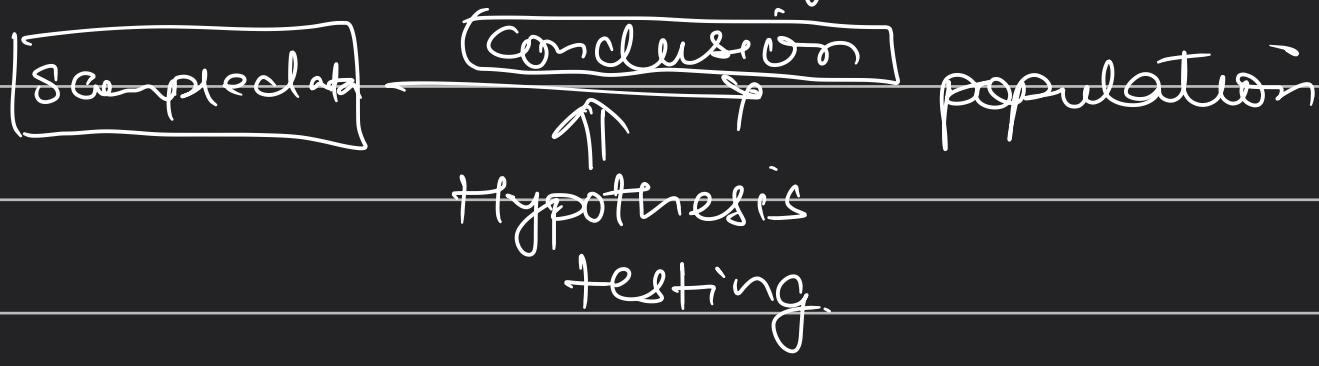
Hypothesis testing

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusion about a population parameter or population probability distribution

Here we test an assumption regarding a population parameter.

This talks about Inferential Statistics



Hypothesis testing : Coin is fair

① Null Hypothesis (H_0) = Coin is fair

⇒ By Default

② Alternate Hypothesis (H_1) = Coin is not fair.

③ Experiment : the experiment decides if our Null Hypothesis is "accepted" or "rejected"

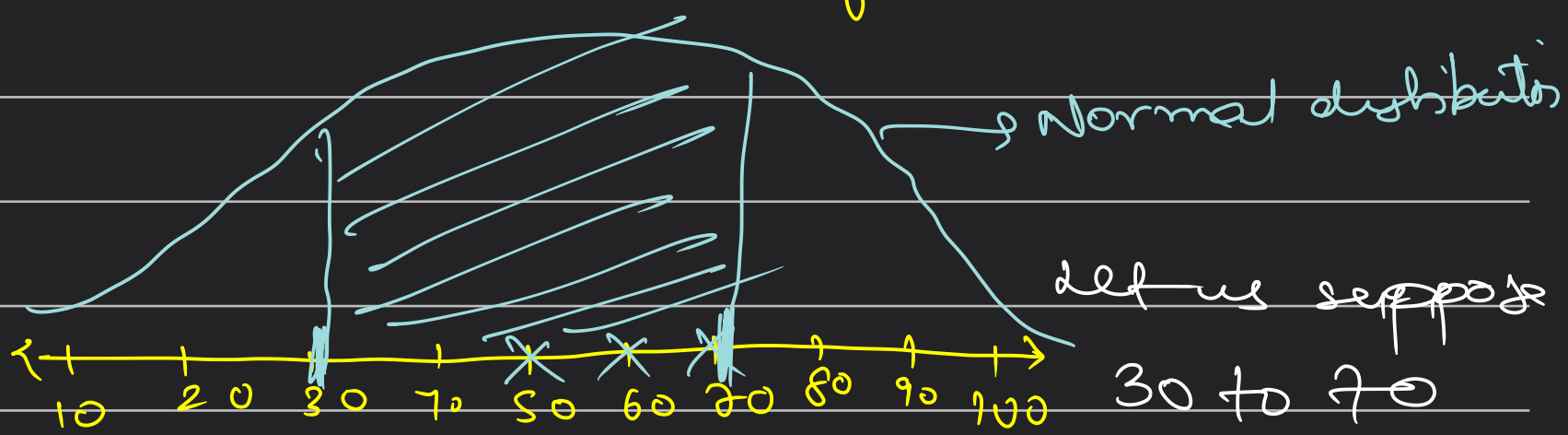
In our case experiment could be tossing a coin 100 times

Getting heads 50 or 60 times we can say coin is Fair. But getting 70 is where we

get our doubts.

