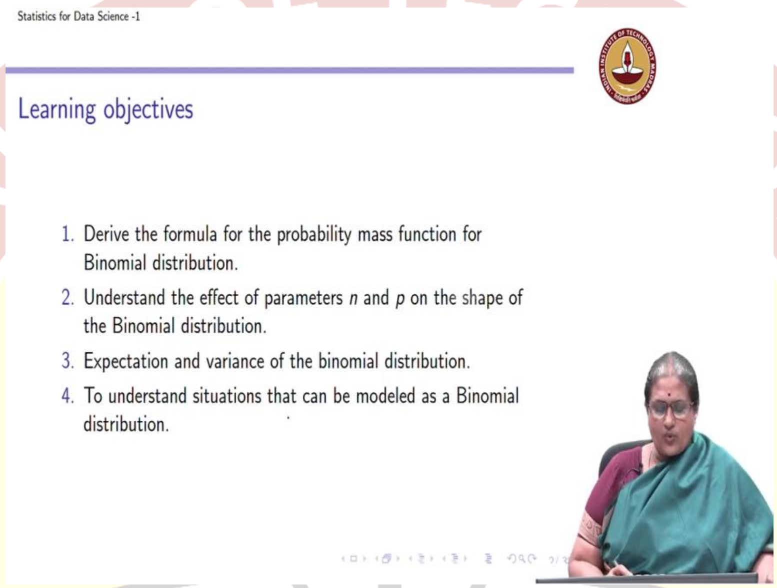
INDIAN INSTITUTE OF TECHNOLOGY MADRAS • सिद्धिर्भवति कर्मजा •

IIT Madras

ONLINE DEGREE

Statistics for Data Science – 1
Professor. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture No. 10.4
Binomial distribution – Modelling Situations as Binomial Distribution

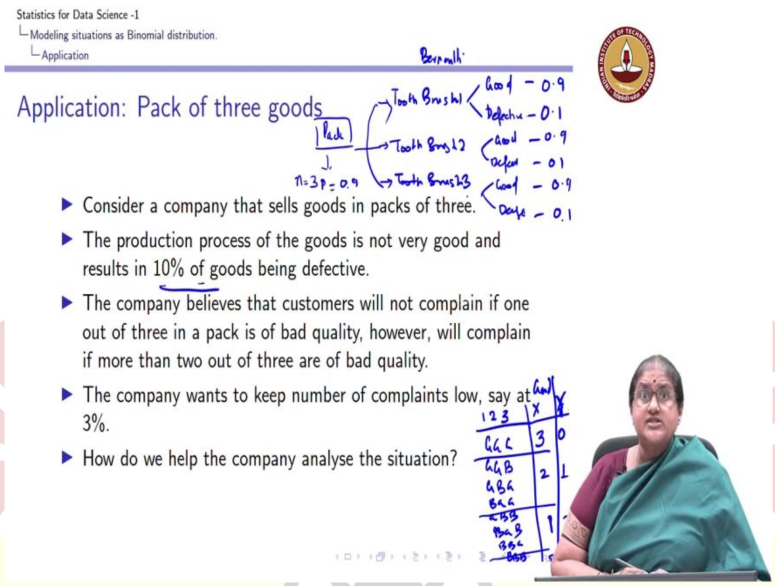
(Refer Slide Time: 00:23)



So, what we have learned so far is that the derived so you go back to our learning objectives, what we saw was to derive the formula for the probability mass function of a binomial distribution, we have derived the probability mass function, the next learning objective is to understand the effect of parameters n and p, we also looked at the effect of the parameters of n and p on the shape of the binomial distribution.

The next thing is we need to understand the expectation and variance of a binomial distribution. But before we understand what about the expectation and variance of a binomial distribution, let us look at and let us look at the case where I want to understand what are the situations that can be modelled as a binomial distribution.

(Refer Slide Time: 01:11)



Now, let us look at an application or real time application of how to model a particular situation using binomial and Bernoulli random variables. Consider a company that sells goods in packs of 3, now, we have many a time we come across situations where I have a pack of 3 toothbrushes, I have a pack of 3 pens and all of this so people or companies bundle these goods together so are having 3 identical toothbrushes which are being sold as a pack of 3.

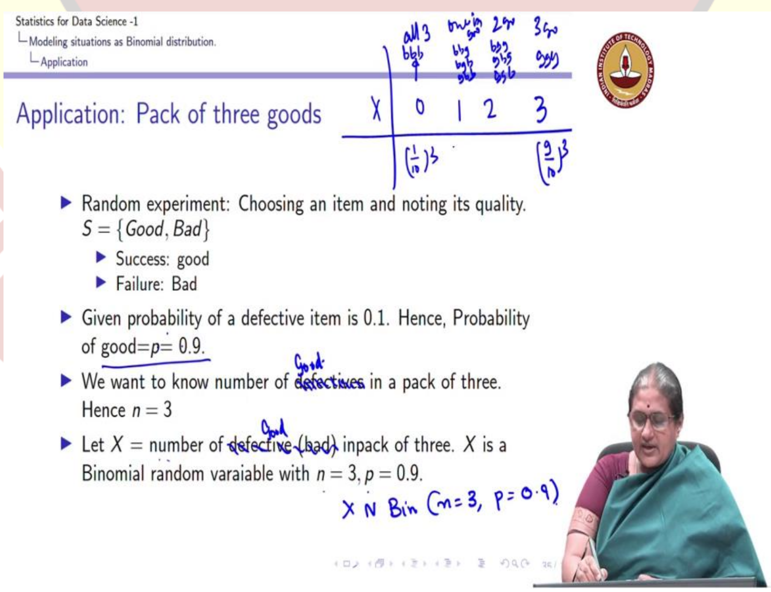
Now, each one of these toothbrushes undergo a manufacturing process or a production process and this production process I know the production process is not very good and it results in a 10% of the goods being defective, that is what I can see, the 10% of the goods are being defective. Now, why I said a Bernoulli trial? I can look at a production process which produces toothbrush, so a Bernoulli trial, the outcome of a Bernoulli trial is the outcome of a good toothbrush which is either good or defective and I know that the chance of it being defective is 0.1 that is what is given to me with the chance of it being good is going to be 0.9.

Now if I am looking at a pack of 3 I have a toothbrush 1, I have a toothbrush 2 and I have a toothbrush 3 and I know that each one of these toothbrushes could be either good or defective, since they come from the same manufacturing process the probability of it being good or defective for these toothbrushes is going to be the same which is 0.0 and 0.1. So, I can look at each one of them as identical and independent trials, so I can look at the entire pack of 3 toothbrushes to have a n=3 with the probability if probability of a good toothbrushes is 0.9 it is going to be 0.9 depends on whether you are counting the number of good toothbrushes or bad toothbrushes.

Now, the company believes that customers will not complain 1 out of the 3 in a pack is of a bad quality, so if I look at the toothbrushes this pack the possible are I could have all the 3 good, I could have 1 of it 2 of it good and that can happen this is the first toothbrush, second toothbrush, third toothbrush 2 of it good, only 1 good and all of them bad or all of them defective. So, I have a possible 8 outcomes, but here you can immediately see that this resembles the toss of a coin 3 times, but here I do not have the probability of getting a good as same as the probability of getting a bad I have different probabilities.

