Variable　　Random variable

$\boxed{x = 2}$

Tossing a coin

x | H | T |

Random variable → C.R.V. → D.R.V

R.V.　infinite

Y ~ Real

$x \rightarrow$ H | T

fin / finite

Dx　f_1　f_2　f_3

Data Sci

Dx　c-id　Income　city

C.R.V.　D.R.V

c-id	Income	city
1	14	T-1
2	15	T-2
3	14	T-1
4	11	
5	18	
-	17	

infinite

T-1
T-2
T-3

→ outliers

y : height of a students

Kano

{122.2, 146.4, 132.5 - - - - , 12.25}

cm

165

i.e outliers

⇓

Reason

⇓

collecting

mistake

"pre processing"

DE

outliers ☑

bad ☑

eg. income

{14, 16, 14, 13, 14, 15, 2140}

Rich

Age = {14, 15, 30, 27, 250}

X

old ☒

Population and Sample (imp)

eg. Task :- estimate Avg (ht) of human earth ?

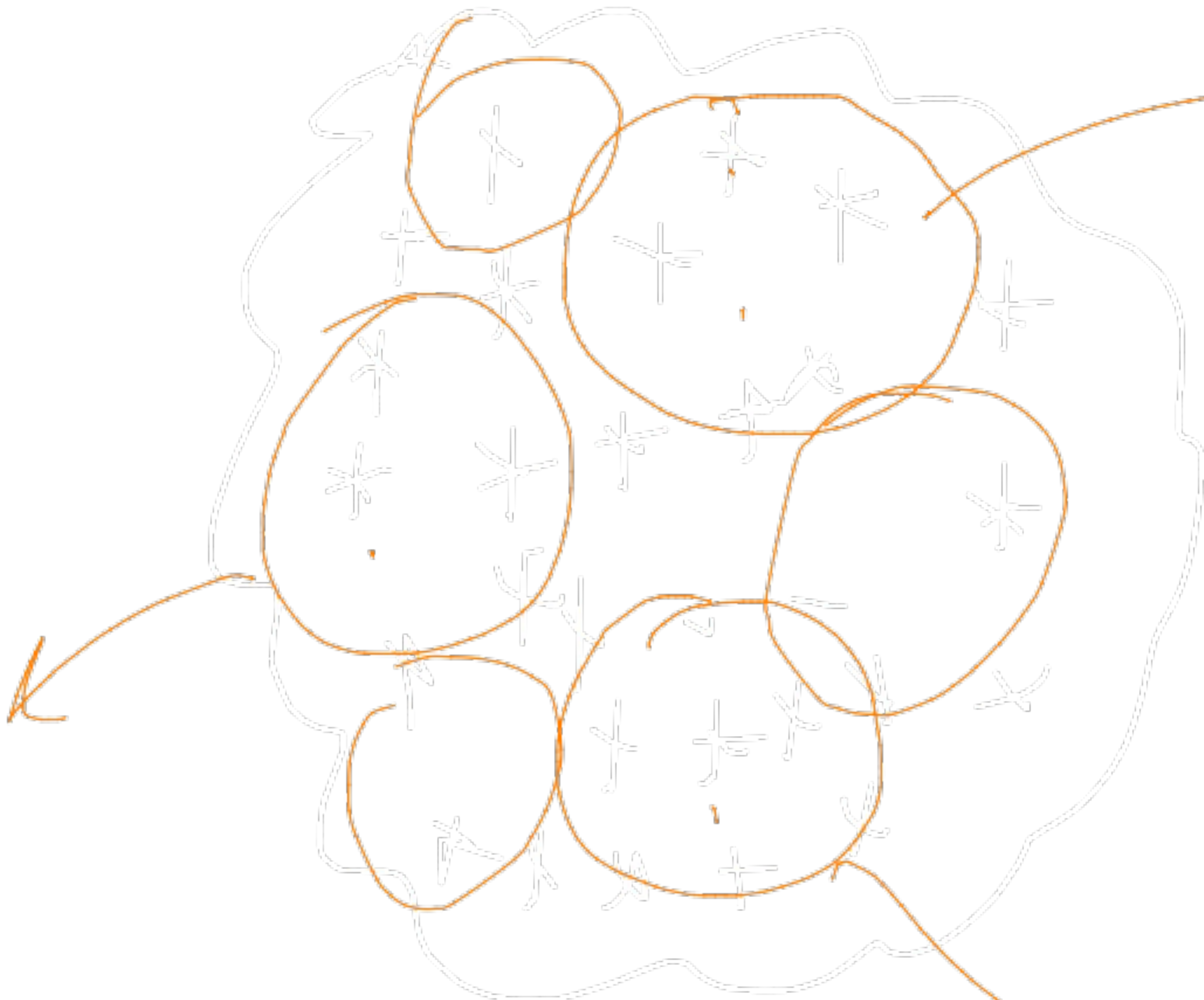
impossible

sample

small 5 ppl → [172, 162, 175, 162 160]

