Discrete Probability Distributions

Infinite trials:

ex:- Let's see who wins?

Lets consider A & B and both throw the dice. whoever gets the first six wins. Consider A starts.

$$P(A \text{ winning}) = \underbrace{\frac{1}{6}}_{\text{in first chance}} + \frac{5}{6}\times\frac{5}{6}\times\underbrace{\frac{1}{6}}_{\text{in second chance}} + \frac{5}{6}\times\frac{5}{6}\times\frac{5}{6}\times\frac{5}{6}\times\frac{1}{6} + \cdots \text{ up to } \infty$$

$$= \frac{1}{6} + \left(\frac{5}{6}\right)^2\times\frac{1}{6} + \left(\frac{5}{6}\right)^4\times\frac{1}{6} + \cdots$$

↓

Infinite geometric progression.

↓

$$\therefore \quad S_\infty \text{ (Sum of infite terms)} = \frac{a}{1-R}$$

$a \longrightarrow$ first term

$R \longrightarrow$ Common Ratio

here

$$a = \frac{1}{6} \qquad R = \left(\frac{5}{6}\right)^2 \rightarrow \frac{3^{rd} \text{ term}}{2^{nd} \text{ term}} \text{ divided by}$$

$$P(A \text{ winning}) = \frac{1/6}{1-(5/6)^2} = \frac{1/6}{1-25/36}$$

$$= \frac{6}{11}$$

So whoever starts 1st will have $\frac{6}{11}$ probability to win.

So person starting 2nd will have $\frac{5}{11}$ probability to win. (i.e $1 - \frac{6}{11} = \frac{5}{11}$)

(or)

$$P(\text{B winning}) = \frac{5}{6} \times \frac{1}{6} + \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{1}{6} + \ldots$$

$$a = \frac{5}{36}$$

$$R = \left(\frac{5}{6}\right)^2$$

$$P(\text{B winning}) = \frac{5/36}{1 - 25/36} = \frac{5}{11}$$

Q) A and B throw a pair of dice A wins if he throws 6 before B throws 7 & B wins if he throws 7 before A throws 6. If A begins find the probability of

(i) A winning (ii) B winning.

$n(S) = 36$

$n(\text{total}=6) = 5$ $\quad$ $n(\text{total}=7) = 6$

$n(\text{not } 6) = 31$ $\quad$ $n(\text{not } 7) = 30$

$$P(\text{A winning}) = \frac{5}{36} + \frac{31}{36} \times \frac{30}{36} \times \frac{5}{36} + \left(\frac{31}{36} \times \frac{30}{36}\right)^2 \times \frac{5}{36} + \ldots$$

$$= \frac{5/36}{1 - \frac{31}{36} \times \frac{30}{36}} = \frac{5/36}{1 - \frac{930}{1296}}$$

$$= \frac{30}{61}$$

$$P(\text{B winning}) = 1 - P(\text{B winning})$$

$$= \frac{31}{61}$$

$P(\text{B winning}) > P(\text{A winning})$ because probability of total = 7 > total = 6.

Bernoulis Trials

- finite in number
- independent of each other
- Each trial has only 2 outcomes (Success, failure)
- probability of success or failure remains same in each trial.

* Bernouli trial is also called binomial trial.

ex:- The product can be defective or not defective

Q) Eight balls are drawn from a bag containing 10 white and 10 black balls. Predict whether the trials are Bernouli trials if the ball drawn is replaced and not replaced.

(1) with replacement the probability of success say white ball remains same so Bernoulis trials.

(2) without replacement the probability of success changes so not Bernoulis trials.

Binomial distribution :

Binomial theorem

General term $(x+y)^n = {}^nC_r\, x^r\, y^{n-r}$.

let us consider $P(\text{success}) = p$

$P(\text{failure}) = q$

we know that $p+q = 1$

If we perform a trial 'n' times in which we have 'r' successes and $(n-r)$ failures.

- these 'r' successes in 'n' trials can be anywhere

so here ${}^nC_r = \frac{n!}{(n-r)!\,r!}$

ex:- if 10 trials, 4 successes then permutations is

$${}^{10}C_4 = \frac{10!}{6!\,4!} = \frac{10^5 \times 9^3 \times 8^2 \times 7}{4 \times 3 \times 2}$$

$$= 210$$

In Binomial - Distribution

$$1 = {}^{n}C_0 p^0 q^n + {}^{n}C_1 p^1 q^{n-1} + {}^{n}C_2 p^2 q^{n-2} + \dots$$

- ${}^{n}C_0 p^0 q^n$: Probability of getting 0 successes in n trials
- ${}^{n}C_1 p^1 q^{n-1}$: Probability of getting exactly 1 successes in n trials
- ${}^{n}C_2 p^2 q^{n-2}$: Probability of getting exactly 2 successes in n trials

$${}^{n}C_r p^r q^{n-r}$$ exactly 'r' successes in n trials

Q) An experiment succeeds twice as often as it fails. Find the probability that in next 6 trials, there will be atleast 4 successes.

$\therefore p = 2q$ $\quad p + q = 1$

$\therefore q = \frac{1}{3}, \; p = \frac{2}{3}$

P(atleast 4 successes) $\therefore \{4, 5, 6\}$

$$= {}^{6}C_4 \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^2 + {}^{6}C_5 \left(\frac{2}{3}\right)^5 \left(\frac{1}{3}\right)^1 + {}^{6}C_6 \left(\frac{2}{3}\right)^6 \left(\frac{1}{3}\right)^0$$

Random Variable : In probability, a real-valued function, defined over the sample space of a random experiment, is called a random variable.

Definition : A Random Variable is a rule that assigns a numerical value to each outcome in a sample space.

Random Variables

- Discrete Random Variable – A random Variable is said to be discrete if it assumes only specified values in a interval
- Continous Random Variable
  - If we want to create discrete values from continous we can consider intervals and bins and these will be discrete

Random variable is denoted by capital 'x'. It takes input as all the outcomes in the sample space as input and gives a number as output

ex:- If we toss 2 coins and assign random variable number of heads

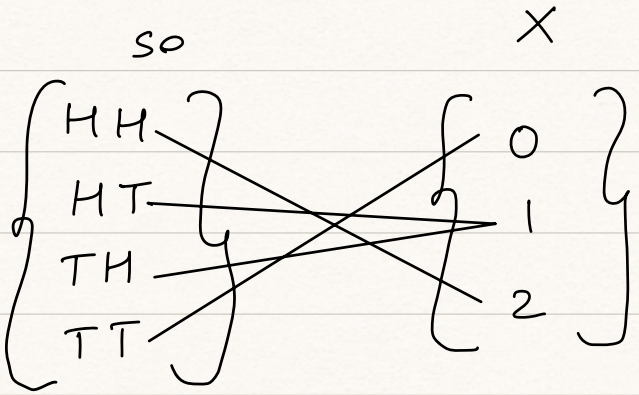
$$S = \{HH, HT, TH, TT\}$$

X — HH $\xrightarrow{\text{output}}$ 2

X — HT $\longrightarrow$ 1

X — TH $\longrightarrow$ 1

X — TT $\longrightarrow$ 0



* Random variable will have many to one kind of relationship.

Discrete Random variable:

- A discrete random variable can take only a finite number of distinct values such as 0, 1, 2,... and so on.
- The probability distribution of a random variable has a list of probabilities compared with each of its possible values known as probability Mass function.

ex:- 1) Toss a coin and count no.of heads
2) Defects in a product produced.

Continous Random variable:

- A numerically valued variable is said to be continous, if the random variable X can assume an infinite and uncountable set of values, it is said to be a continous random variable.
- when X takes any value in a given interval (a, b), it is said to be a continous random variable in that interval.
- A continous random variable is such that it has constant cumulative function throughout.
- ex:- pressure, height, weight, volume etc...

Probability distribution Function:

A function which is used to define the distribution of a probability is called a probability distribution function.

Probability Distribution Table:

A probability distribution is a table that displays the probability that a random variable takes on certain values

X	X_1	X_2	X_3	.....	X_n
$P(X)$	P_1	P_2	P_2	....	P_n

Properties of probability Distribution table:

1) All the probabilies must add up to 1.

i.e $P_i \geq 0$ and $P_1 + P_2 + P_3 + \cdots + P_n = 1$

2) Mean of random variable:

In statistics mean $= \frac{\Sigma x_i f_i}{\Sigma f_i}$

So here mean $= \frac{\Sigma x_i P_i}{\Sigma P_i}$ but $\Sigma P_i = 1$

$\therefore$ mean $= \Sigma x_i P_i$

For population mean is μ

3) Variance of a Random Variable

The variance tells how much is the spread of random variable x around the mean value. The formula of the variance of a random variable is given by:

$$Var(X) = \sigma^2 = E(X^2) - [E(X)]^2$$

where $E(X^2) = \sum x_i^2 p_i$ and $E(X) = \sum x_i p_i$

or $\sigma^2 = \mu(x_i^2) - (\mu(x_i))^2$

here E(X) is expectation of x.

or $$\sigma^2 = \sum (x_i - \mu)^2 * p(x_i)$$

x_i : The ith value

μ : The mean of distribution

$p(x_i)$: The probability of i^{th} value.

4) Standard deviation is $\sqrt{\text{Variance}}$.

Q1) In the toss of 2 coins example from above.

X	0	1	2
P(X)	1/4	2/4	1/4

- The top row shows the values of Random Variable X.
- The bottom row shows the probability of the X. This is similar to frequency tables in statistics.

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

$$\text{Variance} = \left(0^2 \times \frac{1}{4} + 1^2 \times \frac{2}{4} + 2^2 \times \frac{1}{4}\right) - 1$$

$$= 2/4$$

Q2) A urn contains 4 white and 6 red balls. Four balls are drawn at random from the urn. Find the probability distribution of the number of white balls.

↓ No. of white balls drawn

X	0	1	2	3	4
P(X)	$\frac{{}^6C_4}{{}^{10}C_4}$	$\frac{{}^4C_1 \times {}^6C_3}{{}^{10}C_4}$	$\frac{{}^4C_2 \times {}^6C_2}{{}^{10}C_4}$	$\frac{{}^4C_3 \times {}^6C_1}{{}^{10}C_4}$	$\frac{{}^4C_4}{{}^{10}C_4}$

Q) If a six-face die is thrown 3 times. what is the probability that 5 appears exactly 2 times.

$n = 3$

success is rolling a 5, failure rolling anything other than 5.

$p = 1/6$ & $q = 5/6$

$r = 2$

$$P(X = x) = {}^{3}C_{2} \left(\frac{1}{6}\right)^{2} \left(1 - \frac{1}{6}\right)^{3-2}$$

$$= 0.0694$$

<u>logically</u>

$p^x (1-p)^{n-x}$ → one specific ordering of one success and $n-x$ failures.

