Wrap up: CLT + CI

↳ Gaussian, know ↳ dataset "Bootstrap"

Hypothesis testing (2 to 3)

↳ tricky

↳ interview "p-value" → later

↳ transf → Normal

Recap:

$X \rightarrow E[X] = \mu, \quad Var[X] = \sigma^2$

→ CLT: Sum of $X_1 + X_2 + \cdots X_n$ → Behaviour

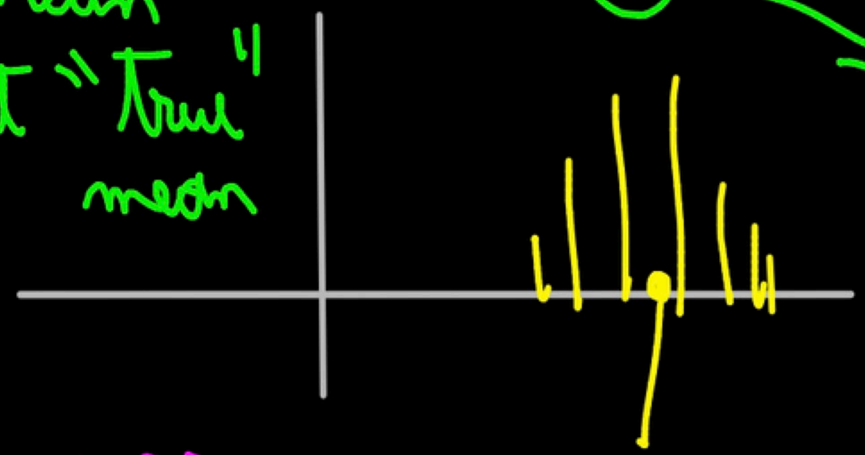
Mean: $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ → Behaviour "Gaussian"

$\bar{X}$ → sample mean

Convert "True" mean

"Stand Gaussian"

n → Sample



$$Z = \frac{\bar{X} - E[\bar{X}]}{\sqrt{Var[\bar{X}]}} \rightarrow E[Z] = 0, \quad Var[Z] = 1$$

$$E[\bar{X}] = E[X] = \mu$$

$$Var[\bar{X}] = \frac{Var[X]}{n} = \frac{\sigma^2}{n}$$

Election

Whole country pop.

$\bar{x} = 52\%$

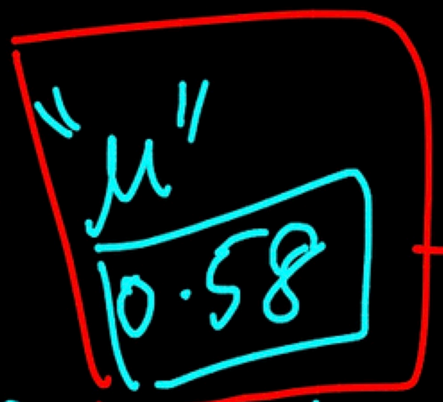
"A" Candidate

Sample "$\bar{x}$" → 47%

↓

fraction of the pop. that support him?

"polls"



