✓ - C.I of metrics — Non Gaussian to Gaussian — Param vs Non param

Today

Remedial Session: Statistics (3 more topics)

Regular Session: Linear Algebra & Co-ordinate Geometry

Geometric / intuitive

Application

ML concepts

Next Sessions (2 more)

Ops:

each section → chat (Q&A)

general / broad / out of the context → Q&A (30 mins)

Parametric vs Non-parametric tests

Parametric

- makes assumptions about the 'distb' of the data
- F-statistic
- ANOVA: each groups observations are sampled from a Gaussian distb

Non-parametric tests

- χ^2-test
- any distb. assumptions about the data?

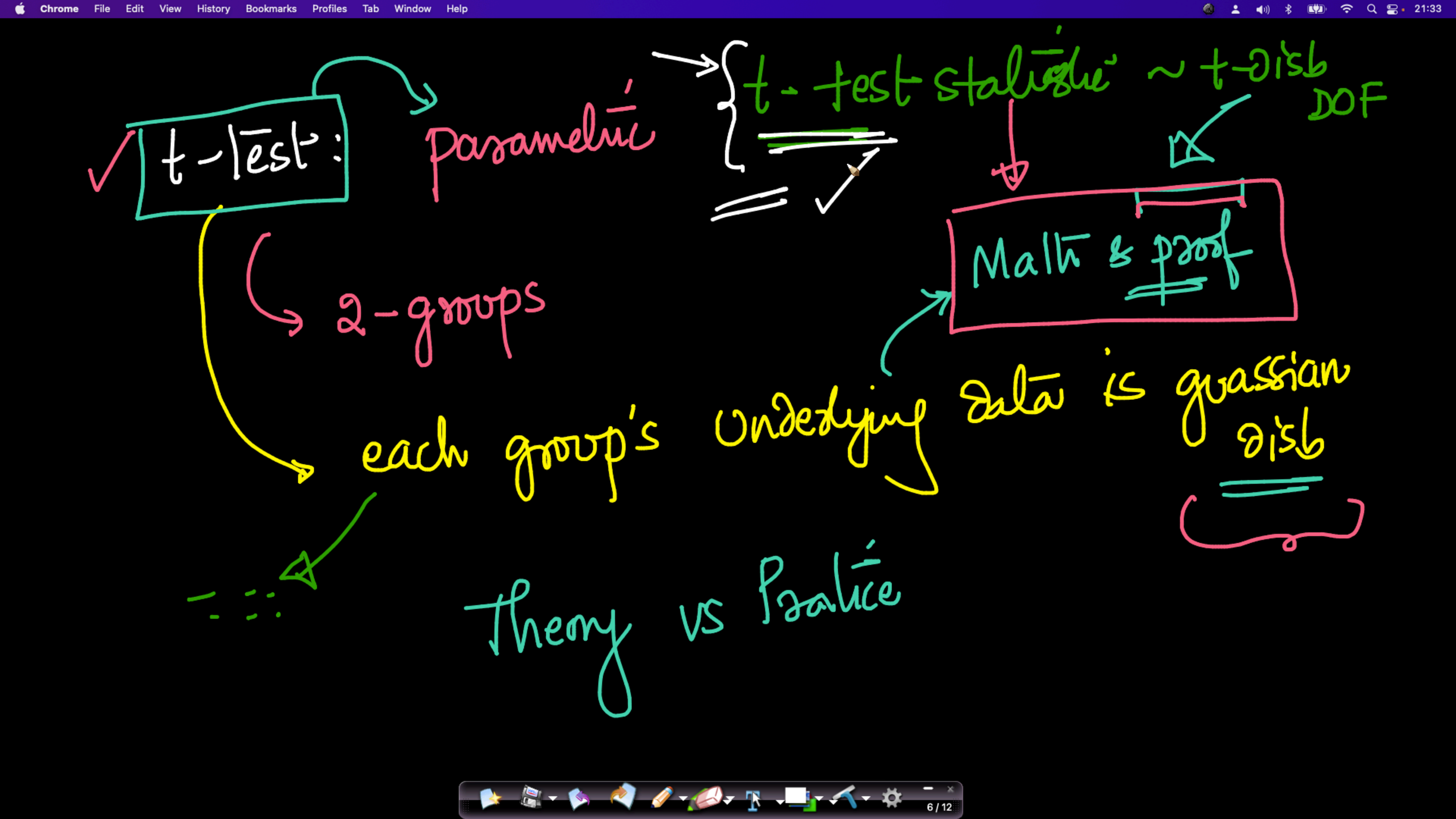
(Q) χ^2-test → No assumptions about underlying data

Common-Confusion

χ^2 test statistic $= \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ ✓

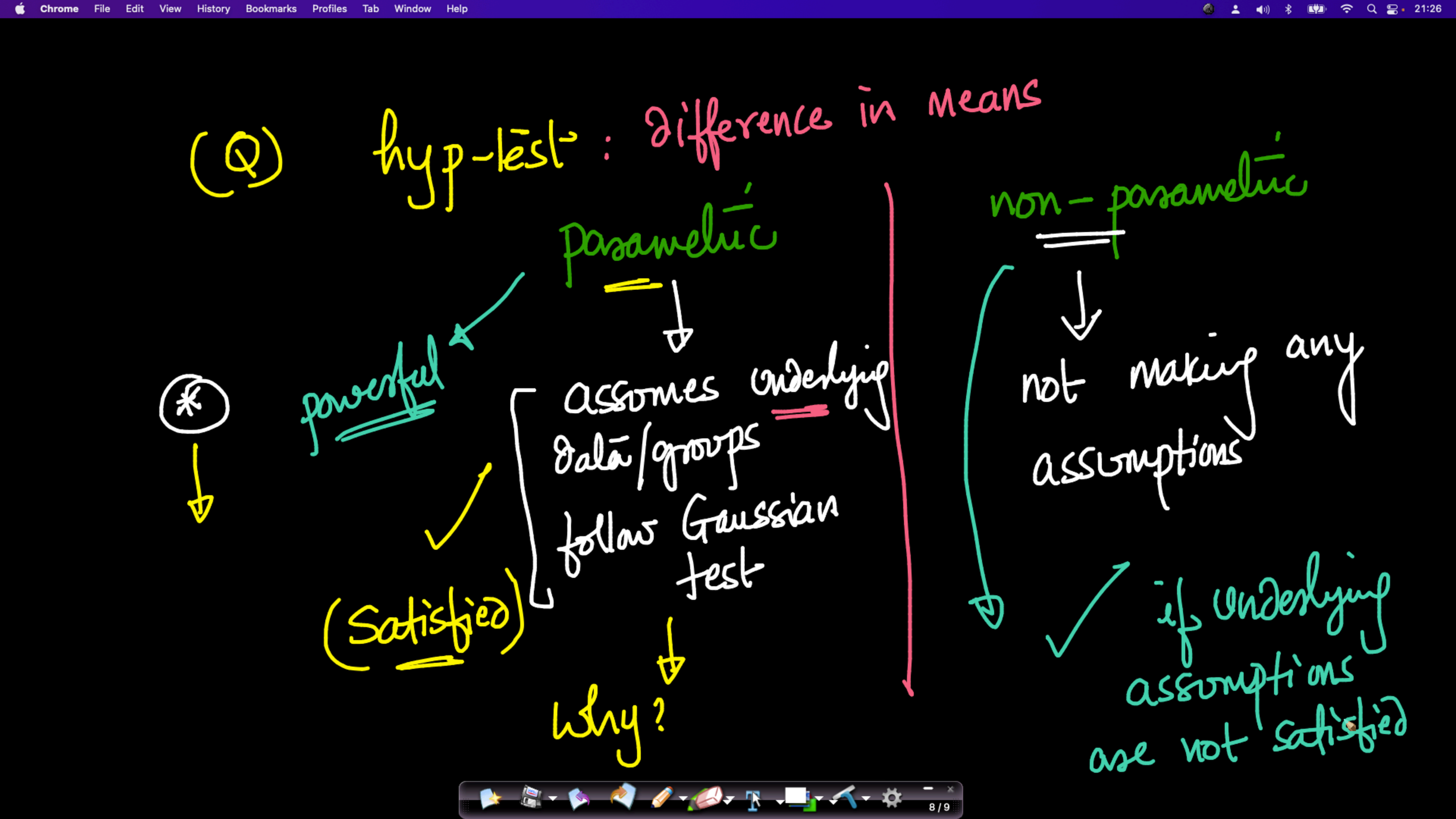
χ^2-distribution → p-value ✓

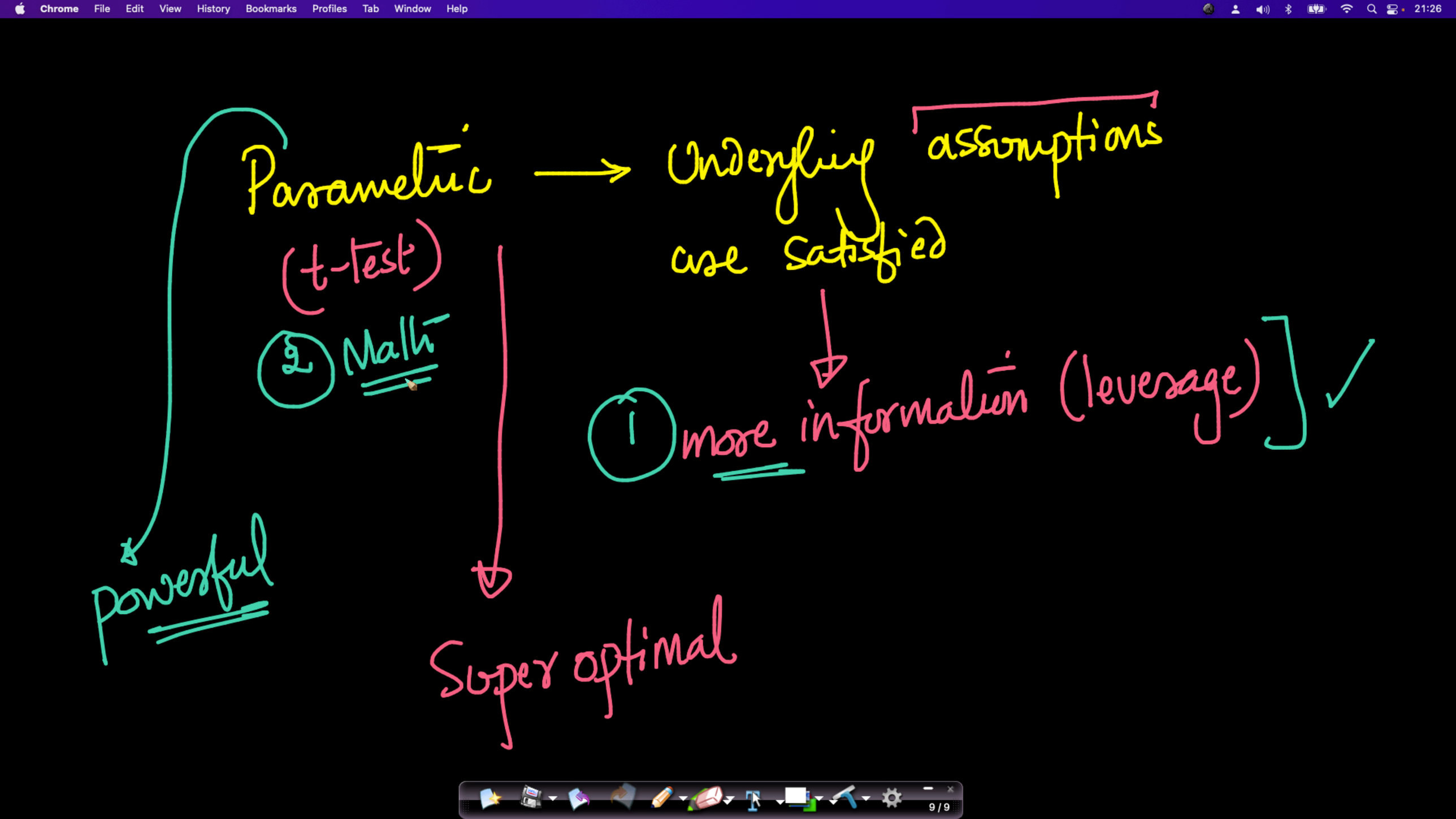
→ why is it non-parametric


t-test:
Parametric
t-test statistic ~ t-dist
DOF
2-groups
Math & proof
each group's Underlying data is gaussian dist
Theory vs Practice

✓ test.
t-statistic ~ t-dist^b

[underlying data ~ Gaussian (ideal-case)


(Q) hyp-test : difference in means
parametric
non-parametric
powerful
assumes underlying data/groups follow Gaussian test
(satisfied)
why?
not making any assumptions
if underlying assumptions are not satisfied
8 / 9


Parametric
(t-test)
2 Multi
Underlying assumptions
case satisfied
1 more information (leverage)
Powerful
Super optimal

(Q) 2-Sample z-test : parametric or non-param

- t-test → fewer sample size ; unknown σ
- z-test → larger sample size ; known σ

each groups underlying dist is Gaussian ✓

{ 100's of hyp. testing

↓ appropriate in the context ; param or non-param

→ 6-step framework

Underlying data

CLT: Sample-means ~ Gaussian dist ✓

Finite μ, σ

t-test → means of samples

Underlying data ≠ Gaussian

Can't prove

t-test statistic ~ t-dist DOF

Underlying data

Uniform-dist

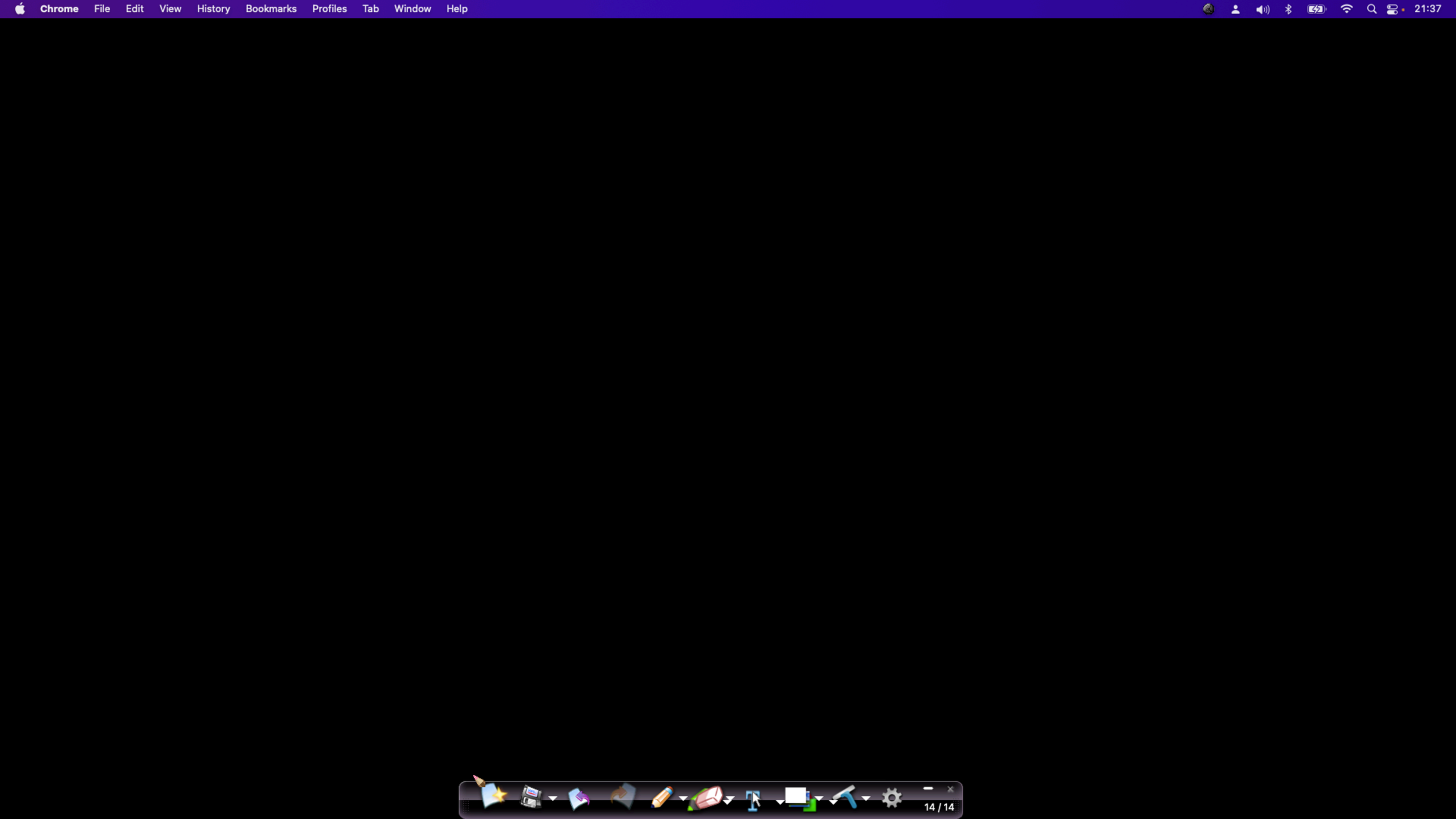
(finite μ, σ)

