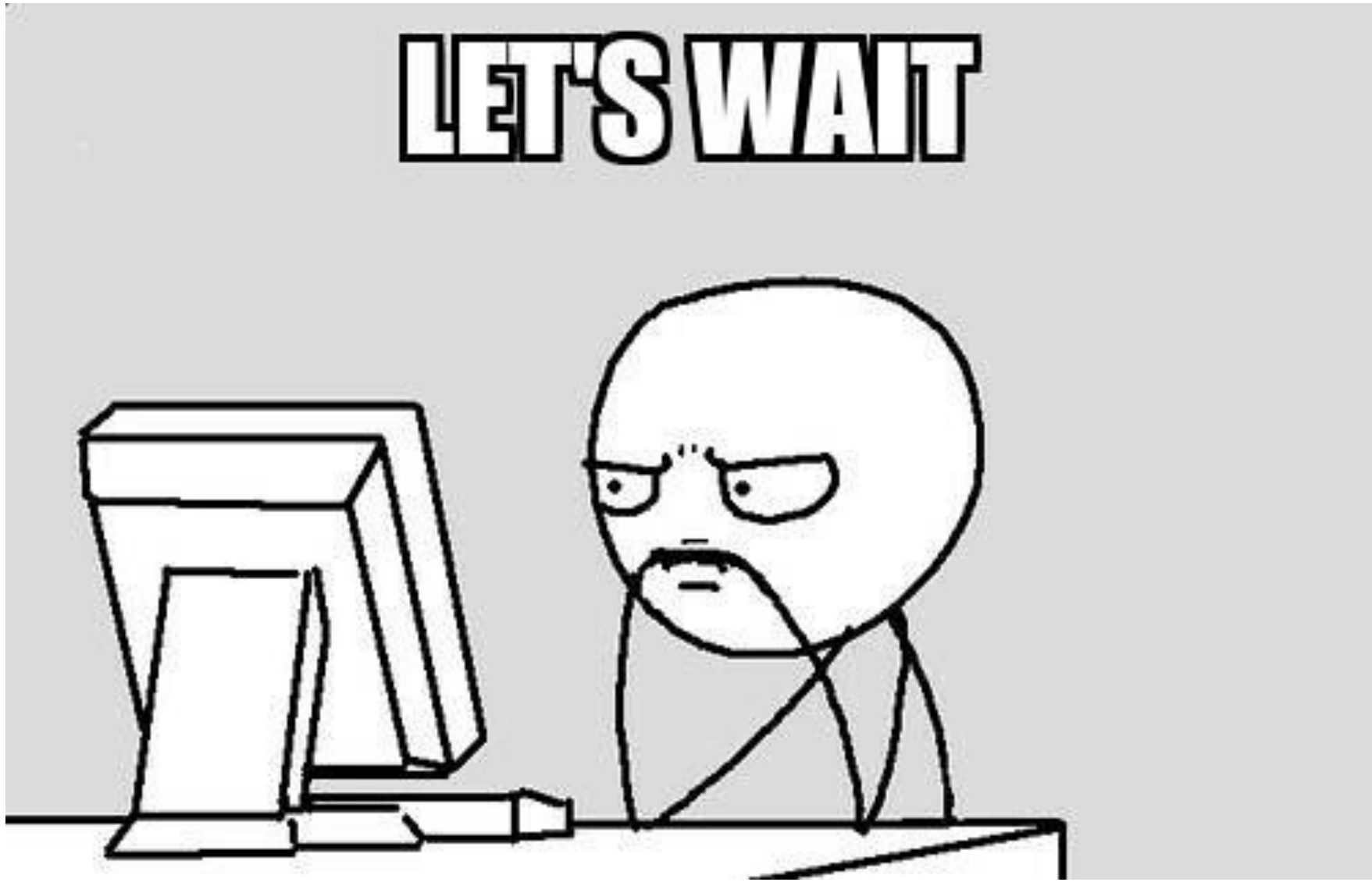

LET'S WAIT

upGrad

Course : Machine Learning

Lecture On :Inferential Statistics

Instructor : Shivam Garg

#LifeKoKaroLift

Data Science Certification Program

upGrad

Today's Agenda

1. Random Variable
2. Distributions – Normal & Binomial
3. Inferential Statistics
4. Central Limit Theorem
5. Doubt Session

Random Variable: Numerical formulation

upGrad

* It is a variable which maps the output of random process to some {numerical values.}

→ Random process: An event whose O/p is completely random.