58%

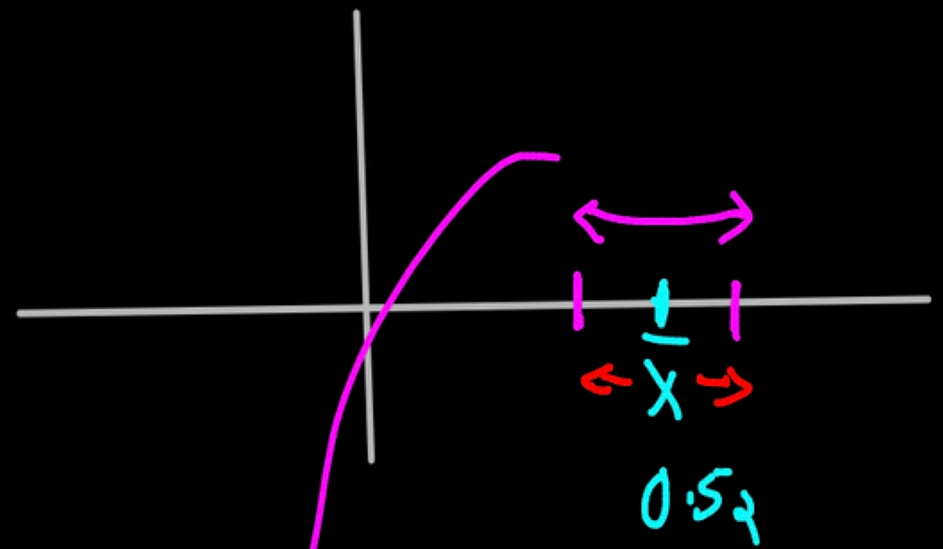
→ comment about this using 52% or 47%

We got sample : $\bar{x}$

"Population" mean : μ

Goal is to say : $\mu \in$ interval

95%.



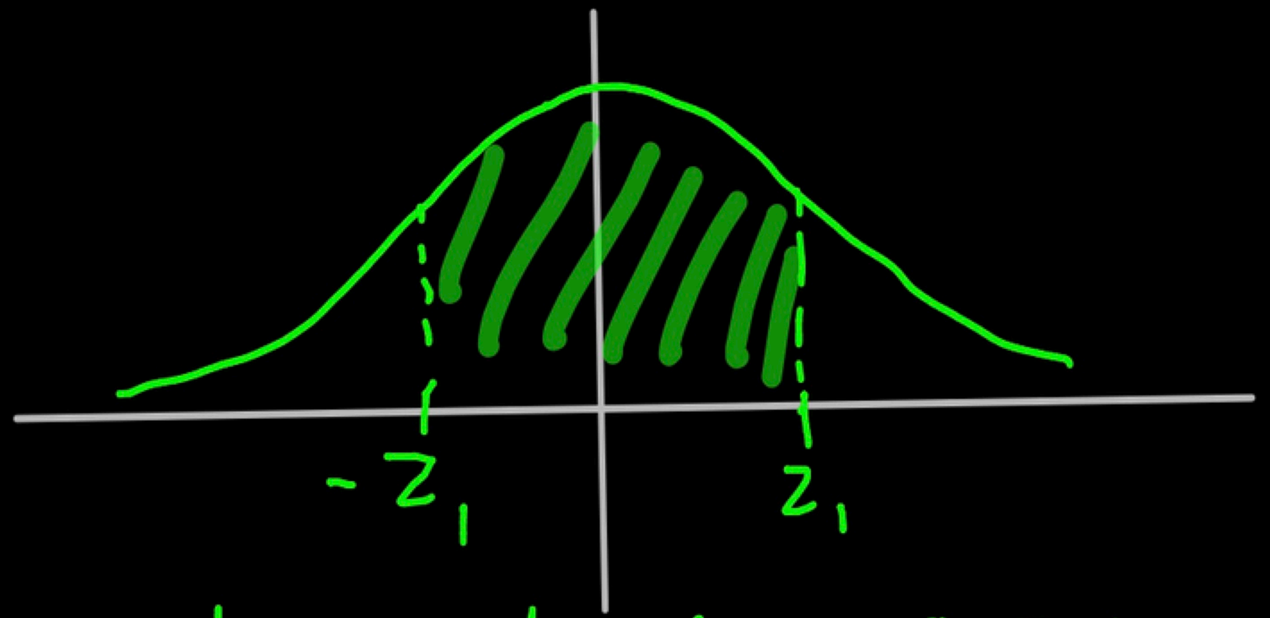
interval

$\bar{x}$ lies in the middle of the interval

$$\mu \in \left[\bar{x} - \frac{1.96\sigma}{\sqrt{n}},\ \bar{x} + \frac{1.96\sigma}{\sqrt{n}} \right]$$