$$\Rightarrow \frac{172+162+175+162+160}{5}$$

$$\mu_p = \frac{\sum_{i=1}^{7B}}{7B}$$



100 ppl

$$S_1 = \text{Avg-}h = \frac{\sum_{i=1}^{n} h_i}{100} = \bar{X}_1$$

100 pple $\quad S_2 = \text{Avg-}h - \bar{X}_2$

$X_1 \quad X_2 \quad X_3 \quad \sim \mu$

Histogram → Analysis → univariate (1D)

plot

eg → $\boxed{\text{Age}}$

24
25
24
✓32
✓35
✓35
✓36
48
41
5
|

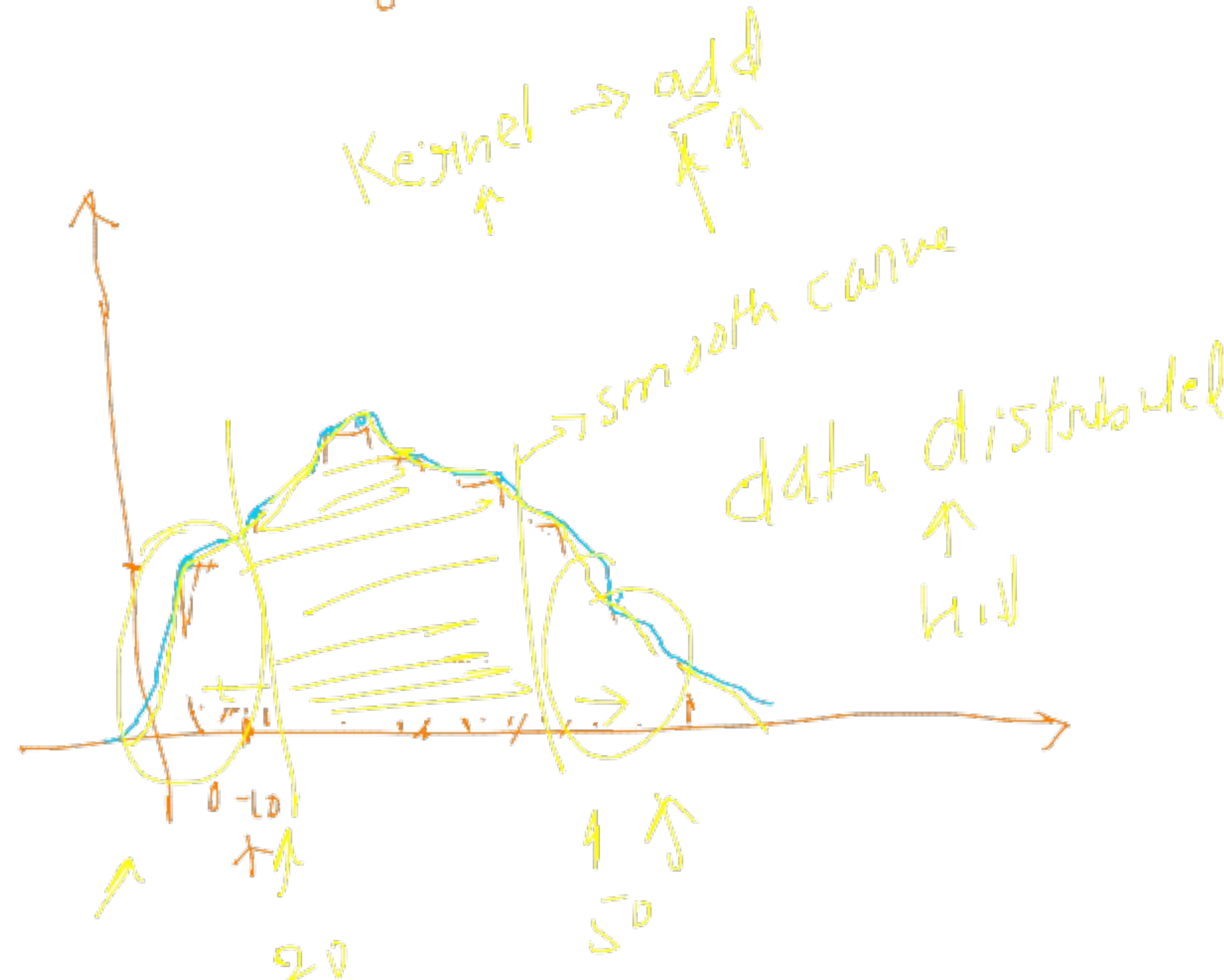
→ Bin

(10)

0 - 100 freq.

