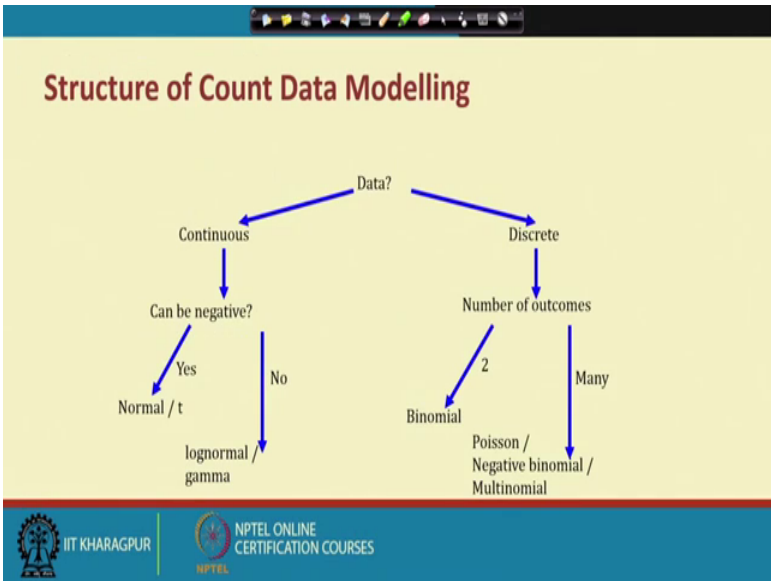
Engineering Econometrics
Prof. Rudra P Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture - 59
Fitting Models to Data (Contd.)

Hello everybody, this is Rudra Pradan here. Welcome to Engineering Econometrics, today will you continue with count data and discrete modeling. In the last couple of lectures, we have highlighted this concept that to understanding the Count Data and Discrete Data. And accordingly we have also identify various models relating to count data and discrete data. So, we can directly go to various structure of count data modelling.

(Refer Slide Time: 01:05)



So, basically the structure is like this; so we have the concept called as count data modeling, and the starting is actually to understand the data.

In the last couple of lectures, we are strictly discussing various models relating to data structure starting with you now time series models, then cross sectional models, then panel data models and then count data models. So, in the count data models we specifically highlighted some of the structure that is, with respect to data types. That means, here the type of data is like this that we have continuous structure; we have discrete structure that is what here.

So that means technically, so the data division will be continuous type and then discrete type. And it can be negative while in the continuous structure and if it is really negative then we will follow the normal TDS tuition, and if not then lognormal and gamma distribution.

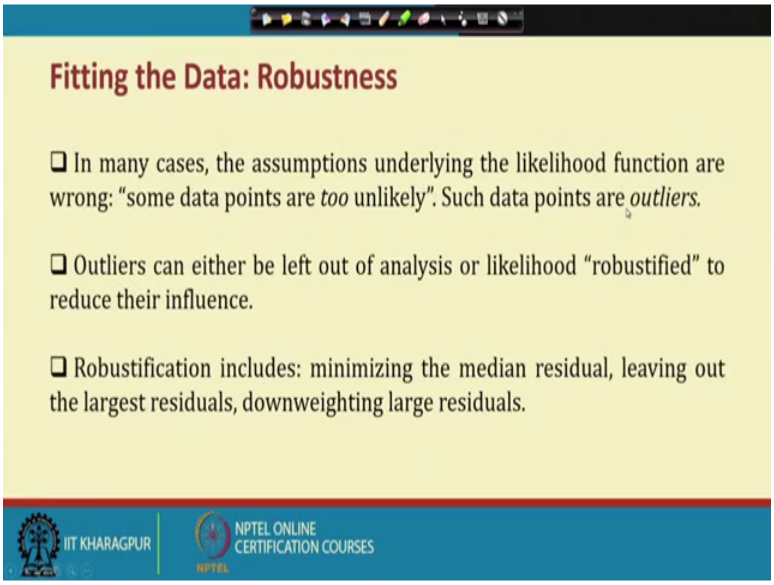
On the other side we have discrete structure and that to number of out comes. If it is with respect 2 that is what which we call as Binomial, and if it is many it is called as Poisson's or Negative Binomial or Multinomial. So, that is what the typical structure about the kind of count data modelling, but what is happening here is; we are not going in details about this modelling structures.

