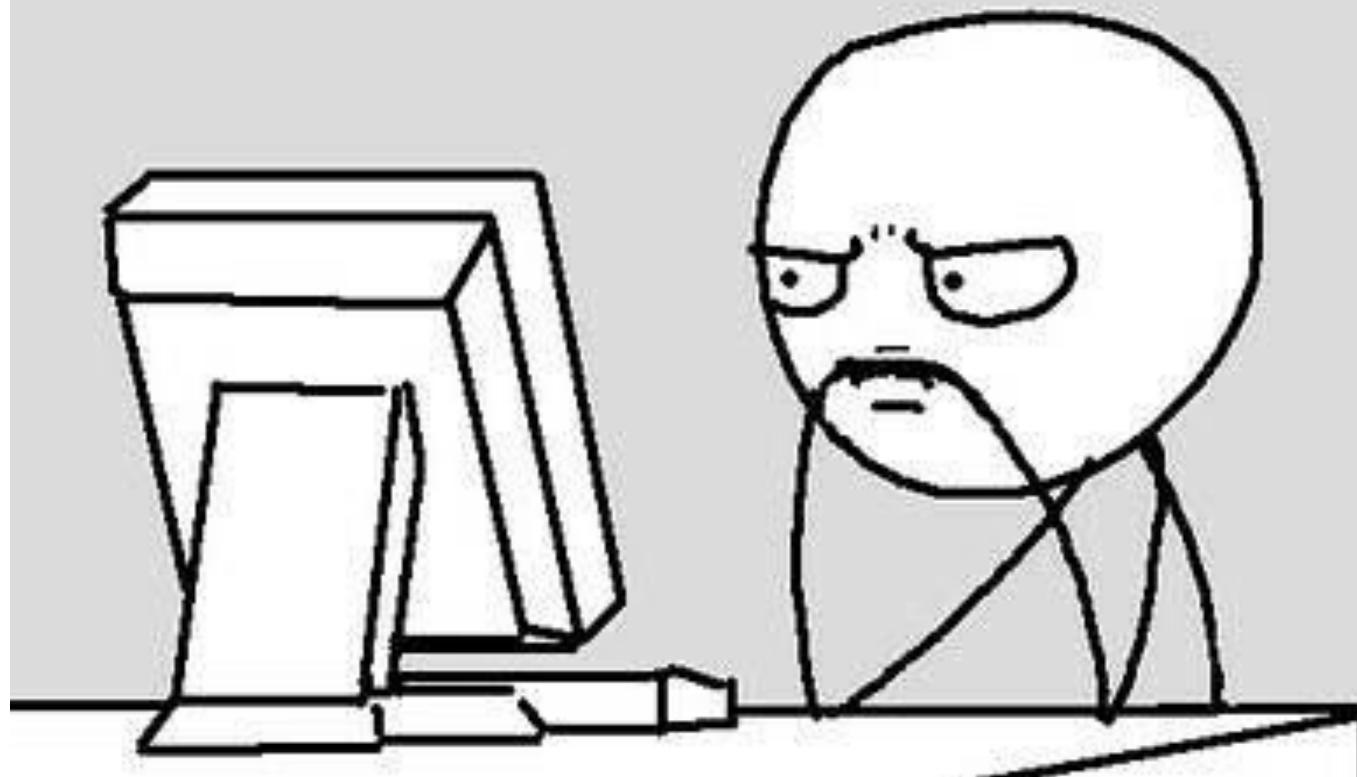
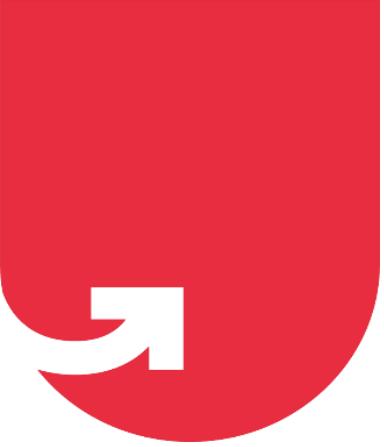


LET'S WAIT





#LifeKoKaroLift

Data Science Certification Program

Course : Machine Learning

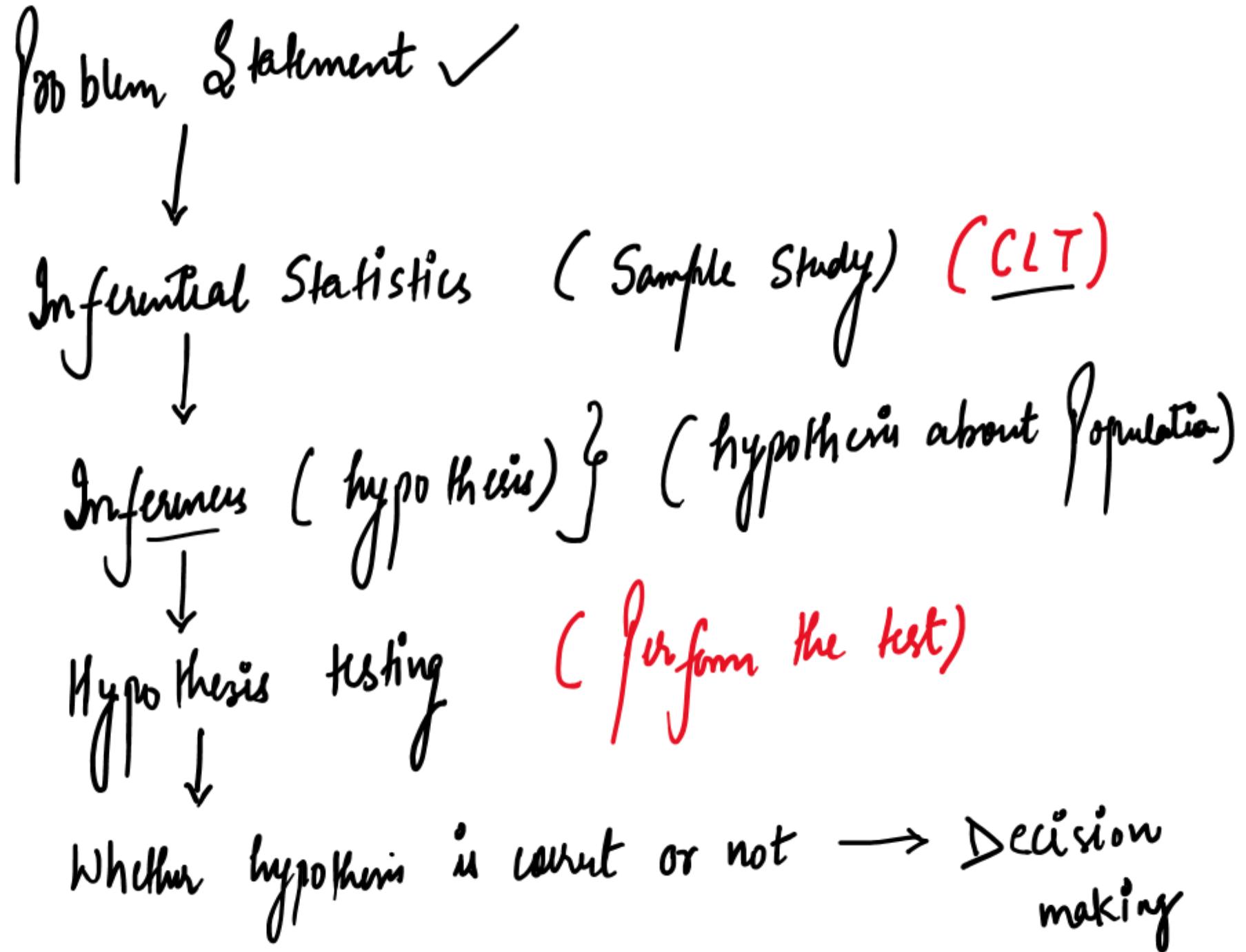
Lecture On : Hypothesis
Testing

Instructor : Shivam Garg



No
validation

Validation



Hypothesis Testing: (Validated)

upGrad

Inferences about population derived with the help of inferential statistics.

H_0 : Null hypothesis : {status quo (Claim given)} } Numericals (Symbol) { $=, \geq, \leq$ }
 $\downarrow H_0$

H_1 : Alternate hypothesis : {Opposite to status quo } } { $\neq, >, <$ }
 $\downarrow H_1$

(Theoretical definition)

always ✓

→ Claim: f_K will generate average ticket size of 2000?

$$T \left\{ \begin{array}{l} H_0 = \mu = 2000 \\ H_1 = \mu \neq 2000 \end{array} \right. \quad S \quad \left. \begin{array}{l} H_0 = \mu = 2000 \\ H_1 = \mu \neq 2000 \end{array} \right\}$$

→ Claim: f_K will not generate average ticket size of 2000.

$$T \left\{ \begin{array}{l} H_0 = \mu \neq 2000 \\ H_1 = \mu = 2000 \end{array} \right. \quad S \neq \quad \left. \begin{array}{l} H_0 = \mu = 2000 \\ H_1 = \mu \neq 2000 \end{array} \right\}$$

Formulating Null and Alternate Hypothesis

Example-1: A restaurant owner installed a new automated drink machine. The machine is designed to dispense 530 mL of liquid on the medium size setting. The owner suspects that the machine may be dispensing too much in medium drinks. They decide to take a sample of 30 medium drinks to see if the average amount is significantly greater than 530 mL.