0-10 → 1
11-20 -
20-30 -
(31-40) → 4
40-50 -
50-60 -
60-70 -
70-80 -
80-90 -
90 above -

Age



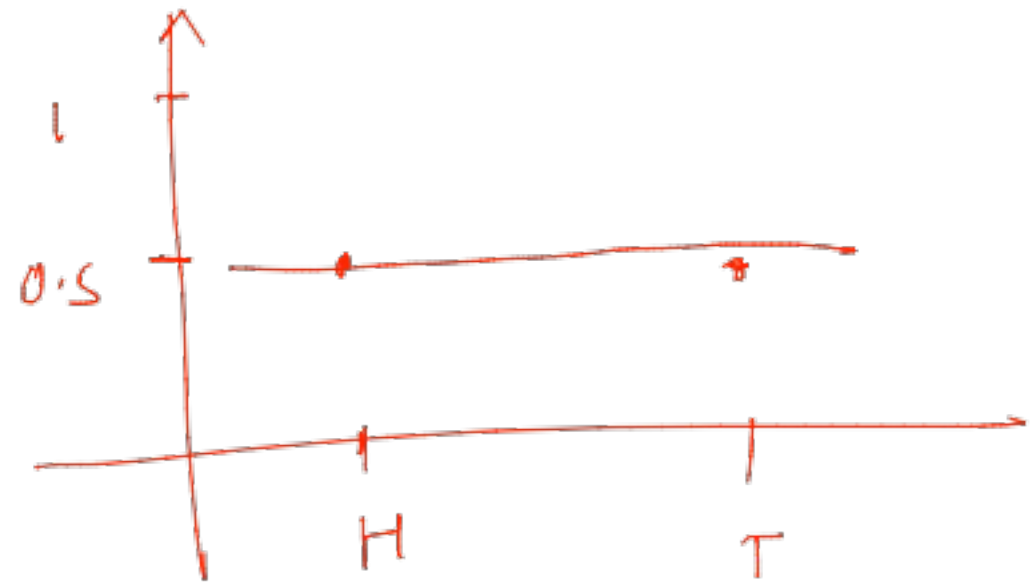
tossing a coin

D.R.V

P.P

possi

X	H	T
$P(X=x)$	$1/2$	$1/2$



p.m.f $\rightarrow$ $\sum P(X=x) = \boxed{1}$

3 times

{ HHH HHT HTH HTT

THH THT TTH TTT }

prob of getting head 3 times

atleast

X	0	1
$P(X=x)$	1/8	7/8

2 ⟶ 3

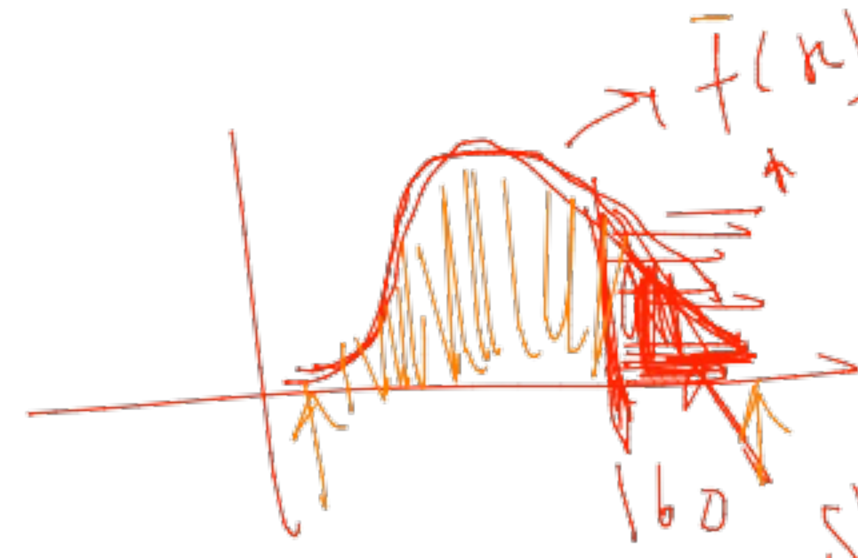
→ p.m.f ← 0, 1, 0

$\sum_{i=1} P(X=x) \to 1$

1

Y → Histogram

C.R.V. → "Data distributn" →



D.R.V

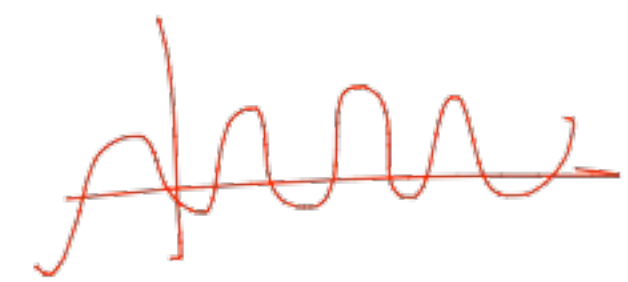
$P(h > 160) = ?$ easily

↑ score

$$\int_{160}^{\infty} f(x)\,dx \rightarrow 21\%$$

20%

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$



p.d.f.