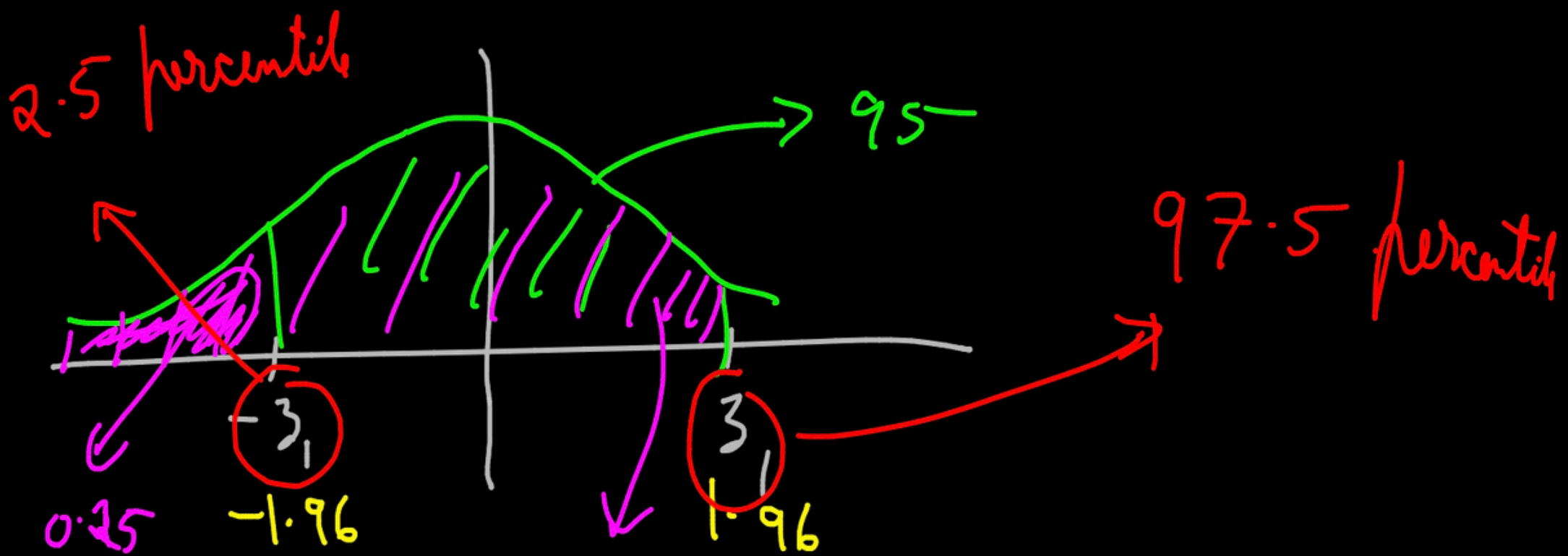
If "σ" is known

What if σ is not known: "Bootstrap"



