upGrad

Course : Machine Learning

Lecture On :Course 1 Revision

Instructor : Shivam Garg

#LifeKoKaroLift

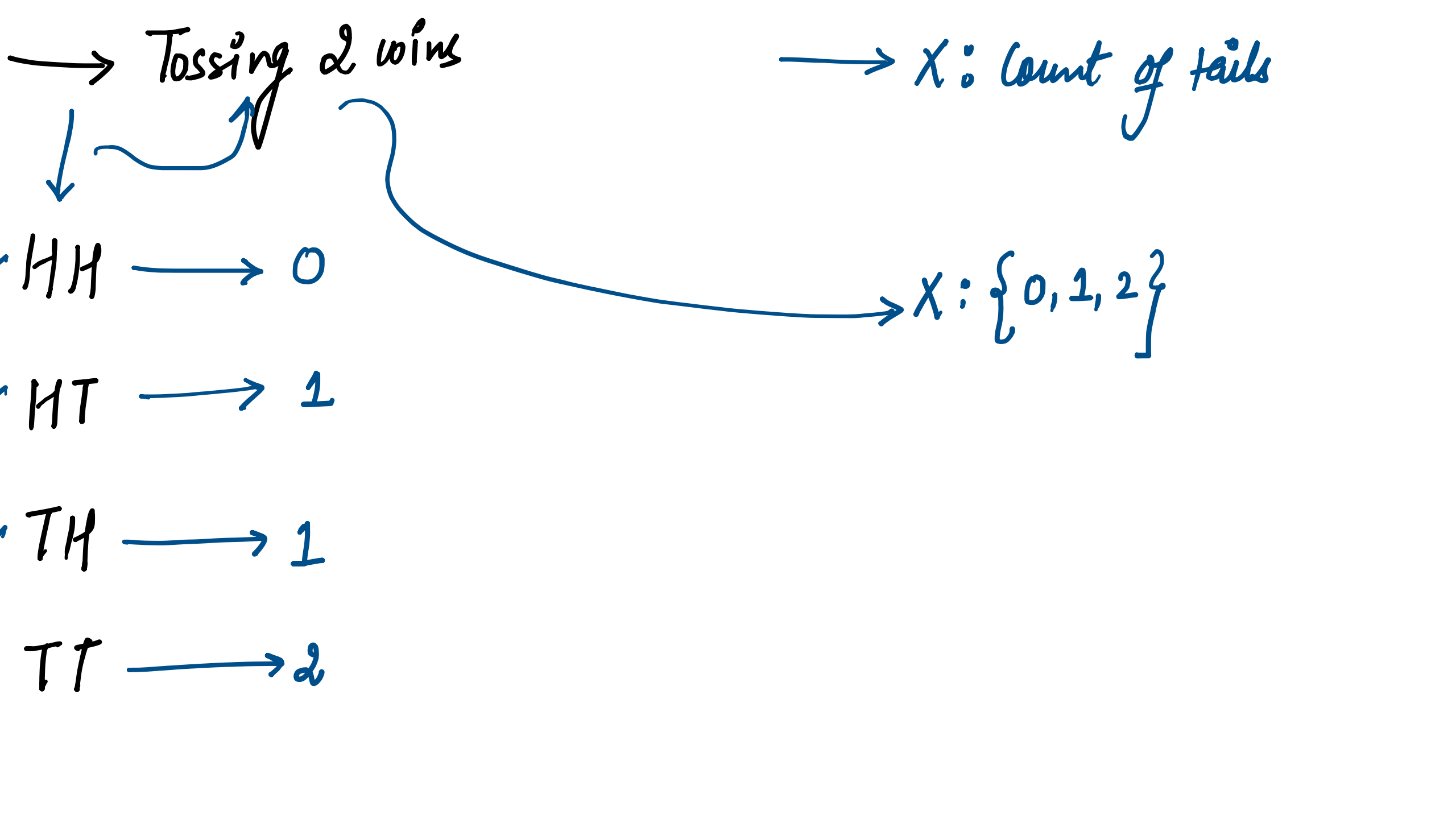
Data Science Certification Program

Random Variable: (Numerical formulation)

upGrad

* It is a variable which maps the o/p of a random process to some numerical value.