We are strictly discussing this aspects only that to playing the game between you know Poisson distributions and Negative binomial distribution. Of course, we have highlighted these 2 models in the last lecture and again we are bringing this particular structure to connect certain examples, where we can use actually Poisson distribution and then negative binomial distribution.

Accordingly, we have 2 different model that we need to highlight under the count data modeling, that is Poisson regression modelling and the binomial regression modeling. In the choice of these 2 models which we have already highlighted this basic structure starts with calculating mean and variance. If mean variance are equal, then we can actually use Poisson distribution or Poisson regression modeling.

If mean variance are not equal, then we may go for negative binomial regression modeling. That is how the beginning of this entry or that to with respect to count data modelling.

(Refer Slide Time: 04:23)



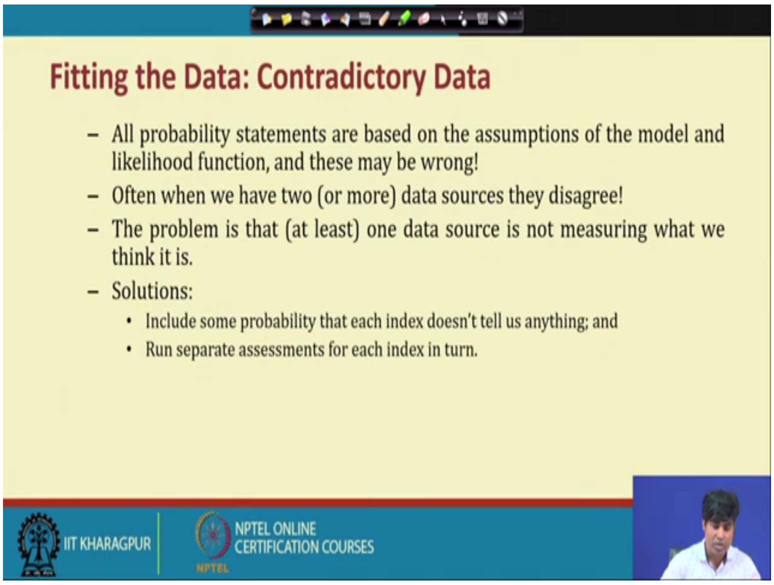
So, what I like to highlight here is I will bring some of the examples, where we can actually connect with you know this a kind of Poisson regression modelling and the kind of negative binomial regression modelling. Before that I like to bring one particular concept; since we are discussing about the fitting models that to relating to count data. So, we have couple of issues here, one is which we have highlighted is called as over dispersion.

And then this is called as robustness; in many instances this assumptions underline the likelihood function are wrong; some data points are too unlikely so that data points are outliers. That means, technically means we have already discussed this concept you know that is the outlier in the header scarcity problem.

It is a very dangerous kind of component which also having actually problem in this count data modeling. Outliers can be either be left out in the analysis or like likelihood you know robustified to reduce their influence; that means, we have lots of instances.

Since, we have lots of possibilities where we can reduce the impact by normalising the data or some kind of restructuring of the models or something like that. But ultimately what is actually requirement there should not be outlier in the final set up. So, that the model may not be actually creates any kind of problems of for us, prediction is concerned or the kind of forecasting is concerned. So, it so the requirement is that minimizing the median residuals, living out the largest residuals and down weighting large residuals this is how the kind of structure.

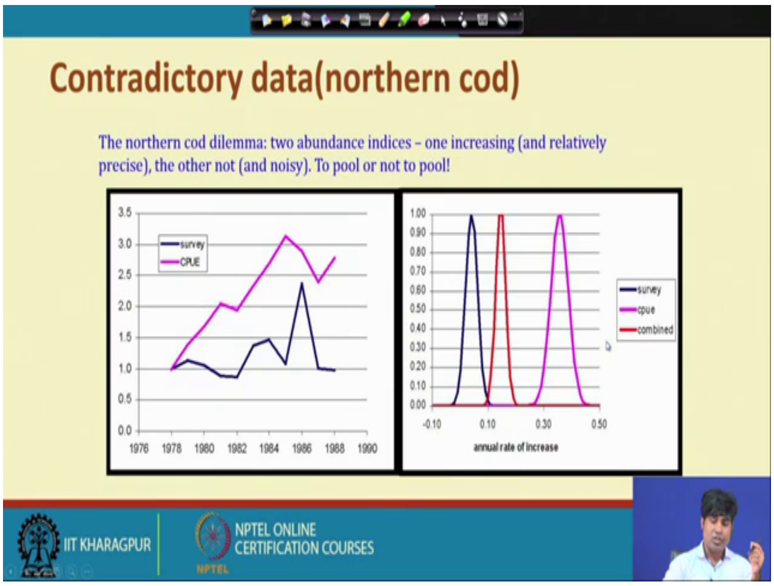
(Refer Slide Time: 06:21)