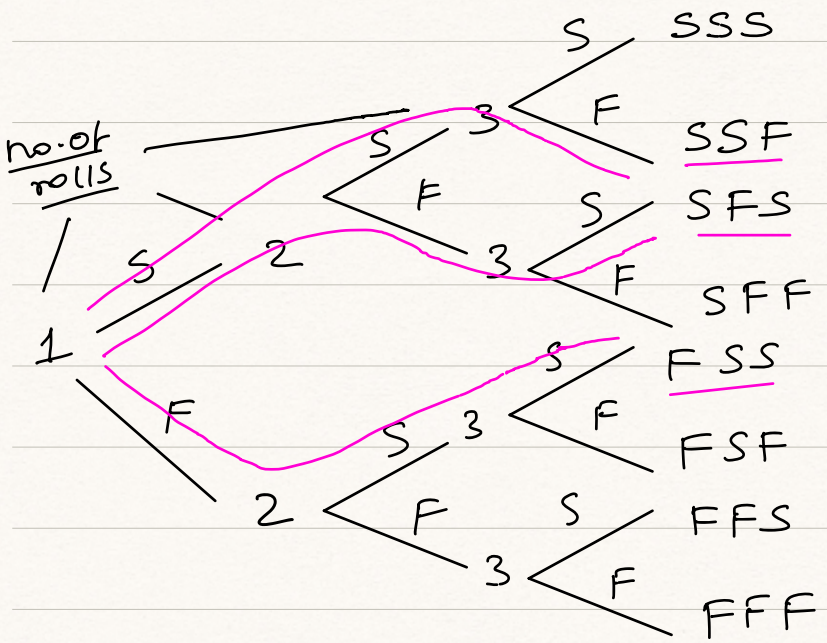
Here we are concerned about 'x' success in 'n' trials so

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

<u>using probability tree with the above example</u>

$n = 3$

$r = 2$ (2 times 5)



here exactly 2 times success is (SSF, SFS, FSS)

$$P(X=2) = p^2(1-p) + p(1-p)p + (1-p)p^2$$

$$= 3p^2(1-p)$$

$$= {}^3C_2\, p^2(1-p)$$

$$\therefore P(X=x) = {}^nC_x\, p^x(1-p)^{n-x}$$

Mean and Variance of Binomial distribution

Probability of success = p

Probability of failure = q

Total number of trials = n

Total number of successes = r

mean = np Variance = npq.

X	0	1	2	....	n
P(X)	$^nC_0 p^0 q^n$	$^nC_1 p^1 q^{n-1}$	$^nC_2 p^2 q^{n-2}$	....	$^nC_n p^n q^0$

Using binomial theorem derivation we get

Since $(p+q)^n = 1^n = 1$

$$1 = {}^nC_0 p^0 q^n + {}^nC_1 p^1 q^{n-1} + \ldots + {}^nC_n p^n q^0$$

mean = np

Variance = npq. i.e $nP(1-P)$

Bernoulli's distribution: Special case of binomial distribution where only one trial is conducted.

So mean = p

Variance = $p(1-p)$ or pq.

- So for Bernoulis distribution mean is always the probability of success.
- Variance is probability of success and probability of failure so Variance is always less than mean.

For Bernouli Distribution PMF

$$P(X = x) = p^x (1-p)^{(1-x)}$$

for $x = 0, 1$ (as success or failure)

$$P(X=1) = p^1 (1-p)^{1-1}$$

$$= p \qquad = P(\text{success})$$

$$P(X=0) = p^0 (1-p)^1$$

$$= 1-p \qquad = P(\text{failure})$$

ex:- Approximately 1 in 200 American adults are lawyers. One American adult is randomly selected. what is the distribution of the number of lawyers?

here

- 1 adult selected → one trial
- either lawyer or not lawyer - only two results
- probability of one selection being a lawyer is $\frac{1}{200}$ and not dependent on other person.

- All Bernouli Trial conditions are satisfied.

$$\therefore P(X=x) = \left(\frac{1}{200}\right)^x \times \left(1-\frac{1}{200}\right)^{1-x}$$

$x = 0, 1.$

$$P(X=1) = \frac{1}{200} \qquad P(X=0) = \frac{199}{200}$$

this we can logically also but Bernouli gives nice explanation.

Other distributions built on Independent Bernoulli trials are:

1) Binomial Distribution – number of successes n Independent Bernoulli trials
2) Geometric Distribution – Distribution of trials to get the first success in Independent bernoulli trials.
3) Negative binomial distribution – It is the distribution of no. of trials to get the rth success in Independent Bernoulli trials.

What is Probability Distribution used for?

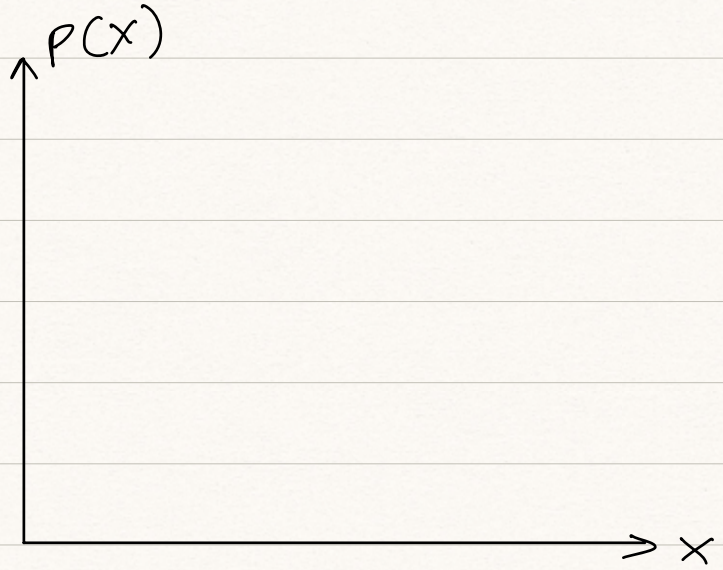
- The probability distribution is one of the important concepts in statistics.
- It has huge applications in business, engineering, medicine and other major sectors
  It is majorly used to make future predictions based on a sample for a random experiment.
- example in business sector it is used to predict if there will be profit or loss to the company using any new strategy or by proving any hypothesis test in the medical field, etc.

What is the importance of Probability distribution in statistics?

- In statistics, to estimate the probability of a certain event to occur or estimate the change in occurence and random phenomena modelled based on the distribution.
- Statisticians take a sample of the population to estimate the probability of occurence of an event.

Probability distribution of a Discrete Random Variable:

We can plot this considering the Random Variable on x-axis and Probability of the Random Variable on y-axis.



ex:- If we toss 2 coins and assign random variable number of heads

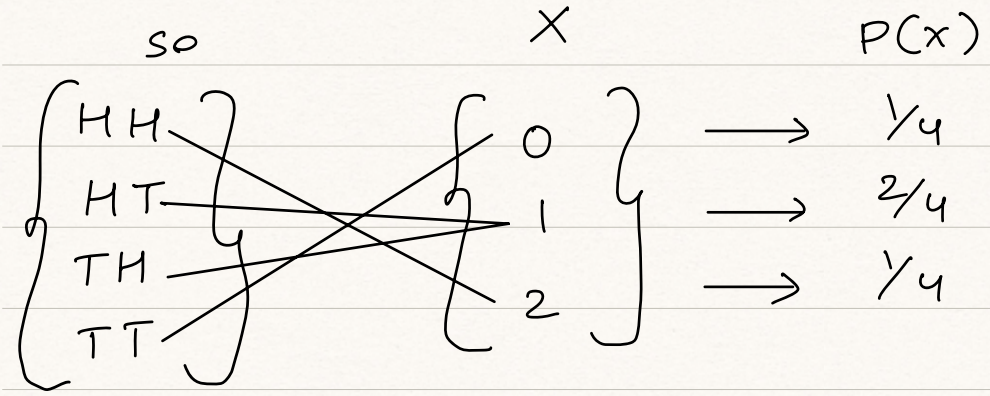
$$S = \{HH, HT, TH, TT\}$$

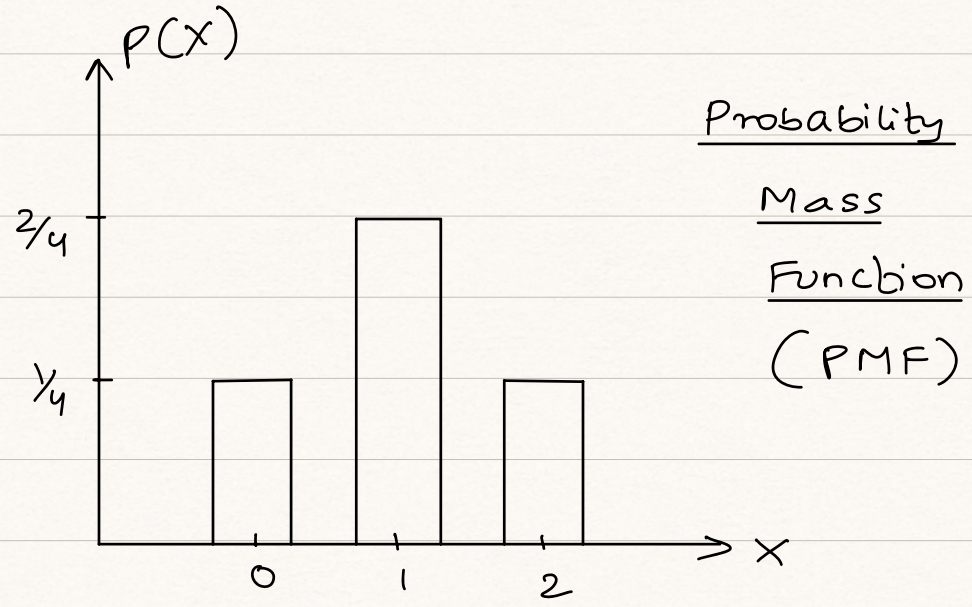
X — HH $\xrightarrow{\text{output}}$ 2

X — HT $\longrightarrow$ 1

X — TH $\longrightarrow$ 1

X — TT $\longrightarrow$ 0





Probability Mass Function (PMF)

For Discrete Random Variable when we plot Random Variable vs Probability of Random Variable we call it Probability Mass Function. It gives the probability that a discrete random

Variable is exactly equal to some value. Sometimes it is also known as discrete density function.

Cummulative Distribution Function : (CDF)

The cummulative distribution function of a real valued random variable X, evaluated at x, is the probability function that x will take a value less than or equal to x.