Sample-means ~ Gaussian

(CLT)

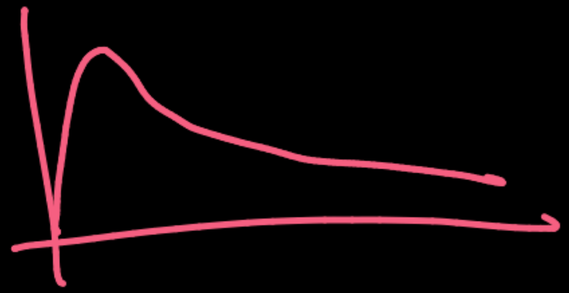
NOT converging

non-Gaussian to Gaussian


Chrome File Edit View History Bookmarks Profiles Tab Window Help
21:37
14 / 14

Non-Gaussian ⟶ Gaussian

$X \rightarrow \{x_1, x_2, \ldots\}$ observations



Q

e.g.: log-normal ⟶

$X \sim$ log-normal

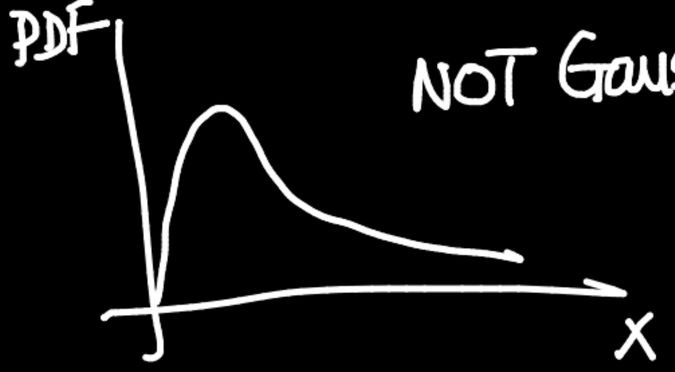
$\log_e(X) = \ln(X) \sim$ Gaussian-dist

Obs

$X \Rightarrow \{ x_1, x_2, x_3, \ldots x_n \}$

PDF

NOT Gaussian

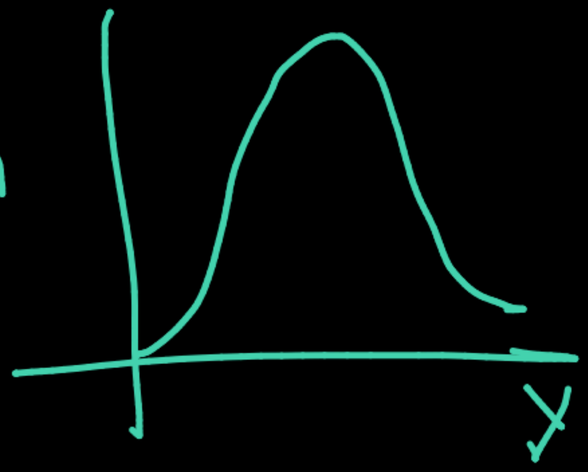


one transforming

$\ln(x)$: log transform

$Y = \{ \ln(x_1), \ln(x_2), \ln(x_3) \ldots \ln(x_n) \}$

$y_1 \quad y_2 \quad y_3 \quad \ldots \quad y_n$



why ?

analogy: cm → feet
°C → °F

10, 10

↳ Non-Gaussian (log-normal)

transforming → log

Gaussian

t-test

tampering

↳ X

123.6	→ 121.6
29.3	→ 213.45

Conclusions

hyp-testing: $H_0: M_1 = M_2$
$H_a: M_1 \neq M_2$

$$y = \ln(x)$$

$$x = e^{y} \Rightarrow \exp(y)$$

✓ Box-Cox :

X $\{x_1, x_2, \dots x_n\}$

non-Gaussian

box-cox (scipy)

λ

optimization

exp

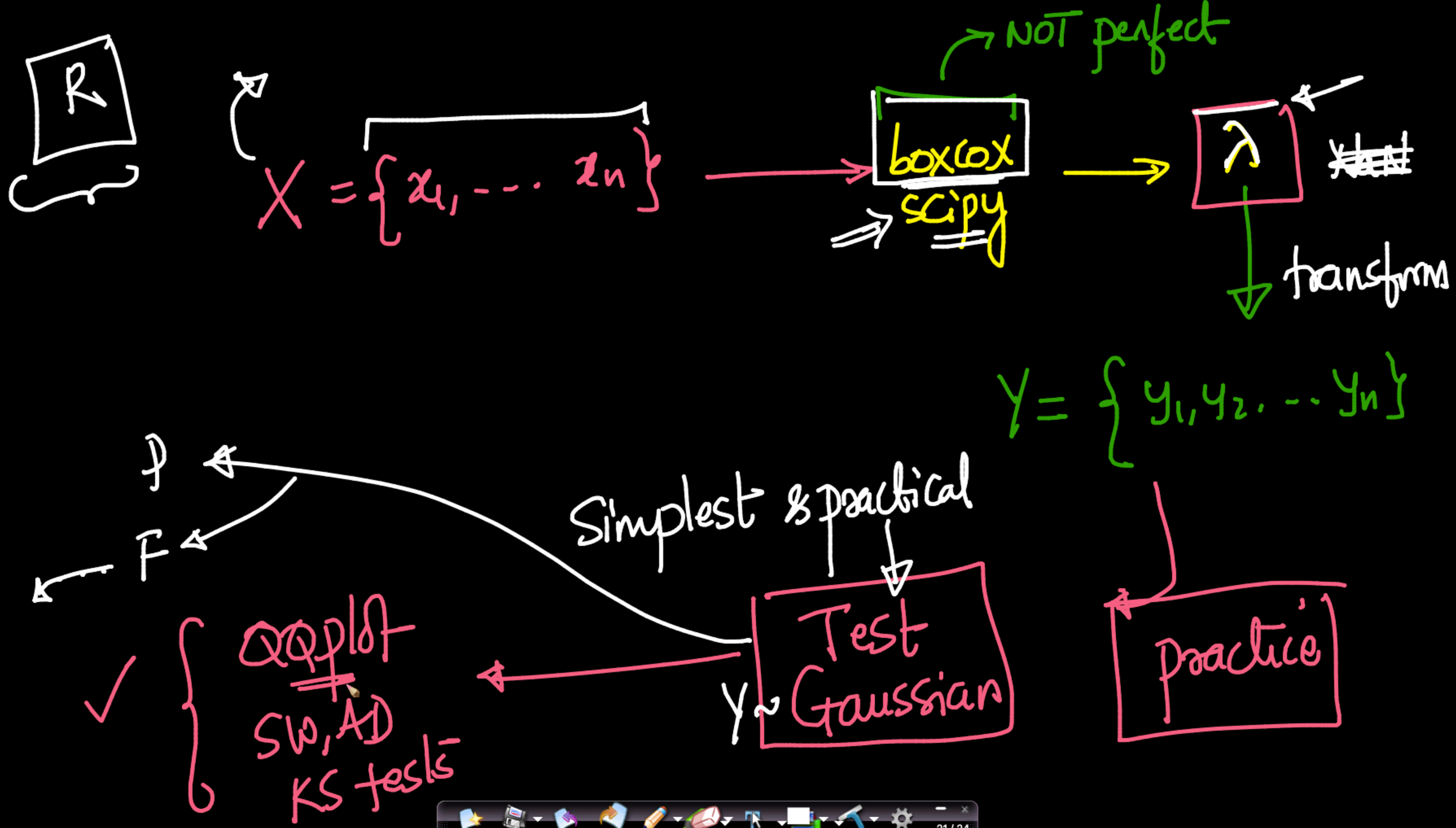
$$\text{Gaussian} \sim Y_i = \begin{cases} \dfrac{X_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(X) & \text{if } \lambda = 0 \end{cases}$$

log

generalization of log-normal

any data → Gaussian [No]

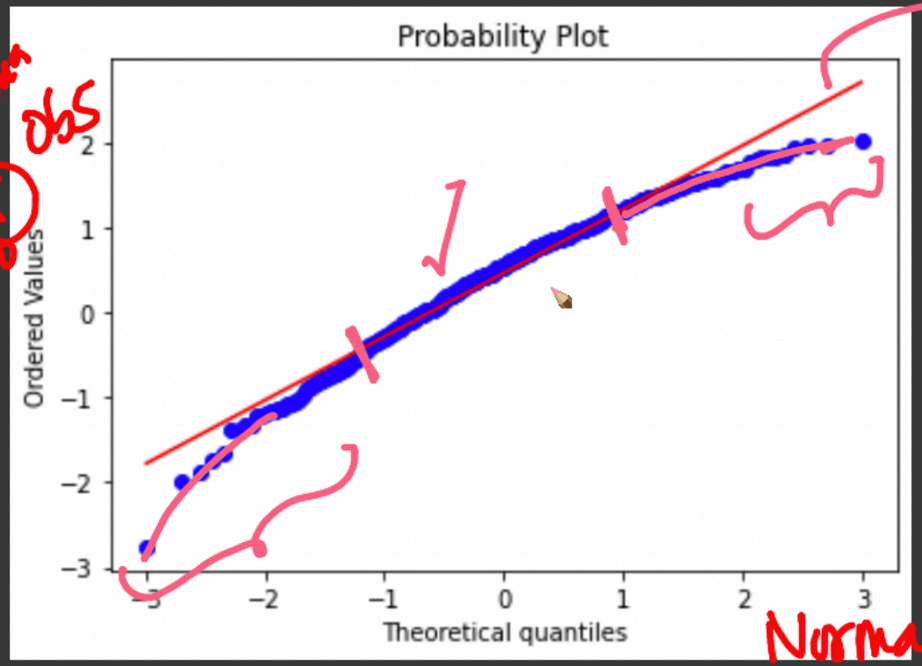
NO



[5]
0s

0s

```
# qq plot of x vs normal
fig, ax1 = plt.subplots()
prob = stats.probplot(x, dist='norm', plot=ax1)
```

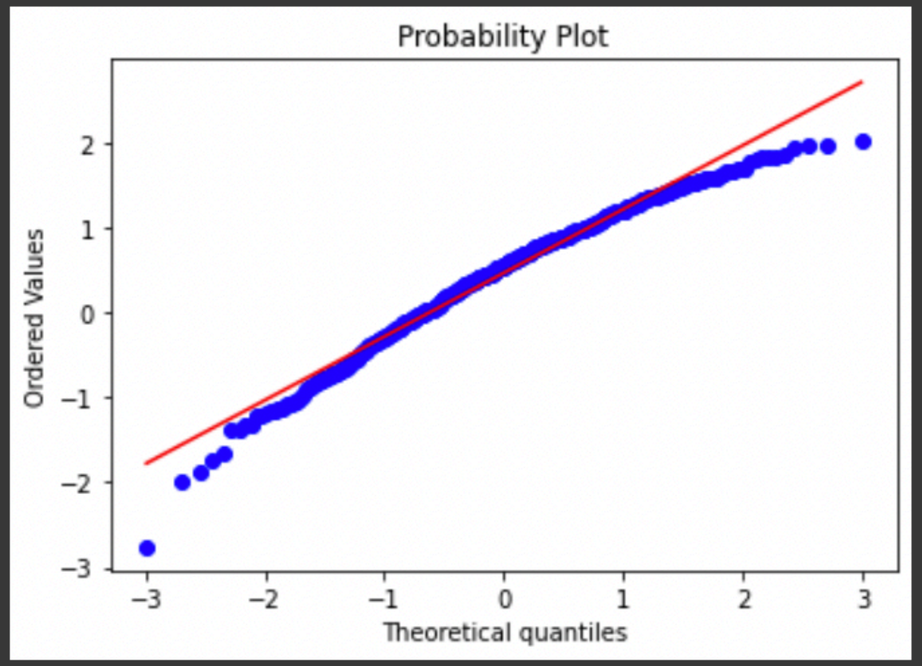


log-gamma
obs
45 degree line
if X ~ Normal
Normal-dist

[5]
0s

0s

```
# qq plot of x vs normal
fig, ax1 = plt.subplots()
prob = stats.probplot(x, dist='norm', plot=ax1)
```



Chrome File Edit View History Bookmarks Profiles Tab Window Help 22:18

InterviewBit Software | Live | DSML Advan | dsml-course/Coordina | dsml-course/01-Scala | dsml-course/15-Hypo | Gamma distribution - | box-cox.ipynb - Colab | Untitled1.ipynb - Colab | scipy.stats.boxcox

colab.research.google.com/drive/1HBSL2rle-pokrz50QIhgWGuZiUaxAeUe#scrollTo=D6-0DIS2T_Pb

+ Code + Text

RAM Disk

Theoretical quantiles

abs →

```
# box cox transform
x1 = x+5
xt, l = stats.boxcox(x,); # returns x_trasnfomred and lambda
print("lambda :" + str(l))

# check if xt is gaussian or not using QQ-Plot
fig, ax2 = plt.subplots()
prob = stats.probplot(xt, dist='norm', plot=ax2)
```

2, -2

$x_1, x_2 \ldots x_{50}$

↓

+5

↓

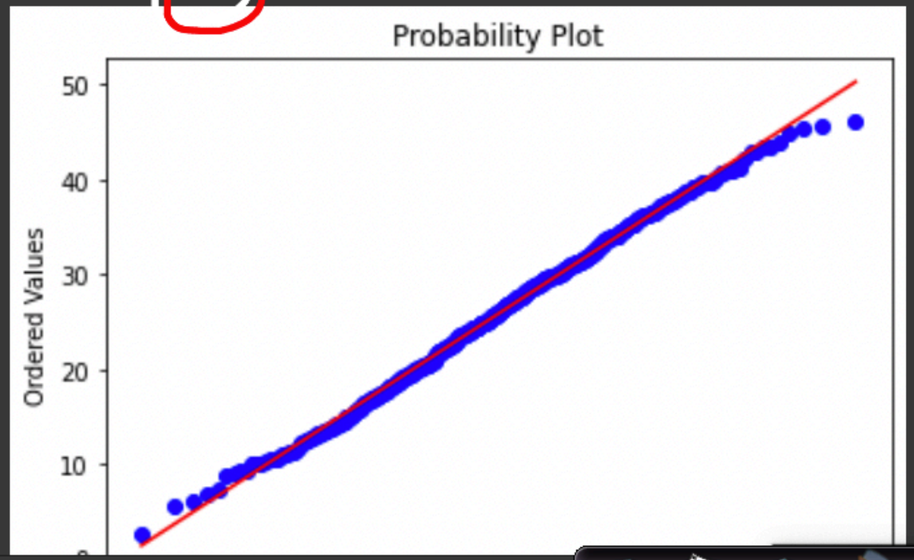
$x_1+5, x_2+5 \ldots x_{50}+5$

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-10-1c3c724c108c> in <module>()
      1 # box cox transform
----> 2 xt, l = stats.boxcox(x); # returns x_trasnfomred and lambda
      3 print("lambda :" + str(l))
      4
      5 # check if xt is gaussian or not using QQ-Plot

