Confidence Interval

Date :
Page No.

distb ← X: height

$\{x_1, x_2, x_3 ---- x_{10}\}$ random sample from X of size 10

Pms estimate population mean of $X = \mu$

$\mu \approx \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ } simpl avg

pop mean

sample mean

as $n \uparrow$, $(\bar{x} \to \mu)$

$\left\{\begin{matrix} 180, 162, 158, 172, 168 \\ 150, 171, 183, 165, 176 \end{matrix}\right\}$ → heights

Point estimate of $\mu = \frac{1}{10}\sum_{i=1}^{10} x_i = \underline{168.5 cm}$

→ $\mu \in \underline{[162.1, 174.9]}$ with $\underline{95\%}$ probability

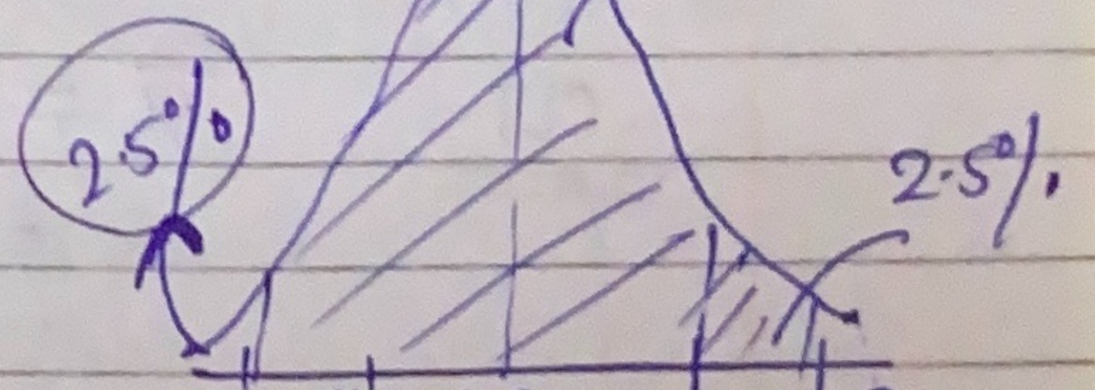
interval

confidence

Computing C.I given underlying dist

Heights ← $X \sim N(\mu, \sigma)$ 1-c/2

Let $\begin{pmatrix} \mu = 168\,cm \\ \sigma = 5\,cm \end{pmatrix}$



Ques C.I of heights

$(\mu - 2\sigma, \mu + 2\sigma)$ ↺ 95% of obs lies in

$\frac{1-0.95}{2} = 2.5$

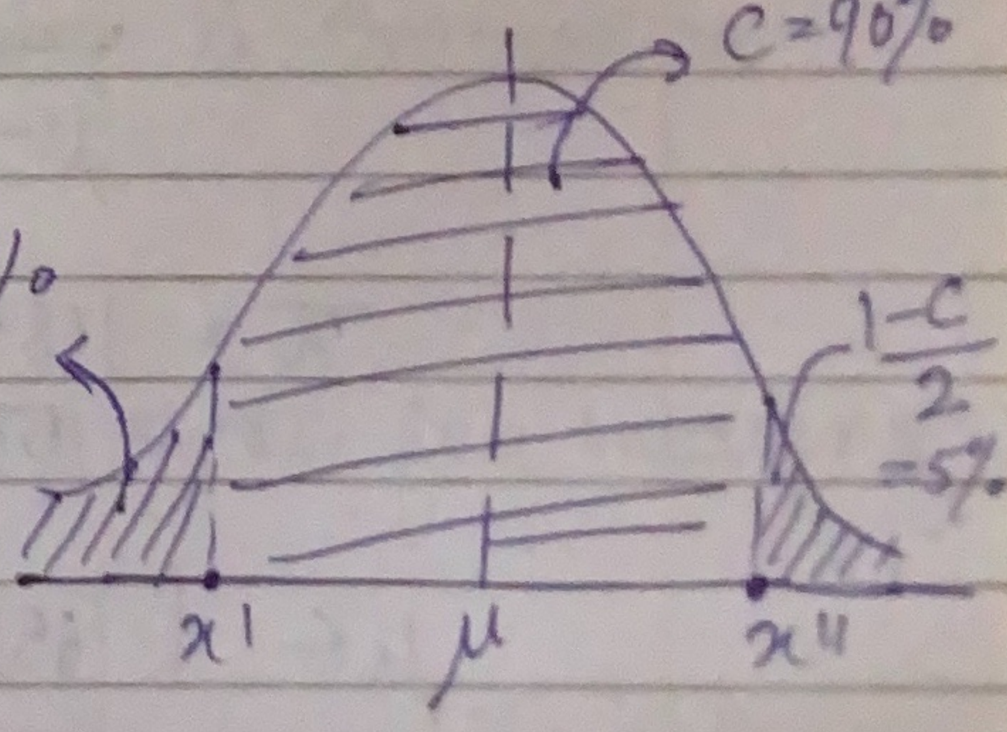
↓

(158, 178) with 95% prob

C.I

lie in [x_I , x_{II}] with 90% confidence 5%

↑ lower bound , ↑ upper bound



C.I for mean (μ) of a size

$x \sim F$ with pop mean of μ & std dev σ

↓

$\{x_1, x_2 \text{---} x_{10}\}$ → sample of size = n=10

{180, 162, 158, 172, 168,
150, 171, 183, 165, 176}

$\bar{x} = 168.5\,cm$

Date :
Page No.

Q What is 95% C.I of μ.