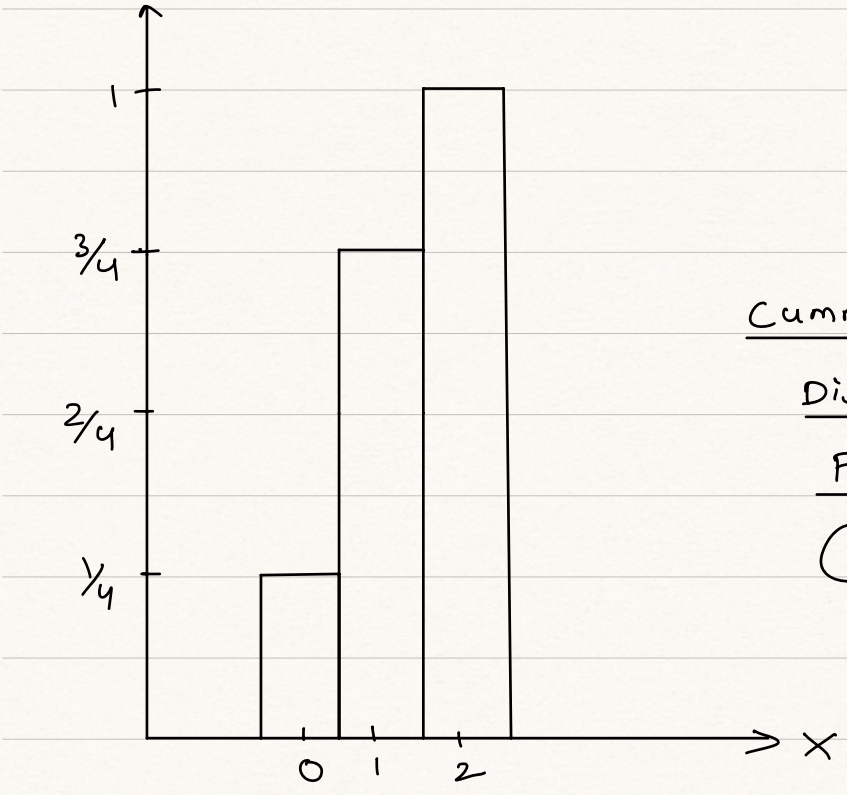
- It is used to describe probability distribution of random variables in a table.
- with the help of these we can plot a CDF plot.
- CDF finds the cummulative probability for given value.
- It can be used for both Discrete Random Variables and Continous Random variables.
- It is used to compare the probabilities between values under certain conditions.
- For discrete distribution functions, CDF gives the probability values bill what we specify.
- For continous distribution functions, it gives the area under the probability density function up

to given value specified.

* Cummulative implies to adding previous values to the current one. Similar to calculating the cummulative frequencies.

ex:- On toss of two coins

X	0	1	2
P(X)	1/4	2/4	1/4
F(X) Cummulative Probability	1/4 = 1/4	1/4 + 2/4 = 3/4	1/4 + 2/4 + 1/4 = 1

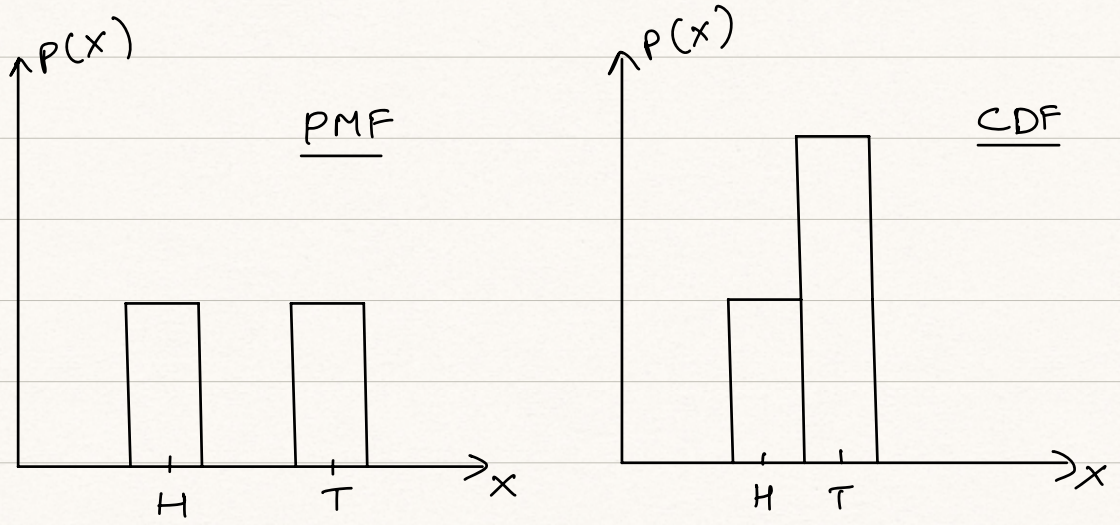


Cummulative Distribution Function (CDF)

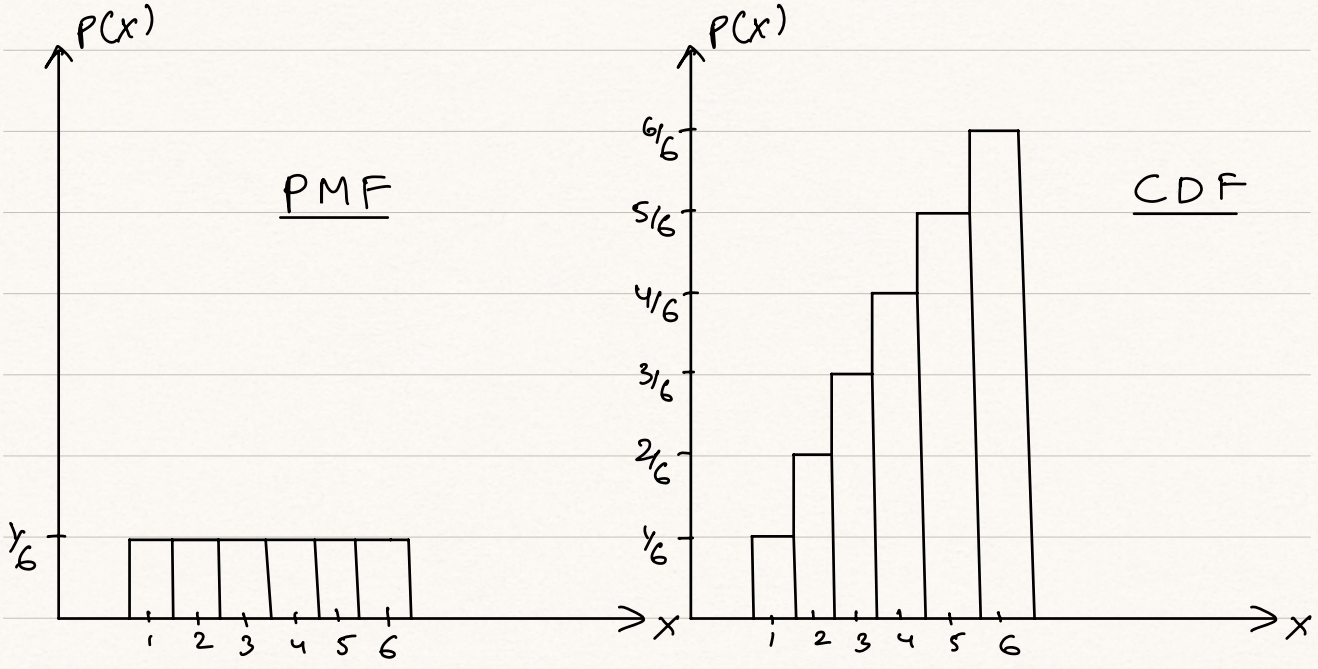
so if we want to find at most or max 1 head i.e $\{HT, TH, TT\} = 3/4$

we can see the same from CDF plot.

ex:- Toss a coin and plot PMF, CDF

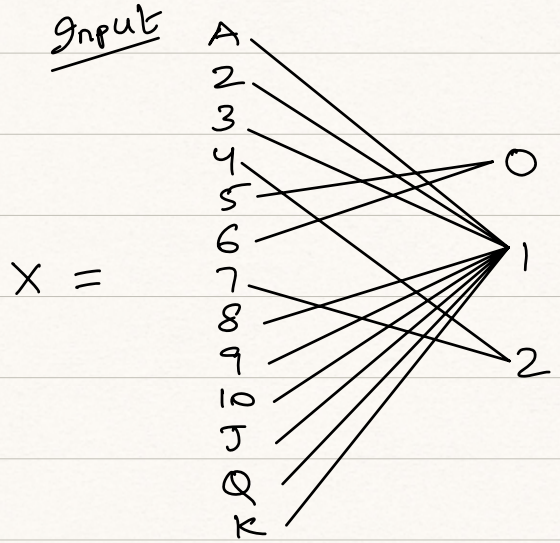


Roll a die



Q) From a deck of spade cards (13) we have removed 5 & 6 and inserted extra 4 & 7.
Draw PMF and CDF for every event in sample space.