Let us say Average is "50", head can comes above 50 or below 50 → this is called "2 Tail Test" → we need to check for both sides.

Next we decide a "Confidence Interval"



as our confidence interval, if my number of Heads comes under "confidence Interval" we will accept Null Hypothesis else if No of Heads comes under 30 or over 70 we will reject Null Hypothesis.

But who decides "Confidence Interval"?

⇒ Its the work of Domain Expert, in this case a Mathematician.

Remember this Hypothesis testing is done on a Sample Data from Sample Data results we will draw conclusions for population Doing a Toss 100 times can come under Sample Data.

Confidence Interval is also called Decision Boundary. we also define Confidence Interval in terms of Percentage, Lower fence value, higher fence value is also used for Hypothesis testing.

If we say Confidence Interval is 95% then

Significance value = 1 − CI

= 1 − 0.95 = 0.05

or 5% is our significant value

- Probability Distribution function:

whenever we talk about distribution of data, we talk about Probability Distribution.

Has 2 types

① Probability Density Function

② Probability Mass Function

If our data is "continous random variable" to check its distribution we use "prob density function [PDF]"

If your data is "discrete random variable" for this distribution check we use "probability mass function [PMF]"

- But now the question comes ⇒ What is random variable [I what is variable]

when ever we calculate values of "x" & "y" in linear equation, we call these X and Y as variables.

* **Random Variable** ⇒ is a process of mapping the output of random process or experiment to a number.

values get changed based upon Experiments

Eg: Tossing A Coin [Experiment]

$X = \begin{cases} 0 & \text{if Head} \\ 1 & \text{if Tail} \end{cases}$

Random variable can be 0 or 1 based upon o/p

Eg 2: Rolling a Dice

$Y = \sum$ Sum of rolling a dice 5 times

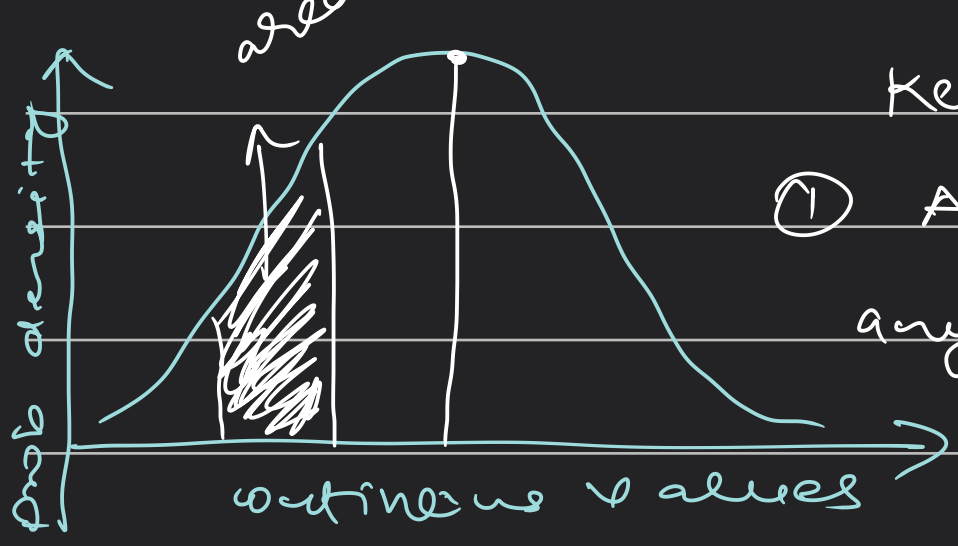
- **Probability Density**: refers to how packed or concentrated the likelihood of an event is around a specific value.

It's like asking "How many possibilities are around this particular point?"

2) **Probability density function**: it help us understand the chances of different values occuring in continous range.

Ex Heights between 4 feet to 6 feet.

Its a way to see where things are more likely to happen along a scale of possibilities.