Case ① $\sigma = 5cm$ {the know population Std dev}

CLT $\bar{x}$ = Sample mean

$$= \frac{1}{10} \sum_{i=1}^{10} x_i$$

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \longrightarrow \text{(CLT)}$$

$\bar{x}$ → Sample mean; N → gaussian; μ → pop mean; $\frac{\sigma}{\sqrt{n}}$ → pop std dev / $\sqrt{n}$

$$\mu \in \left[\bar{x} - \frac{2\sigma}{\sqrt{n}}, \bar{x} + \frac{2\sigma}{\sqrt{n}}\right] \text{ with 95\% confidence}$$

$\bar{x} - \frac{2\sigma}{\sqrt{n}} = \mu - 2\sigma$; $\bar{x} + \frac{2\sigma}{\sqrt{n}} = \mu + 2\sigma$

$\bar{x} = 168.5$ $\quad \sqrt{n} = \sqrt{10}$

$\sigma = 5cm$

$\mu \in [165.34, 171.66]$ with 95% confidence

Case ② If we don't know σ (pop std dev)

Sample (n)

t-distb → Student's t-distb

$\bar{x} \sim t(n-1)$

$\bar{x}$ → Sample mean ; $(n-1)$ → degree of freedom

Date :
Page No.

Estimate C.I of μ of r.v

↳ Case I: σ is known CLT; $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

↳ Case II: σ is unknown t-dist$(n-1)$

(Q) estimate C.I of σ of a r.v
median
90th percentile

(Computer) → (bootstrap C.I)

C.I using empirical bootstrap → computational

→ C.I for median, var, std dev, 90th percentile

→ computer → programming & simulation
↳ 60-70 yrs.

→ $X \sim F$ task: estimate 95% CI for median of X

X:

Sample of size n: $S = \{x_1, x_2, \cdots x_n\}$ $x=10$

↓
using only this
↳ CI of median of X

S_1: $x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \cdots x_m^{(1)}$ $m \leq n$

↳ random sample of size m

↓
Sampling with replacement from S

$S = \{x_1, x_2, \ldots x_n\}$

↓ using sampling with replacement

bootstrap samples

S_1 : $(x_1^{(1)}, x_2^{(1)}, \ldots x_m^{(1)}) \rightarrow m_1 \rightarrow$ median of S_1

S_2 : $x_1^{(2)}, x_2^{(2)}, \ldots x_m^{(2)} \rightarrow m_2$

S_3 :

S_k : $x_1^{(k)}, x_2^{(k)} \ldots x_m^k \rightarrow m_k$

$m_1, m_2, \ldots m_{1000}$ ← 1000 medians generated using bootstrap samples

↓ sort

$m_1' \leq m_2' \leq m_3' \leq \ldots \leq m_{1000}'$

↓ CI (95%)

(25) m'_{25} (950) m'_{975} (25)

95% CI of median of x is

$[m'_{25}, m'_{975}]$

Non-parametric technique

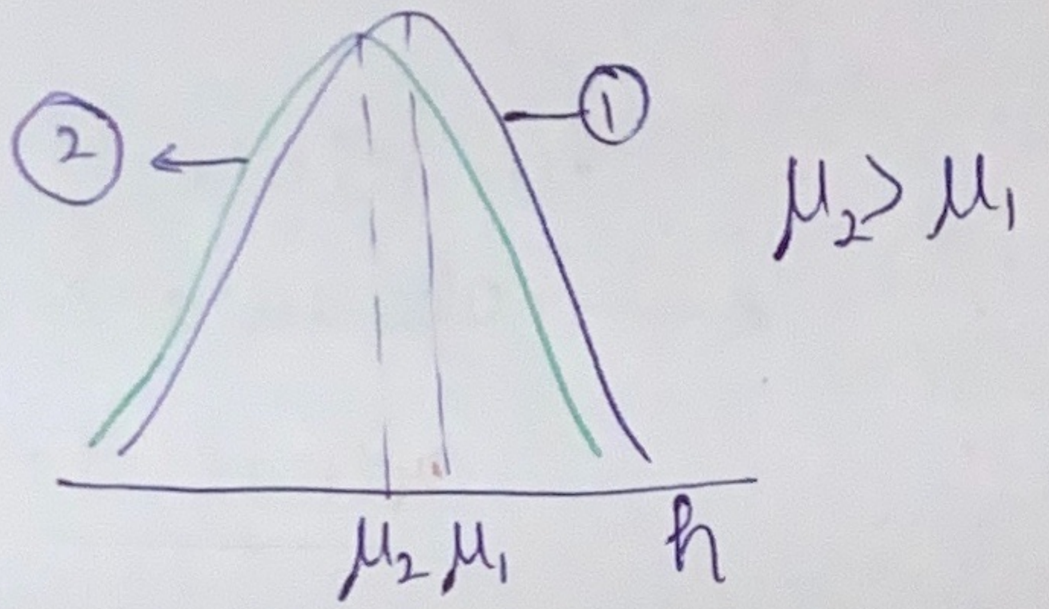
↑ doesn't make any assumptions about

Probability And Statistics

Hypothesis Testing

(Q) Is there a difference in heights of students in Cl_1 and Cl_2

	Cl_1	Cl_2
1	160	162
2	152	156



① Choosing a test - Statistic

$$x = (\mu_2 - \mu_1)$$

$$x = 10cm$$

② Null hypothesis (H_0) (Proof By Contradiction)

H_0: no difference in μ_1 & μ_2

Alternative hypothesis (H_1): diff in μ_1 & μ_2

③ p-value : probability of observing $(\mu_2 - \mu_1)$ if null hypothesis is true

assume H_0 is true.

