Central limit Theorem

DATE __/__/__

The central limit theorem can also be explained as the "distribution of sample "Mean" which approximated the Normal distribution when this is apply, when sample size larger.

here is the assumations all the samples are similar and shape of population distribution could anything.

e.g - Let assume that there are ten teams of cricket in your school, every team has a total 100 students in it.

if we want to measure the average height of all the students in the ~~step~~ sports team, then that would be humongous task.

1- One way, go to the about this task and find of all the students height and divide by total no of student

$$\text{Mean} = \frac{\text{All Students height}}{\text{Total No. of student}}$$

DATE / /

but if 2000 student then?

every time not possible to find the ~~student~~ height of All students

The Central limit theorem come into the picture

— Take a Random 30 sample of every Team-Team.

— Take all the sample and try to find the "Mean" for every single sample

— Once done, then find the mean of Means of the samples

— The value Recieve approx height of All the Student

20 Team, every Team has 100 students

$20 \times 100 = 2000$ students

— Now — Take a sample of

DATE ___/___/___

every Team at list 30 or more at this Stats we take 30 (Reandom sample)

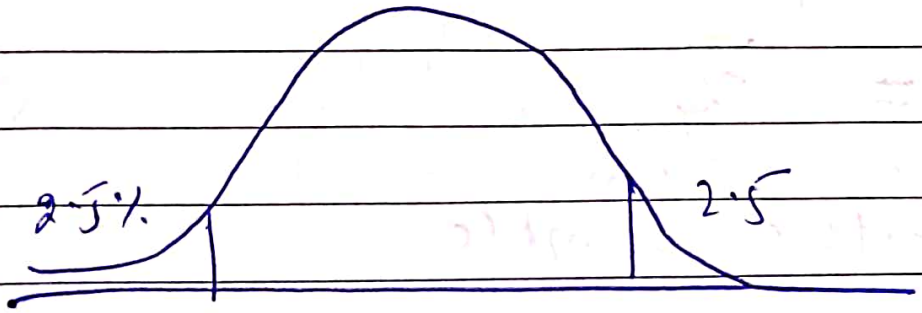
30 X 20(Team)

= 600 (sample)

$= \frac{6000}{20000}$ (600 = Total sample

2000 = Total No. of Sample

$= .3 = 30\%$



Mean → Avearge value

Standard deviation → is a measure how spread out of the values are.

SD increase → Normal distribution curve Widen

DATE ___/___/___

Gaussian distribution

$X \sim GD(\mu, \sigma^2)$

mean, variance

Population size

Random variable

We took some Random Sample

Sample size ≥ 30 ($n \geq 30$) CLT

$S_1 = x_1 \ldots x_{30}$

$= \bar{x}_1$

Take another Random Sample

mean of every Random samples

$S_2 = x_1, x_2 \ldots x_{30} = \bar{x}_2$

$S_3 = x_1, x_3 \ldots x_{30} = \bar{x}_3$

$\vdots$

$S_{100} = x_1 \ldots x_{30} = \bar{x}_{100}$

(100 samples)

DATE ___/___/___

Now

$$\bar{x} \simeq GD\left(\mu, \frac{\sigma^2}{n}\right)$$

Where the sample mean = population Mean

$$\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \cdots \bar{x}_{100} = \mu$$

So

GD (μ is same)

it mean whatever random Sample Mean would be same sample for population 100.

Sample mean = population

Variance is differen = $\frac{\sigma^2}{n}$ (Total no. of sample size)

~~Total~~

Imp

→ Once this plot in Histogram Pen it would convert in Normal distribution like "bell curve"

DATE / /

Cummulative Distribution function

- PMF (probability distribution function) is one way to describe the distribution of "discreate random variable"
- PMF can not defined for "continuous random variables"
- CMF (Cumulative Distribution function) This is an another method describle "distribution of random variable"
- Advantage of CMF, it can defined any kind of random variable (discrete, continuous, and Mixed)

DATE ___/___/___

CDF formula →

CDF is Real value random variable X is function given by

$$F_X(x) = P(X \leq x)$$

where $F_X(x)$ = function of x

X = real value variable

P = probability that X will have a value less than or equal to x

$$F_X(x) = \int_{-\infty}^{x} f_x(t)\,dt$$

incase of random variable X which has distribution having a discrete component at a value b

$$P(X=b) = F_X(b) - \lim_{z \to b^-} F_X(z)$$

where F_X is continuous is b.

