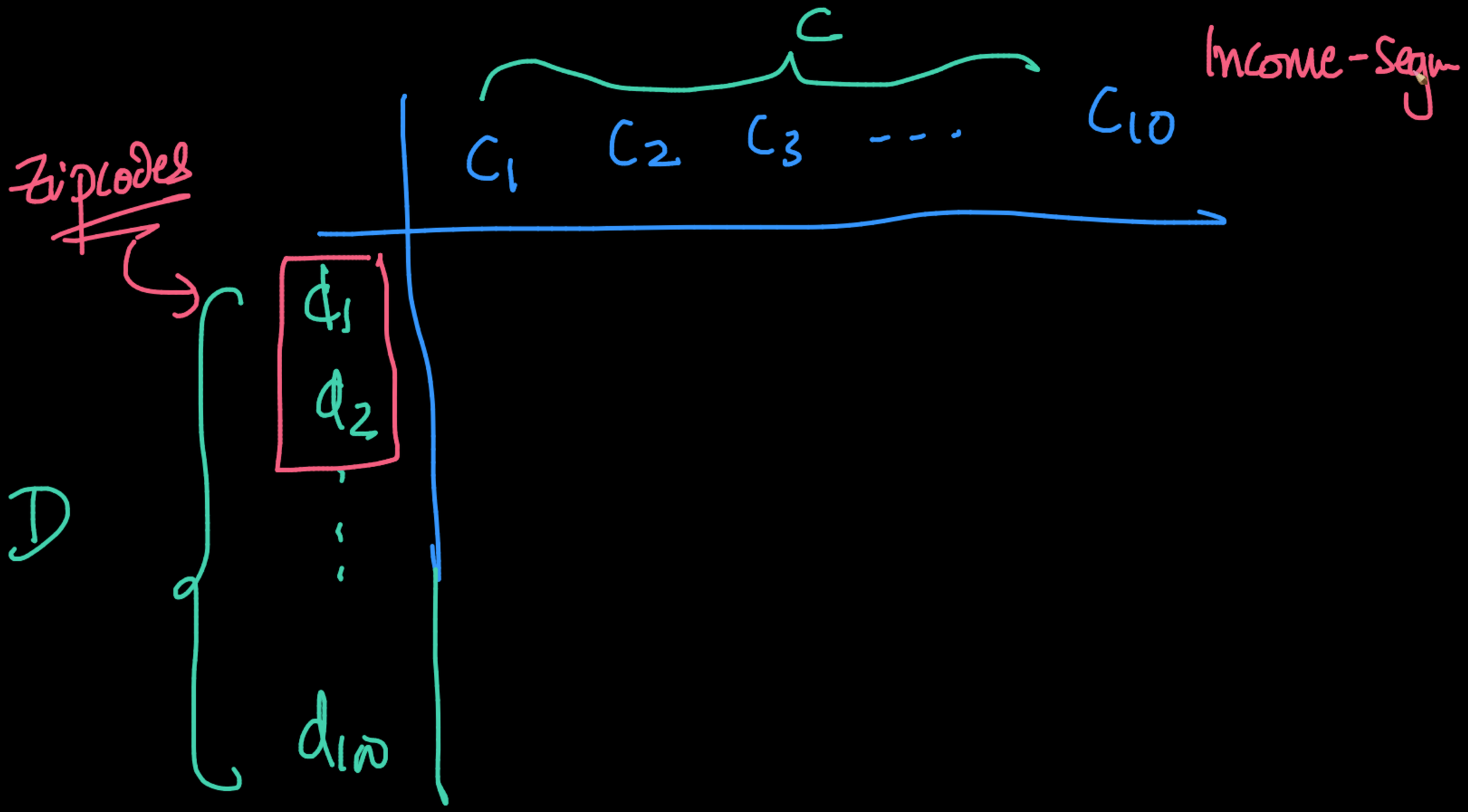
(4) p-val vs $\alpha = 5\%$

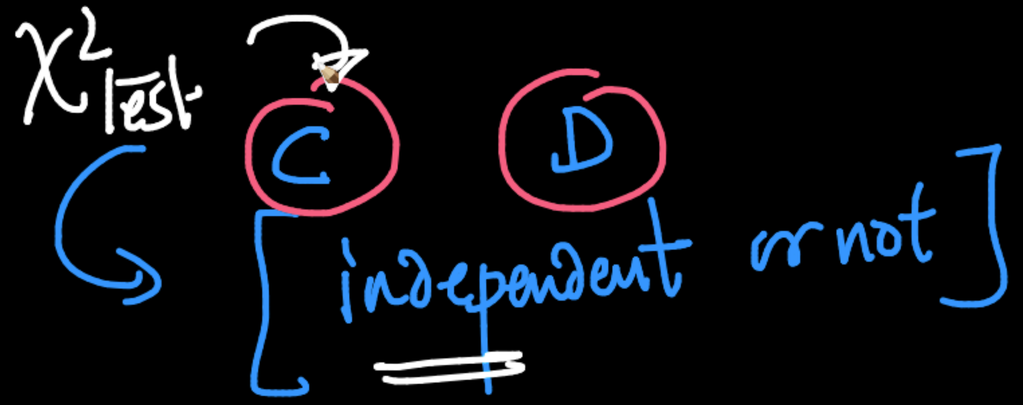
p-val 40.7%

p-val > α ⇒ accept H_0 ✓

p-val = P(observing T as extreme as T_{obs})

IF R/L/2-$\$$ailed ⟶ H_0; H_a / T under H_a


C
Income-Segm
C1
C2
C3
...
C10
Ziplodes
D
d1
d2
d10


χ^2 test
C
D
[independent or not]

SRCC / PCC → fi vs y → numeric

pcc (fi, y) ∀i

f1, f2 ... f10

numeric

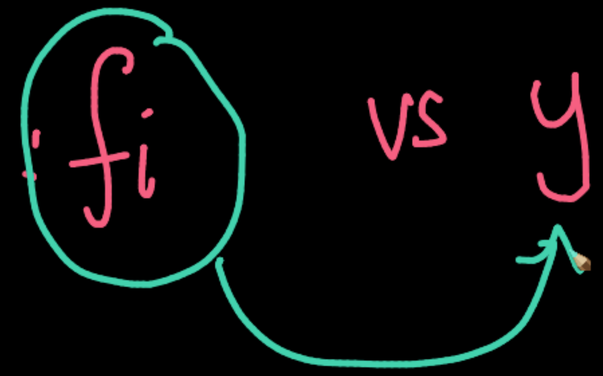
non-linearity (RF/DT)

PCC → numericals

χ^2-test-statistic → Categorical

$$T = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$ } large

Category : f_i vs $y = \{0 \text{ or } 1\}$



Univariate

$f_1, \ldots, f_{10}$ Categorical

$y = \{0, 1\}$

$T_{\chi^2}(f_i, y) = T_i \quad \forall i = 1 \to 10$

largest one → Most imp feature

ANOVA

Z/T-test

[2-sample test]

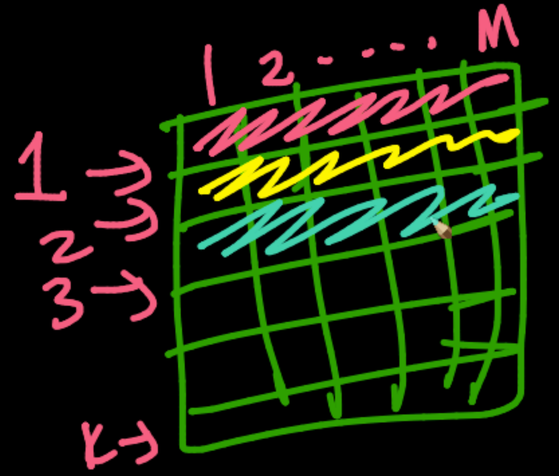
covid

↓

Medicine 1 vs Medicine 2

k-sample Test

- (k-drugs) → group
- m patients per group
- $n = m \times k$



X_{ij} = rec. time for j^{th} patient in group i

$i : 1 \to k$

$j : 1 \to m$

H_0: no-difference in recovery times in group means

$\mu_1 = \mu_2 = \cdots = \mu_k$

H_a: $\exists$ difference "