Cl_1	Cl_2
(50)	(50)

if p-value = 0.9

$\Rightarrow$ prob of 10cm is 0.9 if H_0 is true

reject H_0 occupt H_1 ← if pvalue is 0.05 → 5% chance that diff is 10cm if H_0 is true

Hypothesis Testing & Permutation Test / Resampling

↳ Same example

→ stressing on important points

① Cl_1 Cl_2

Step 0

50 50

μ_1 μ_2

← we already made observation

Actual data

Sample size = 50

Δ / X: $\mu_2 - \mu_1 = 10cm$ ← Observation

② What is prob of observing a value of $X \geq 10cm$ if there was no difference in class heights ?

prob($X \geq 10cm$ | H_0)

obs → null hypothesis

↳ p-value

if p-value is small = 0.01 or 1%

↳ < 5% → customary

↓

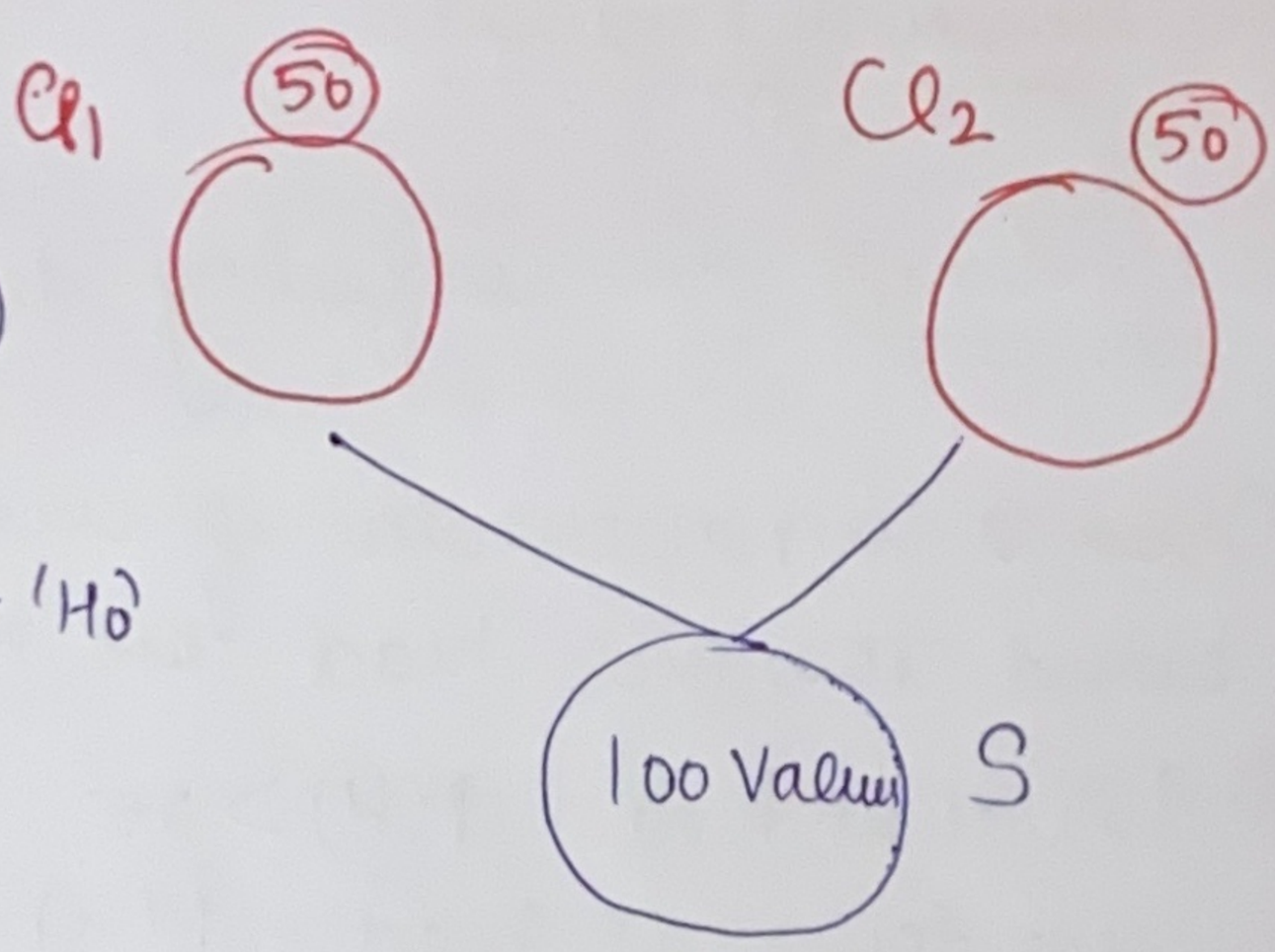
$p(X \geq 10cm | H_0) = 0.01 / 1\%$

↓

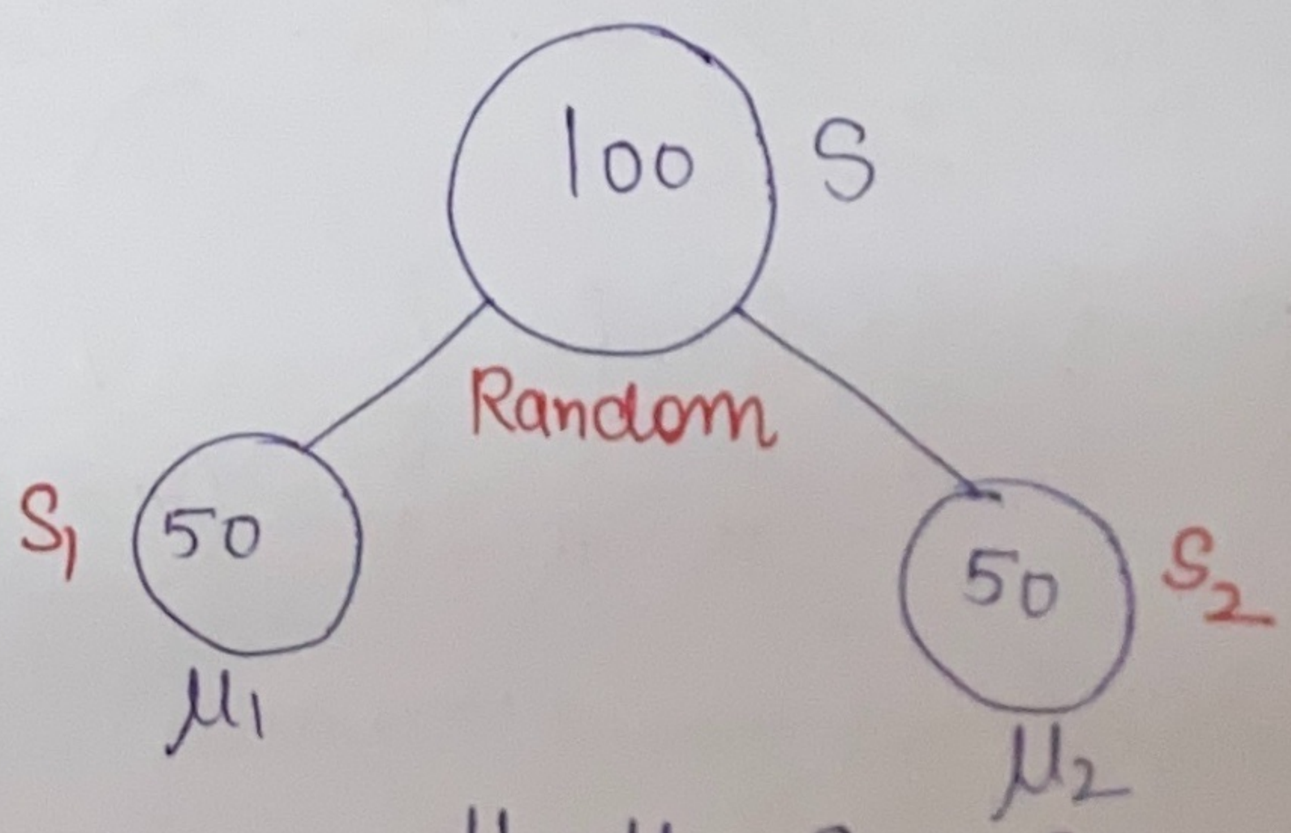
$x = 10cm$ or $x \geq 10cm$ ← True

→ H_0 is less probable or H_0 may not be true

pValue



if H_0 was true → no diff in class heights



$\mu_2 - \mu_1 = \delta_1 =$ Simulated diff in class heights with sample size of 50 & H_0 true

$\delta_1, \delta_2, \delta_3, ---- \delta_{10K}$ ← Simulated diffs with H_0 being true

↓ Sort

$\left(\delta_1', \delta_2', \delta_3' - 10cm \; \delta_{10K}'\right)$ (8K ←, 2K →)

$p(x \geq 10cm | H_0) = \frac{2000}{} = 20\%$

$S' \subseteq S_2' \subseteq S_3' \subseteq \cdots \subseteq 10cm \subseteq S_{10K}'$

(9900 over $S' \subseteq S_2'$; 100 over $10cm \subseteq S_{10K}'$)

$$P(X \geq 10cm \mid H_0) = \frac{100}{10K} = 1\%$$