Sample space = {A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K}



$P(5) = 0$

$P(6) = 0$

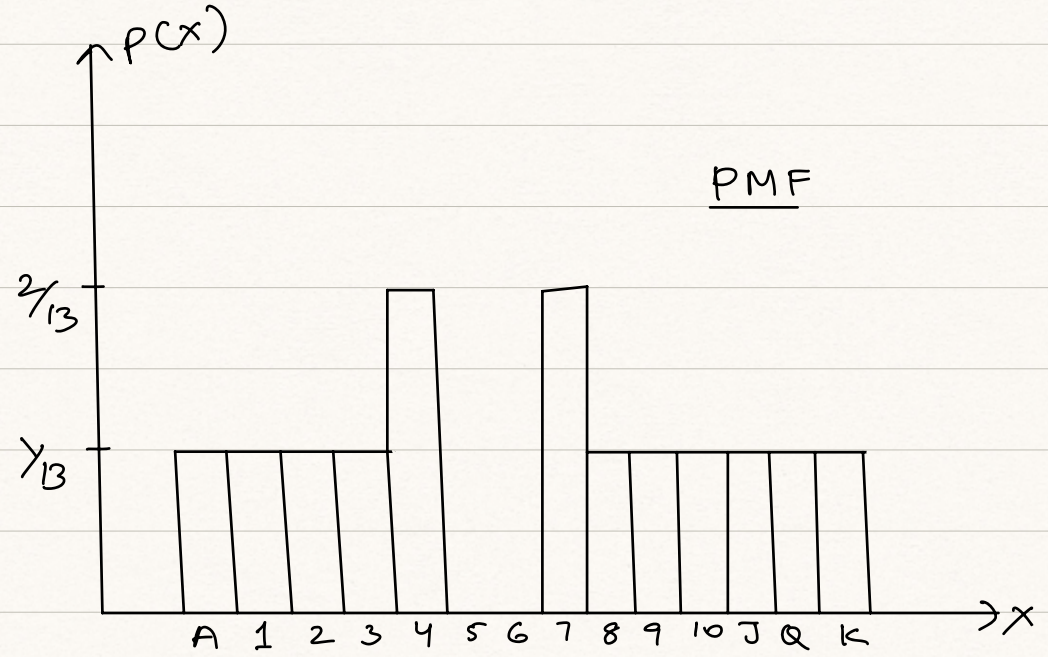
$P(4) = 2/13$

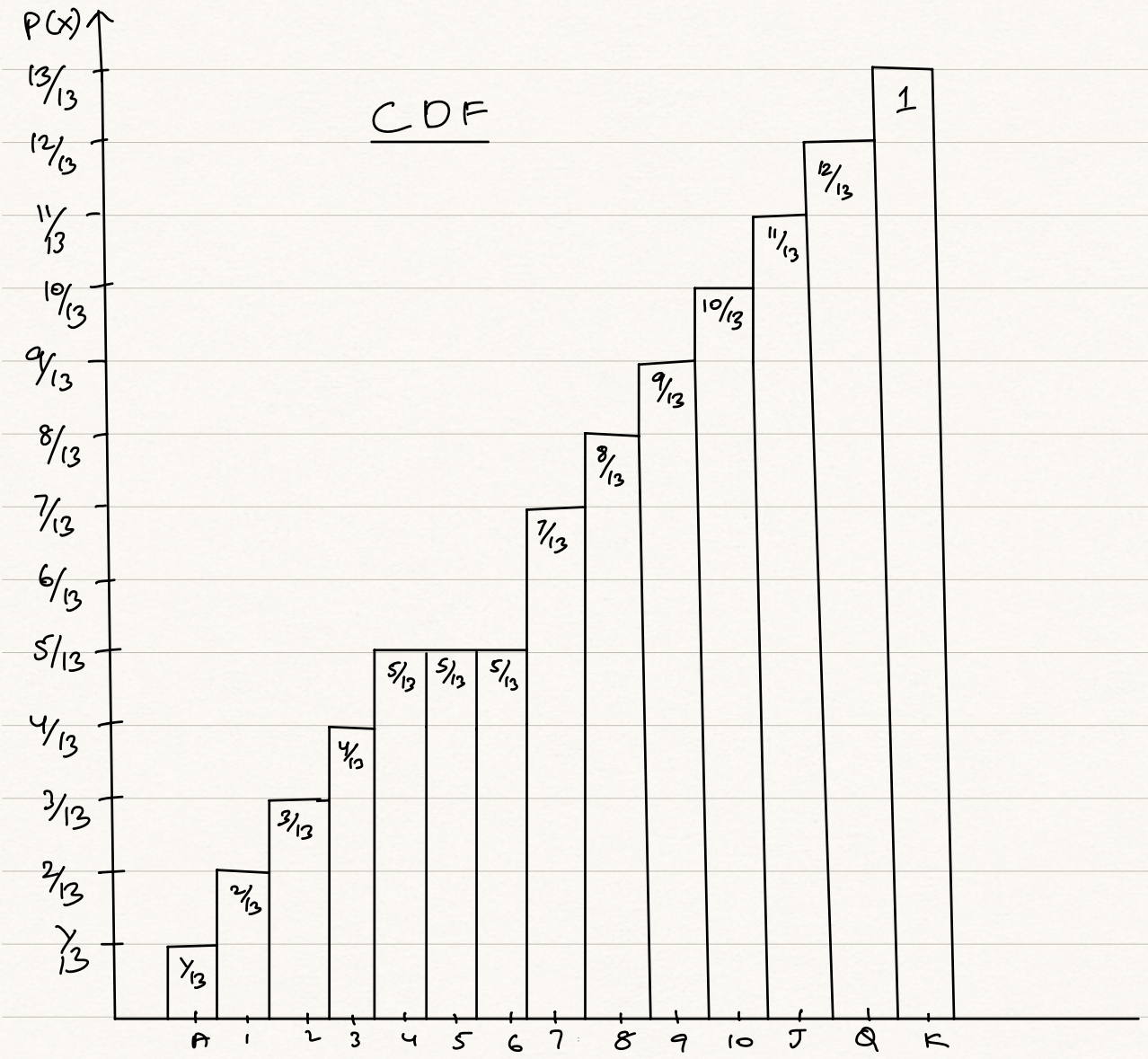
$P(7) = 2/13$

for rest of cards

$P(x) = 1/13$

total cards = 13




P(x)
CDF
13/13
12/13
11/13
10/13
9/13
8/13
7/13
6/13
5/13
4/13
3/13
2/13
1/13
1/13
2/13
3/13
4/13
5/13
5/13
5/13
7/13
8/13
9/13
10/13
11/13
12/13
1
A
1
2
3
4
5
6
7
8
9
10
J
Q
K

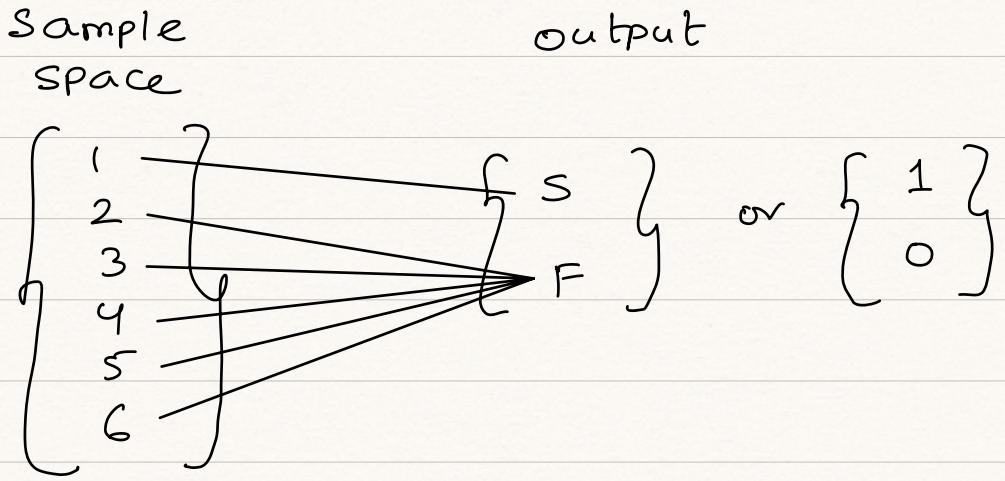
Bernoulli's Distribution PMF & CDF

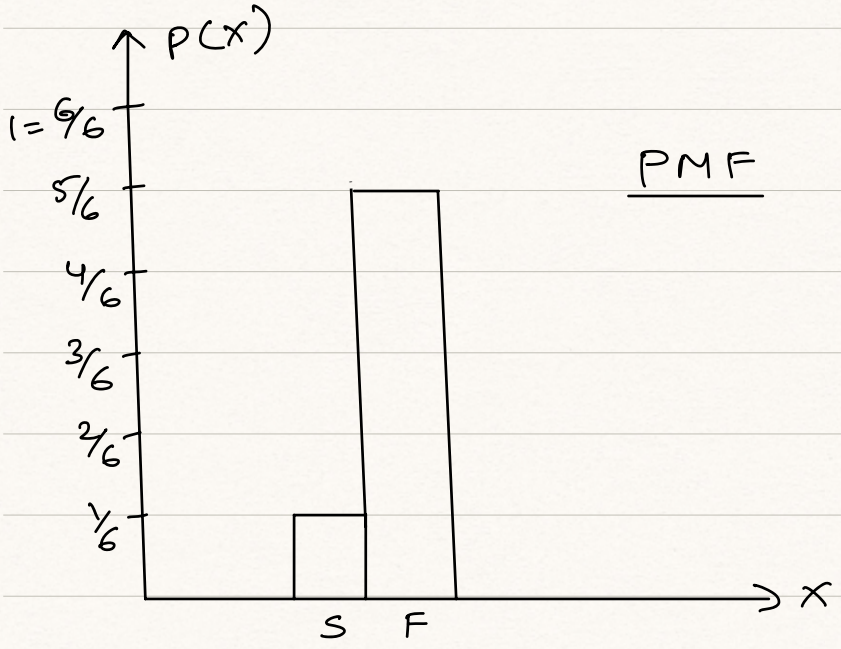
Here the random variable has only two values success (1) or failure (0).

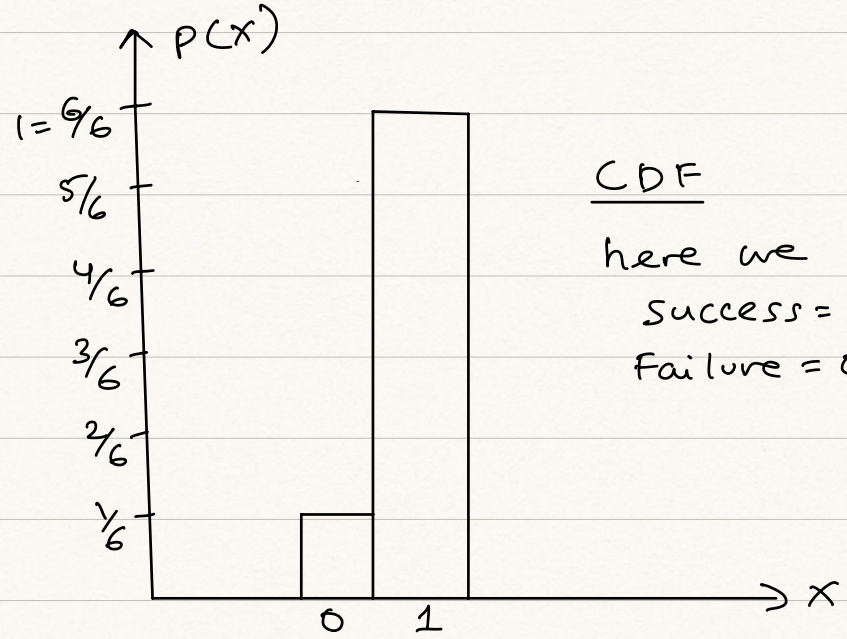
ex:- Getting 1 in toss of a die

$$P(\text{success}) = 1/6$$

$$P(\text{failure}) = 5/6$$







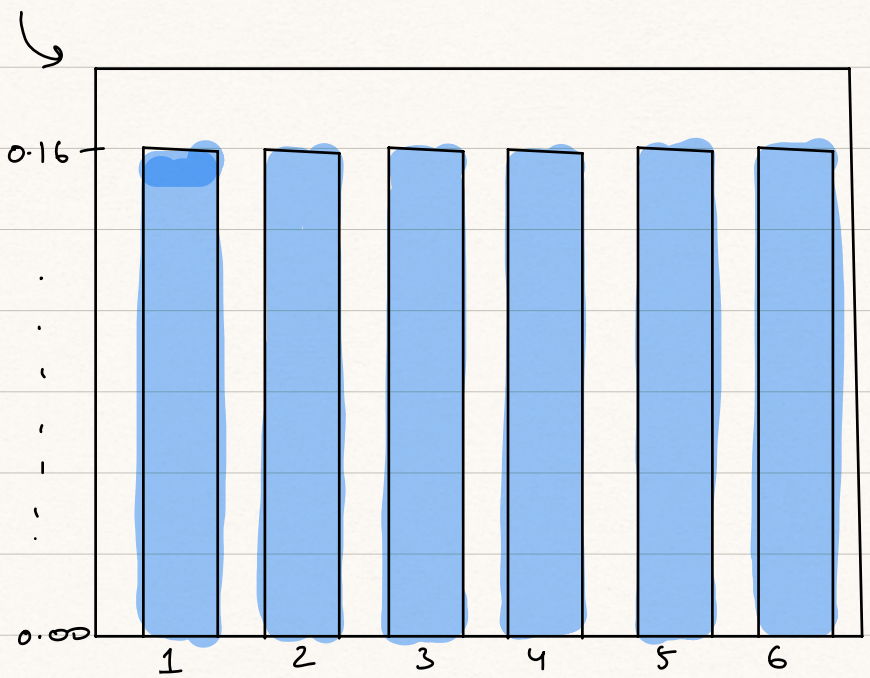
CDF

here we use
Success = 1
Failure = 0

* we can plot this using python.

ex:- faces, pFace = np.arange(1,7), np.repeat(1/6,6)

