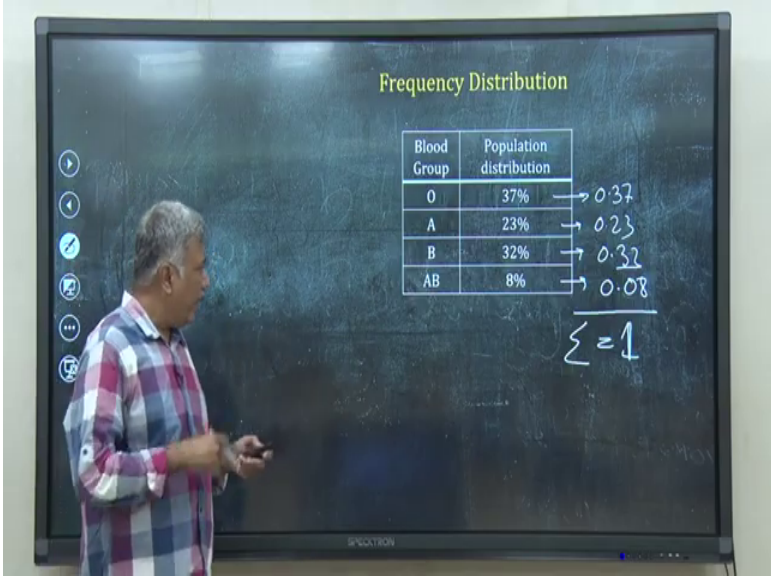
Data Analysis for Biologists

Professor Biplab Bose
Department of Biosciences & Bioengineering
Mehta Family School of Data Science & Artificial Intelligence
Indian Institute of Technology, Guwahati

Lecture: 2
Discrete Probability Distribution

Welcome back, this is the second lecture of first week of our course. In this lecture, we will discuss about Discrete Probability Distribution.

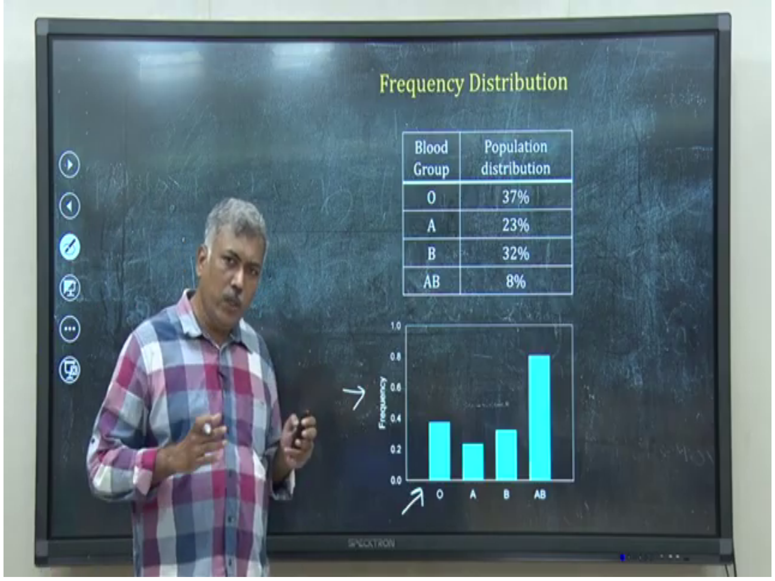
(Refer Slide Time: 00:44)



Frequency Distribution

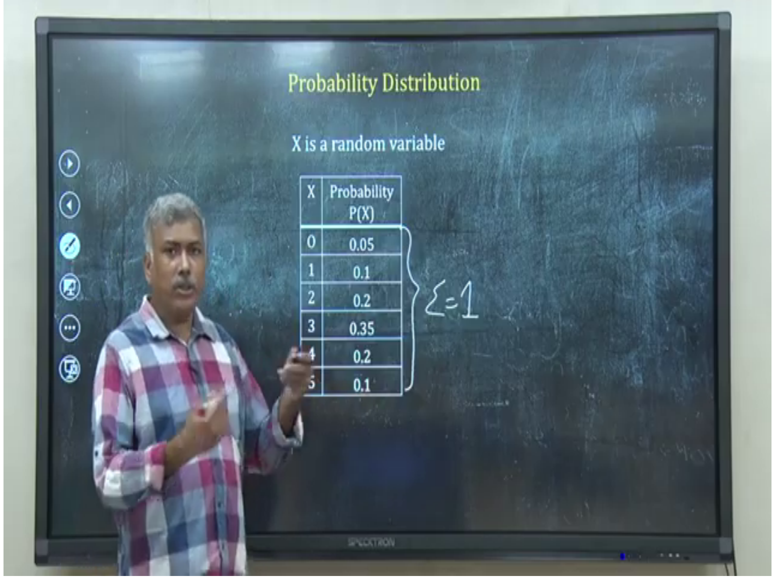
Blood Group	Population distribution	
O	37%	→ 0.37
A	23%	→ 0.23
B	32%	→ 0.32
AB	8%	→ 0.08