→ Random process is a process whose o/p is completely random


Tossing 2 coins
X : Count of tails
HH → 0
HT → 1
TH → 1
TT → 2
X : {0, 1, 2}

Expected Value: It is the average value of random variable if random process is repeated multiple times.

upGrad

Mathematically,

$$E[x] = \sum x_i p_i$$

$x_i \rightarrow$ Value of random variable

$p_i \rightarrow$ Corresponding probabili

$$E[x] = x_0 p_0 + x_1 p_1 + x_2 p_2 + x_3 p_3 + \text{- - - - - -}$$

Ex- Tossing 2 coins (X: Count of tails)

	x_i (h)	p_i
HH	0	1/4
HT	1	2/4
TH		
TT	2	1/4
		$\sum p = 1$

$$E[X] = x_0 p_0 + x_1 p_1 + x_2 p_2$$

$$= 0 \times \frac{1}{4} + 1 \times \frac{2}{4} + 2 \times \frac{1}{4}$$

$$\boxed{E[x] = 1}$$

Question-1: The 2010 U.S. Census found the chance of a household being a certain size. The data is in the table ("Households by age," 2013). What is the probability that a household will have at least 5 members?

Size of household	1	2	3	4	5	6	7 or more
Probability	26.7%	33.6%	15.8%	13.7%	6.3%	2.4%	1.5%

A ● 0.33
B ● 0.10
C ● 0.26

$P(S \geq 5) = P(5) + P(6) + P(7+)$

$= 6.3 + 2.4 + 1.5 = 10.2\%$

$= .102$

Question-2: A discrete probability distribution of scoring runs in one throw of a ball by a particular batsman in a cricket match is given in the table. Find the missing probability.

F	0	1	2	3	4	5	6
$P(F)$	0.15	0.27	0.35	0.15	0.04	0.01	x

$0.15 + 0.27 + 0.35 + 0.15 + 0.04 + 0.01 + x = 1$

A ● 0.3
B ● 0.02
C ● 0.03

$\sum p_i = 1$

$x = 0.03$

Probability:

upGrad

100 − 92 = 8%

92%

Question-3: A hockey goaltender has a save percentage of [redacted]. What would be the expected number of goals scored on this goaltender in a game where she faced 35 shots?

35 X 0.08 = 2.8

A ● 5.2
B ● 2.8
C ● 2
D ● 2.5

Question-4: Which of the following is not a property of a Binomial Experiment?

A ● All trials are identical.
B ● Each trial has only two possible outcomes.
C ● The probability of success may change from trial to trial. (not the correct property)
D ● The purpose of the experiment is to determine the number of successes that occurs during the n trials.