So, against contradictory data is that all probability statements are based on the assumption of the model. And likelihood function and these maybe actually wrong; upon when we have 2 data sources they disagree.

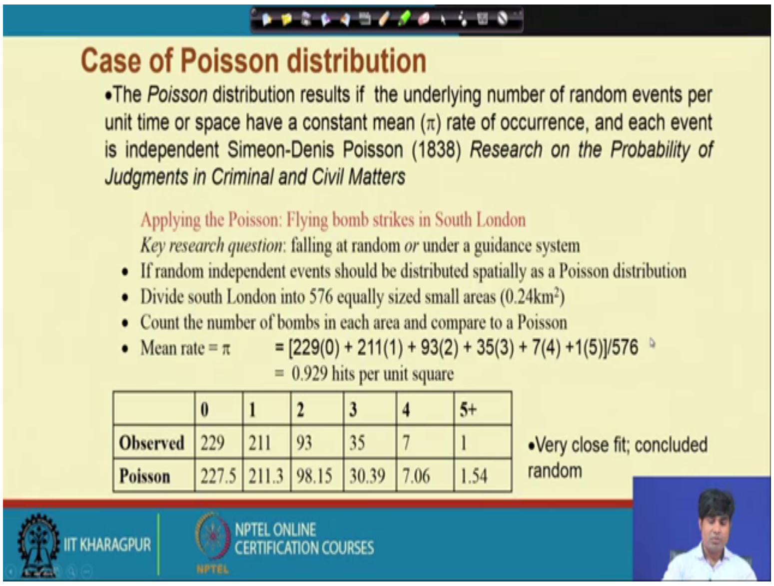
The problem is that one data source is not actually measuring what we think it should be. What is exactly the solution that include some probability that each index does not tell us anything. And along separate assessments for each index in turns; that means, we have a kind of different structure of in handling the data and then the kind of modelling.

(Refer Slide Time: 07:02)



So, let us have a look here a contradictory data the one part of this particular structure follow kind of simple patterns and the other one is actually having some kind of different. So, that means, actually one way you can say that one increasing the others not. So that means there is a kind of huge obstacles in the setting up the data and the kind of modelling.

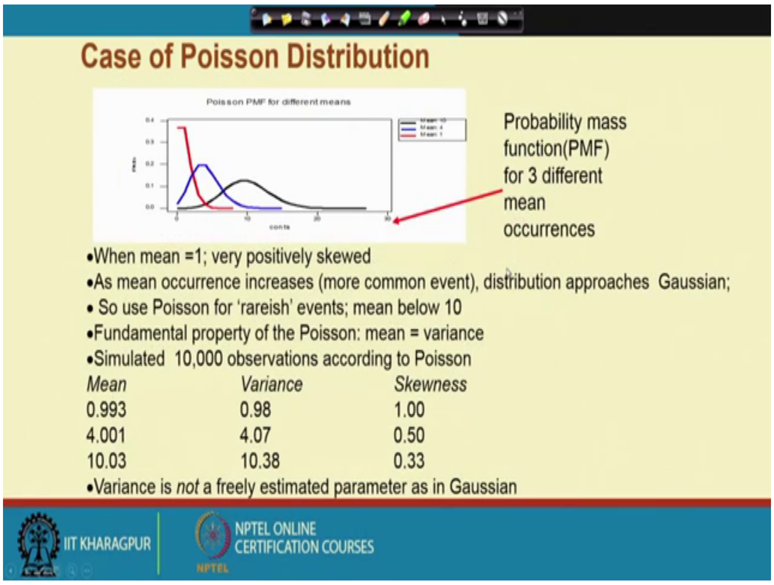
(Refer Slide Time: 07:34)



So, let us bring some kind of cases relating to Poisson distributions or Poisson regression modelling and then the binomial regression modelling. So, here actually we have some sample occurrence in the observe frequency. So that means, technically so how we actually apply here these are all observed structure corresponding to the possible values.