So, I have the outcomes of the experiment which if I can if I am counting the outcome to be number of good here I have 3, I have 2, I have 1, I have 0, if I am counting that this is good; if I am counting the number of bad suppose that is why this is going to be 0, this is going to be 1, this is going to be 2 and this is going to be 3. So, it depends on what you are counting the model of count, so if this x equal to 3, 2, 1, 0, y could be 0, 1, 2, 3 it is all counting. Now, the company wants to keep the complaint slow say at 3%, so how do you help the company analyse a situation?

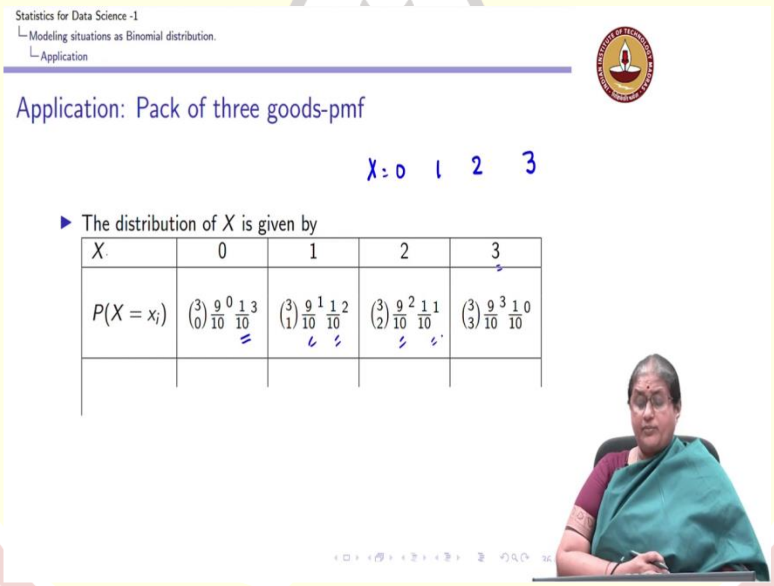
(Refer Slide Time: 05:10)

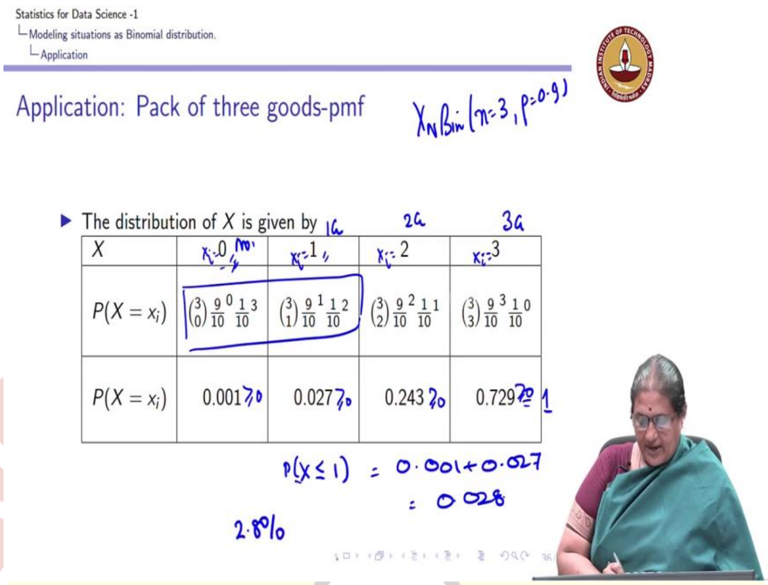


The first thing is I realized is a Bernoulli trial with good and bad or good and defect, the probability of good is 0.9, now I might want to know his of looking at number of defectives I want to look at the number of good items, I am going to put it as number of good items in a pack of 3, I realize because X is the number of good items a probability with each I get a good item is 0.9, I realized quickly that X is a binomial random variable with n = 3 and p = 0.9. These are the parameters of my distribution.

Now, if X is a binomial random variable X takes the value 0, 1, 2 and 3, X takes the value 0 means all the 3 are defective, so it corresponds to bad bad bad, here I have X = 1 only 1 is good none good or here so this corresponds to be bbg bgb gbb. 2 good this corresponds to bgg gbg and ggb, all 3 good corresponds to ggg. Now, if I am looking at the probability distribution I know all the 3 are bad is going to be $\frac{1}{10}^{3}$, this is going to $\frac{9}{10}^{3}$.

(Refer Slide Time: 06:43)





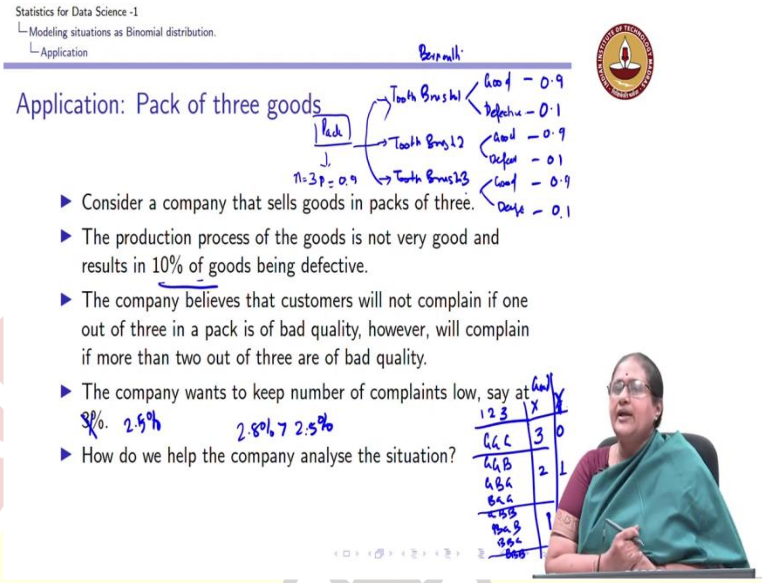
And here this is going to be $\binom{3}{1}\frac{1}{10}^{2}\frac{9}{10}^{1}$ and $\binom{3}{1}\frac{1}{10}^{1}\frac{9}{10}^{2}$, which on simplification I can see that probability X =xi, where xi takes the value 0 xi take the value 1, xi takes the value 2 and xi take the value 3 is given by this you can verify it is a probability mass function by checking that this adds up to 1 and all of them are greater or equal to 0.

But that is not what we are seeking we have got we have recognized that this X is a binomial random variable with n=3 and p=0.9 where I am finding p is the probability of finding a good teeth brush or a non-defective toothbrush. Now, people are going to complain if they find more than 1 toothbrush which is bad, so what is when would people complain?

So, people I know are going to complain if they find more than 1 which is bad, when do I have more than 1 which is bad? That would read or that would translate to probability X is less than or equal to 1, more than 1 is bad tells that again you see that more than one is going to be bad would be when all the 3 are bad or 2 out of 3 or bad, here I have all 3 good, here I have 2 good, here I have 1 good, here I have no good.