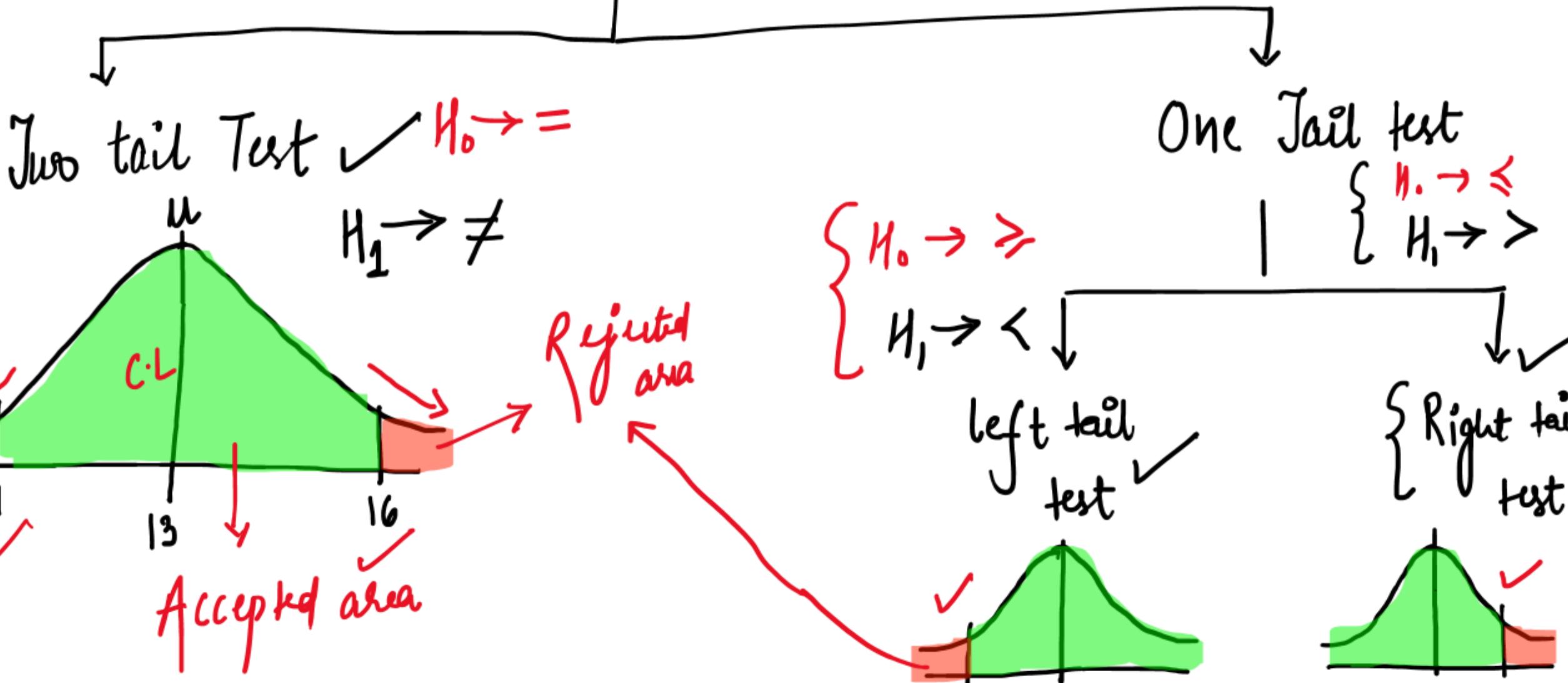
$$\left\{ \begin{array}{l} H_1 \rightarrow \mu > 530 \text{ mL} \\ H_0 \rightarrow \mu \leq 530 \text{ mL} \end{array} \right. \quad \left. \begin{array}{l} \text{formulation of} \\ \text{hypothesis} \end{array} \right\}$$

$$\begin{aligned} &\rightarrow H_0: =, \geq, \leq \\ &\uparrow \quad \uparrow \quad \uparrow \\ &\rightarrow H_1: \neq, <, > \\ &\downarrow \quad \downarrow \quad \downarrow \\ &\rightarrow \text{Right tail test} \end{aligned}$$

→ Hypothesis testing :- (Problem)

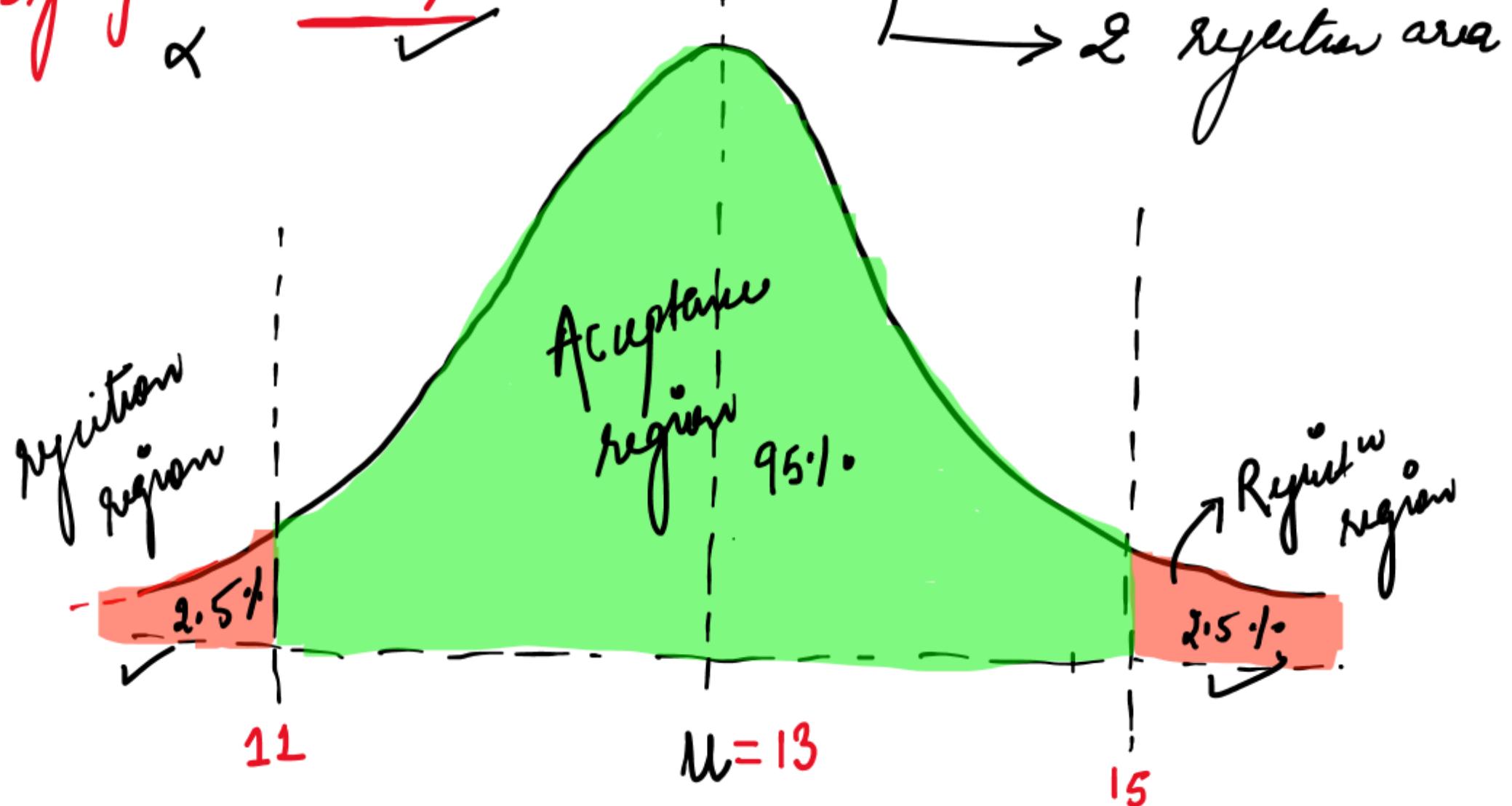
- ① formulation of hypothesis (what) {
- ② Which test ?
 - 2-Tail
 - Left (which)
 - Right}
- ③ Methodology (CVM / p-value method) (how)
- ④ Decisioning

Type of test



Confidence level = 95%. Two Tail Test $\rightarrow \{ H_0 \rightarrow = H_1 \rightarrow \neq \}$