z_1 : That value such that 95-1. area is in $[-z_1, z_1]$

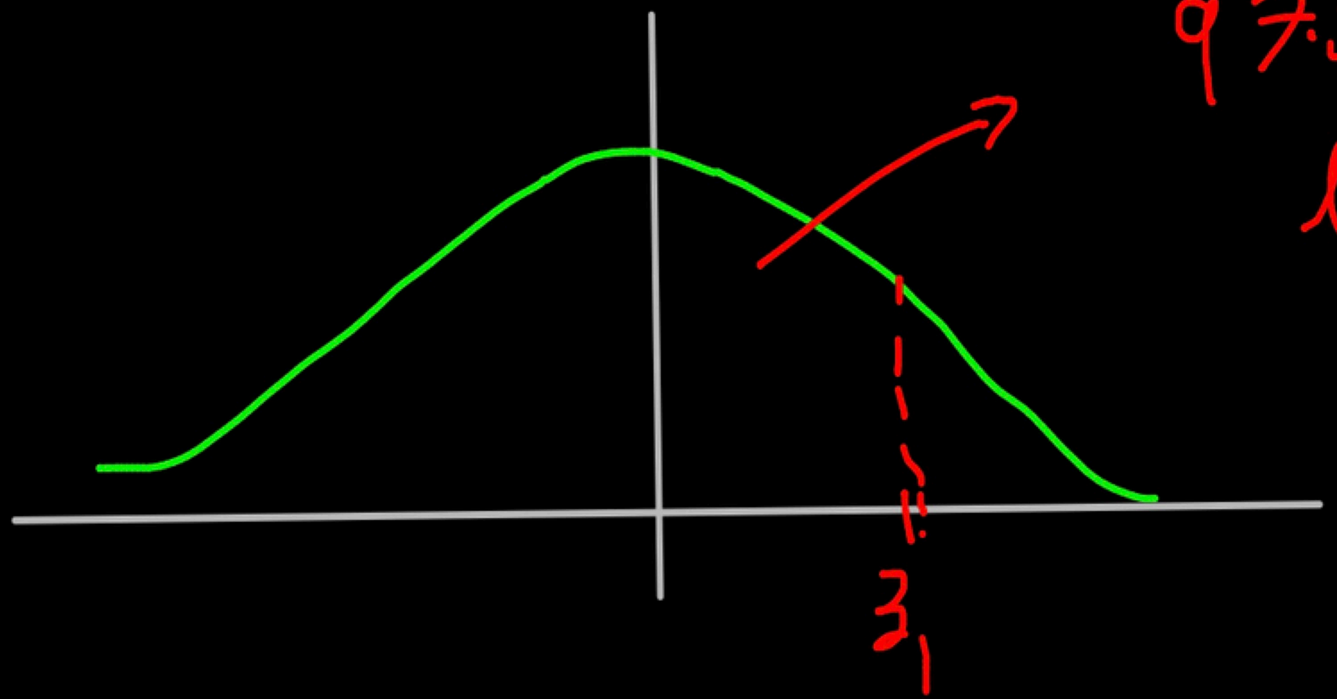
norm.ppf (0.975)



$$95 + 0.25 = 0.975$$

z_1 is that value such that $\text{norm.cdf}(z_1) = 0.975$

$$z_1 = \text{norm.ppf}(0.975)$$



97.5% of values are below 3,

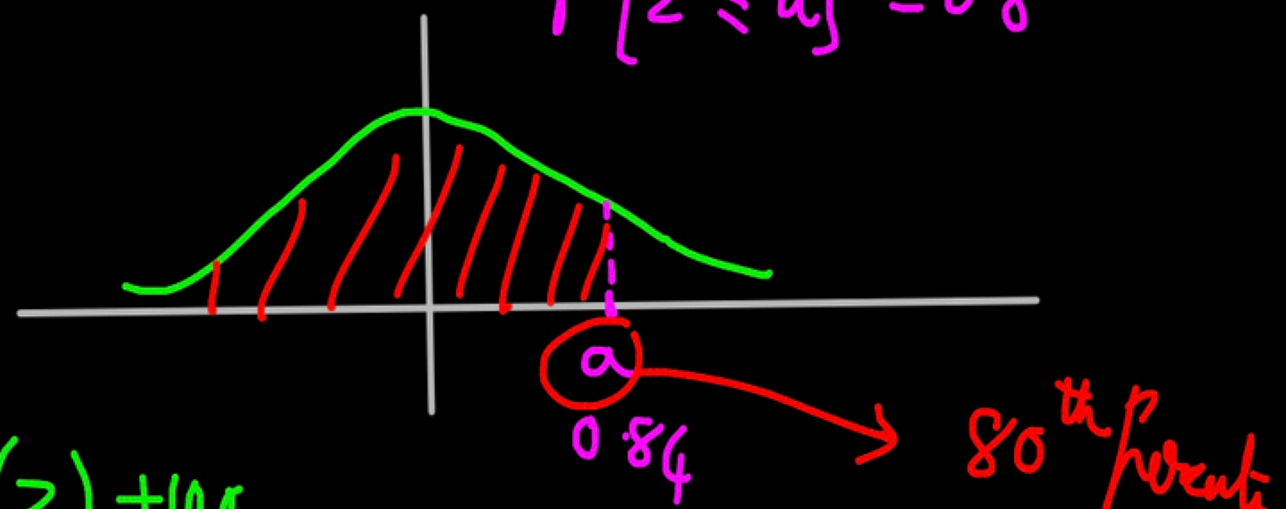
IQ Scores

$\mu = 100$, $\sigma = 15$ Var $\rightarrow 15^2$

what should a person IQ be such that he is in the top 20%.