Ex: ① Tossing a coin {H, T}

② Rolling a die {1, 2, 3, 4, 5, 6}

Tossing of 2 coins:-

(Random Process)

o/p	X_i
HH	0
HT	1
TH	1
TT	2

X : Count of tails

Random variable

$$X_i = \{0, 1, 2\}$$

Expected Value :- Average value of random variable if the random process is repeated multiple times.

Mathematically,

$$E[X] = \sum X_i p_i$$

X_i = Values of Random Variable

p_i = Corresponding prob.

$$E[X] = X_0 p_0 + X_1 p_1 + X_2 p_2 + \text{-----}$$

Tossing of 2 coins:-

(Random Process) X: Count of tails

o/p	x_i	p_i
HH	$x_0 = 0$	$1/4$ p_0
HT	$x_1 = 1$	$2/4$ p_1
TH	1	
TT	$x_2 = 2$	$1/4$ p_2
		1

$$E[X] = \sum x_i p_i = x_0 p_0 + x_1 p_1 + x_2 p_2$$

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

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

If you toss 2 coins on an average you will be getting one tail.

Problem Statement :- Game of catching the ducks (100)

	p_i	X_i
Small	0.60	$0.5
+		
Medium	0.30	$1
+		
Large	0.10	$5
	1	

$$E[X] = 0.6 \times 0.5 + 0.3 \times 1 + 0.1 \times 5$$

$$= 0.3 + 0.3 + 0.5 = \$1.1$$

Player

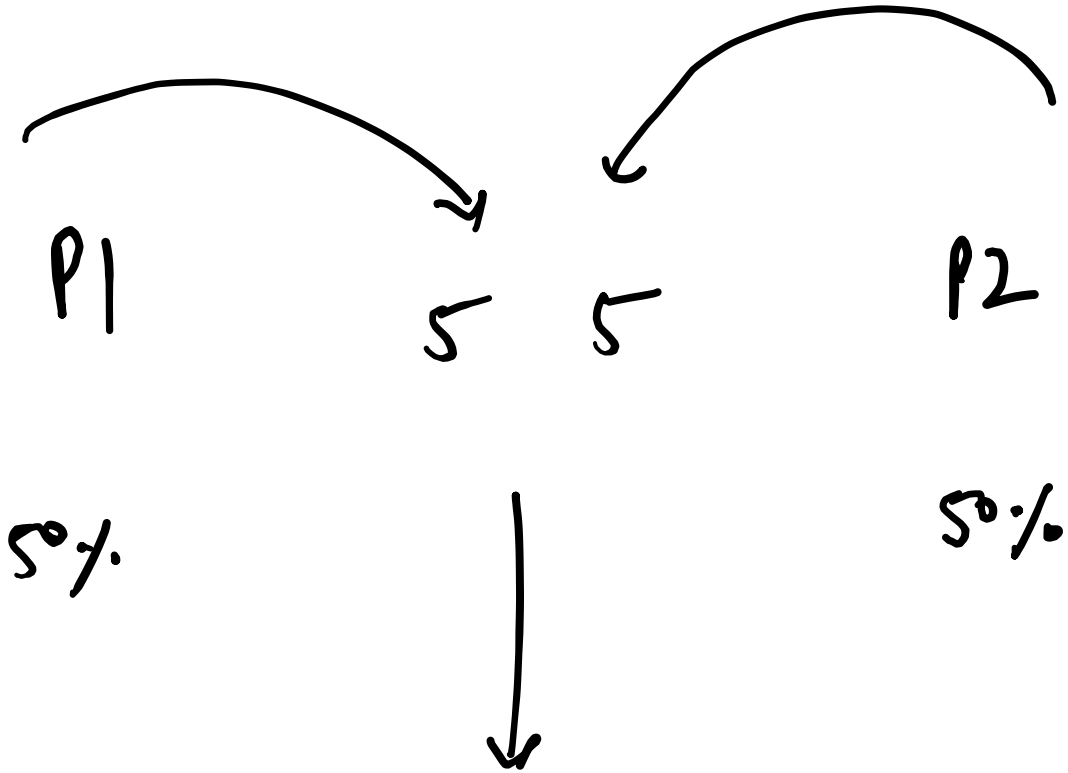
Case 1:- No knowledge of statistic

= $1

Case 2:- You have knowledge

= > $1.1

X = Winning amount



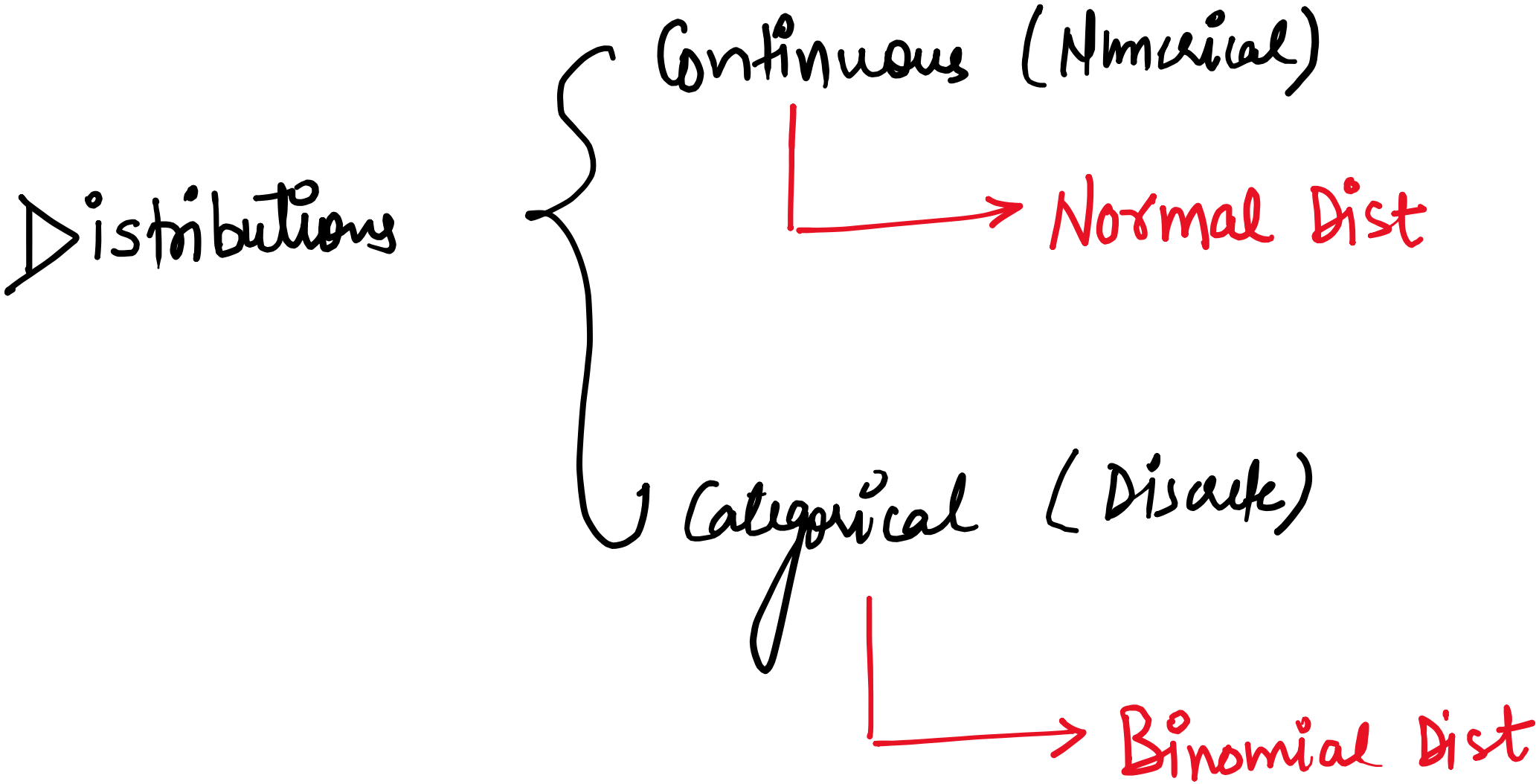
