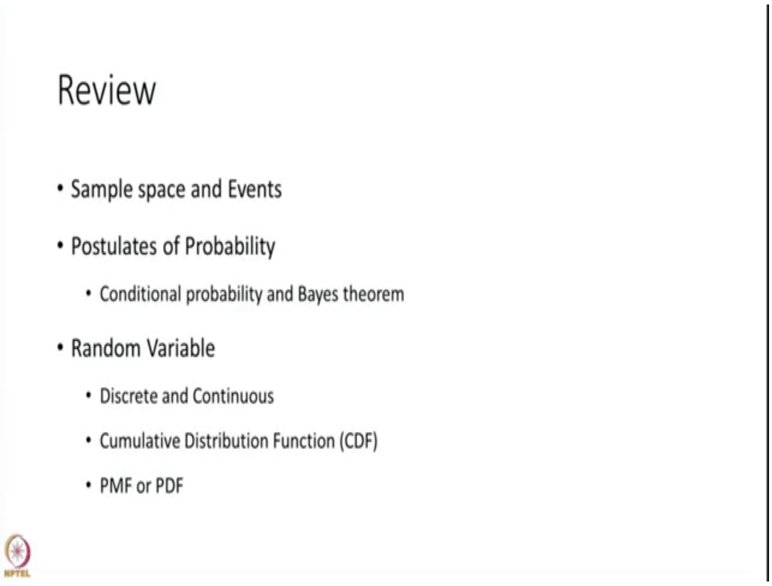
Dealing with Materials Data
Collection, Analysis and Interpretation
Professor. Hina A Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology Bombay
Lecture 31
Special Random Variables I

Hello and welcome to the course on Dealing with Materials Data. In today's session we are going to talk about certain special random variables.

(Refer slide Time: 00:33)



Let us, review what we have done in the past we introduced formally the definition of probability through the definition of sample space and events. We said that any random experiments the collection of all its possible outcomes is called a sample space and any subset of a sample space is called an event.

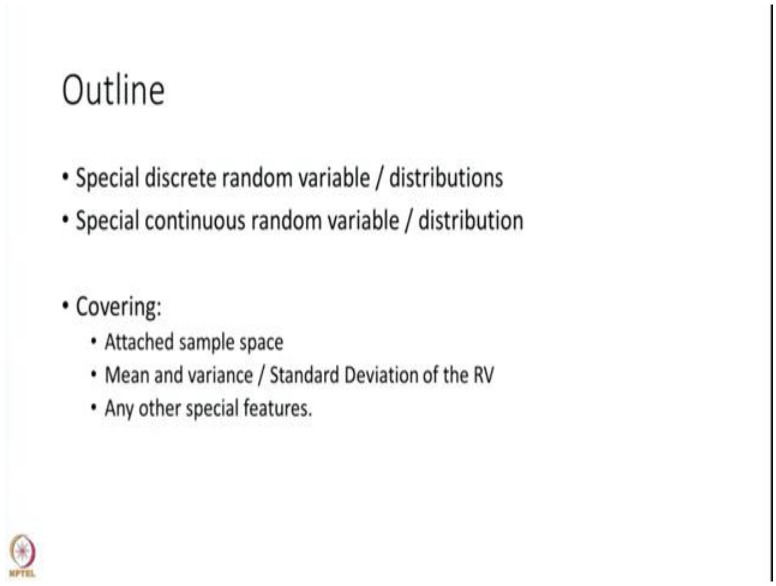
Then we introduced three postulates of probability which basically says that for any event probability is positive for the complete event that is for sample space probability is one and if there are mutually exclusive events finite or infinite of them. The probability of union of all of them is the sum of all the probability of sum of probability of all the events.

Then, we also introduced what is known as conditional probability that instead of sample space if you have a knowledge of one particular event and in the light of having the knowledge of that event if you look at probability of other events it is called conditional probability and from that we introduced base theorem and we briefly talked about the importance of base theorem in today's life.

Then, we went further and we also introduced what is known as random variable which is a function a real function defined from any probability space, which is a sample space and the probability measure from probability space to a real line and then we said that it could discrete or continuous depending on kind of values it takes. If he takes a countable values finite or infinite it is called discrete and if he takes a different kinds of a continuous values continuous numbers it is called continuous random variable.

With every random variable we introduced that is cumulative distribution function attached which is probability of that random variable taking a value given than a specified value x. Now, depending on the discrete a random variable you have a probability mass function attached to it and if it is continuous random variable you have probability density function attached to it.

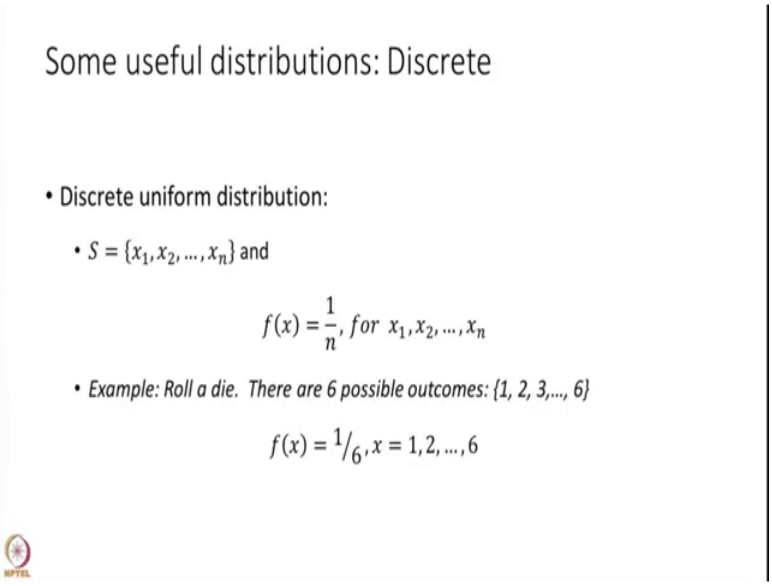
(Refer slide Time: 03:08)



This time what we want to do is, we want to introduce certain special discrete random variables or discrete distributions which we come across more frequently in our studies of material data. Similarly, we would there will be some distribution, some random variables which are discrete, so will have discrete distributions some random variables will be continuous we will have some continuous distribution. While covering this, what we would like to do is we would like to see very briefly what is the attached samples space very important to know where it is define you must know.