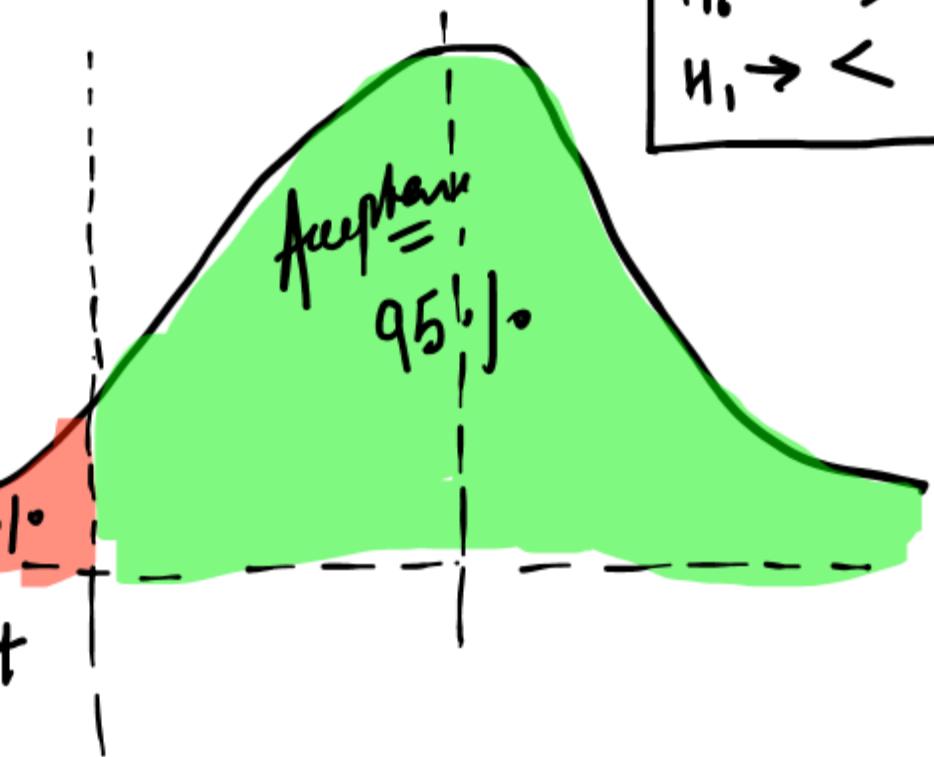
Significant level = 5%. α



left Tail

C.F = 95 %.

$$\begin{array}{l} H_0 \rightarrow \geq \\ H_1 \rightarrow < \end{array}$$



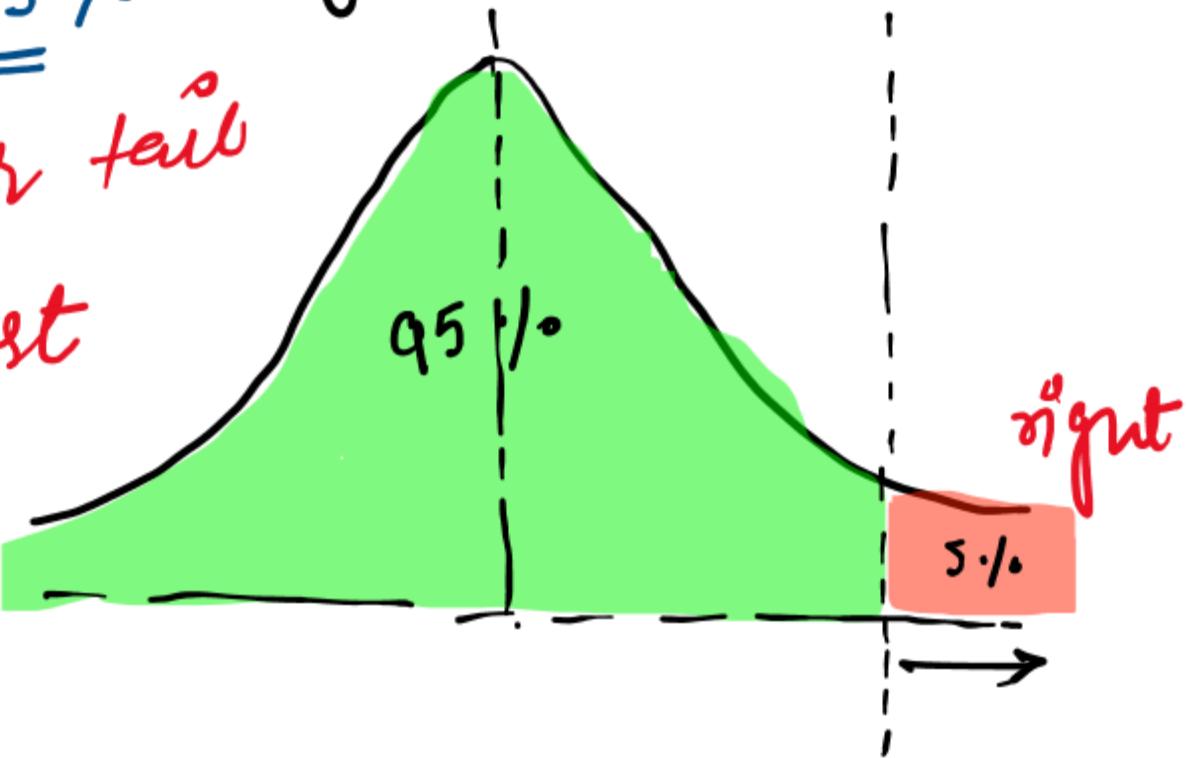
Lower tail test

Right Tail

C.L = 95 %

S.L = 5 %

Upper tail
test



$H_1 \rightarrow \neq \rightarrow \text{2-tail test}$

$H_1 \rightarrow < \rightarrow \text{Left tail test}$

$H_1 \rightarrow > \rightarrow \text{Right tail test}$

Two methods



→ Critical Value method :-

$$L \cdot C \cdot V = U - Z \times \frac{\sigma}{\sqrt{n}}$$

- ① formulation of hypothesis

② which test?

③ Calculate lower critical value and/or upper critical value
(L.C.V) | (U.C.V) / (H.C.V)

χ^2 -tail test \rightarrow L.C.V and U.C.V

Left tail test \rightarrow L.C.V

Right tail test \rightarrow U.G.V

A simple random sample of 50 adults womens is obtained, and each person's red blood cell count (in cells per microliter) is measured. The sample mean is 4.63. The population standard deviation for red blood cell counts is 0.54. Construct the 95% confidence interval estimate for the mean red blood cell counts of adults.

$$n = 50 \checkmark \rightarrow CL = 95\% \quad (\text{Green Area})$$

$$\mu = 4.63 \checkmark \rightarrow \alpha = 5\%$$

$$\sigma = 0.54 \checkmark$$

Significance level ($100 - CL$)



Activate Windows
Go to PC settings to activate Windows.

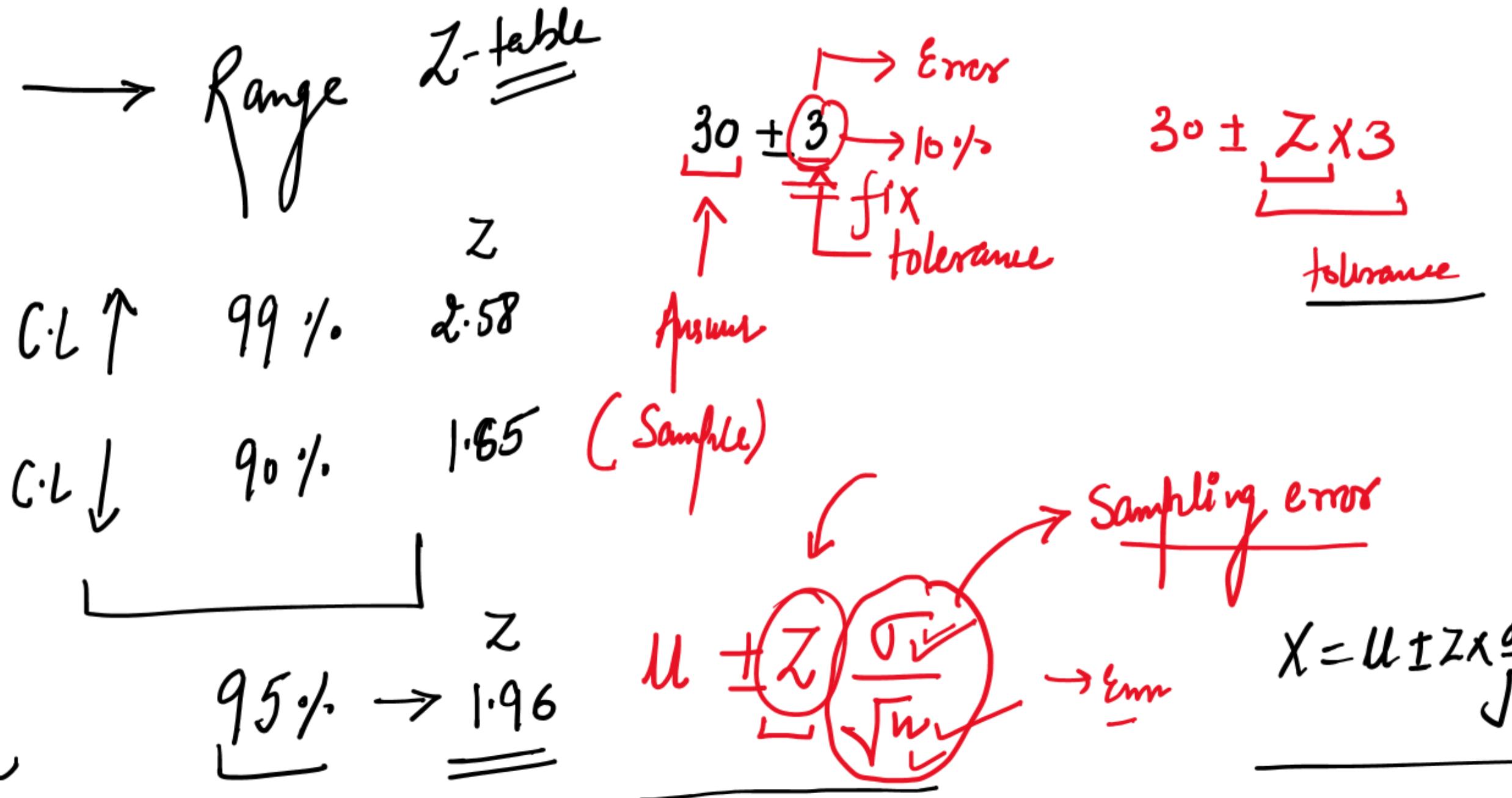
Q: Analyst claims that average red blood cell count is 4.63 cells/μL
Validate the claim using population parameter of ~~4.53~~ cells/μL