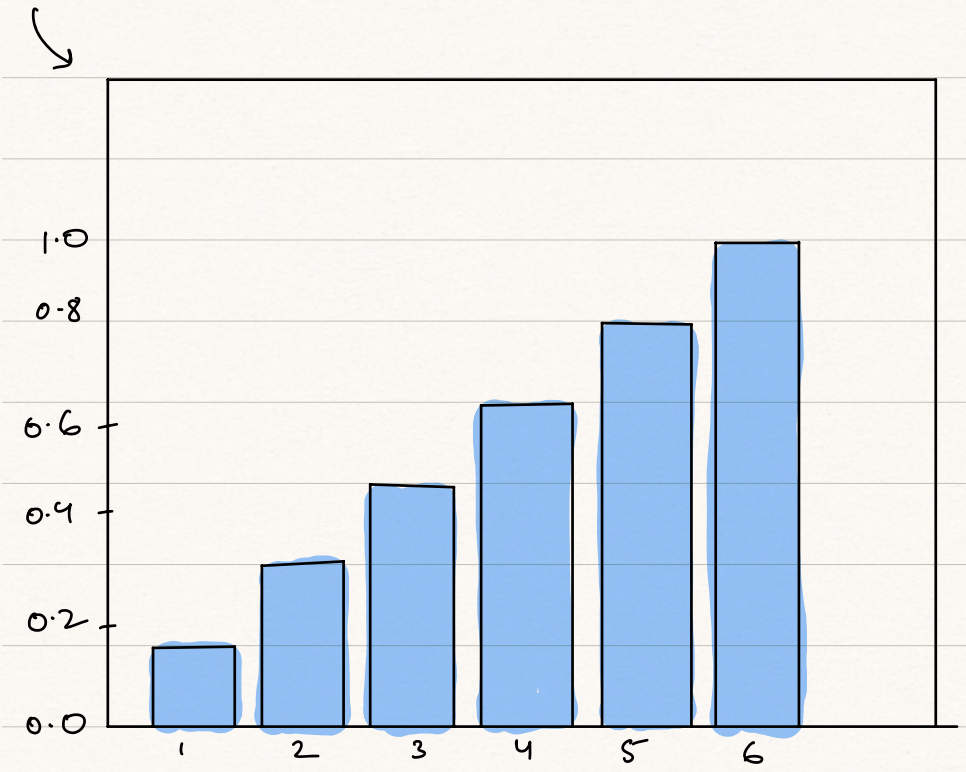
sns.barplot(x=faces, y=pFace, color="blue")



PMF

Cummulative Distribution function

sns.barplot(x=faces, y=np.cumsum(pFace), color="blue")



If we want to plot PMF & CDF and find mean variance and standard deviation using python.

ex:-

```
a = pd.DataFrame({'goals' = [0, 1, 2, 3, 4],
                  'P(x)' = [0.18, 0.34, 0.35, 0.11, 0.02]})
```

a

	goals	P(x)
0	0	0.18
1	1	0.34
2	2	0.35
3	3	0.11
4	4	0.02

$b = a \cdot T$

b

↳

	goals	P(x)
0	0	0.18
1	1	0.34
2	2	0.35
3	3	0.11
4	4	0.02

$mean = \sum x * p(x)$

```
# To calculate mean
# we use zip function and create zip object and
# then create list using the zip object
# use for loop and iterate over the list and
# find mean using Summation of goals * P(x).

mean = 0
for i, j in list(zip(b['goals'], b['P(x)'])):
    mean += i*j
print(mean)
```

↳

1.4500...2

Variance = $\Sigma (x-\mu)^2 + P(x)$

Here μ is mean calculated above

```
Variance = 0
Mean = 1.4500..2
for i,j in list(zip(b['goals'],b['P(x)])):
    Variance += ((1-mean)**2)*j
Print(variance)
```

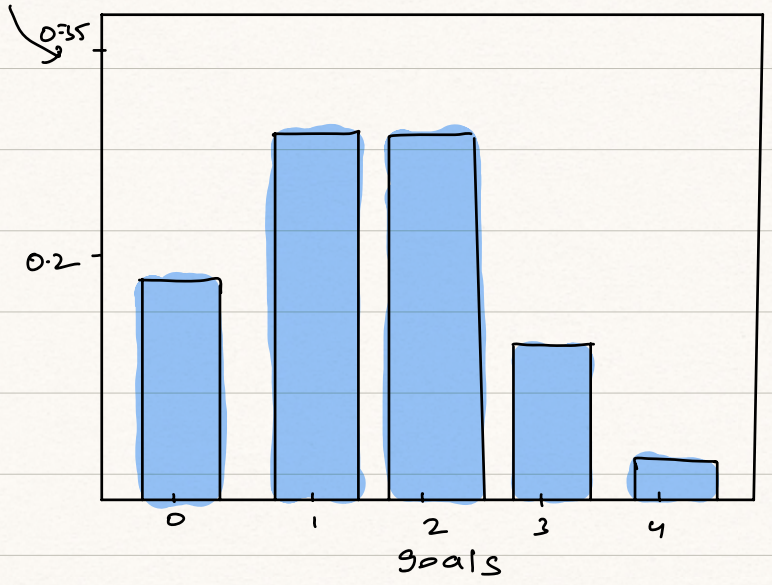
↘ 0.9475

```
std = abs(variance** 0.5)
print(std)
```

↘ 0.973396..

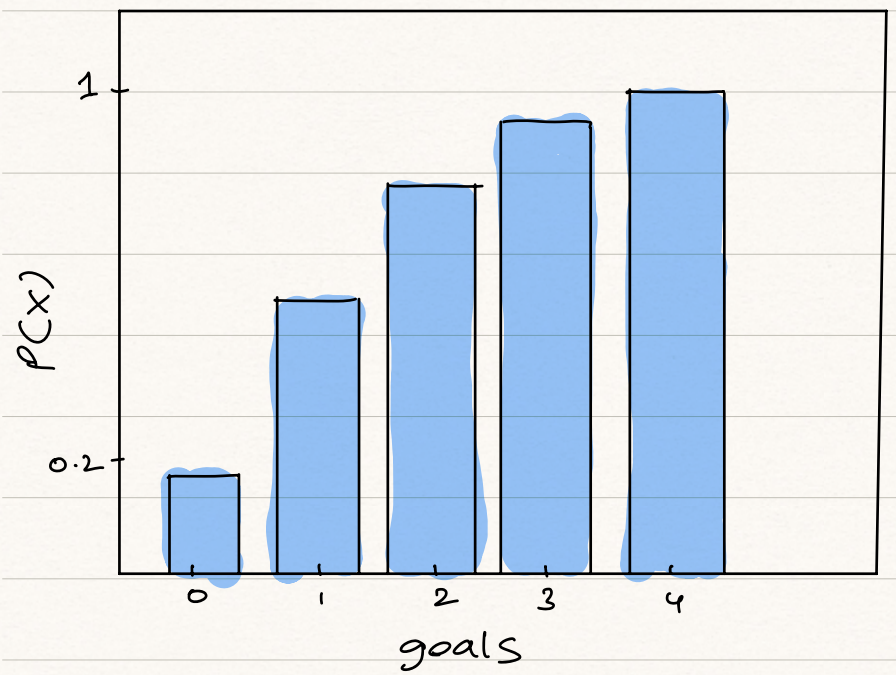
PMF Plot

```
sns.barplot(data=a, x="goals", y="P(x)", color='blue')
```



CDF plot

sns.barplot(x=a["goals"], y=np.cumsum(a['P(x)']), color='blue')



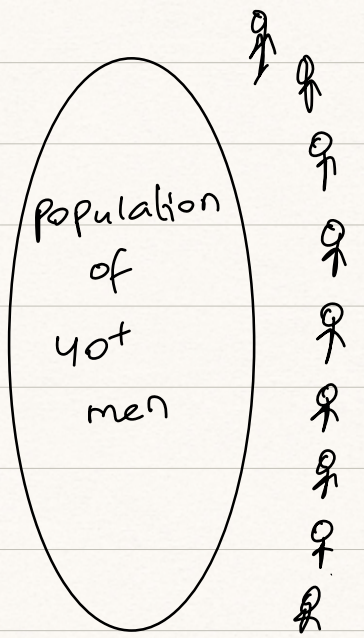
Binomial distribution - refers to population so not finite and with replacement.

Hypergeometric distribution - refers to a finite sample and without replacement

ex :- Binomial Distribution with coding

Probability of obesity in age 40+ men is 0.014

(i) if we take a sample of 10 people what is the probability of all of them being obese



Since we are dealing sample of any 10 people from the population. No replacement so independent trials.

$P(e_1 \cap e_2 \cap e_3 \cap \dots e_{10}) = (P(e))^n$

as $P(e_1) = P(e_2) = \dots = P(e)$

So $P(\text{all 10 being obese}) = (0.014)^{10}$

$= 2.8925 \times 10^{-19}$

(ii) if we pick 2 people out of 10 people what is the probability of one being obese and other being not obese

here choose r from n so ${}^{n}C_{r}$

$$= {}^{10}C_{2} = \frac{10!}{8!\,2!} = 45$$

So we can pick 2 people from 10 people in 45 ways.

Now in these one being obese and other not being obese is

$$45 \times (0.014)$$??

From 10 People:

$$P(\text{'0' being obese}) = {}^{10}C_{0}\,(0.014)^{0} \times (1-0.014)^{10-0}$$

$$= 1 \times 1 \times (0.986)^{10}$$

$$= 0.8684$$

$$P(\text{'10' being obese}) = {}^{10}C_{10} \times (0.014)^{10}\,(1-0.014)^{10-10}$$

$$= (0.014)^{10}$$

$$= 2.8925 \times 10^{-19}$$

$$P(\text{1 being obese}) = {}^{10}C_{1}\,(0.014)^{1} \times (1-0.014)^{9}$$

$$= 0.1233$$

$$P(\text{exactly 3 being obese}) = {}^{10}C_3 (0.014)^3 (0.986)^7$$

$$= 0.00029833$$

$$P(2 \text{ being obese}) = {}^{10}C_2 (0.014)^2 (0.986)^8$$

$$= 0.007879$$

$$P(\text{min 3 being obese}) = 1 - (P(0) + P(1) + P(2))$$

$x = 3, 4, 5, \ldots 10$

$$= 1 - (0.8684 + 0.1233 + 0.007879)$$

$$= 0.00042077$$

$$P(7 \text{ men being obese}) = 1 - P(\text{min 3 being obese})$$

$x = 7, 8, 9, 10$

$$= 1 - 0.00042077$$

$$= 0.99957923$$

$$P(\text{exactly 7 being obese}) = {}^{10}C_7 (0.014)^7 \times (0.986)^3$$

$$= 0.12126 \times e^{-10}$$

Coding of Binomial distribution using python classes:-

```
import math
from math import factorial as fact
Class Binomial:
    def __init__(self, population_size, P_S):
        self.n = population_size
        self.p = P_S
        self.dtable = self.binomialDistTable()
        self.dft = self.dtable.T
    def binomialDistTable(self):
        l=[]
        for x in range(0,self.n+1):
            choice = fact(self.n)/((fact(self.n-x))*fact(x))
            ps = self.p**x
            pf =(1-self.p)**(self.n-x)
            p_x = choice*ps*pf
            l.append((x,p_x))
        return pd.DataFrame(l, columns=['x', 'prob_x'])

    def min_of(self, num):
        if num>0 and num<=self.n:
```

```
            return self.dtable.iloc[num:,1].values.
                                        cumsum()[-1]

        else:
            Print(f'num {num} does not qualify')

    def max_of(self,num):
        if num>0 and num<= self.n:
            return self.dtable.iloc[:num+1,1].
                                values.cumsum()[-1]

        else:
            Print(f'num {num} does not qualify')

    def pmfplot(self):
        return sns.barplot(data=self.dtable, x='x',
                                y='prob_x')

    def cdfplot(self):
        return sns.barplot(data=self.dtable, x='x',
                y=self.dtable['prob_x'].values.cumsum())

# Creating Binomial class objective
Bd = Binomial(10,0.014)
# Probability of min 3 being obese
Bd.min_of(3)
    ↳ 0.00030587
```

```
Bd.max_of(3)
```

↳ 0.999992459

```
Bd.dtable
```

↳

	X	Prob-X
0	0	$8.68.. \times e^{-01}$
1	1	..
2	2	..
3	3	$2.9833.. \times e^{-04}$
4	4	..
5	5	..
6	6	..
7	7	..
8	8	..
9	9	..
10	10	..

