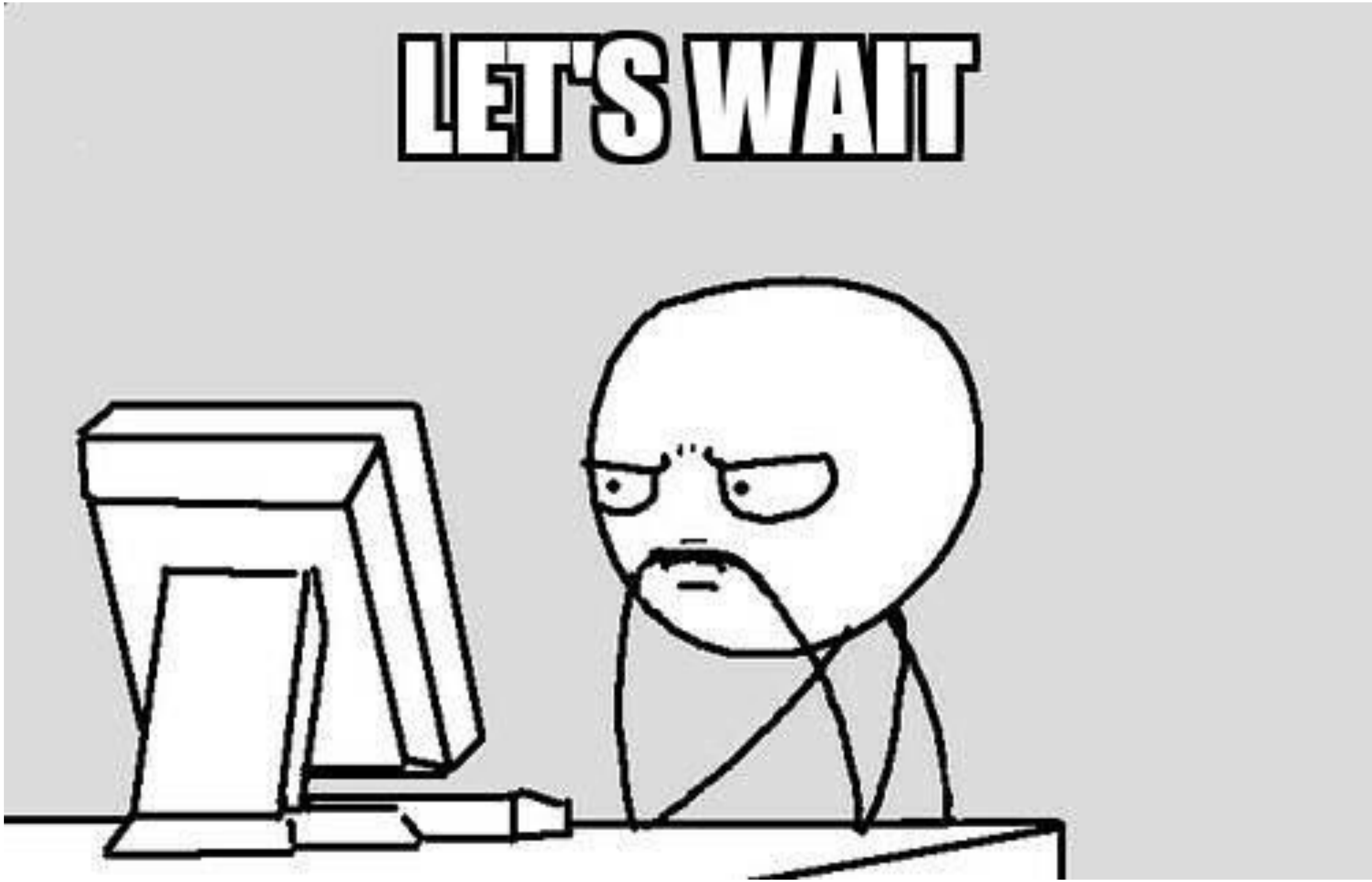
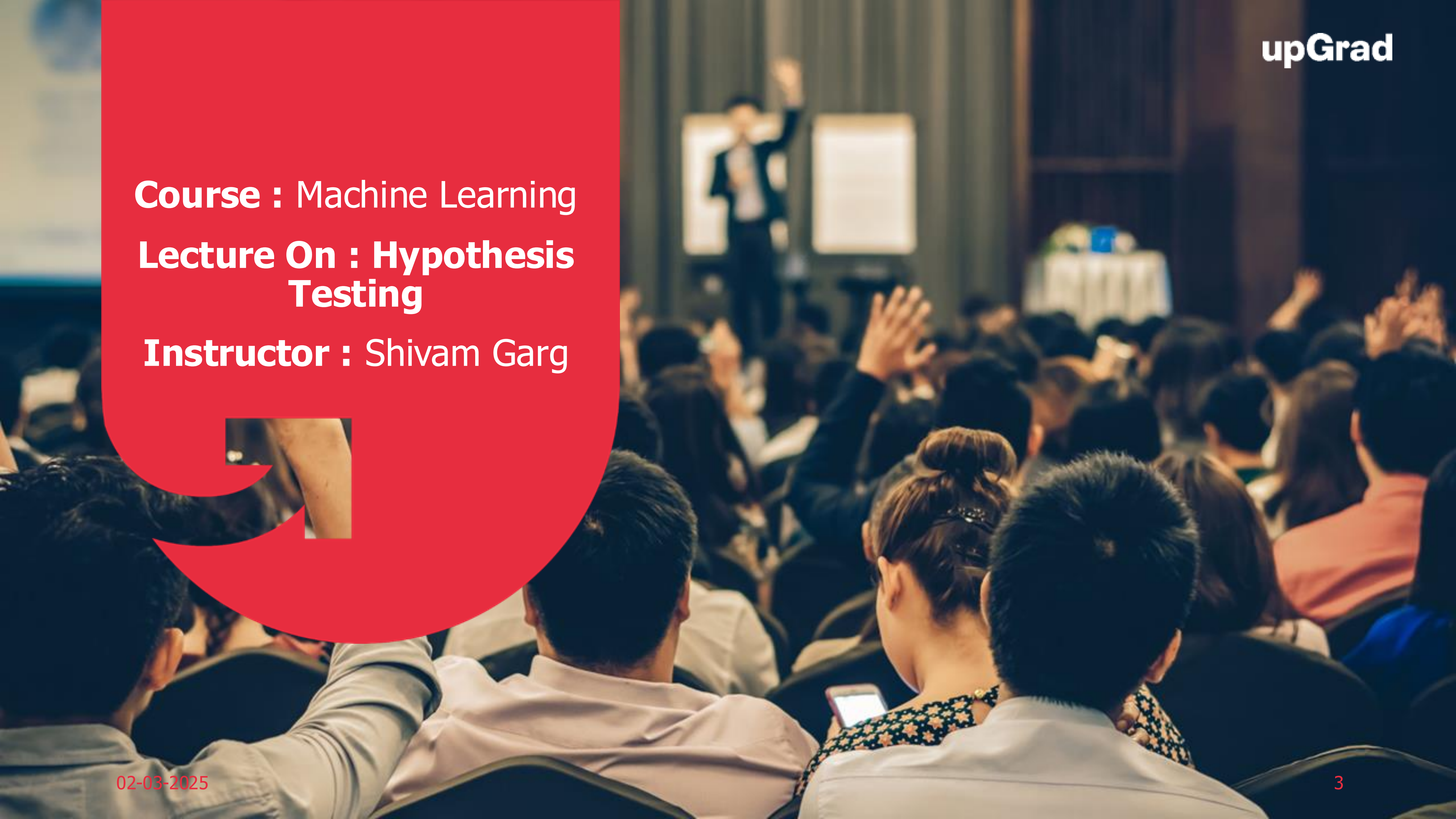