H_0 ← easier to simulate

Hypothesis Testing ← confusing idea

example: Given a coin determine if coin is biased towards head or not ?

biased towards head :- $P(H) > 0.5$

not biased towards head :- $P(H) = 0.5$

experiment: flip a coin 5 times and count no of heads $= X$ ← Test Statistic

X : r.v

Perform exp

f, f, f, f, f

↓ ↓ ↓ ↓ ↓

H H H H H

$X = 5$ ← obs

$P(X=5 \mid \text{coin not biased towards heads}) = P(\text{obs} \mid H_0)$

assumption

null hypothesis (H_0)

H_0 : coin is not biased towards head

$P(X=5 \mid H_0) = 1/2^5 = 1/32 \approx 0.03 = 3\%$

$X=5$: five heads in five tosses

H_0 : coin is unbiased toward heads

$P(X=5 \mid H_0) = 3\%$

There is a 3% chance of getting 5 heads in 5 flips if coin is not biased towards heads

hyp testing

$P(\text{Obs by exp} \mid \text{assumption}) = 3\%$

↓ pvalue — ↙ Small

if $P(\text{Obs} \mid H_0) < 5\%$

then H_0 may be incorrect

⇓

assumption / H_0 is not true

⇓

reject H_0 ⇒ reject coin is not biased towards head

⇓

Coin is biased

Null hypto

H_0: coin is not biased towards head

Alt hypot

H_1: coin is biased towards head

{ rejecting H_0 ⇒ accepting H_1
rejecting H_1 ⇒ accepting H_0 }

expt — flip 3 times
↳ count no of heads = X ← test Statistic

perform — B · B · B
H H H

$X = 3 \leftarrow$ Obs

$P(\text{obs} \mid \text{assumption}) = 1/2^3 = \frac{1}{8} = 12.5\% > 5\%$

↳ can not biased

accept H_0

★ Imp key points:

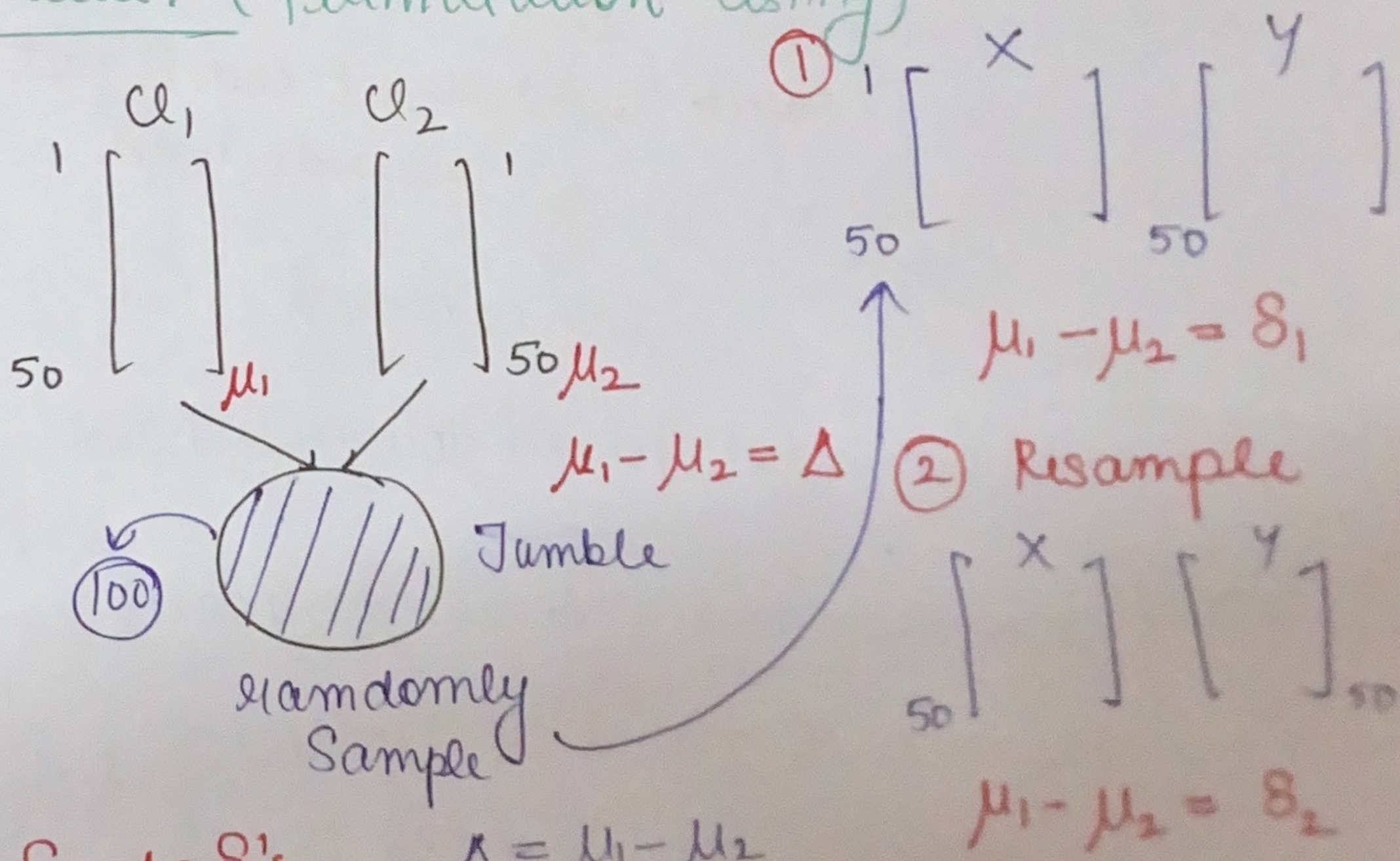
① design of experiment (Sample size)

② H_0: null hypothesis

$P(\text{obs} \mid H_0)$ is easy & feasible

★ p value is probability of observation given assumption

pValue: (Permutation Testing)



③ Sort δ's $\quad \Delta = \mu_1 - \mu_2$

$\delta_1', \delta_2', \delta_3' - \underline{\Delta} - \delta'_{10K}$

5%

⟶ Central (mid)

p value = 0.05

pre computing

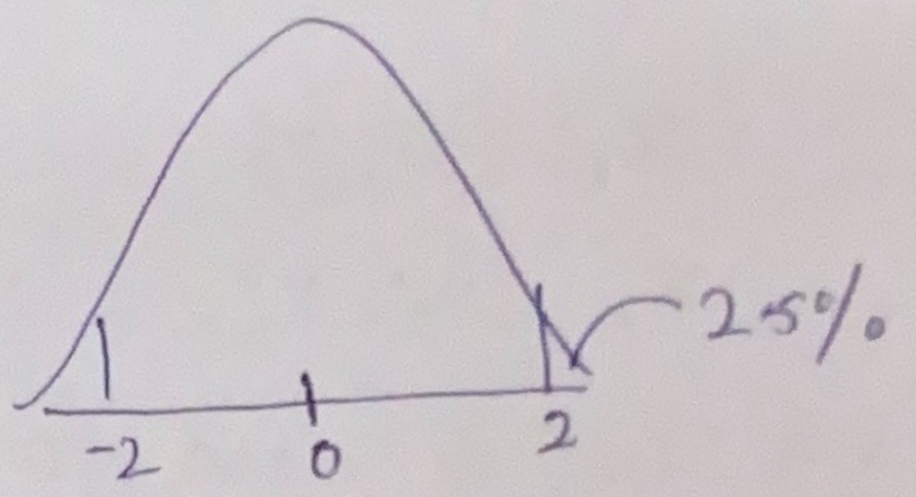
δ's ⟶

$\delta \sim N(0,1)$

$\Delta = 2$

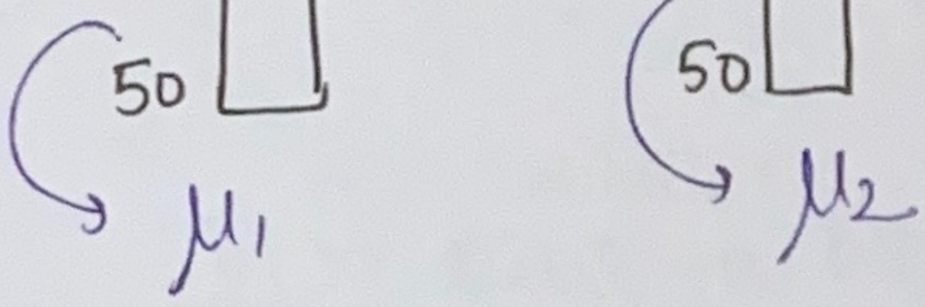
$p(\delta \geq 2) = 0.025$

p value $= 0.025$



↳ Same example in more detail

① Cl_1 Cl_2 ← we already made obs (actual data)



Sample size = 50

Steps

Δ on X: $\mu_2 - \mu_1 = 10\,cm$ ← observation → truth

② What is prob of observing a value of $\geq 10cm$ if there was no diff in class height?