" "

$\mu_1 \neq \mu_2 \neq \mu_3 \cdots \neq \mu_k$

$\mu_1 \neq \mu_2 = \mu_3 \cdots$

2-sided ✗

Safari File Edit View History Bookmarks Develop Window Help 10:57

Assumptions:

- each group's rec-times → population are gaussian
- each gp's variance is same / similar
- all observations are indep of one another

$k = 10$

$M = 100$

1 →

4

Test-Statistic

$$f = \frac{MSB}{MSW}$$

MSB → mean-sum-squares between gps

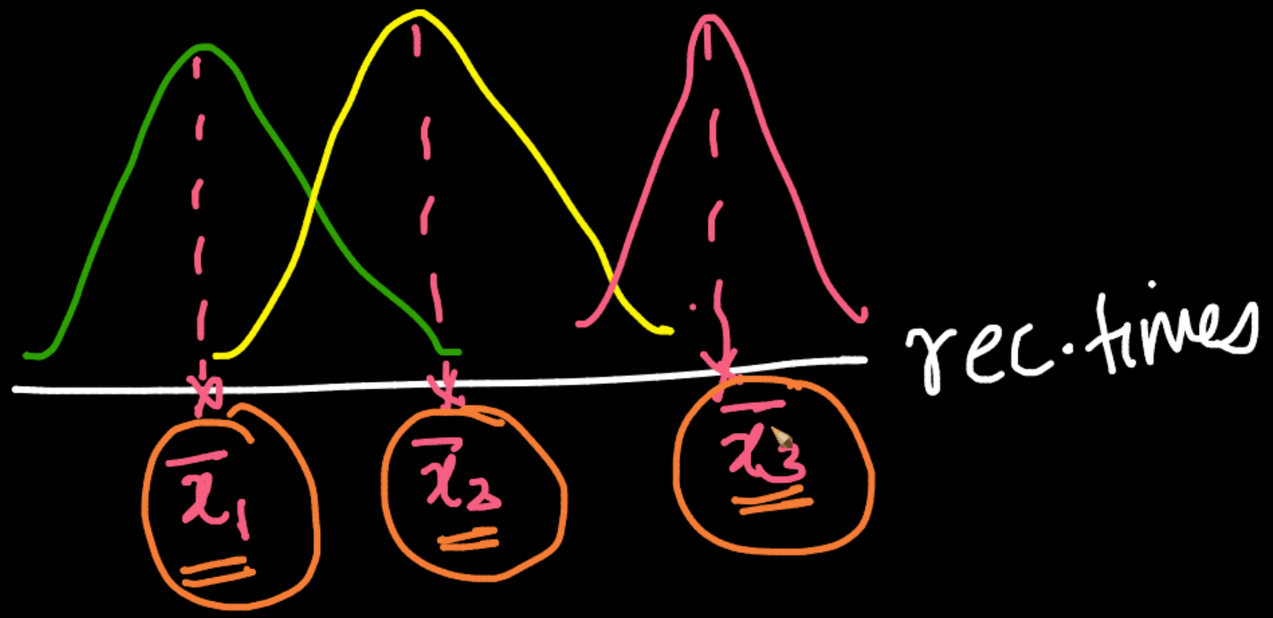
MSW → mean-ss-within gps

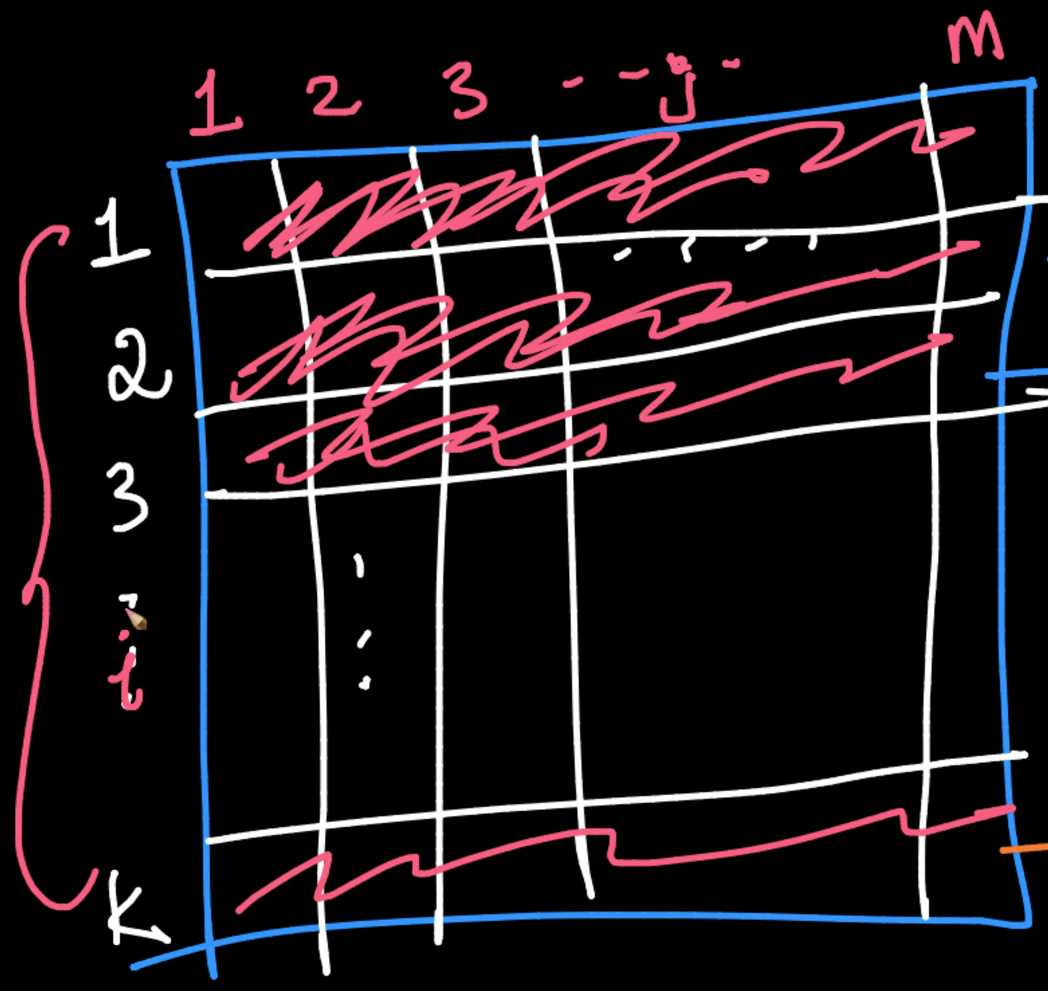
Under H_0

$$\sim F\text{-dist}(k-1, n-k)$$

$n-k$ → $m \times k$

$f = \frac{MSB}{MSW}$ → Mean of $(\bar{x}_i - \bar{x})^2$ $i : 1 \to k$





X_{ij} :