LET'S WAIT

#LifeKoKaroLift

Data Science Certification Program



Course : Machine Learning

Lecture On : Hypothesis Testing

Instructor : Shivam Garg

No validation

Problem Statement ✓

↓

Inferential Statistics (Sample Study) (CLT)

↓

Inferences (hypothesis) } (hypothesis about Population)

↓

Validation

Hypothesis testing (Perform the test)

↓

Whether hypothesis is correct or not → Decision making

Hypothesis Testing: (validated)

Inferences about population derived with the help of inferential statistics.

H_0: Null hypothesis : { Status quo (Claim given)

H_1: Alternate hypothesis : { Opposite to status quo

(Theoritical definitⁿ)

{ Numericals (Symbol

$\{=, \geq, \leq\}$ $\rightarrow$ H_0

$\{\neq, >, <\}$ $\rightarrow$ H_1

always ✓

$\rightarrow$ Claim: FK will generate average ticket size of 2000?

T
$$\begin{cases} H_0 = \mu = 2000 \\ H_1 = \mu \neq 2000 \end{cases}$$

S
$$\begin{rcases} H_0 = \mu = 2000 \\ H_1 = \mu \neq 2000 \end{rcases}$$

$\rightarrow$ Claim: FK will not generate average ticket size of 2000

T
$$\begin{cases} H_0 = \mu \neq 2000 \\ H_1 = \mu = 2000 \end{cases}$$

S
$$\begin{rcases} H_0 = \mu = 2000 \\ H_1 = \mu \neq 2000 \end{rcases}$$

Hypothesis Testing:

upGrad

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 mi.

$$\left\{\begin{aligned} H_1 &\rightarrow \mu > 530\,\text{ml} \\ H_0 &\rightarrow \mu \leq 530\,\text{ml} \end{aligned}\right\}$$

formulatⁿ of hypothesis

$\rightarrow H_0: =, \geq, \leq$

$\rightarrow H_1: \neq, <, >$

$\rightarrow$ Right tail test

→ Hypothesis testing :- (Problem)

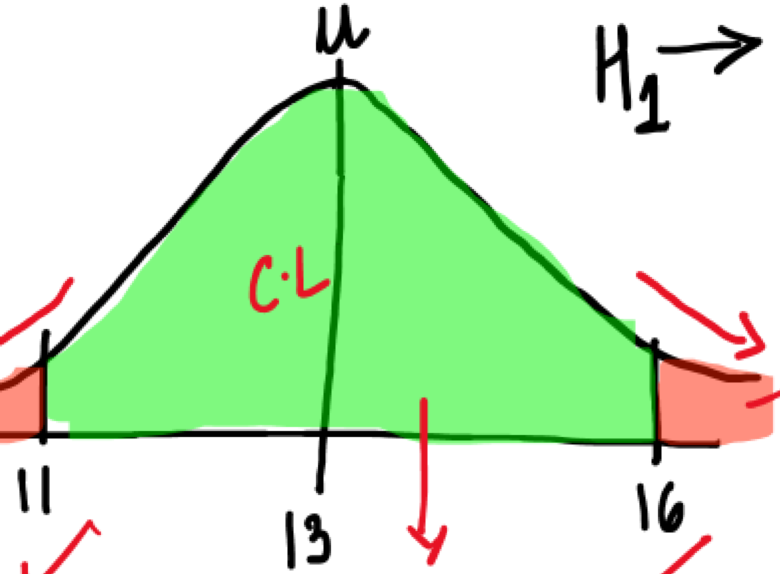
1. formulation of hypothesis (what)
2. Which test? (which)
   - 2-Tail
   - Left
   - Right
3. Methodology (CVM / p-value method) (how)
4. Decisioning