C.R.V

mean

eg: income → 14, 15, 14, 13, 14, 15, 2000 outliers → effects

$$\text{Avg} = \frac{14 + 15 + 13 + 14 + 15 + 2000}{7}$$

160

$$\text{Avg} = 2075/7 \rightarrow \boxed{150}$$

150

me

Gov → policy

Avg –

1000 (100 - 500)

10 15

→ 150

$X = \{1, 1.1, 1.2, \boxed{1.4}, 1.6, 1.6, 1.8\}$ → eg.

$\text{median}(X) = \frac{n+1}{2} = \frac{7+1}{2} = 4$

odd elements

$\text{median}(X) = 1.4$ find

50%-tile value

np.median(X)

$X' = \{1, 1.1, 1.2, 1.4, 1.6, 1.8, 56\}$

→ median

# of elements → n

len(X)

8

$\left[\left(\frac{n}{2}\right), \left(\frac{n+2}{2}\right)\right] = \text{Odd}$

$\frac{8}{2}$, $\frac{8+2}{2}$

4th, 5th

$\left(\frac{1.4+1.6}{2}\right)$

$\boxed{1.5}$

median

↑

50%tile value

↑

e-commerce

↑

amazon → order

{10th, 20th, ...}

to you

↓

time

↓

delivery-time → [3, 4, 5, 3, 4, $\boxed{3}$, 4, 5, 6 - - -] 10K

time (day)

[- - -]

95%

↑ 3 days

50th

↑

median (delivery)

↑

80th percentile

↓

3

50 time

10th → 10

70th → $\boxed{20 \text{ days}}$

↑

time

on time → 3 days

↑

95th time

X

↑

sorts → median

↓

10M →

np.percentile(X, 10)

↑

value

'3' → 10% time → 3 day

~~7~~

median abs deviation

for detecting outliers

MAD

0.2

100

|14 - 14.2|

|15 - 14.21|

2.6

income

14

15

14

13

14

15

56

np.median(income) → $\boxed{14.2}$ (say)

Dx → income

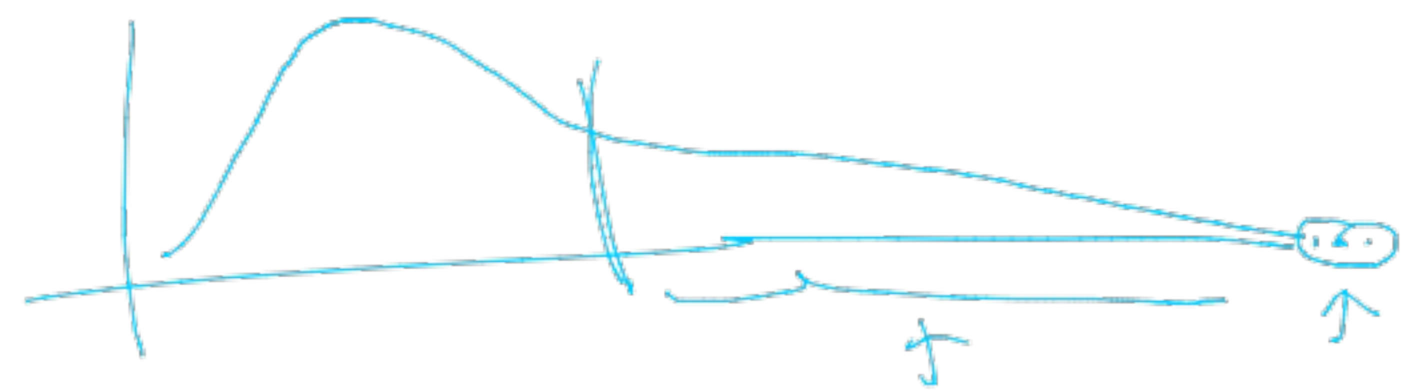
100k

→ Remove

Detect

method

plot it (dis



IQR

(inter quantile range)