Sol:- ① formulation of hypothesis : $H_0: \mu = 4.63 \text{ cells}/\mu\text{L}$ } ✓
 $H_1: \mu \neq 4.63 \text{ cells}/\mu\text{L}$

② Which test? → Two Tail test

$$L.C.V = \mu - Z \times \frac{\sigma}{\sqrt{n}}$$

$$H.C.V = \mu + Z \times \frac{\sigma}{\sqrt{n}}$$



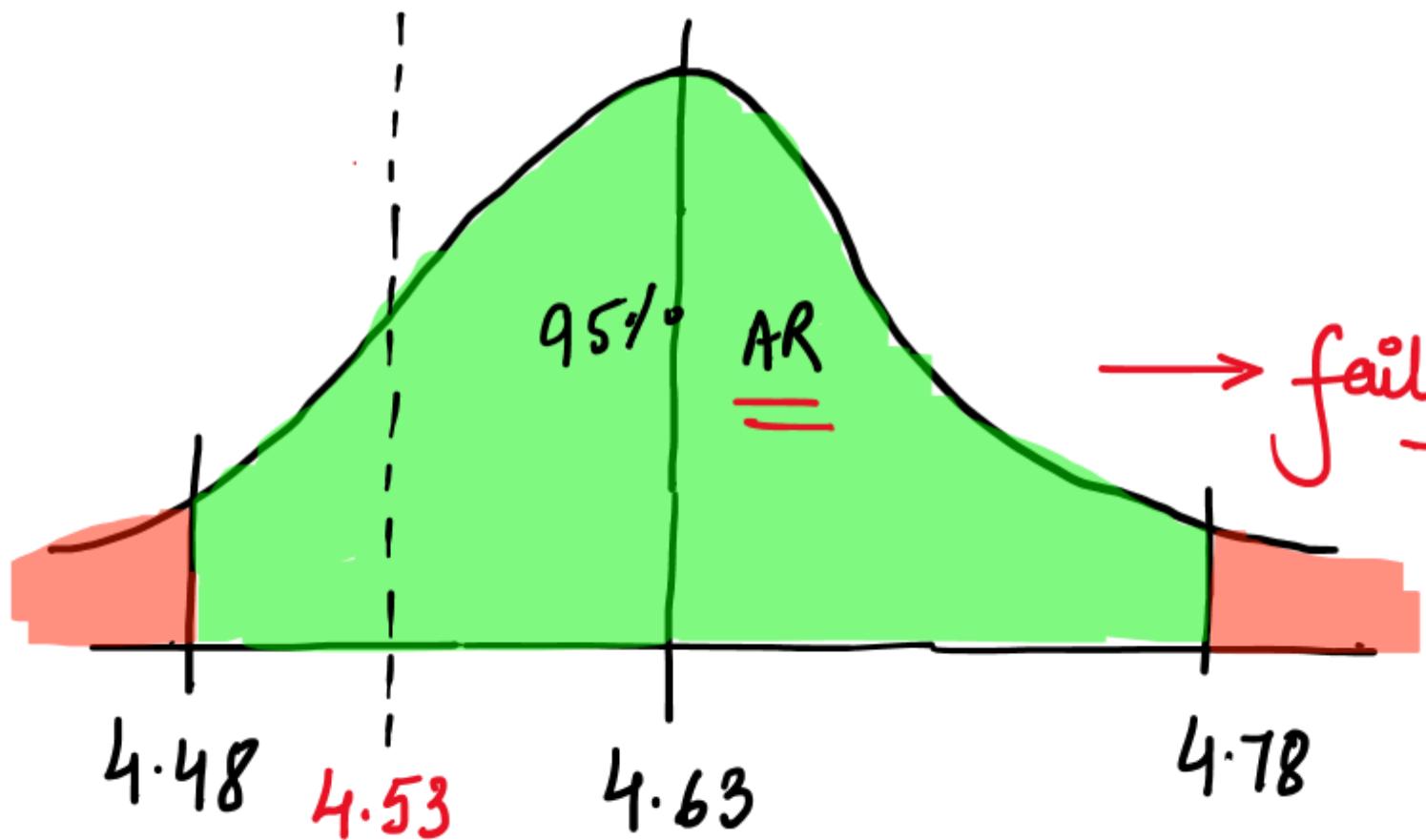
$$L.C.V = 4.63 - 1.96 \times \frac{0.54}{\sqrt{50}} \quad \checkmark$$

$$= 4.48$$

$$U.C.V = 4.63 + 1.96 \times \frac{0.54}{\sqrt{50}}$$

$$= 4.78$$

④



4.63 ± 0.15

→ fail to reject null hypothesis

→ p-value method :-

① formulation of hypothesis ✓

② which test ✓

③ Calculate Z-Stat → $Z = \frac{x - u}{\sigma / \sqrt{n}}$

④ Calculate the p-value from Z-stat using z-table.