↳ p value

prob$(X \geq 10 | H_0)$ ↳ null hypot

if p-value is small = 0.01 or 1%

↳ (<5%) → customary

↓

$p(X \geq 10cm | H_0) = 0.01$ or 1%

↓

H_0 is less probable or H_0 may not extreme

⇓

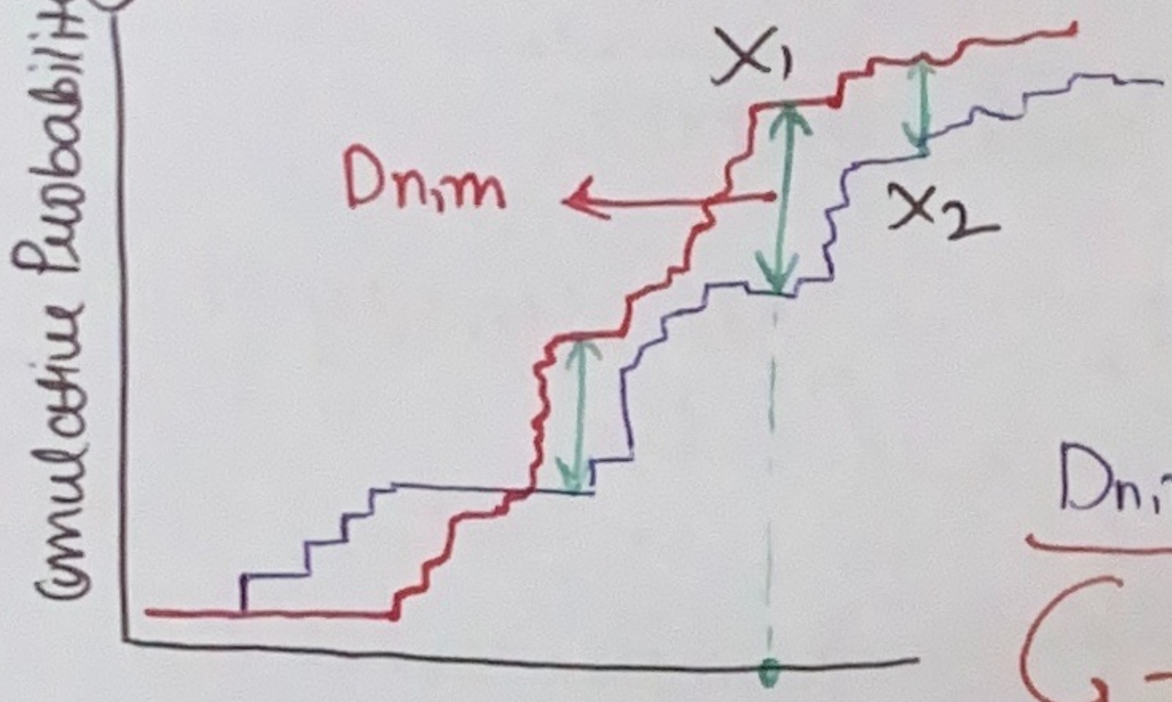
Reject H_0

Two Sample Kolmogorov - Smirnov Test (KS Test)

X_1 : $[x_1', x_2', \ldots x_n']$

X_2 : $[x_1^2, x_2^2, \ldots x_m^2]$

Q KS Test is used to determine whether X_1 & X_2 have same distribution



H_0 : X_1 & X_2 have same same distribution

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|,$$

↳ Test Statistic

The null hypothesis is rejected at level α if

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}}$$

$$c(\alpha) = \sqrt{\frac{-1}{2}\ln\alpha}$$

α	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95

look up table

eg ① If $n = 1000$ $m = 5000$

$\alpha = 0.05$

$D_{n,m} > 0.047$

then we reject null hypothesis at 0.05 level

② If $n = 50$ $m = 30$

$\alpha = 0.05$

$$1.36 \times \sqrt{\frac{50+30}{50(30)}} = 0.31$$

$D_{n,m} > 0.31$

reject H_0

① hypot Testing

H_0 is rejected at a significance level α if p-value $< \alpha$

② KS Test

H_0 is rejected if $D_{n,m} > C(\alpha)\sqrt{\frac{n+m}{nm}}$

Let $D_{n,m} = D$

reject H_0 if $D\sqrt{\frac{nm}{n+m}} > c\alpha = \sqrt{-\frac{1}{2}\ln\alpha}$

$\Rightarrow D\sqrt{\frac{nm}{n+m}} > \sqrt{-\frac{1}{2}\ln\alpha}$

$\Rightarrow D^2\left(\frac{nm}{n+m}\right) > -\frac{1}{2}\ln\alpha$

$= -D^2\left(\frac{nm}{n+m}\right) < \frac{1}{2}\ln(\alpha)$

reject H_0 if

$= \exp\left[-2D^2\left(\frac{nm}{n+m}\right)\right] < \alpha$

↳ p-value for KS Test

Hypothesis Testing: Difference of means

eg ① Coin toss → simple

$P(obs \mid H_0)$: very easy

eg ②

Determine if population means of heights of people in these two cities is same or not?

