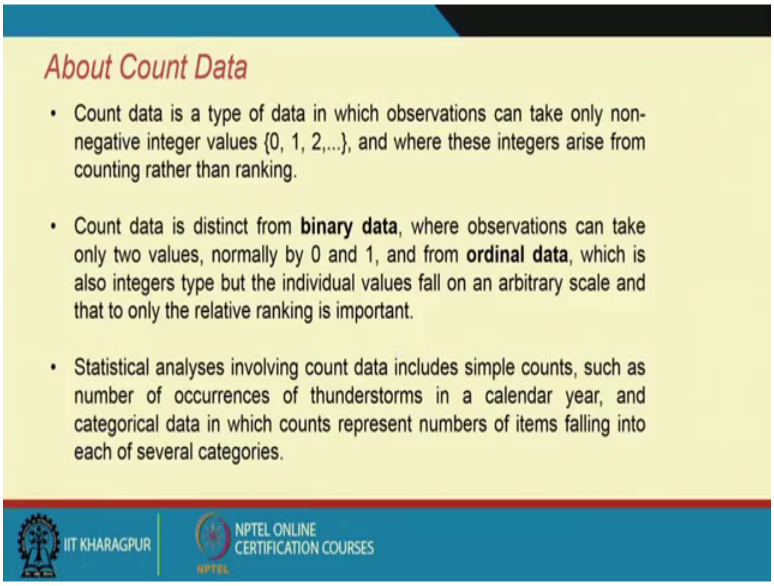
Engineering Econometrics
Prof. Rudra P. Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture - 58
Fitting Models to Data (Contd.)

Hello everybody, this is Rudra Pradhan here, welcome to Engineering Econometrics. Today we will continue with count data and discrete modeling. In the last lecture we have highlighted this particular concept with lots of you know what we can called as you know lots of difference corresponding to different types of you know data that too putting or bringing count data. And differentiating how count data can be different from other types of you know data.

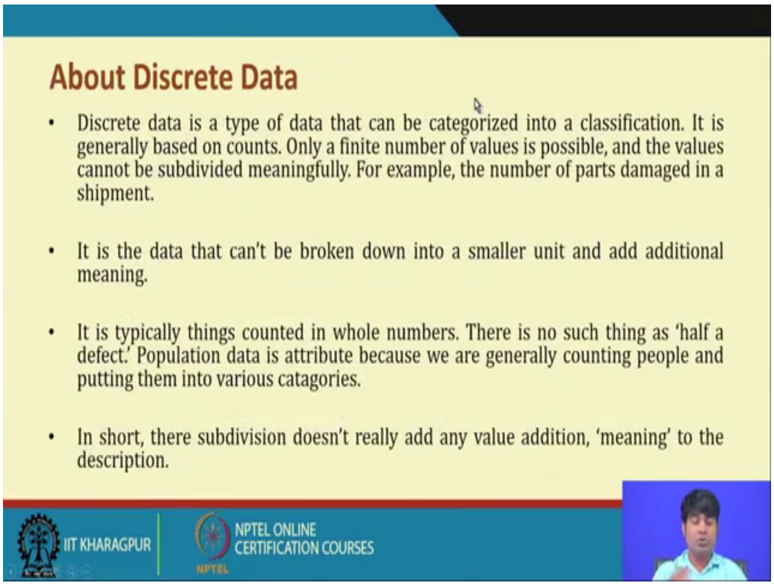
So, it is a kind of you know specialized kind of you know data through which we can analyze some of the engineering problems. As a result we have a modeling structure or a different kind of you know clusters engineering econometrics clusters that is what called as you know count data and discrete modeling.

(Refer Slide Time: 01:25)



So, in the last lecture we have already highlighted this you know count data, what is exactly count data and how you can understand and how you can do the modeling with respect to count data. Once again I am bringing this structure, count data is a kind of you know data where information should be positive in nature and it should be integer type only that is the only 2 requirements which is actually for the count data. It is a slightly means it is explicitly different from binary data and the kind of you know ordinary data.

(Refer Slide Time: 02:03)



So, now corresponding to you know count data so, the discrete data is a type of data that can be grouped into classification right. It is generally based on you know some kind of you know counts only, only a finite number of values is possible and the values cannot be subdivided meaningfully. Like you know there are 2 things, if the data is not actually positive in character and not in a kind of you know integer type then it cannot actually represent the facts as per the particular you know requirement means, in the case of you know count data modeling.

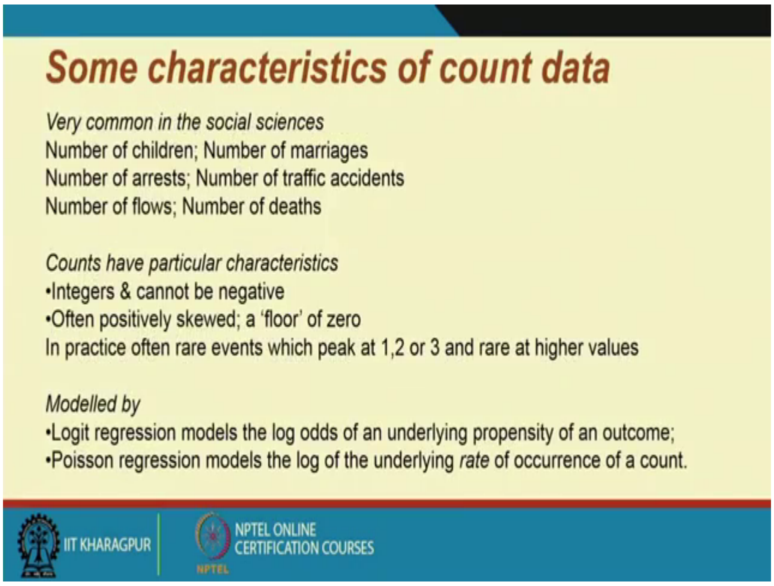
Again, so, another you know constraint to this you know modeling is that you know you cannot so, easily actually divide the samples into you know different subsamples. So; that means so, sub clustering is not so, easy here. So, that means, technically now there are 3 things you have to be very careful.