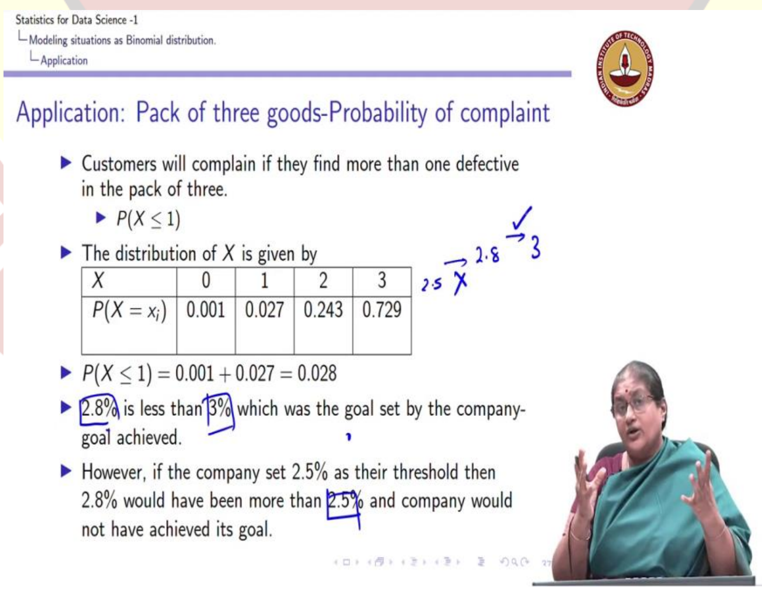
So, these two cases are going to be the case where a person is going to complain. In other words I would interested in knowing what is the $P(X = 0) + P(X = 1) = 0.001 + 0.27 = 0.028$, which I can say is about 2.8%.

(Refer Slide Time: 08:42)



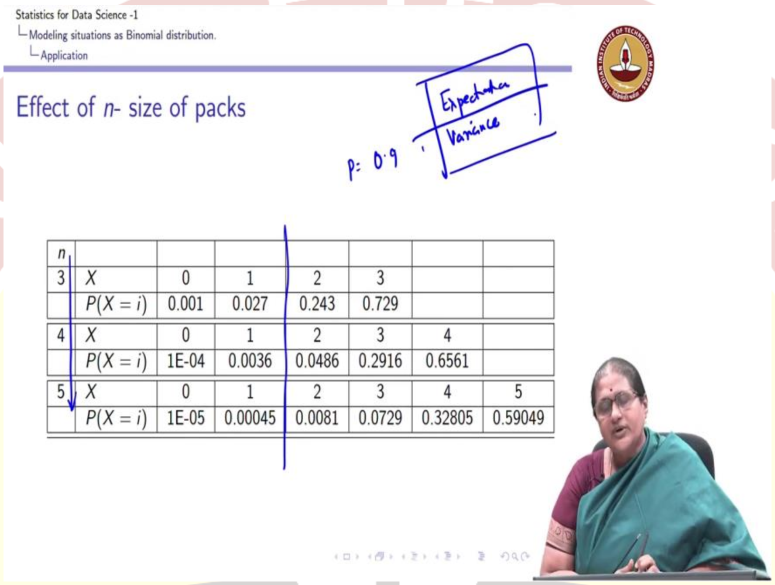
Again you go back and see that the company wants to keep this at less than 3% it is at 2.8%, so I am meeting my goal. But however, if the company had kept this as 2.5% this is a threshold which a company sets, you see that 2.8% > 2.5% and the company misses its goal of keeping customers happy. So, we have seen that a simple example of a production process can be modelled using a normal random variable.

(Refer Slide Time: 09:20)



So, we can see that this 2.8 again whether it was 2.5 or whether it was 3% was again what a company decided as a threshold and depending on whether this threshold is whether what I achieve from my data is 2.8 if my company sets at 3, I have achieved my goal, if the company sets at 2.5 I have not achieved my goal. So, now what would happen if I had packs of 4, is a pack of 3, again if a customer says that okay 1 out of 4 is acceptable to me, I will not complain, you can see there following.

(Refer Slide Time: 10:02)



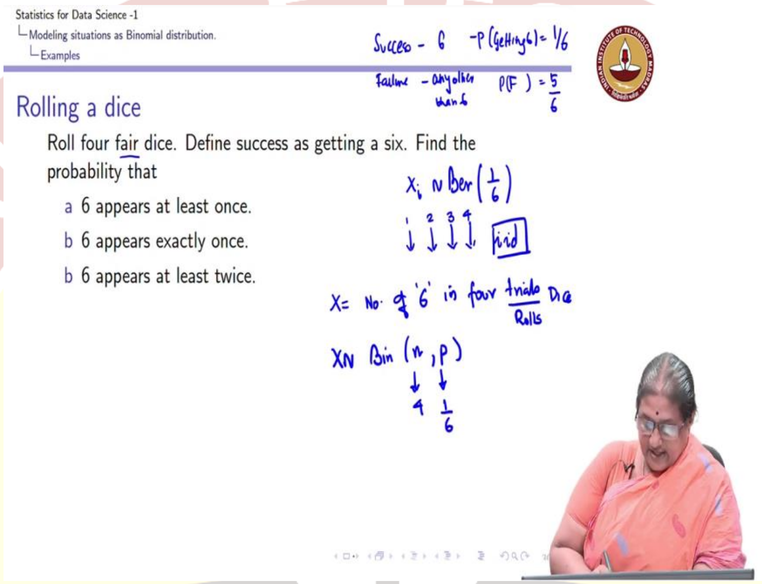
For the same P and what was that P? 0.9. If I keep varying my n, you can see that what is happening to my probability X <= 1, which is again very in intuitive, the chance of finding a defective as my pack sizes are going to go higher the chances of me getting a defective or which chances that I am going to complain is going to be much lesser.

So, is the answer to increase the pack size and look for a defective or if I am going to find 1 defective in a pack size of 1, is definitely going to in a pack size of 3 is going to the chance of finding a defective in a pack size of 3 would be higher than the chance of finding a defective in a pack size of 4 and that is what this demonstrates. So, we have taken a simple example of how we model a situation using a binomial random variable.

Now, we are going to next we have already introduced concepts of expectation and variance of a random variable, discrete random variable, recall the introduced expectation as a long-run average and variance as some measure of dispersion of a random variable. So, now we are going to find out what is expectation and variance of a binomial random variable and how we can use these quantities to help us take decisions in real life.

Now, what we have shown here is how we can model a particular decision or a process using a binomial random variable. Now, we would next is we are going to introduce a concept of expectation and variance and see how we are going to use these concepts to answer or help decision making, this is what we are going to do next.

(Refer Slide Time: 12:11)



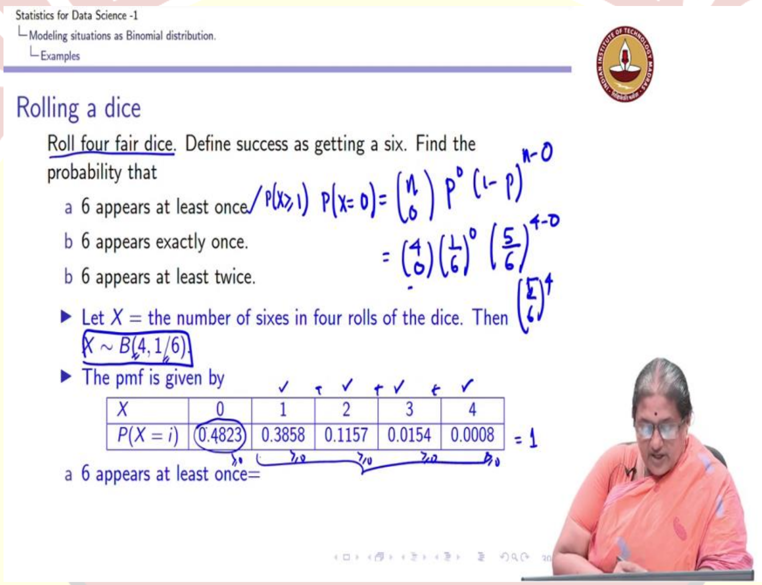
So, we continue by looking at a few examples where we are going to model real life situations or situations that we are going to that arise in day-to-day life as a binomial distribution. So, we look at a few examples before we go forward. So, now let us take the chance again we go back and we understand what these binomial distribution is by looking at rolling 4 dies.