→ Binomial Distribution :-

$$P(X=r) = {}^{n}C_{r}\, p^{r} (1-p)^{n-r}$$

$n \longrightarrow$ no. of trials

$r \longrightarrow$ random variable

$p \longrightarrow$ probability

$$^{n}C_{r} = \frac{n!}{r!\,(n-r)!}$$

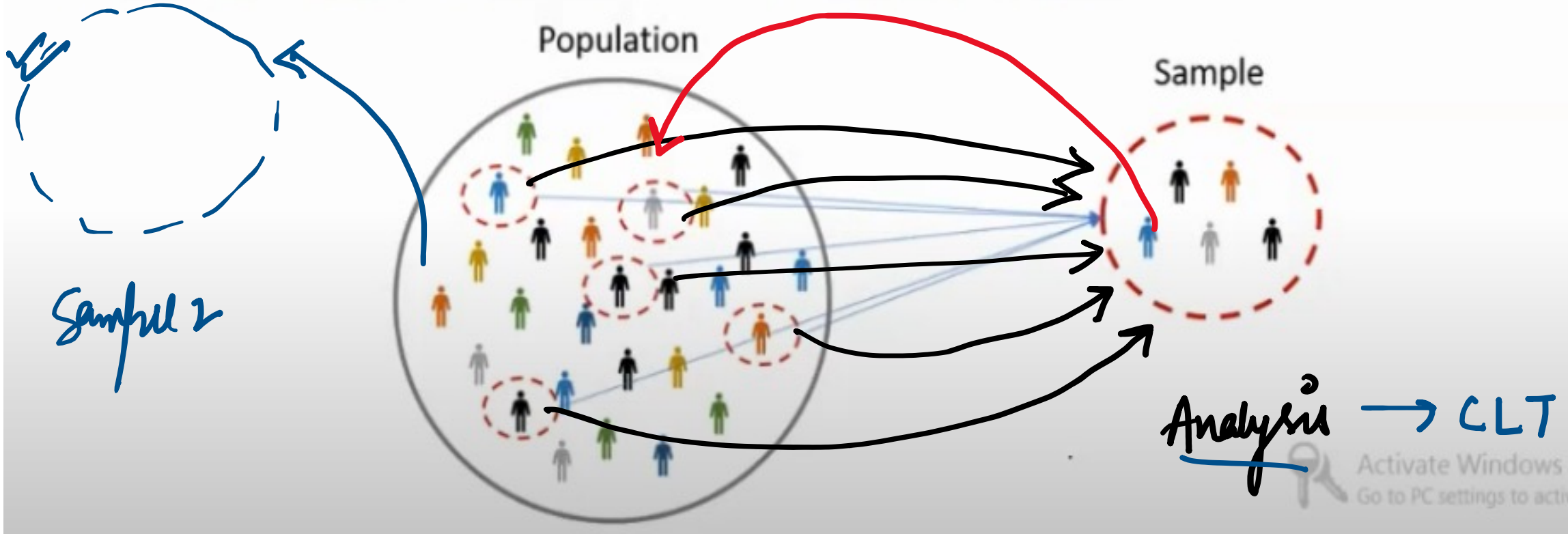
$n! = n \times n-1 \times n-2 \times \cdots \times 1$

$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

Central Limit Theorem: Approximate Analysis

upGrad

The Problem: Let's say that, for a business application, you want to find out the average number of times people in urban India visited malls last year. That's 400 million (40 crore) people! You can't possibly go and ask every single person how many times they visited the mall. That's a costly and time-consuming process. How can you reduce the time and money spent on finding this number?



Central Limit Theorem:

Range → Confidence Interval

upGrad

→ Population mean is given by:

300 ± 20 ✓ (tolerance / error)

$$C.I. = \left[\mu - z \times \frac{\sigma}{\sqrt{n}}, \mu + z \times \frac{\sigma}{\sqrt{n}} \right]$$

- μ → sampling distribution mean
- $\frac{\sigma}{\sqrt{n}}$ → standard error
- z → standard normal variate
- σ → pop. std
- n → sample size

Z-table

Confidence level → Z

300 ± 10 → 290 - 310

Confidence level	Z	$300 \pm Z \times 10$	
90%	1.65	$300 \pm 1.65 \times 10$	283.5 - 316.5
95%	1.96	$300 \pm 1.95 \times 10$	280.5 - 319.5
99%	2.58	$300 \pm 2.58 \times 10$	274.2 - 325.8

Central Limit Theorem:

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.