$\bar{x}_1$: mean rec. time of d_1

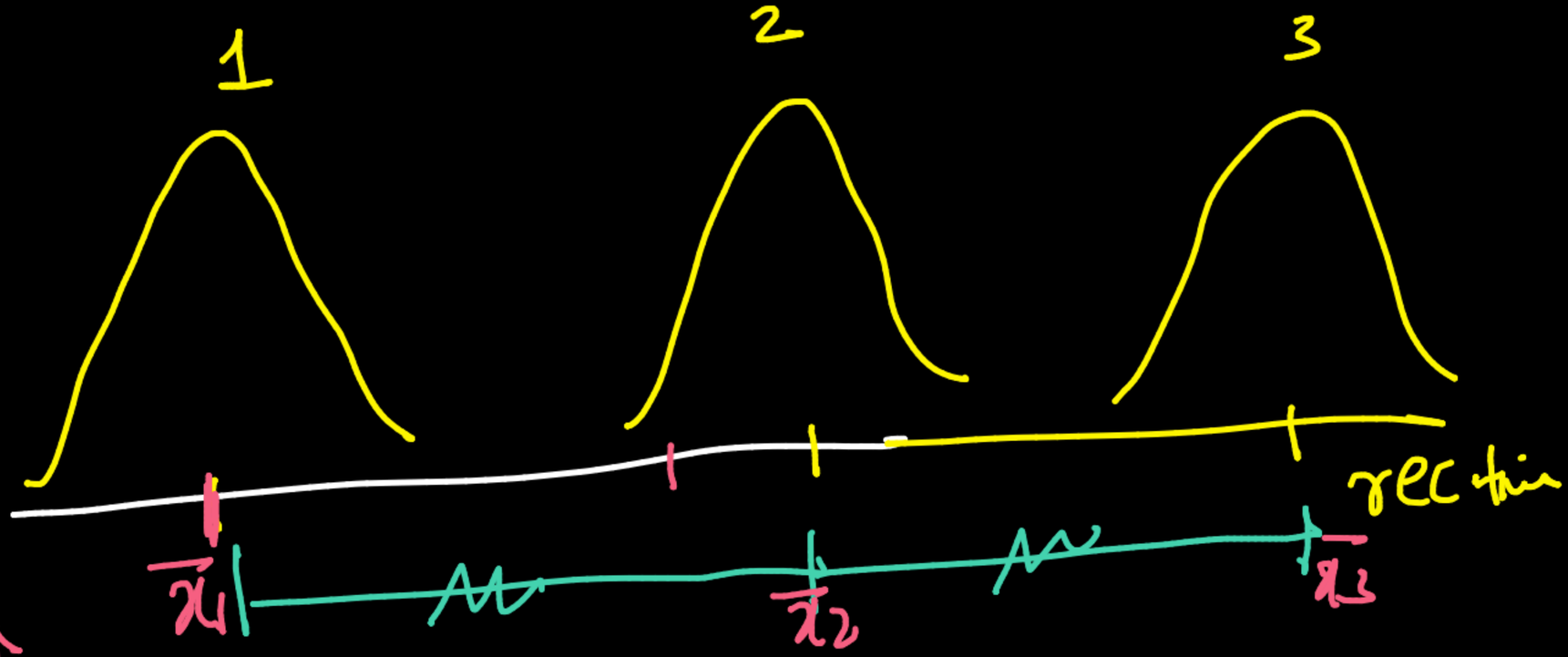
$\bar{x}_2$

$\bar{x}_3$

$\bar{x}_k$

$\bar{x}$: mean of overall all sections

$n = m \times k$

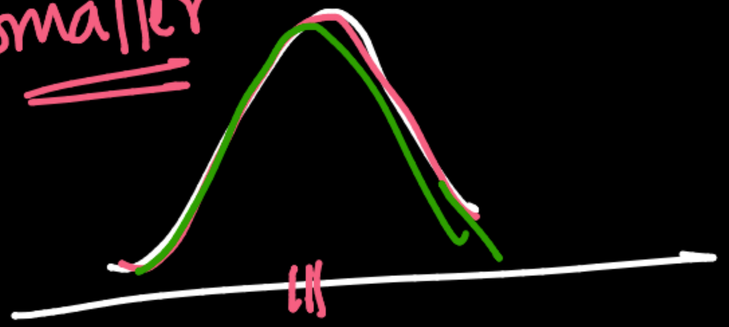


$\uparrow f = \frac{MSB\uparrow}{MSW}$

Variance of gp-Means

{ Under H_0: $f \downarrow$ → smaller

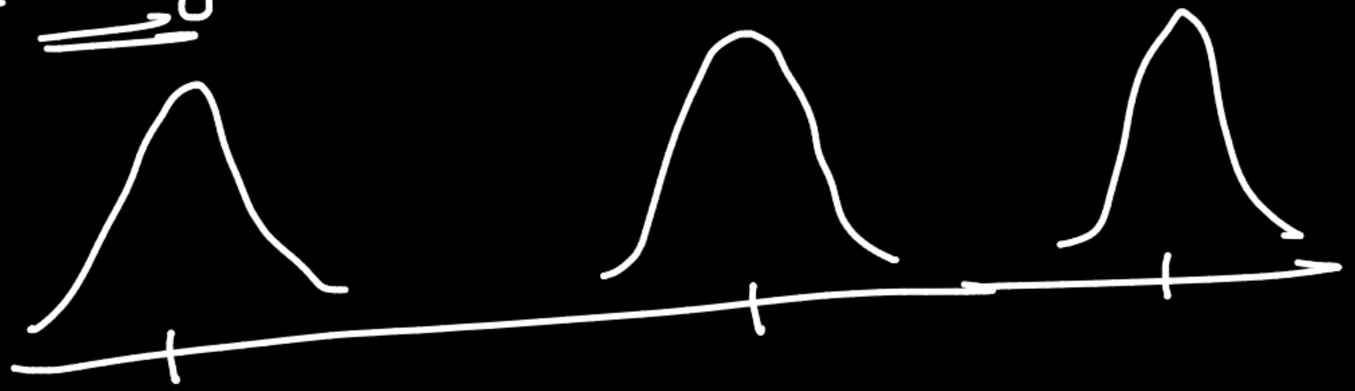