So, we defined a sample space we give mean and variance formulae for the given distribution. We, also gives standard deviation because it this are the two very basic parameters that one would like to know to describe the data. And, if there any other special features we will talk about it as the time comes, wherever possible we will give an example or we will discuss those examples.

(Refer slide Time: 04:31)



So, let us move on, the first distribution that we would like to talk about is the first discrete uniform distribution. This discrete uniform distribution is a simplest of the distribution so if you rolling a die. Then we all know that there are six possible outcomes and if die is fair then probability of any outcome occurring that is whether the it will turn a face 2 face 3 or face 4 whatever it has a probability of 1 over 6. So, this also comes from the postulate the total number of say if you are looking for probability that the face the die faces turns up 2.Then the event can occur 2 is only 1 the number of times it comes is 1 and the total number of such a faces can turn up are 6 so the probability is 1 over 6 is the example we have given here.

Discrete uniform distribution:

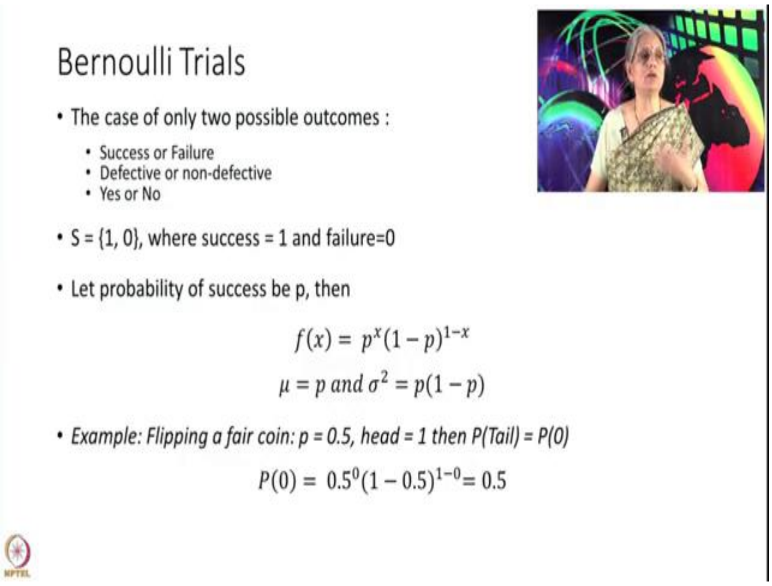
$S = \{x_1,\ x_2,\ \ldots,\ x_n\}$ and

$$f(x) = \frac{1}{n},\ for\ x_1,\ x_2,\ \ldots,\ x_n$$

Example: Roll a die. There are 6 possible outcomes: {1, 2, 3,…, 6}

$$f(x) = {}^{1}/_{6},\ x = 1,\ 2,\ \ldots,\ 6$$

(Refer slide Time: 05:50)



The next we would like to talk about is Bernoulli trails this is one of the important trails. Is any experiment is called a Bernoulli trails if it lends up with exactly two outcomes two possible outcomes it could be success or failure. It could defective or non-defective or could be answer yes or no.So, whenever you have an experiment which throws up only two answers it is called a Bernoulli trail. So, if you are tossing a coin it is also a Bernoulli trail and if you are talking rolling a die but saying that your outcome is success if you get number between 1 and 3 and if, you it is a failure if you get number 4, 5, 6 than also it is a Bernoulli trials.So, in that case the sample space is only 0 and 1 because there are only two outcomes, generally we say success is 1 and failure is 0 but this is a convention.

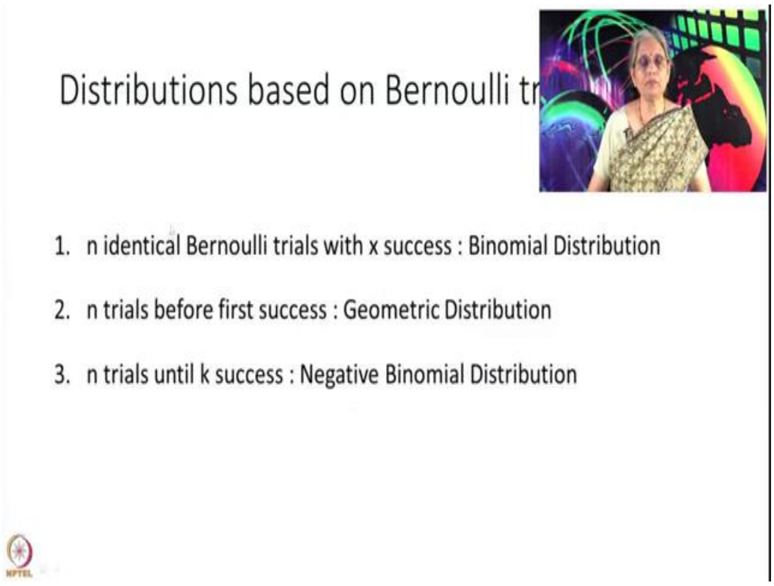
Let probability of success be p, then

$$f(x) = p^x(1-p)^{1-x}$$

$$\mu = p \ and \ \sigma^2 = p(1-p)$$

I am going to leave the proof to the students to try it out for them self. Here is a very simple example of tossing a coin.

(Refer slide Time: 7:46)



Next, there are distributions, which come out of Bernoulli trails. How you conduct the Bernoulli trails? How you conduct your experiment counting the Bernoulli trails give you give rise to different distributions. So the first one, we would like to consider binomial distribution. In which you throw N identical Bernoulli trails. With an you count you want to know the probability that there will be exactly X successes.