Type of test

Two tail Test

$H_0 \to =$

$H_1 \to \neq$



Rejected area

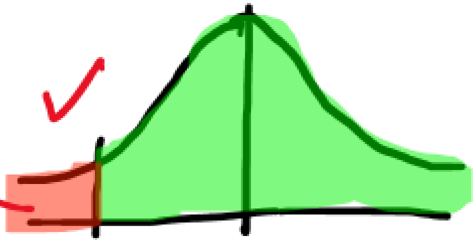
Accepted area

One Tail test

left tail test

$H_0 \to \geq$

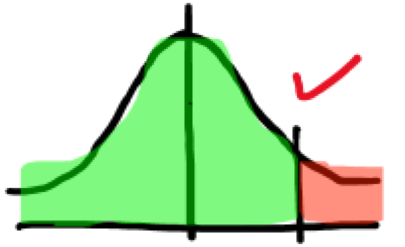
$H_1 \to <$



Right tai test

$H_0 \to \leq$

$H_1 \to >$



Confidence level = 95%

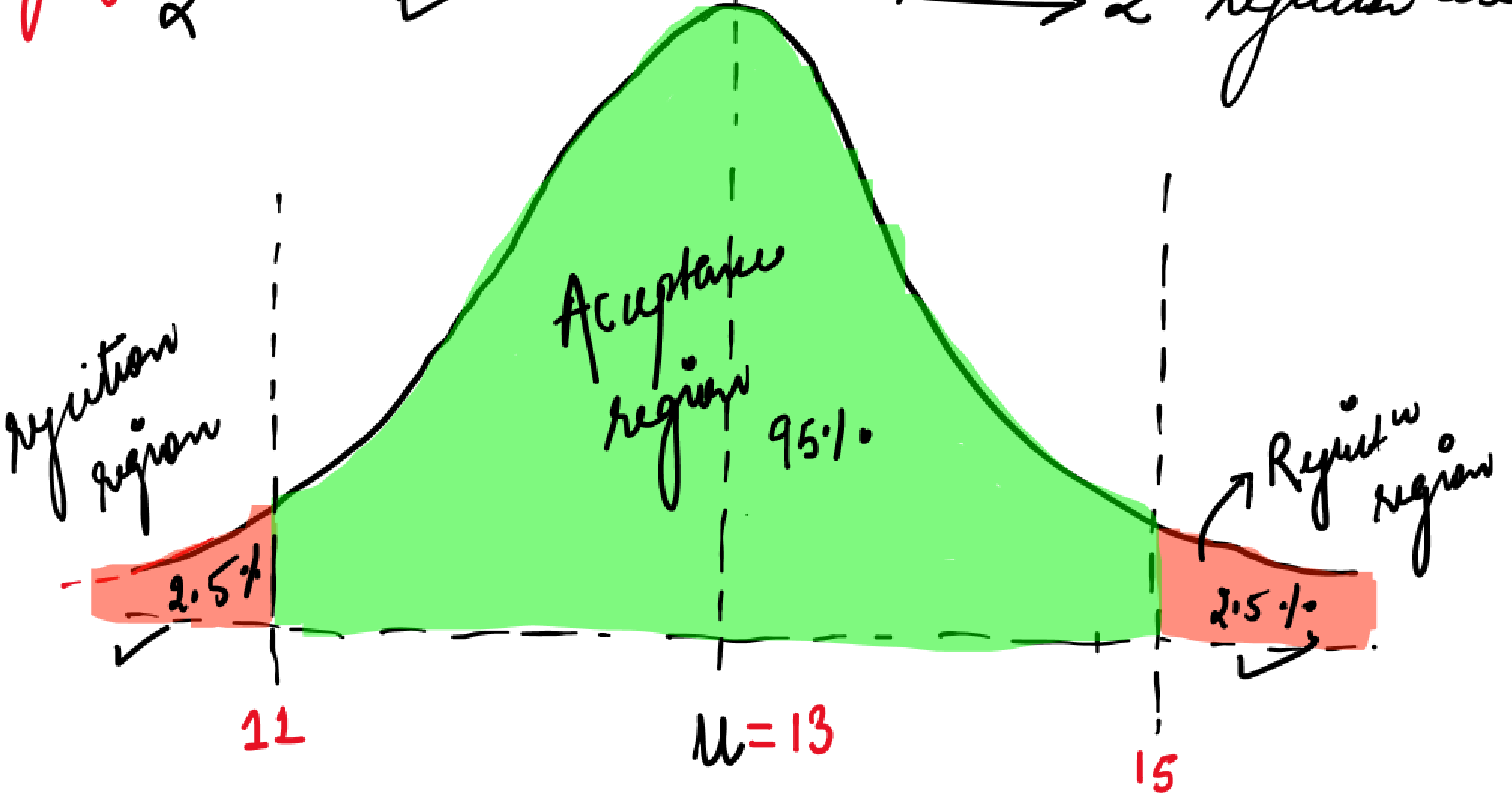
significance level = 5%

α

Two Tail Test

$\rightarrow \left\{ \boxed{H_0 \rightarrow = \quad H_1 \rightarrow \neq} \right\}$