↓

75% − 25th percentile

⇓

50% data

"Boxplot"

income

14

15

16

17

step

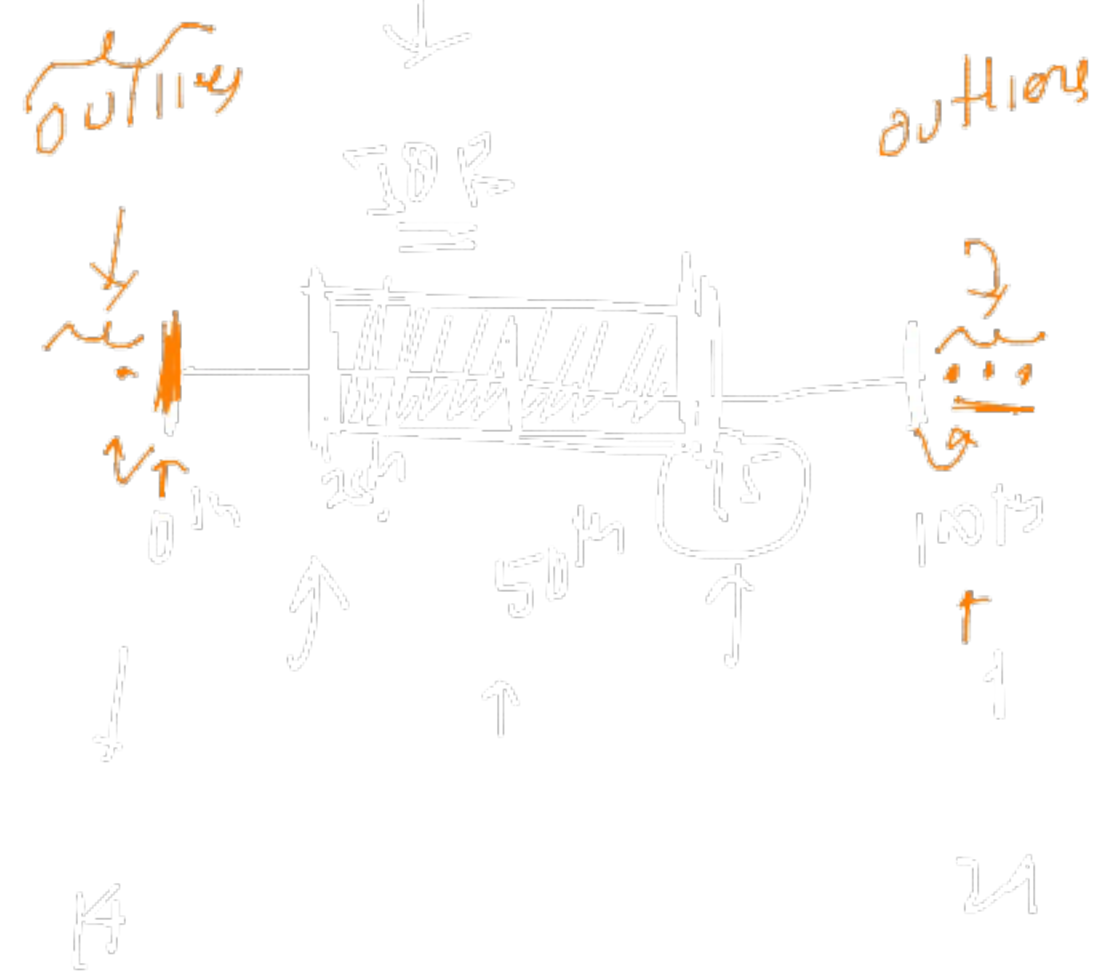
income

14

15

16

29



Sampling distribution → CLT

→ interview

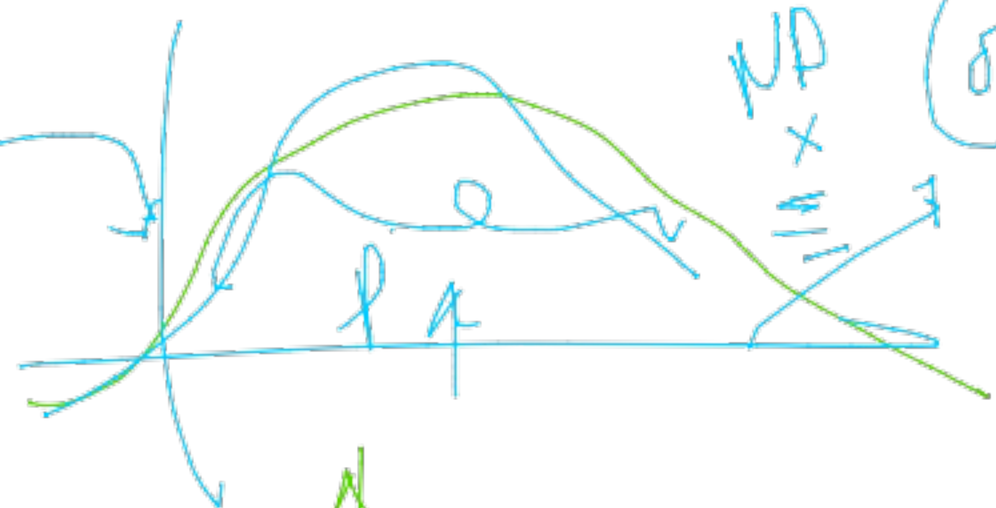
X

(pop.)