So, I have 4 dies, I am going to roll them and we have already seen that the Bernoulli trial when I roll a dice is I define success as getting a 6 and this I get a 6 in a fair dice again because it is a fair dice the probability of getting a 6 in a fair dice is 1 by 6. So, my Bernoulli trial is where I define a success as getting a 6, a failure is getting any other number other than a 6, any other than 6, this is a failure.

And obviously the failure is going to be probability of 1 minus 1 by 6, this probability of a failure equal to 5 by 6. So, my Bernoulli trial xi which is going to be the outcome or whether I am getting a 6 in my first or the i_{th} trial is a Bernoulli with parameter 1 by 6. I am rolling 4 fair dice, so I am having independent rules. So, I know that the outcome of the first role is independent of the outcome of the second role which is again independent of the outcome of the third role which is independent of the outcome of the fourth role and each one of them have the same distribution of a 6 coming.

So, I have iid I satisfy the condition I require for a binomial experiment and I can see that define success as getting a 6, so X is the number of 6 in 4 trials or 4 rows of fair dice, I recognize this X is a binomial I need to know what am are my parameters n and p, I have for independent trials, so n equal to 4 and p would be 1 by 6.

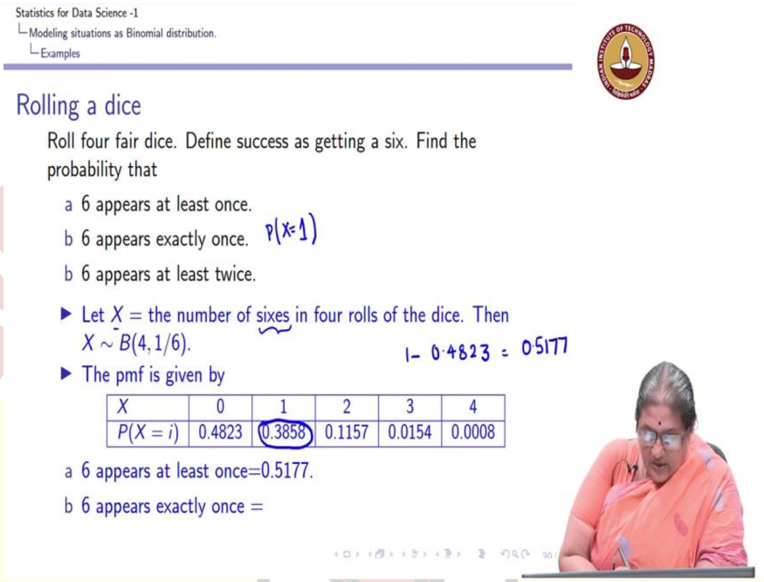
(Refer Slide Time: 14:53)

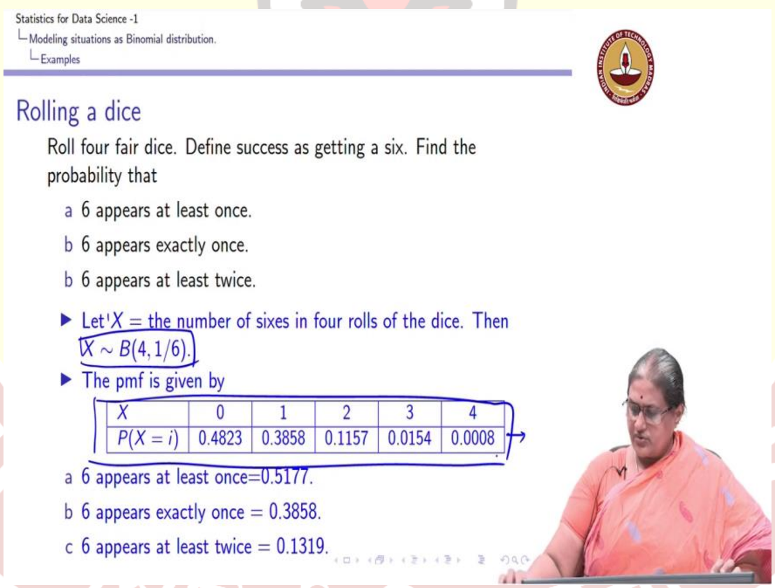


So, the first thing is to understand that this experiment of rolling four dice is a binomial experiment with p equal to 4 and with n, so it is a binomial experiment with n equal to 4 and p equal to 1 by 6. So, I can get the probability mass function, by probability mass function I can verify that probability X equal to 0, I know is n choose 0, p to the 0, 1 minus p to n minus 0, n is 1, so it is 4 choose 0, my p = 1/6 to 0, 5/6 4 minus 0 and you can verify that P(X=0)=$\frac{5}{6}^{4}$ and $P(X = 4) = \frac{1}{6}^{4}$.

Can verify these and you can check that this is my probability mass function, if it a probability mass function it indeed as I can see that all my probabilities are greater than or equal to 0 and they sum up to 1. So, now I am asking the question 6 appears at least once, so 6 appears at least 1 is equivalent to P(X>=1). So, it is the sum of this plus the sum of this plus the sum of this plus the sum of this or it is 1-P(X=0). So, either you add the sub because I know that the addition of these is equal to 1 minus probability 0.4823.

(Refer Slide Time: 16:54)