First thing, the values of the variable should be positive in character, then it should be exclusively in numbers; that means, integer type. And again this informations or this data cannot be you know subdivided into another group because, it will not give you some kind of you know meaningful representation you know kind of you know representation like we have already discussed half death or you know half suicide these are all not the actually clear cut kind of you know information. So, far as you know count data is concerned.

So, similarly you cannot just divide these you know data into different kind of you know clustering that is what you know called as you know subdivision. If the it is the data that cannot be broken down into a smaller unit and will not add any value or it will not bring any kind of you know meaningful kind of you know interpretation.

It is typically you know in a situation where you know we count information like you know whole numbers only. There is no such thing as you know half defect or you know half death or something like that. In brief the particular subdivision does not really add any further value addition again that too. So, far as you know description is concerned.

(Refer Slide Time: 04:47)



Now, I will bring some of the classic examples relating to count data and discrete data. And very common in social science environment or you know engineering science environment number of children's, number of marriages, number of arrests, number of traffic accidents, number of flows, number of deaths. So, these are all you know common examples where you know the information exclusively count data type. Integers and positive in character simply integers and positive in character right. So; that means, non negativity and the information should be in numbers only. So, no fractional information it is exclusively integer type.

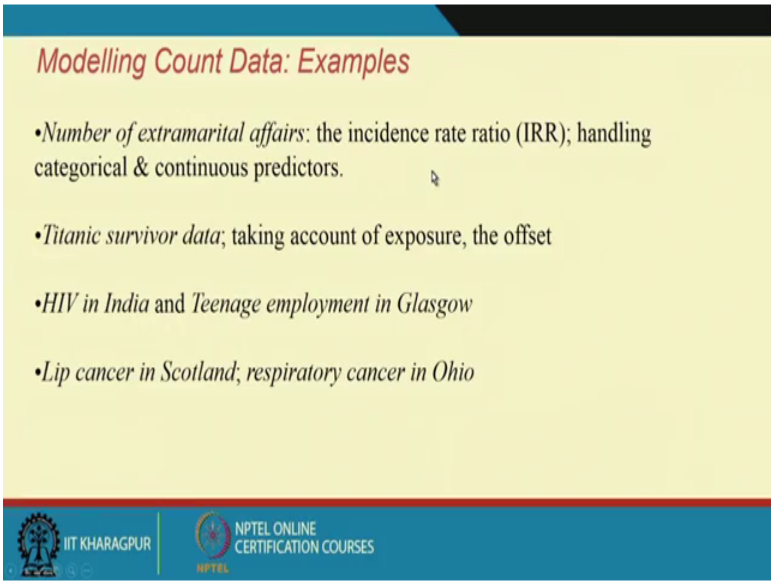
So, count data have you know typical features that is integers and cannot be negative. Sometimes, it may be positively skewed, but in reality we have no issue whether it is a positively skewed or you know something you know exclusively normalized. So, we have no issue, but ultimately we should specifically follow that you know whether you know the structure is actually non negativity in character and integer type. So, far as a modeling is concerned corresponding to the you know count data. So, we have a couple of models that is you know starting with logit regression then Poisson regression. So, we have lots of you know kind of you know variety all together.

But, in fact, logit model to the count data is not so, good. Of course, you know if the particular counting or the particular informations by default will be in a 0, 1 you know structure then we have no issue it may be coming under the count data. But, that may not be the actually generalized case that is very very exceptional case. So, because ultimately we like to understand here variables and the corresponding information to the variable which we call as you know data.

So, we have 2 restriction here so, the information should be positive in character and integer information should be in numbers only that too in integers only. So, if it is actually 0, 1 since, the 0, 1 itself is a integer so; that means, if the information will be only in 0, 1. So, that can be also you know called as you know count data type, but the reverse is not true; that means, you know you cannot just generalize that you know 0, 1 means it is actually the entire integer type.

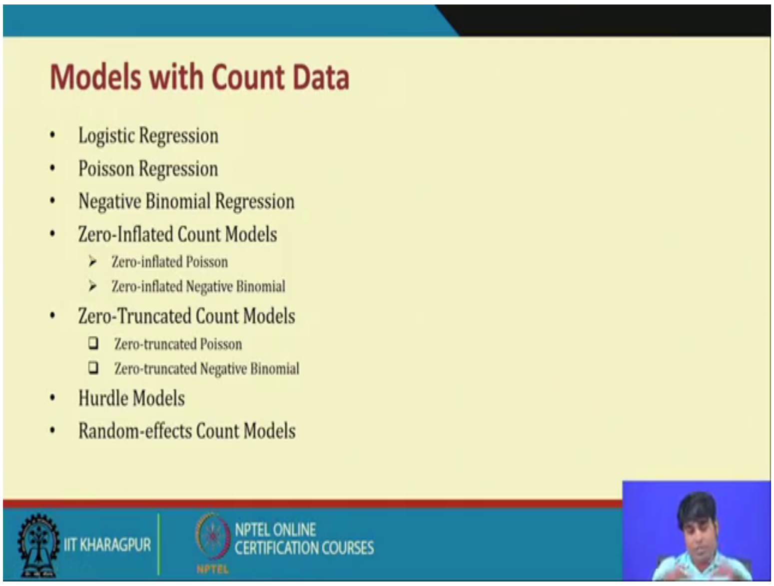
So, it is the special case of you know integers type that is called sometimes you know in the optimization problem we called as you know 0, 1 integer programming. So, here actually also within the integers if it is a 0, 1 kind of you know information only the particular problem is different and we can use you know various types of you know model including the logit models ok. But, if the information is not 0, 1 only then you know you cannot exclusively use this kind of you know binary model, binary chase logit model or the kind of you know probit model. So, what is happening here? So, we can still you know use something like that.