x_i	p_i
0	0.5
7	0.5

fair

$$E[x] = 0.5 \times 0 + 7 \times 0.5$$

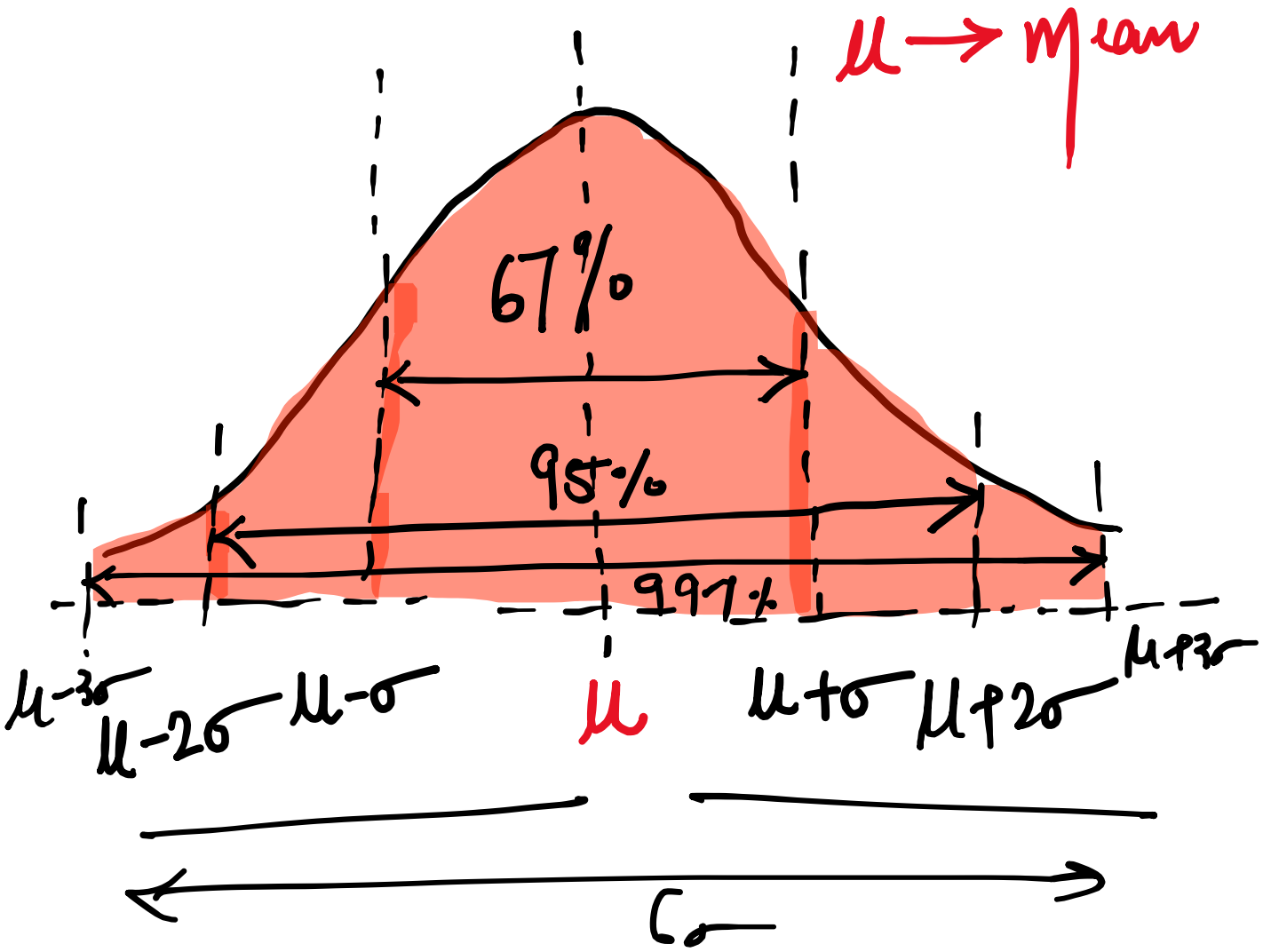
$$= ₹3.5 - ₹5$$

$$E[x] = -₹1.5$$


Distributions
Continuous (Numerical)
Normal Dist
Categorical (Discrete)
Binomial Dist

→ Normal Dist :-

* Bell shape curve
* Symmetry at centre



σ → std (Standard deviat^n)

(sq root of variance)

> 99.7

→ Binomial Distribut^{n} :- (Discrete distribut^{n})

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

* There should be only 2 categories (binary)
* No. of trials should be known.
* All trials should be identical to each other.