DATE __/__/

Probability Distribution function

Discreate → probability mass function PMF

Continuous → Probability density function PDF

Both → Cummulative distribution function CDF

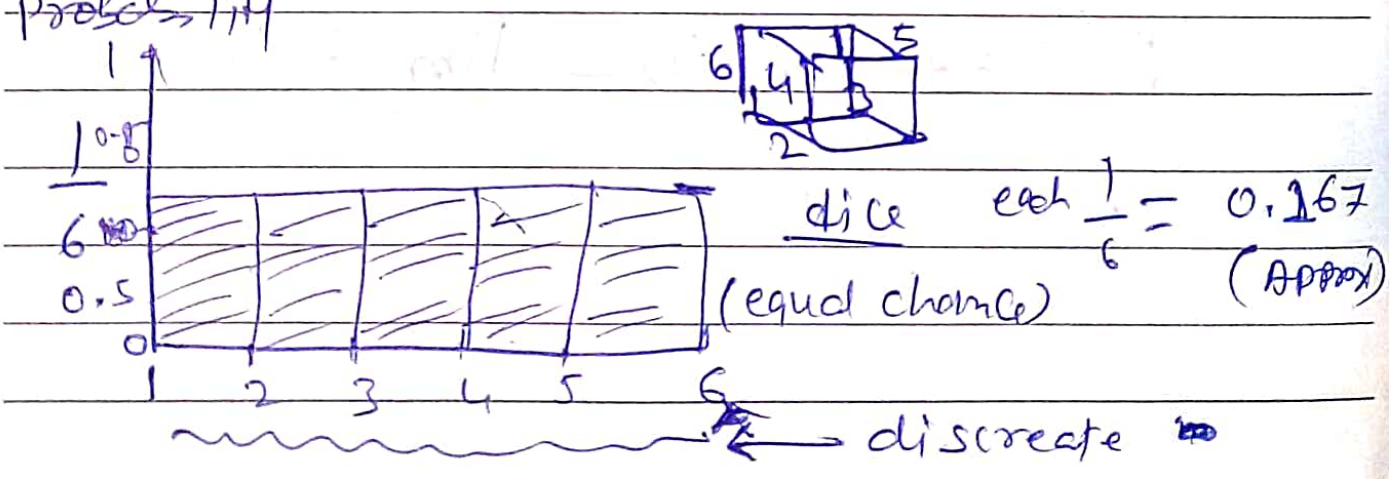
PDF (Probability distribution function)