(Refer Slide Time: 08:18)



Before that you know starting these are all actually different kind of you know example say again where we can actually use to understand the count data and the kind of you know modeling ok. So, I will bring these examples after you know highlighting some of the count data model.

(Refer Slide Time: 08:37)

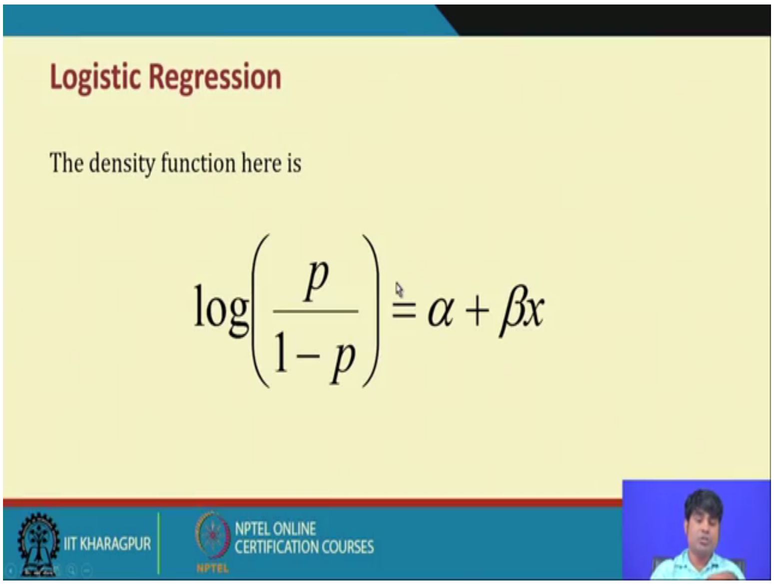


So, far as a classification is concerned that too with respect to count data; that means, regression modeling with count data or model fitting with you know count data. So, it is actually very you know straight forward clustering. So, first one is the logistic regression type where you can use, but provided it should be actually some kind of you know informatic that too integer types and positive one. Then the second model connecting to this count data is the Poisson regression modeling, third one is the negative binomial regression.

We have also models like you know zero- inflated count models and under this we have 2 different forms zero-inflated Poisson distribution and zero-inflated negative binomial distribution then zero-truncated counted models. So, here again we have a actually with respect to Poisson distribution again and the negative binomial distribution. Then we have 2 other models, hurdle models and random effects count models. So; that means, technically these are the complete you know you know model structure. So, far as you know count data is concerned.

So, now, now 2 things are very important first thing the data which we like to use in these models are exclusively count data type, that is the values of the variable means values of the variable means it is a kind of you know information that toward data is a positive in character and should be integer in nature. Now, we can see how we can actually use these models. So, far as you know regression kind of you know structure is concerned and the kind of you know particular engineering requirement is concerned.

(Refer Slide Time: 10:38)



So, starting with you know simple logistic regression this is what the means here every times means whatever models we are going to discuss all has actually specific you know density functions right. So; that means, logistic has it is density functions probability density functions then Poisson probability density functions, negative binomial density functions.

So; that means, technically we have a particular functional form and we can just connect to that particular function form and then go for the estimation as usual you know flow. Of course, you have to check the balance if it is actually non-linear in character then you know you may you may have to think about whether to use OLS technique or you know maximum likelihood estimate technique. If it is linear you know kind of you know structure.

For instance, let us say binary choice model simply then you can go for you know simple OLS kind of you know estimation, but if you say starting with Poisson distribution and binomial negative binomial distribution. So, in this kind of you know modeling. So, it is a non-linear in character and you know maximum likelihood estimation can be deploy there.

But alternately, what is more important in this kind of you know modeling framework that too data modeling in other words regression modeling with count data so, is the density functions. Because, we strictly use you can say you know for 5 you know different kind of you know modeling structures starting with you know simple logistic regression to Poisson distribution to the negative binomial distribution and so, on.

So, every times you like to check how is the density function and what kind of you know data and how is the problem flow and how to start the process. It is as usual you know simple you know regression modeling or simple time series modeling or panel data modeling. But, ultimately it is the means here it is the move is actually with respect to the data and the kind of you know connecting point is the density functions.

So, if the data is positive in character and integer type then you check you like to check the flow once you go through the problems you can understand whether it is the Poisson type distributions or it is the negative binomial distribution type. And accordingly you can bring that density function and then connect with this you know modeling and finally, you can you can analyze the problem as per the particular you know requirement.

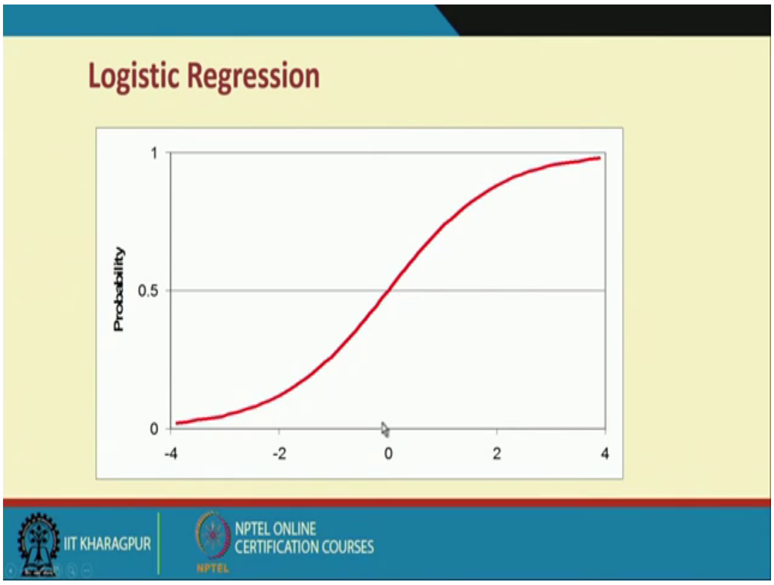
So, ultimately 2 things very important, so, the type of you know data and that too you have to think whether this particular you know data structure of course, it is with respect to positive in a information and non negativity and then integer type then again you think actually what you know distribution exactly it follows.