$\Sigma = 1$



Frequency Distribution

Blood Group	Population distribution
O	37%
A	23%
B	32%
AB	8%

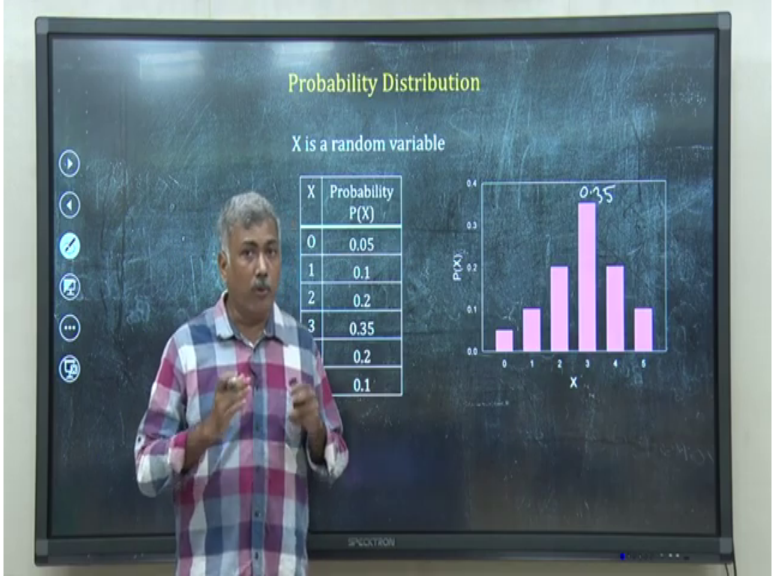


Probability Distribution

X is a random variable

X	Probability P(X)
0	0.05
1	0.1
2	0.2
3	0.35
4	0.2
5	0.1

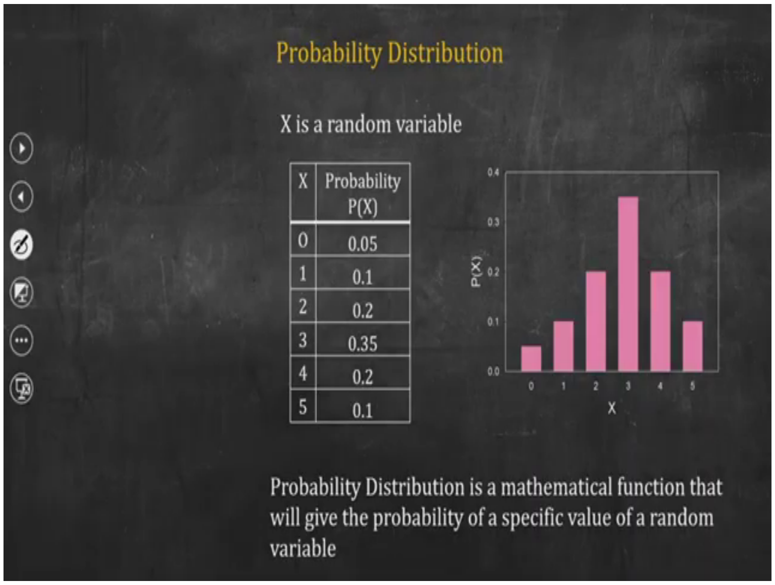
$\Sigma = 1$



Probability Distribution

X is a random variable

X	Probability P(X)
0	0.05
1	0.1
2	0.2
3	0.35
	0.2
	0.1



Probability Distribution

X is a random variable

X	Probability P(X)
0	0.05
1	0.1
2	0.2
3	0.35
4	0.2
5	0.1

Probability Distribution is a mathematical function that will give the probability of a specific value of a random variable

In the last lecture, I left with a particular problem to solve, where I have shown the population distribution of blood group O A B and AB and how much abundant they are in a population. So, we have percentage of each of those. I can represent this percentage also as a fraction, for example, I can write this 37 percent is nothing but 0.37.

Similarly, I can write the other one 0.23, 0.32 and 0.08. And, the summation of all these will be equal to 1, because remember these are frequencies and proportion so, their summation should be 1. Now, I can represent the same thing also as a histogram right. I have the frequency in the vertical axis and those blood groups O A B AB are in the horizontal axis.

So, this is the frequency distribution or frequency histogram for this blood group. So, what we are showing here by this data on the histogram is the frequency distribution of blood groups in our population. Now, similarly, suppose X is a random variable and it can take some discrete 6 values 0, 1, 2, 3, 4, 5 right.

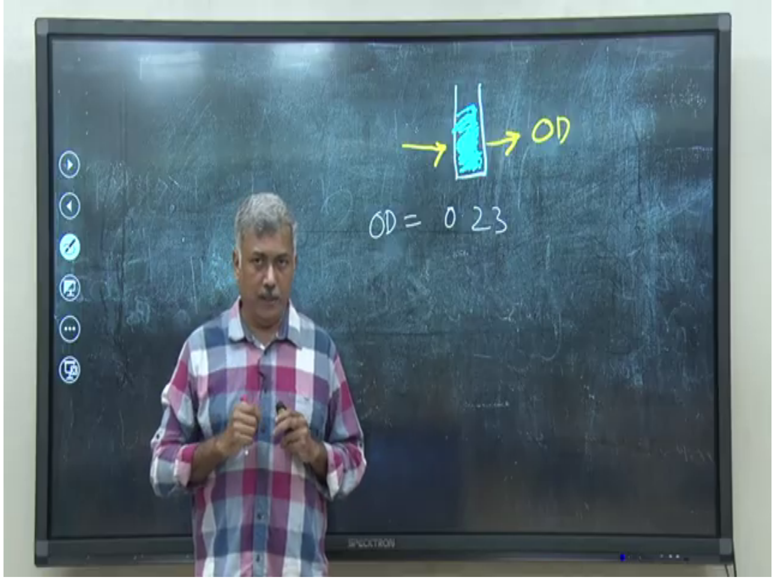
And somehow suppose somehow, we know the probability that, what is the probability that X will be equal to 1, probability of X equal to 0, something like that, and those are all listed here. As these are probability and if you remember summation of all these probability should be equal to 1. So,

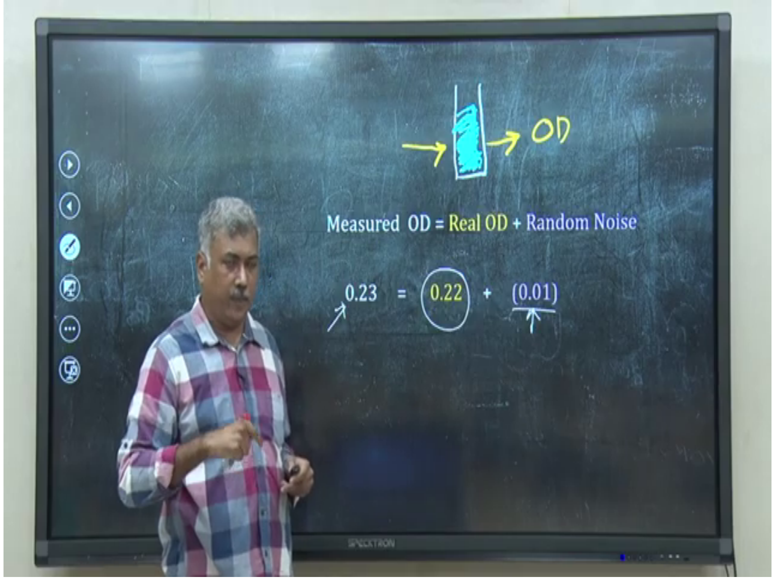
$$\Sigma P(X) = 1$$

Now, I can represent this probability distribution, this probability values also as a histogram right, just like we did for frequency histogram of the blood group. So, this can look like this. So, this third one is this value is 0.35 right, it's the largest one, highest frequency. Now, suppose, I do not know these values.

I do not have this figure either. I want you to give me a mathematical function, which will tell me what will be the probability that X will be equal to 3, or X will be equal to 5. So, I want a mathematical function and that mathematical function is called Probability Distribution. So, probability distribution is a mathematical function that will give us the probability of a specific value a random variable takes. Now, you must be wondering this is Data Analysis Course, and why we are talking about random variables, probability distributions. So, let me explain it bit. Suppose you are measuring the OD of a DNA solution.

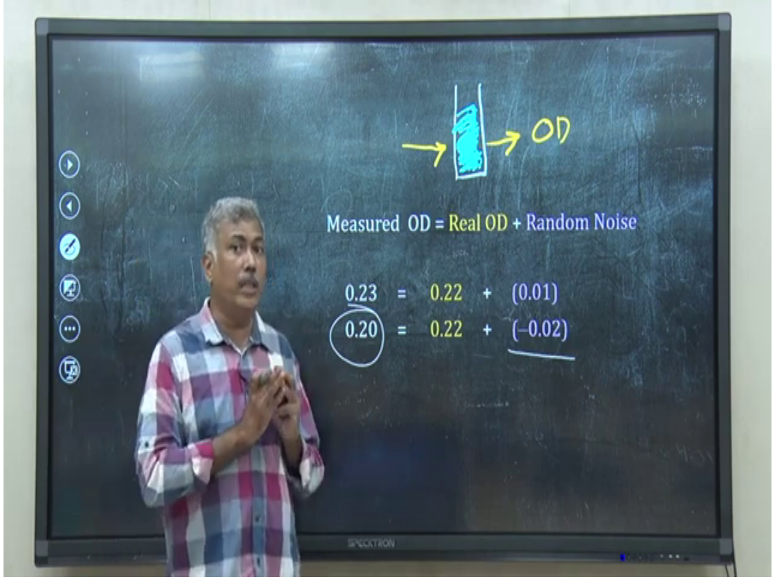
(Refer Slide Time: 03:29)