μ_1 / μ_2 are same or different

exp:

randomly

C_1: $h_1, h_2, \ldots, h_{50}$ — C_2: $h_1, h_2, \ldots, h_{50}$

↳ sample heights of 50 people

$$\mu_1 = \frac{h_1 + h_2 + \cdots h_{50}}{50} \longrightarrow 162\,cm$$

$$\mu_2 = \frac{h_1' + h_2' + \cdots h_{50}'}{50} \longrightarrow 167\,cm$$

Test Statistic: $|\mu_1 - \mu_2| = x = 162 - 167 = 5\,cm$

Null hypothesis: H_0 = no difference in population means

Compute: $p(x = 5cm \mid H_0)$

probability of observing a difference of 5cm in sample mean heights of sample size 50 b/w C_1 & C_2 if there is no difference in mean-heights

Case 1: $p(\bar{x}=5 \mid H_0) = 0.2 = 20\%$

There is 20% chance of observing a diff of 5cm in sample mean heights of C_1 & C_2 with sample size 50 if there is no pop mean diff

$\Rightarrow$ assumption must be true

$\Rightarrow$ accept H_0

Case 2: $p(\bar{x}=5 \mid H_0) = 0.03\% = 3\%$

$p(\text{obs} \mid \text{assumption}) = 3\%$ → Small

$\Rightarrow$ assumption incorrect

$\Rightarrow$ reject $H_0 \Rightarrow$ accept H_1

Resampling & Permutation Testing:

$\bar{x} = \mu_1 - \mu_2$ ← diff in sample mean with sample size of 50

H_0: no diff in population means

Step 1

$S = \{h_1, h_2, h_3 \dots h_1', h_2', h_3' \dots h_{50}'\}$

$C_1 \cup C_2$

Step 2

$S \rightarrow S_1, S_2$ randomly Selecting

$S_1 \rightarrow \mu_1$, $S_2 \rightarrow \mu_2$ → Resampling

repeat resampling

1. $\mu_2 - \mu_1 = 3\text{cm}$ → δ_1
2. $\mu_2 - \mu_1 = -2\text{cm}$ → δ_2
3. $\mu_2 - \mu_1 = +1\text{cm}$ → δ_3
4. $\mu_2 - \mu_1 = 6\text{cm}$ → δ_4

k = 1000

Step 3 Sort S_i's

$$S_1' \leq S_2' \leq S_3' \leq S_4' - - - - \leq S_k'$$

inc order

Case 1 obs-difference $= x = 5cm$

$P(\text{diff} \geq 5cm \mid H_0) = ?$

$S_1' \leq S_2' \leq S_3' - 5cm \quad S \leq S_{1000}'$

$80\% \leq 5cm$ $\quad$ $20\% \geq 5cm$

pvalue $P(\text{obs diff} \mid \text{assumption}) = 20\% \searrow > 5\%$

Case 2 $S_1' \leq S_2' \leq - - - 5cm - - - - \leq S_{1000}'$

97% $\quad$ 3%

$P(\text{obs diff} \geq 5cm \mid H_0) = 3\%$

→ assumption incorrect

→ reject H_0

How to use hypothesis - testing

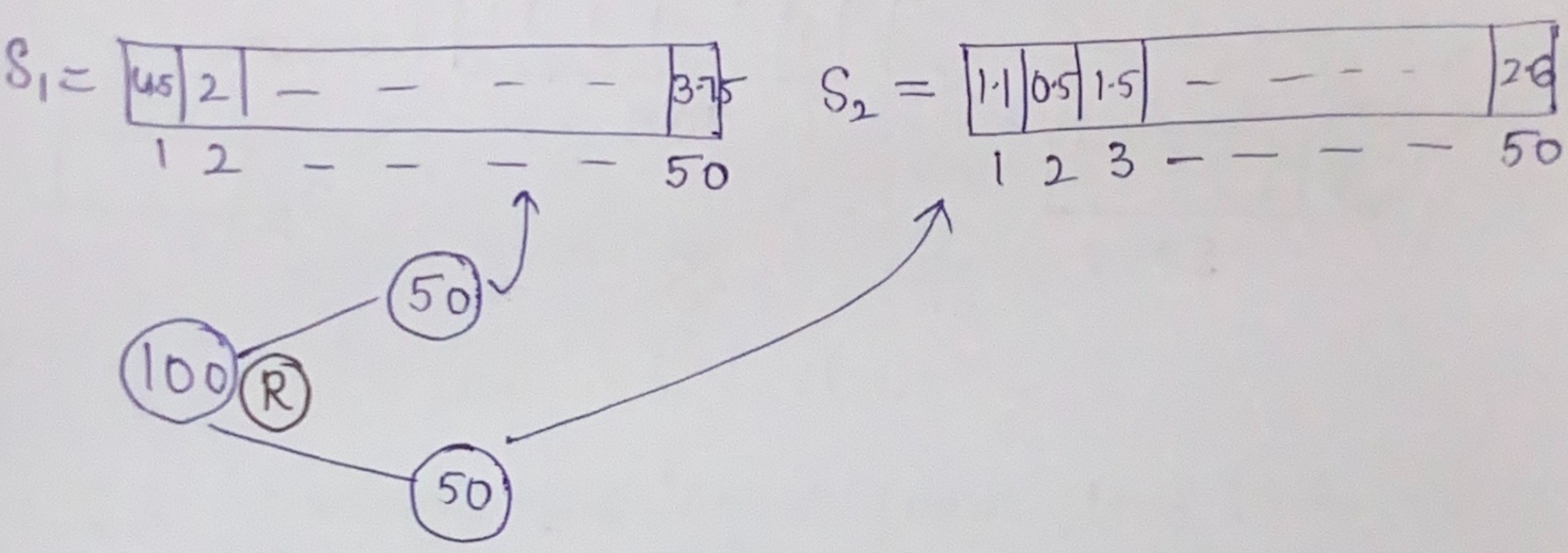
KS test → used if two r.v have same distribution or not.