$\rightarrow$ 2 rejection area

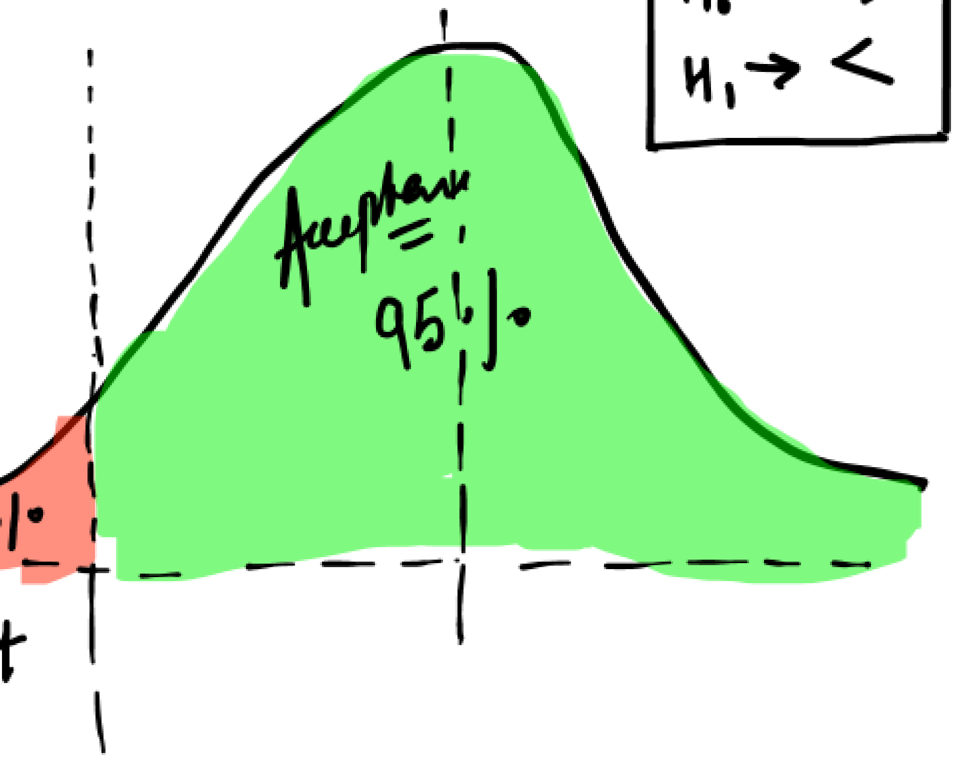


left Tail

$C.f = 95\%$

$H_0 \rightarrow \geq$

$H_1 \rightarrow <$



Lower tail test

$C.L = 95\%$

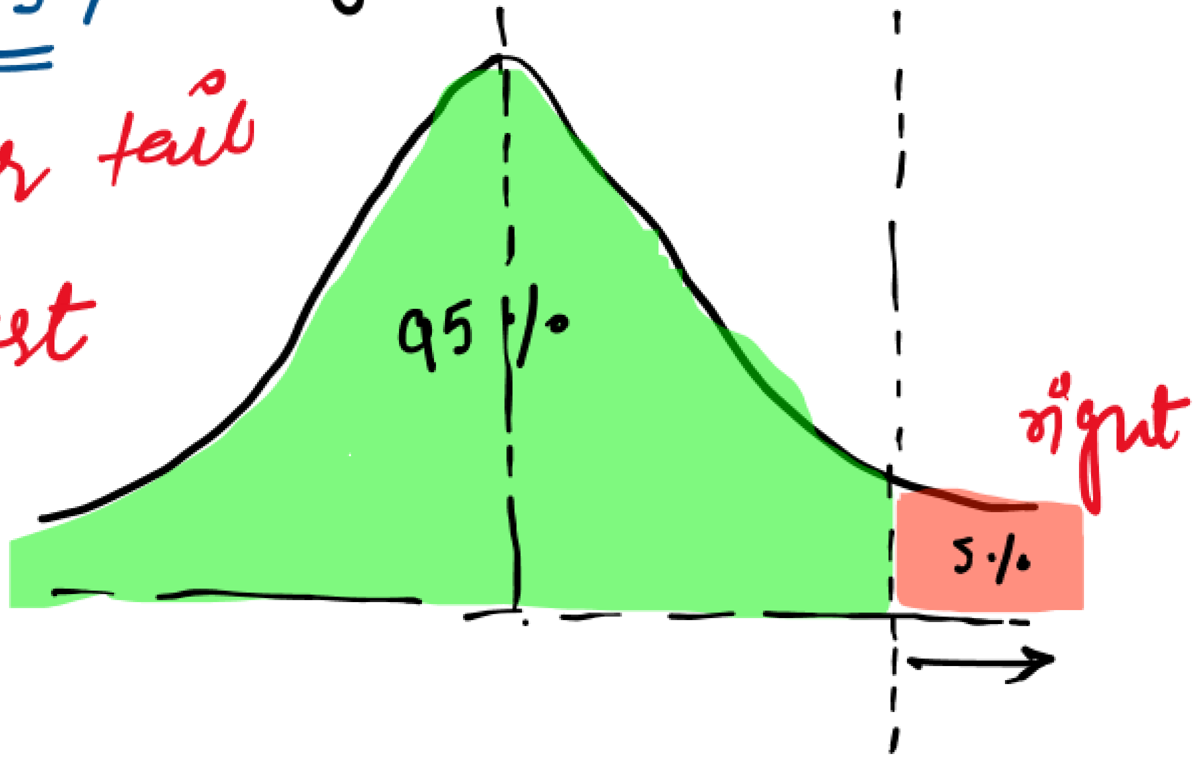
$SL = 5\%$

upper tail test

Right Tail

$H_0 \rightarrow \leq$

$H_1 \rightarrow >$



$H_1 \longrightarrow \neq \longrightarrow$ 2-tail test

$H_1 \longrightarrow < \longrightarrow$ Left tail test

$H_1 \longrightarrow > \longrightarrow$ Right tail test

Two methods

- Critical Value method
- p value method

→ Critical Value method :-

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

$$U.C.V = \mu + z \times \frac{\sigma}{\sqrt{n}}$$

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)

2-tail test → L.C.V and U.C.V

Left tail test → L.C.V

Right tail test → U.C.V

Inferential Statistics:

upGrad

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.

$n = 50$ ✓ → $C.L = 95\%$ (Green Area)

$\mu = 4.63$ ✓ → $\alpha = 5\%$

$\sigma = 0.54$ ✓

Significance level (100 - CL)

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 : $H_0 : \mu = 4.63$ cells/μL }

$H_1 : \mu \neq 4.63$ cells/μL

② Which test ? $\longrightarrow$ Two Tail test

③ $L \cdot C \cdot V = \mu - Z \times \frac{\sigma}{\sqrt{n}}$

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

→ Range

Z-table

		Z
C.L ↑	99%	2.58
C.L ↓	90%	1.65

95% → 1.96

$\underbrace{30}_{\text{Answer (Sample)}} \pm \underbrace{3}_{\text{fix tolerance}}$

→ Error

→ 10%

$30 \pm \underbrace{\underbrace{Z}\times 3}_{\text{tolerance}}$

$\mu \pm \underbrace{Z} \frac{\sigma}{\sqrt{n}}$

Sampling error

→ Error

$X = \mu \pm Z \times \frac{\sigma}{\sqrt{n}}$

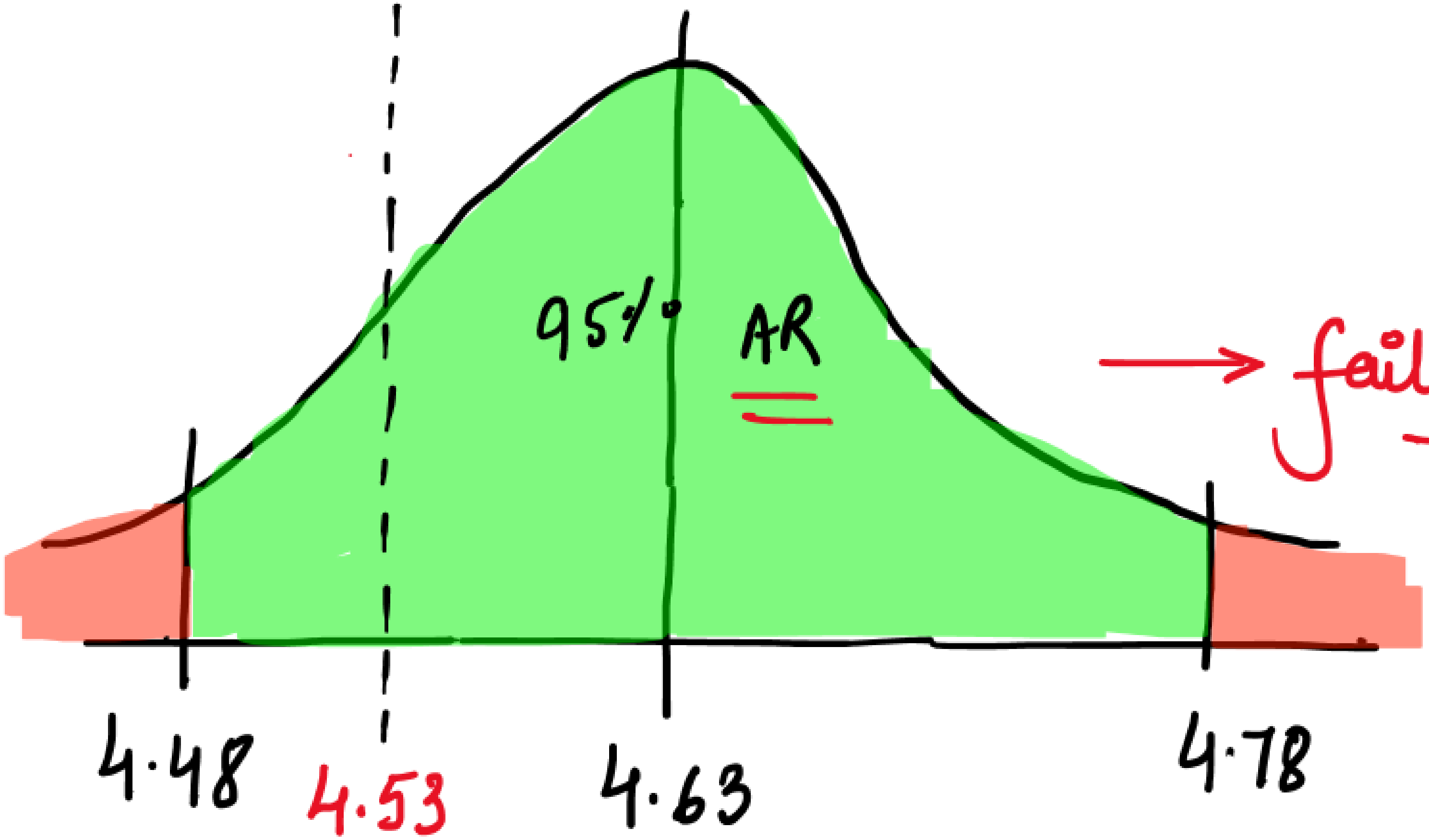
$$L.C.V = 4.63 - 1.96 \times \frac{0.54}{\sqrt{50}}$$

$$= 4.48$$

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

$$= 4.78$$

(4)