Measured OD = Real OD + Random Noise

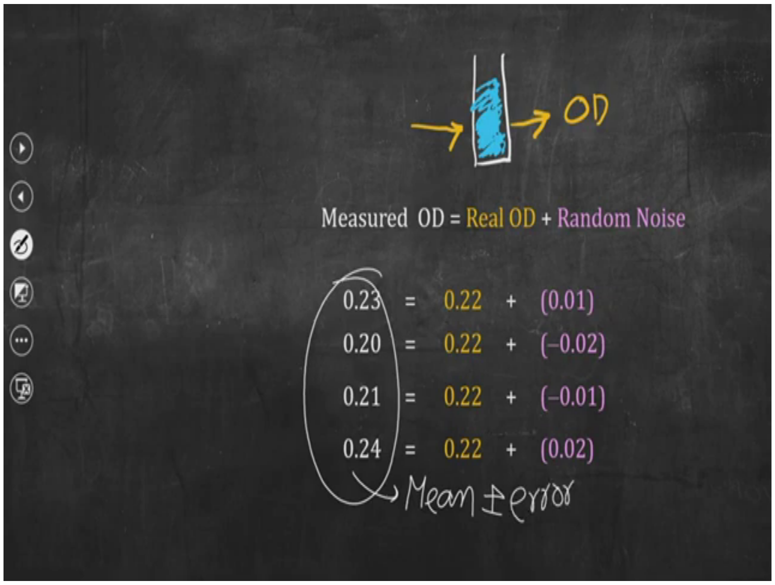
0.23 = 0.22 + (0.01)



Measured OD = Real OD + Random Noise

0.23 = 0.22 + (0.01)

0.20 = 0.22 + (−0.02)



Measured OD = Real OD + Random Noise

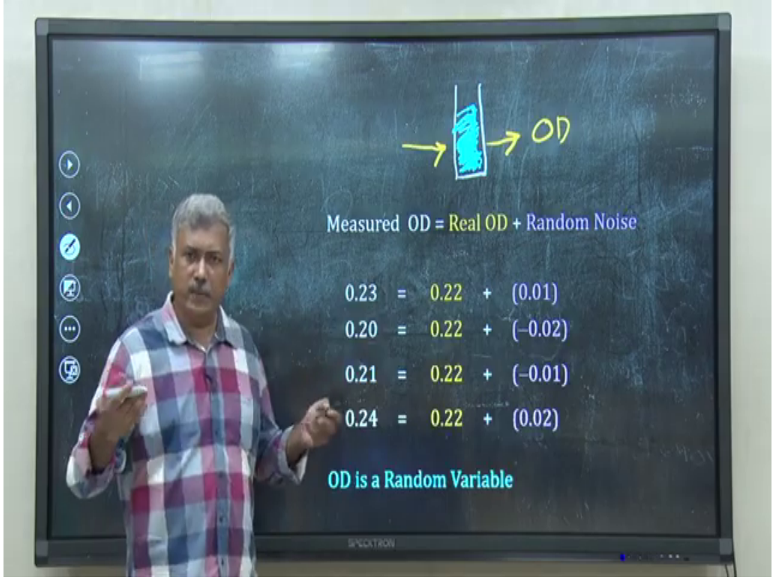
0.23 = 0.22 + (0.01)

0.20 = 0.22 + (−0.02)

0.21 = 0.22 + (−0.01)

0.24 = 0.22 + (0.02)

Mean ± error



Measured OD = Real OD + Random Noise

0.23 = 0.22 + (0.01)

0.20 = 0.22 + (−0.02)

0.21 = 0.22 + (−0.01)

0.24 = 0.22 + (0.02)

OD is a Random Variable

So, what you will do, you will take the DNA sample solution in a cuvette, shine with a light and then you will measure the OD and suppose the OD you have got is 0.23. This is the OD you have got suppose. Is this the correct measure of OD, optical density of this sample? Not really. Because we have to remember that every machine, every instrument that we use to measure something has some sort of noise.

So, what we are getting is not the exact result, not the exact OD, that is a real OD. So, what we are getting? We are getting something like this, we are getting the measured OD that I am measuring and reporting is nothing but, summation of the Real OD which I do not know, and a random noise.

So, a random noise get added when I make a measurement in any instrument. So, why do I call this noise random? Because first of all, I am quite uncertain about its value. I do not know it is most of the time do not know the origin of it and I cannot control it. So, we may assume some sort of distribution again probability distribution for this noise considering noise as a random variable.

But we do not, we are quite uncertain about its specific value. For example, suppose the right value of this, the real value of the OD for this particular sample is 0.22 suppose, although we do not know suppose we assume that this is the value. Then, when I make a measurement, suppose in the first measurement, the random noise was 0.01, then that get added to my real value.

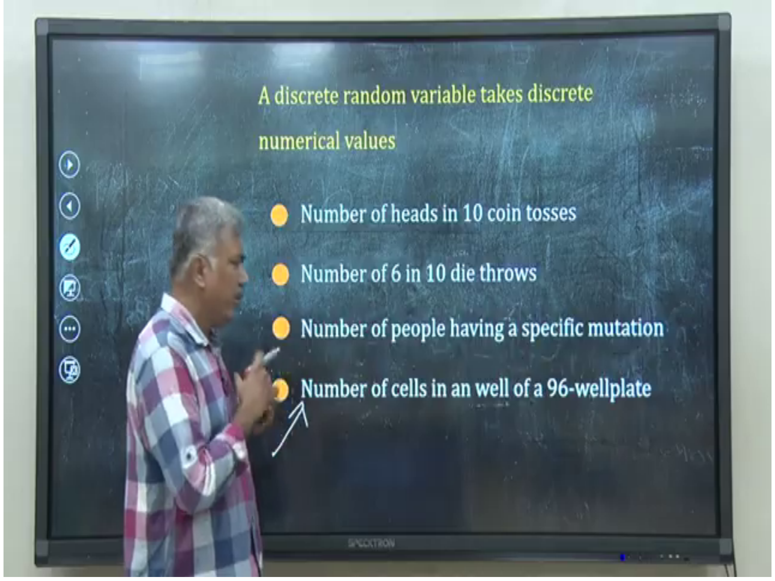
And what I get reported by the machine is 0.23. I repeat this experiment again. Now again, although the real OD should remain same, this value should remain same, because I am using the same sample, I am just repeating the experiment, but the noise value should change because it is the random variable. So, now suppose I have got a noise of minus 0.02.

So, then my reported value, my measured value, becomes 0.20. You can see this changing. So, what I am measuring is changing although the real value of the OD, the measured thing is constant, but the noise is varying, and it is varying randomly. So, I suppose repeated it two more times. So, I have four data. And that is what you do in experiment in reality in lab, so you measure 3 times, 4 times, 5 times for the same sample, and then what you do, you take this measurement and report the mean, you calculate the average of these.

And sometimes you may also add, the error term, which may be the standard deviation you have calculated from this data. So, you can see here that my optical density is actually a random variable, the measured optical density is a random variable, because it includes a random noise coming out of my measurement process.

So, all measurements that we do in science are actually, everything that we measure can be considered as a random variable. And that is why we have to understand their probability distribution. Now, in this case, if OD is a random variable, you can easily see OD can take any value bigger than 0, it can be 0.1 it can be 1.2. So, it is a continuous variable.