⑤ Decisioning

x : population parameter

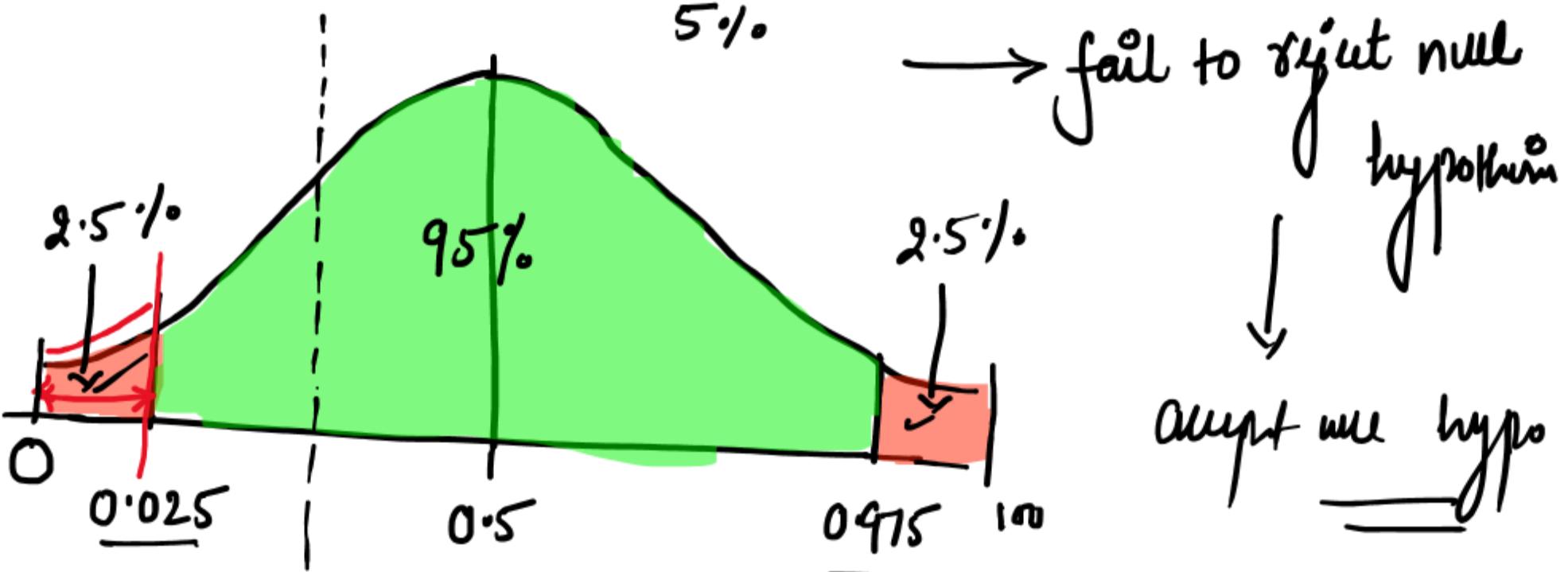
u : sample mean

σ : pop. std

n : sample size

$\rightarrow H_0 \rightarrow \mu = 4.63, \quad H_1 \rightarrow \mu \neq 4.63$	$x = 4.53$
\rightarrow Two-tailed test	$\mu = 4.63$
$\rightarrow Z\text{-Stat} = \frac{x - \mu}{\sigma / \sqrt{n}} = \frac{4.53 - 4.63}{0.54 / \sqrt{50}}$	$\sigma = 0.54$
$= \frac{-0.1 \times \sqrt{50}}{0.54} = -\frac{10 \times \sqrt{50}}{54} = -1.31$	$n = 50$
$\rightarrow p\text{-value} (@ -1.31) = \underline{\underline{0.0951}}$	-1.31

→



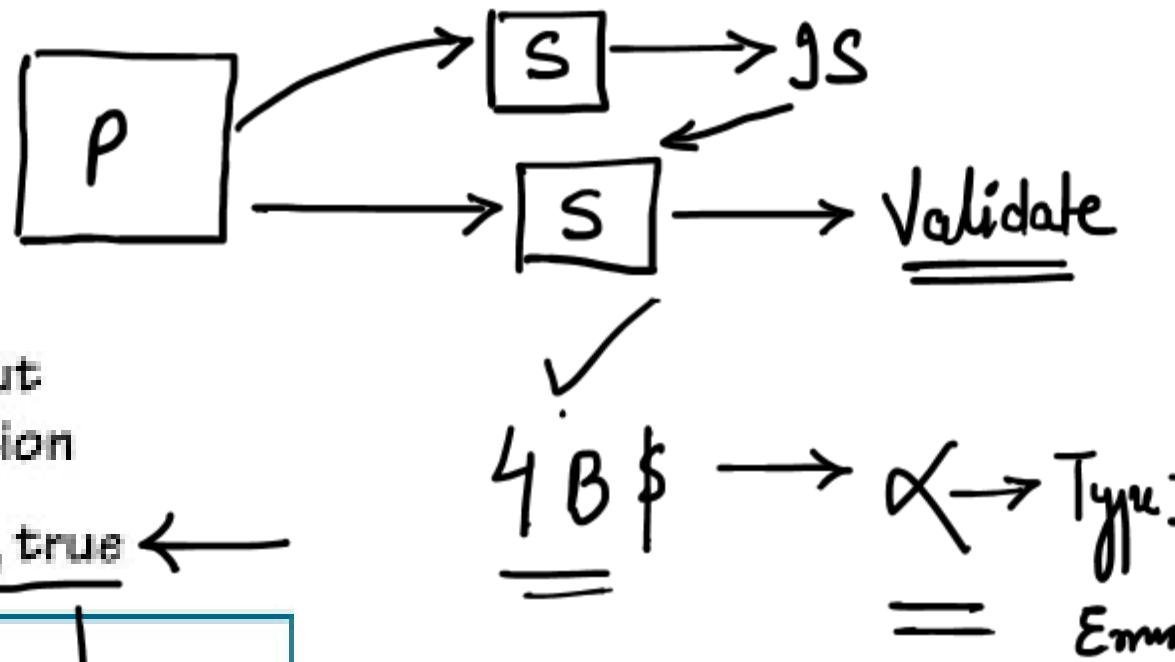
• 0.0951

* fail to reject null hypothesis } → Accept null hypothesis ?

* Reject null hypothesis ✓

→ Type of Errors :-

		Truth about the population
Decision based on sample	Reject H_0	H_0 true
	Accept H_0	H_a true
	Type I error	Correct decision
	Correct decision	Type II error

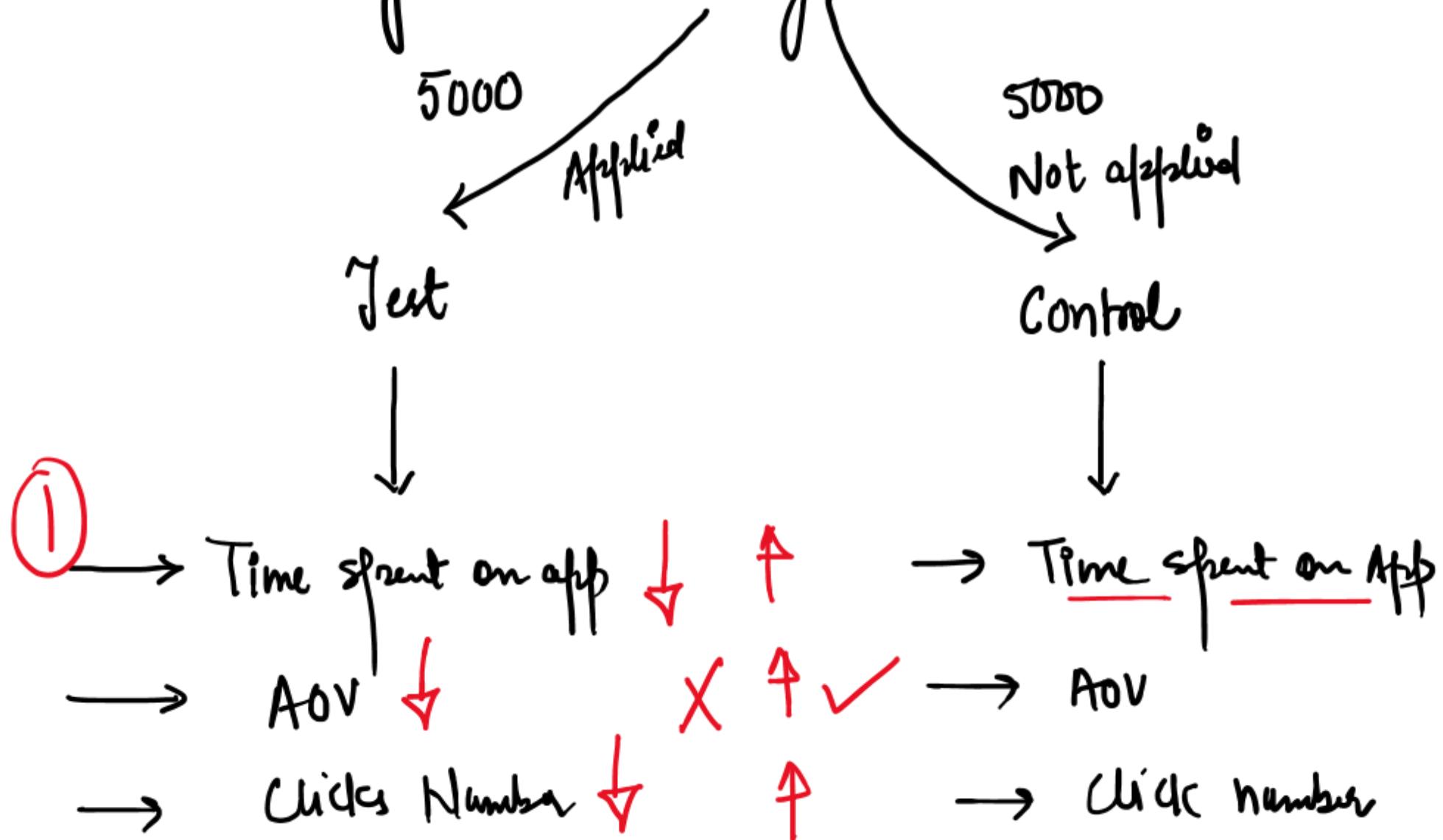


Claim ✓ → Rejected
 Claim ✗ → Accepted

Women → Self-kit → Not Pregnant → Type I error =
(Pregnant) = }
↓
Non-Pregnant → Self-kit → Pregnant → Type II error = }