$\mu_1 = \mu_2 = \mu_3$

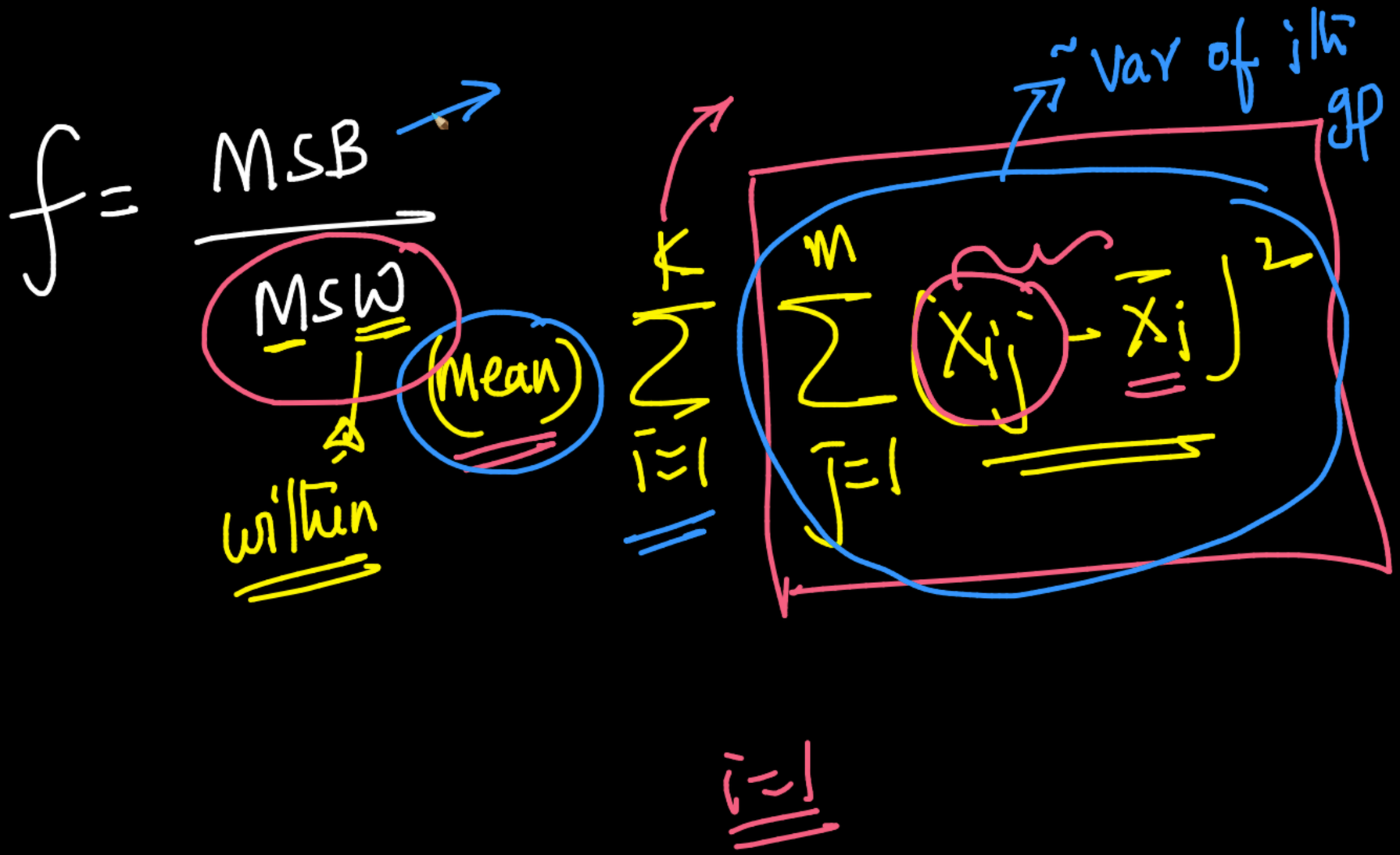


$\downarrow f = \frac{MSB\downarrow}{MSW}$

{ Under H_a: → f is large

$\uparrow f = \frac{MSB\uparrow}{MSW}$



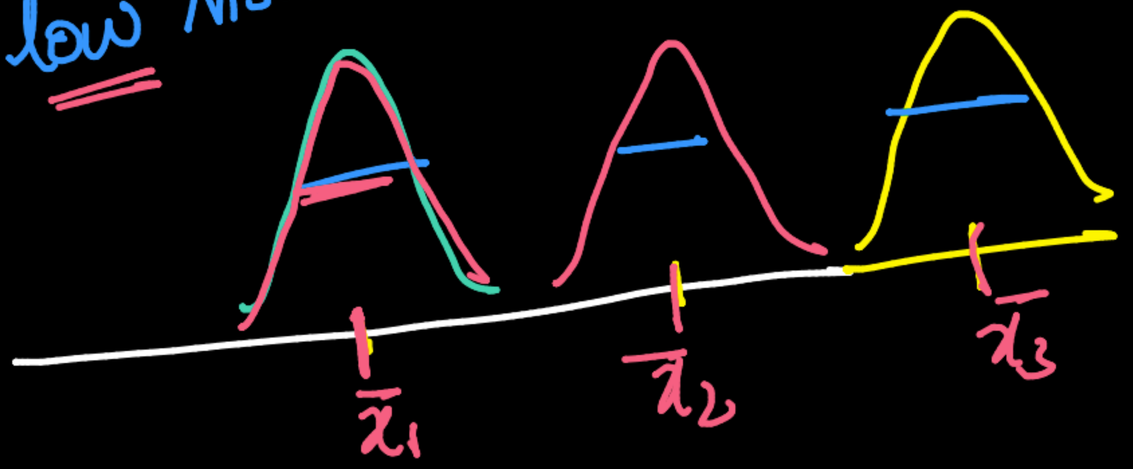


$f = \frac{MSB}{MSW}$ → constant

Case 1:

H_0

low MSW

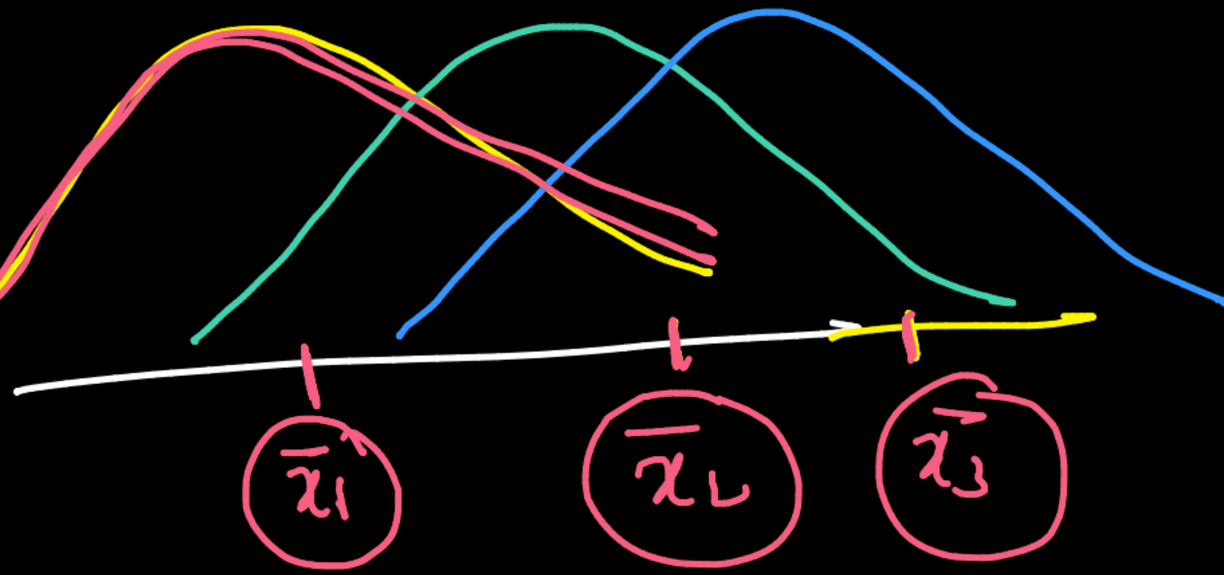


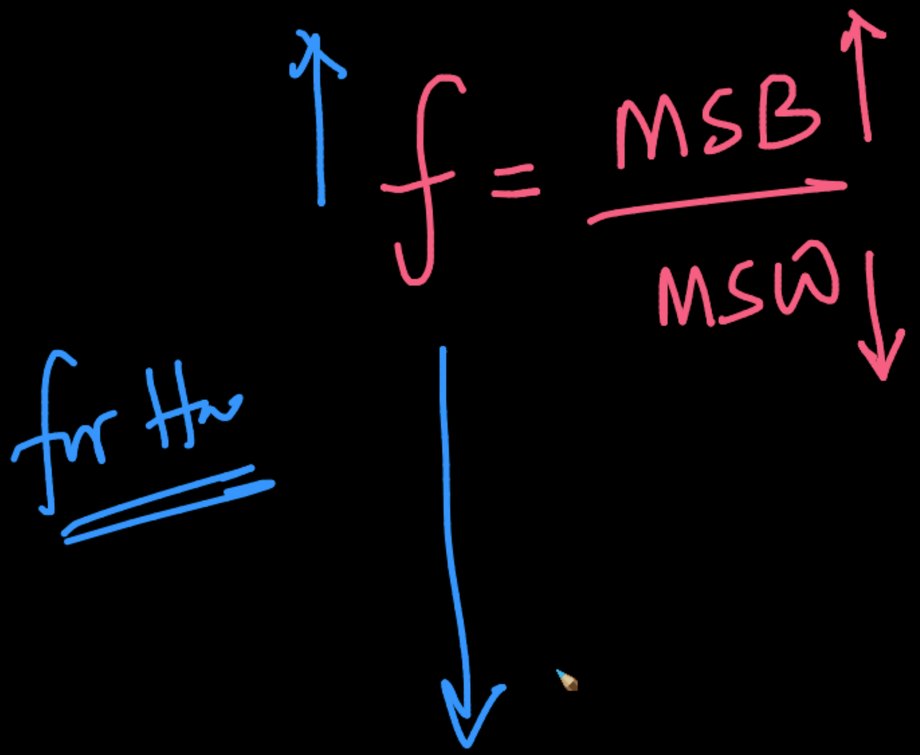
m-obs

Case 2:

H_0 ✓

high MSW



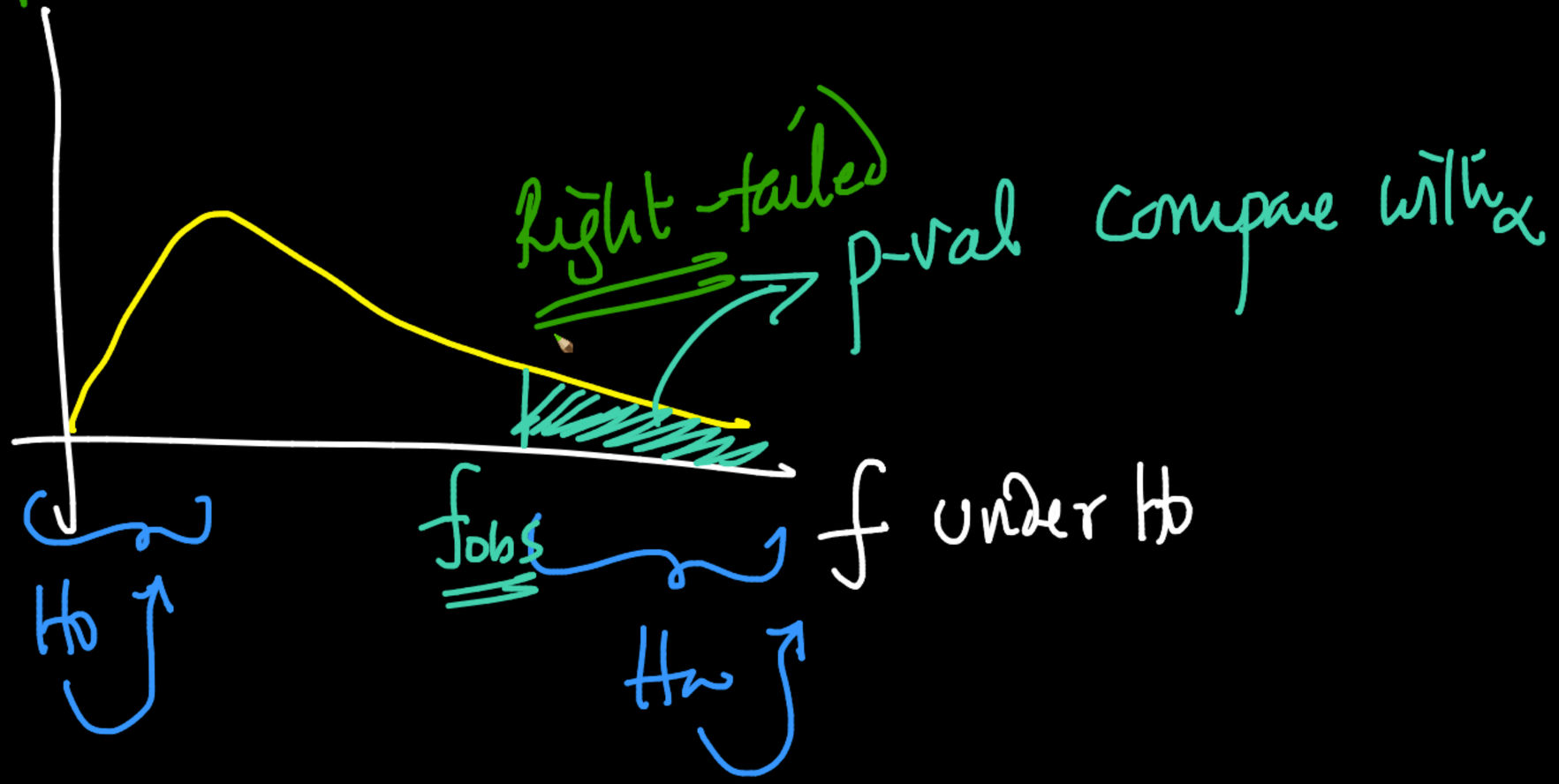

f = MSB ↑ / MSω ↓
for Hw

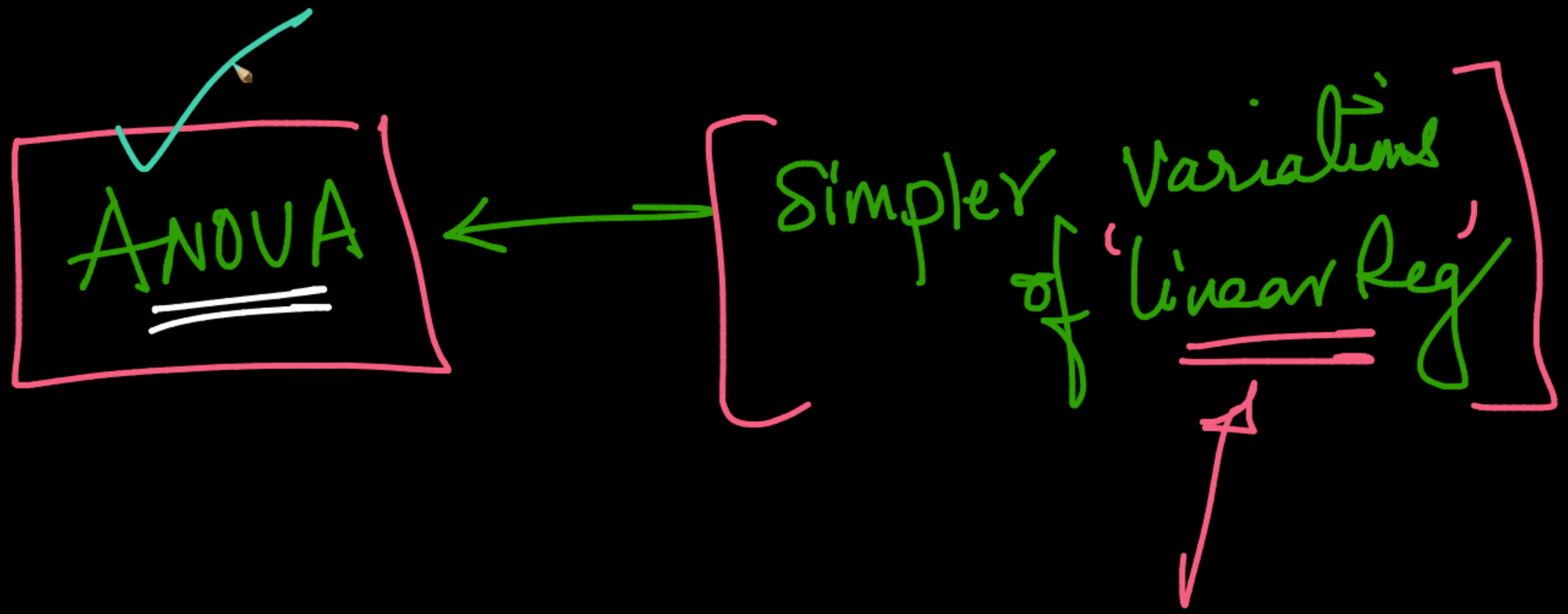
f under $H_0 \sim f\_disb(x-1, n-x)$

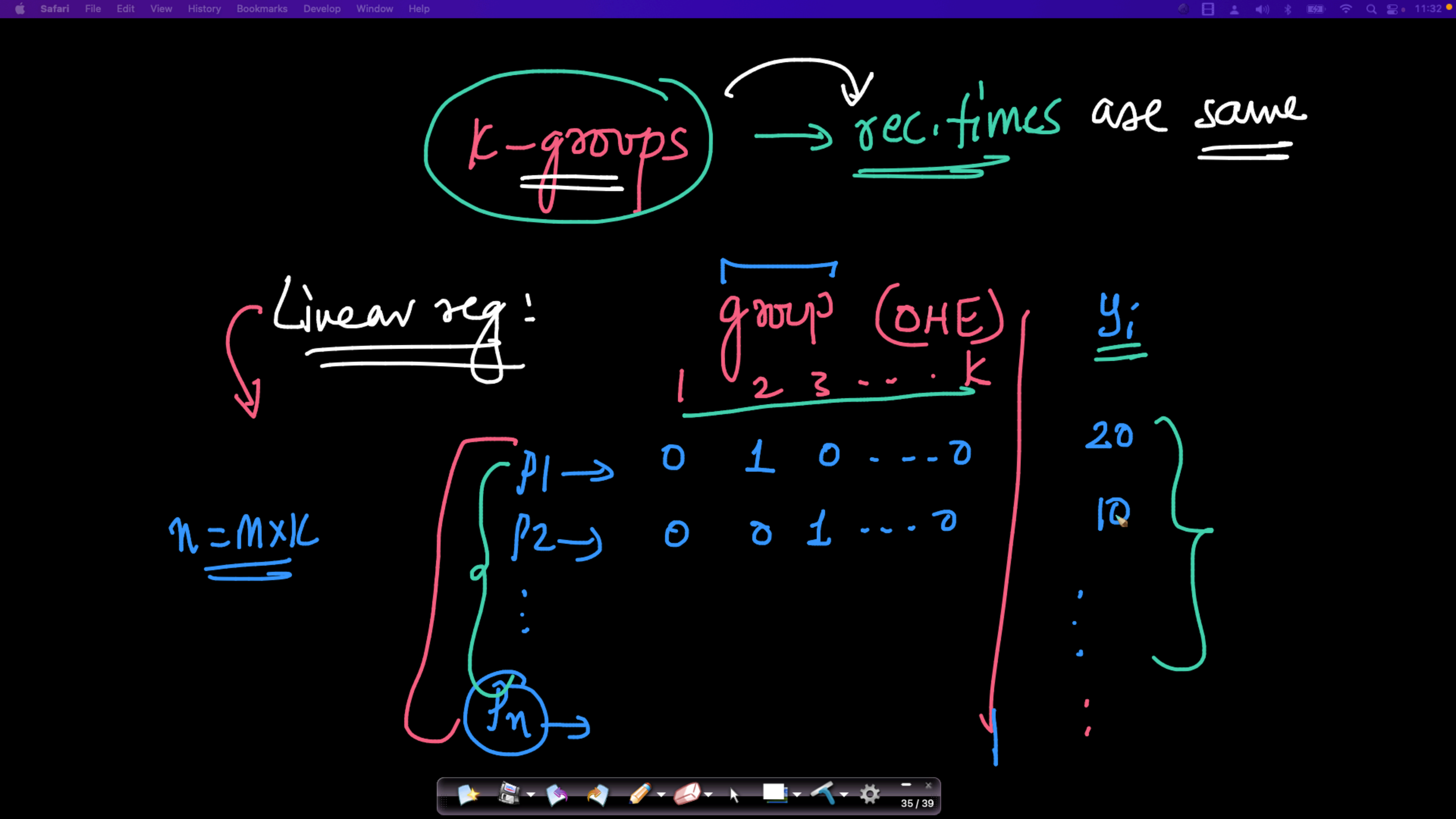
$\{$ H_0: all means are same
H_a: " " not same