(Refer Slide Time: 06:53)



A discrete random variable takes discrete numerical values

- Number of heads in 10 coin tosses
- Number of 6 in 10 die throws
- Number of people having a specific mutation
- Number of cells in an well of a 96-wellplate

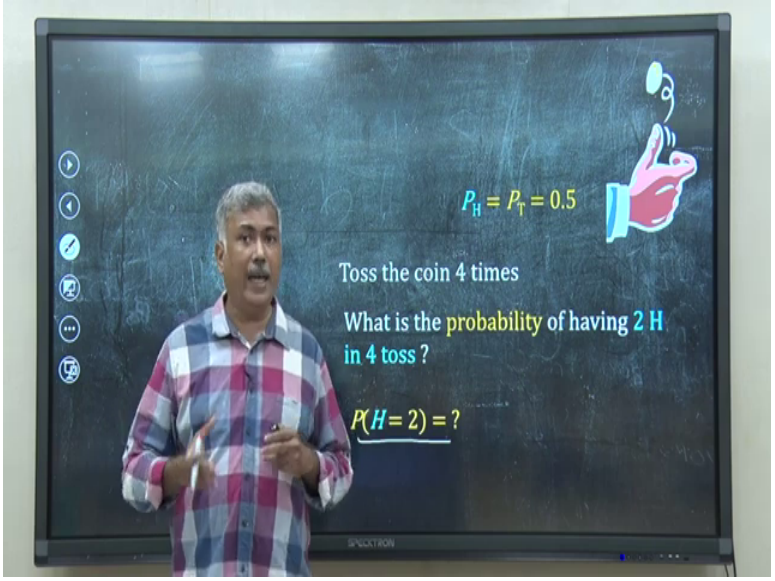
But there are some random variable which are discrete. And in this lecture, we will discuss about the probability distribution of those discrete random variable. Now, I have listed a few of the discrete random variables, check the last one. Here, suppose what you are doing, most of the biologists must be doing it regularly in lab.

Suppose I have some mammalian cell culture, I have made a cell suspension in such a way that if I take 100 microliters of the cell suspension, and then if I dispense it in a single well of a 96-wellplate, I will get roughly suppose 20 cells. Now, if I have 5 wells and if I dispense 100 microliter, one after another from the same sample. Do you expect that in every well, you will get exactly 20 cells? No.

Our common sense says no, in some well it may be 20, in some well it may be 25, in some other well, it may be 19, 10. Even it can be 0 also, because there are, the way we have diluted, we have only 20 cells in 100 microliters. So, you can see here the number of cells that you are seeding in this 96-wellplate is a random variable.

And this is a discrete random variable, because cell number cannot be 12.35 it can be either 12 or 13. So, this is a example of discrete random variable. And there are many of them. And we want to study the probability distribution, some of the commonest probability distribution that we know and very frequently use in data analysis, that we will discuss in this lecture.

(Refer Slide Time: 08:38)



$P_H = P_T = 0.5$

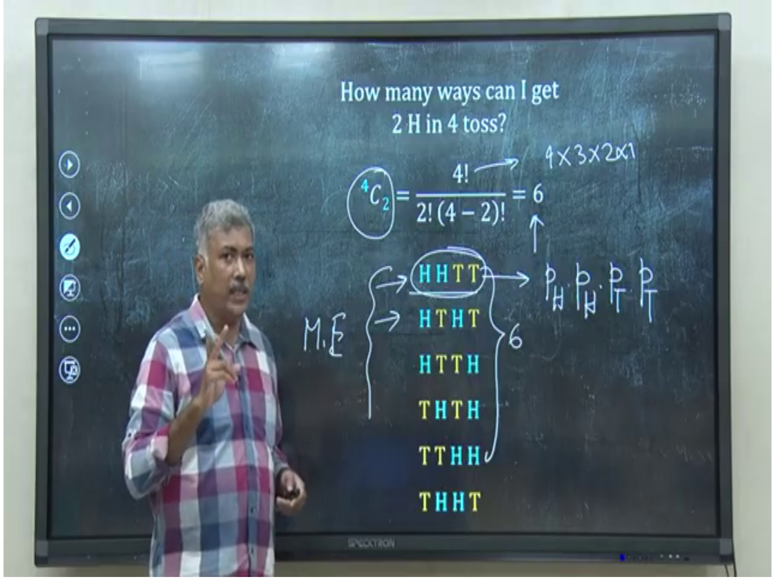
Toss the coin 4 times

What is the probability of having 2 H in 4 toss ?

$P(H = 2) = ?$

So, let us start with the first most common discrete probability distribution. Suppose I have a coin, and I have tossed it, how many times I have tossed it, I have tossed it 4 times. So, if I have 4 times, tossed it 4 times, and then if I ask you the question that what is the probability of having 2 heads out of these 4-coin tosses. So, I will represent it symbolically like this, I want you to calculate P(H = 2). Now, to answer this question, the first thing I have to do is to calculate how many ways I can get 2 heads out of 4-coin toss. So, let us check that.

(Refer Slide Time: 09:19)



How many ways can I get 2 H in 4 toss?

$${}^4C_2 = \frac{4! \rightarrow 4\times3\times2\times1}{2!\,(4-2)!} = 6$$

HHTT → $P_H \cdot P_H \cdot P_T \cdot P_T$

HTHT

HTTH

THTH

TTHH

THHT

I have listed the ways by which I can actually get 2 heads out of 4-coin tosses. Those are 6, HHTT, HTHT, these are sequence of coin toss. So, I have 6 of them. You can explicitly write them the way I have written here, or you can use the concept of combinatorics. And from that you can say, the number of ways I can get 2 heads out of 4-coin tosses would be 4 choose 2.

Which is equal to,

$$= \frac{4!}{2!\,.\,(4-2)!}$$

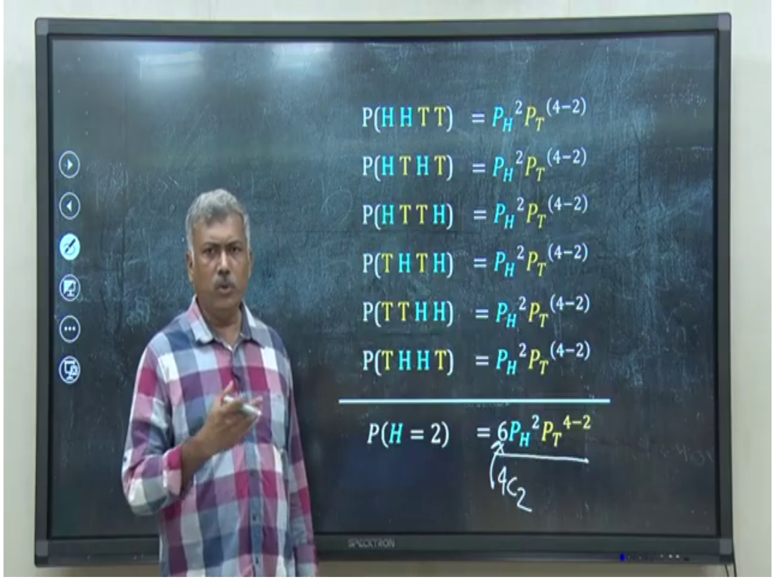
$$= \frac{4!}{2!\,.\,2!}$$

So, what is factorial 4, factorial 4 is nothing but 4 into 3 into 2 into 1. All factorials we get it by multiplying this number sequentially up to 1. So, this is equal to 6. So, 6 way I can get 2 heads out of 4-coin tosses.

Now, remember these are all mutually exclusive event, these are all mutually exclusive event, because if you have got this HHTT, you cannot get the other one. And also, if you remember, we have discussed about independent events in the previous lecture, each of these coins toss is independent.