95% → $z = 1.96$

$n = 50$

$\mu = 4.63$

$\sigma = 0.54$

$$C.I = \left[\mu \pm z \times \frac{\sigma}{\sqrt{n}}\right]$$

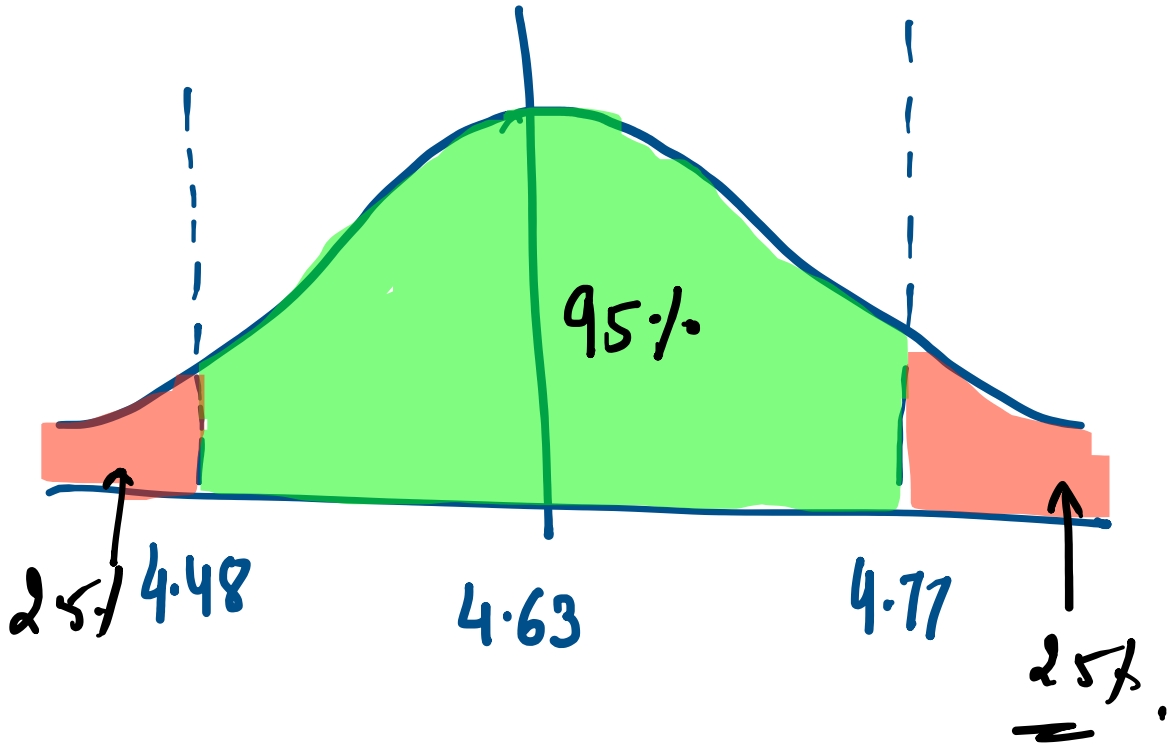
$$= 4.63 \pm 1.96 \times \frac{0.54}{\sqrt{50}}$$

Central Limit Theorem:

upGrad

$$C.I = 4.63 \pm 0.15$$

$$\boxed{C.I = 4.48, 4.77}$$



→ Hypothesis testing :- It is responsible for validating the hypothesis (claim) made during the inferential statistics.

H_0 → Null hypothesis → Status quo

H_1 → Alternate hypothesis → opposite to status quo

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

$H_1 \rightarrow >, <, \neq$

} Symbol basis decision in numericals

Hypothesis Testing:

H_0 & H_1 are disjoint events.
↳ nothing in common

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.