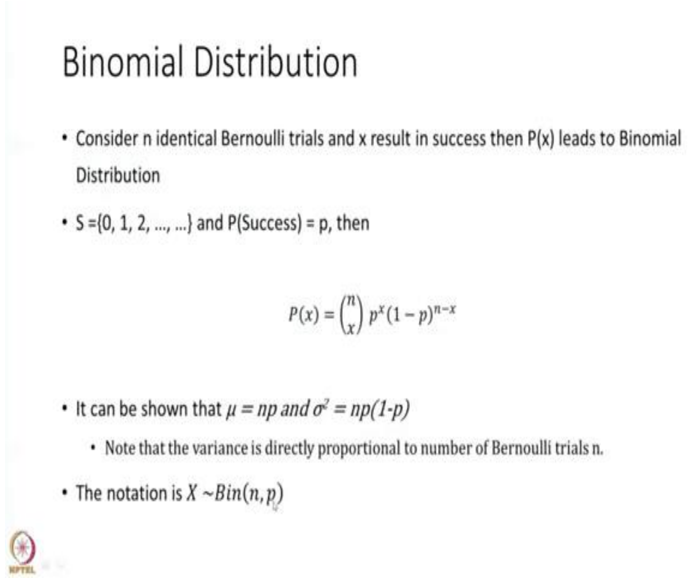
Another way to look at is that you throw n trails before you encounter the first success this is called a geometric distribution. And then you have you have n Bernoulli trails until you get K successes here there is spelling mistake it should be K successes and this is called a negative binomial. Why the adjective negative and why still binomial we will look when we look in to the distribution.

S ={0, 1, 2, …,n} and P(Success) = p, then

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\mu = np \text{ and } \sigma^2 = np(1-p)$$

(Refer slide Time: 8:56)



So, first let us go to binomial distribution, consider n identical Bernoulli trails you have conducted and exactly x have resulted in a success. The probability of success is p then your sample space is there can be no success or there can be exactly n successes here it should be it should not have dotted it should be n please correct it.

Probability of success is p then the probability mass function p of x is out of n choose out of n you choose x successes because it can occur anywhere. p to the power x multiplied 1 minus p to the power n minus x please recall p to the power x tells you the probability of successes in x x successful trails.

This, is probability of n minus x unsuccessful trails and this is the different ways the success can occur in a sequence of n trails. I will again leave it to you it can been shown that the mean value for such a p m f for such a distribution is in p and the variance is n times p times 1 minus p sometime 1 minus p is also denoted by q. So, it is also called sigma square is n p q. Few thing we need to notice number one Why it is called a binomial distribution?

Because these are binomial coefficient please recall, your expansion of a plus b to the power n then this is a binomial coefficient and there for this is called binomial distribution.

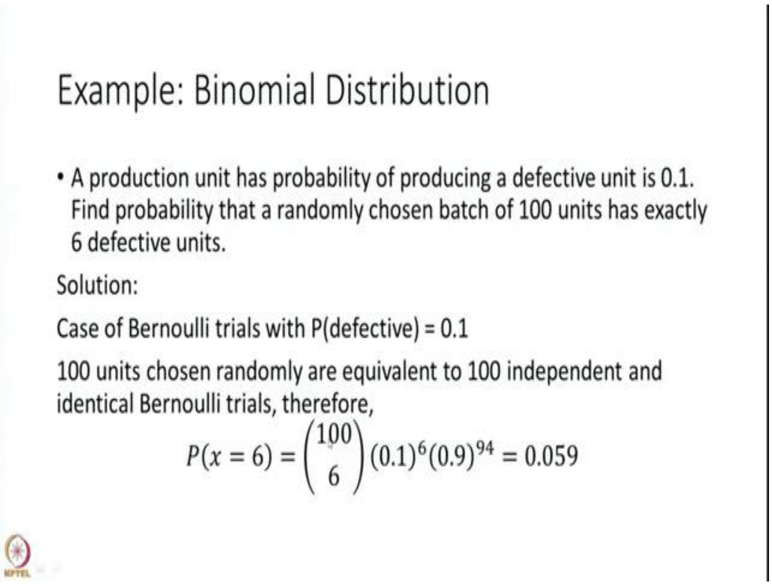
$$\mu = np \text{ and } \sigma^2 = np(1-p)$$

It very clearly says that whenever you see that the variance is directly proportional to the mean value you have a chance that you actually playing with a Bernoulli trails.

So, if you coming up with some data in which you see this this is a property very specific to the Bernoulli sorry n Bernoulli trails or binomial distribution. The notation is when we say that random variable x is following binomial distribution with two parameter n and p

The notation is $X \sim Bin(n, p)$

(Refer slide Time: 11:37)

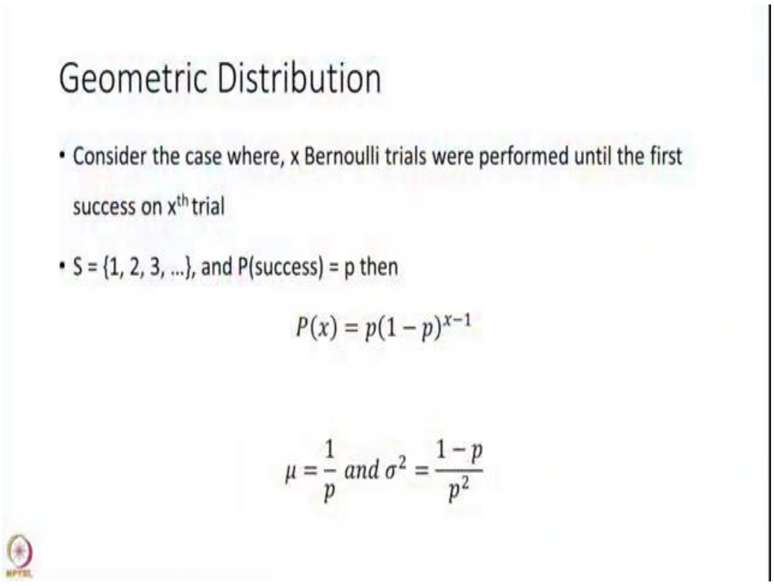


Here, is an example we will work out, suppose a production unit has a probability of producing a defective unit is 0.1 and a batch of 100 unit is chosen randomly. What is a probability that exactly six defective unit will be found? So, this becomes a case of Bernoulli trail with a probability of defective 1. You have taken 100 independent Bernoulli trails and then once you have the 100 independent and identical Bernoulli trails.

$$P(x = 6) = \binom{100}{6}(0.1)^6(0.9)^{94} = 0.059$$

So, there please remember the probability it is 0.059. Because we are going to work out the same problem using Poisson distribution by taking approximation. But that will be in future. Let us, continue with some of the Bernoulli trails.

(Refer slide Time: 13:02)



Now we go to next distribution which we called a geometric distribution. What we are going to do there? We consider x Bernoulli trails are performed until the first success has happened on x^{th} trail. So, it means that the first success can happen on the first trail or second trail or third trail or any trail in future.