Key features of PDF graph

① Area under PDF curve btw any points represents probability that a random variable will take a value between those points.

② Height of curve at any point represents the likelihood of random variable taking that value.

③ Total area under the curve is 1

- Cumulative Distribution function

Hp

Cumulative Probability → is like keep running total of how often something happens as time or event progress.

Its like adding up probabilities as we go along.

Key features:

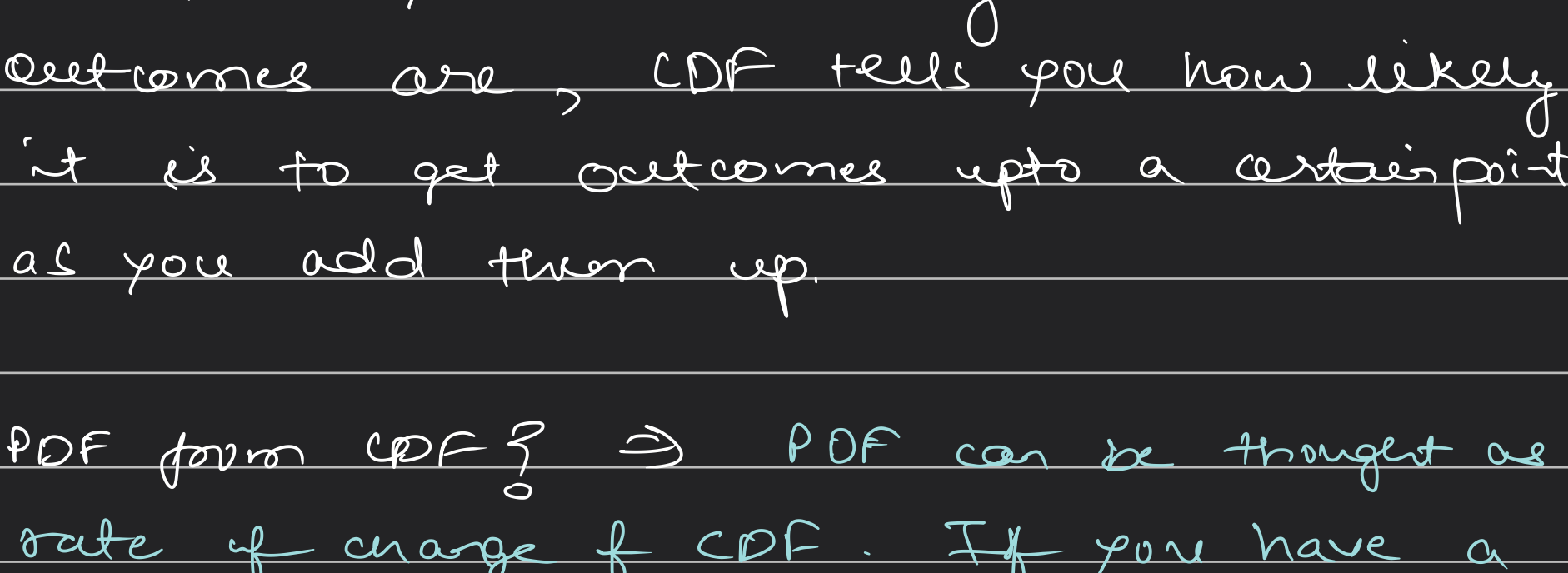
X axis represents possible continous / discrete values