ML → logistic Reg

→ A/B Testing :-



Interv
→ Def \xrightarrow{HT} Not-def (Concern) Entire
Agent ✓

→ Non-Def \xrightarrow{HT} Def (Very less concern) Interest
Agent ✓

Telcom

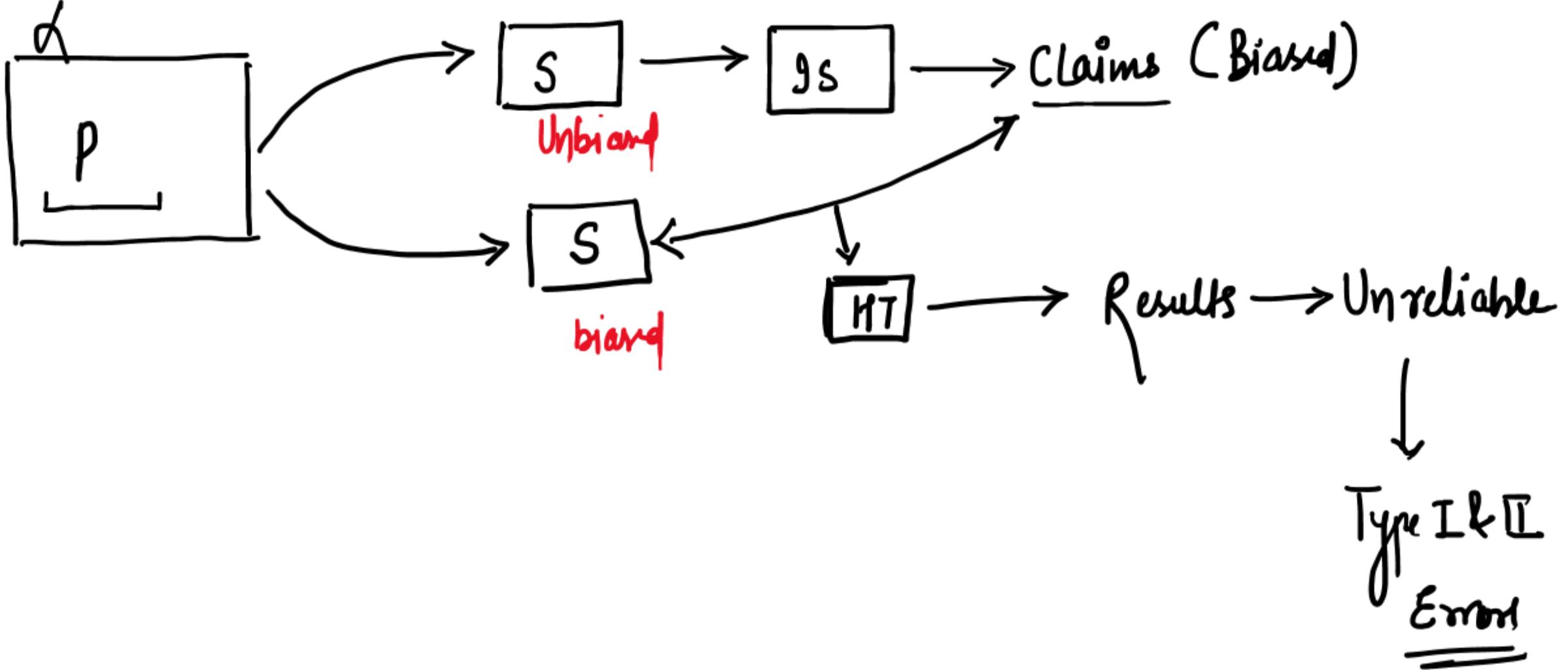
Churn \xrightarrow{HT} Non-Churn (Concern) ✓

Non-Churn \xrightarrow{HT} Churn (Not a issue) of

H

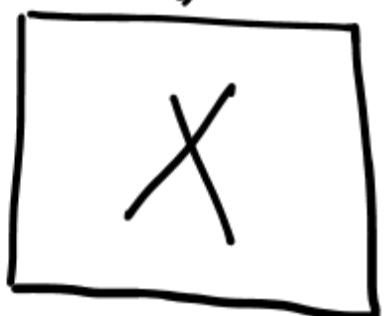
C → NC (Concern)

NC → C (Concern)