Because what will happen in my second coin toss is not decided by what has happened in my first coin toss. So, let me use the idea of independent event to calculate the probability of this one, HHTT. So, probability of getting head is suppose P H. In the first one it will be P H, the second one is also probability of head, multiplied by probability of getting tail, into multiplied by probability of tail. So, I have used the product rule that we have discussed in the last lecture. In this way, I can get the probability of each of these 6 mutually exclusive events where in all cases, I have 2 heads out of 4-coin toss.

(Refer Slide Time: 11:30)



$$P(HHTT) = P_H^2 P_T^{(4-2)}$$

$$P(HTHT) = P_H^2 P_T^{(4-2)}$$

$$P(HTTH) = P_H^2 P_T^{(4-2)}$$

$$P(THTH) = P_H^2 P_T^{(4-2)}$$

$$P(TTHH) = P_H^2 P_T^{(4-2)}$$

$$P(THHT) = P_H^2 P_T^{(4-2)}$$

$$P(H = 2) = \underbrace{6P_H^2 P_T^{4-2}}$$

$$({}^4C_2$$

So, that is what it will look like. I have written them in a reduced form so, the first one is PH square into PT square also because 4 minus 2 is again square. So, now what is the total probability that if I toss the coin 4 times, I will get the head 2 times. Now, these are all mutually exclusive event and in either of this way, I will get 2 heads out of 4-coin tosses.

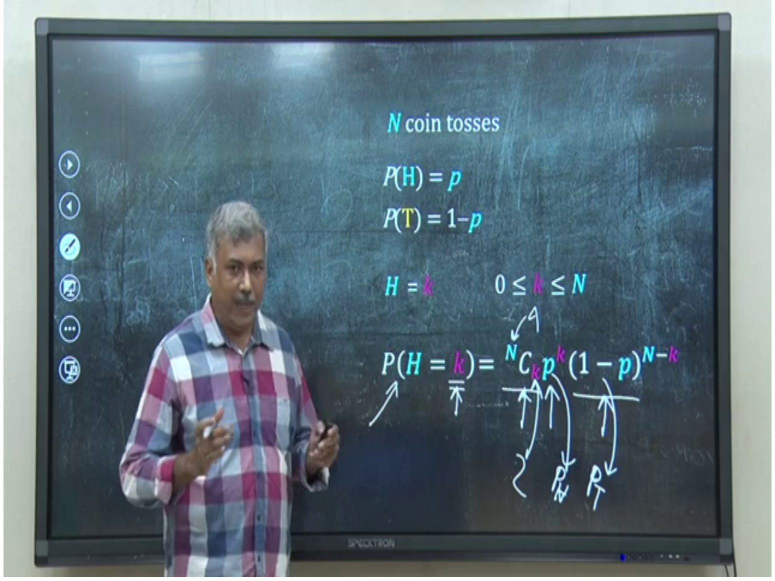
So, now, I will use the ADDITION RULE that again we discussed in the last lecture. So, I will add this all probability and that will give me,

$$= 6.\, PH^2 \,.\, PT^{4-2}$$

$$= 6.\, PH^2 \,.\, PT^2$$

So, what is this 6, where this 6 coming from. Because I have 6 mutually exclusive event and it can be also said this is 4C2. So, now, let me generalize this rather than saying 4, 2, something like that.

(Refer Slide Time: 12:31)



Let us consider like this. I have done N coin tosses, probability of getting head is P. So, the probability of getting tail is (1 – P) because summation of these two probabilities will be equal to 1. And I want to know what is the probability that I will get k number of heads out of the N coin tosses.

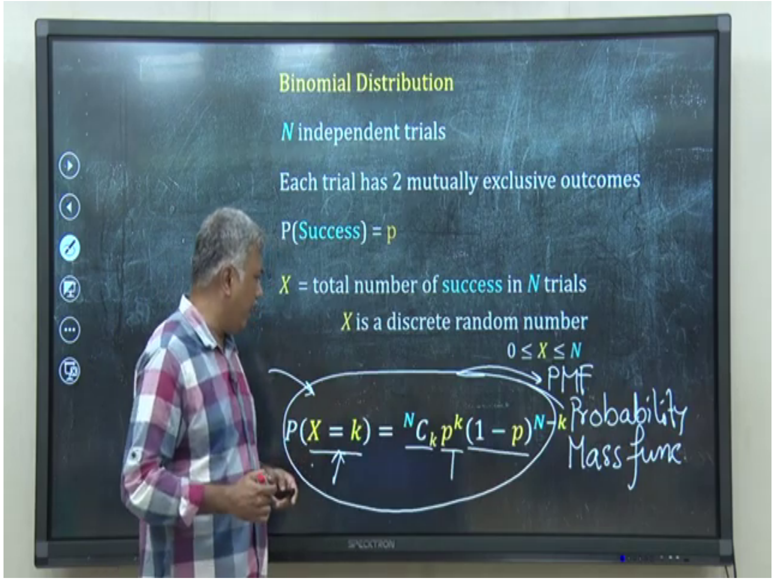
So, that will be using, if I use the way we have calculated, the method that we have used earlier for specific value, that 2 heads out of 4 coin toss I will get, it will be equal to, this probability will be equal to,

$$^{N}C_{k} \,.\, P^{k} \,.\, (1 - P)^{N-k}$$

You can easily identify the similarity, what we just derived one slide back.

So, in the earlier example, N was 4, k was 2 and this P was, we were writing as probability of head and this one is nothing but probability of tail. So, let us move further. Rather than just keeping ourselves bounded by coin toss let us make it generalized.

(Refer Slide Time: 13:50)



Binomial Distribution

N independent trials

Each trial has 2 mutually exclusive outcomes

P(Success) = p

X = total number of success in N trials

X is a discrete random number

$0 \leq X \leq N$

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

So, suppose if I have N independent trials and in each trial there are two mutually exclusive outcome, either you will get a head or a tail, either you will get a 0 or 1, either it is true or false, success failure, either it is mutated not mutated, either the cell died or the cell is not died, something like that.

If I have this type of problem, and I have done N independent trials, and then suppose the probability of success, probability of death, probability of getting head, probability of getting 1, something like that is equal to P. Then suppose X is a random variable and that represent total number of success in N trials. So, X is a discrete random number.

So, the probability that X will take a particular value k will be equal to,

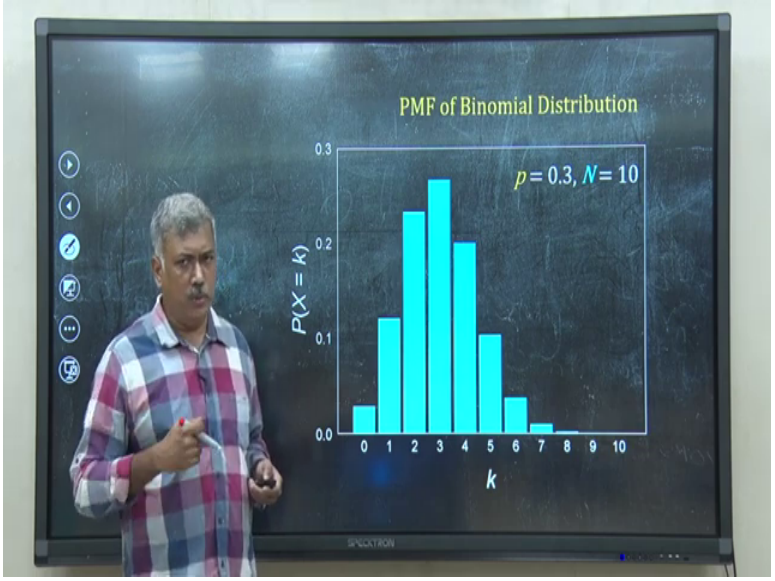
$${}^{N}C_k \,.\, P^k \,.\, (1 - P)^{N-k}$$

So, I have generalized what we have learned from our coin toss problem. In our coin toss problem also we had the same relationship, here also we have the same one.