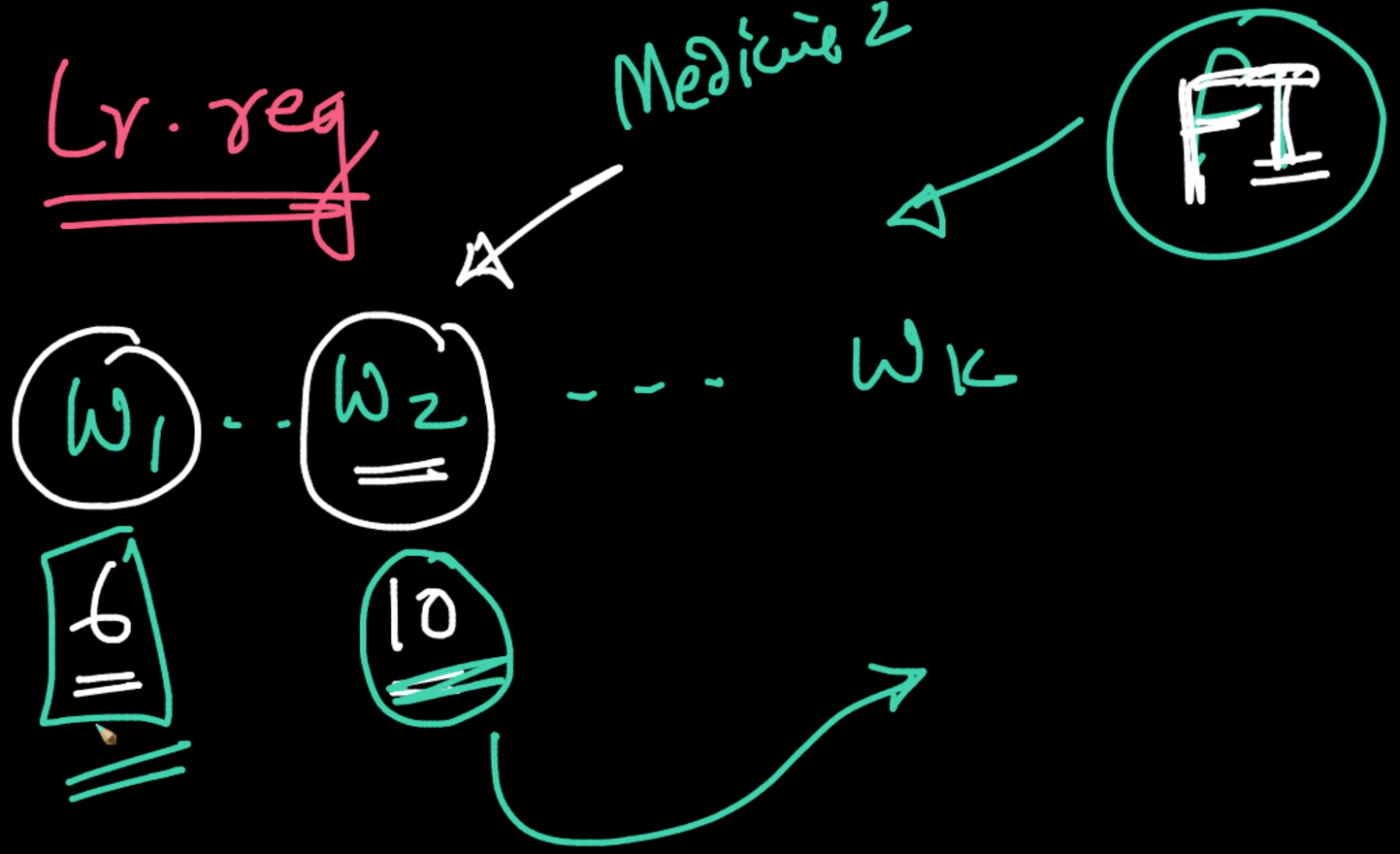
$k-1$ $n-k$

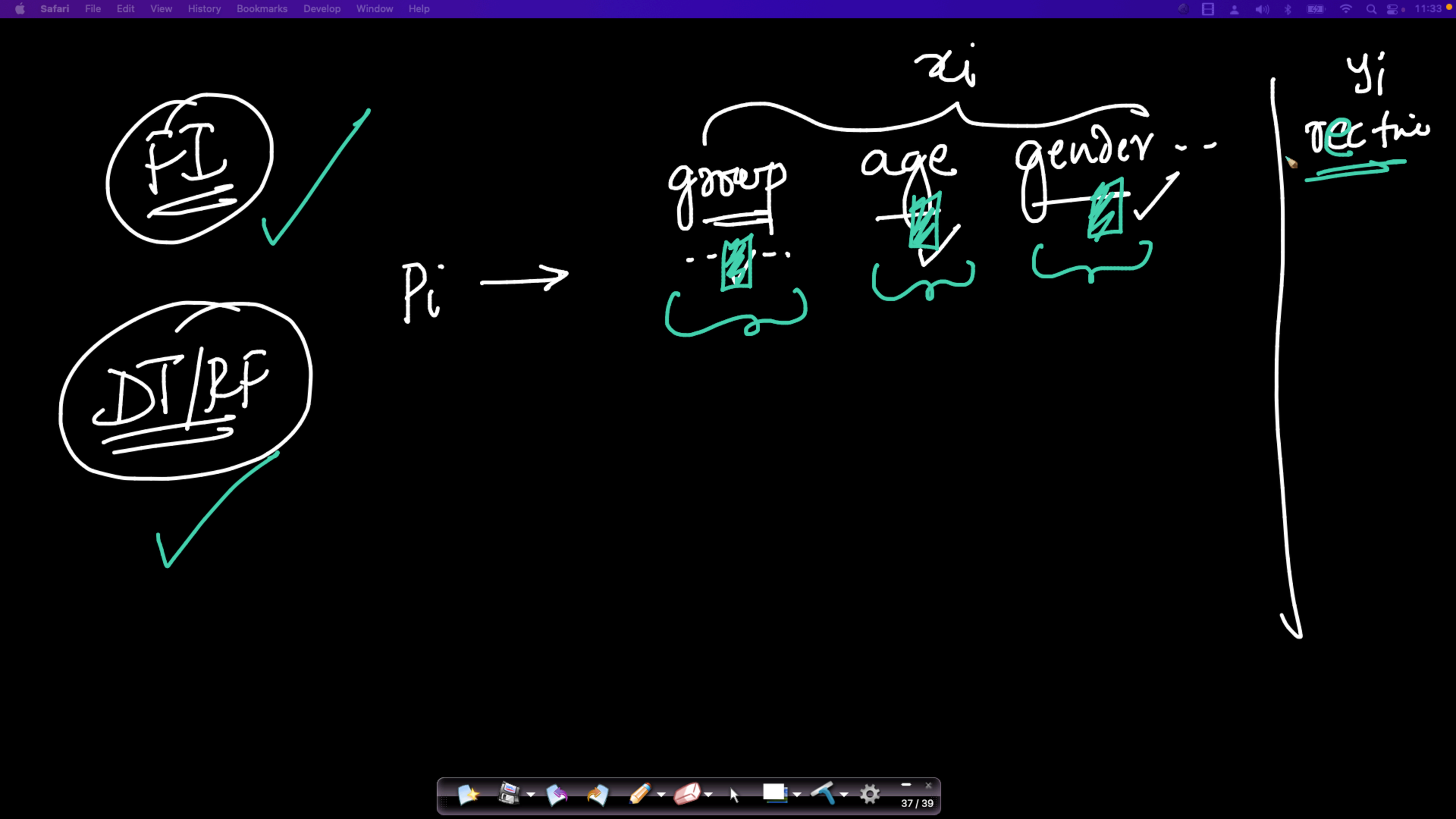