$P[z \leq a] = 0.8$

$X \sim N(\mu=100, \sigma=15)$



$z = \frac{X - 100}{15} \rightarrow X = 15(z) + 100$

How do we find a? norm.cdf(a) = 0.8 $\rightarrow$ norm.ppf(0.8) = a

$a = 0.84$ $15(0.84) + 100 = 112.6$ $\rightarrow$ percentile cut-off

80th percentile has two meaning

Distribution

norm.ppf

Sample

Use the dataset

np.percentile

95% Confidence

→ 2.5 percentile "Sample"

→ 97.5 percentile

norm.coff(0.975)
= 1.96

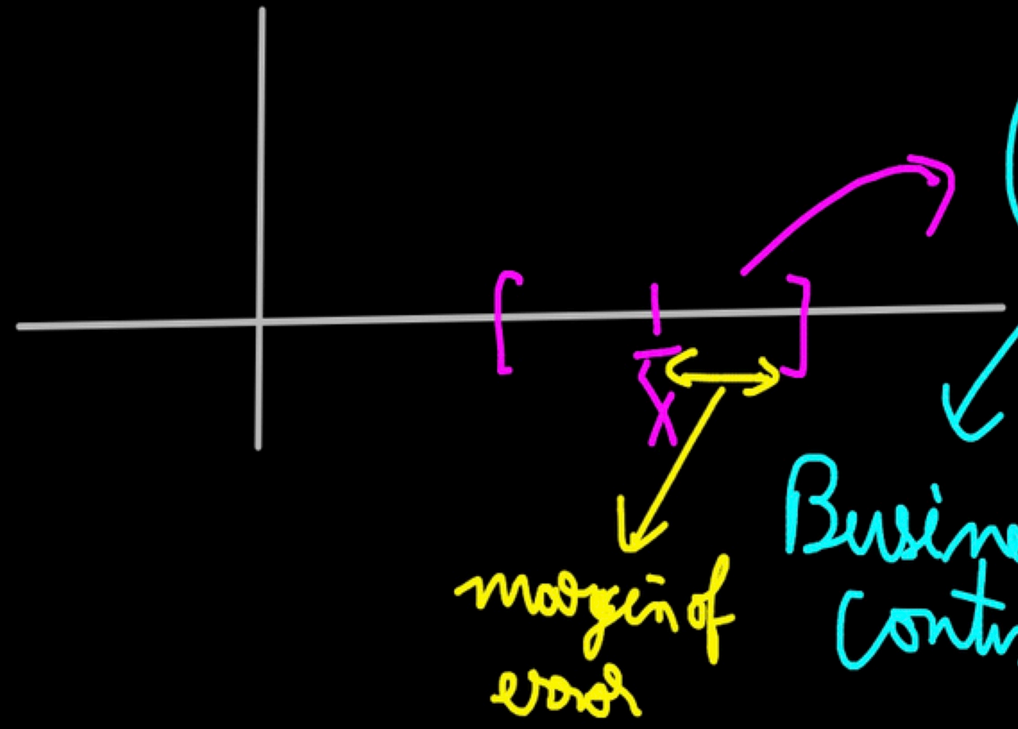
$n = 10,000$

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

→ n large ~ Gaussian (CLT)

Sample mean

How to compute conf. interval? for true mean!



10,000 samples : 97.5 percentile & of the sample mean
"row" 2.5 percentile

→ "Bootstrap"

Use the 10,000 samples → sample with replacement

One iteration
1) pick one row randomly → 17th row
2) pick one row randomly → 585th row
in a loop
k
k ≤ 10,000