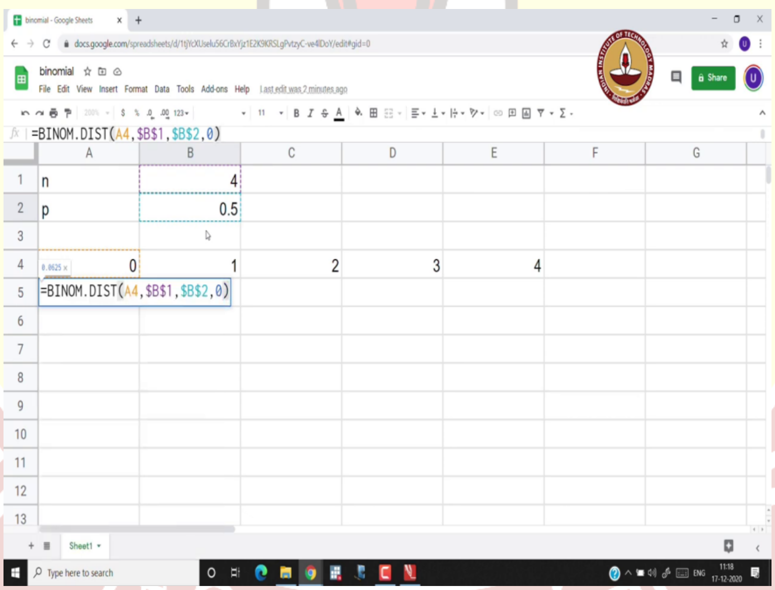


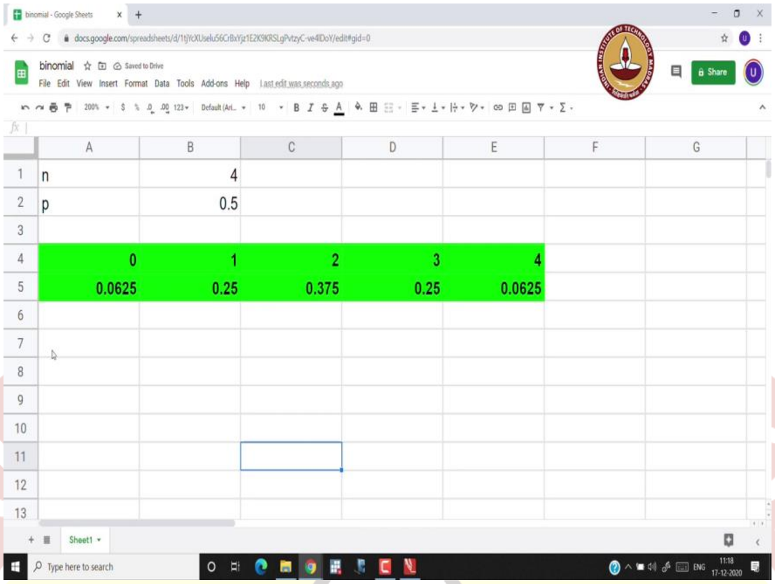
And I can see that the probability that X appears at least 1 is 1 minus 0.4823 which is going to be 0.5177 or in other words I can articulate it by saying that when I roll 4 fair dice and I define success as getting a 6 the chance that I am successful at least 1 of the 4 times is there is about a 51% or 52% approximately 52% chance that I am successful at least once, which is not a bad thing.

What is a chance that 6 appears exactly 1? So, the thing I am asking is what is the chance of exactly one success, remember X is the number of sixes or number of success in my 4 rows, so I could have no success, I could have 1 success, I could have 2 success, I could have 3 success, of I could have all the 4 to the sixes, so that is the success and this is just P(X=1)=0.3858 as shown in my probability mass function. Now, probability X appears at least twice is probability X is greater or equal to 2, so I add up these and I can find that that is 0.1319.

So, you can see that given the binomial so we identified this as a binomial distribution and from that we obtain the probability mass function. So, what is the next thing which we need to do is to understand how we got the probability mass function. Towards this I will just show how to get hold of the probability mass function we can go to Google, in our Google we can just type in a Google sheet, so let us go back to a Google sheet, I am just going to allow this Google sheet to be just some Google sheet, you are going do it.

(Refer Slide Time: 19:24)





So, you can see that I can open a Google sheet, I just type in the Google sheet, I type a binomial just for the sake of our understanding and within the Google sheet I will set up the parameters of my distribution, by parameters of the distribution I know I have to set up a n I need to give what is p. So, for example let us look at the case when I just have n equal to 4 and the p equal to 0.5. So, this is a binomial with n equal to 4 and p equal to 0.5, so I know the values my random variable would take our 0, 1, 2, 3 and 4.

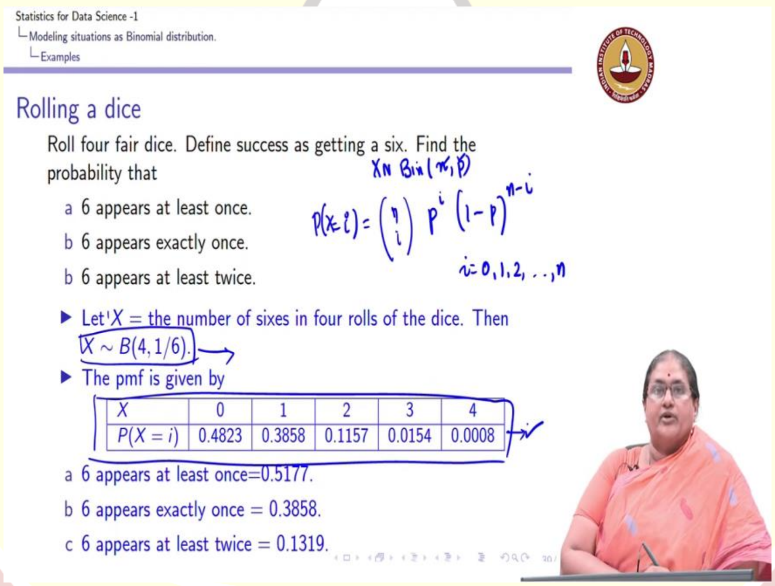
So, those are the values my random variable would take, so let me expand this so let us have a, so these are the values my random variable would take 0, 1, 2, 3, 4 and here what you are going to see is the distribution, so I a distribution which is binomial distribution probability, click on the binomial distribution probability, it asks for the number of successes, so what it means is the first thing it tells you the number of successes for which to calculate the probability, so it would take the values 0.

That is here I am just going to give it or I am just going to link it to the number of successes is going to be linked to the value which is a 4. The next thing I need to give is the number of trials that is going to be 4 and once I do that, I am going to freeze this because that is something which is going to be there for everything, the third thing is the probability of success, which I am again going to freeze and you can see that when once I freeze that and the final thing is I am not interested in cumulative, I am interested in just the probability mass function, I do that and you can see that this is 0.0625.

0.0625 I increase all of it you can see that for n = 4 and p = 0.5 this gives me or the shaded region whatever I am shading or filled it in green gives me nothing but the probability distribution function of a binomial distribution. The function we use to as a by binom.dist the first argument is the value that the random variable takes, so the value that random variable takes is either 0, 1, 2, 3, 4.

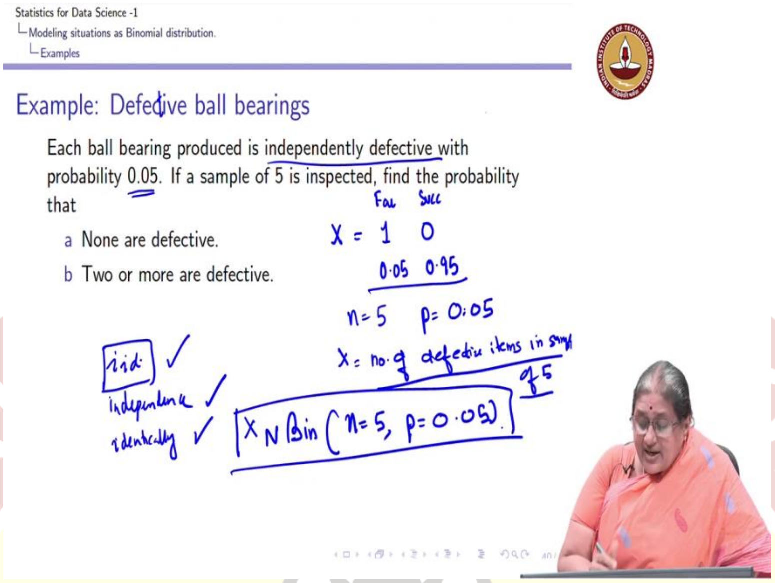
The second argument is the parameter n, the third argument is the probability p and I put 0 because I am interested in the probability mass function and I am not in the cumulative distribution function. So, you can obtain the probability mass function of a binomial distribution using the Google function by binom.dist.