So; that means, what is happening here; the key research question is that falling at random or under the guidance systems, if random independent events should be distributed specially as a Poisson distribution. Then you have to follow the particular structure by calculating mean of this particular series and then variance of this particular series. So, after calculating the mean so we will find here 0.93 having the particular data availability.

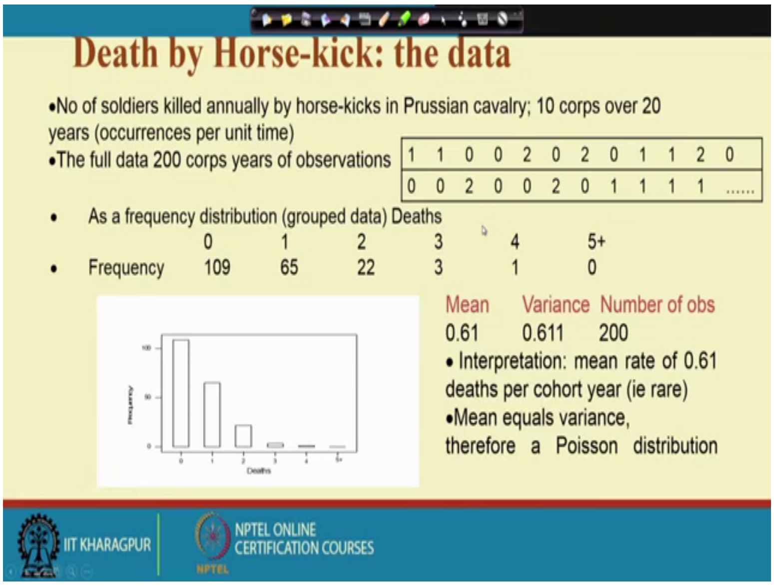
(Refer Slide Time: 08:34)



Now if you go to next slides, where you can actually report the kind of variance; sometimes what is happening it may not be closely tie, but if it is a slightly converge to each other the equal the equality may not come because of round of errors.

But somehow if it is actually equal and converge to equality by mean and variant, then we can easily apply this kind of regression modelling.

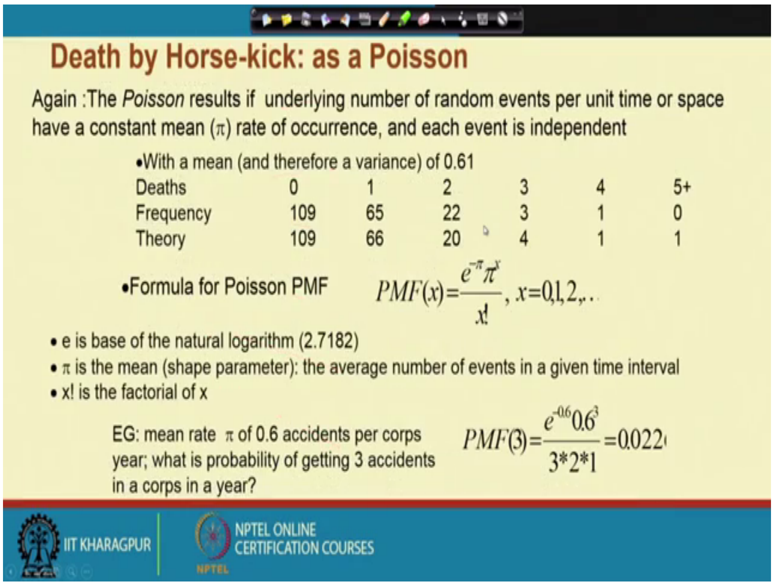
(Refer Slide Time: 09:00)



So, similarly there is another kind of situation here number of soldiers killed annually by horse kicks. So, here we have actually data structures structures that is, that is actually with respect to frequency distribution and that to group data of death occurrence. And the frequency corresponding to each frequency available here like these ones. So that means, technically again so this is what the frequencies and the corresponding observations, and then we will again calculate mean and variance and corresponding to this total observations.