1 — probability mass function

2 — Probability density function

3 — Cummulative distribution function

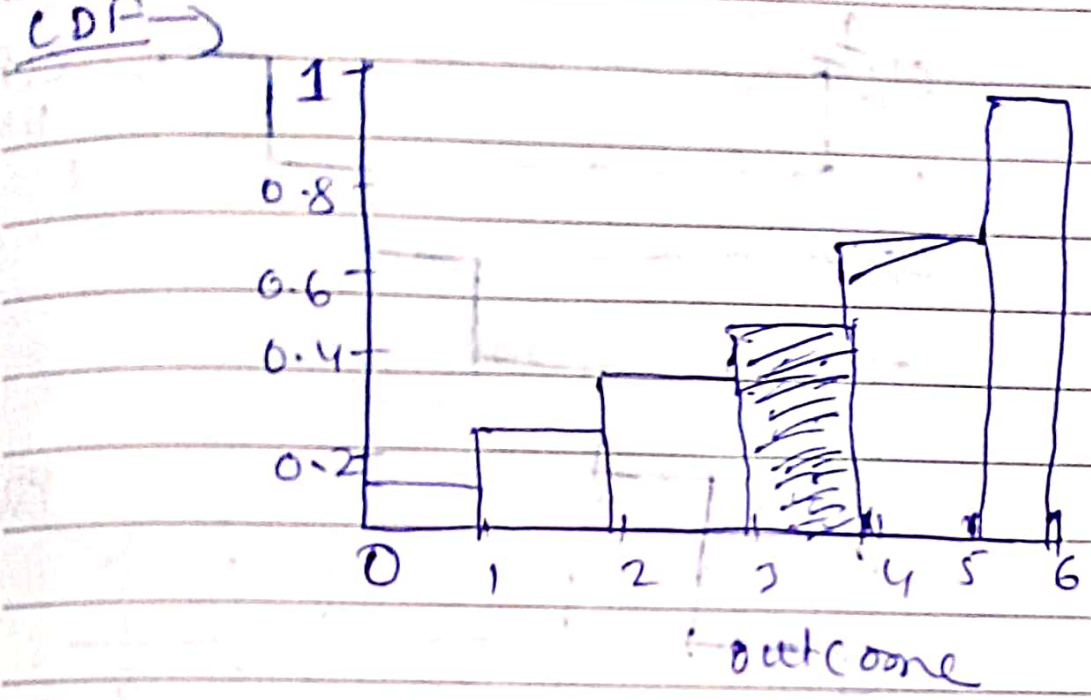
(1) Probability Mass function (PMF) →



DATE __/__/__

because Why discreate? it outcome is 1, 2, 3 ...

CDF



$$P(x \leq 4) = P(x=1) + P(x=2) + P(x=3) + P(x=4)$$

~~Rolling Probability~~

Why doing PDF and CDF?

both the part of probability function.

Probability distribution

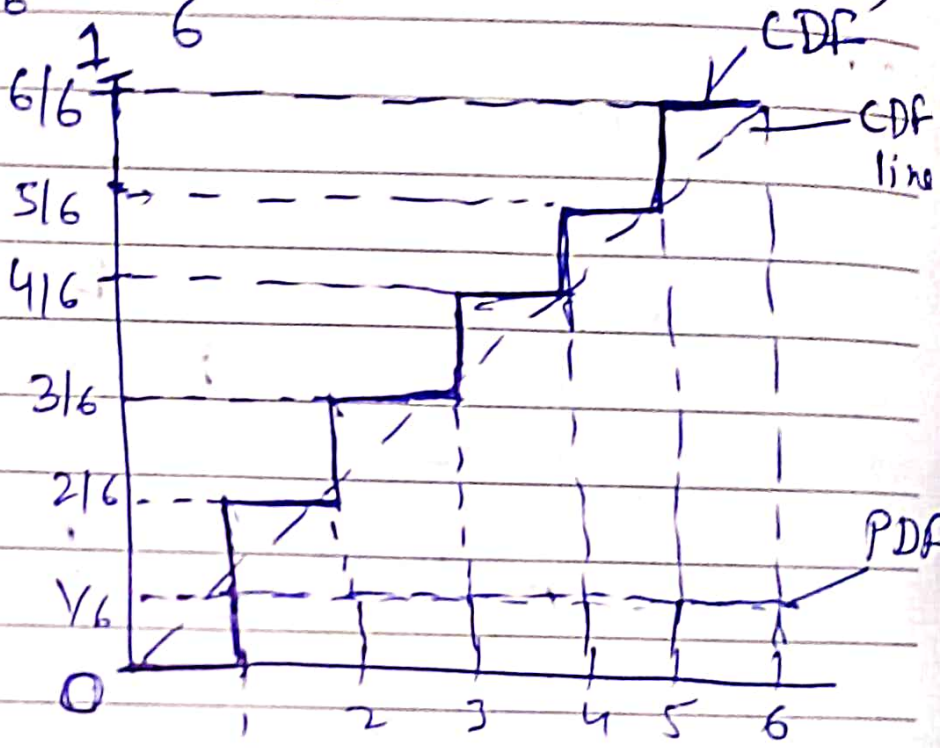
- CDF
- PDF

DATE ___/___/___ $F_x(x)$ (Capital F)

CDF → denoted by $F_x(x) = P(X \leq x)$

if we have dice.