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

$n \rightarrow$ no. of trials

$r \rightarrow$ value of random variable

$p \rightarrow$ probability

→ Prob. of getting 2 tails while tossing 2 coins.

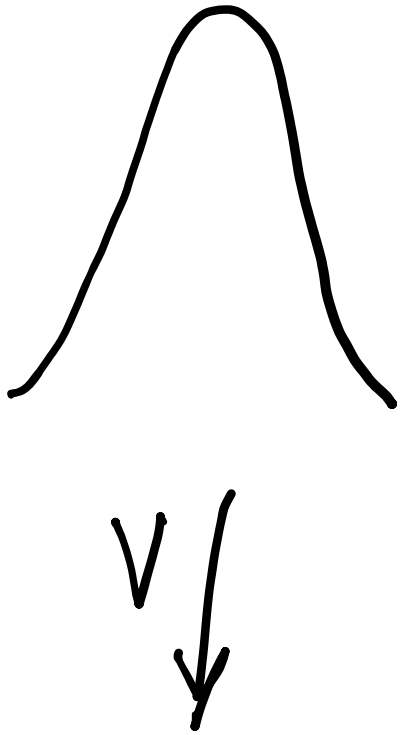
$n = 2$

$r = 2$

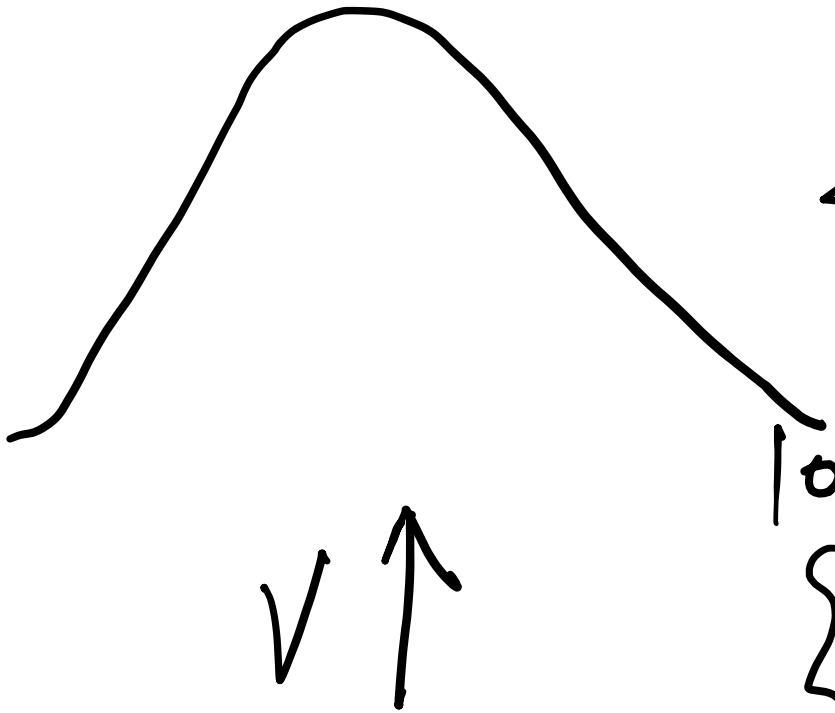
$p = \frac{1}{2}$

$$P(X=2) = {}^{2}C_{2}\left(\frac{1}{2}\right)^{2}\left(1-\frac{1}{2}\right)^{2-2}$$