/usr/local/lib/python3.7/dist-packages/scipy/stats/morestats.py in boxcox(x, lmbda, alpha)
   1043
   1044     if any(x <= 0):
-> 1045         raise ValueError("Data must be positive.")
   1046
   1047     if lmbda is not None:  # single transformation
```

24 / 24

```
# box cox transform
x1 = x+5
xt, l = stats.boxcox(x1,); # returns x_trasnfomred and lambda
print("lambda :" + str(l))

# check if xt is gaussian or not using QQ-Plot
fig, ax2 = plt.subplots()
prob = stats.probplot(xt, dist='norm', plot=ax2)
```

```
lambda :2.4288828199512202
```



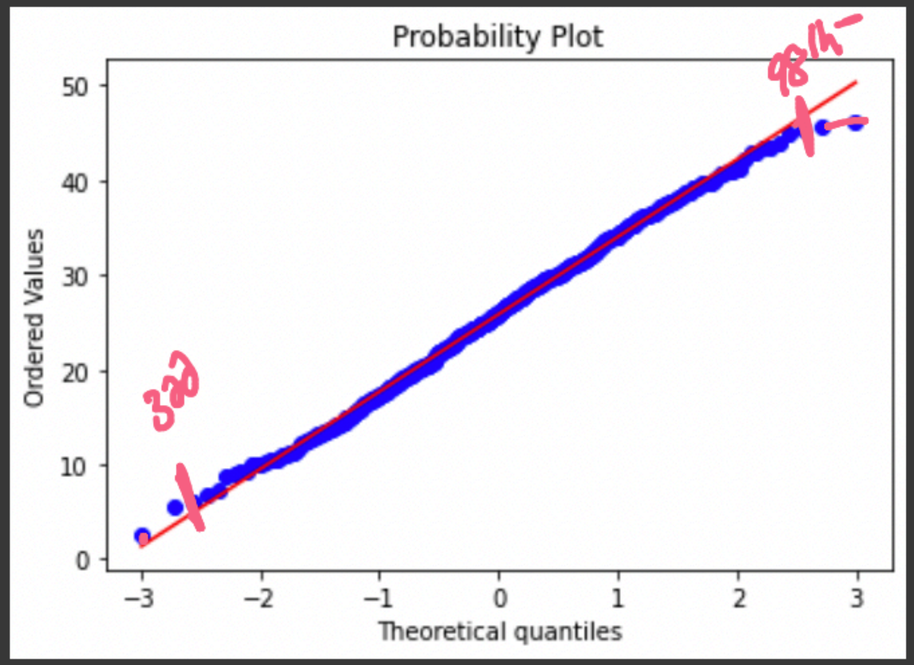
$$\text{box cox}\left(x_i + 5\right)$$

$$\lambda = 2.4 \neq 0$$

$$x_t = \frac{(x_i+5)^{2.42} - 1}{2.42}$$

```
xt, l = stats.boxcox(x1,); # returns x_trasnfomred and lambda
print("lambda :" + str(l))

# check if xt is gaussian or not using QQ-Plot
fig, ax2 = plt.subplots()
prob = stats.probplot(xt, dist='norm', plot=ax2)
```