Y axis represents Cumulative Probability range is 0 to 1

Its either a flat graph or increases as x → random variable increases

" The value of CDF at a particular point

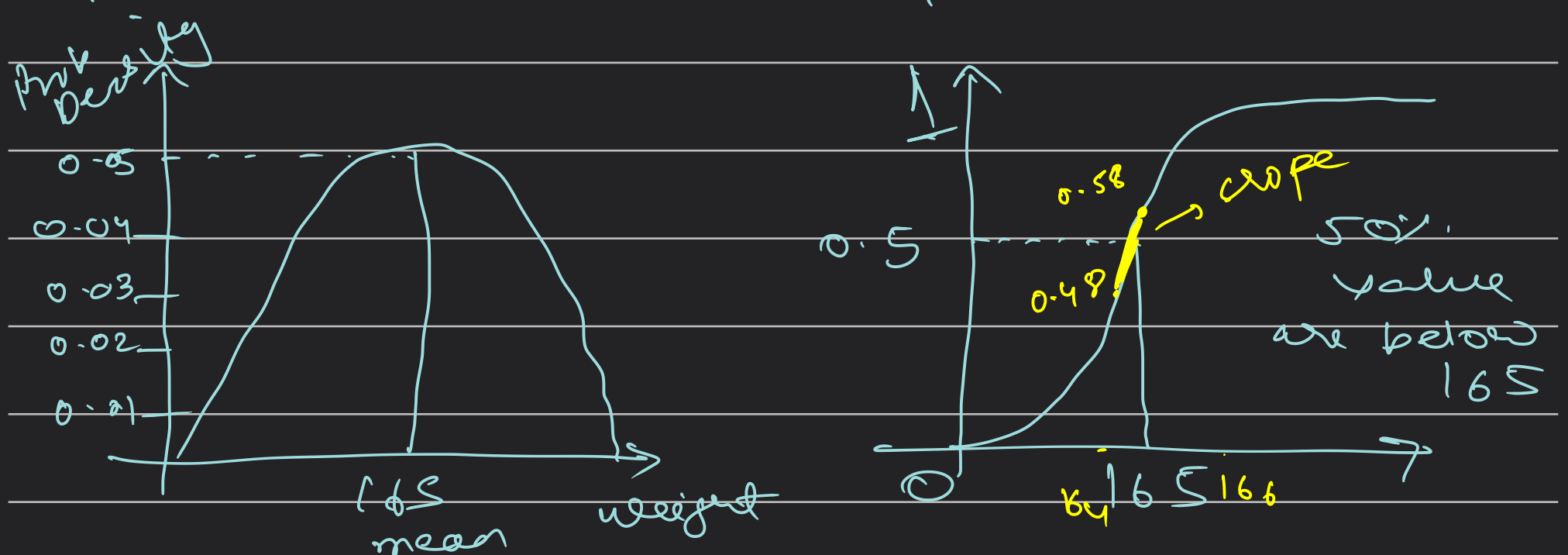
represents the probability that random variable x is less than equal to that value



CDF at a point is area under PDF curve upto to point x

PDF tells you how likely individual outcomes are, CDF tells you how likely it is to get outcomes upto a certain point as you add them up.

PDF from CDF? ⇒ PDF can be thought as rate of change of CDF. If you have a CDF or graph, then slope or tangent or gradient at exact point of CDF will give you pdf value at that point



Now we calculate derivative for the slope.

$$\text{Gradient} = \frac{\text{Change in } Y}{\text{Change in } X} = \frac{0.58 - 0.48}{166 - 164}$$