soon will repeat

Every sample → "new" dataset

1st set of sample → $\bar{X}_1$

2nd set of sample → $\bar{X}_2$

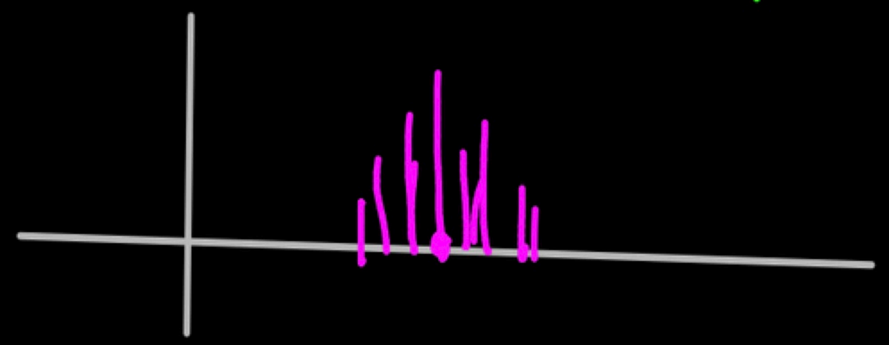
.

.

.

100th → $\bar{X}_{100}$

} Histogram → looks like a gaussian



Dataset 1

$\mu = 111$

[110, 110, 111, 109, 110, 111, 115]

for → Sample with replacement → sample

bootstrapped means → $\bar{X}_1$, $\bar{X}_2$ → "long version"

= smaller CI "

Dataset 2

$\mu = 100$

[102, 97, 111, 98, ...]

→ $\bar{X}_1$, $\bar{X}_2$ } more variance → percentile → "longer CI"

[1, 2, 3, 4, 5, 6]

→ sample with replace 6 times

chosen

[3 4 3, 5, 1, 2]

→ sample without replace 6 times

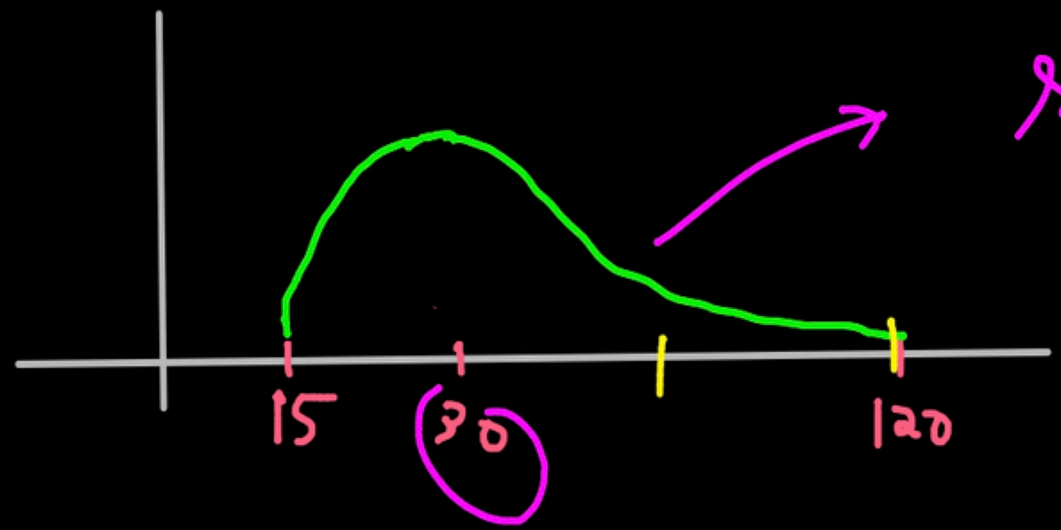
[3, 1, 4, 2, 5, 6]

chosen do not "replace"

same sample more everytime

Swiggy delivery

30 min → 100 min



skewed

X_1, X_2, X_3 → data

$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$

Close to μ

$\mu \in [27, 33]$ → 95% CI

↳ Publicize

μ

1) Conf. Int for max
2) Conf Int for 90th percentile

[45 - 55] CI 90th percen

$(45-55] \rightarrow 90^{th}$ percentile

Only 10% of times will you have to wait more 55 mins.