```
lambda :2.4268828199512202
```



better ~ Gaussian approx

Confidence Interval using bootstrapping

per-item-sale price

e.g:

$S \rightarrow \{S_1, S_2, S_3, \ldots S_{100}\}$ (n=100)

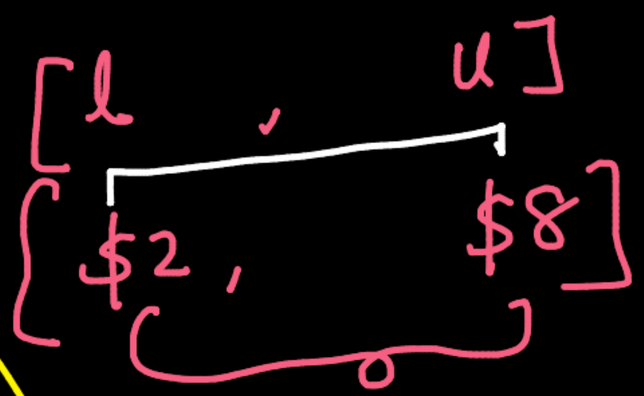
→ randomly sampling Millions

not gaussian dist

only data

(Q) find the 95% C.I of the median sale price @ Amazon (population)

→ bootstrapped samples



Gaussian μ ≈ median

95% C.I of the median of the population

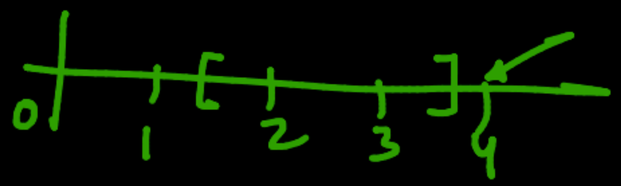
bootstrapped - Sampling

↓ all we have → random Sample of 100 obs

1

↓

≈ population



median → $4 = median(sample)

point estimate

Do Not populate

estimate population-median

[$2, $6] 95% C.I

[$4.5, $4.5] let

pop-median

drugs for COVID

median recovery time

population median

IS

1000

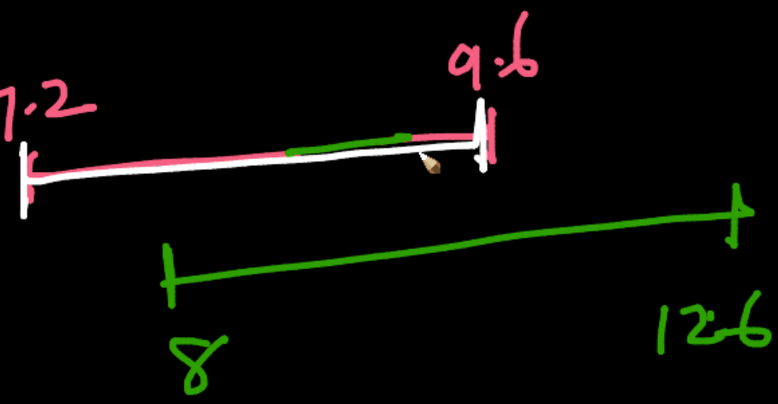
2000

d_1 → [7.2, 9.6]

d_2 → [8, 12.6]

d_3

d_4



Q

Sample-size is fixed

95% C.I

[2 , 3.2]

99% C.I

[1.2 4.6]

wider or narrower

$[-\infty \quad \infty]$

medians = [ ]

$S = \{s_1, s_2, \ldots s_{100}\}$

For i = 1 to 1000 # resampling

1. S_i ← sample 100 prices from S with replacement ✓ (50)
2. medians[i] = median(S_i)

Q why not without replacement

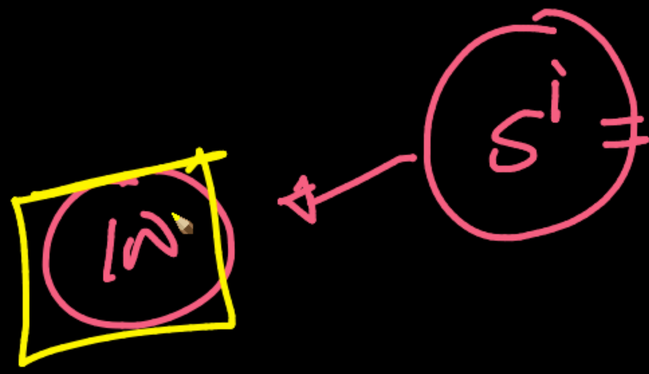
Sort medians

→

1000

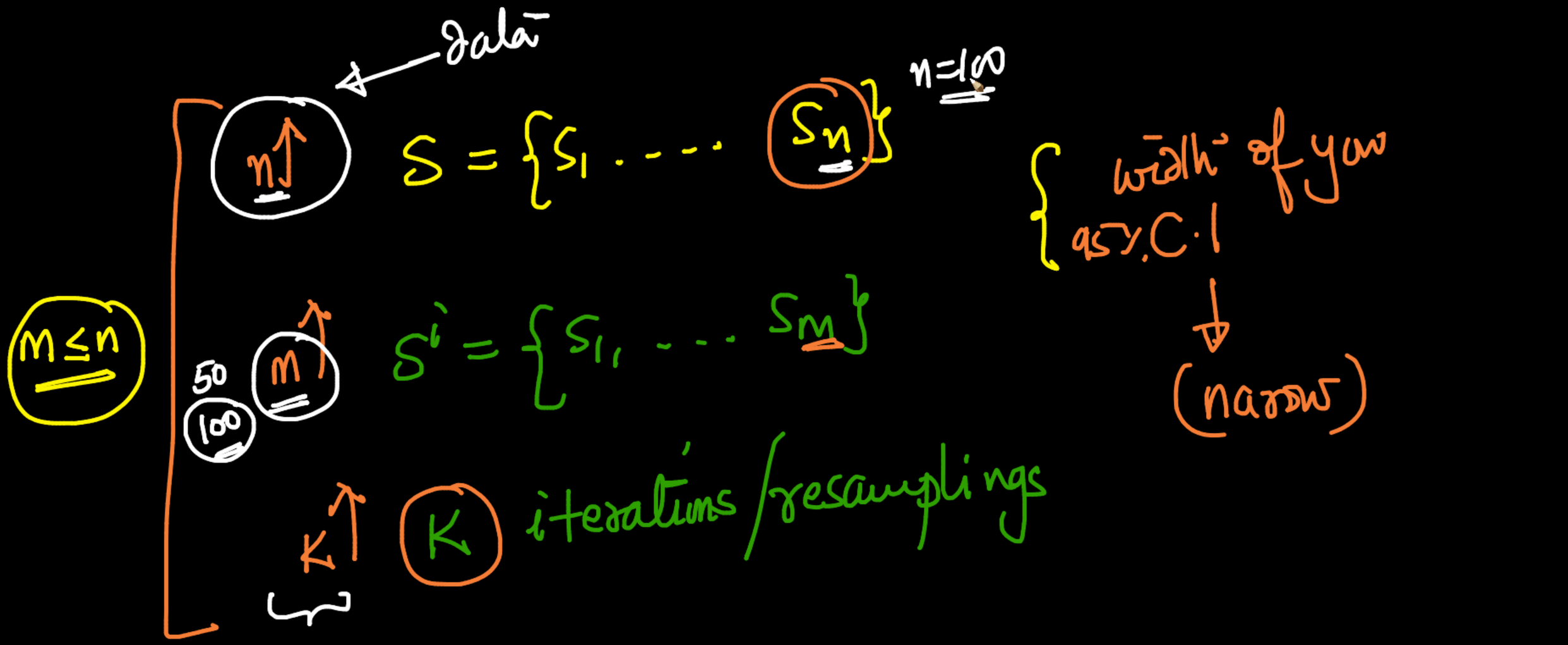
$S = \{s_1, s_2, \ldots s_N\}$

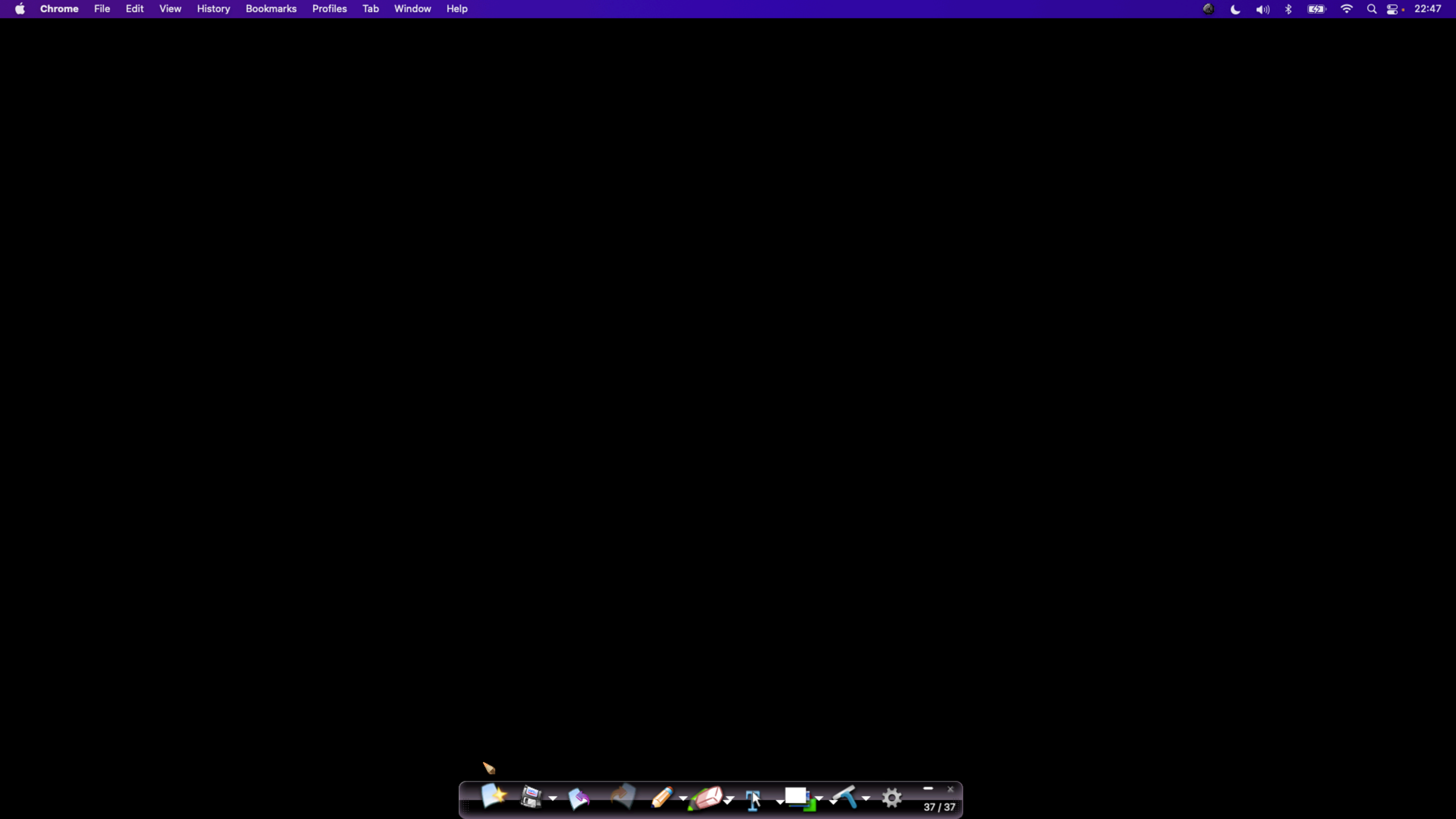
$S^1 = \{s_4,$

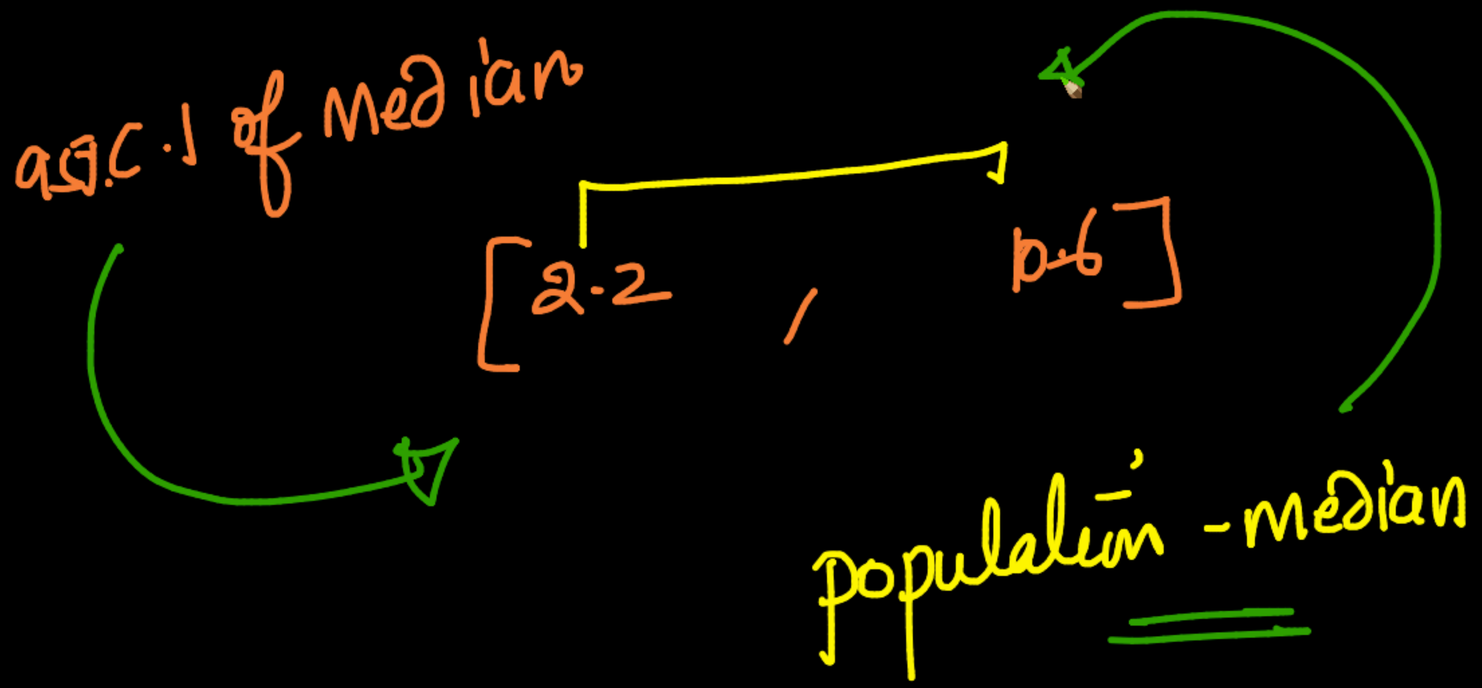


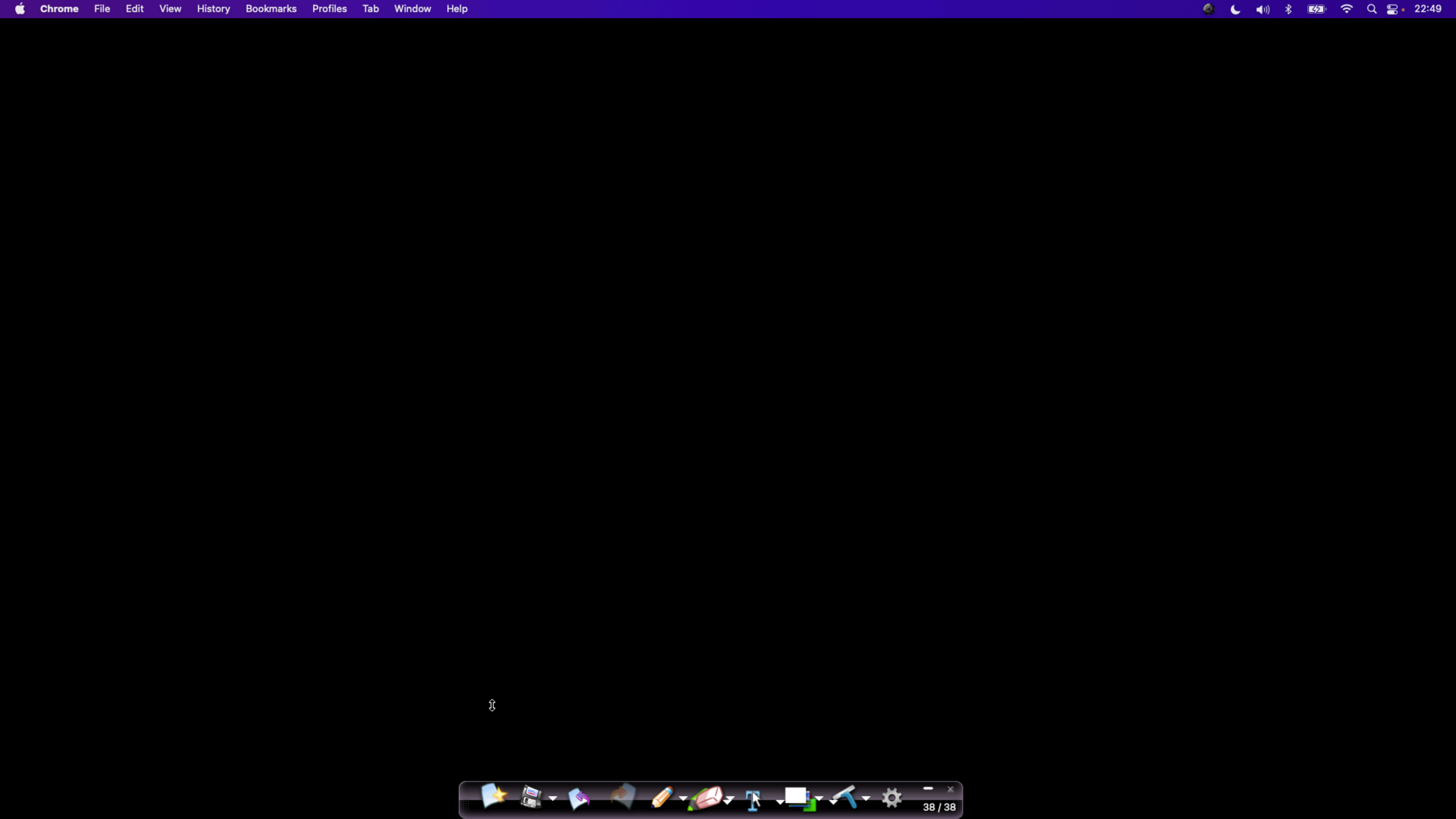
= randomly sample 100 obs from S without replacement

$S^1 = S^2 = S^3 = S^4 \ldots$


Jalan
n
S = {S1 .... Sn}
n=100
{ width of your
95% C.I
(narrow)
m ≤ n
50
100
m
S' = {S1, ... Sm}
K
K iterations / resamplings


Chrome File Edit View History Bookmarks Profiles Tab Window Help