Probability of exactly getting x=3

```
Bd.dtable.loc[3,'Prob-x']
```

↳ 0.00029833

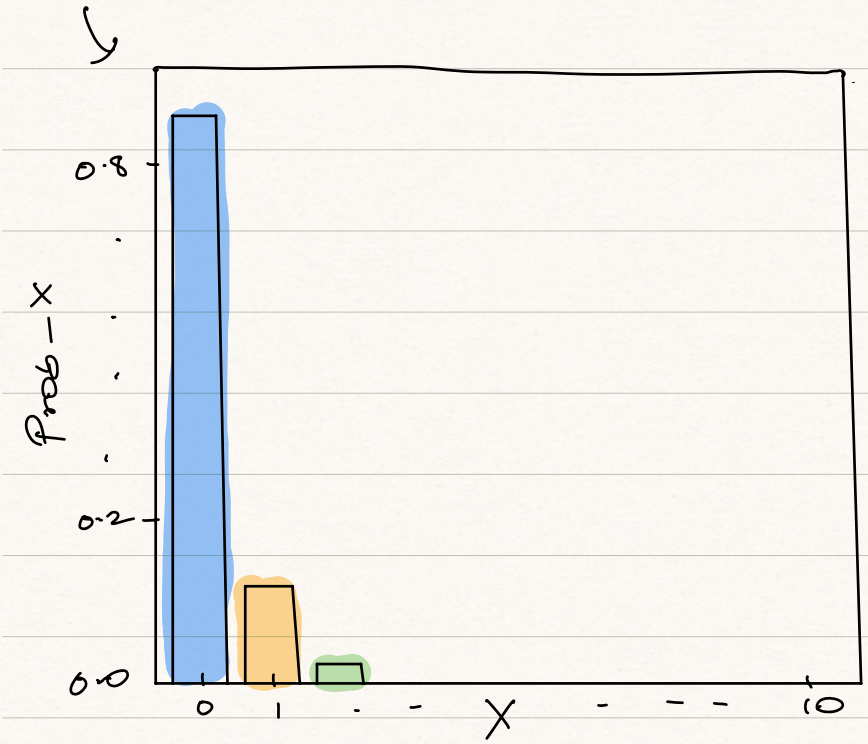
Probability Distribution table

```
Bd.dft
```

	0	1	2	3	4	5	6	7	8	9	10
X	0.00..	1.00..	.	.	.	.	.	.	.	.	.
Prob_X	0.86..	0.12...	.	.	.	.	.	.	.	.	.

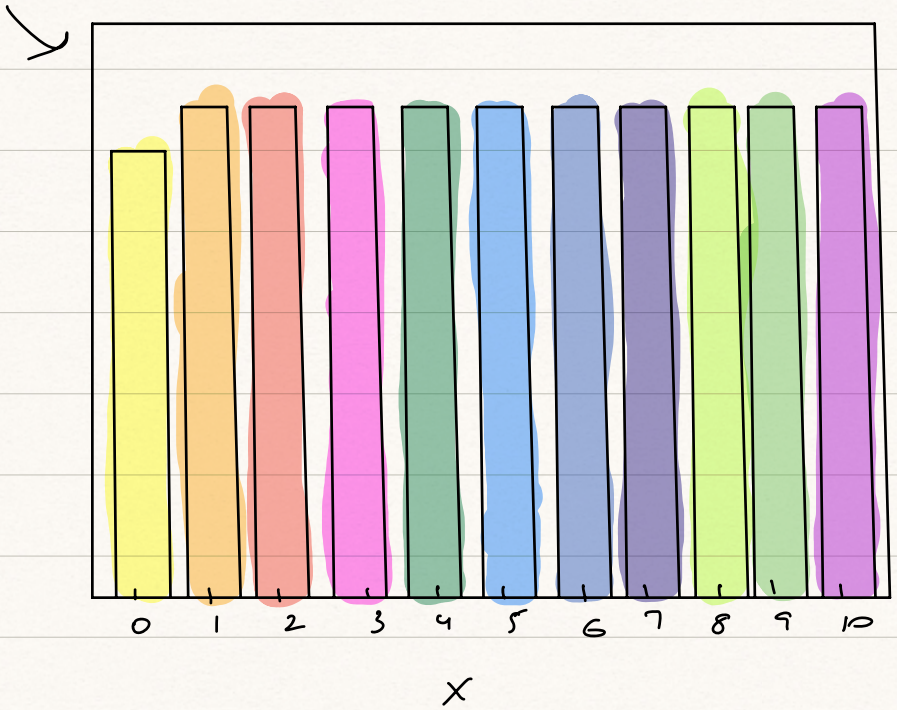
Pmf plot for Binomial

Bd.pmfPlot( )



Cdf plot for Binomial

Bd.cdfPlot( )



we have a scipy stats function called

scipy.binom.pmf(x, n, p)

or

scipy.binom.cdf(x, n, p)

Hypergeometric distribution.

- Another Discrete probability Distribution
- we should already know

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

- Hypergeometric distribution is not based on bernoulli trials as without replacement so no longer independent trials.
- Main difference from Binomial distribution is this is done without replacement.

Let us consider an example first for better understanding.

ex:- An urn contains 6 redballs and 14 Yellow balls. 5 balls are randomly drawn without replacement. what is the probability exactly 4 red balls are drawn.

So once the ball is drawn then it is placed aside. So the probability of the next draw depends on the first one so no longer independent trials so Binomial cannot be used.

let us Calculate using logic before deriving formulas.

If we are randomly selecting 5 balls then any sample of 5 balls is equally likely, so:

P(Exactly 4 red balls)

$$= \frac{\#\text{ of samples that result in 4 red balls and 1 yellow ball}}{\#\text{ possible samples of size 5}}$$

$$= \frac{{}^{6}C_{4} \times {}^{14}C_{1}}{{}^{20}C_{5}}$$

(${}^{6}C_{4}$: from 6 red balls pick 4 red balls; ${}^{14}C_{1}$: from 14 yellow balls pick 1 yellow (not red) ball.)

$$= 0.01354$$

here if we think logically

on first trial $P(\text{Red}) = 6/20$

on second trial if first ball is Red $P(\text{Red}) = 5/19$

So P(success) keeps changing so not independent.

For us to use Binomial distribution the Question should be

ex:- An urn contains 6 redballs and 14 Yellow balls. 5 balls are randomly drawn with replacement. what is the probability exactly 4 red balls are drawn.

$P(\text{success}) = \frac{6}{20}$ and it will not change

$$P(4 \text{ red balls}) = {}^{5}C_{4} \left(\frac{6}{20}\right)^{4} \times \left(\frac{14}{20}\right)^{1}$$

$$= 0.02835.$$

different from Hypergeometric probability.

Hypogeometric distribution

- Suppose we are randomly sampling n objects without replacement from a source of N objects that contains a successes and N-a failures
- X represents the number of successes in the sample.

Then x has a hypergeometric distribution:

(PMF)

$$P(X=x) = \frac{{}_{a}C_{x} * {}_{(N-a)}C_{(n-x)}}{{}_{N}C_{n}}$$

or

$P(x)$

Here we see
$N-a+a = N$
$x+n-x = n$
we can use for double check

for $x = Max(0, n-(N-a)), \ldots, min(a,n)$

i.e first we need to pick x from total successes of a i.e ${}_{a}C_{x}$

then we need to pick n-x failures from total N-a failures

Denominator is total samples i.e picking n objects from N objects.

Here x is # of successes

it cannot be bigger than # of objects sampled n and also cannot be bigger than the total successes i.e $\min(a, n)$

The min value x can take cannot be lesser than 0 or n minus total failures i.e $n-(N-a)$

so min value of x is $\max(0, n-(N-a))$

mean of Hypergeometric distribution

$$\mu = n \times \frac{a}{N}$$

↑ no. of samples

↑ proportion of successes in the total population N

$$\text{Variance} = \sigma^2 = V(X) = \left(\frac{N-n}{N-1}\right) * n * \left(\frac{a}{N}\right)\left(\frac{N-a}{N}\right)$$

ex:- suppose a large high school has 1100 female students and 900 male students.

A random sample of 10 students is drawn.

what is the probability exactly 7 of the students selected are female?

Ans) Though not mentioned this is without replacement

$P(X=7) = ??$

logically:- denominator : total samples

so picking 10 from 2000

i.e $^{2000}C_{10}$.

Numerator: probability of picking exactly 7 female students is from 1100 we pick 7 i.e $^{1100}C_7$ and remaining 3 students we pick from 900 male students

i.e $^{900}C_3$ i.e $^{1100}C_7 \times {}^{900}C_3$

$$\therefore P(X=7) = \frac{^{1100}C_7 \times {}^{900}C_3}{^{2000}C_{10}}$$

$1100+900 = 2000$

$7+3 = 10$

double check satisfied

$$= 0.166490$$

gf formula $N = 2000$

$n = 10$

$a = 1100$

$x = 7$

$$P(X=7) = \frac{{}^{1100}C_7 \times {}^{(2000-1100)}C_{(10-7)}}{{}^{2000}C_{10}}$$

$$= 0.166490.$$

If we assume this is with replacement and use Binomial distribution instead

So $P(\text{success}) = \frac{1100}{2000}$

$$P(X=7) = {}^{10}C_7 \left(\frac{1100}{2000}\right)^7 \times \left(1 - \frac{1100}{2000}\right)^{10-7}$$

$$= 0.166478$$

we can see that this is pretty close to probability using Hypogeemetric distribution.

* The binomial distribution can sometimes provide a reasonable approximation to the hypogeometric distribution.

Rough guidline: If we are not sampling more than say 5% of the population, the binomial provides a reasonable approximation.

So we use binomial distribution where ever we can approximate for easier calculations.

Multi-variate hypergeometric

Example: Suppose a business employs 12 democrats, 24 Republicans, and 8 independents.

i.e $12+24+8 = 44$ total.

If a random sample of 6 employees is drawn, what is the probability there are 3 democrats, 2 republicans and 1 independent in the sample.

$$P(x) = \frac{{}^{12}C_3 \times {}^{24}C_2 \times {}^{8}C_1}{{}^{44}C_6}$$

So we can apply hypergeometric methods to more than two types of object and this is sometimes called multi-variate Hypergeometric distribution.