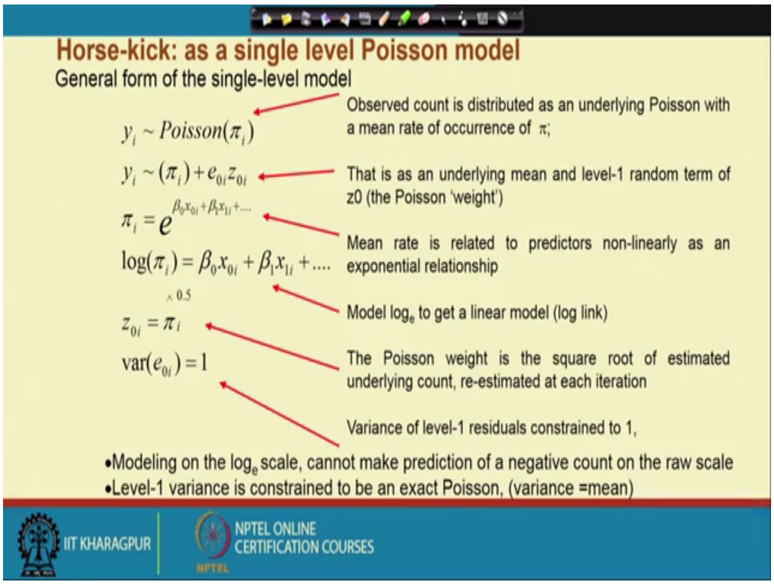
And then we come to we come to conclusion that since mean and variance are equal then you can easily again apply the kind of by say regress regression modelling that to with respect to Poisson distribution.

(Refer Slide Time: 09:56)



Now, or likewise another kind of examples here similarly see the deaths occurrence frequency. So, then again is the structure is actually to connect with Poisson distributions, and then check whether the mean, and variance are you now coming each other. If they are same then again you can go for Poisson regression modeling, if not then you have to go back to the binomial negative; binomial regression modeling.

(Refer Slide Time: 10:25)



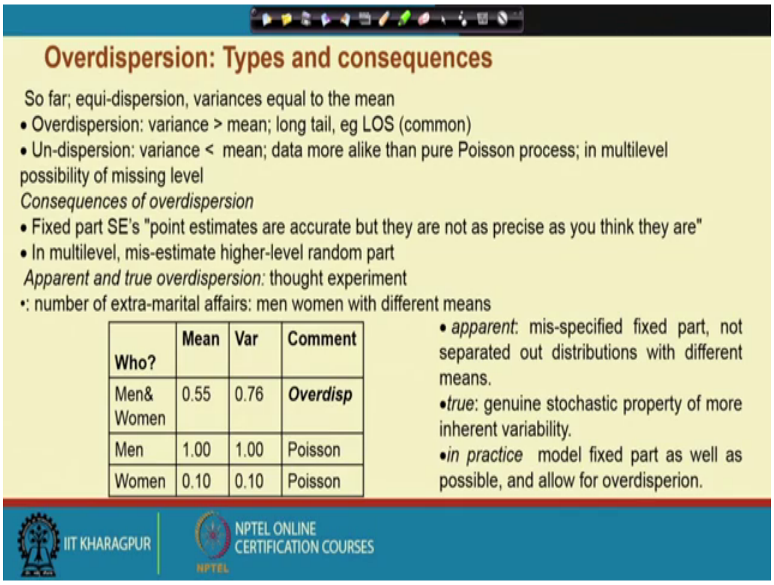
So, basically the Poisson regression modelling is a like this. So, it is exclusively you know exponential type which we have already discussed in some extent in the case of regression modelling with you know different functional form; that to the game between linear regression modelling and non non-linear regression modelling.

So, here the these kind of models what we are discussing you know related to Poisson regression modelling in binomial regression modeling, where actually we are actually strictly focusing on you know non-linear regression modelling. So, the only difference we are bringing here is with respect to you know data structure only because here the data structure is that the similar kind of model.