22:47
37 / 37




Chrome File Edit View History Bookmarks Profiles Tab Window Help
22:49
38 / 38

$S^1 = S^2 = S^3 \ldots$ if sampling without replacement

$n = M$

$S_{i=5}$ = {1, 2, 3, 4, 5}

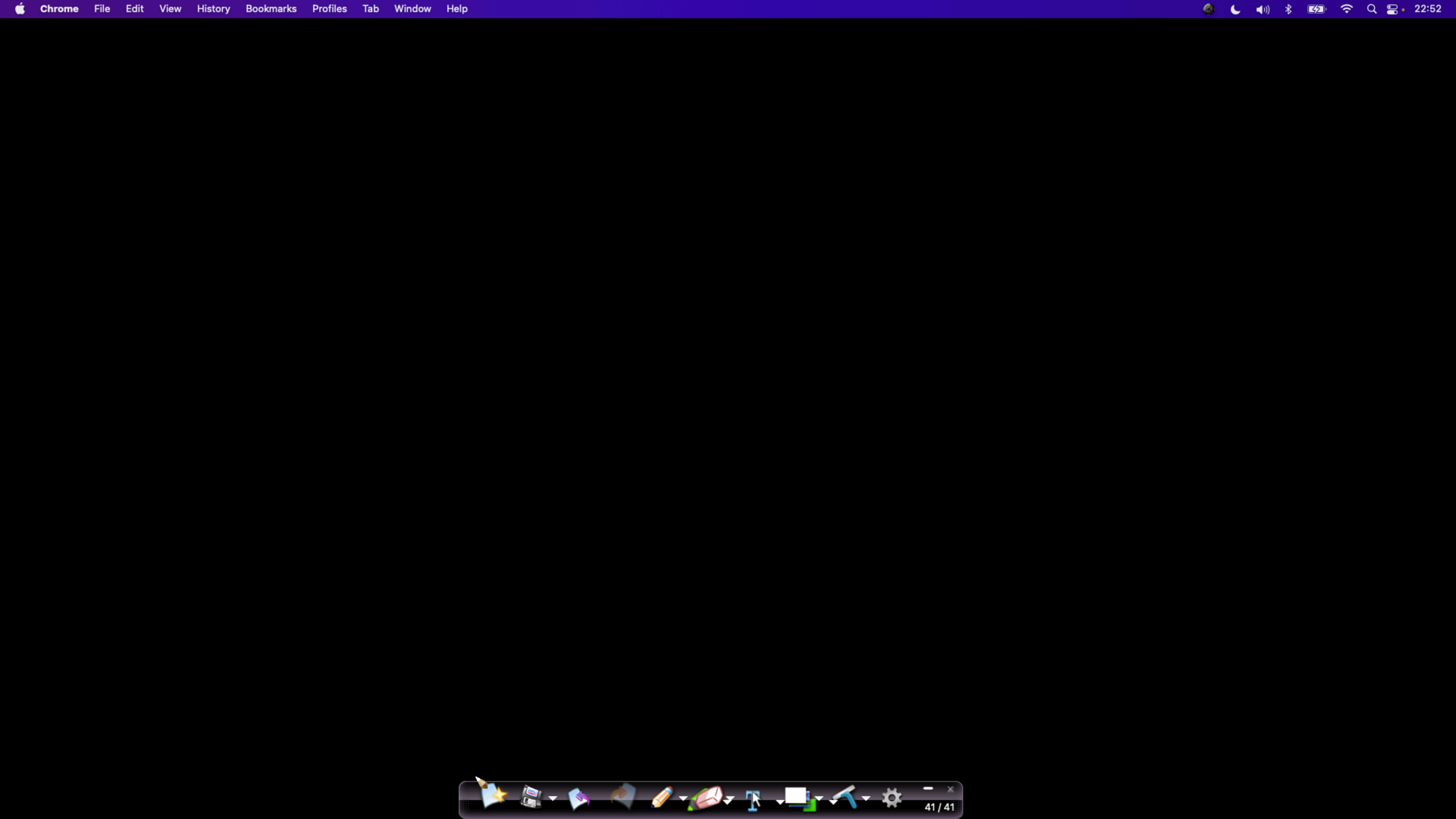
$S^1_{M=5}$ = {2, 3, 4, 5, 1}

with repl:

S = {1, 2, 3, 4, 5}

S^1 = {3, 4, 3, 5, 1}

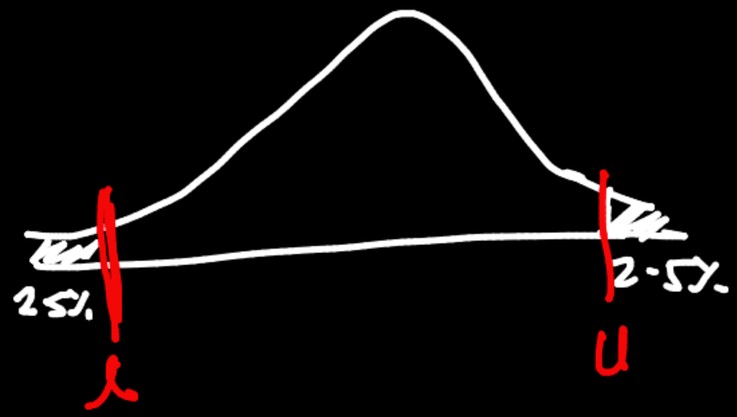
✓ bootstrapping → Means
Medians
percentiles


Chrome File Edit View History Bookmarks Profiles Tab Window Help
22:52
41 / 41

bootstrapped

1000 resamplings

Median



Sort them

2.5%

1 2 3

2.5%

2.5%

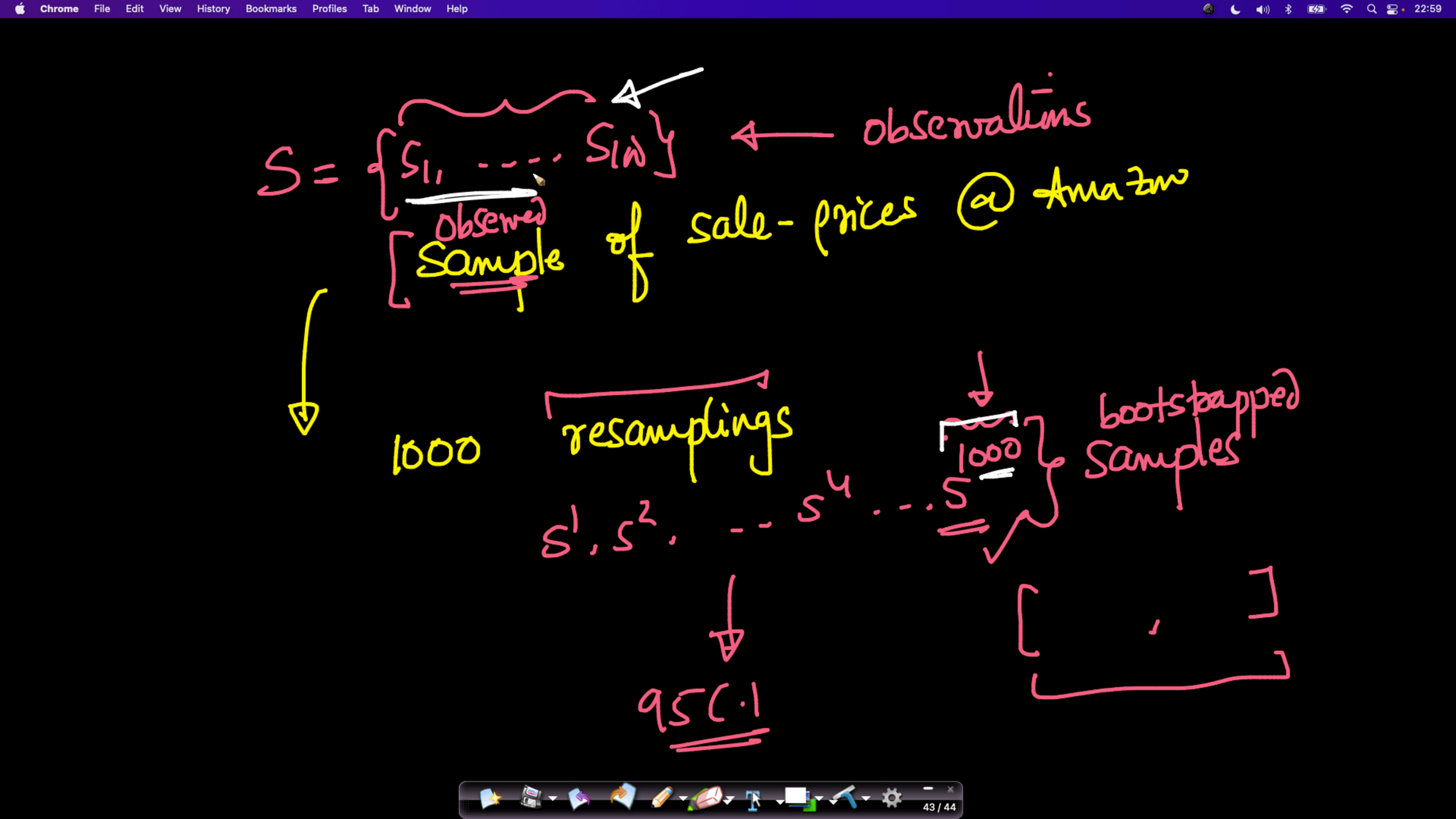
1000

2.5th

97.5th

non-parametric stat-tech

percentiles


S = {s1, ---- s1000}
observations
observed sample of sale-prices @ Amazon
1000 resamplings
s1, s2, -- s4 ... s1000
1000 bootstrapped samples
95C.I
[,]
43 / 44

{ CLT → means

96.231

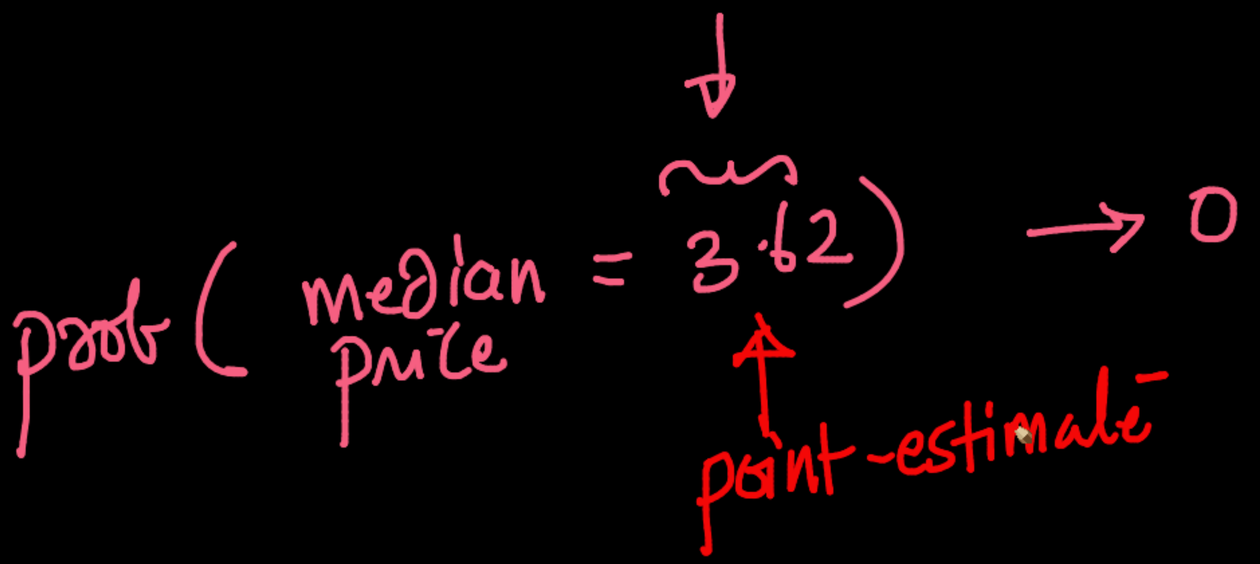
Prob(a person who weighs 96.23 kgs)

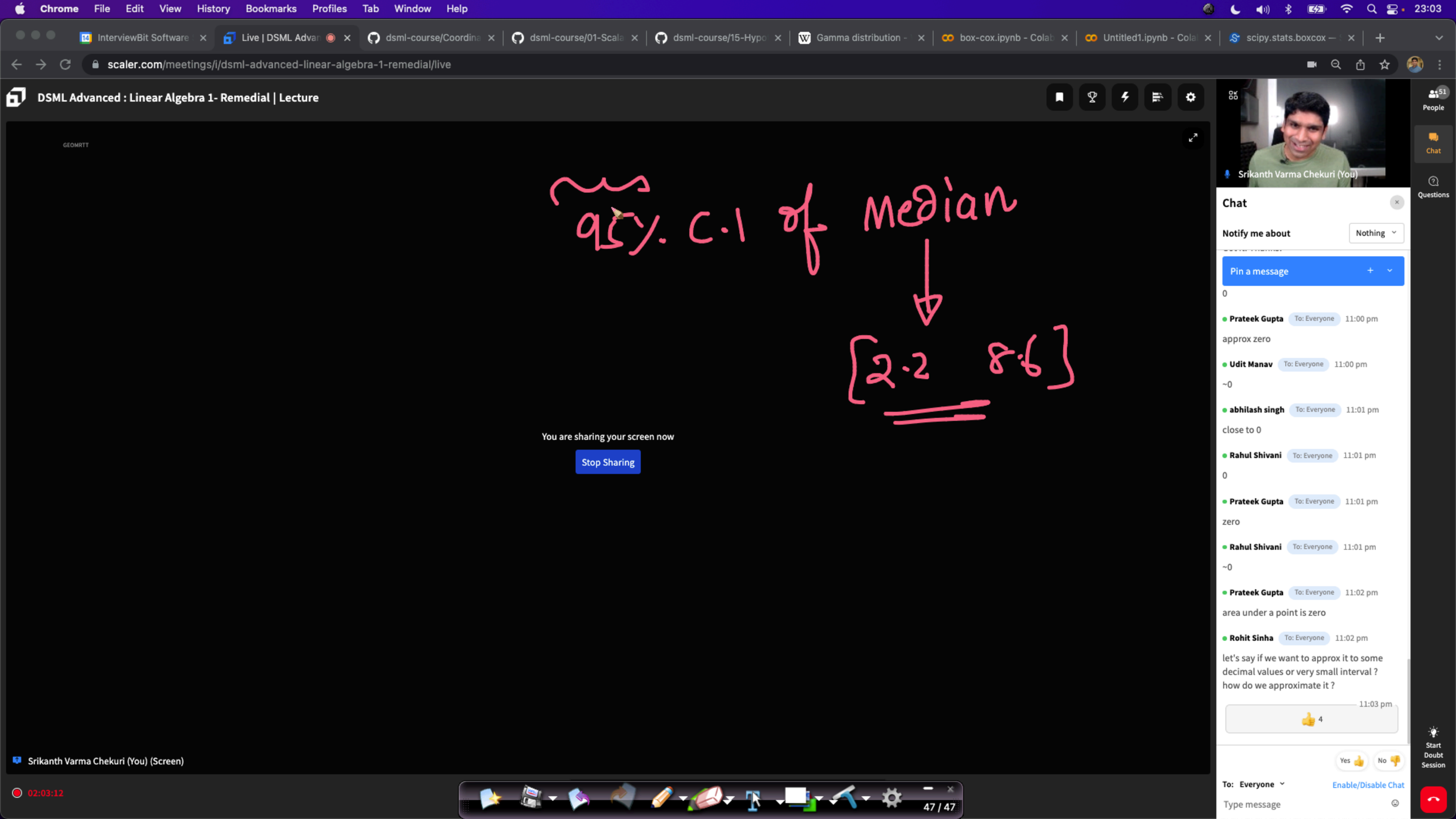
96.23000000...

Continous r.v

PDF

0


prob (median price = 3.62) → 0
point-estimate

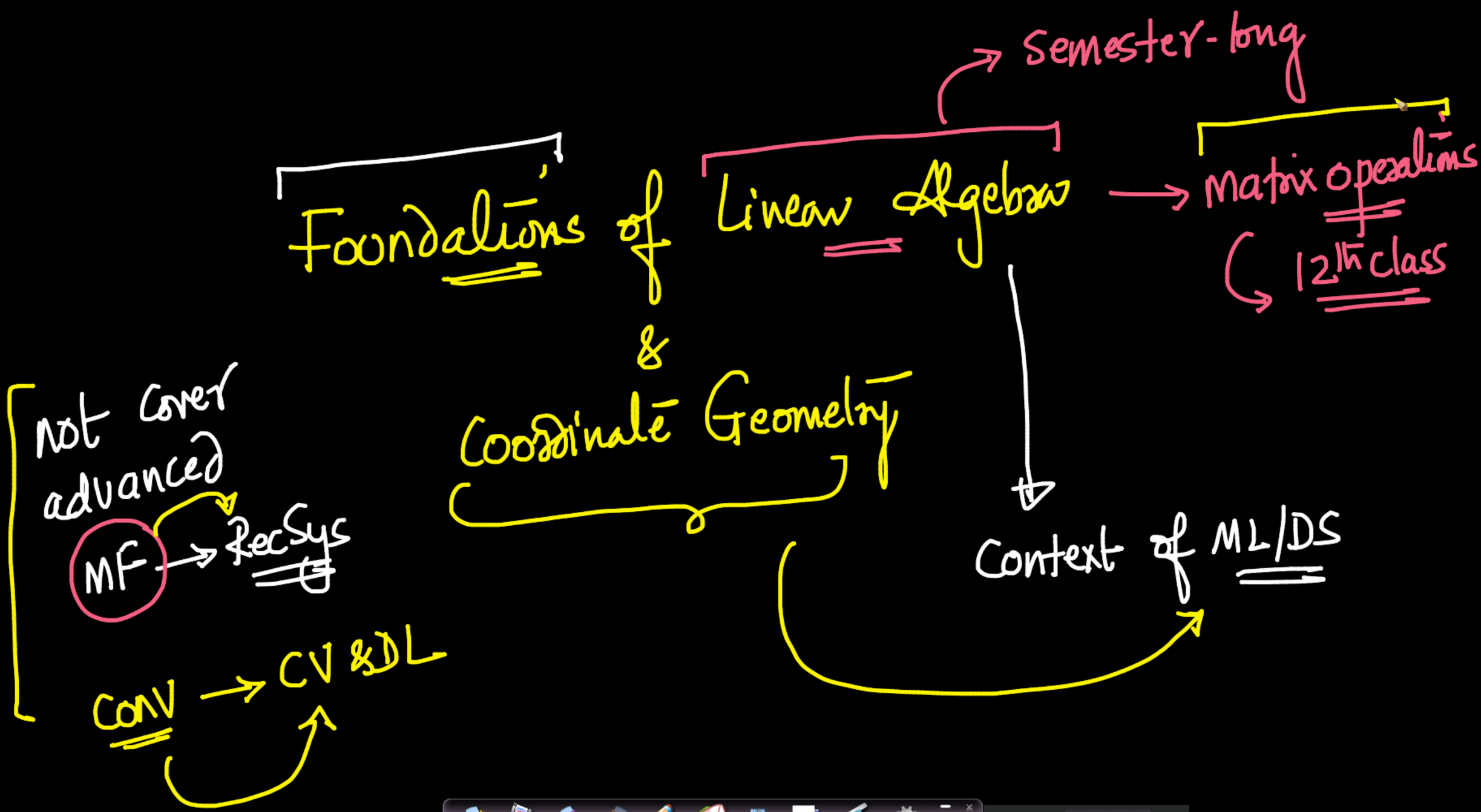

Chrome File Edit View History Bookmarks Profiles Tab Window Help
23:03
InterviewBit Software
Live | DSML Advan
dsml-course/Coordina
dsml-course/01-Scala
dsml-course/15-Hypo
Gamma distribution -
box-cox.ipynb - Colab
Untitled1.ipynb - Cola
scipy.stats.boxcox —
scaler.com/meetings/i/dsml-advanced-linear-algebra-1-remedial/live
DSML Advanced : Linear Algebra 1- Remedial | Lecture
GEOMRTT
qty. c.1 of median
[2.2 8.6]
You are sharing your screen now
Stop Sharing
Srikanth Varma Chekuri (You) (Screen)
02:03:12
47 / 47
Srikanth Varma Chekuri (You)
People
Chat
Questions
Chat
Notify me about