$P_{1,2,3,4,5,6} = \frac{1}{6}$ (probability occurs)



Calculate

$F_x(0) = 0$ * PDF is straight line

$F_x(1) = P(1)$

$F_x(2) = P(1) + P(2)$

$F_x(3) = P(1) + P(2) + P(3)$

$F_x(4) = P(1) + P(2) + P(3) + P(4)$

⋮

$F_x(6) = P(1) + P(2) + P(3) + P(4) + P(5) + P(6)$

This cummulative Addition

DATE ___/___/___

$F_X(0) = 0$

$F_X(1) = P(1) = 1/6$

$F_X(2) = P(1) + P(2) = 1/6 + 1/6$

$F_X(3) = P(1) + P(2) + P(3)$

$= 1/6 + 1/6 + 1/6$

$F_X(3) = \frac{3}{6}$

⋮

$F_X(6) = P(1) + P(2) + P(3) + P(4) + P(5) + P(6)$

$= 1/6 + \frac{1}{6} + 1/6 + 1/6 + \frac{1}{6} + \frac{1}{6}$

$F_X(6) = \frac{6}{6}$

Properties of CDF →

1) CDF is Alway greater than equal to zero

$$F_X(x) \geq 0$$

2) Range of CDF is

$$0 \leq \underbrace{F_X(x)}_{\text{CDF is 0 to 1 (Range)}} \leq 1$$

DATE ___/___/___

3) It is Alway increasing function
min 0 to max 1.

PDF

denoted $f_x(x)$ (small f)

$$f_x(x) = P(x \leq x)$$

CDF is cummulative Addition of probability
PDF is Magnitude Addition of probability

if we want PDF

$$f_x(x) = P(x \leq x)$$

$$P(5) = F_x(5) - F_x(4)$$

PDF

$$f_x(x) = \frac{d F_x(x)}{dx}$$ Relation of

CDF

$$F_x(x) = \int_{-\infty}^{+\infty} f_x(x)\, dx$$

This is the Relation of CDF & PDF

DATE __/__/____

Properties of PDF

1) PDF range Always 0 to 1

$$0 \leq f_x(x) \leq 1$$

2) Total Area of PDF = 1

$$1 = \int_{-\infty}^{\infty} f_x(x)\,dx$$

e.g →

In a experiment, a trial consist of three successive tosses of coin. If we define random variable x as the no of heads appearing in trial, determine $f_x(x)$ and $F_x(x)$ (PDF and CDF)