So, whether it is a Poisson type then you use starting with you know Poisson regression modeling, if it is actually normal flowing with you know binomial distribution then you can start with you know negative binomial regression modeling. So, like that that is what the actually checkpoint and the kind of you know turning point. Rest of the process is as usual actually you know similar.

So, now starting with logistic function this is how the estimation and here actually the problem is actually the kind of you know estimation. Of course, if you find out the odd ratio and after getting the odd ratio means technically if you start with actually p value here then the requirement of the p values is that you know it should be within 0 to 1. But, ultimately if you start with you know count data modeling that too in a kind of you know means logistic model with count data framework then the odd ratio should be actually integer type integer type. And then the kind of you know the kind of you know structure through you can go for the estimations.

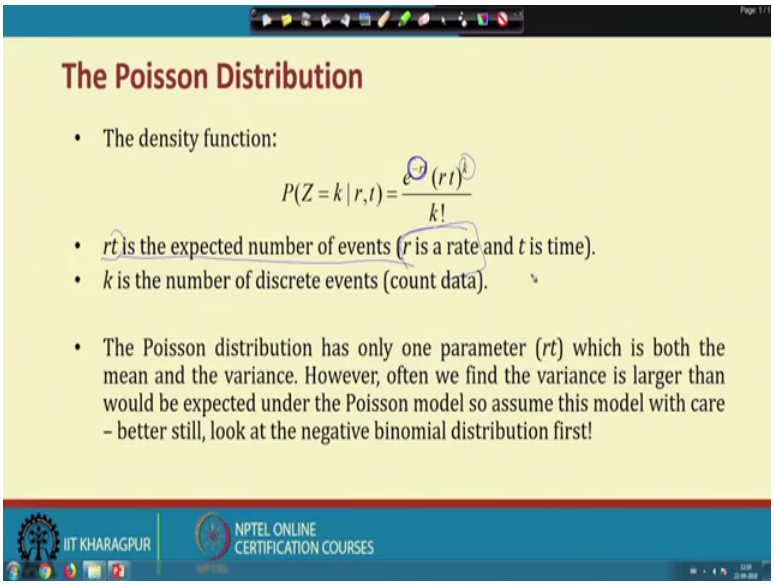
But, in a reality a most of the count data models you know means most of the instances the particular framework we will follow you know Poisson distribution type and the kind of you know binomial distribution type. And then we may have some situation where there is a clear cut integration between binomial distribution means Poisson distribution in binomial negative binomial distributions.

(Refer Slide Time: 15:20)



So, this is how the structure of the logistic functions looks like you know the graphical look.

(Refer Slide Time: 15:27)



And let us you know start with the Poisson distribution, this is one of the heart of the count data modeling. So, what I have already mentioned this kind of you know modeling most important is the density functions with means which kind of you know models you like to finally, use depending upon the type of data and the kind of you know natures that too non negativity and the integer type.

So; obviously, the one of the such kind of you know modeling is the flow of you know Poisson distributions and the density function that too for the Poisson distribution is like this.

So, it is actually exponential type of things e to the power minus lambda lambda to the power k by k factorials that is what the exact actually a flow. So, what is happening here it is on how this actually the density function is all about right, like we have already discussed logistic functions and normal distribution functions for you know probit model. So, it is also similar kind of you know structure it is not highly different.

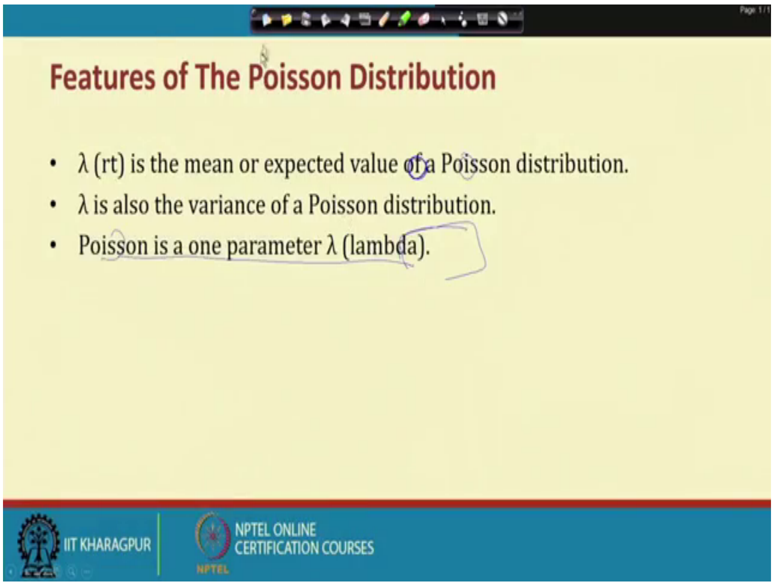
So, here r t is the expected number of you know events that too this one and k is the number of you know discrete events that is what the count data. And then we have the we have here actually let us bring here that kind of you know concept; that means, here technically. So, this is a one of the you know which is very important in the Poisson distribution minus r t sometimes we call as you know lambda and then you know this k k component ok.

So, r t that is the lambda is the expected number of events that is the that is calculated with the help of you know r and t that is the rate with respect to time and k is the number of you know discrete events that is what the count data is all about. Now, so; that means, how the mean operates you know within a number of you know counts or the kind of you know discrete events. So, Poisson distribution has only one parameter that is the r t that is the lambda which is which is both the mean and the variance; that means, technically it is a kind of you know special kind of you know distributions where mean and variance are you know coincide.