So, S is an infinite set probability of success again is p as in Bernoulli trails and in that case the probability under geometric distribution probability of x is actually.

Consider the case where, x Bernoulli trials were performed until the first success on x^{th} trial

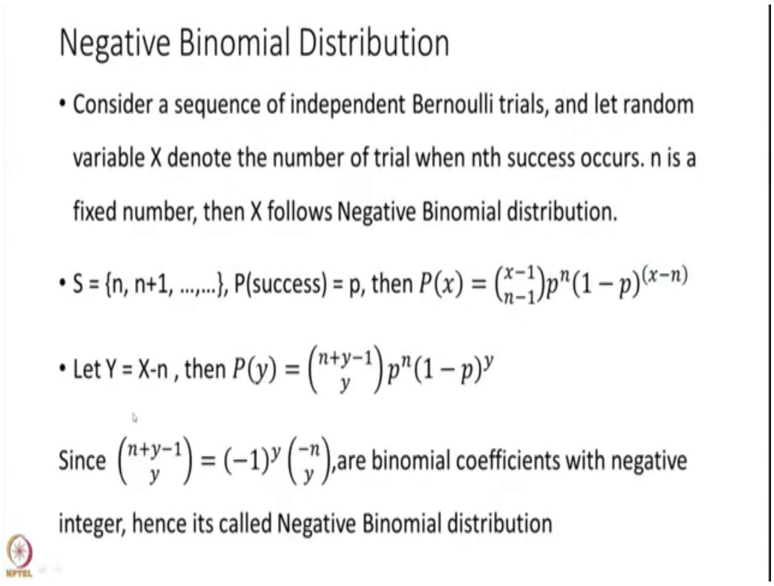
S = {1, 2, 3, …}, and P(success) = p then

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

$$\mu = \frac{1}{p} \; and \; \sigma^2 = \frac{1-p}{p^2}$$

this also you are welcome to work out by yourself.

(Refer Slide Time: 13:59)



Now, we come to negative Binomial trails or negative Binomial distribution. So, now we are considering a sequence of independent Bernoulli trails and the random variable x will denote the number of trails when the nth success will occur. Please remember in geometric distribution we said on the xth trail the first success occurs here we say that on xth trail, nth success will occur n is a fixed number. So write in the, like we said that first trail first success occurs nth is a fixed number is a nth trail which occurs. Then, we say that x follows a negative binomial distribution, therefore

S = {n, n+1, ……}, P(success) = p, then

$$P(x) = \binom{x-1}{n-1} p^n (1-p)^{(x-n)}$$

Why it is called negative binomial is reasoned out here.Instead of looking at the successful trail you look at the unsuccessful trails then the total number of unsuccessful trails have occur when you have taken x trails is y

Let Y = X-n , then $P(y) = \binom{n+y-1}{y} p^n (1-p)^y$

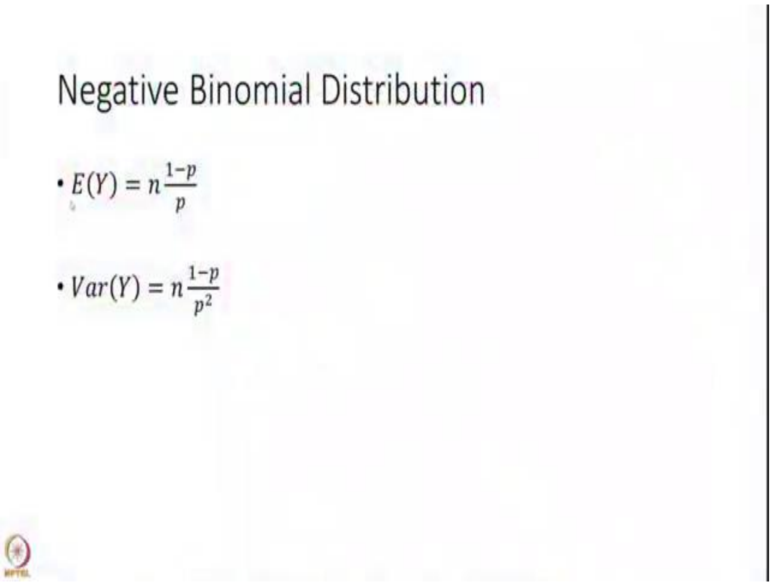
Now, this binomial coefficient n plus y, y minus 1 choose y. If you simplify it, it turns out to

$$\binom{n+y-1}{y} = (-1)^y \binom{-n}{y}$$

These, are also Binomial coefficients with negative integers T therefore this distribution is called negative Binomial distribution.

Please do not worry about going in to so many details if you can understand what exactly is the phenomena rest of it, it follows by itself. So, as far as we understand that there are sequence of independent Bernoulli trails, in which a random variable X will denote the number of success number of trails we had to go through until we arrive at the nth successful trail n is a fixed number than we say that x follows a negative binomial distribution, so if this is understood rest of the mathematics will follow by itself.

(Refer slide Time: 18:09)



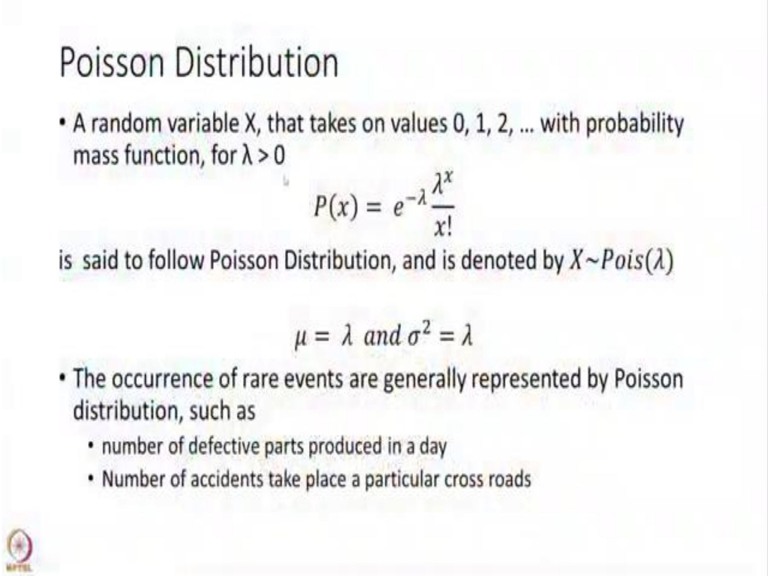
In this case the it is easier to calculate the expected value of y and expected value and variance of y, from which you can off course workout the expect value of

$$E(Y) = n\frac{1-p}{p}$$

$$Var(Y) = n\frac{1-p}{p^2}$$

Now, we move to another distribution very commonly used to distribution which is called Poisson distribution.