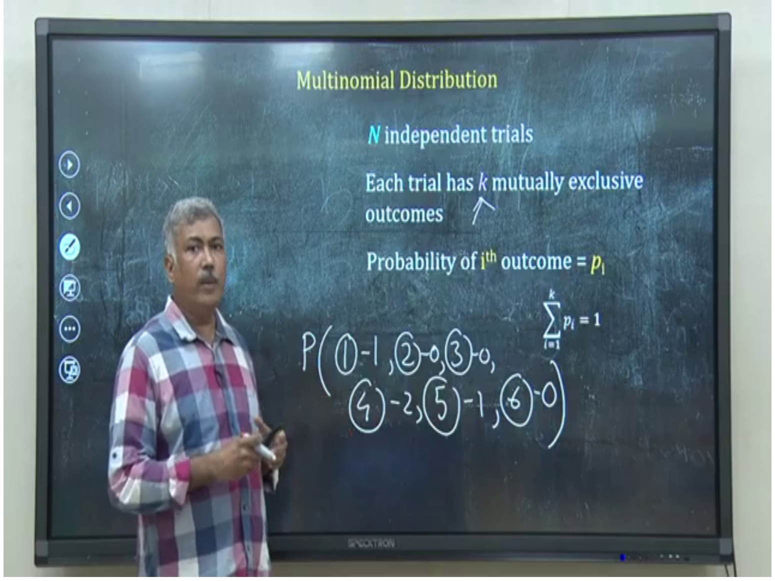
Because coin toss is exactly the same problem. I have N independent trial, in each case I have two possible mutually exclusive outcome and a probability of success, that is probability of head is given to me. Then the probability that X will take a particular value is equal to N choose k, p to the power k into 1 minus p to the power N minus k, this is called the PMF of binomial distribution. What is PMF? PMF is Probability Mass Function. This equation is probability mass function of binomial distribution, which is very frequently used in many calculations.

Now, remember this gives me the probability of taking a particular value by this particular random number X. Now, coin toss has two options. I will get head or tail, or if you are studying mutation either it is mutated or not mutated, but always it will not be like that. So, there are some other cases also. But before I go into those other cases, let me represent this PMF, probability mass function graphically. So, what I have done, I have taken a specific value of N and specific value of p and then I plotted a diagram.

(Refer Slide Time: 16:39)



PMF of Binomial Distribution



Multinomial Distribution

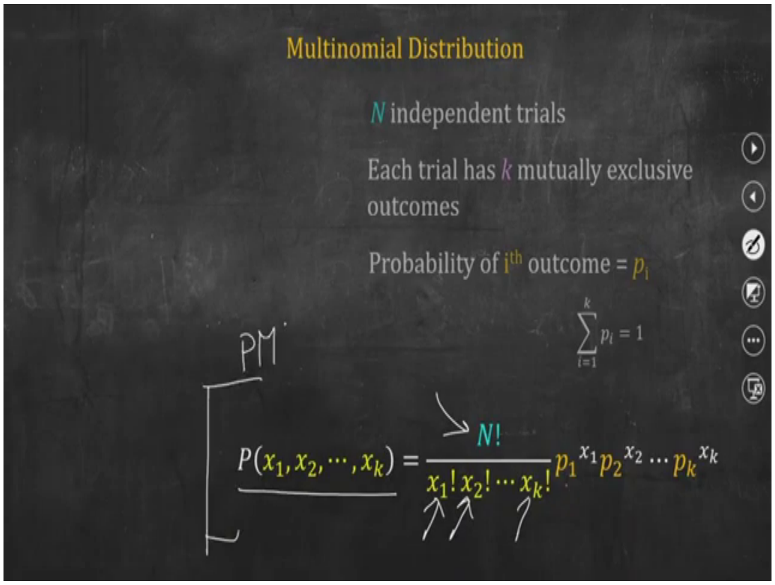
N independent trials

Each trial has k mutually exclusive outcomes

Probability of i^th outcome = p_i

$$\sum_{i=1}^{k} p_i = 1$$

$$P(\textcircled{1}-1, \textcircled{2}-0, \textcircled{3}-0, \textcircled{4}-2, \textcircled{5}-1, \textcircled{6}-0)$$



Multinomial Distribution

N independent trials

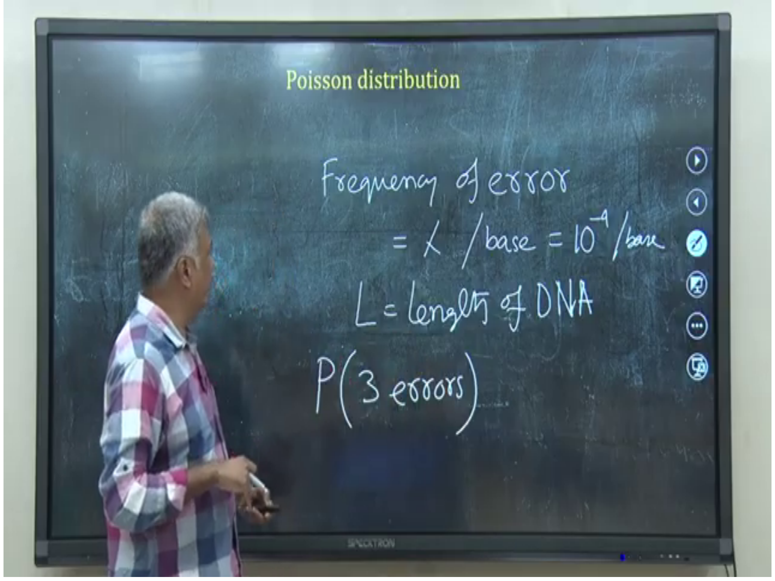
Each trial has k mutually exclusive outcomes

Probability of i^th outcome = p_i

$$\sum_{i=1}^{k} p_i = 1$$

PM

$$\underline{P(x_1, x_2, \cdots, x_k)} = \frac{N!}{x_1!\,x_2!\cdots x_k!}\, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$



Poisson distribution

Frequency of error

$= \lambda$ /base $= 10^{-4}$/base

L = length of DNA

P(3 errors)

So, it looks like this. So, for N = 10 and P = 0.3, that is probability of success, that is probability of head or something like that is 0.3. Then for each k, I have calculated the probability using the relationship the PMF that we have discussed just now for binomial distribution and it looks like this graph.

This looks like a bell curve, inverted curve, it looks like what you commonly say Normal or Gaussian, but remember this is not that. This is the Probability Mass Function PMF of binomial distribution, and in this case, k is discrete value. So, the random variable that we are dealing with is a discrete variable.

Now, let us move into the next most commonly used discrete probability distribution that is multinomial distribution. What is that? Suppose, I have N independent trial and in this case, each trial does not have 2 outcomes, it has more than 2 outcomes and suppose, I represent that by k, k mutually exclusive outcomes and each of these has a associated probability P i.