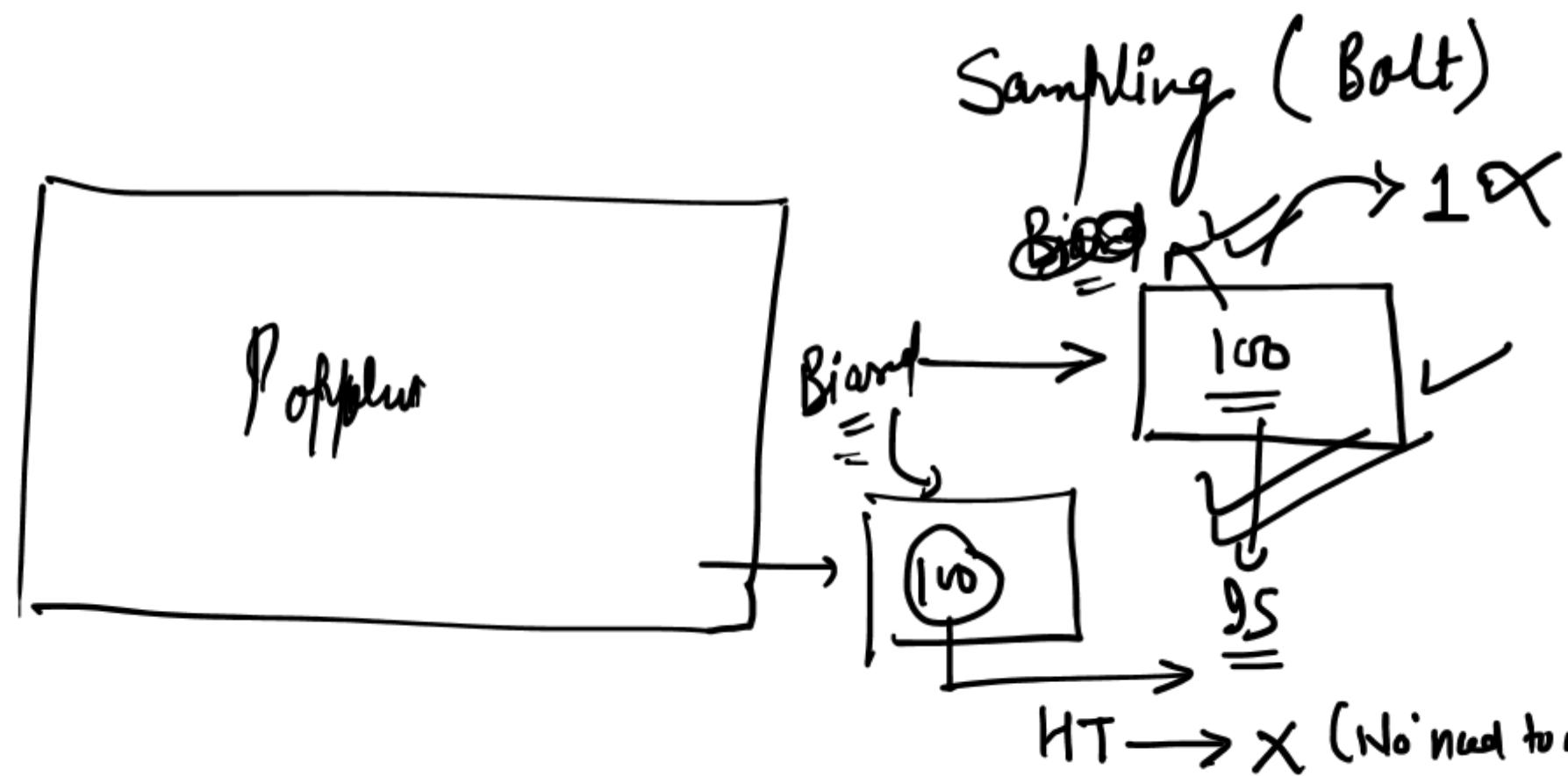


$\downarrow W = 5\text{kg} \rightarrow$ hypothesis is perfectly right

$H_1 \checkmark$
 α



Block \rightarrow claim \rightarrow up to 10kg



① $\mu = 4.63 - H_0$

$\mu \neq 4.63 - H_1$

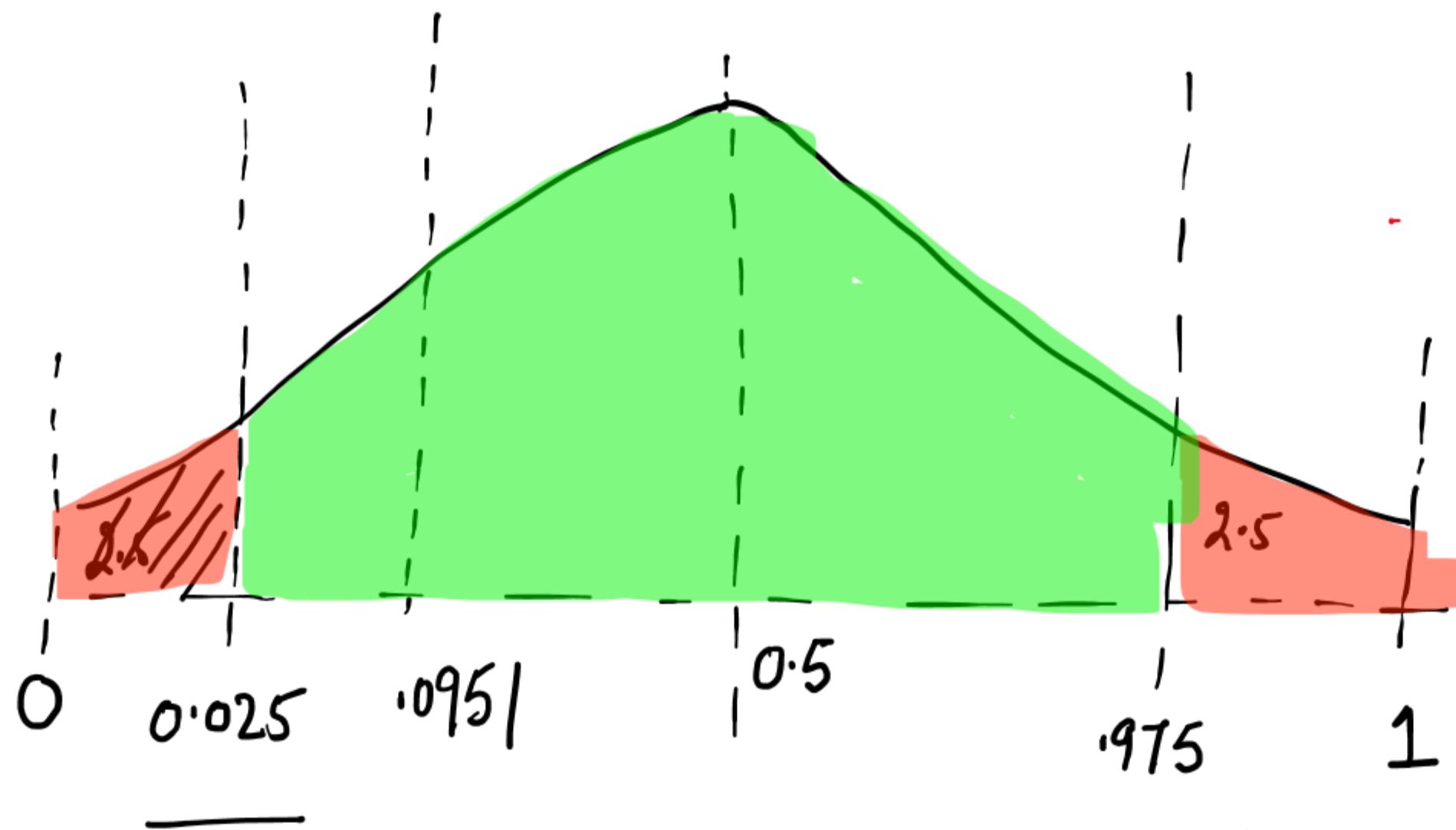
② 2-tail test

③ Z-Stat = $\frac{x - \mu}{\sigma / \sqrt{n}}$ = $\frac{4.53 - 4.63}{0.54 / \sqrt{50}} = -1.31$

(Z-table)

④ p-Value = 0.095
(Z-table)

$$CL = 95\%$$
$$\alpha = 5\%$$



fail to reject the null hypo

$\alpha = 5\%$ p-value method (a) formalat: $H_1: \mu > 120$

SUPERMARKET LOYALTY PROGRAM

Video: Speaker

- A supermarket plans to launch a loyalty program if:
Avg. spending per shopper > ₹120 per week
- A random sample of 80 shoppers enrolled in the pilot program spent an average of ₹130 in a week with a standard deviation of ₹40.

$H_0: \mu \leq 120$

(b) which test \rightarrow Right tail test

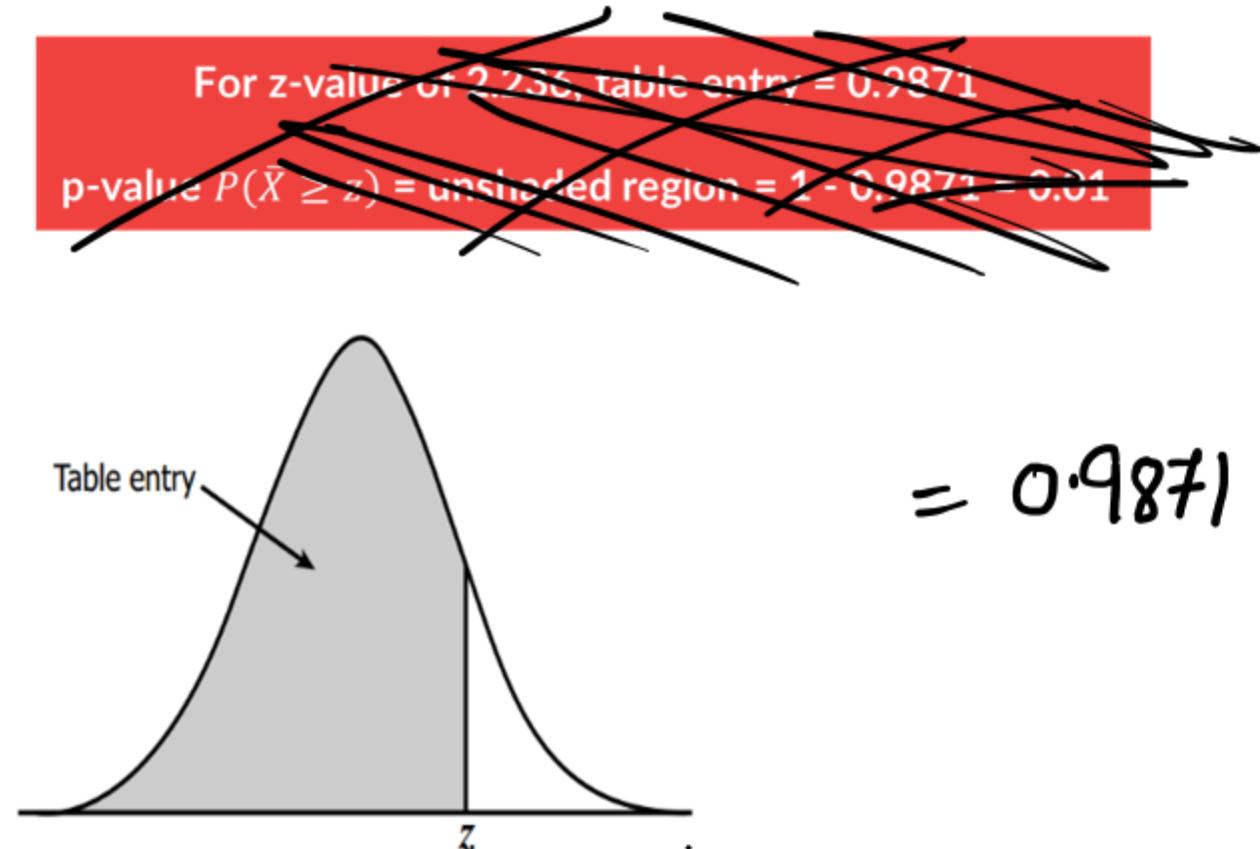
Should the supermarket launch the loyalty program?