(Refer slide Time: 18:57)



This is a random variable which takes values 0, 1, 2, 3 onwards and has a mass probability mass function defined in such a way where lambda it is parameter which is a positive value. Which takes on only positive value,

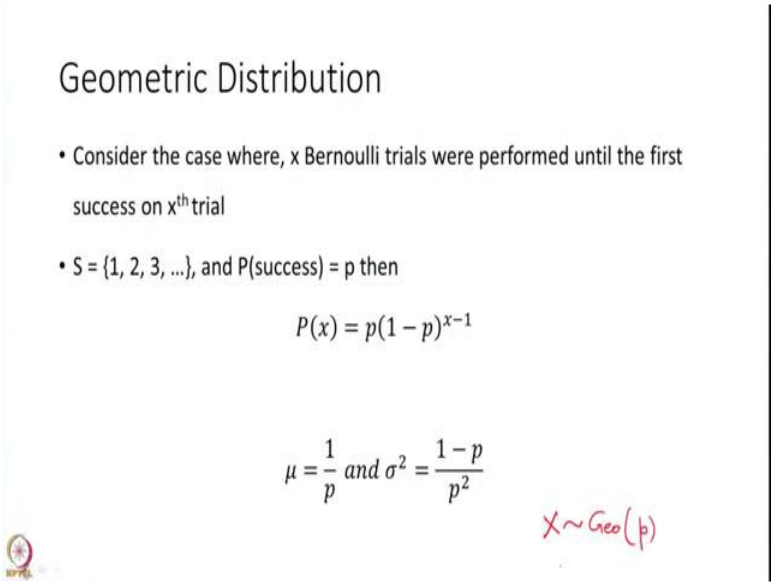
- A random variable X, that takes on values 0, 1, 2, … with probability mass function, for $\lambda > 0$

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

$$X\sim Pois(\lambda)$$

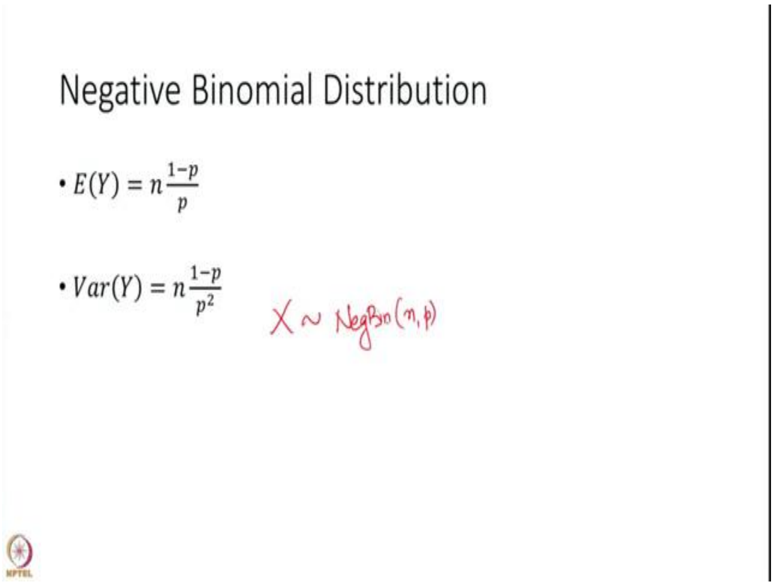
$$\mu = \lambda \ and\ \sigma^2 = \lambda$$

(Refer slide Time: 19:31)



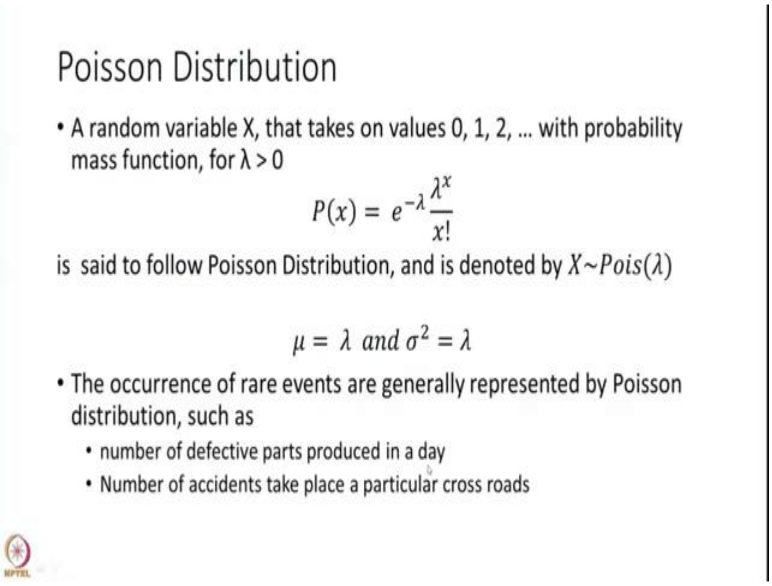
I think I forgotten to mention the notation for distribution, so let me do it here for your information. If X follows a geometric distribution, we are going to say that X follows geometric and it has only one parameter which is p it is called geometric distribution.

(Refer slide Time: 20:04)



If it follows negative binomial distribution the notation is X follows negative binomial n, p. Remember the parameters are same as in binomial so it is n and p.