So, P 1 P 2 something like that, for example, if you take the die, in this die, each face has a probability to appear if I throw. So, those probabilities are this P and summation of all this probability must be equal to 1. Now, if I throw this die, suppose 4 time, and I ask tell me what will be the probability that I will get 1, 1 times, face 2, 0 times, face 3, 0 time and face 4, 2 time, face 5, 1 time and face 6, 0 time. I have thrown the dice 4 times. So, obviously, I will not get all the faces. So, I am asking you that, I have thrown it 4 times.

And I want to know the probability of getting 1 coming up once, 4 coming up twice, and 5 coming up once. 2, 3 and 6 does not appear in this four throws. So, I want to calculate the probability of that. So, if I have to calculate the probability of that I have to use something called the PMF of multinomial distribution, and it looks almost similar to the binomial one but only in this case, I have multiple other terms because number of outcomes is more than 2.

So, the probability of a particular event, that is particular result that you have got out of 4 or 5 or 6 or N number of dice throw, N number of trial is equal to,

$$\frac{N!}{x1!\,.\;x2!\,......\;xk!}\;P1^{x1}P2^{x2}.....\,Pk^{xk}$$

Where,

Factorial of x1 is factorial the number of times that has appeared

P1 is the probability of getting the first face for example in case of die throw, the first outcome to the power x 1 that is the number of times that outcome has happened.

So, this is the PMF of multinomial distribution, probability mass function of multinomial distribution. Now, let us move into the third and the one of the very common discrete probability distribution that we deal usually in data analysis. That is called Poisson distribution. What is that? Before I go into it, let us take an example.

Suppose, I am doing DNA sequencing, some method, we do not we will not go into that some method you are doing DNA sequencing, and suppose obviously, any method will have some sort of error in that, mistakes in that so, suppose the probability or the frequency of error, the frequency of error.

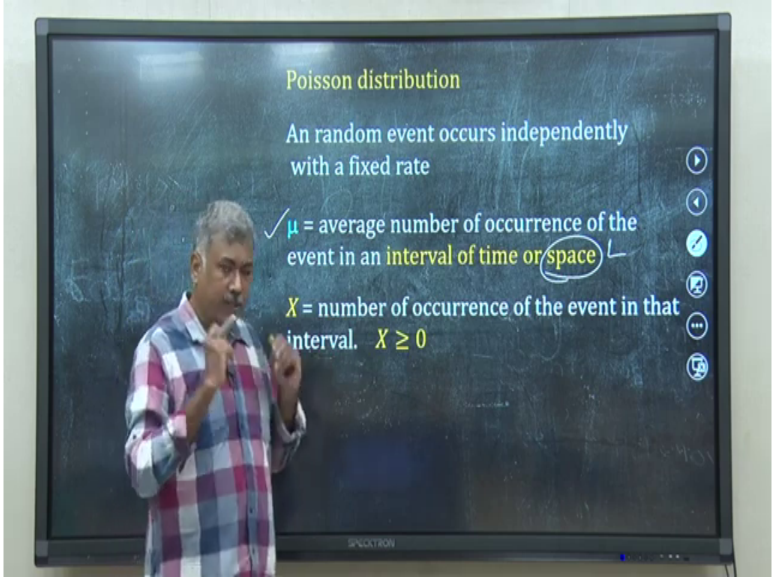
Suppose, let me write something, frequency of error in my DNA sequencing may be suppose something like lambda number of times per base. And suppose that is very low, suppose that is something like 10^{-4} or 10^{-5} something per base suppose. So that means, on an average if I sequence 10^{4} bases, I will get 1 error. So, that is my frequency of error and suppose, I have taken a DNA sequence of length L, that is the length of the DNA right, length of my sample DNA sequence DNA and I have sequenced it.

So, now, I want you to calculate the probability that if I sequence this DNA of this length L which is very big. L is very big suppose, and what is the probability that there will be 3 errors. I do not bother about where the errors are, somewhere in this length L the errors are there. What is the probability that 3 errors will come? Now, remember number of errors are discrete random variable, it can be 0 1 2 something, up to the length L was possible.

So, it is not a continuous variable, you cannot have an error of 2.3. So, I want you to calculate the probability that 3 errors will appear in this length L, where lambda is the frequency of error which is a very small value and length of the DNA is L which is a big value. To answer this question, we have to use the probability mass function of Poisson distribution and poisson distribution is a special case of binomial distribution.

Where the probability of success is very low and number of trial is very big, just like this one, the error frequency of error is low, but the length of the DNA is much big. So, what is poisson distribution, let us look into that.

(Refer Slide Time: 23:16)

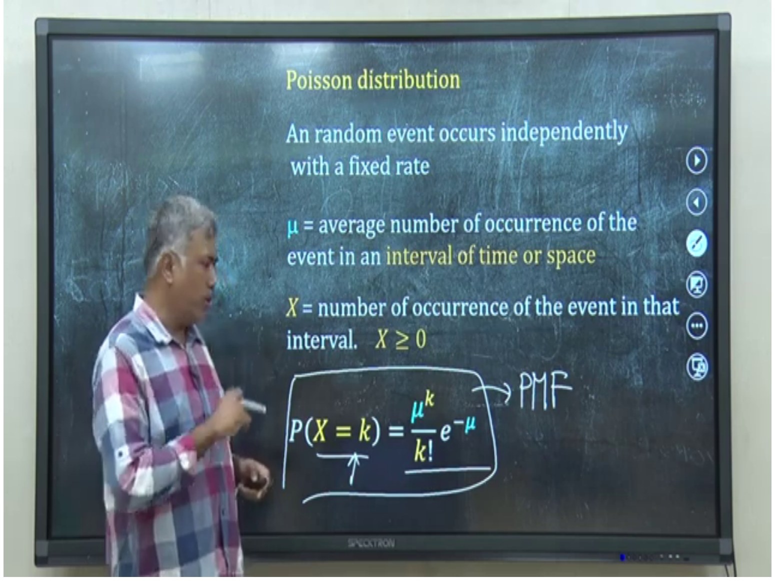


Poisson distribution

An random event occurs independently with a fixed rate

✓ μ = average number of occurrence of the event in an interval of time or space

X = number of occurrence of the event in that interval. $X \geq 0$



Poisson distribution

An random event occurs independently with a fixed rate

μ = average number of occurrence of the event in an interval of time or space

X = number of occurrence of the event in that interval. $X \geq 0$

$$\boxed{P(X = k) = \frac{\mu^k}{k!} e^{-\mu}} \rightarrow \text{PMF}$$

In poisson distribution we are dealing with a random event that occurs independently repeatedly with a fixed rate, that rate is fixed and we have a parameter we call it μ. μ is the average number of occurrence of that event in a particular interval of time or space. What do I mean by space? Space could be in a length for example, in case the example that I discussed just now, length of the DNA is the space. It can be volume, suppose you are measuring some something number of cells in a particular volume that you have in media that you have dispensed. So, that can be volume or suppose some cases it can be area also.

And why are bothered about time because, suppose, we are trying to calculate the probability of some death of bacteria in presence of a particular antibiotic. So, in that case, the we can imagine each of the bacteria die with a, independently with a fixed rate of death and μ is the average number of death in a particular time interval.