4.63 ± 0.15

→ fail to reject null hypoth

→ p-value method:-

① formulation of hypothesis ✓

② which test ✓

③ Calculate z-stat → $z = \dfrac{x - u}{\sigma / \sqrt{n}}$

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

$\longrightarrow H_0 \rightarrow \mu = 4.63, \quad H_1 \rightarrow \mu \neq 4.63$

$\longrightarrow$ Two-tailed test

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

$$= \frac{-0.1 \times \sqrt{50}}{0.54} = -\frac{10 \times \sqrt{50}}{54} = -1.31$$

$x = 4.53$

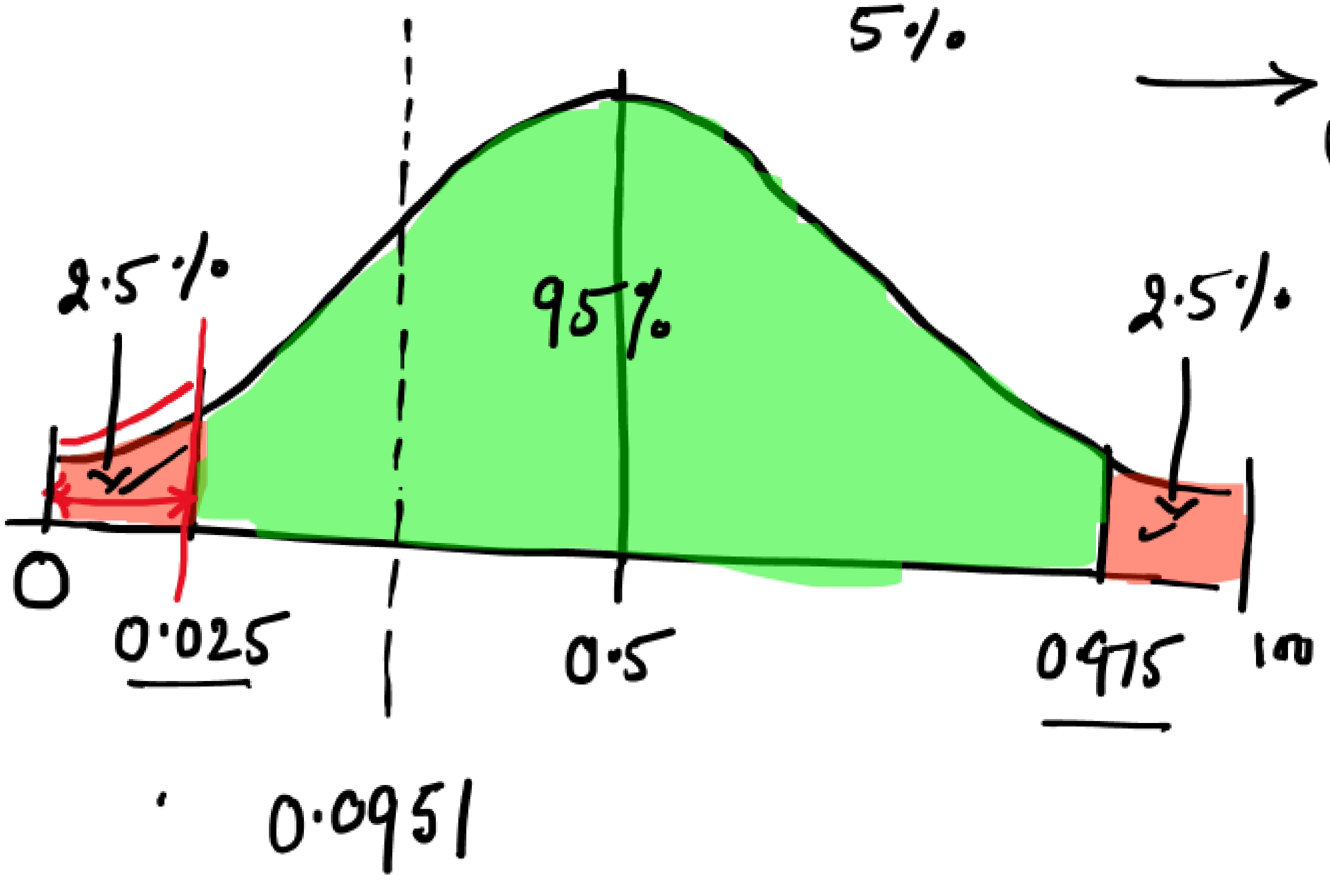
$\mu = 4.63$

$\sigma = 0.54$

$n = 50$

$\longrightarrow$ p-value (@ -1.31) $=$ $\underline{\underline{0.0951}}$

→



→ fail to reject null hypothesis

↓

accept null hypo

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

→ Type of Errors :-

$$\begin{tikzcd} & S \arrow[r] & gs \arrow[dl] \\ P \arrow[ur] \arrow[r] & S \arrow[r] & \text{Validate} \end{tikzcd}$$

✓

4B\$ → α → Type
Erro

Truth about the population

Decision based on sample

	H_0 true	H_a true
Reject H_0	Type I error	Correct decision
Accept H_0	Correct decision	Type II error

Claim ✓ → Rejected

Claim X → Accepted

Women (Pregnant) $\longrightarrow$ Self-kit $\longrightarrow$ Not Pregnant $\longrightarrow$ Type I error

$\downarrow$

Non-Pregnant $\longrightarrow$ Self-kit $\longrightarrow$ Pregnant $\longrightarrow$ Type II error

ML $\longrightarrow$ Logistic Reg

→ A/B Testing :-

Changes (Website UI)

5000 Applied → Test

5000 Not applied → Control

Test ↓

① → Time spent on app ↓ ↑ → Time spent on App

→ AOV ↓ ✗ ↑ ✓ → AOV

→ Clicks Number ↓ ↑ → Click number

Control ↓

fintech

$\longrightarrow$ Def $\xrightarrow{HT}$ Not-def (Concern) Entire Amount ✓

$\longrightarrow$ Non-Def $\xrightarrow{HT}$ Def (Very less concern) Interest Amt ✓