R.V	f(x)
0	1/8
1	3/8
2	3/8
3	1/8

TTT
TTH
THT
THH
HTT
HTH
HHT
HHH
= 8

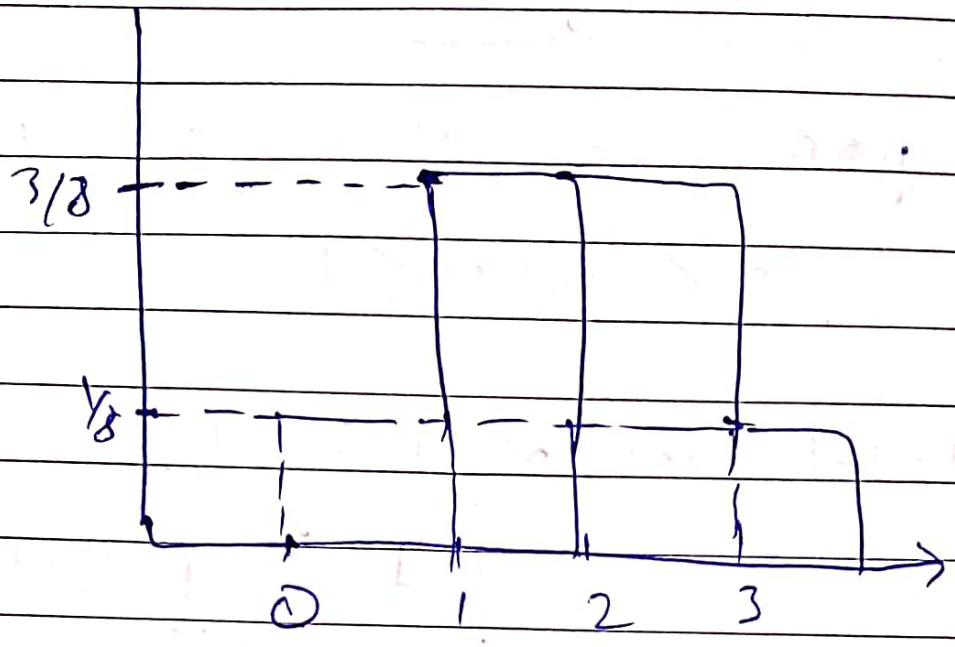
Toss to get head

0 0 0 = No heads = 0
0 0 1 = One head = 1
0 1 1 = Two head = 2
1 1 1 = Three head = 3

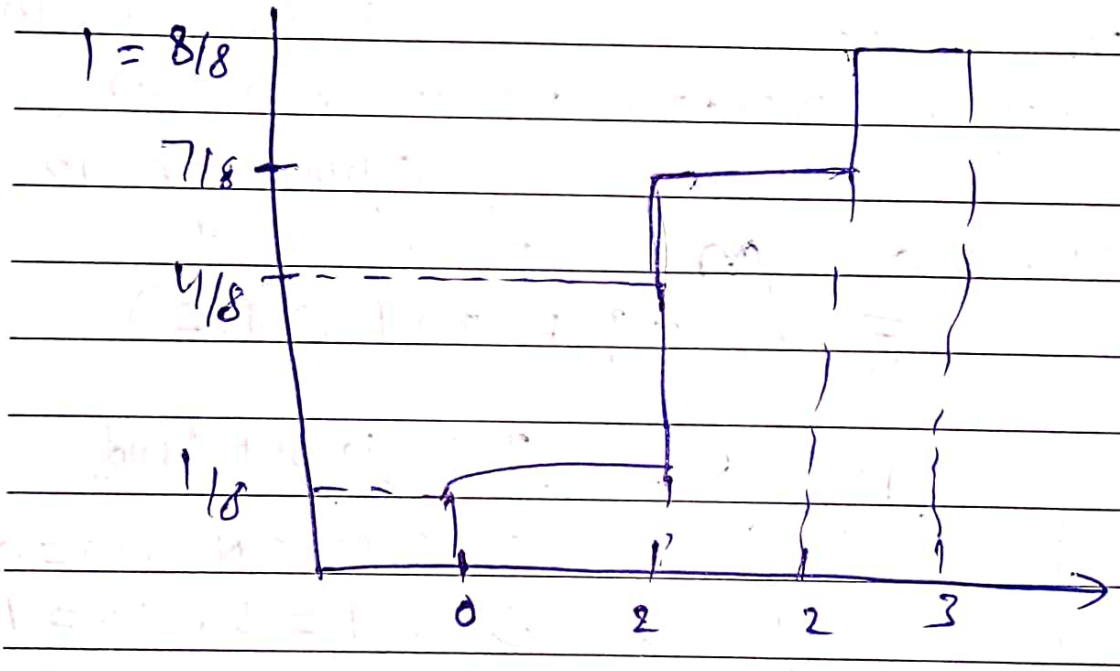
It mean of Random Variable

DATE ___/___/_____

Draw PDF-



Draw CDF



DATE __/__/__

Statistics Estimation

- Parametric
  - PDF, Normal distribution, CDF
- Non-Parametric
  - Kernal density function

Kernal density function

Formula

$$\hat{f}(x;h) = \frac{1}{nh} \sum_{i=1}^{n} K\left\{\frac{(x - X_i)}{h}\right\}$$

Where K = Kernal function (non-negative)

h → smoothing parameter (bandwidth) → Smoothing the Curve

X_i = given value of x

DATE ___/___/_____

~~KDE & Histogram~~
~~are similar~~

Bandwidth → (Smoothing Parameter)

Smoothing of Curve

1) Cross validation
2) Normal optimal Smoothing

Kernal density estimates are closely Related to Histogram !