(Refer Slide Time: 23:16)



So, going back to what we have done here, so you can see that I our binomial distribution I get the pmf, I can get the pmf either by my by looking at my n choose i, so probability x equal to i for x which is a binomial with parameter n and p the probability x equal to i is n choose i, p to the I, 1 minus p to n minus i, i takes the value 0, 1, 2, up to n. So, the first key thing in any application is to identify the parameters n and p, once you identify the parameters you set up the probability mass function and once you set up the probability mass function, you can answer the questions.

(Refer Slide Time: 24:13)



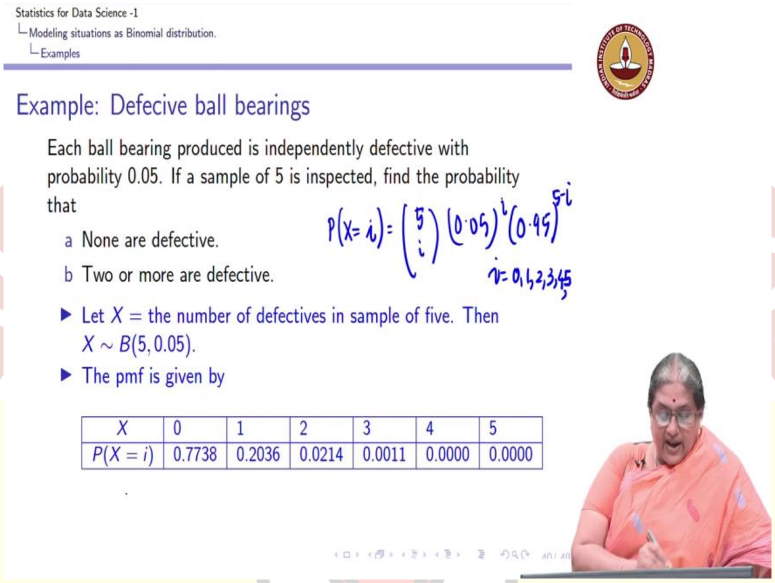
Now, let us look at the next thing where I am looking at defective ball bearings, so again I have a production process or a manufacturing process where each ball bearing produced independently defective with a probability 0.5. So, the Bernoulli trail here is the chance of a ball bearing being defective and I know that the success is being good and the failure is being defective or I can given the independently defective we can define this.

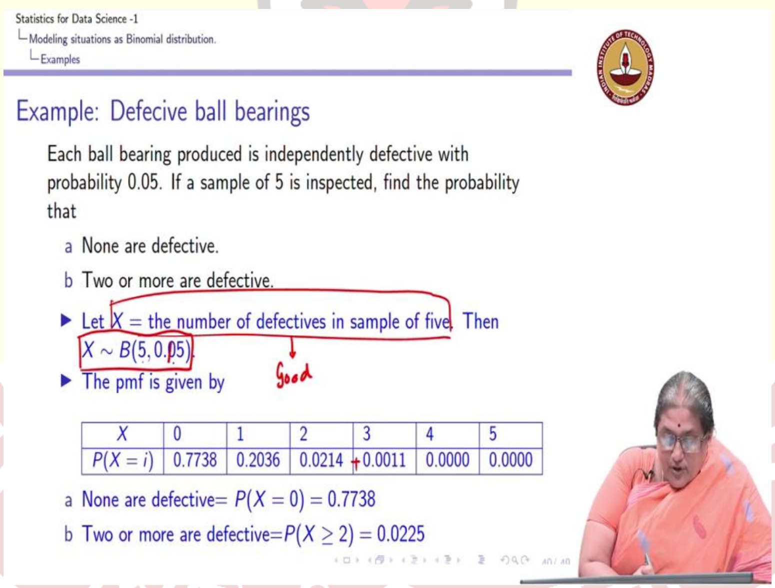
So, this is if failure is with p then successes with 1- p, you need to know what you are looking at, so here I can define because I am interested in knowing how many are defective, I can define X takes the value 1 and 0 this is failure, this is success, it does not matter to me, this it is taking the value 0.05, so this it is successful probability 0.95. Now, I am taking a sample of 5, so my n = 5 my p remains to be 0.5, so what is the question I see that the number of so how when I defining it number of defective items in my sample of 5.

So, first we need to check the iid, so that I take my first sample, if I take the first defective the chance that the first ball bearing is defective is with a probability 0.5, the second ball bearing is defective, it is independent whether the first was defective or not. So, independence is satisfied and the chance with which it is defective or not is again with the same probability 0.05, so it is identically distributed, so the iid assumption is what I satisfy or verified first, once I had check that then I can say that this X which is the number of defective items in a sample of 5 is a binomial distribution with parameter n equal to 5 and p = 0.05.

So, this is the most crucial part is to identify verify the conditions of a binomial distribution and to identify the parameters.

(Refer Slide Time: 27:00)





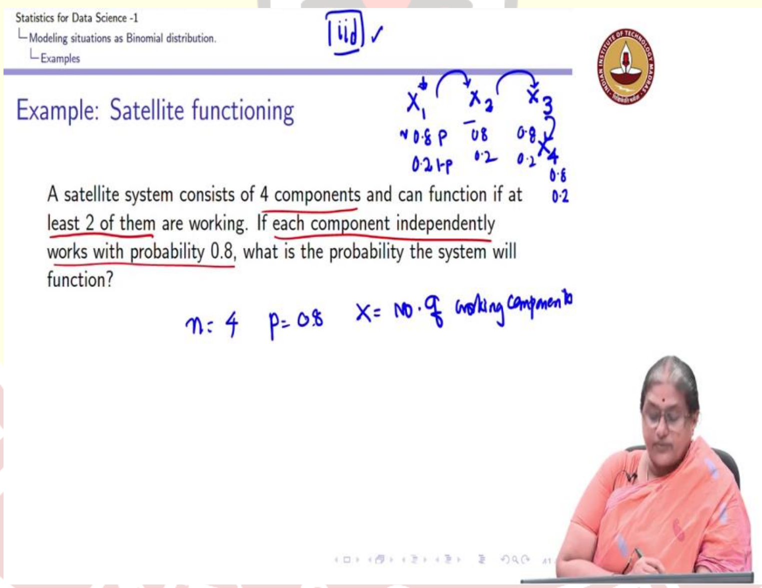
Once I identify the parameter, so I have defined the random variable and I have identified that it is a binomial distribution, I can either write it as B or B i n, once I have identified it, I get the probability mass function. Again recall the probability mass function here is probability X equal to i is 5 choose i 0.05 to the power of i into 0.95 to the power of 5 minus i, i takes value 0, 1, 2, 3, 4, 5. Those are the values my i take, so you can see that these are the value the i takes. So, now I am interested in knowing the probability that none of them are defective.

Now, X is the number of defective, so the probability none of them are going to be defective is equivalent to probability X takes the value 0, 2 or more a defective is same as P(X>=2), the first city straightforward none are defective is the probability which X = 0 which is given by 0.7738 that is something which I can see here, probability that none are defective is just 0.7738.

Two or more defective I can see that, that would be this probability plus this plus this plus this the last two probabilities are not actually 0, but they are very very small I have rounded up to four decimal places here, so you can see that that probability is 0.225, which is given here that is probability this plus this and these two are 0 so it just 0.0225.