But the type of data the information contents is relating to the variable must be a count data that to integer types and then avoiding the kind of negativity; that means, we need actually simple non negativity integers only. If the data structure is like that then by default we will go for Poisson regression modelling and the kind of you now negative binomial regression modeling; subject to check point of the mean and variance if they equal then we will move to Poisson modeling, if not then we will go for negative binomial modeling.

So, it is similar kind of flow here. So, this is how the kind of structures which you like to apply in the case of count data modelling.

(Refer Slide Time: 12:11)



So now, coming to this particular structure again we have a kind of structure what we called as double over dispersions, that is the case where you know mean and variance are not actually is same; that means, that there is a question of over identify case. So, again so we have to you know the same problems may be whatever we have discussed earlier.

Of course, we have sited you know problem which where the case is actually mean exactly equal to variance because our intention is to connect with the Poisson regression modelling. And here again our intention is to connect with you know negative binomial regression modelling. So, as a result we have to bring one such problems where we like to justify or bring the scenario that, you know the mean and variance are not actually equal.

So, if it is actually equal then will we go for Poisson distribution or Poisson regression modelling if there you know mismatch then we will go for some kind of other structure, what we call is a negative binomial regression modelling.

So, the classic example is here is where you now, if you clop these 2 samples that is the men groups and women groups. So, where we find mean and variance are unequal that is the case of over dispersion. And if you go by clustering that men clusters and women clusters separately what is happening here so mean and variance. And again mean and variance all are same so; that means, if you go for; that means, technically in this kind of problem.

So, we have 3 different kinds of models if you have a clopping and that to group samples. So, then we can use actually negative binomial regression modeling, and if you go by you know sample specific men separately and women separately corresponding to the data and the kind of inference between, mean of this particular series and variance of these particular series then you can apply Poisson regression modelling.

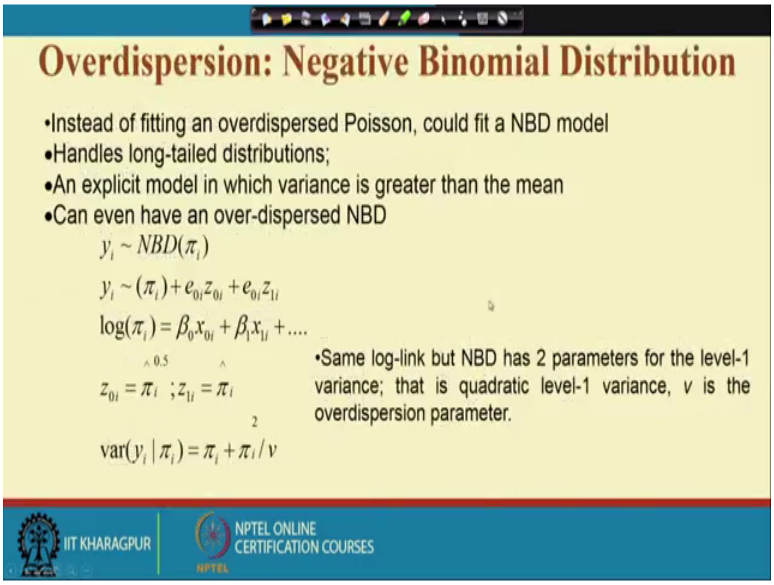
That is what the interesting about the count data modeling. The same dataset if pull then you will have actually structure where we can apply negative regression; negative binomial regression modelling and again if you go by clustering. So, you can apply actually Poisson regression modelling, but ultimately this kind of modelling the clue is actually somehow same.

Of course, we have already discussed the case that where if you increase actually a sample points you know indefinitely; that means, technically n stands to infinite, then by default they will eat equal each other's; that means, there is no such big difference between Poisson distribution and negative binomial distribution.

Of course the where density function of Poisson distribution and negative binomial distributions are different, but ultimately the outcomes of regina regression modelling is concerned followed by density function may not actually so much difference; which we have already highlighted in the case of dummy modelling that to in the context of non-linear regression modelling with the help of logic and profit the same structure, will follow here with respect to Poisson regression modelling and binomial regression modelling.

Now, to continue again so this is the case where we can apply both the model simultaneously.