$n = M \times k$

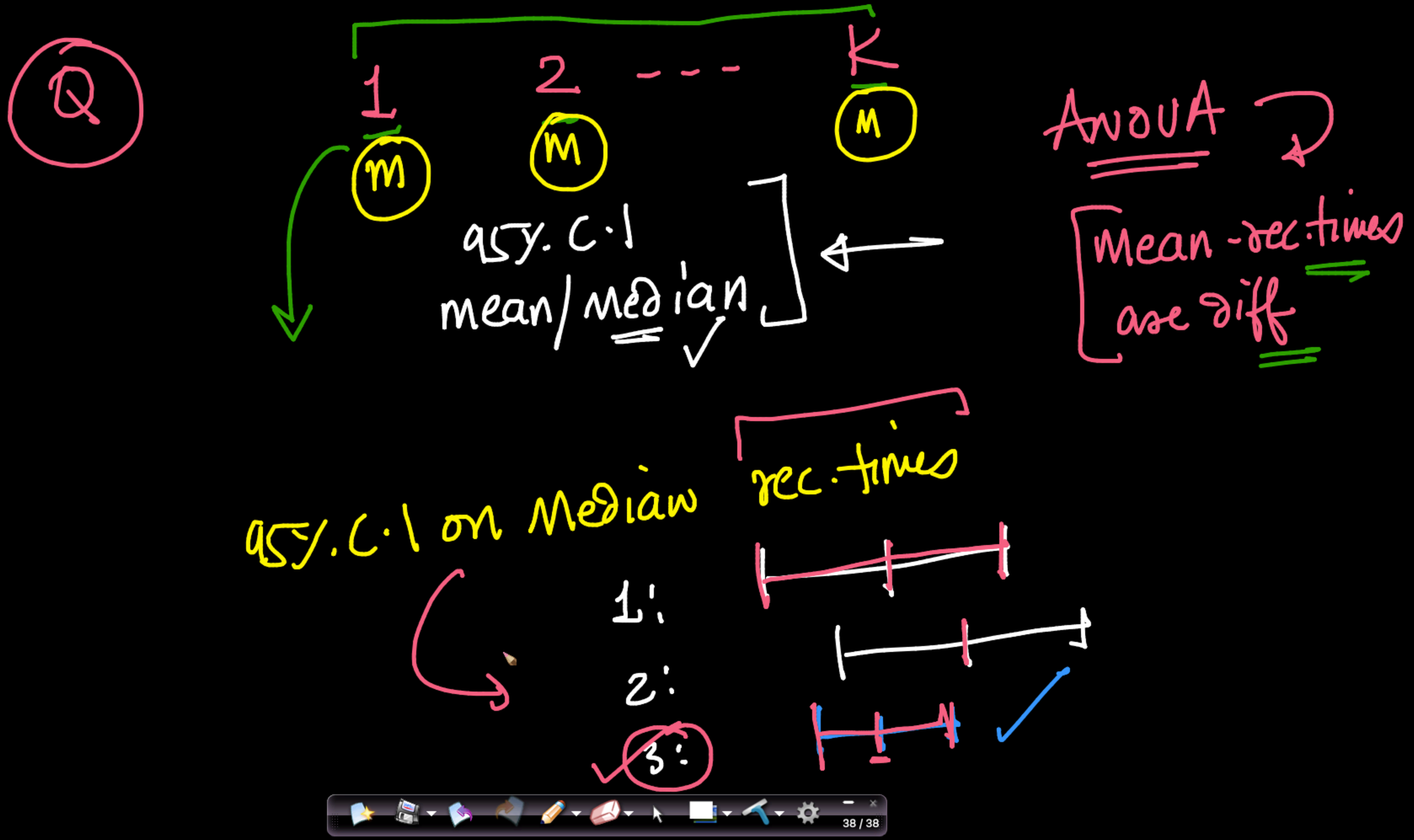



ANOVA
Simpler Variations of 'linear Reg'