median

Earlier [ ] → mean

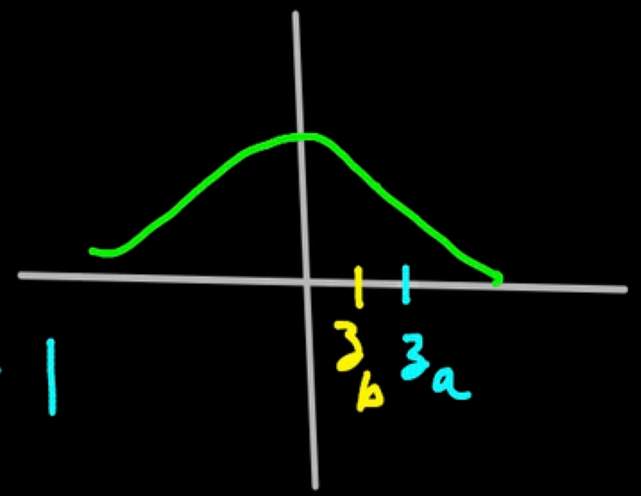
Now [ ] → 90^{th} percentile

Two types of Exams - Range of marks

$$z = \frac{x - \mu}{\sigma}$$

different range

	Test 1	Test 2
mean	1500	21
std dw	300	5



A: test 1 → 1800 → $z_a = \frac{1800 - 1500}{300} = 1$

B: test 2 → 24 → $z_b = \frac{24 - 21}{5} = \frac{3}{5}$

z-score

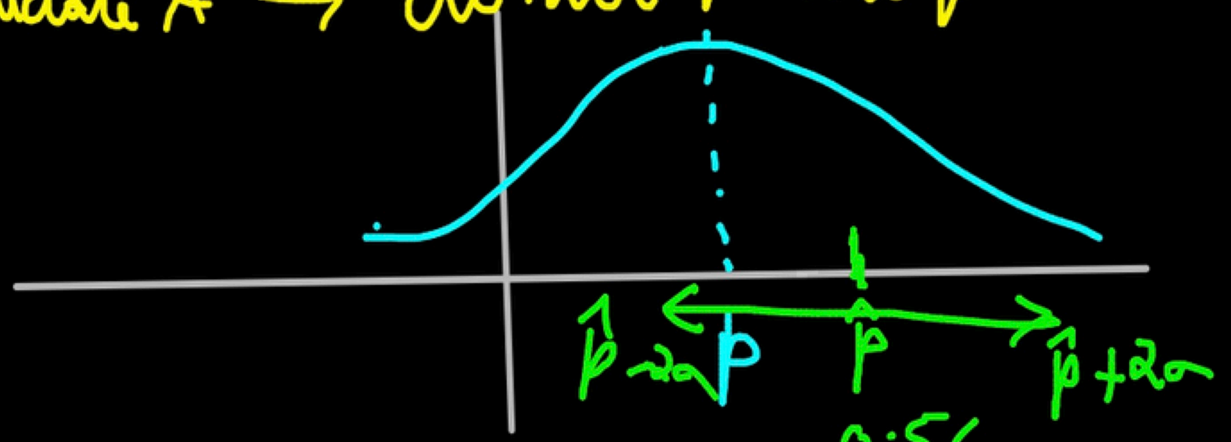
Election

: proportion of support for candidate "A"

Population → "p" for candidate A → do not know p

$\hat{p} = 54\%$

$\hat{p} = 58\%$



Standard Error: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

↳ Skip the proof

$[\hat{p} - 2\,SE \;,\; \hat{p} + 2\,SE]$

Eg: $[52, 56]$

p lies in CI around $\hat{p}$ with 95% Conf.

Hypothesis testing → tricky

Coin: Cricket captain ⇒ "Heads" always

1) 10-match series
Won 7 tosses
not suspicious

2) 100-match series
Won 70 times
not so sure?

3) 1000-match series
Won 700 times
very suspicious

Is there math to guide this?