Telecom

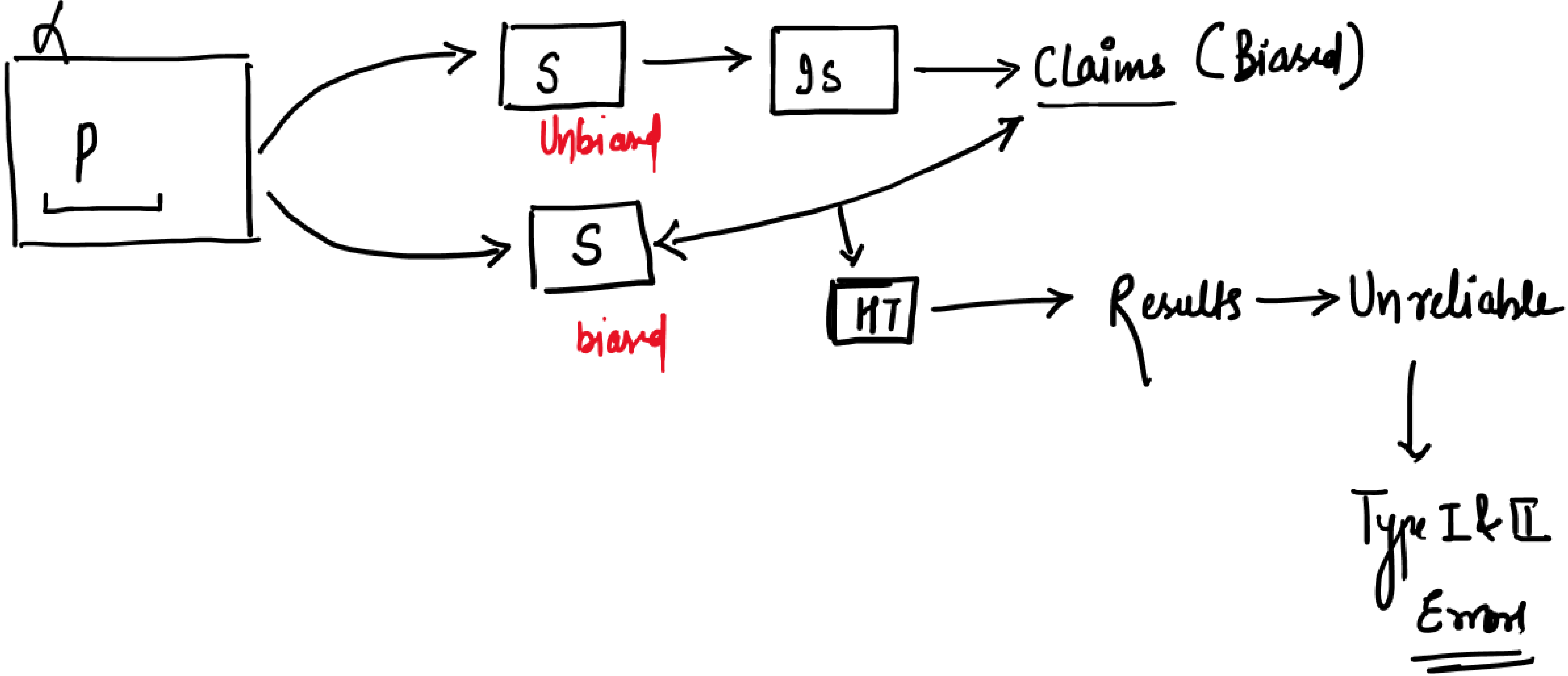
Churn $\xrightarrow{HT}$ Non-Churn (Concern) ✓

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

H

C $\longrightarrow$ NC (Concern)

NC $\longrightarrow$ C (Concern)



α
P
S
Unbiased
gs
Claims (Biased)
S
biased
HT
Results
Unreliable
Type I & II
Error

W = 5kg → hypothesis is perfectly right

X

Block → claim → upto 10kg

HT ✓

α =

Population

Sampling (Bolt)

Biased → 1α

Biased → 100 ✓

100

95

HT → X (No need to w

① $\mu = 4.63 - H_0$

$\mu \neq 4.63 - H_1$

② 2-tail test

$x = 4.53$

$\mu = 4.63$

$\sigma = 0.54$

$n = 50$

③ $Z\text{-stat} = \dfrac{x-\mu}{\sigma/\sqrt{n}} = \dfrac{4.53-4.63}{0.54/\sqrt{50}} = -1.31$

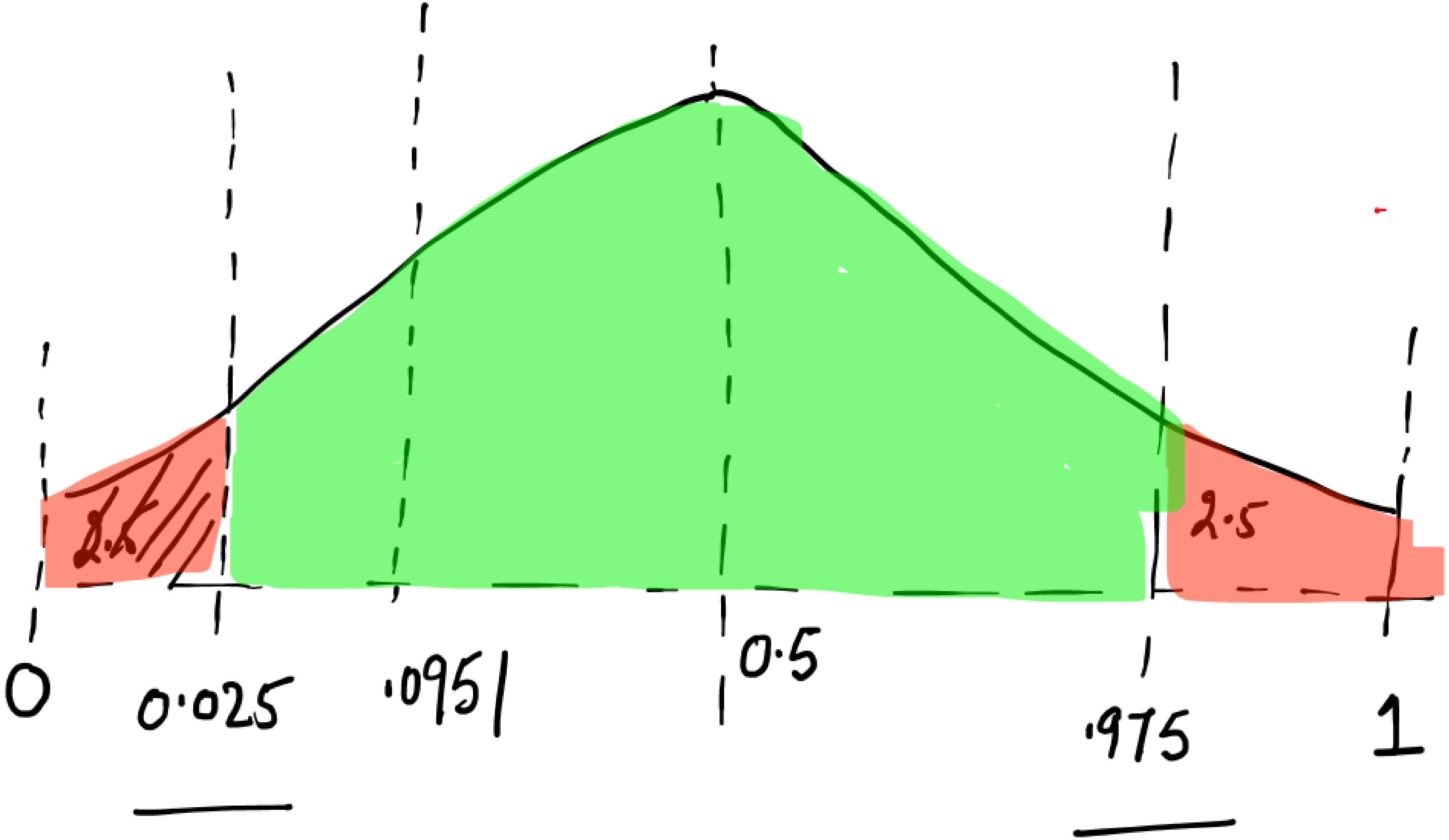
(Z-table)