$$= 0.05$$

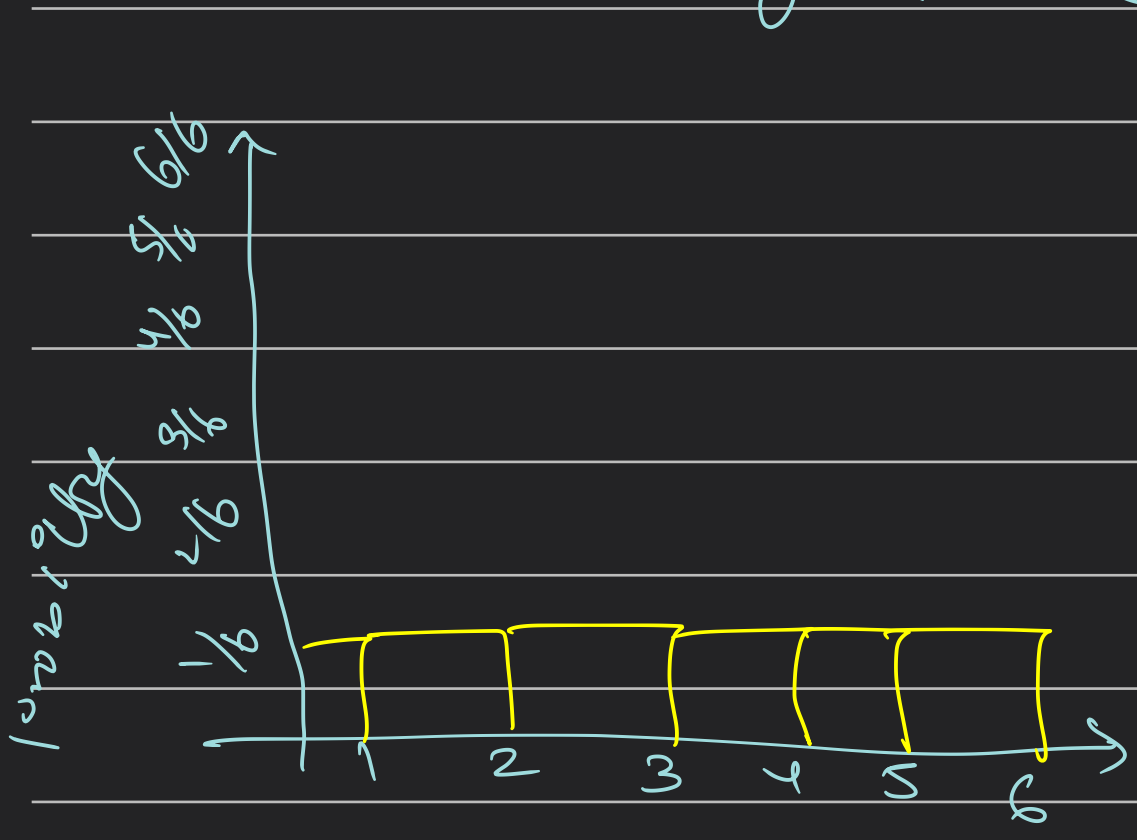
$\therefore$ pdf for 165^{th}

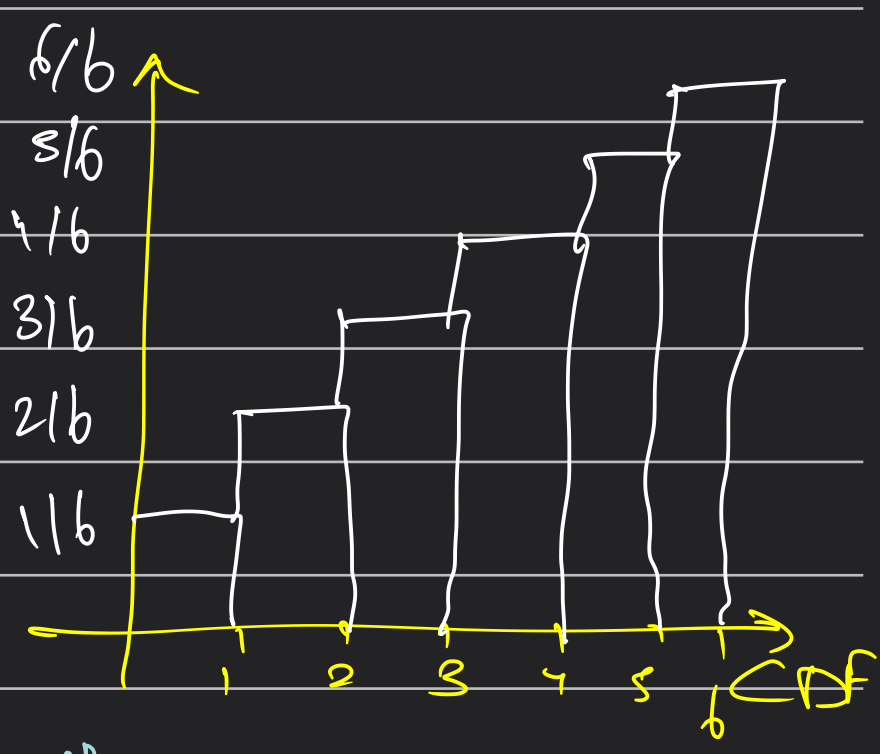
PDF = Gradient Decent of [slope] Cumulative curve

- Probability Mass functions

Same as PDF but for discrete values, Tell us how probable each outcome is when dealing with things that can only take on certain values. like dice rolls and counting number of heads

PMF for Rolling a dice





Prob of Each item is 1/6

CDF here is like telling what is probility of dice rolling $\leq x$ for $x \leq 6$ then prob is 1, ≤ 5 then prob is 5/6.