Hypogeometric Distribution with coding

Simplified formula used for coding is

$$P(X=x) = \frac{n!\,k!\,(N-n)!\,(N-k)!}{N!\,x!\,(k-x)!\,(n-x)!\,(N-k-n+x)!}$$

here

N - Total size of sample/population

k - total size of interest group

n - total number picked

x - input of random variable.

Hypogeometric distribution using python classes:-

```
import math
from math import factorial as fact
Class Hypogeometric:
    def __init__(self, population_size, P_S):
        self.n = population_size
        self.p = P_S
        self.dtable = self.hypogeometricDistTable()
        self.dft = self.dtable.T
```

```
def hypogeometricDistTable(self):
    l=[]
    for x in range(0, self.k+1):
        P_x = (fact(self.n)*fact(self.k)*
              fact(self.N-self.n)*fact(self.N-
              self.k))/(fact(self.N)*fact(x)*
              fact(self.k-x)*fact(self.n-x)*
              fact(self.N-self.k-self.n+x))
        l.append((x,P_x))
    return pd.DataFrame(l, columns=['x','prob_x'])

def min_of(self,num):
    if num>0 and num<=self.n:
        return self.dtable.iloc[num:,1].values.
                                    cumsum()[-1]
    else:
        Print(f'num {num} does not qualify')

def max_of(self,num):
    if num>0 and num<=self.n:
        return self.dtable.iloc[:num+1,1].
                                values.cumsum()[-1]
```

```
        else:
            print(f'num {num} does not qualify')

    def pmfplot(self):
        return sns.barplot(data=self.dtable, x='x',
                           y='prob-x')

    def cdfplot(self):
        return sns.barplot(data=self.dtable, x='x',
                           y=self.dtable['prob-x'].values.cumsum())

# we create a class object
hG = Hypogeometric(230,5,10)
# probability of min 3 being obese
.min_of(3)
    ↳ 0.0005721...

.max_of(3)
    ↳ 0.999990...
```

hg.dtable

↳

	X	Prob-X
0	0	7.99... xe^{-01}
1	1	--
2	2	--
3	3	5.6307.. xe^{-04}
4	4	--
5	5	--

Probability of exactly getting x=3

.dtable.loc[3, 'Prob-x']

↳ 5.6307 x e^{-04}

Probability Distribution table

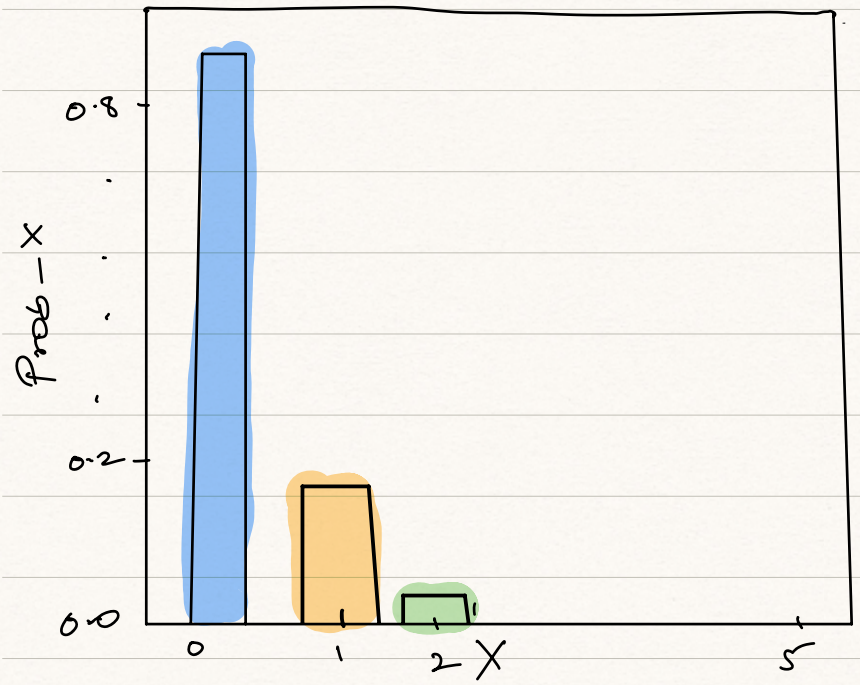
hg.dft

	0	1	2	3	4	5
X	0.00..	1.00..	.	-	-	.
Prob_X	0.799...	0.184..	.	.	-	-

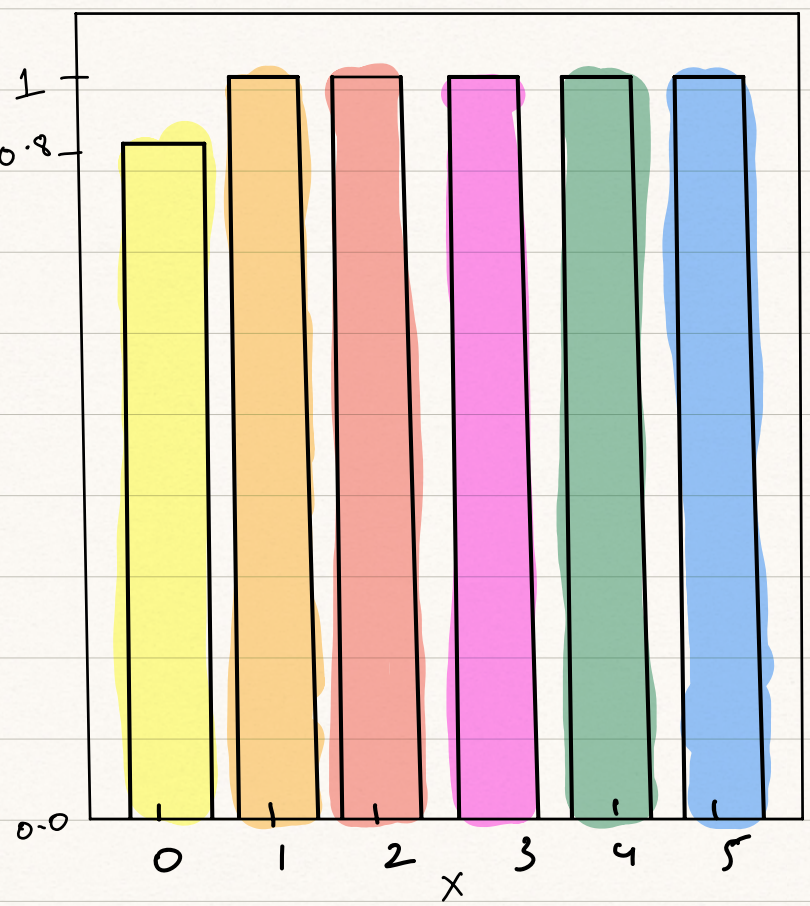
Pmf plot for hypogeometric

.pmfPlot()

↳



#cdfplot for hypogeometric



In scipy we have a stats hypergeometry function called

scipy.stats.hypergeom( )

logical explanation of pmf function of hypergeometric function.

PMF of hyper geometric function $= \frac{\{a \text{ successes}\} \cap \{n-a \text{ failures}\}}{\text{Choices of 'n' out of N.}}$

$= \frac{\text{no.of ways of choosing a group of 'n' with 'a' successes}}{\text{no.of ways to choose 'n' from 'N'.}}$

Geometric distribution: To better understand let us consider an example:

- In a large population of adults, 30% have received CPR training.

If adults from this population are randomly selected, what is the probability that the 6th person sampled is the first that have received CPR training?

we can use geometric distribution to calculate this probability.

The geometric distribution is the distribution of the number of trials needed to get the first success in independent repeated Bernoulli trials.

Repeated independent Bernoulli's trials:

Suppose:

- There are independent trials
- Each trial results in one of two possible mutually exclusive outcomes, labelled success and failure.

on any individual trial:

- P(success) = p and this stays constant from trial to trial
- P(Failure) = $1-p$
- X represents the number of trials needed to get the first success.

pmf

for the first success to occur on x^{th} trial:

1. The first $x-1$ trials must be failures $\rightarrow (1-p)^{x-1}$
2. The x^{th} trial must be success $\rightarrow p$

since trials are independent.

$$P(X = x) = (1-p)^{x-1} p$$

$x = 1, 2, 3, \ldots$ (no upper bound)

$$\text{mean} = \mu = \frac{1}{p}$$

$$\text{Variance} = \sigma^2 = \frac{1-p}{p^2}$$

so in geometric distribution we just need probability of success.

so if we go back to the example.

- In a large population of adults, 30% have received CPR training.

If adults from this population are randomly selected, what is the probability that the 6th person sampled is the first that have received CPR training?

$p = 30\% = 0.3$ on any individual trial

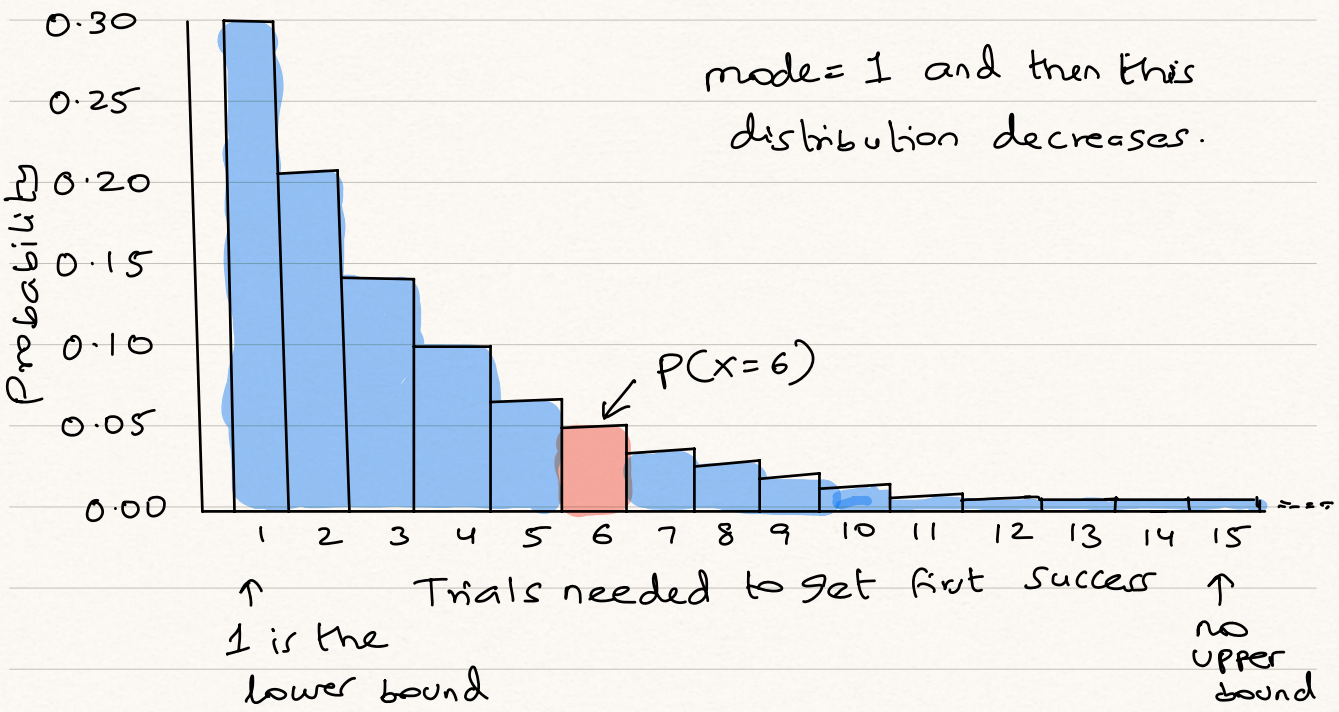
$x = 6$.