④ p-value $= 0.0951$

(Z-table)

$CL = 95$

$\alpha = 5\%$



fail to reject the null hypo

$\alpha = 5\%$ p-value method

SUPERMARKET LOYALTY PROGRAM

(a) formulat^n: $H_1: \mu > 120$

$H_0: \mu \leq 120$

- A supermarket plans to launch a loyalty program if:
  $Avg.\, spending\ per\ shopper > ₹120\ per\ week$

(b) Which test → Right tail test

- 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.

Should the supermarket launch the loyalty program?

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

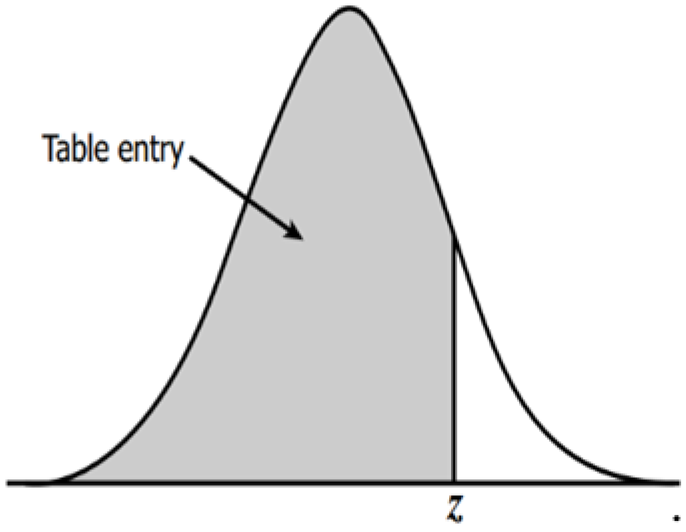
Z table.

LOOKING UP Z-TABLE

z	.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

~~For z-value of 2.236, table entry = 0.9871~~

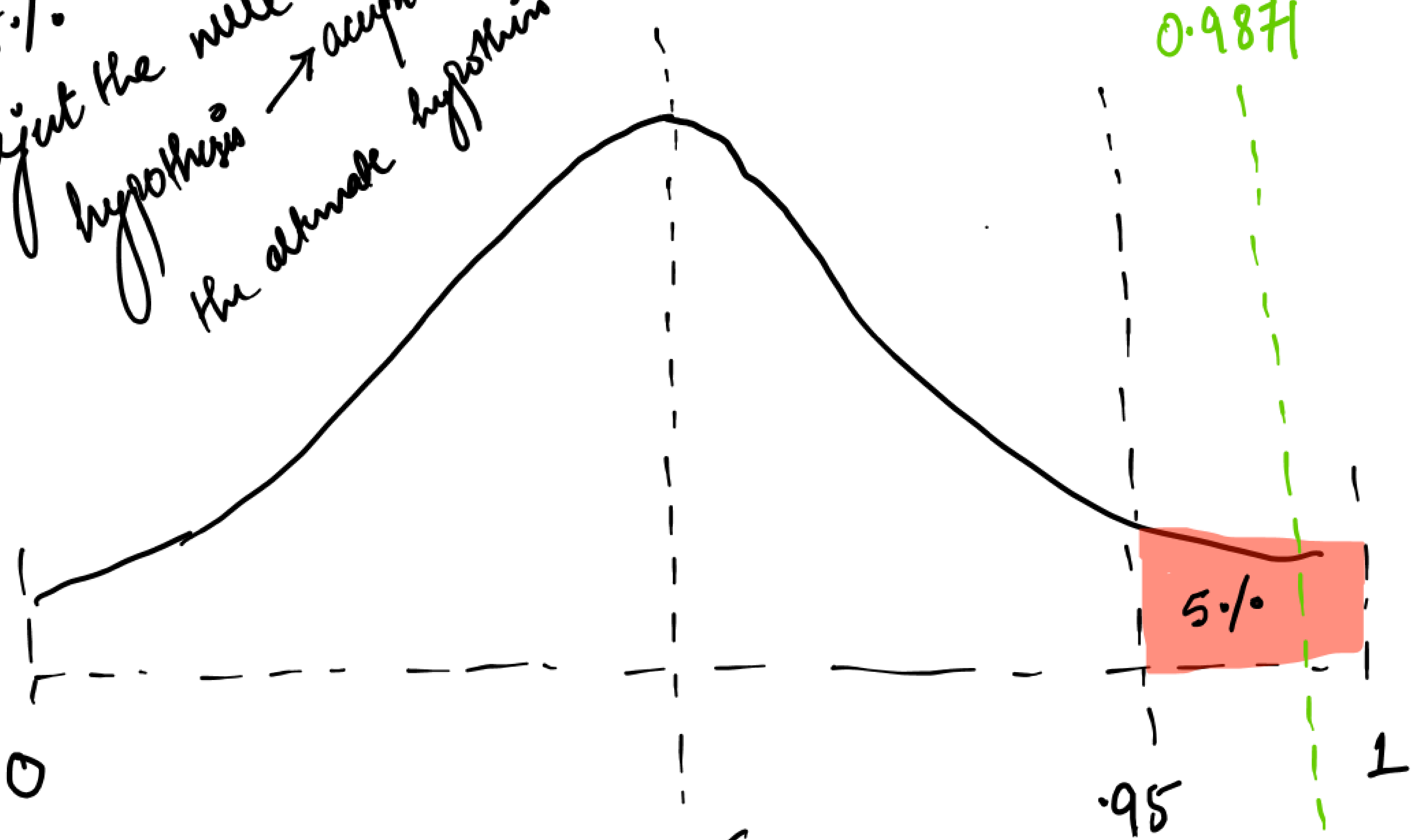
~~p-value $P(\bar{X} \geq z)$ = unshaded region = 1 - 0.9871 = 0.01~~



= 0.9871

$\alpha = 5\%$

reject the null hypothesis → accept the alternate hypothesis



→ 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.