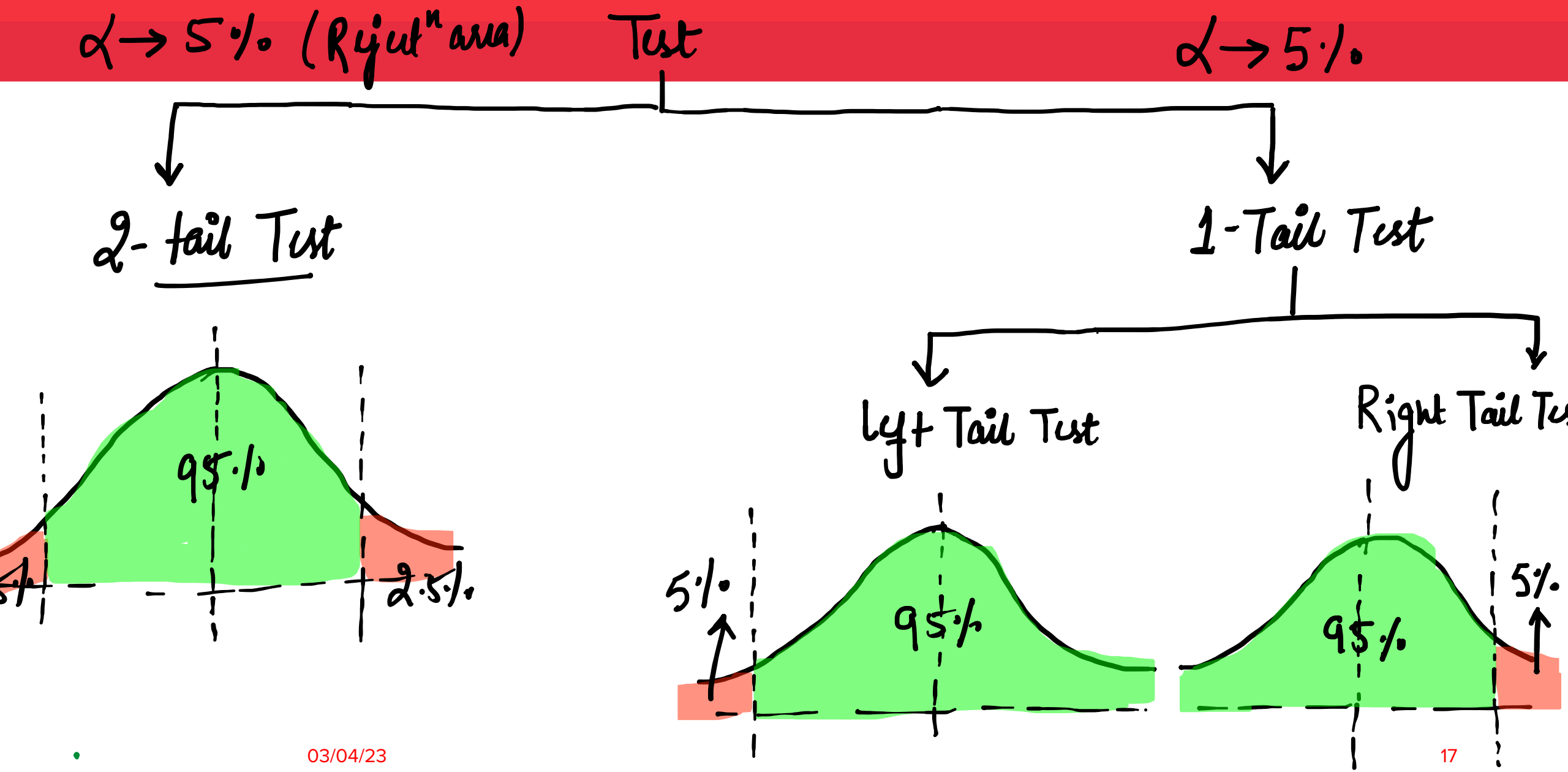
$\mu > 530$ ml $\longrightarrow H_1$

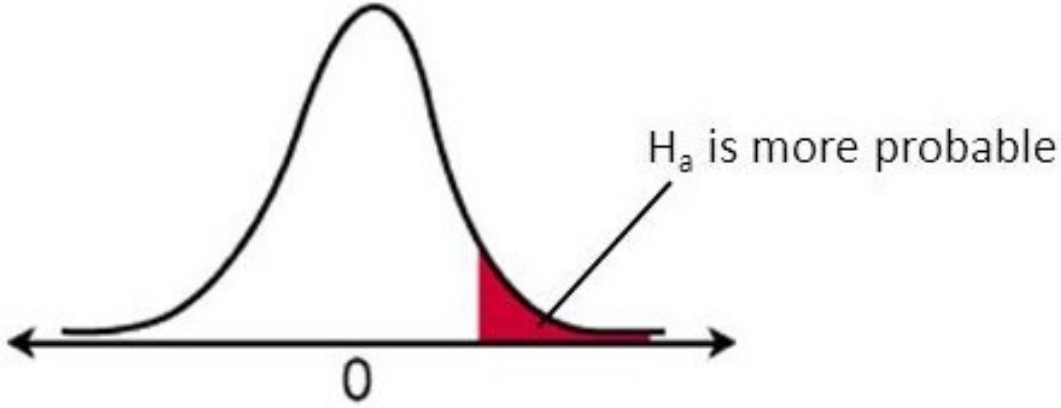
$\mu \leq 530$ ml $\longrightarrow H_0$

} Right Tail Test

Z-value and Area Under Curve:

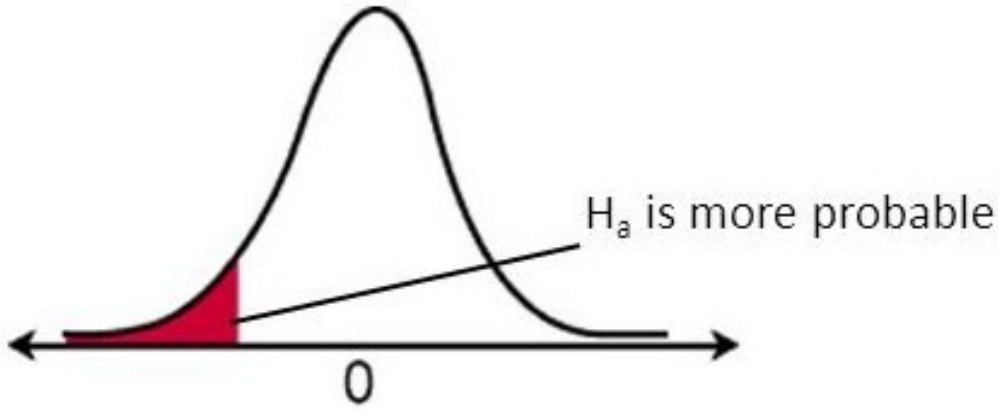
upGrad





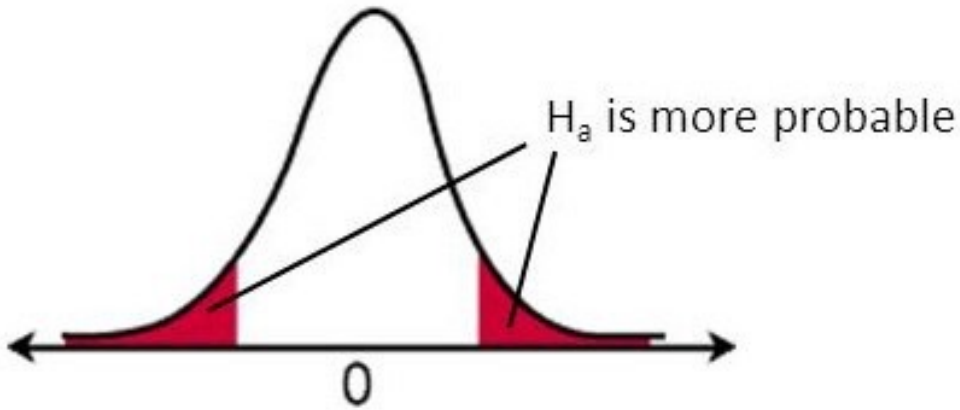
Right-tail test

$H_a: \mu > \text{value}$



Left-tail test

$H_a: \mu < \text{value}$



Two-tail test

$H_a: \mu \neq \text{value}$

Hypothesis Testing:

upGrad

Critical Value Method

A sample of 40 new baseballs had a bounce height mean of 92.67 inches and a SD of 1.79 inches. Use a .05 sig. level to determine whether there is sufficient evidence to support the claim that the new balls have bounce heights with a mean different from 92.84 inches. (a previous test figure).