Nothing
Pin a message
0
Prateek Gupta To: Everyone 11:00 pm
approx zero
Udit Manav To: Everyone 11:00 pm
~0
abhilash singh To: Everyone 11:01 pm
close to 0
Rahul Shivani To: Everyone 11:01 pm
0
Prateek Gupta To: Everyone 11:01 pm
zero
Rahul Shivani To: Everyone 11:01 pm
~0
Prateek Gupta To: Everyone 11:02 pm
area under a point is zero
Rohit Sinha To: Everyone 11:02 pm
let's say if we want to approx it to some decimal values or very small interval ? how do we approximate it ?
11:03 pm
4
Yes
No
To: Everyone
Enable/Disable Chat
Type message
Start Doubt Session

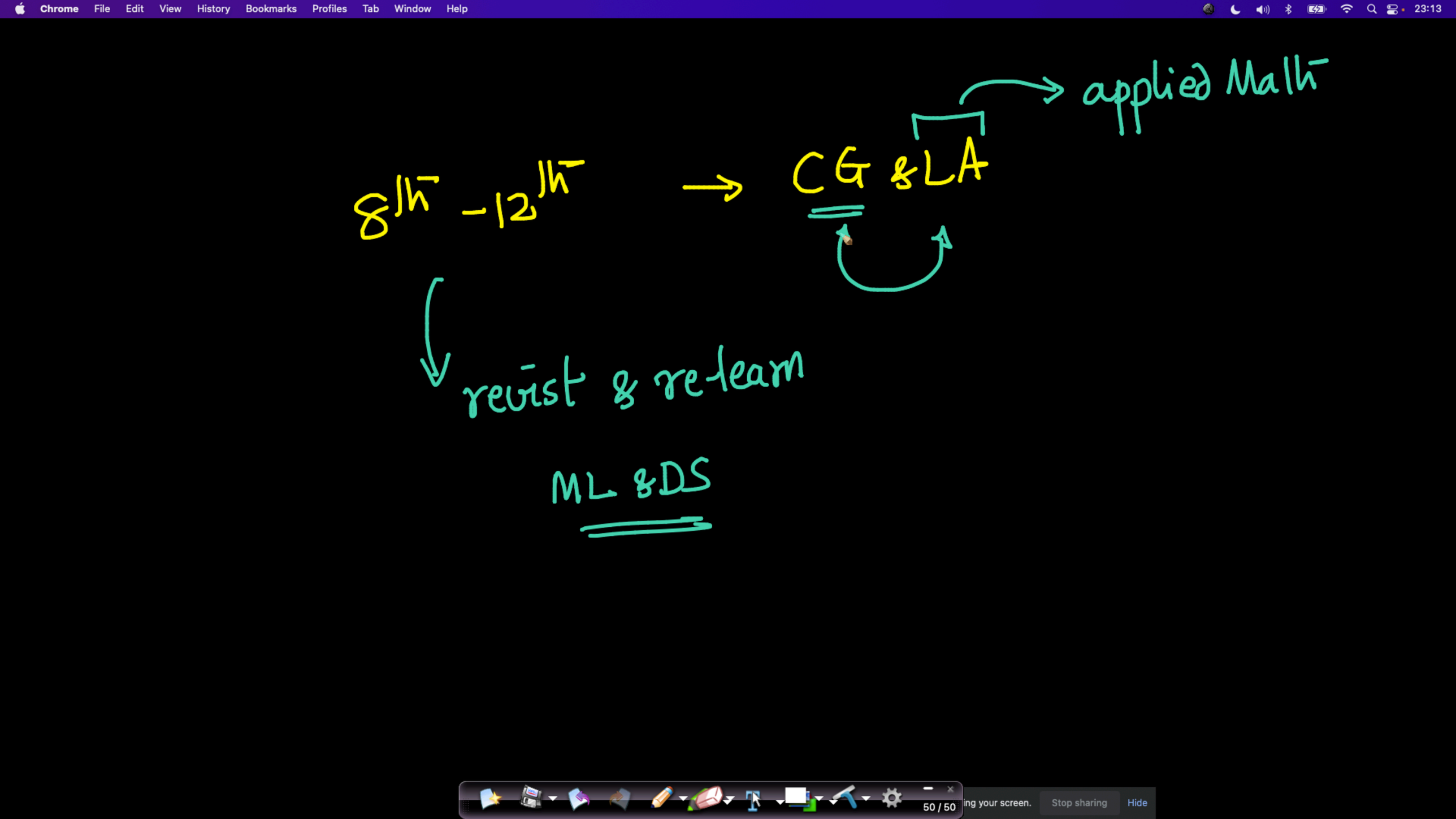
Chrome File Edit View History Bookmarks Profiles Tab Window Help 23:05

Break:

23:07

48 / 49




8 lh - 12 lh
CG & LA
applied Malh
revist & relearn
ML & DS

Lines, plane → high-school & college

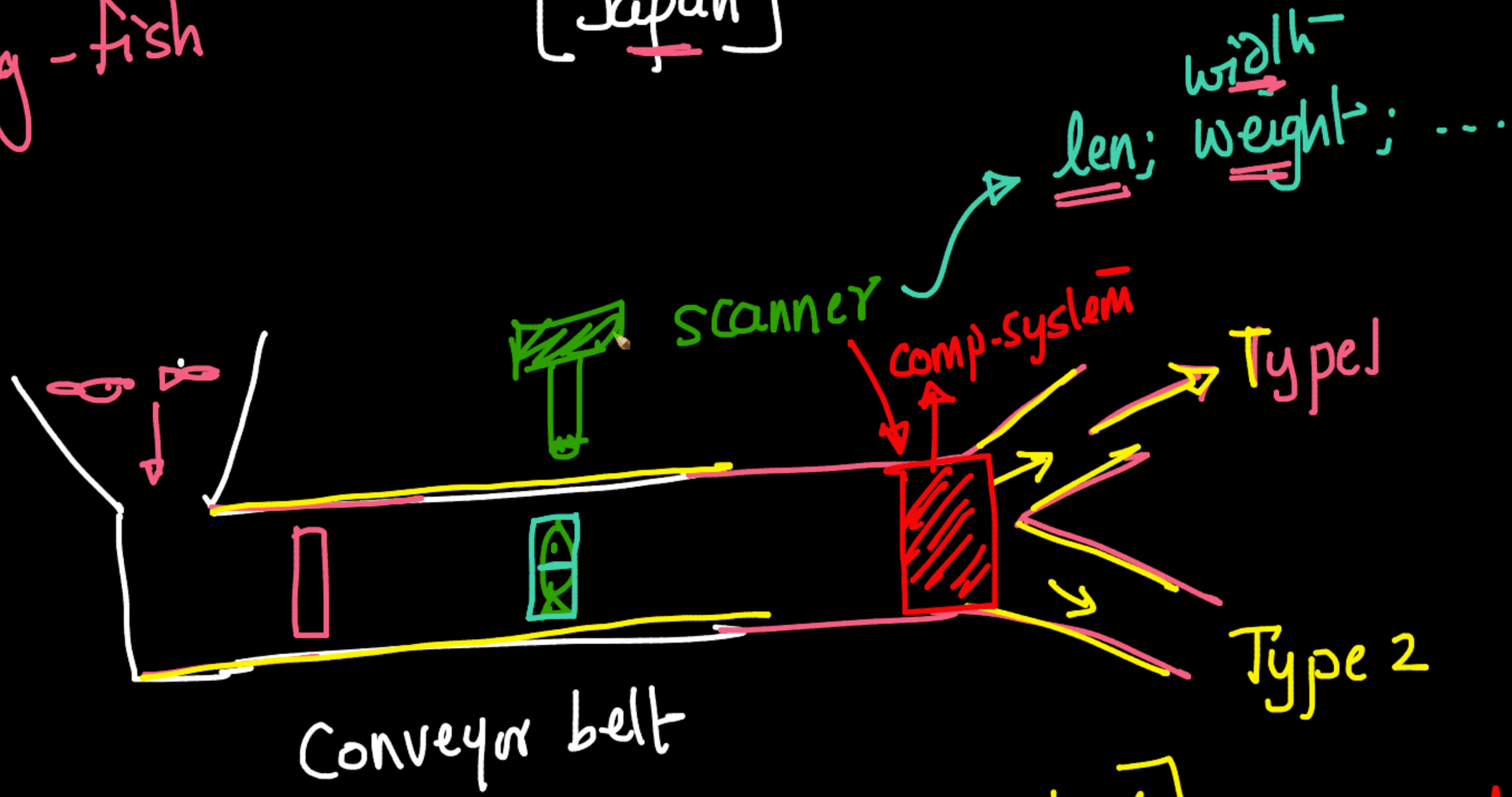
Vectors; matrices; dot-product → Physics X

real-world

circle; ellipses; spheres

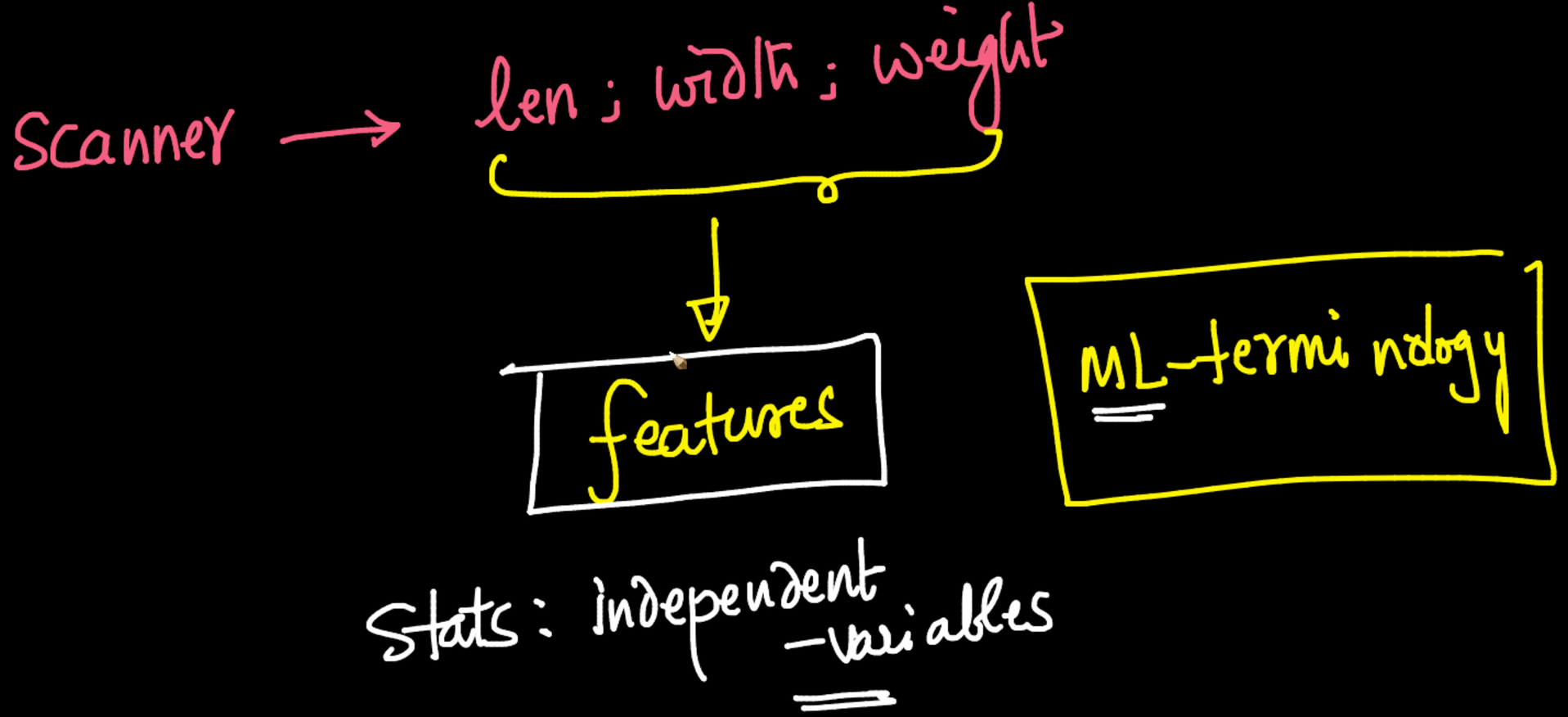
Sorting - fish

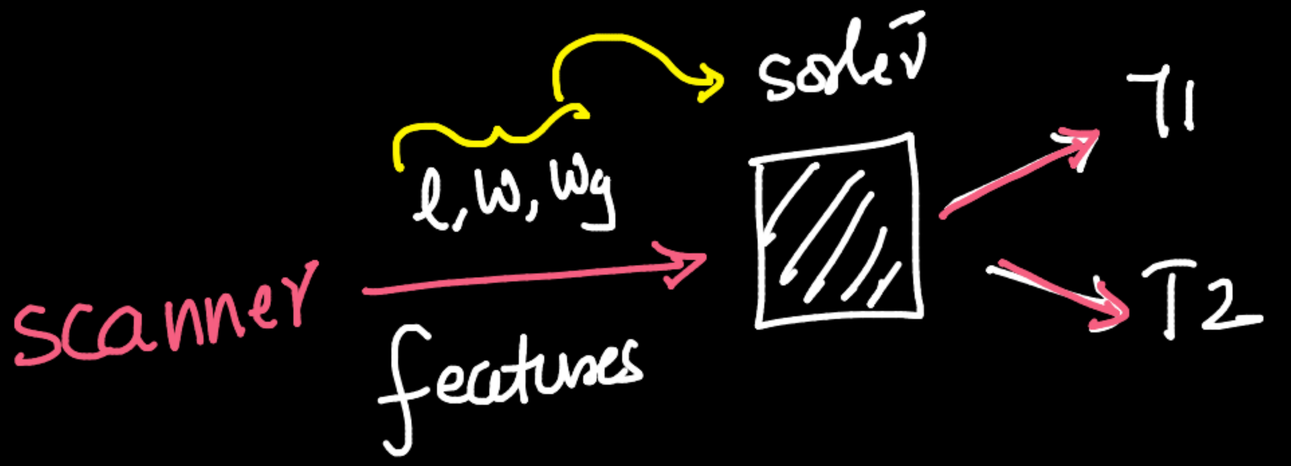
[Japan]



[Credits: Duda - hart]

automated sorter



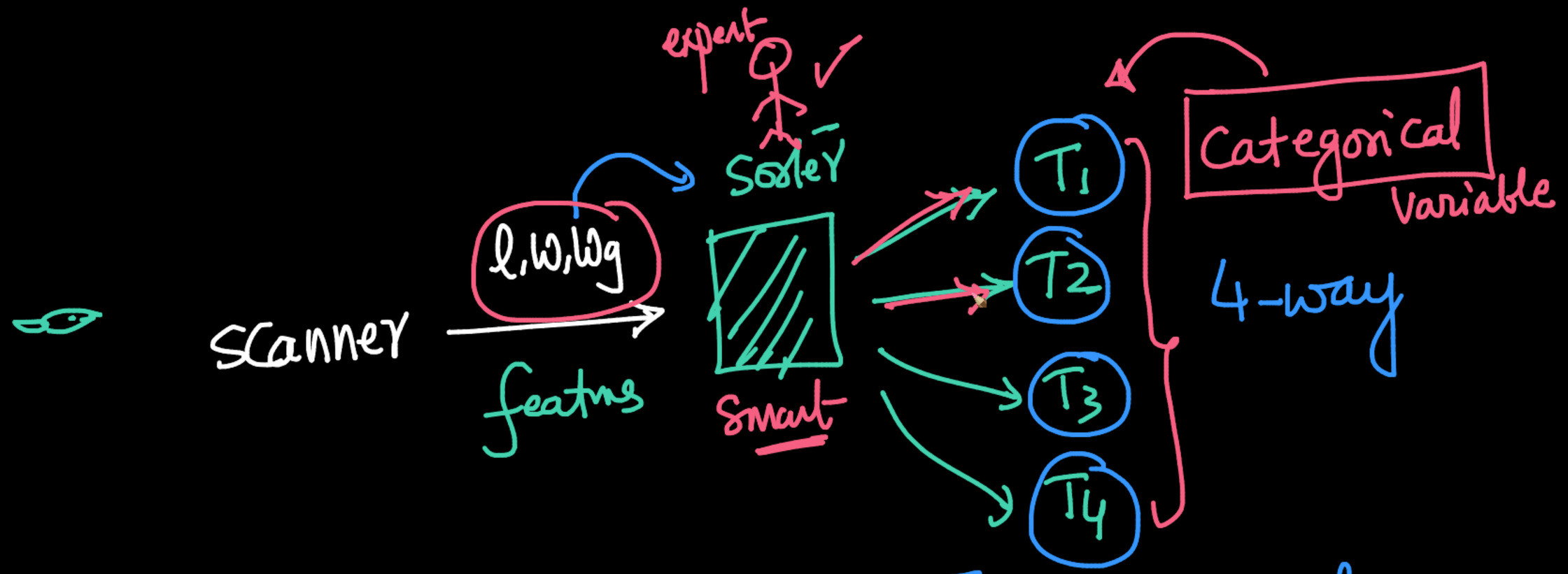


binary

categorical

classifying

[Binary - classification]



multi-class Classification

> 2 classes

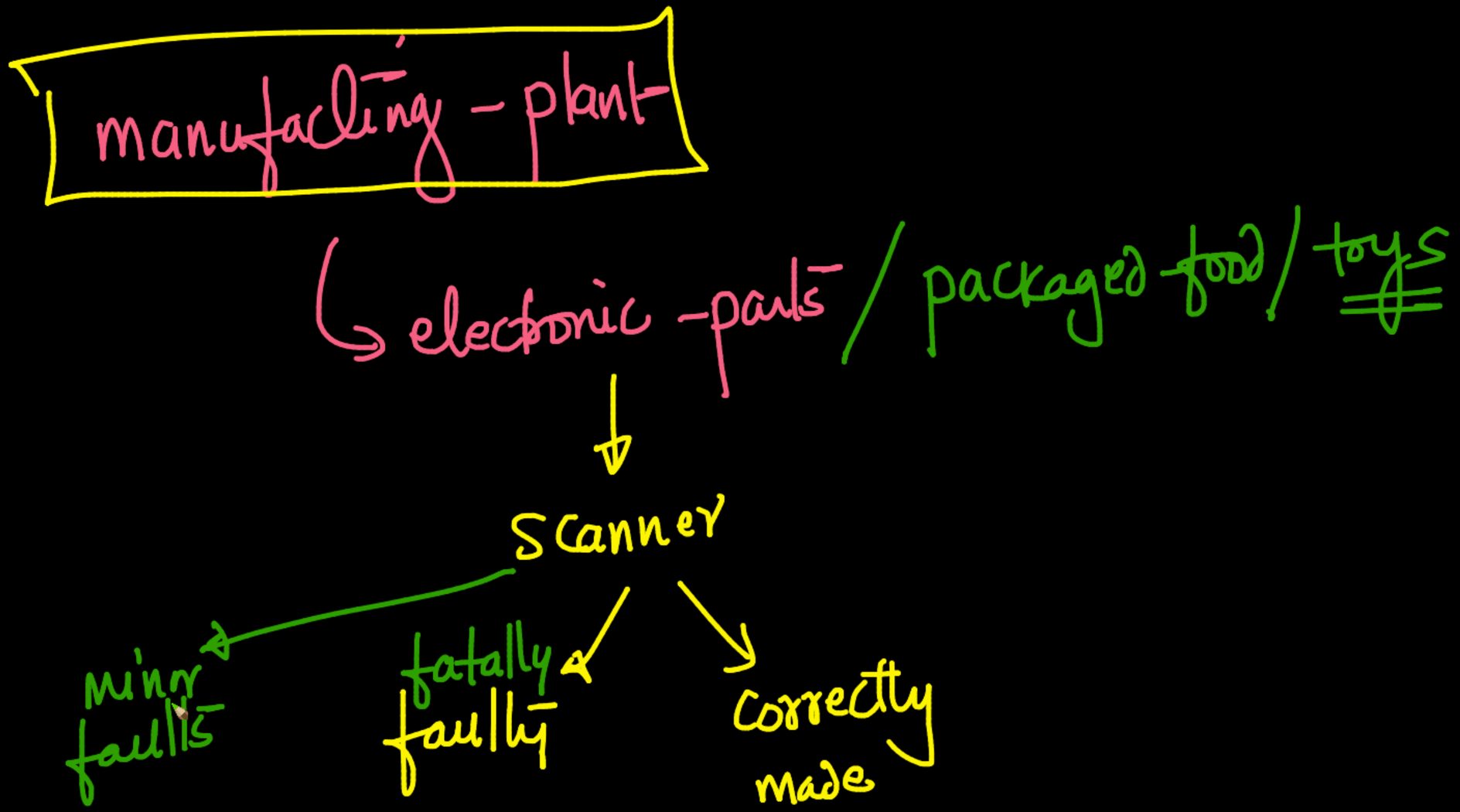
natural ordering amongst classes?

Predict age of fish using l, w, wg

↓

numerical

regression



Sorting fish ⟶ Linear Alg & Co-ordinate

(Q) predict Marks = {1, 2, 3, 4 - - 100} → discrete ; features

regression or multi-class classification

practical →

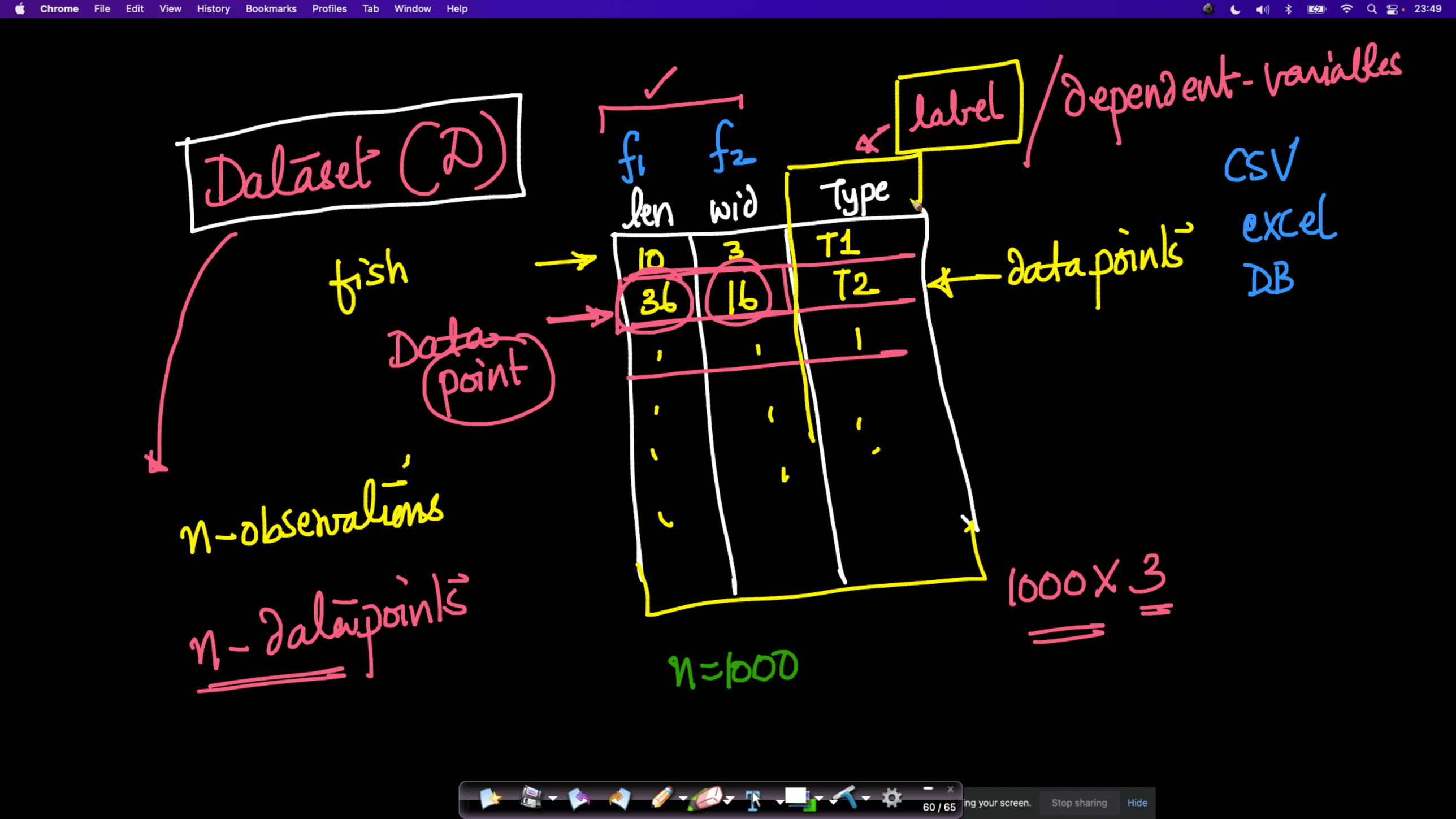
natural ordering

no-ordering amongst the classes

round (3.2)

3.2

3

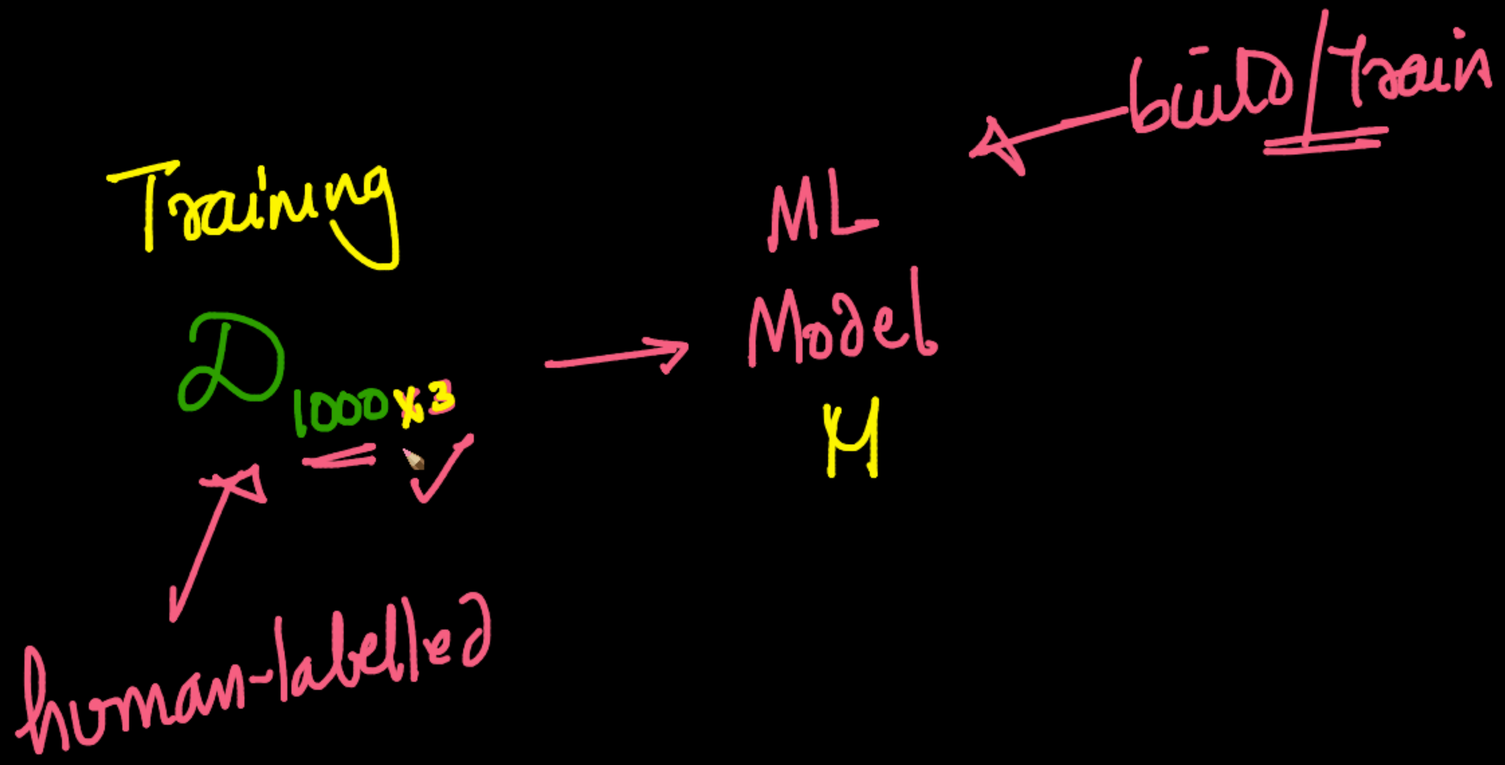

Dataset (D)
fish
Data point
n-observations