(Refer slide Time: 20:32)



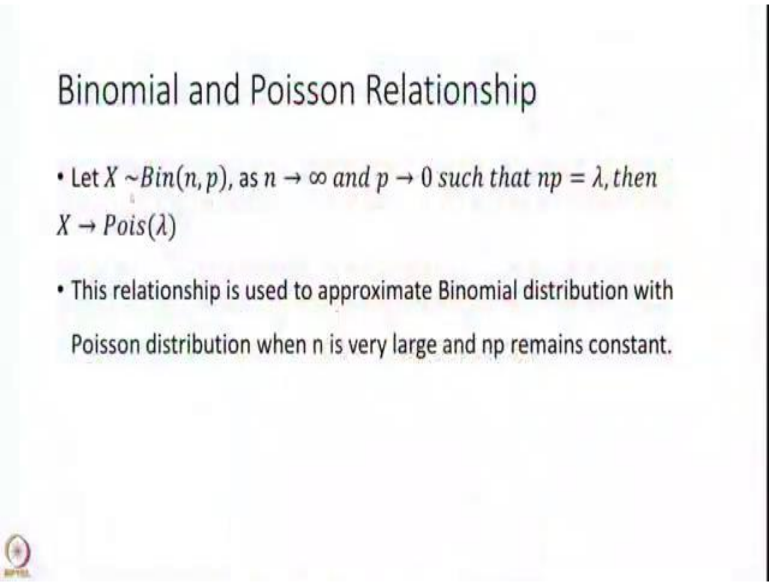
Please note this, now we move on so Poisson distribution you denote it by X following Poisson with lambda the mean value is lambda and variance is also lambda. This is the specialty of this distribution where the mean value and the variance are the exactly the same, this can be one of the characteristics of this distribution when you trying to identify a distribution.

Were, does this distribution occur you see we started with a Bernoulli trail uniform distribution Bernoulli trail and the distribution derived from Bernoulli trail we had some genesis. Poisson distribution genesis is such that when you are talking about a very occurrence of a very rare event that is an event which occurs very, after a long period of time it occurs and it occurs with very small probability it is generally tends to follow Poisson distribution.

And, therefore for example the number of if you have good edited book number of typographical error in a page can be can follow a Poisson distribution. A number of defective parts produce in a day if you are system is very well defined and it follows in a very quality assured way then the possibility of getting a defective part is very, very, very low and therefore it it tends to follow Poisson distribution.

A number of accidents that can be take place on a particular crossroad in a day is a very rare event fortunately and therefore it is called Poisson distribution. The Poisson distribution is also derived with this philosophy and from binomials, so there is relationship between a binomial and Poisson distribution.

(Refer slide Time: 22:46)



This relationship is

$$X \sim Bin(n,p)$$
$$n \to \infty \; and \; p \to 0 \; such \; that \; np = \lambda$$

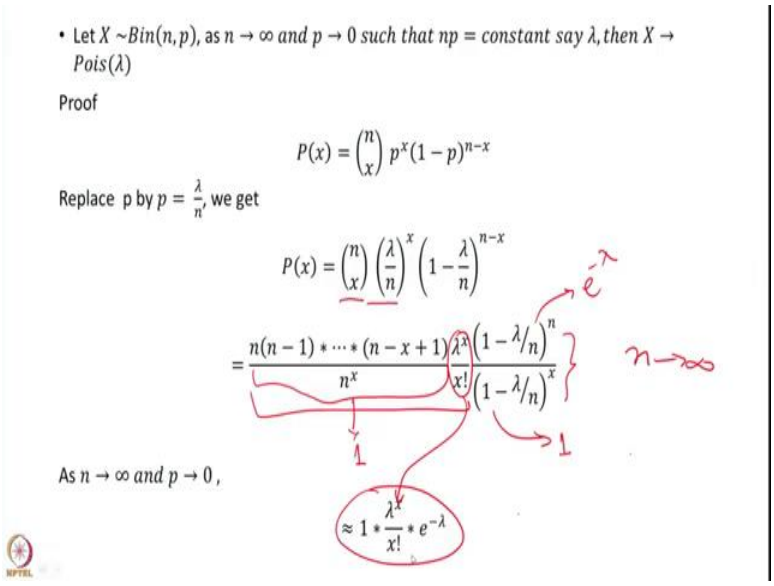
It means that you have conducted that the probability of success is p and you have conducted n independent Bernoulli trails, then as n tends to infinity the number of trails tend to infinity and the probability of success the probability of occurrence tend to 0. In such a way that the np that is the, you remember the average of the binomial distribution remains constant. Then, X as n tends to infinity tends to Poisson distribution with parameter lambda which is the average value of Binomial distribution.

Now, you can see why we do not want to called probability p as probability of success here. Because we do not want to have success tending to zero. So, therefore number of times we simply say the probability and it is probability of occurrence of error or defective or something.

So, here there is slide change in the definition of p because otherwise it is sounds very funny that you have so many trails and you have no success that what not we are trying to say.

This relationship is used to approximate binomial distribution with Poisson distribution when n is large and np remains constant. When n is large I have forgotten to write n is large and p is getting smaller and np remains constant.

(Refer slide Time: 24:32)



Will, I will show you how it can be prove mathematically

- Let $X \sim Bin(n,p)$, as $n \to \infty$ and $p \to 0$ such that $np = constant$ say λ, then

$$X \to Pois(\lambda)$$

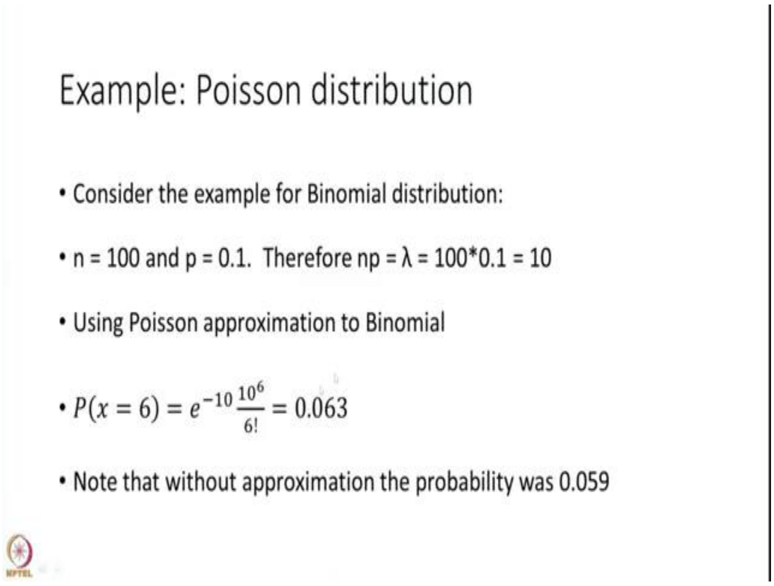
- $P(x) = \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x}$