$n = 40$

$\sigma = 1.79$

$\alpha = 5\%$ } $Z = 1.96$

$\mu \neq 92.84 \longrightarrow H_1$

$\mu = 92.84 \longrightarrow H_0$

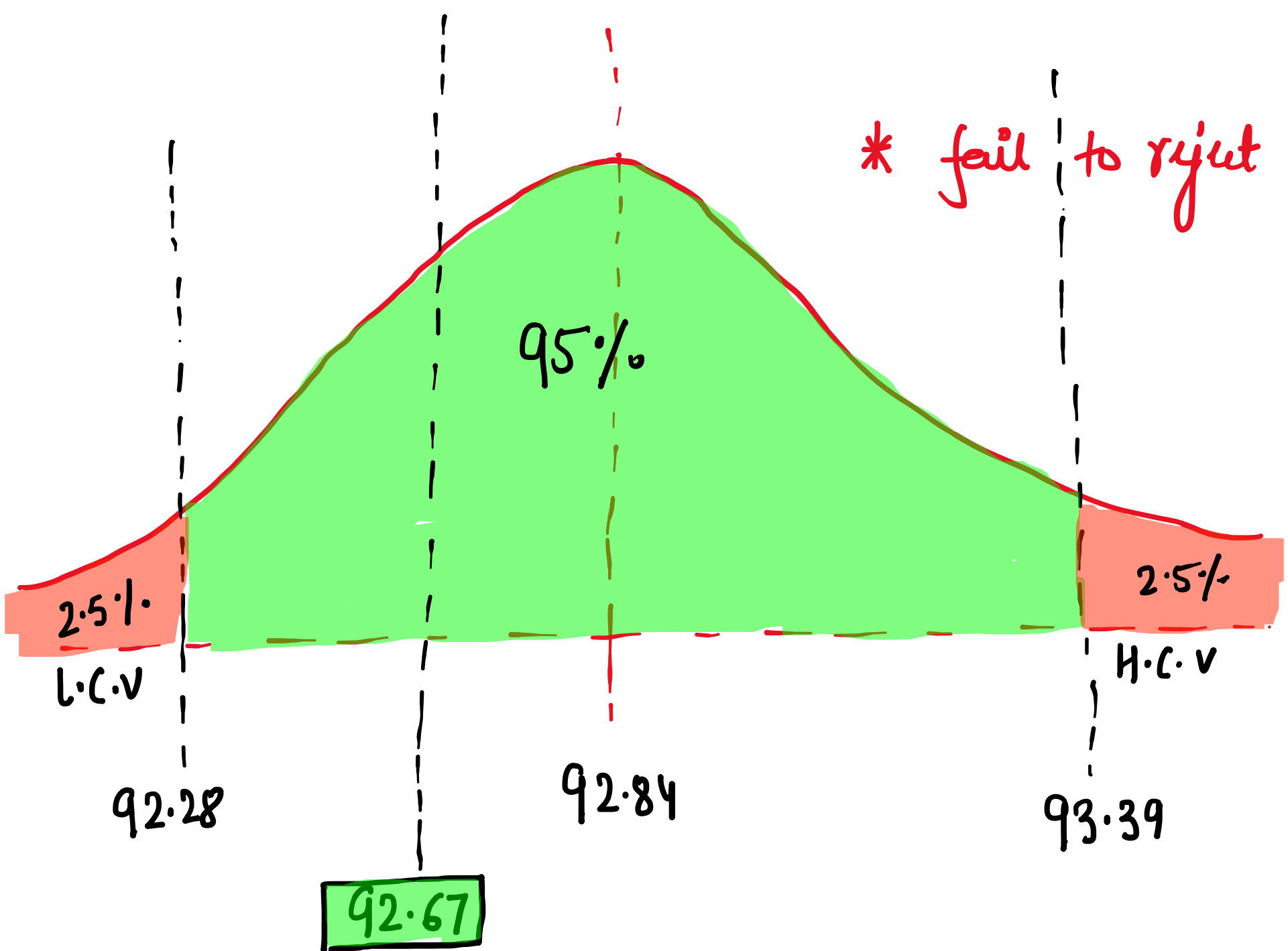
} Two tail Test

$$L.C.V = \mu - Z \times \frac{\sigma}{\sqrt{n}} = 92.84 - 1.96 \times \frac{1.79}{\sqrt{40}}$$

$$L.C.V = 92.28$$

$$H.C.V = \mu + Z \times \frac{\sigma}{\sqrt{n}} = 92.84 + 1.96 \times \frac{1.79}{\sqrt{40}}$$

$$H.C.V = 93.39$$


* fail to reject null hypothe
95%
2.5%
2.5%
L.C.V
H.C.V
92.28
92.67
92.84
93.39

EDA:

upGrad

Expenditures of a Company (in Lakh Rupees) per Annum Over the given Years.

Year	Item of Expenditure				
	Salary	Fuel and Transport	Bonus	Interest on Loans	Taxes
1998	288	98	3.00	23.4	83
1999	342	112	2.52	32.5	108
2000	324	101	3.84	41.6	74
2001	336	133	3.68	36.4	88
2002	420	142	3.96	49.4	98

Refer to the above table and answer the following questions

Question-1: The total amount of bonus paid by the company during the given period is approximately what percent of the total amount of salary paid during this period?