n-datapoints
f1
f2
len
wid
Type
label
dependent-variables
10
3
T1
36
16
T2
datapoints
CSV
excel
DB
n=1000
1000 X 3

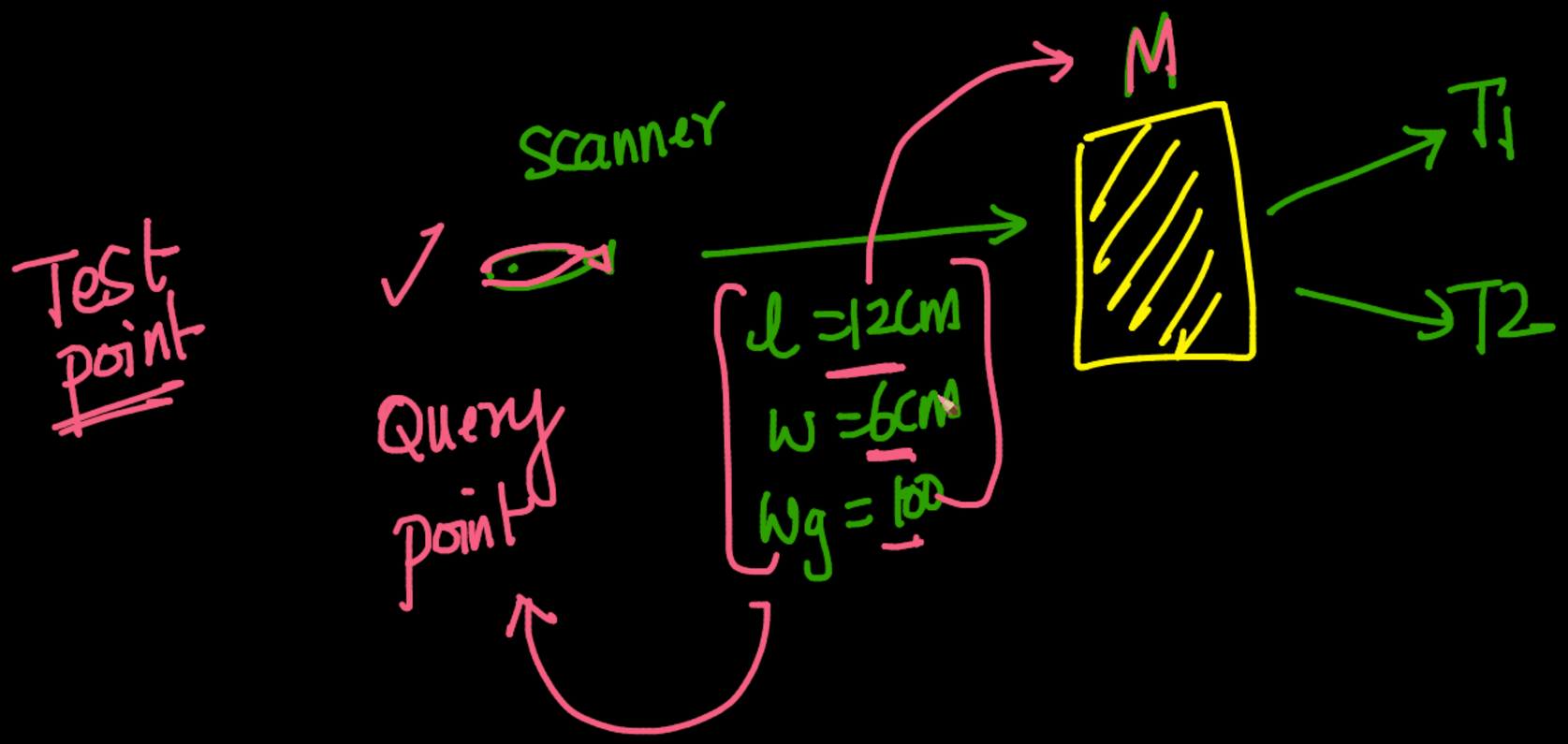
(Q) How do we get the data ?

↳ Sensors : $f_1, f_2, \ldots$; label → human expert

f_1, f_2 : l, w

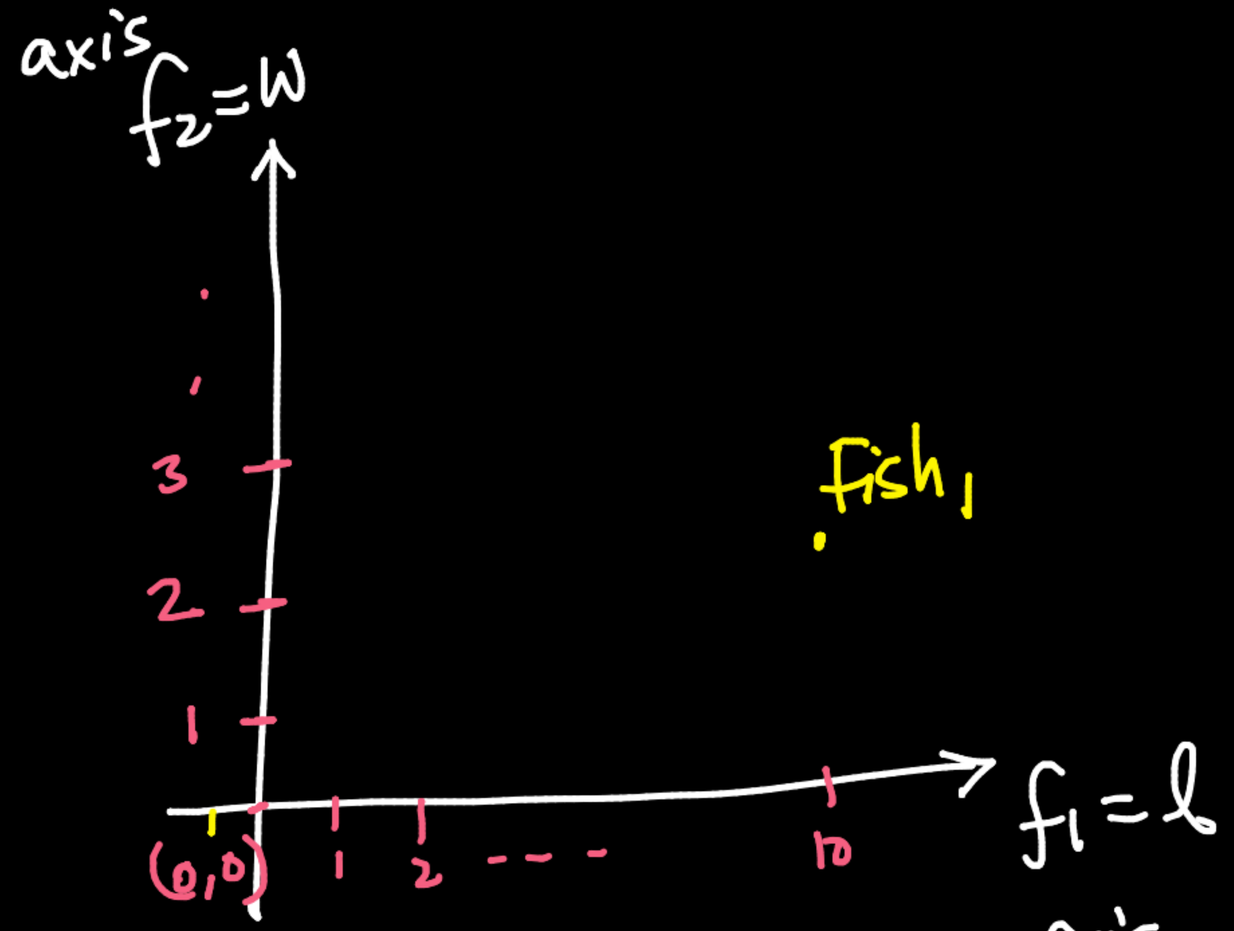
human-labelled (annotated-data)





data - point

data - vector

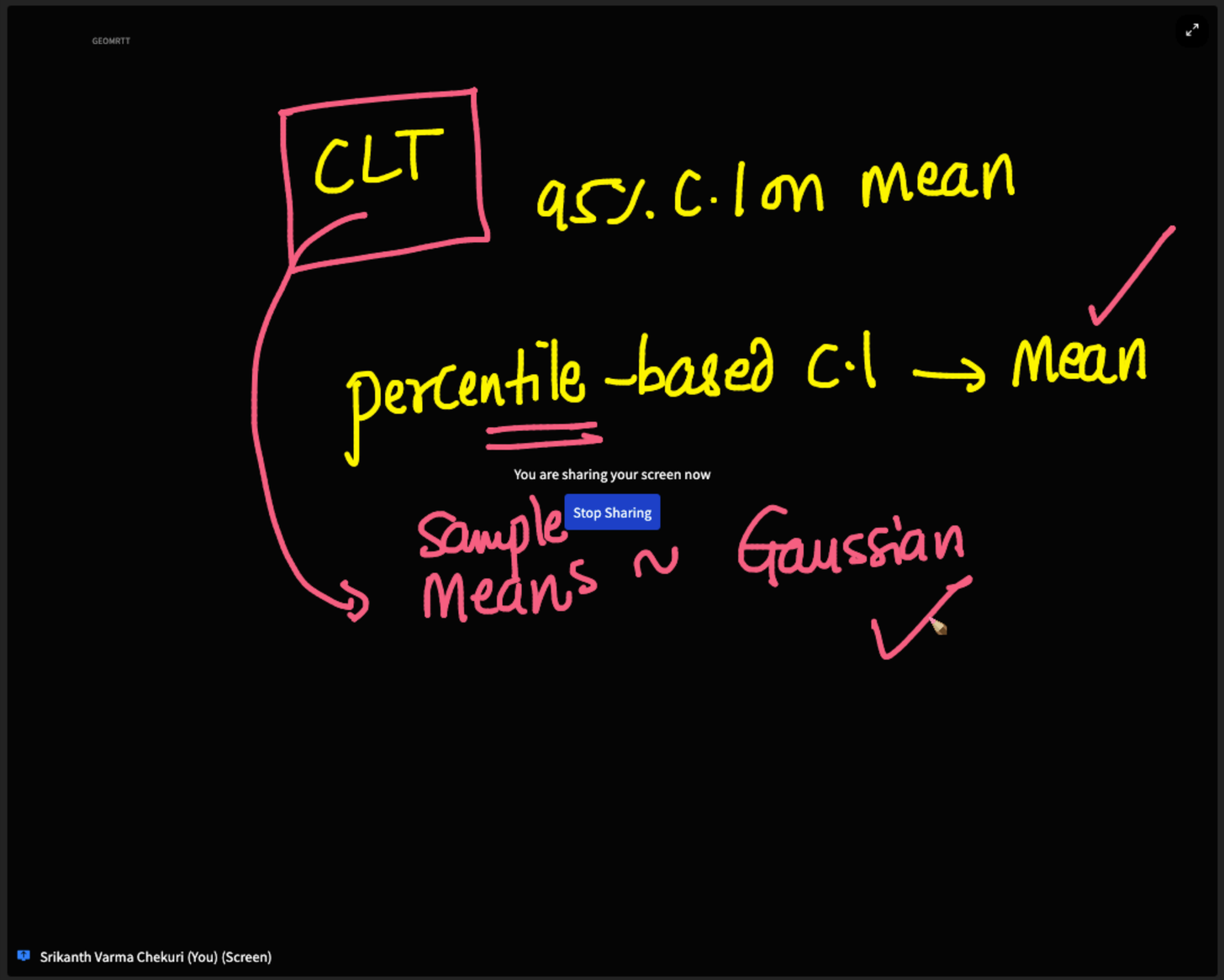


Fish$_1$

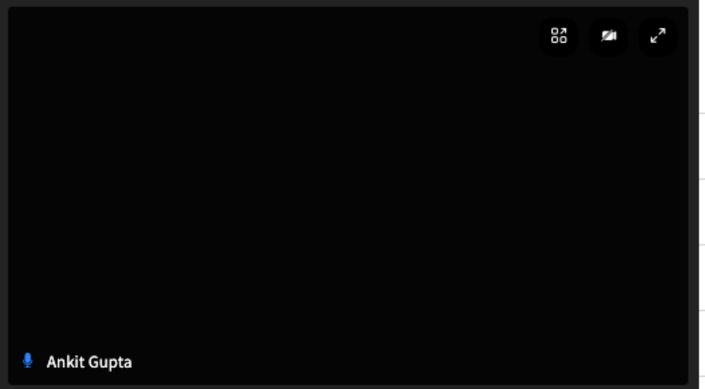
2D - geometry

Fish$_1$ = [10cm$_l$, 2cm$_w$]

DSML Advanced : Linear Algebra 1- Remedial | Lecture



Srikanth Varma Chekuri (You) (Screen)



People

Search

- Srikanth Varma Chekuri (Host, You)
- Ankit Gupta
- Rishav Kumar
- Adarsh Vinayak
- Aditya Yadav
- Amit Srivastava
- Arijit Bhowmick
- Atin Gupta
- Avijit Das
- Gowtham
- Hemant Parashar
- MANISH PATEL
- Mohammad Rashid
- Muthu kamalan .M
- Navroop
- Nikunj Prajapati
- Raghu Vamsi Yaram
- Rahul Maramreddy

People 30 | Chat 1 | Questions | Start Doubt Session

02:59:05

65 / 65

ing your screen. Stop sharing Hide

DSML Advanced : Linear Algebra 1- Remedial | Lecture

large
k-resamplings → Median

CLT ≈ percentile based ✓

You are sharing your screen now

Stop Sharing

Sample-Size = n = 100

Size of bootstrap sample = M = 1MM

Sameer Udgirkar has raised hand Unmute

Ankit Gupta

Srikanth Varma Chekuri (You)

People

Search

- Srikanth Varma Chekuri (Host, You)
- Ankit Gupta
- Rishav Kumar
- Adarsh Vinayak
- Aditya Yadav
- Amit Srivastava
- Arijit Bhowmick
- Atin Gupta
- Avijit Das
- Gowtham
- MANISH PATEL
- Mohammad Rashid
- Muthu kamalan .M
- Navroop
- Nikunj Prajapati
- Raghu Vamsi Yaram
- Rahul Maramreddy
- Rahul Shivani

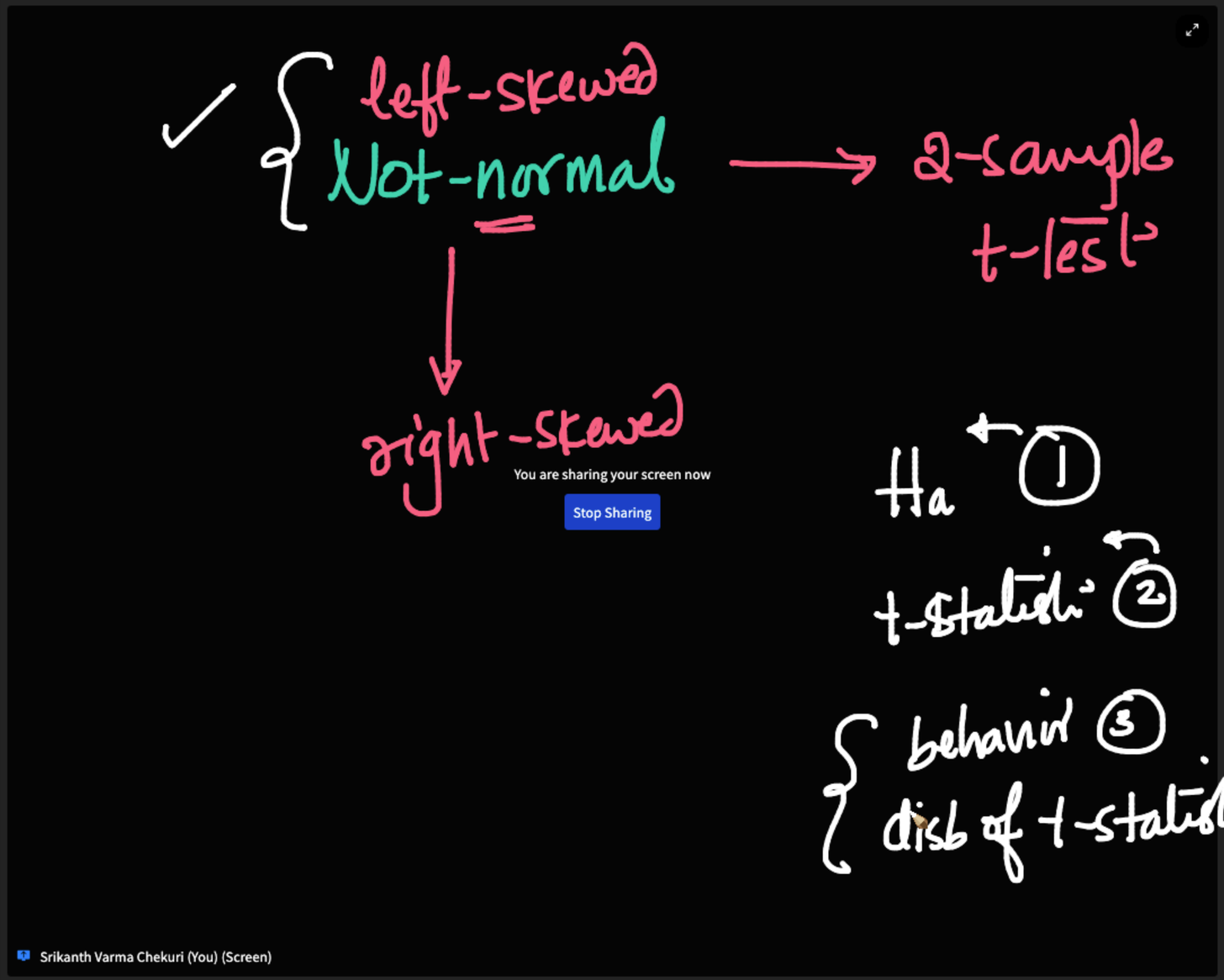