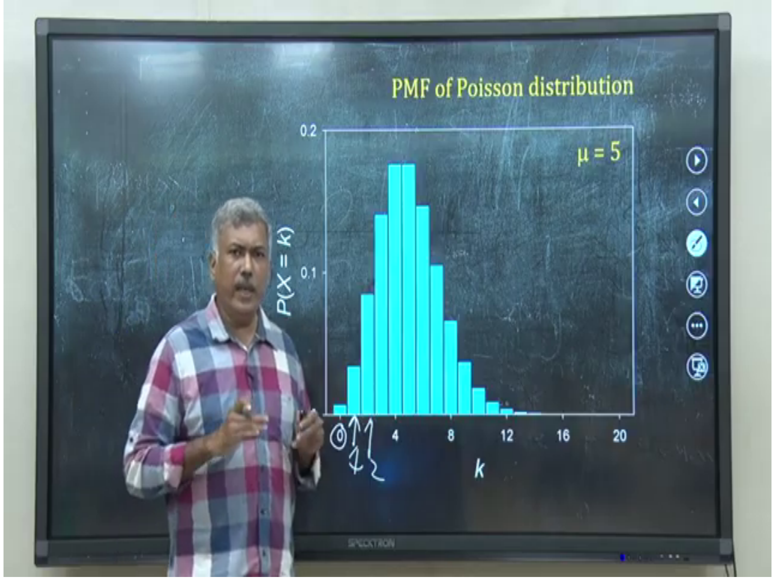
So, what I have, I have a parameter μ, which is average number of occurrence of that random event in an interval of time or space. And then X is a random variable for this particular distribution, X is the number of occurrence of the event in that interval, that interval of time or space and obviously, X is bigger equal to 0.

So, now poisson distribution PMF give me the PMF, the probability that X will take a particular value k and that is given by this relationship,

$$= \frac{\mu^{k}}{k!} \cdot e^{-\mu}$$

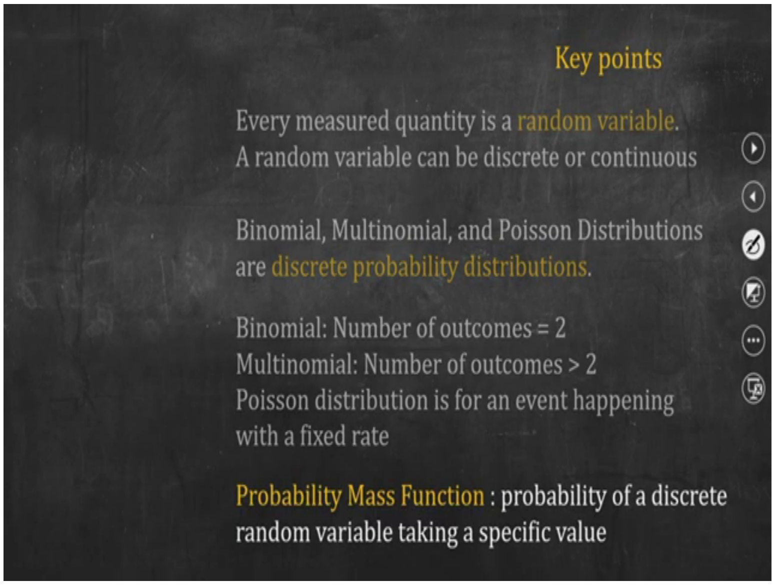
So, this is called the PMF of poisson distribution. So, now let me draw this PMF, this PMF is the PMF for poisson distribution. So, just like I have drawn the PMF for binomial distribution, let me draw the PMF for poisson distribution.

(Refer Slide Time: 25:27)


PMF of Poisson distribution

And here it is, I have taken a particular value for μ, μ equal to 5, and then again, I have plotted each value each probability, P(0), P(1), P(2) using the PMF poisson distribution. And you can see that this histogram looks just like that a binomial distribution. Again, this is a bell-shaped curve, but remember, this is not normal distribution, this is poisson distribution, and the diagram is for the PMF of poisson distribution. So, with this we will end our lecture today. In this lecture, we have learned about three probability distribution, which are very common discrete probability distribution.

(Refer Slide Time: 26:08)



Key points

Every measured quantity is a random variable.
A random variable can be discrete or continuous

Binomial, Multinomial, and Poisson Distributions are discrete probability distributions.

Binomial: Number of outcomes = 2
Multinomial: Number of outcomes > 2
Poisson distribution is for an event happening with a fixed rate

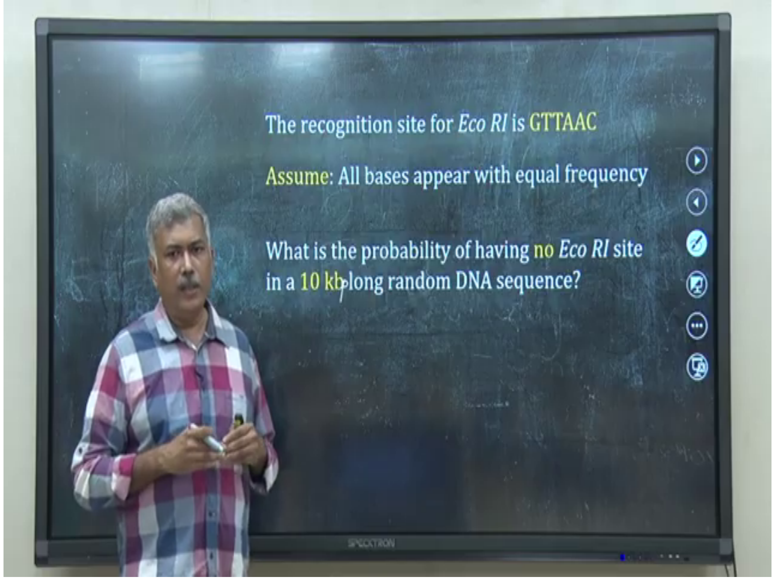
Probability Mass Function : probability of a discrete random variable taking a specific value

So, what we have learned here is that every measurement, every measurement that we do, every data that we collect is a random variable, and a random variable can be discrete or continuous. Now, in this lecture, we have discussed about binomial, multinomial, and poisson distribution, all these three are discrete probability distribution.

If your number of outcomes in a trial is 2, then it is actually binomial. If the number of outcomes is bigger than 2, then we have to use the multinomial PMF. And if we are dealing with something where we have some event, random event, a particular random event, random event happening independently repeatedly with a fixed rate, and we know the average number of times that will happen in a particular interval of time and space.

Then we can use a poisson distribution to calculate the probability of a particular number of times that event will happen in that interval. And we have to remember that probability mass function, each of these three distributions that we have shown, is nothing but the probability of a discrete random variable taking a specific value. That is all for this video, but before I leave, let me give you a problem to solve.

(Refer Slide Time: 27:26)



The recognition site for *Eco RI* is GTTAAC

Assume: All bases appear with equal frequency

What is the probability of having no *Eco RI* site in a 10 kb long random DNA sequence?

Suppose, we are analyzing a DNA sequence and I have written the recognition site for ECoR1 GTTAAC and we assume that in a random DNA sequence given to you all the four bases appear with equal frequency. Now, suppose I have given you a random DNA sequence of 10 kb, the length is 10 kb. Sorry, it will be kbp.

So, suppose I have given you a random DNA sequence of 10 kilo base pair, then what will be the probability that this random sequence of 10kbp does not have any EcoR1 site. You have to calculate this probability, and using the concept of PMF of three different discrete probability distribution that we have discussed, you should be able to answer this question. So, try this one. till then happy learning. Thank you.