$P(X=6)$ = first 5 not CPR × 6th CPR training.

$= (1-0.3)^{6-1}(0.3)$

$= 0.0504$

if we plot this



Geometric distribution always has Right skewed distribution.

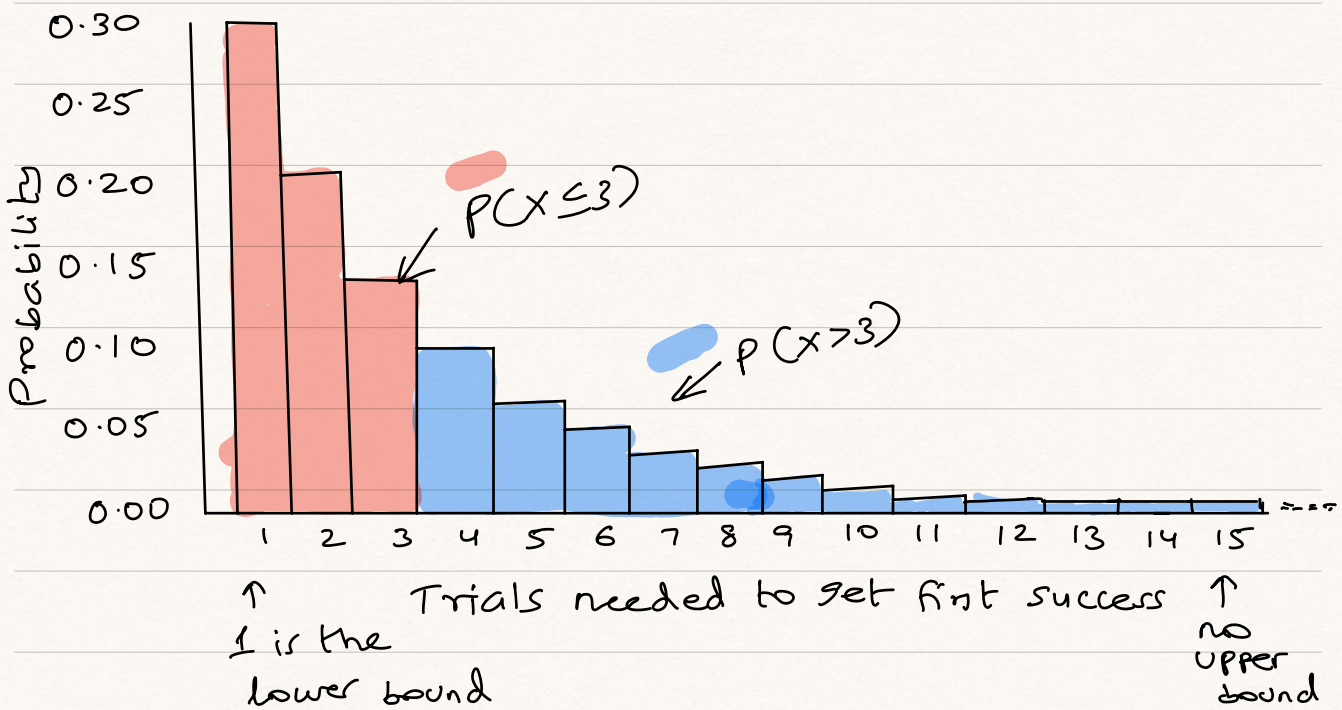
$$\text{mean} = \frac{1}{0.3} = 3.\dot{3}$$

$$\text{Variance} = \frac{1 - 0.3}{0.3^2} = 7.\dot{7}$$

$$\sigma = \sqrt{7.\dot{7}} = 2.79$$

Q) what is the probability that the first person trained in CPR occurs on or before the 3rd person sampled?

$$P(X \leq 3) = ?$$



$$P(X \leq 3) = P(X=1) + P(X=2) + P(X=3)$$

$$= 0.7^0 \times 0.3 + 0.7^1 \times 0.3 + 0.7^2 \times 0.3$$

$$= 0.657$$

or (we use when there are large no. of probabilities)

$$P(X \leq 3) = 1 - P(X > 3)$$

↳ i.e getting three failures.

$$= 1 - 0.7^3$$

$$= 0.65$$

The cummulative distribution function for the geometric distribution:

$$F(x) = P(X \leq x) = 1 - (1-p)^x$$

for $x = 1, 2, 3, \ldots$

Poisson distribution

Suppose we are counting the number of occurences of an event in a given unit of time, distance, area or volume.

For example:

- The number of car accidents in a day.
- The number of dandelions in a square metre plot of land

The number of events is going to be a random variable that may or may not have the Poisson distribution, depending on the specifics of the situation.

* But a poisson distribution is a count of the number of occurences of an event.

If we consider in terms of time

Suppose:

- Events are occuring independently i.e knowing when one event happens has absolutely no information on when another event will occur.
- And the probability that an event occurs in a given length of time does not change

through time i.e the theoretical rate at which the events are occurring does not change through time.

so we can say the events are occuring randomly and independently.

* If these conditions hold then the random variable x which represents the no. of events in a fixed unit of time has poisson distribution.

* Till not we were looking at random variables that take finite values. In poisson distribution the random variable can take values up to ∞. Though the probability near infinity would be minute.

* Here the mean and variance are same and represented as lambda(λ)

$$\mu = \sigma^2 = \lambda .$$

PMF of poisson distribution that we will use to calculate probabilities is

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$x = 0, 1, 2, \cdots, \infty$

e is base of natural logirithm

$e \simeq 2.71828$

It is an irrational number that has infinite decimal places

ex:- Plutonium-239 is an isotope of plutonium that is used in nuclear weapons and reactors.

One nanogram of plutonium-239 will have an average of 2.3 radioactive decays per second, and the number of decays will follow a poisson distribution.

what is the probability that in a 2 second period there are exactly 3 radioactive decays?

let x represent the number of decays in a 2 second period.

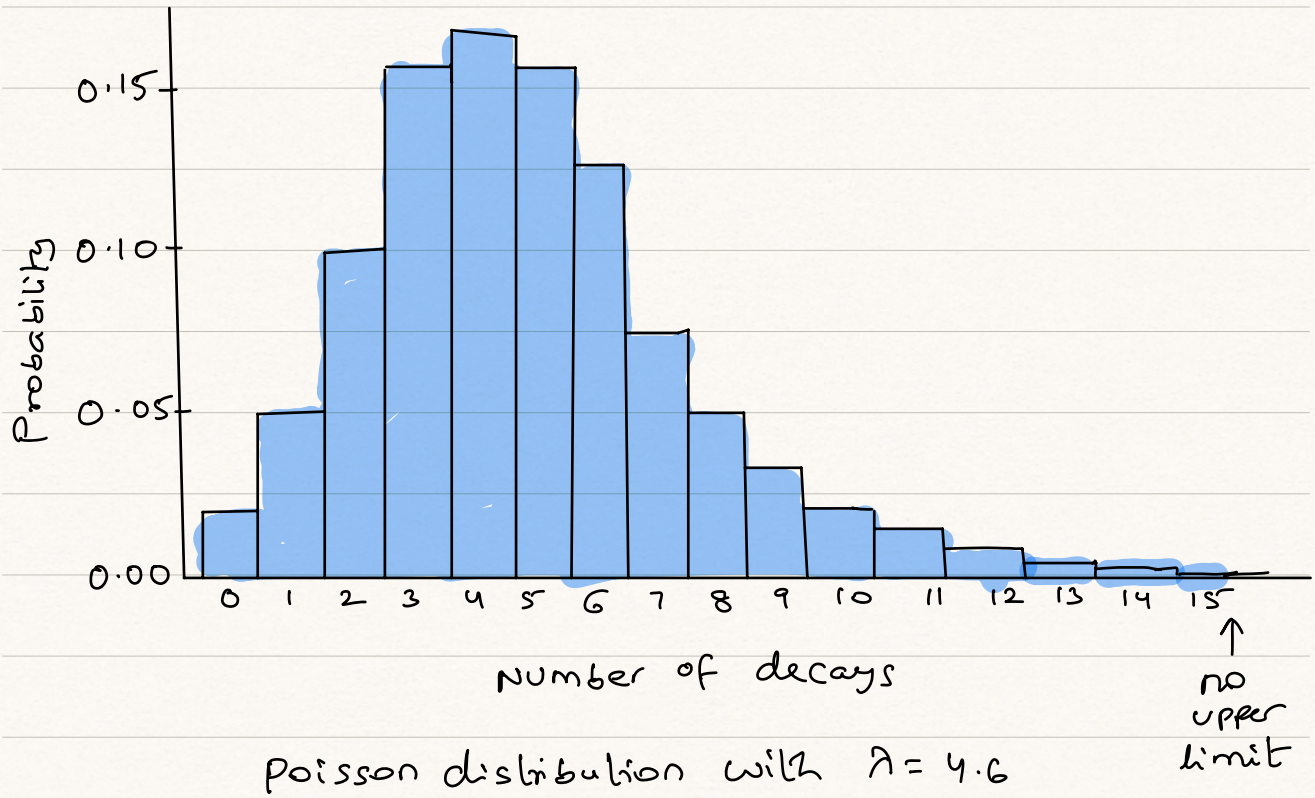
mean for 2 second = 2.3×2

$\lambda = 4.6$

$$P(X=3) = \frac{4.6^3 e^{-4.6}}{3!}$$

$= 0.163$

if we were to calculate probabilities for the different possible values x can take and plot them we will get the probability distribution plot



Poisson distribution with $\lambda = 4.6$

$$\mu = \lambda = 4.6$$

$$\sigma^2 = \lambda = 4.6$$

$$\sigma = \sqrt{4.6}$$

* when λ is large the distribution is close to symmetric.

* when λ is close to 0, the right skewness can

be pretty strong

$P(X \leq 3) = ?$

$= P(X=0) + P(X=1) + P(X=2) + P(X=3)$

$= \frac{4.6^0 e^{-4.6}}{0!} + \frac{4.6^1 e^{-4.6}}{1!} + \frac{4.6^2 e^{-4.6}}{2!} + \frac{4.6^3 e^{-4.6}}{3!}$

$= 0.010 + 0.046 + 0.106 + 0.163$

$= 0.326$

The relationship between Binomial and Poisson Distributions

- The binomial distribution tends toward the poisson distribution as $n \to \infty$, $p \to 0$ and np stays constant.
- The poisson distribution with $\lambda = np$ closely approximates the binomial distribution if n is large and p is small.

→ this is why the radioactivity decays of plutonium has a poisson distribution. Even for a tiny bit of plutonium, there are a very large number of atoms, and each one has a tiny probability of experiencing a radioactive decay in a two second period. So in the example we just worked through, it was in its underlying nature a binomial problem with a very large n and very small p. And thats why the number of radio active decays is very well approximated by poisson distribution.

- The poisson distribution is typically used as an approximation to the true underlying reality.
- It can be difficult to determine whether a

random variable has a Poisson distribution.