Again the key thing was to identify the X is a binomial distribution with parameters n and 5 but prior to that how you are defining X is going to be very important. Suppose you had defined this as number of good items then this probability would have been 0.95 and these probabilities would have changed accordingly.

(Refer Slide Time: 29:31)

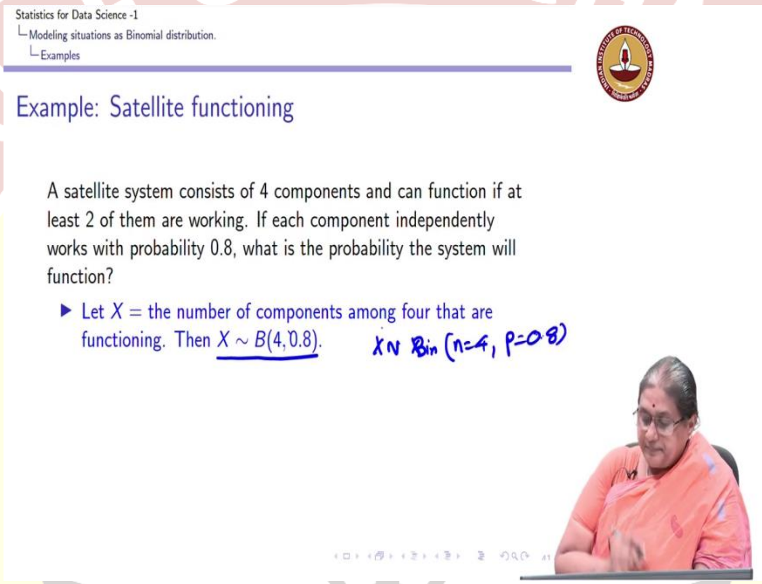


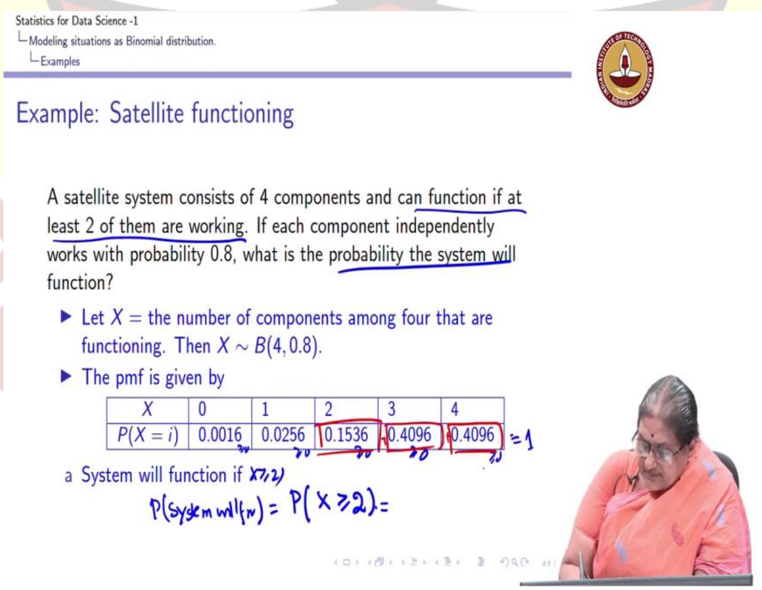
Now, let us look at a third example or application where I am interested in knowing about the functioning of a satellite system. Now, the satellite system consists of 4 components and I know that it can function as at least 2 of them are working. Now, what is the Bernoulli trial here? It says that each component independently works with probability 8.

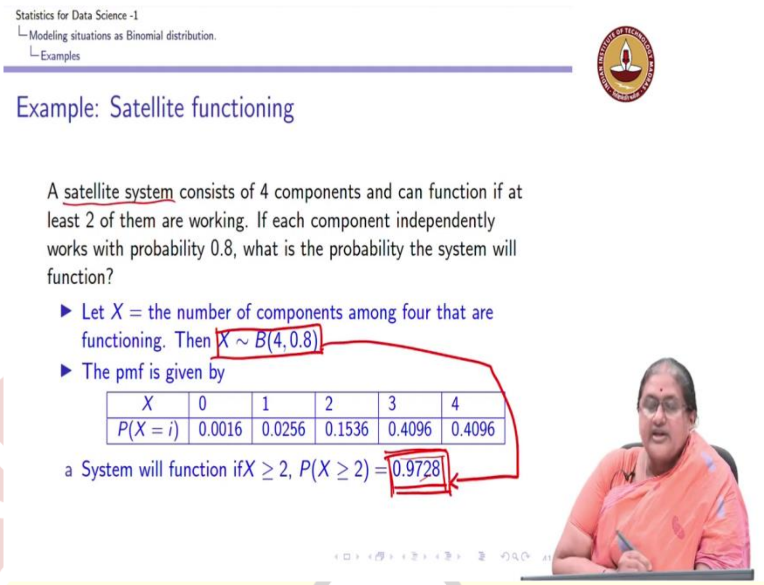
So, the Bernoulli trial I have here is I have 4 components, I can say X1 X2 X3 and X4 and each one of them independently works or does not work with probability 8, it does not work with probability 0.2, the chance that the first component work is independent of second is independent of third is independent of forth, because each component is working independently the probabilities are the same, so the iid condition is satisfied, I verify it.

Now, what is n? n equal to 4 components say n equal 4, p equal to 0.8, x equal to number of working components among the 4 components.

(Refer Slide Time: 31:02)





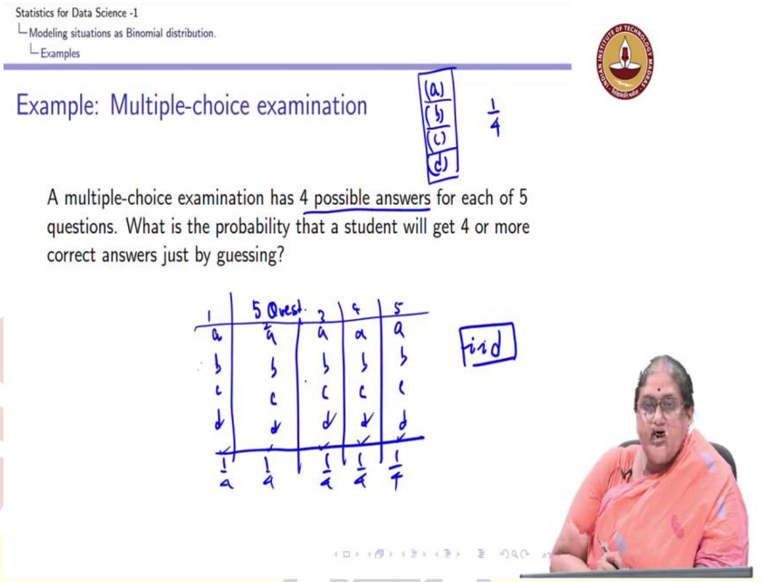


So, the first thing to realize is the number of components among 4 that are functioning, it is a binomial with parameter n equal to 4 and p is equal to 0.8. Again get the probability mass function, the probability mass function is given by the following, you can again check this is a probability mass function all of them are greater or equal to 0 and they add up to 1, so it is a probability mass function.