$$= \frac{2!}{2!\,0!} \times \frac{1}{4} \times 1 = \frac{1}{4}$$



V ↓



V ↑

8	2 ✓	
9	3	
10	10	
10	10	10
12	22	
11	23 ✓	

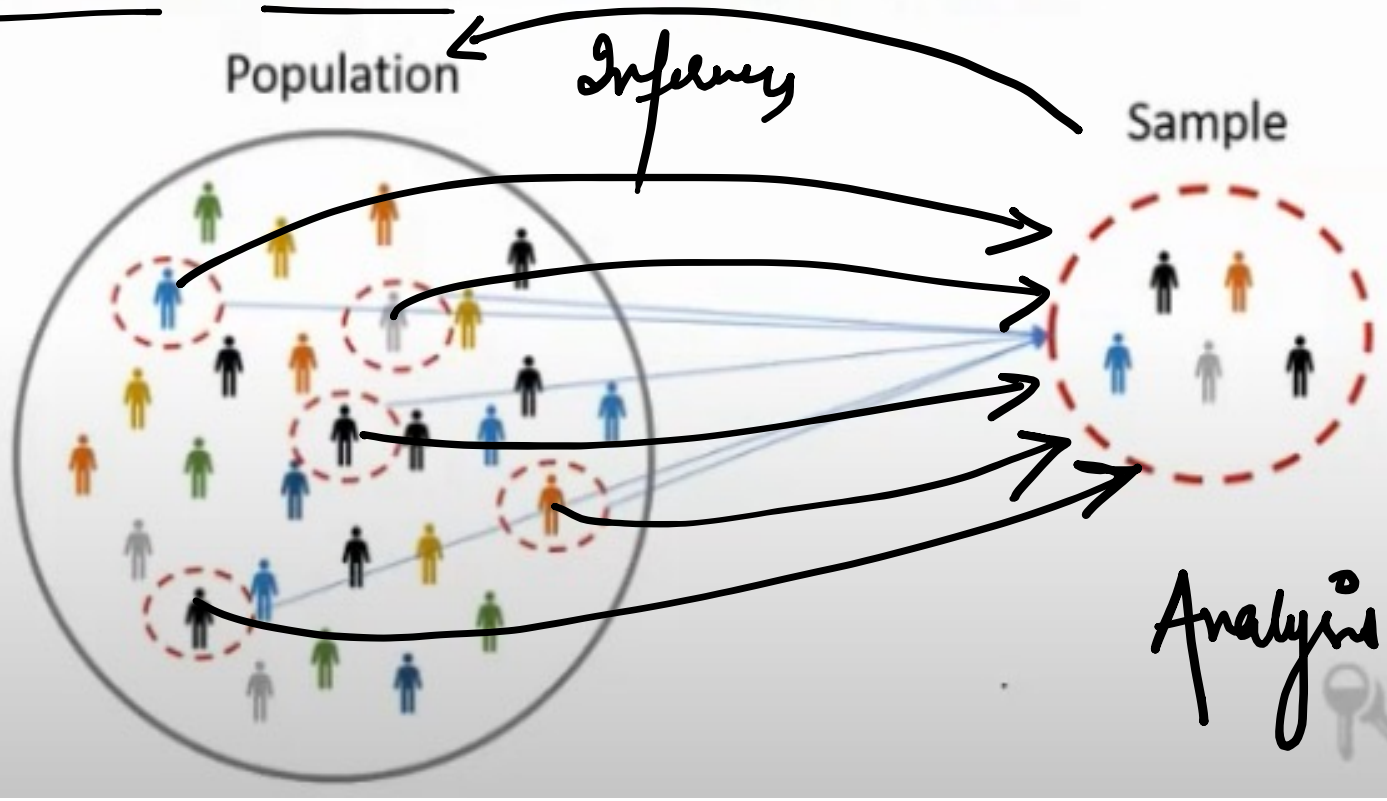
15

Inferential Statistics:

Approximate analysis → Range → Confidence Interval

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?



→ Population mean will lie in confidence interval of

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

σ → pop. std.

n → Sample size

$\frac{\sigma}{\sqrt{n}}$ → Sampling error

μ → Sampling distribution mean

Z → Standard normal variate

Z → confidence level

CL

Z

→ 90% → 1.65

{ by default } → 95% → 1.96

→ 99% → 2.56

hypo → Why ?

Z table

Value = $x + \frac{(100-x)}{2}$

$= 90 + \frac{10}{2} = 95$

0.95

Central Limit Theorem: (Assume → normal distribut^n)

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.