$11/100 = 11\%$

- Assume significance level α = 0.05
- $\sigma = 0.4$

Should the company buy the spam filter?

Yes

① formulating hypothesis :-

H_1 : spam < 0.2

H_0 : spam ≥ 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

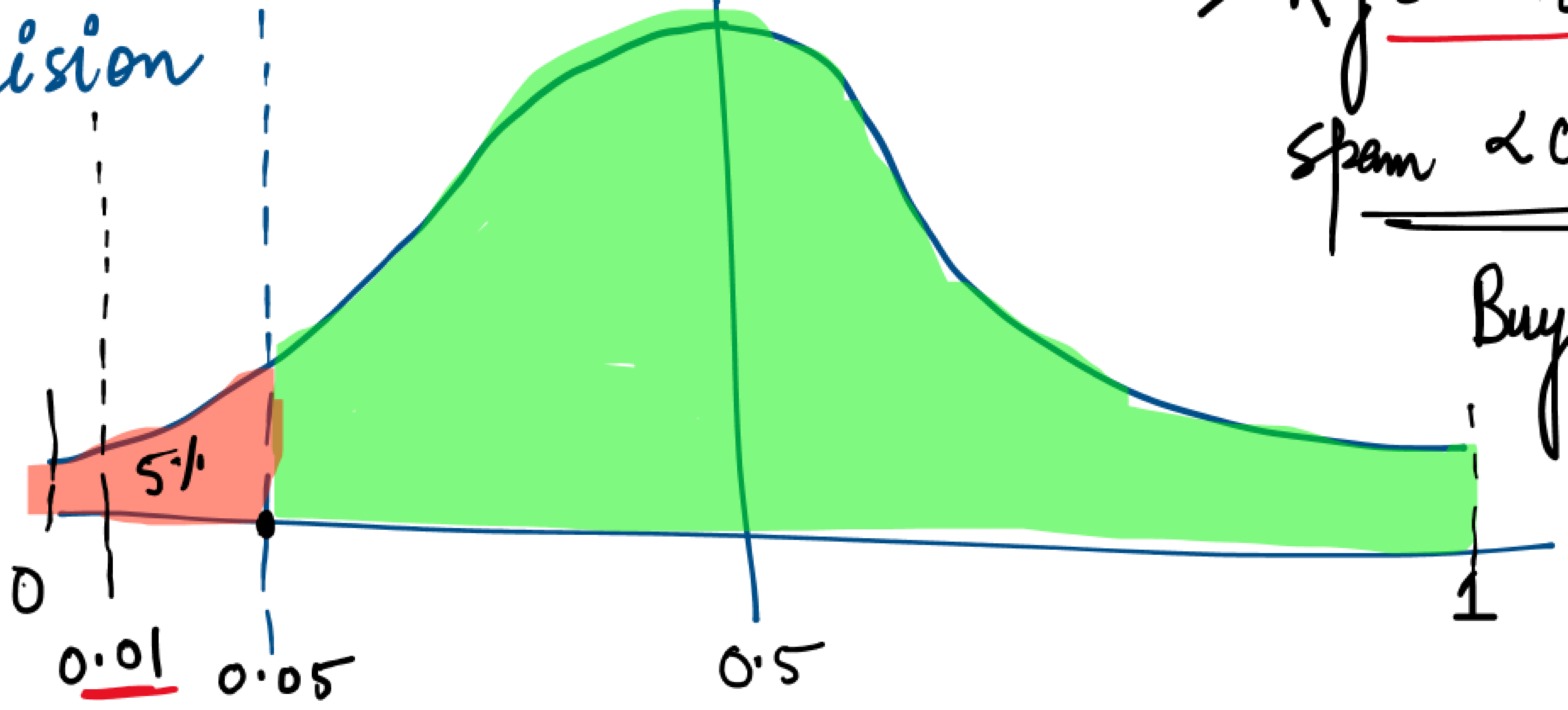
(Z-table)

→ Reject H_0

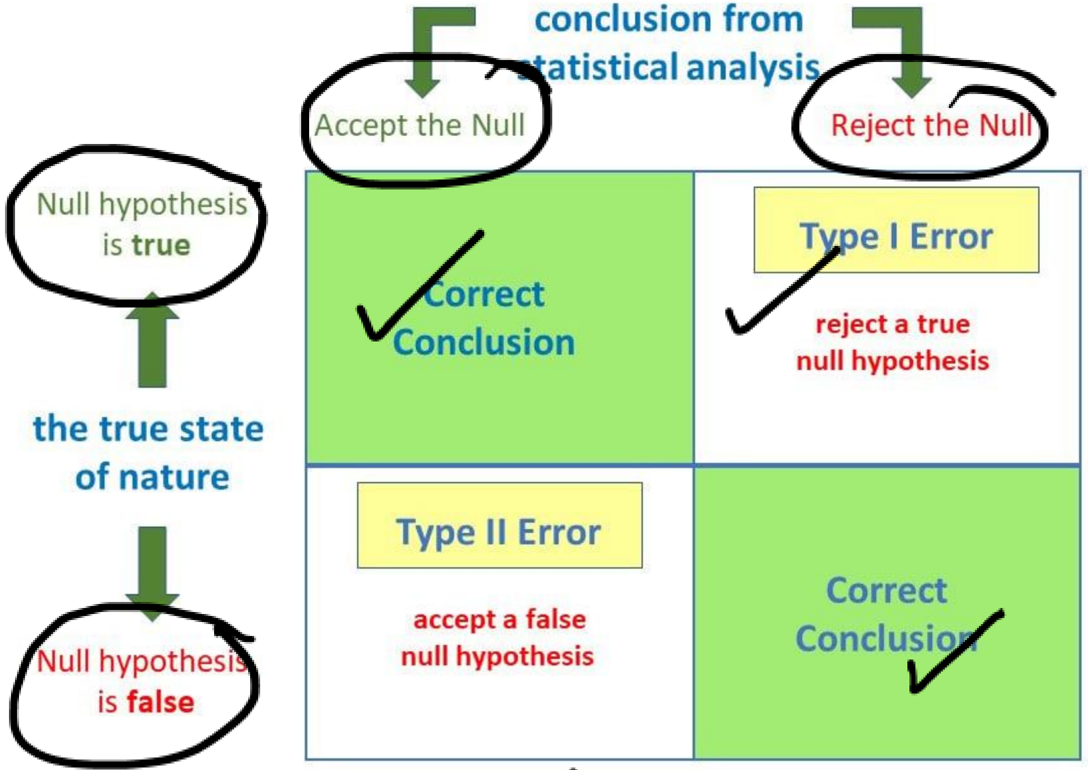
⑤ Decision

spam < 0·2

Buy the softw



→ Errors :-



N — True → Reject

N → False → Accept