Now, the system will function if at least 2 of them are working, so what is the probability that the system will function? The system is going to function if x is greater or equal to 2, so the probability that the system would function yes if probability system will function is the same as probability x is greater or equal to 2 which is equal to you can verify that that is equal to this plus this plus this plus which is 0.9728 or the chance that the system will function is pretty high, which is 97.28.

Now, this 97.28 might seem to be a very high chance but remember we are talking about a satellite system and we also might want to have a very high level of precision or accuracy, so depends on what is the threshold a person might hold and whether 97.28 is a good threshold or not, but from the using the laws of probability and recognizing saying that X is a binomial I get the probability that it is going to function to be 0.9728.

(Refer Slide Time: 33:09)

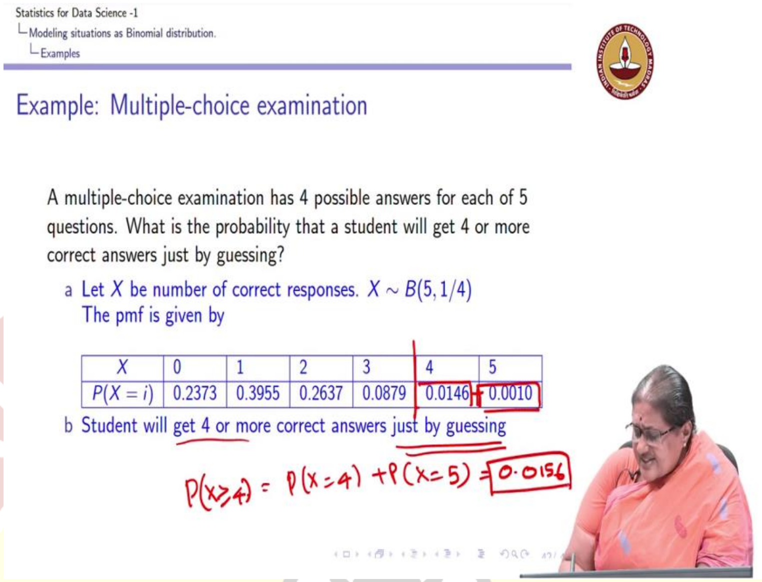


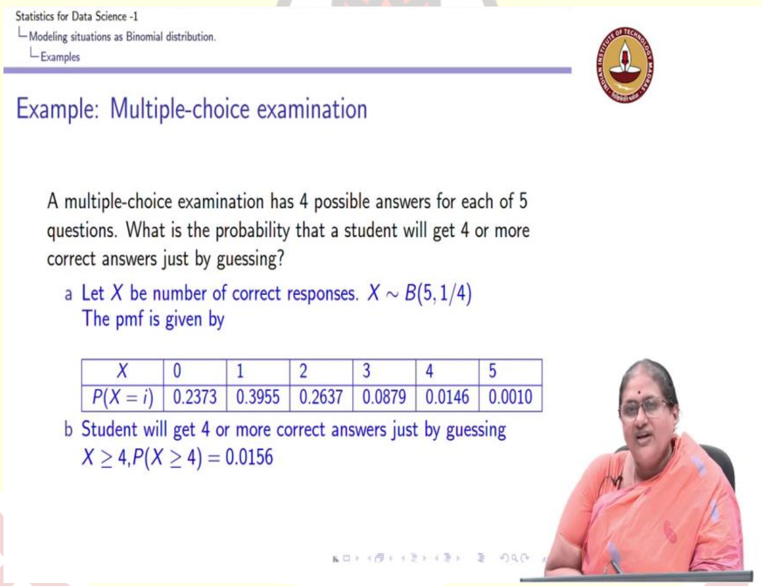
So, now let us look at a last application of a situation that is modeled as a binomial distribution, this is something all of us are very very familiar with that is when we are given a multiple choice quiz typically a multiple choice quiz you have the choices in a multiple choice quiz are typically most of the time you will have 4 choices and if you are going to guess it has 4 possible answers and the chance that one of them would be right is 1 by 4.

Anyone could be right, this is a multiple choice question and anyone could be right as 1 by 4. So, suppose I have 5 questions and each question has 4 choices and I am going to guess all of them I have 5 question each question has 5 choices, question 1, question 2, question 3, question 4, question 5. First of all is it a Bernoulli trial? Yes, here the chance of getting a right answer is 1 by 4, the chance of getting a right answer as 1 by 4, identically distributed each one of them have a chance of getting a right answer which is 1 by 4.

And it is independent also whether you are getting a right guessing a right answer in the second question is independent of your first is independent of third which is independent of four which is independent of five, so I am verifying the iid and I know each one is a Bernoulli trial to you me that if X is the number of correct responses then this X is a binomial distribution, again my parameter n equal to 5, p equal to 1 by 4.

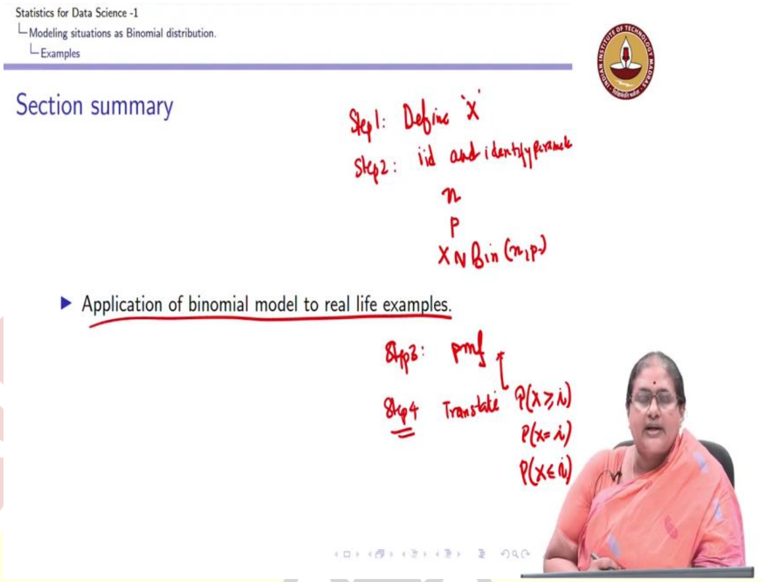
(Refer Slide Time: 35:10)





So, the next step I work out the probability mass function and I am interested in knowing whether in if a question of 5 the chance that a student would get 4 or more just by guessing the answer is going to be probability of X equal to 4 plus probability of X equal 5 that is probability X is greater or equal to 4, which I see is the sum of this plus the sum of this which is 0.0156, I see that this is a very very low number, so I can state that the person getting 4 or more just by guessing has a rather low chance. So, guessing does not get you a very very the chance is very low that you will get 4 or more.

(Refer Slide Time: 36:06)



So, what we has seeing so far is what is a binomial distribution and where how do we apply the binomial model to real life examples the key thing we need to do is step 1, define the variable X, so it was the number of correct answers it was the number of components working or it was a number of defective items or the number of times you get a 6.

So, this X was the definition that variable you are defining verify the iid property and identify parameters, by parameters identify the n and p so that you get X which is a binomial with the parameters n and p, the third step is to come up with a probability mass function or the distribution of X and then the last step is translate the questions asked in terms of the probabilities.

So, probability X is greater or equal to i or probability X equal to i of probability X is less than or equal to i, whatever it is translate them into this, use the information given in the probability mass function and answer the questions. So, you can see that a lot of real-time applications can be modelled as a binomial random variable.