→ $Z = 1.96$

$\mu = 4.63$

$\sigma = 0.54$

$n = 50$

$$CI = \left[\mu \pm Z \times \frac{\sigma}{\sqrt{n}}\right]$$

$$= \left[4.63 \pm 1.96 \times \frac{0.54}{\sqrt{50}}\right]$$

Central Limit Theorem:

upGrad

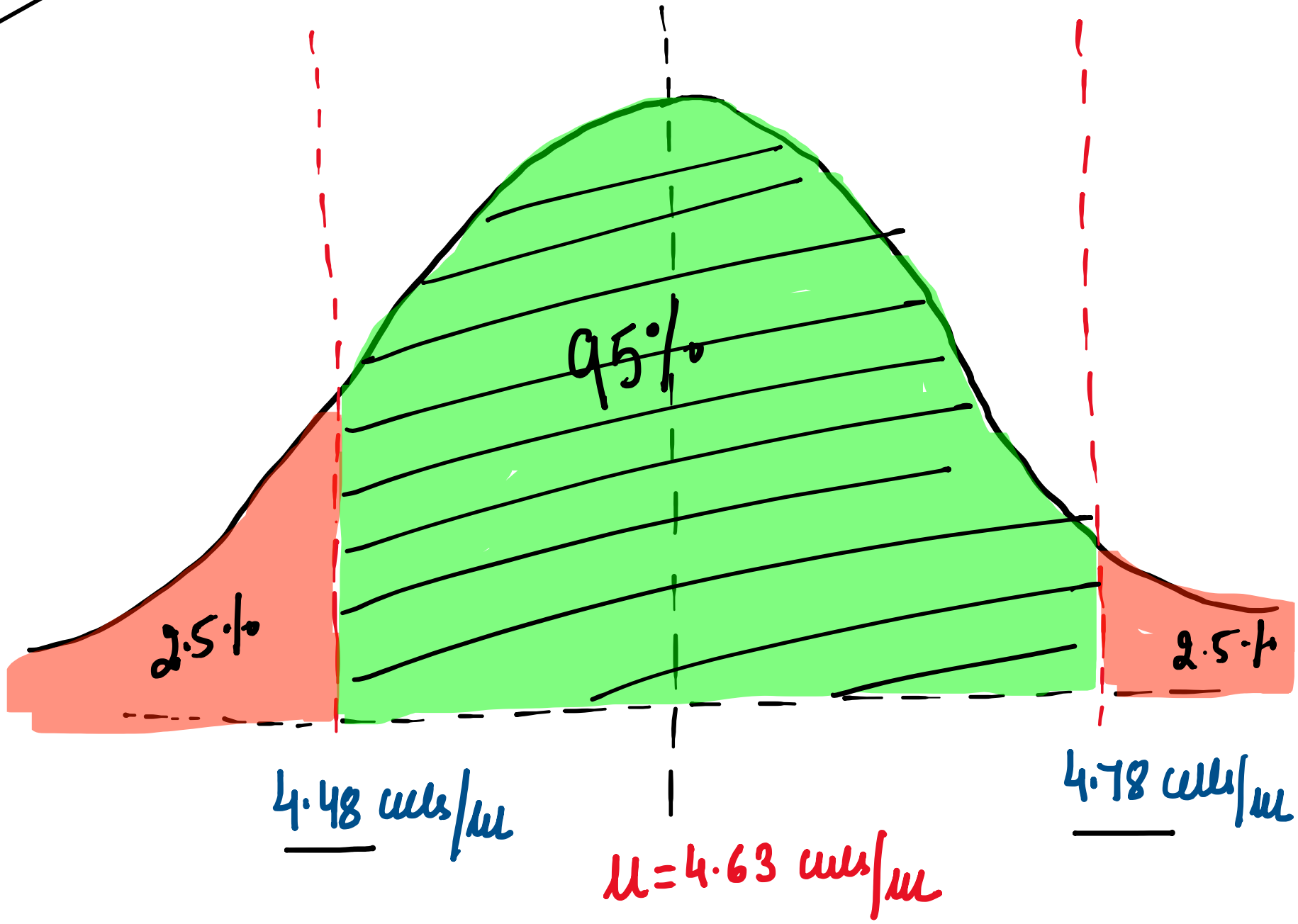
$$CI = 4.63 \pm .15$$

$$CI = [4.48, 4.78]$$

upGrad

$C.I = [4.48, 4.78]$

→ Population mean red blood cell count per micro liter will lie in the range of 4.48 mills/ml to 4.78 mills/ml & we are 95% confident for the same.


95%
2.5%
2.5%
4.48 cells/μl
μ=4.63 cells/μl
4.78 cells/μl

→ Drawback of Inferential Statistics :-

* Whatever inference is given as output of inferential statistics, is not validated anyway of the concept itself.

Doubts:

upGrad