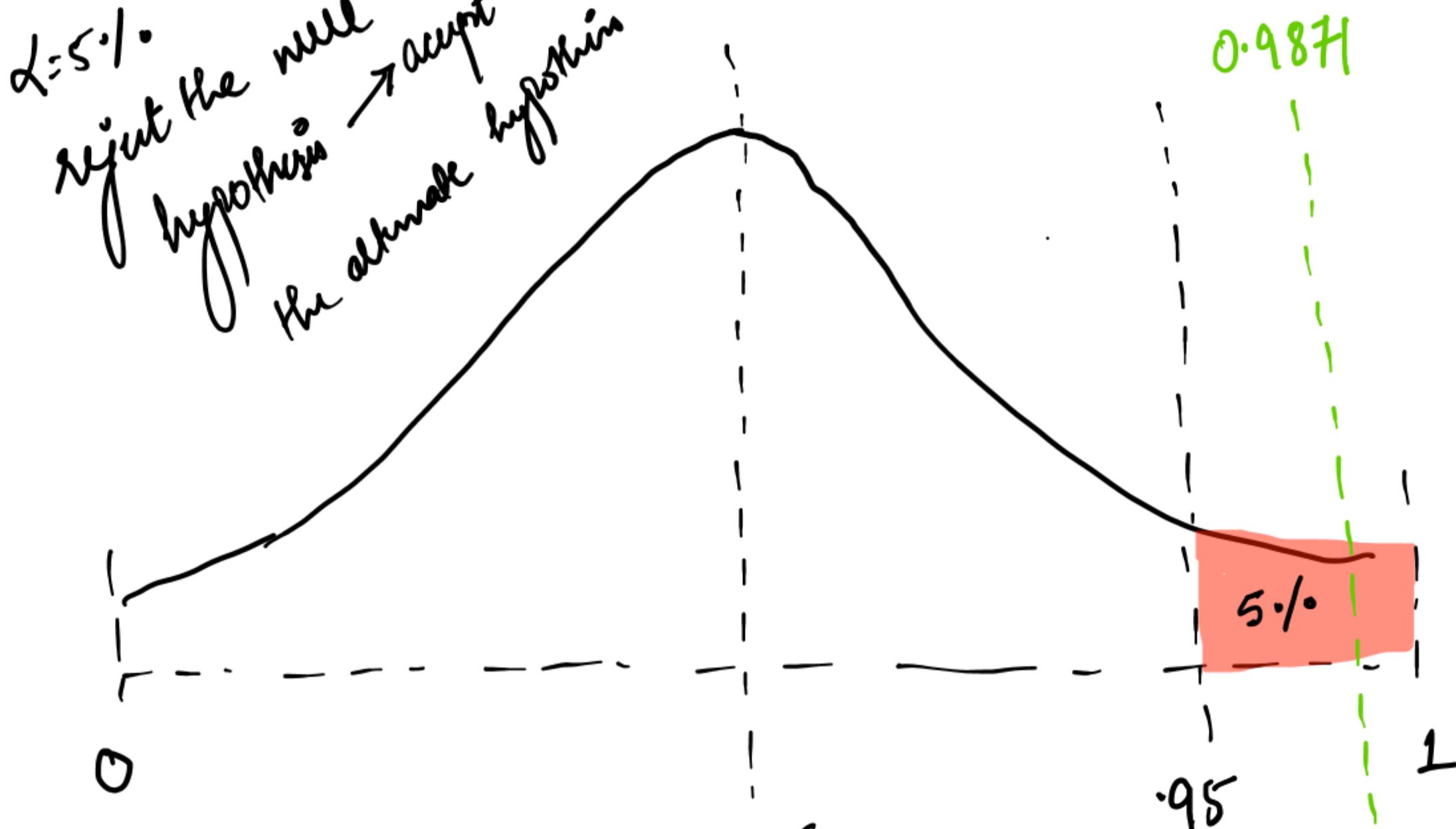
$$(c) Z_{\text{Stat}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{130 - 120}{40 / \sqrt{80}} = \frac{10}{40 / \sqrt{80}} = \frac{10}{40 / 4\sqrt{5}} = \frac{10}{4\sqrt{5}} = 2.23$$

Z tabl.

LOOKING UP Z-TABLE

<i>z</i>	.00	.01	.02	.03	.04	.05
0.0	.5000	.5040	.5080	.5120	.5160	.5199
0.1	.5398	.5438	.5478	.5517	.5557	.5596
0.2	.5793	.5832	.5871	.5910	.5948	.5987
0.3	.6179	.6217	.6255	.6293	.6331	.6368
0.4	.6554	.6591	.6628	.6664	.6700	.6736
0.5	.6915	.6950	.6985	.7019	.7054	.7088
0.6	.7257	.7291	.7324	.7357	.7389	.7422
0.7	.7580	.7611	.7642	.7673	.7704	.7734
0.8	.7881	.7910	.7939	.7967	.7995	.8023
0.9	.8159	.8186	.8212	.8238	.8264	.8289
1.0	.8413	.8438	.8461	.8485	.8508	.8531
1.1	.8643	.8665	.8686	.8708	.8729	.8749
1.2	.8849	.8869	.8888	.8907	.8925	.8944
1.3	.9032	.9049	.9066	.9082	.9099	.9115
1.4	.9192	.9207	.9222	.9236	.9251	.9265
1.5	.9332	.9345	.9357	.9370	.9382	.9394
1.6	.9452	.9463	.9474	.9484	.9495	.9505
1.7	.9554	.9564	.9573	.9582	.9591	.9599
1.8	.9641	.9649	.9656	.9664	.9671	.9678
1.9	.9713	.9719	.9726	.9732	.9738	.9744
2.0	.9772	.9778	.9783	.9788	.9793	.9798
2.1	.9821	.9826	.9830	.9834	.9838	.9842
2.2	.9861	.9864	.9868	.9871	.9875	.9878
2.3	.9893	.9896	.9898	.9901	.9904	.9906
2.4	.9918	.9920	.9922	.9925	.9927	.9929
2.5	.9938	.9940	.9941	.9943	.9945	.9946
2.6	.9953	.9955	.9956	.9957	.9959	.9960
2.7	.9965	.9966	.9967	.9968	.9969	.9970





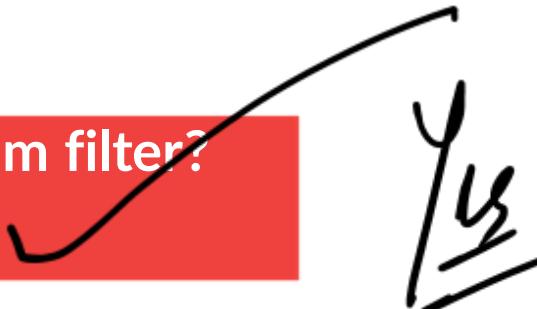
→ Supermarket should launch the loyalty program

SPAM FILTER

- ABC company is considering buying a commercial filtering software costing \$15,000/annum that the vendor claims will reduce spam to less than 20%.
- The company believes that if the software can reduce spam to less than 20%, then it can help improve employee productivity.
- The company collected a sample of 100 incoming mails with the spam filter and found that 11 of them were spam.
- Assume significance level $\alpha = 0.05$
- $\sigma = 0.4$

$$11/100 = 11\%$$

Should the company buy the spam filter?



① formulating hypothesis :-

$H_1: \text{spam} < 0.2$ ✓

$H_0: \text{spam} \geq 0.2$ ✓

② which test :- left tail test ✓

③ Z-Stat :-

$$\frac{x - \mu}{\sigma / \sqrt{n}} = \frac{0.11 - 0.2}{0.4 / \sqrt{100}} = \frac{-0.09}{0.04} = -\frac{9}{4} = -2.25$$

④ p-value corresponding to Z-value of -2.25

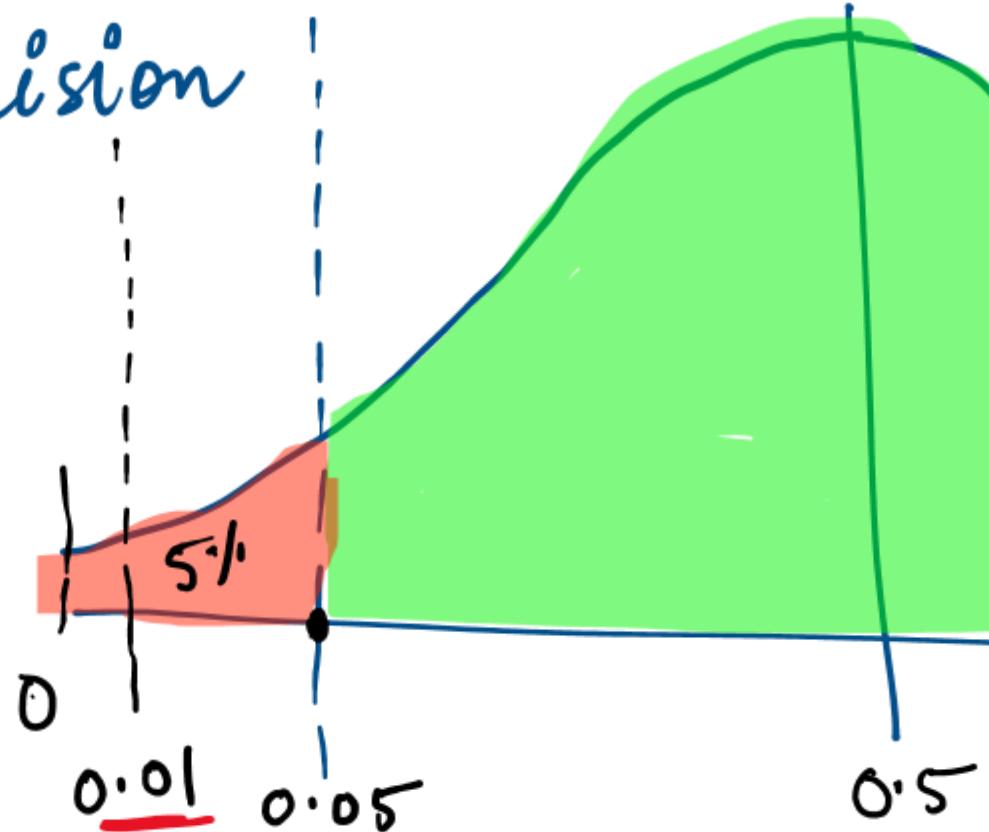
p-value = 0.0122

(Z-table)

→ Reject H_0 ✓

Spam < 0.2

Buy the software



→ Errors :-