- 0.1%
- 0.5%
- 1%

$$\frac{\Sigma B}{\Sigma S} \times 100 = \frac{17.1}{1700} \times 100 \approx 1\%$$

EDA:

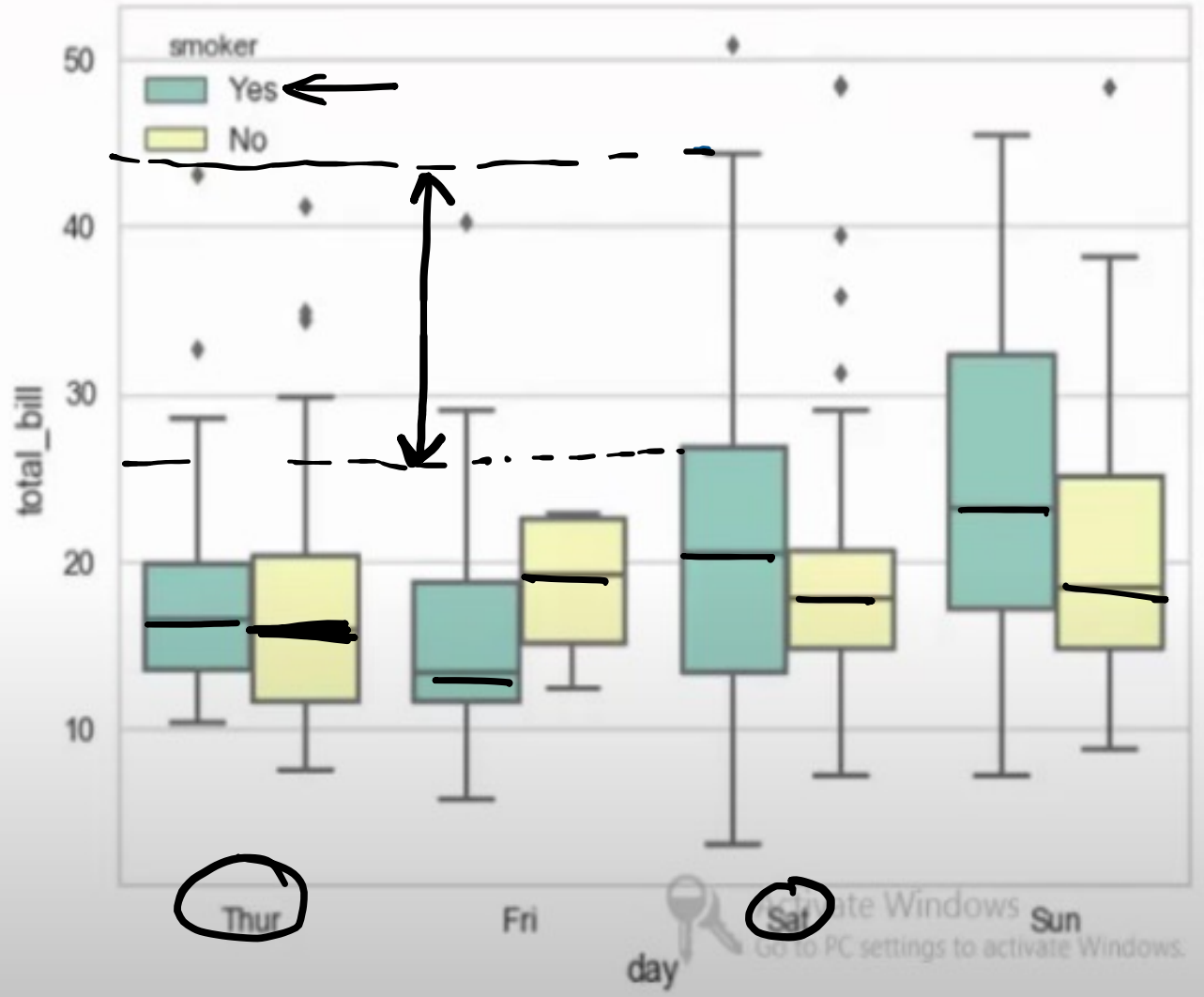
upGrad

Question-1: What is the upper bound range of the total bill for the Smokers on Saturday?

A ● 38-45
B ● 30-50
C ● 38-50

Question-2: On which particular day the median total bill for both Smokers and Non-Smokers is approximately same?

A ● Thursday
B ● Friday
C ● Saturday
D ● Sunday



Doubts:

upGrad