NP

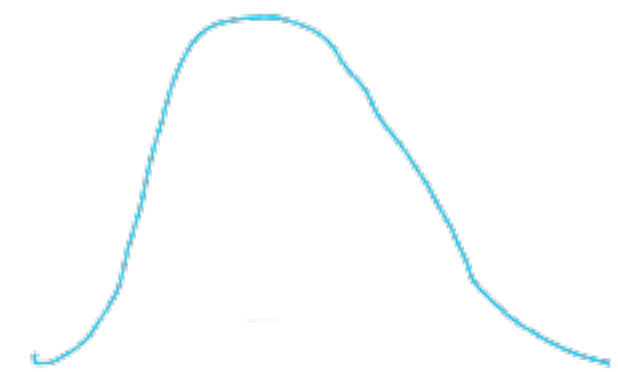
σ^2

μ



{Normal distribution

plot



μ

↓ random sample → size n

(but $n = 30$)

CLT

(30) $S_1 = \bar{x}_1$ (mean of S1)

$S_2 = \bar{x}_2$

$S_n \Rightarrow \bar{x}_n$

at the end → $\bar{x}_1 \quad \bar{x}_2 \quad \bar{x}_3 \quad - \quad --- \quad \bar{x}_n$

$\bar{x}_i \sim N\left(\mu, \dfrac{\sigma^2}{\sqrt{n-1}}\right)$ $\sqrt{n}$

varia

$n \to \infty$

population

$\left(\dfrac{\sigma^2}{\sqrt{n}}\right)$

diff!

X → ND →

income

plot → ND → tell

find

'k' → ND

w

X → ND

⇓

soln

Q-Q plot

Quantile-Quantile plot

⇓

(X) → ND

objective

⇓

client (X) : $X_1\ X_2\ X_3\ X_4$ ---- X_{500}

ND

Step1 Sort X_i and compute percentile

⇓

[x_1' x_2' x_3' x_4' ---- x_{500}']

⇓

percentile

[x_1' x_{11}' ---------- x_{500}]

1 2 100

ND

Step2 $Y \sim N(0,1)$

numpy

($Y_1\ Y_2\ Y_3\ Y_4$ --- Y_{500})

⇓

sort

⇓

$Y_1'\ Y_2'\ Y_3'$ --- Y_{500}'

⇓

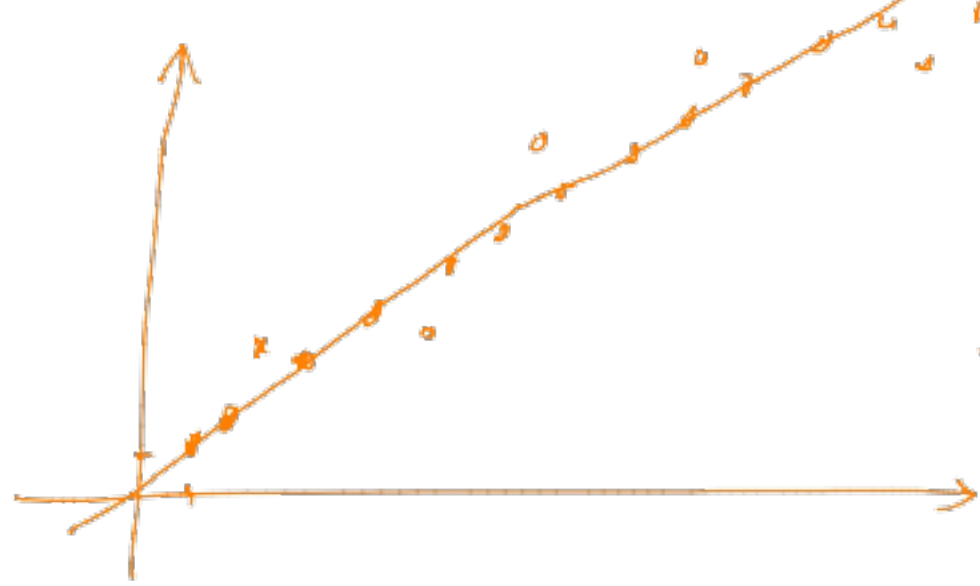
percentile

$Y_1'\ Y_2^2\ Y^3$ --- Y_{500}'