People 28 | Chat 1 | Questions | Start Doubt Session

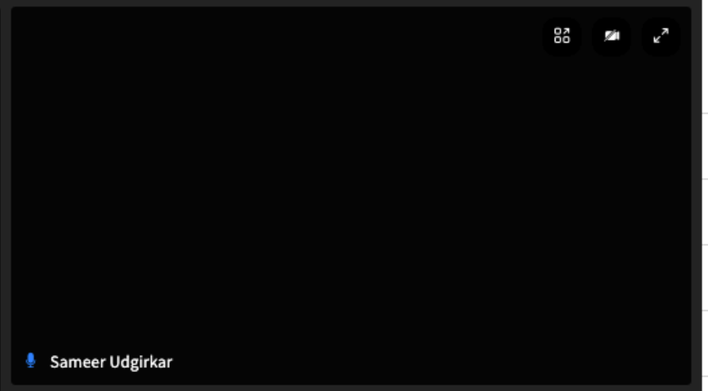
03:01:41

66 / 67

ing your screen. Stop sharing Hide

DSML Advanced : Linear Algebra 1- Remedial | Lecture





People

Search

- Srikanth Varma Chekuri (Host, You)
- Ankit Gupta
- Rishav Kumar
- Sameer Udgirkar
- Adarsh Vinayak
- Aditya Yadav
- Amit Srivastava
- Arijit Bhowmick
- Atin Gupta
- Avijit Das
- Gowtham
- MANISH PATEL
- Mohammad Rashid
- Muthu kamalan .M
- Navroop
- Raghu Vamsi Yaram
- Rahul Maramreddy
- Rahul Shivani

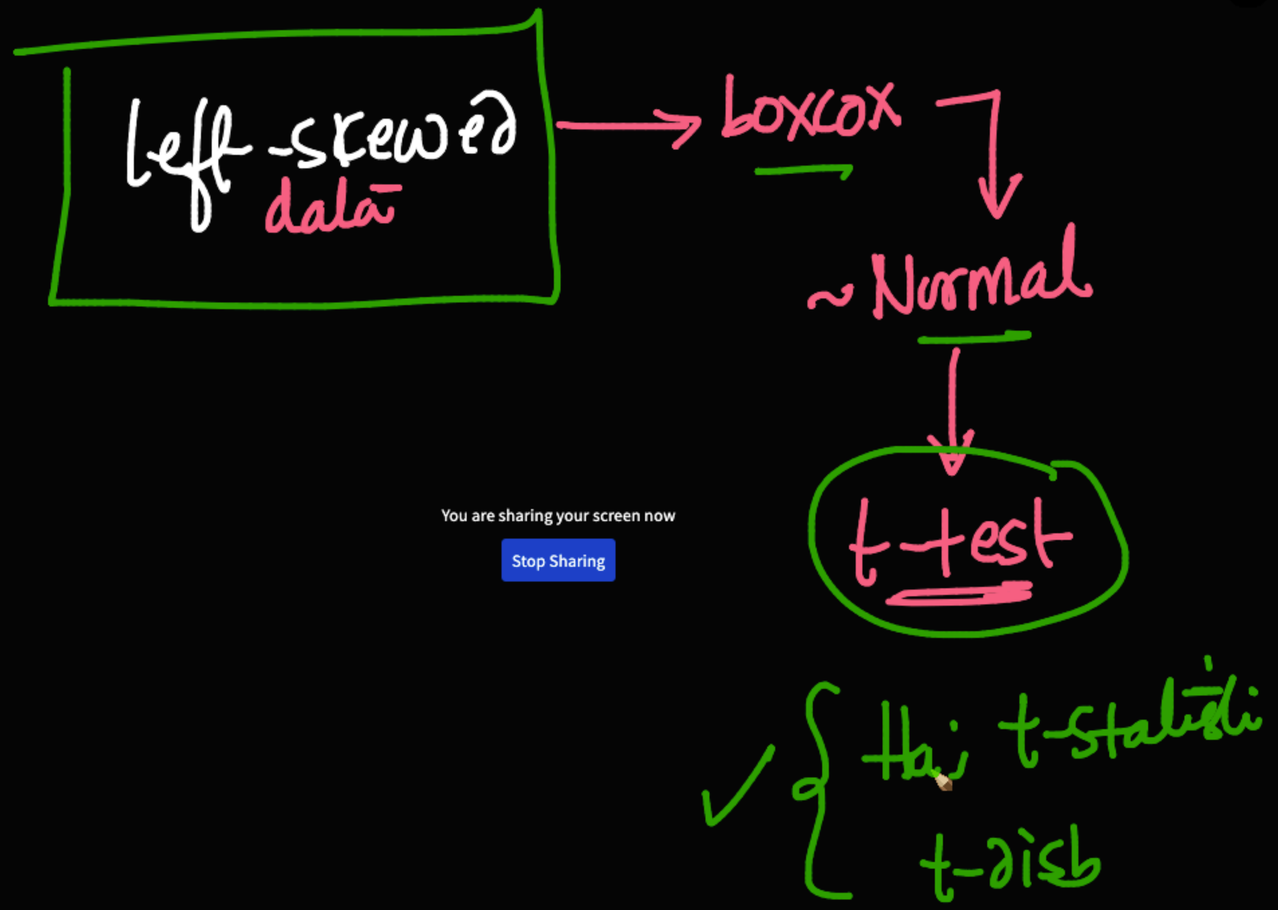
People 27 | Chat 1 | Questions | Start Doubt Session

03:03:48

ing your screen. Stop sharing Hide

67 / 68

DSML Advanced : Linear Algebra 1- Remedial | Lecture



Sameer Udgirkar

Srikanth Varma Chekuri (You) (Screen)

03:05:25

68 / 69

ing your screen. Stop sharing Hide

People

Search

- Srikanth Varma Chekuri (Host, You)
- Ankit Gupta
- Rishav Kumar
- Sameer Udgirkar
- Adarsh Vinayak
- Aditya Yadav
- Amit Srivastava
- Arijit Bhowmick
- Atin Gupta
- Avijit Das
- Gowtham
- MANISH PATEL
- Muthu kamalan .M
- Navroop
- Raghu Vamsi Yaram
- Rahul Maramreddy
- Ranita Bhattacharya
- Rishabh Singh

People 25 | Chat 1 | Questions | Start Doubt Session

You have left the meeting

We get frequent requests for your notes!

Notes written by you helps in understanding the topic better. You can upload the notes in two simple steps mentioned below

1. **Scan the QR code with your iPad**
   Scanner should be present in the top menu on your iPad
2. **Upload Notes on the generated link**
   All notes uploaded will be visible in the saved version of this session

OR

Drag and drop files or **click here to upload**

Files Uploaded from your computer appear here