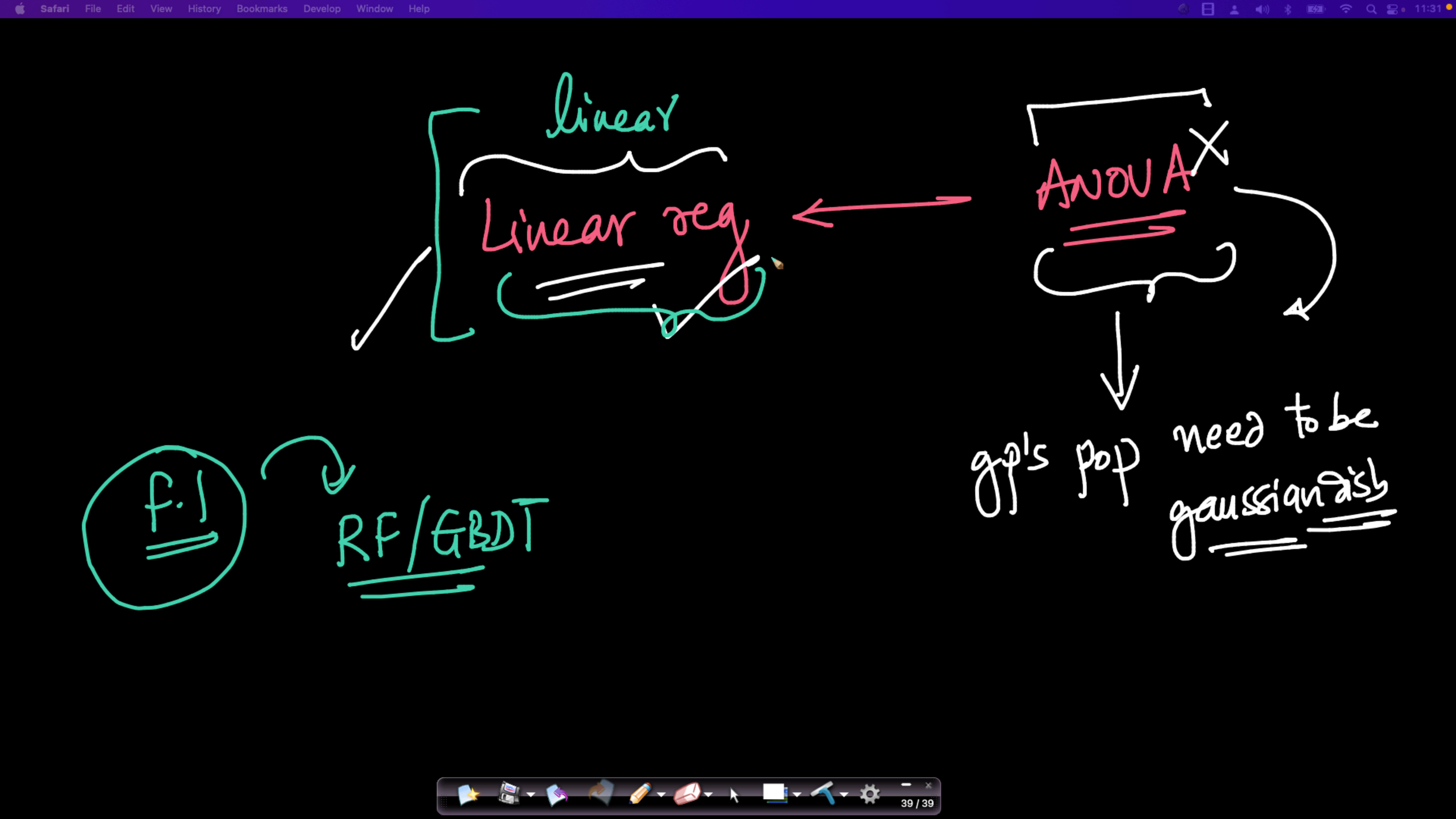

K-groups
rec. times are same
Linear reg :
group (OHE)
1 2 3 . . . k
yi
P1 → 0 1 0 . . . 0
P2 → 0 0 1 . . . 0
Pn →
n = M x K
20
10

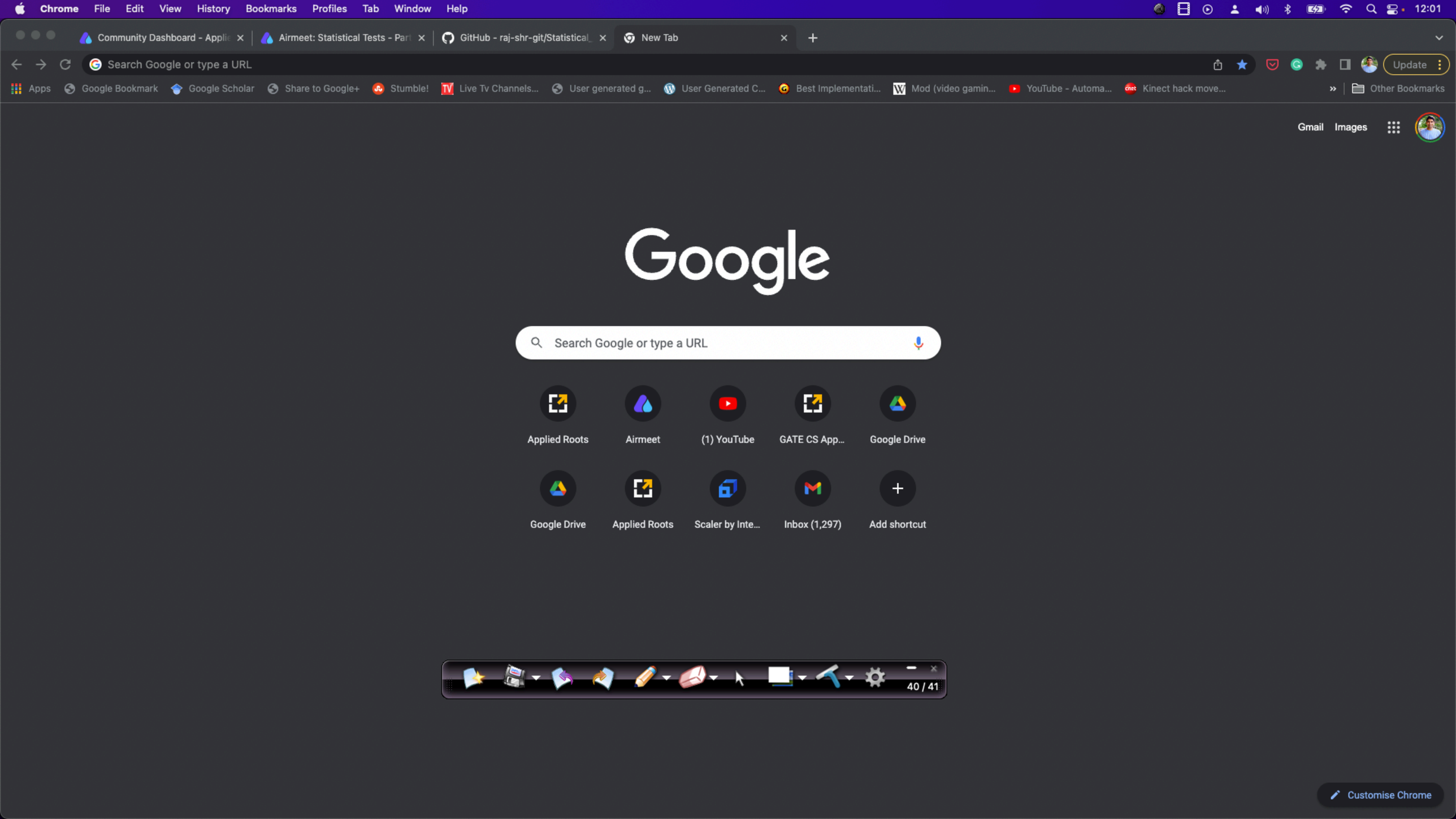

Lr. reg
Medicine 2
FI
w_1
w_2
w_k
6
10


FI
DT/RF
Pi
xi
group
age
gender
yi


Q
1
2

K
M
M
M
95% C.I
mean/Median
ANOVA
Mean - rec.times are diff
95% C.I on Median
rec.times
1:
2:
3:


linear
Linear reg
ANOVA
gp's pop need to be gaussian ish
f.1
RF/GBDT


Chrome File Edit View History Bookmarks Profiles Tab Window Help
12:01
Community Dashboard - Applic
Airmeet: Statistical Tests - Part
GitHub - raj-shr-git/Statistical_
New Tab
Search Google or type a URL
Update
Apps
Google Bookmark
Google Scholar
Share to Google+
Stumble!
Live Tv Channels...
User generated g...
User Generated C...
Best Implementati...
Mod (video gamin...
YouTube - Automa...
Kinect hack move...
Other Bookmarks
Gmail
Images
Google
Search Google or type a URL
Applied Roots
Airmeet
(1) YouTube
GATE CS App...
Google Drive
Google Drive
Applied Roots
Scaler by Inte...
Inbox (1,297)
Add shortcut
40 / 41
Customise Chrome