(case-i) →

X

45°

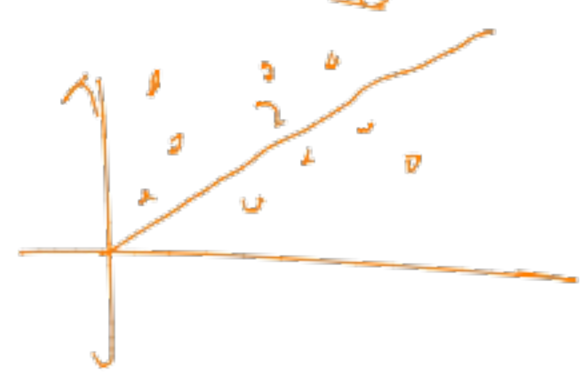


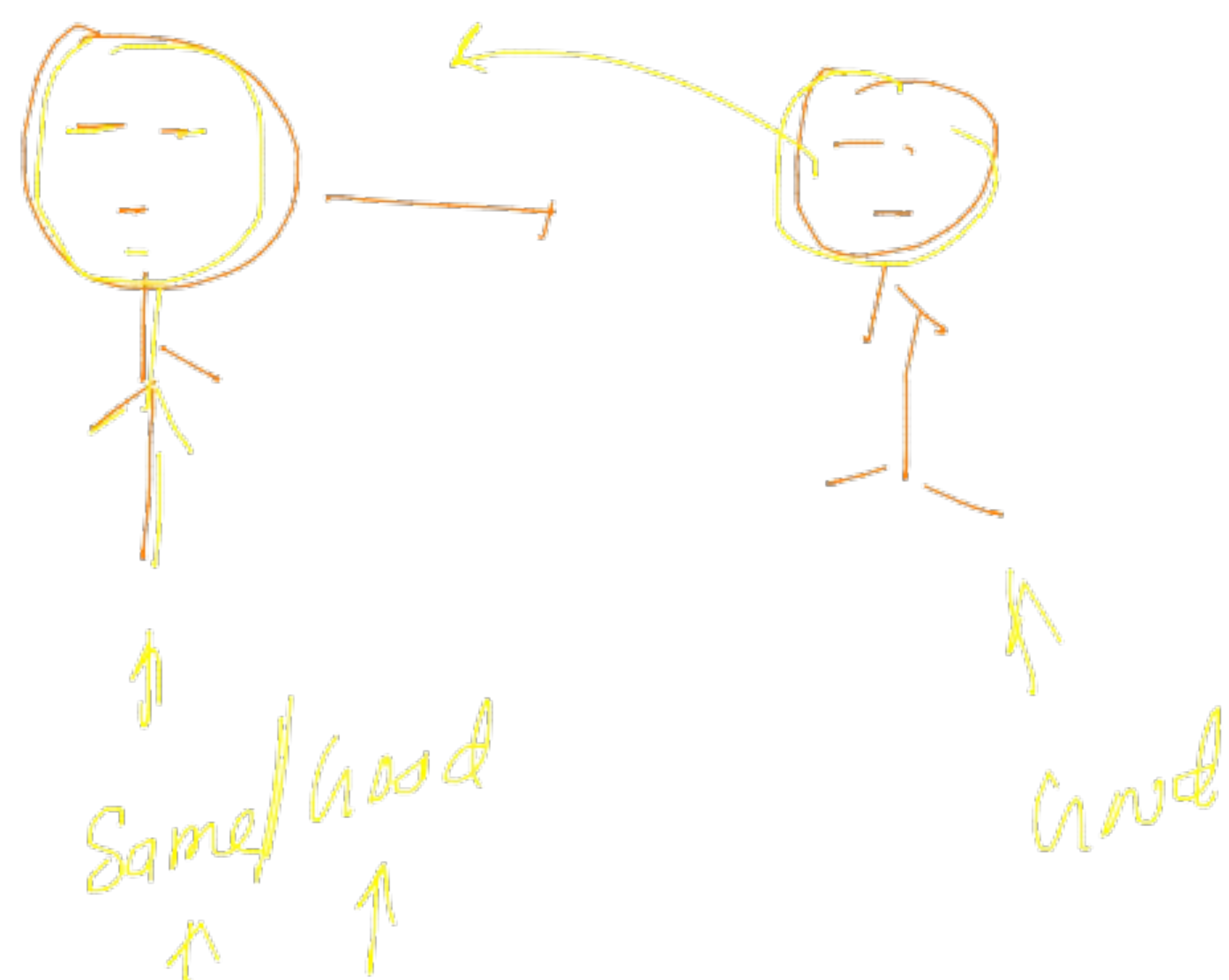
Y

Y

X

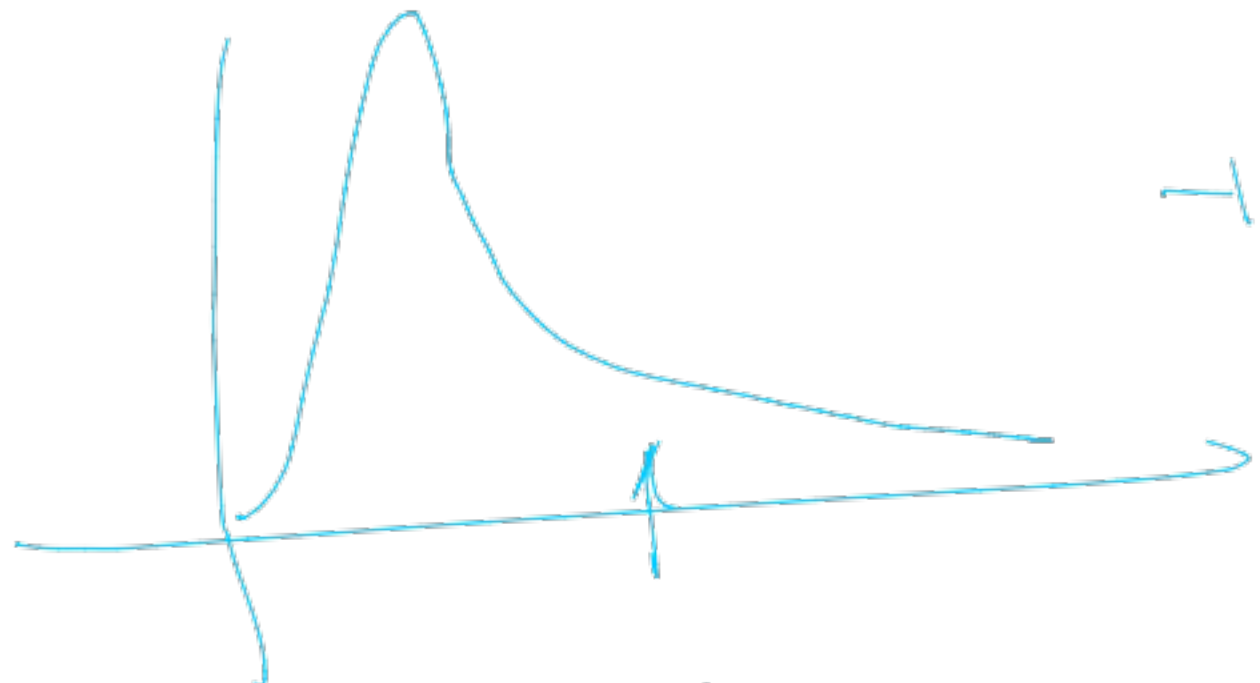
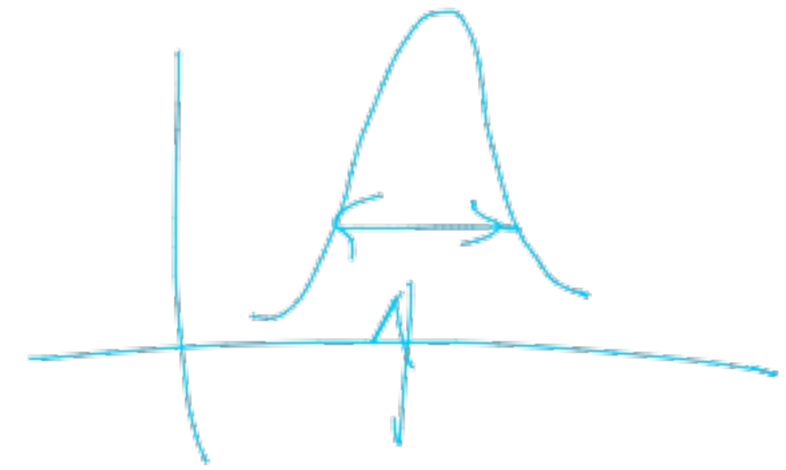
$X \sim$ ND




Same/Good
Good

$N\left(\mu, \frac{\sigma^2}{\sqrt{n}}\right)$

$\sqrt{\text{variance}}$

 $\rightarrow$ 

variance

$\downarrow$

diff'n

$\sqrt{\downarrow}$

$=$