So,; however, often we find the variation variance is larger than you know sometimes you know it may be larger it may be lesser with you know with mean actually. In that case Poisson distribution have some kind of you know issue so; that means, ok. So, the technically 2 things are very clear here you know before you start applying this kind of you know modeling. So, your first thing data should be very specialized that too count data types where information should be positive and integer.

And again while you know doing this one so, you know you know it is the kind of you know parameters that too lambda and we like to check what is the game between mean and variance of this parameter. If the mean and variance of this parameter are actually coincide then the particular choice means we have to make a choice about the Poisson you know Poisson regression modeling. if mean and variance are not equal then you can go for you know negative binomial distribution that is what the turning point about the choice of these 2 models. So, far as you know in the first instance the count data modeling is concerned. So, now, what is happening here you have to go to this you know structure.

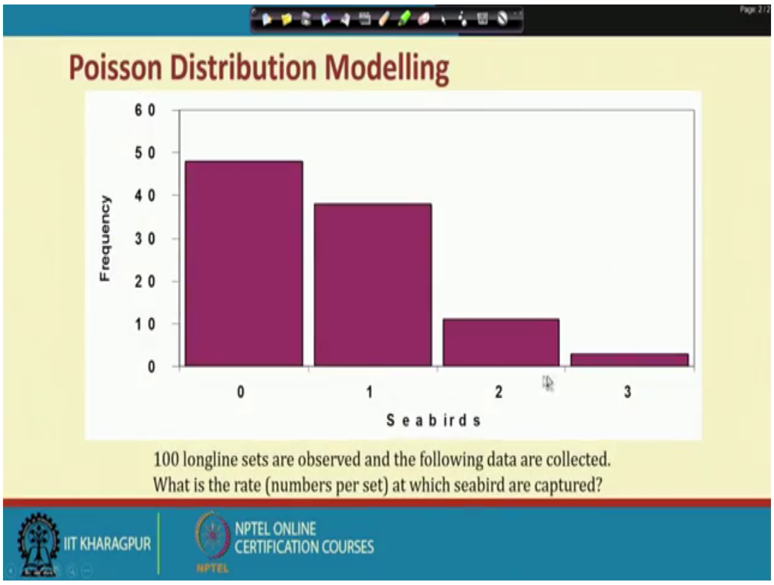
(Refer Slide Time: 19:29)



Then so, what is exactly you know situation here which I already highlighted. So, lambda is the you know kind of you know Poisson distribution parameter and we like to find out you know mean of this parameter and variance of this parameter. And we like to check whether you know this mean of this parameter and variance of these parameters are coincide or not that is the first checkpoint you have to do.

Of course, the first checkpoint is the data structure, then second checkpoint is the mean and variance of the parameter and then it will take you to the clue whether to move towards the Poisson regression modeling or negative binomial distribution modeling. So, now you will find you know mean variance of this parameter will not in the same then you know by different you will you will differentiate and you have to choose a particular technique as per the you know requirement.

(Refer Slide Time: 20:32)

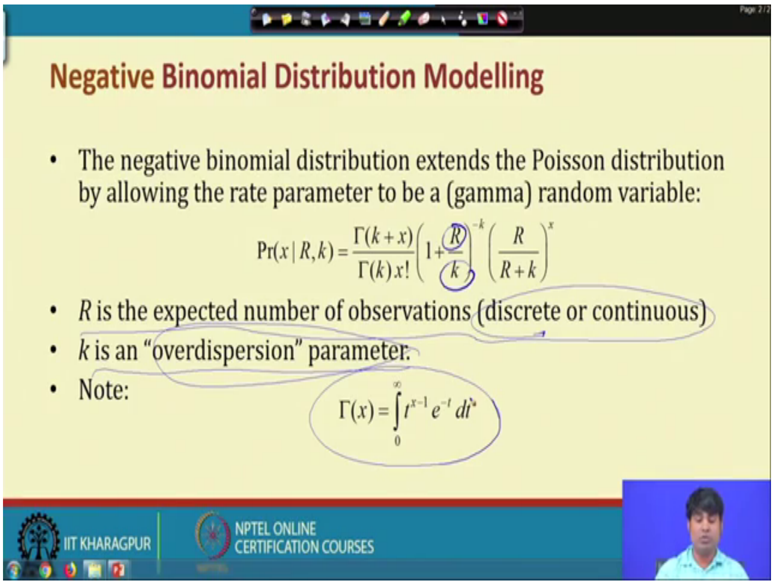


So, now moving towards this I will just give you the you know simple examples how what are the you know situation of the Poisson distribution. So, you see here so, the Poisson distribution the graph will be actually non-linear in character a slightly somehow you know exponential type because, the model a model if you look to this you know models here it is you know starting with you know exponential type. So; obviously, if you plot then the structure will become like this and that itself we will see whether you know you are in a position to use you know Poisson distribution or not.

So; that means, like as usual regression modeling we like to choose a particular functions. What we can what we supposed to do actually we go for you know graphical check which give which can give you some kind of you know indication about the exact you know functional forms whether to fit with you know linear structure and or non-linear structure. Or even if in the non-linear structure what type of you know functional form you have to bring.

So, so here in the count data modeling we can also follow the similar kind of you know structure in you know just check how is the behaviour and then you can actually go for the particular choice of the modeling and then you can go for the estimation.