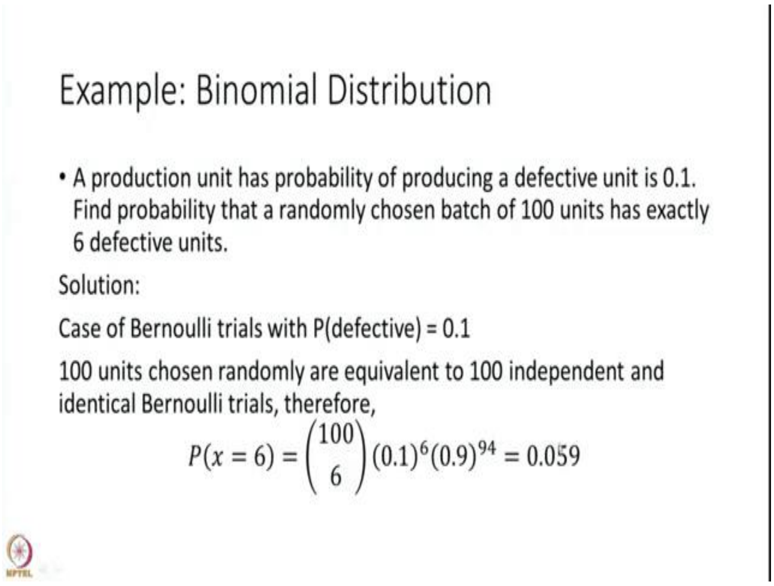
- $= \frac{n(n-1)*\cdots*(n-x+1)}{n^x} \frac{\lambda^x}{x!} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^x}$

As $n \to \infty$ and $p \to 0$, $\approx 1 * \frac{\lambda^x}{x!} * e^{-\lambda}$

The distribution the random variable tends to Poisson distribution with n p, which is a constant as it is average value.

(Refer slide Time: 26:26)





Here it is an example of Poisson distribution.

n = 100 and p = 0.1. Therefore, np = λ = 100*0.1 = 10

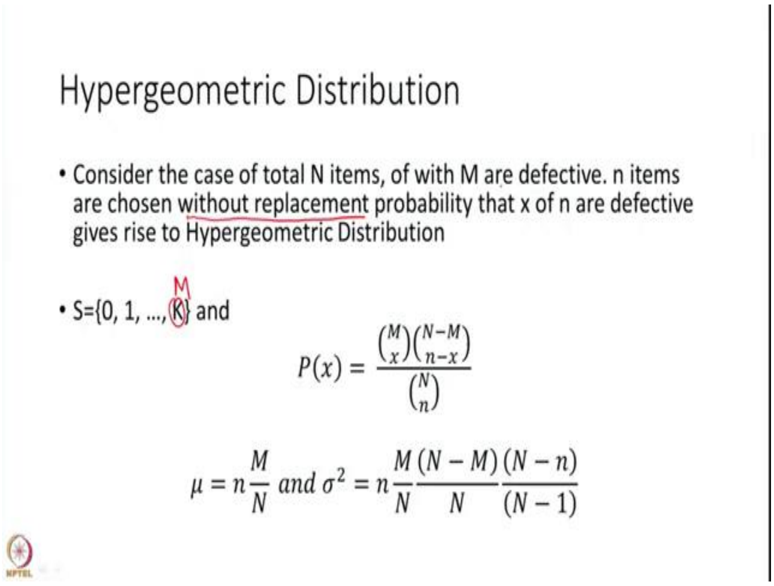
Using Poisson approximation to Binomial

$$P(x = 6) = e^{-10}\frac{10^6}{6!} = 0.063$$

Please recall that the same value in the past sorry the same value in the past had been 0.059 instead we got approximated value as 0.063, see this approximation was 0.059 and 0.063 which are almost the same. So, this is how Poisson approximation in number of problems that you solve. It is much easier to solve it through Poisson method than finding the 100 choose 6 kind of parameter. It is much easier to as calculate e to the power 10, 10 to the power 6 divided by 6 factorial.

The next distribution that I would like to talk about is a hypergeometric distribution. This is not coming out of Bernoulli trails, it is very different one but remember that when we are considering the Binomial distribution. We are considering the case of choosing N items with replacement. While, here you are choosing the N items without replacement.

(Refer slide Time: 28:41)



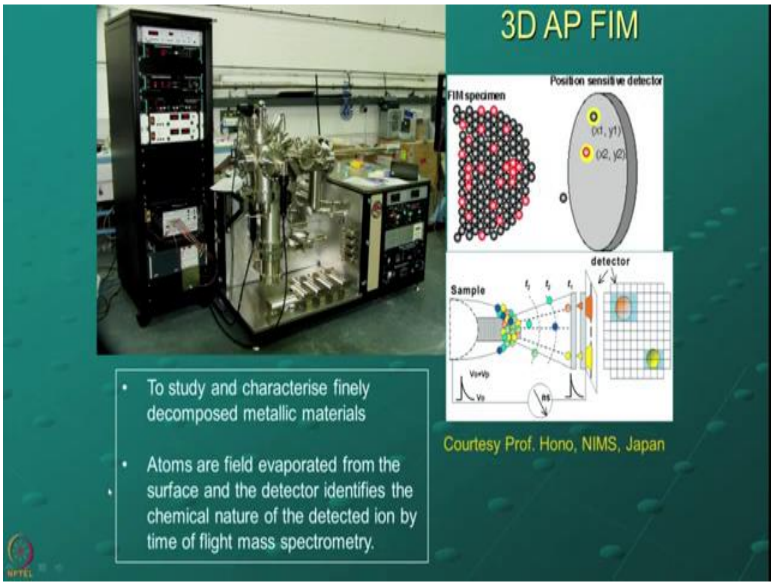
This, is a main difference, so here we consider a case of total N items that you have of which M are defective. n small n items are chosen without replacement with a probability that x of n are defective gives rise to hypergeometric distribution. So, your sample space is 0, 1, 2, 3 K where does the K come from I think so I have taken a mistake. It should not be K it should be M. Let me correct it, it should be M, so you have x is equal s your sample space is 0, 1, 2, 3 up to M.

$$P(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

$$\mu = n\frac{M}{N} \text{ and } \sigma^2 = n\frac{M}{N}\frac{(N-M)}{N}\frac{(N-n)}{(N-1)}$$

So, you can see that it n multiplied M divide by N and sigma square is this. This, comes as I said this is because we are considering without replacement binomial distribution comes with replacement.