eg designing drug / medicine

(in market) D_1 → reduces fever in 4 hours

D_2 (new drug) → claim: reduces fever faster than drug 1

experiment → to determine if true or not (claim)



μ_1 = mean time = 4.0 hrs

μ_2 = mean time = 2 hrs

Hyp testing : ① H_0 : D_2 & D_1 are not different
have same time taken to reduce fever

② Test - Statistic : $X = \mu_2 - \mu_1 = 4 - 2 = 2$ hrs (obs value)

↳ $P(x \geq 2 \mid H_0)$ = very small (1%) let

↳ If there is no difference in D_1 and D_2 (H_0), the probability of observing mean diff $x \geq 2$ is very small (1%)

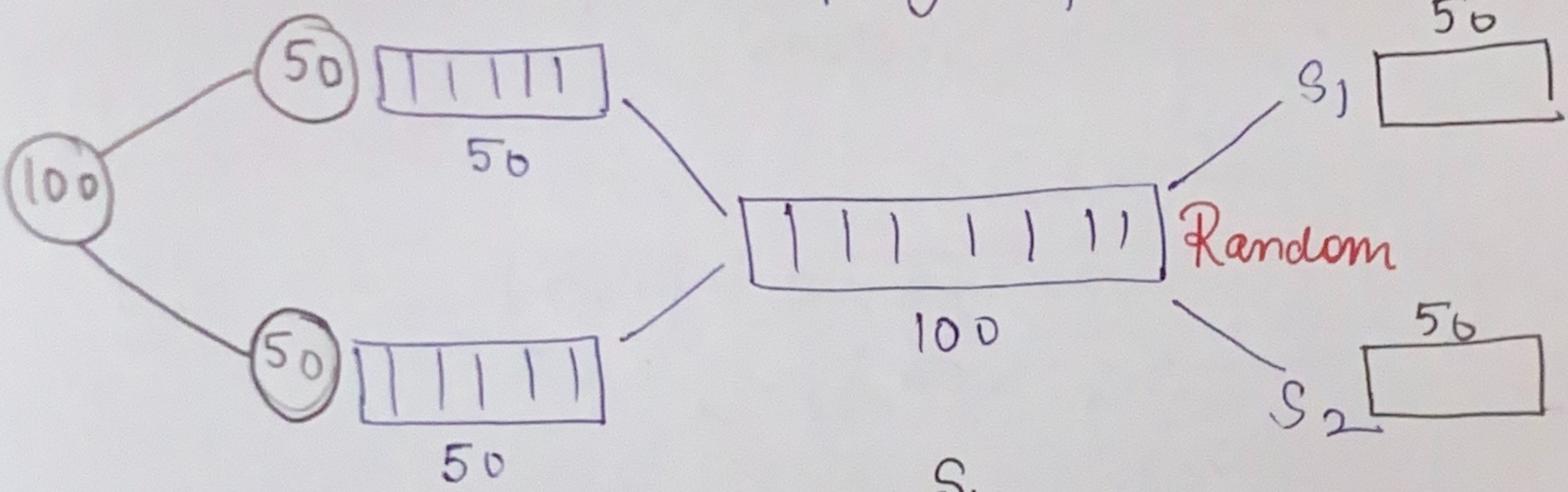
⇓

H_0 & obs do not agree with each other

⇓

H_0 must be wrong

$P(X \geq 2 \mid H_0)$ ⟶ resampling & permutation test



S_1
S_2
⋮

Proportional Sampling

$d =$

d_1	d_2	d_3	d_4	d_5
2.0	6.0	1.2	5.8	20.0
1	2	3	4	5=n

randomly

Task: pick an element amongst n elements s.t the prob of picking an element is proportional to the d_i's value

Step 1 (a) $S = \sum_{i=1}^{n} d_i = 35$ ← compute the sum

(b) $d_i' = d_i / S$ ← normalizing values using sum

$d_1' = 0.0571$
$d_2' = 0.171428$
$d_3' = 0.0343$
$d_4' = 0.1657$
$d_5' = 0.5714$

} 0 to 1, Sum to 1

$\sum d_i' = \sum \frac{d_i}{S} = 1$

(c) cumulative normalized Sum

$\tilde{d}_3 = \text{sum}$ { $d_1 = 0.0571$, $d_2' = 0.1714428$, $d_3' = 0.0343$, $d_4' = 0.1657$, $d_5' = 0.5714$ } →

$\tilde{d}_1 = d_1' = 0.0571$

$\tilde{d}_2 = \tilde{d}_1 + d_2' = 0.2285528$

$\tilde{d}_3 = \tilde{d}_1 + \tilde{d}_2 + d_3' = 0.2628728$

$\tilde{d}_4 = 0.4285528$

$\tilde{d}_5 = 1.00$

$\tilde{d}_i$ = cum-nor- values

Step 2 Sample one value (γ) from unif (0.0, 1.0)

γ = numpy.random.uniform(0.0, 1.0, 1)

range ↓ , no of values ↓

let $\gamma = 0.6$

Step 3 proper-sampling

```
if γ ≤ d̃1        d̃1
    return 1     d̃2
else if γ ≤ d̃2   d̃3 ] ← γ
    return 2     d̃4 ]
else if γ ≤ d̃3   d̃5
    return 3
    :
    :
```

prob of γ lying b/w $\tilde{d}_3$ & $\tilde{d}_4$ = $d_4' \propto d_4$

$d_4' = d_4/S$