(Refer Slide Time: 21:50)

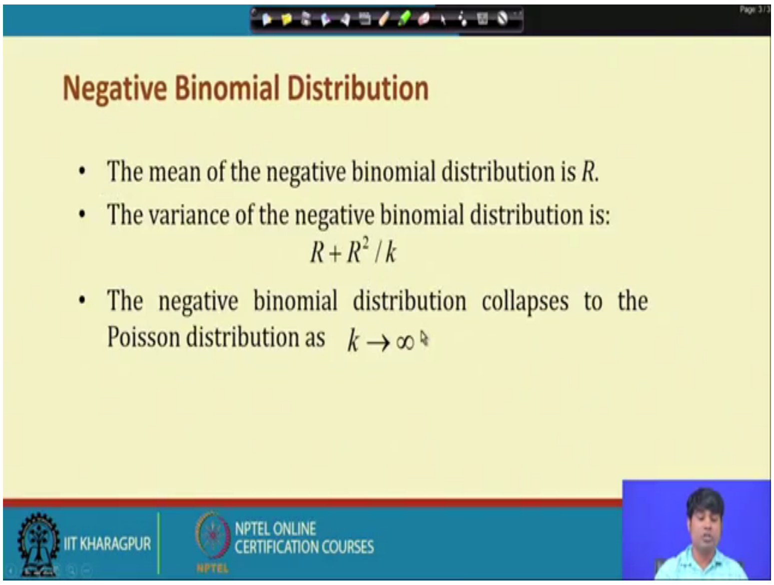


And the second model related to the a count data modeling is the negative binomial distribution modeling. And it is just you know extensions of you know Poisson distribution. And here we allow the parameter to be a instead of you know lambda here we call as you know gamma and that is one of the you know parameter for you know negative binomial distributions. And hear the issue is actually with respect to 2 things R and r and k corresponding to their you know lambda which R and t small R and t.

So, rate with you know time so, here also r is the expected number of observation that is called as you know discrete or you know continuous. And k is the overdispersion that is very that is very actually a big issue in the kind of you know negative binomial distribution.

So; that means, technically so, the it is not a uniform kind of you know variations we call as you know dispersion means you know if the variance are not actually equal over the time or something like that. So, if it is not you know equal or you know not constant or not stable then by default we will come to you know follow the binomial distribution modeling rather than you know Poisson distribution modeling. So, the gamma function actually follows the you know the kind of you know structure like this and then we will actually go for the kind of you know estimations.

(Refer Slide Time: 23:28)



So, again in order to understand actually you know the concept negative binomial distribution. So, here the mean of the negative binomial distribution is r and the variance of the negative binomial distribution is R plus R square by k. So, which is actually the differentiating point between the Poisson distribution and binomial distribution; in the case of you know Poisson distribution the mean distribution will coincide that is actually called as you know lambda.

But, here the gamma gamma parameter that too you know the mean variance are not actually same. So, it is actually a corresponding to the this you know functional form which we have already highlighted here R and k and that too with this particular you know exponential type function.

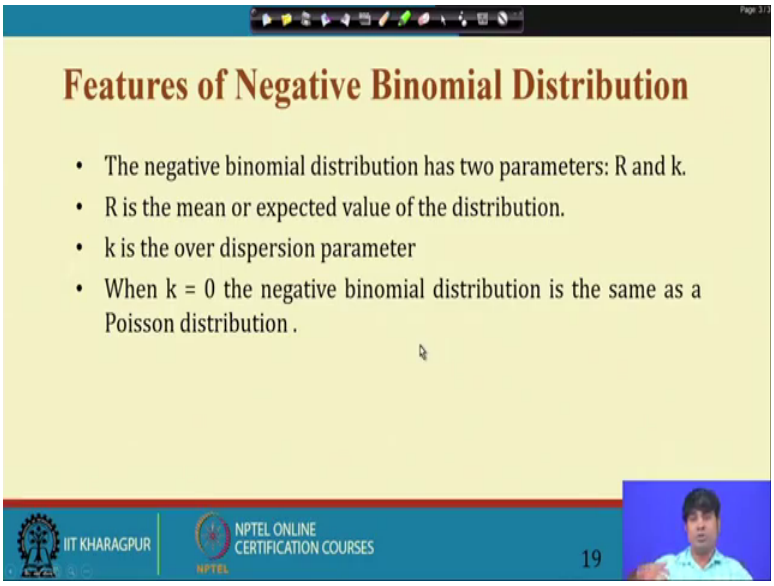
So, we have the mean called as you know R and then the variants of this particular distribution is called as R plus R square by k. And since the mean and variance are not same of course, the distribution density function is different, but by the way when we will calculate you know mean and you know you know variance that is the expected value means expected values of this happenings and the variance of this particular happening.

So, if they will not coincide then by default we declare that you know it is the it is the pattern of you know binomial distribution otherwise, it is the kind of you know flow of you know Poisson distribution.

The negative binomial distribution you know can you know converts to the Poisson distribution if you now k stands to you know infinite. So, that is that is the kind of you know the structure like you know we have discussed the concept called as you know central limit theorem where you know if you start you know increase you know you know sample size; that means, when n tends to infinity every distribution will converts to normal distribution.

So, this in this case also it is similar kind of you know flow, when k will start you know moving or you know approaching towards you know infinite. So, then this binomial distribution negative binomial distribution will converts to the Poisson distribution. So; that means, technically.

(Refer Slide Time: 25:47)



So, if your you know data set is very huge and huge. So, ultimately so the difference between the 2 model will not so, much you know significant. So, but ultimately there is no harm if you will start with you know building you know regression modeling with you know Poisson density function and negative binomial density function.