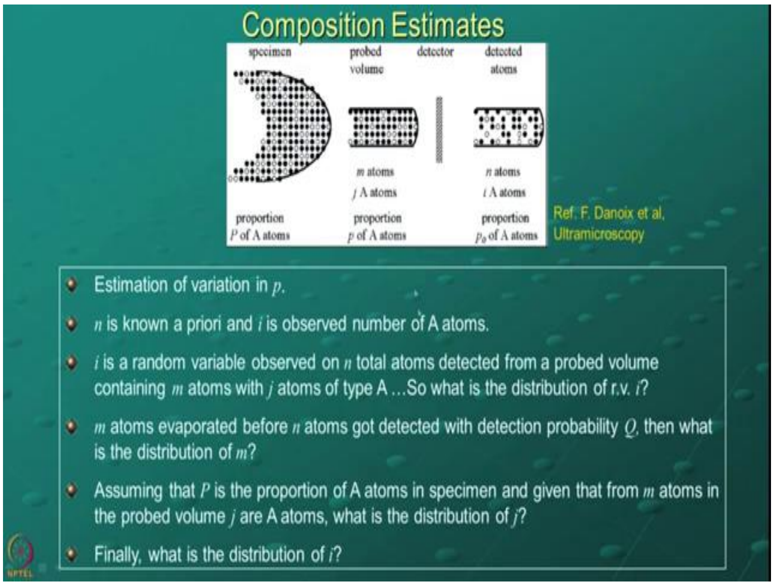
(Refer slide Time: 30:30)



Now, where is the application for all this distribution in material science we say that dealing with material science data. So, here it comes the all answers I would like you to find out but here is case of three dimensional atom probe field ion microscopy. This is the machine and the purpose is that you take any specimen any metallic or any material specimen.

Which has two kinds of atoms in it and you take a very, very thin sort of a thin cylinder like you know in a pencil what you have you put in the microscope and then it uses the it probes and it finally takes out the atoms from the specimen and it detects over the detector. So, the atoms are field evaporated from the surface and the detector identifies the chemical nature of the detected ion by the time of light in the mass spectroscopy. So, how much time it takes to fly to reach depending on that it actually detects the item the atom.

(Refer slide Time: 32:00)



You can refer to this paper but it is very beautifully describing that if this is your whole specimen the probed volume will be a piece of that so suppose you assume that you are looking for atom A which is the dark one. And you are looking for proportion P of atoms A this is your main purpose.

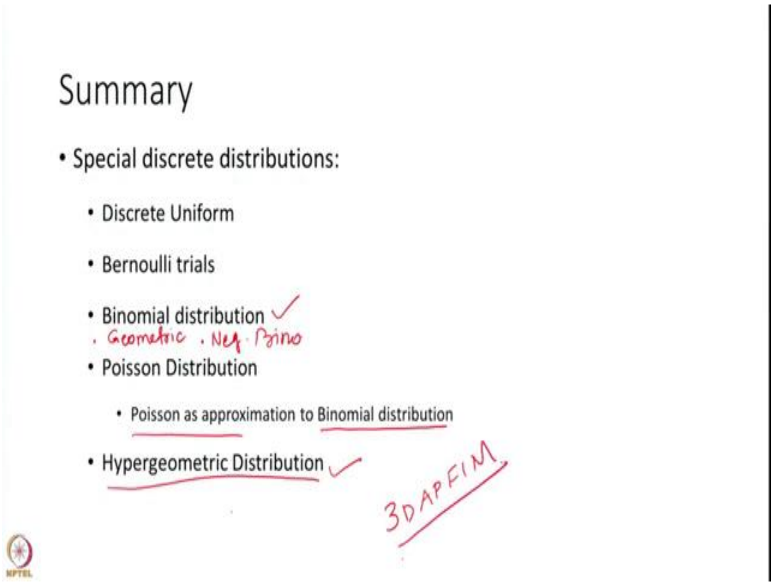
What you are getting out is a probed volume which as a total number of says small m atoms and j are the A type atoms and the proportion of p of A atom is you are looking for. Now, what really comes is through detector is you see on detected atoms in which you see small n number of total atoms of which I is the A type of atoms and then you only know the proportion P_0 of atoms A which has come out in the detected atoms.

And this paper actually describes how to estimate the variation in this p value. This this p is the best estimator for your capital P here and this are the questions that you have to answer you have all the knowledge that you need in this particular area in this for the discrete distributions.

So, let us take one after the other, n is known a priori and "i" is observed value of a atoms. "i" is a random variable observed on n total atoms detected from a probed volume containing total of m atoms with j atoms of type A. So, this distribution will be hypergeometric distribution. If m atoms evaporated before n atoms got detected with detection probability of then what will be the distribution of "m"? It is like coming to the nth success you conduct the Bernoulli trails till you come up to end success. So, this is going to lead to negative binomial distribution.and, then assuming that capital P is the proportion in the specimen given that small m atoms in the probed volume and of which j are the atoms what is the distribution of "j" , I leave it to you. And finally given all that what is probability of "i" ? See we have already said one distribution of random variable i. I am asking the distribution of random variable i again, it is good question for conditional probability negative pro negative binomial distribution and hypergeometric bino hypergeometric distribution.

You are welcome to refer to this particular paper, it is having the solution but it will give you the idea that all distribution has some role or other to play in the field of material science data analytics.

(Refer slide Time: 35:24)



So, let us summarize quickly we introduce special discrete distributions in today's lecture. We introduce discrete uniform distribution which is a very basic kind of a distribution. Then we introduce Bernoulli trails I have missed out a few things I will note them down here. We introduce Bernoulli trails and then we introduced three distributions arising out of Bernoulli trail binomial distribution geometric distribution and negative binomial distribution.

Then we introduced Poisson distribution and it is relationship with the binomial that is in some sense as the number of trail increases to very large quality. It tends to become Poisson distribution provided they average of binomial distribution tends to constant value lambda as p also tends to 0.

Hypergeometric distribution binomial distribution occurs without replacement this occur with replacement. I am sorry this occurs with replacement this hypergeometric distribution occurs without replacement. We took an example of 3D APFIM and raised a few questions and give you the reference of the paper to understand that, all these distributions do occur in the world of material science and materials engineering,

Thank you.