))(((10 - match 7 wins)))

Invite 1000 people to toss 10 times

→ How many got ≥ 7 heads

"n"

→ p - value : $\frac{n}{1000}$

→ ~ 160 out of 1000 0.16 16%

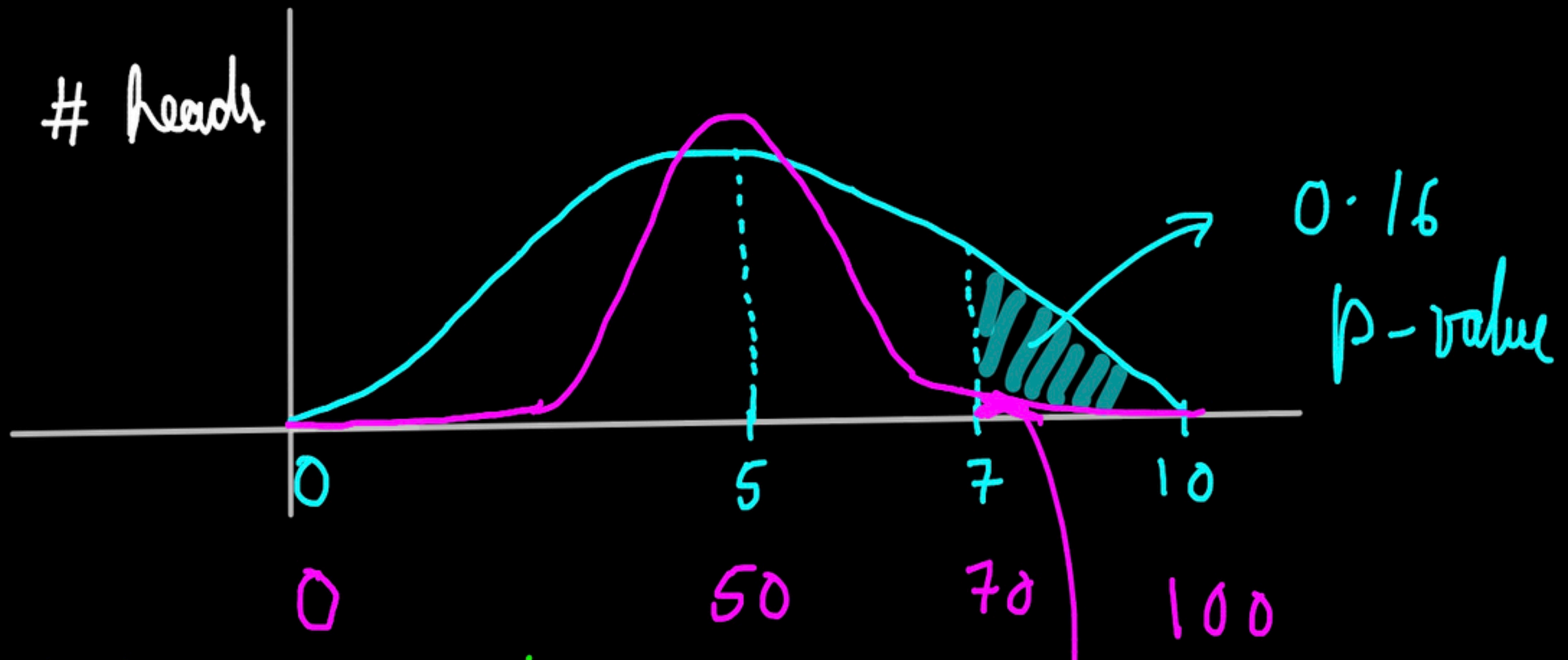
2) (100 matches - 70 heads)

1000 → toss 100 times

→ # people ≥ 70 heads → n

$\frac{n}{1000}$ p-value

"Under Null hypothesis" → fair coin



p-value ≤ 0·05 → significance level

strong evidence to reject the H_0

very very small p-value

"Coin is fair" → Null Hypothesis (H_0)

"Coin is biased towards Heads" → alternate Hypothesis (H_a)

"Reject" null hypothesis?

We reject only if we have a very strong evidence

Court Judge: Crime has happened. There is a suspect.

What is the null hypothesis made by the judge regarding the suspect?

H_0: "Innocent until strongly proven guilty"

Reject only if you have strong evidence

Cricket: 3rd umpire DRS

H_0: On-field umpire is correct

Reject if we have strong evidence

CLT $P\left[\bar{X} - \left(\frac{\sigma}{\sqrt{n}}\right), \bar{X} + (\ )\right]$

$\bar{X}$ ↓ sample mean only

median, any percentile → bootstrap