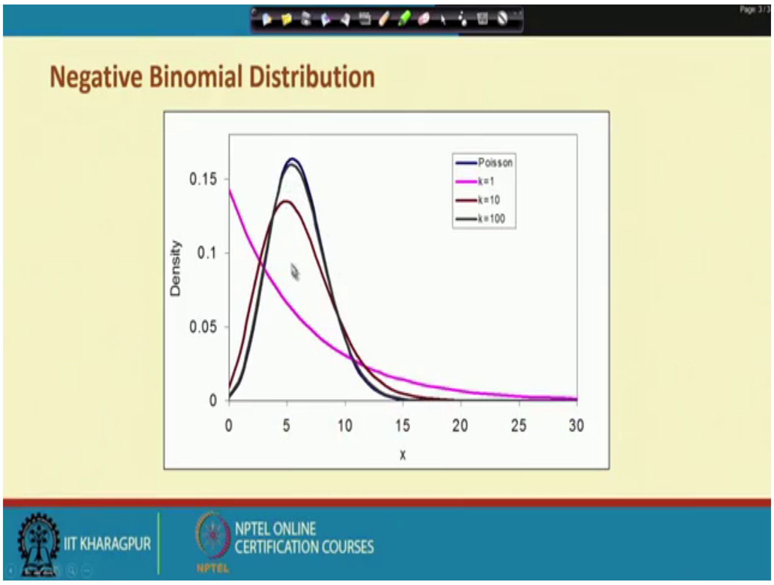
Ultimately, we have no actually clear idea where you know where the n will be approaching the infinite; that means, technically you can increase you know sample size subsequently then ultimately at a particular point of time they will actually converge. But, if that is not the case; that means, if we the sample size is lesser to that particular situation then by default the model can be is a with you know particular choice only either you can use Poisson distribution or you can use actually negative binomial distribution.

So, both cannot go simultaneously because, in the case of you know Poisson regression modeling one of the indication that you know the mean of the parameter and variance of the parameter should be same, but in this case. So, the mean of the parameter and variance of the parameter are you know different that is what is called as you know overdispersions.

And in the case of you know means in the situation where n approaching infinite then you know there is high chance that you know they can converge; that means, technically we start with assumption that you know mean and variance are not equal. And as a result you can start with you know using normal you know negative binomial negative binomial distribution that too negative binomial regression. And then when you know starting you know in you know start increasing sample size and then ultimately you will find the deviation or the difference between mean and variance over the time will be minimized and then coincide.

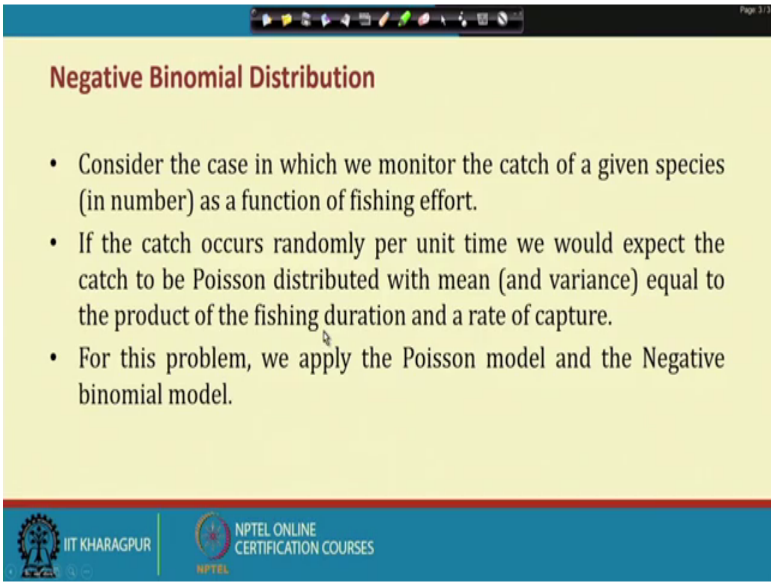
So, in that case so, the particular distribution will converge to Poisson distribution and then finally, you can actually use Poisson regression modeling. That is actually kind of you know simulated process or you can call as you know one kind of you know sensitivity structure in the regression modeling, but ultimately if that is not the case then we have a 2 different models all together. So, one kind of you know model is the Poisson regression, another kind of is a negative binomial distribution and the kind of you know.

(Refer Slide Time: 28:26)



Function if you graphically plot the behavior will be look like this and by default it will give the signal that you know you will go for you know negative binomial regression.

(Refer Slide Time: 28:35)

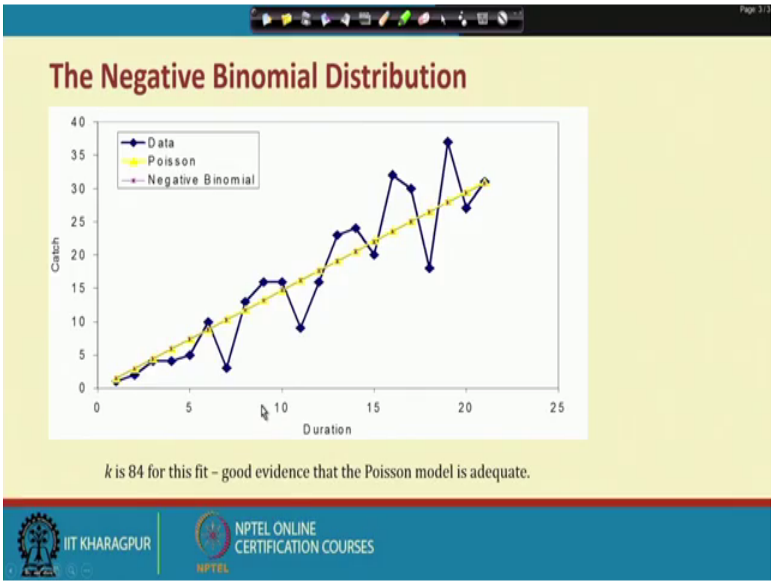


So, now considering the case which we call as you know, the kind of you know binomial structure. The catchment is that you know what is the kind of you know occurrence point or you know the kind of you know durations and the kind of you know rate of capture that is how the big deal between the mean and variance.

For these particular problems we apply the Poisson models and the negative binomial models, but ultimately; that means, technically you know both models will be you know very close to each other like you know we have the closeness between logit model and probit model. because, there it iself also when sample size is substantially very high and a high then the difference of the estimated outcome or output will not actually so, much whether to use actually logit model or probit model.

The same thing also here when your sample size is substantially very high then the outcome between a Poisson regression modeling and binomial regression modeling will not actually so, high so, that is that is actually big deal.

(Refer Slide Time: 29:52)

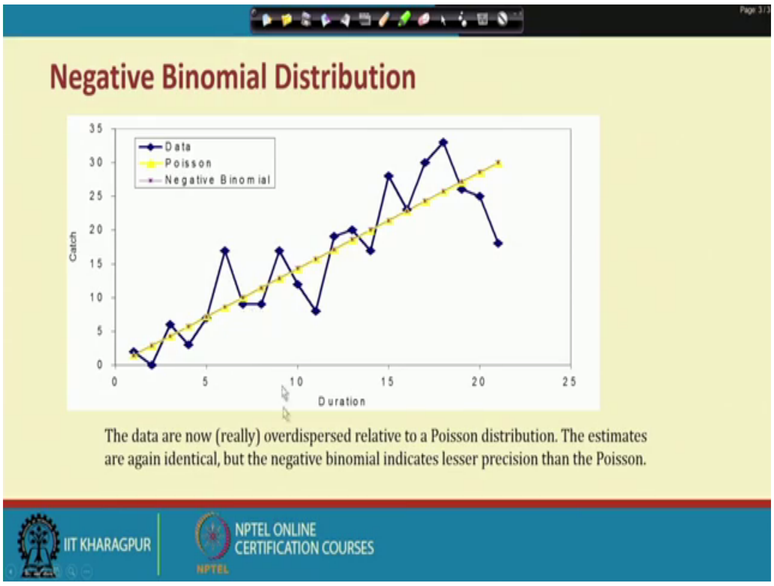


That means it clearly gives the signal that you know when you are actually the information is you know having high then ultimately you have lots of flexibility. So, that times actually either you can use you know say you know Poisson regressions or you know binomial negative binomial you know regression like we use actually logit or you know probit. So, we have no issue, but when your sample size is very small one even if it is follows with you know count data patents still you know you have to make a choice. If it is very less sample then by default you cannot use both the model simultaneously.

But, if the sample size is substantially very high then you can use both the model simultaneously that is that is how the a big deal or you know big understanding. So, this how the kind of you know little bit you know differentiating point between you know

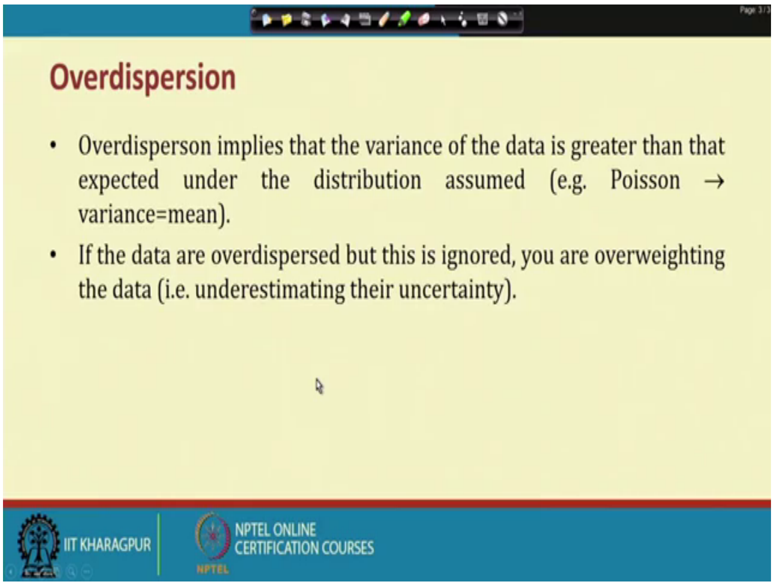
Poisson distribution negative. So, the yellow mark is following the Poisson distribution and the kind of you know blue one is actually showing give an indication about the negative binomial distribution ok.

(Refer Slide Time: 30:59)



So, likewise there are different kind of you know flow here.

(Refer Slide Time: 31:04)



And that is how the kind of you know means it is exclusively which you can call as you know overdispersions which implies that you know variance of the data is greater than that expected under the distribution; that means, technically it is the mean actually.

So, if the variance is a somehow you know greater than to mean then that is declaration is the overdispersion say that too the situation where we can use actually a negative binomial regressions. And if the data are you know overdispersed; that means, about that is ignored you are you know means this kind of you know underestimating the their you know uncertainties.

So; that means, technically before you start the process about this kind of you know count data modeling. So, we have 2, 3 checkpoints. So, the first checkpoint is the kind of you know information that too data which would like to check that whether it is actually a exactly count data type where we have where the checkpoint is again the negative non negative values and then the integer type values.

And it will be more interesting the result will be more interesting and the model will be more interesting if it is not exactly the 0, 1 it may be you know above than 0, 1 and that is how the count data starts you know bringing actually the kind of you know reality. And then the second checkpoint is actually the parameters value which is the differentiating point between the 2 different models that too Poisson regression modeling and then the negative binomial distribution modeling.

So, where one case mean variance will be equal say then the you know clear cut choice will be the Poisson regression modeling. If a the mean and variance are not equal then the situation will be called as you know overdispersion and in that case the particular choice will be the negative binomial regression modeling.

So, that is how the difference and so, far as a count data model is concerned. So, you can actually one you can use the Poisson regression modeling, the other way you can use the negative binomial regression modeling. So, the choice ultimately depends upon the type of data and the parameters actually reaction that too with respect to mean and variance.

And in the next lectures we will discuss some of the problems; that means, we will bring some kind of you know live problems that too application where we can use this Poisson regression modeling. And, the kind of you know binomial regression modelling negative binomial regression modeling. With this we will stop here.

Thank you very much, have a nice day.