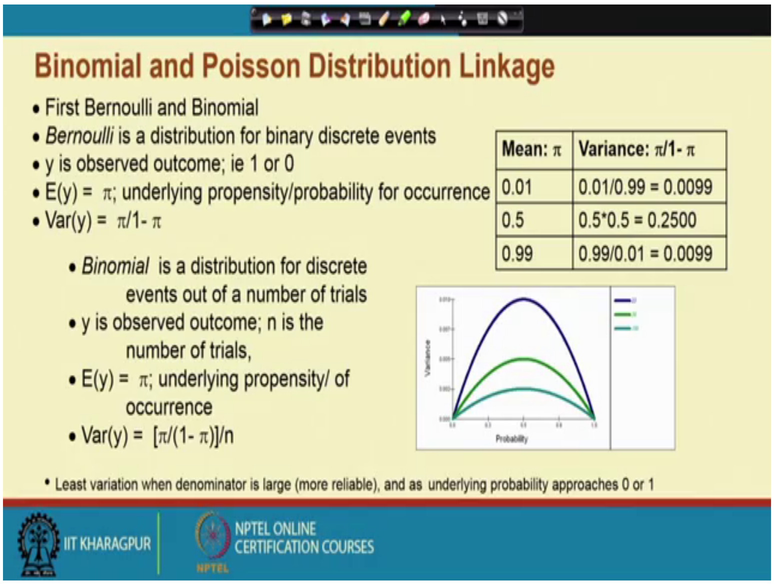
(Refer Slide Time: 15:41)



And they this is how the structure and in the case of negative binomial distribution which is in the case for over dispersion the same structures. We have to start with you know log of this density function subject to all the independent variables which we have already highlighted here that is the case here and same way.

So, the typical variance vector is here, and here the link is that the variance and you know mean are not same. So, variance is greater than mean so that is how the over override infrastructure and by default we will use actually negative binomial regression modelling ok.

(Refer Slide Time: 16:24)



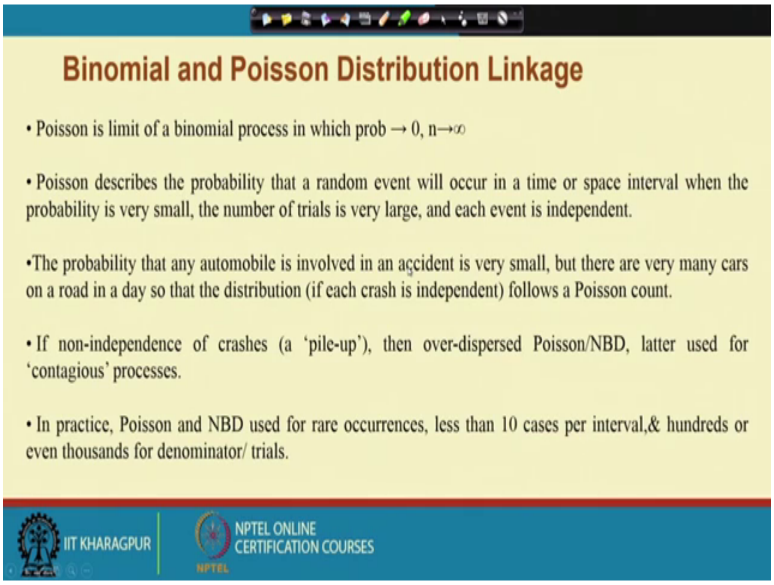
So, that is the case where we apply negative binomial distributions. Now we will highlight you know some kind of linkage between binomial and Poisson distributions in that to binomial negative binomial regression modelling with Poisson's regression modelling. So, somehow actually mean and variance again so means you see it is a game is mean and variance.

So, if you increase the sample size than by default they converge each other, there is a high chance and as a result then both can be actually tie. And ultimately our objective is it not to you know check whether to use actually Poisson regression modelling or negative binomial regression.

Of course that is the requirement subject to you know data structure, ultimately we you know if you increase the data size data size or sample size. So, we have twin you know twin advantage because higher the data or higher the sample size model accuracy will be much higher. And then again is we may be in a position where you now that will be normally distributed and then finally, this kind of in environment both the distribution can close each other.

As a result, we can instead of going for 2 different kind of setups it is better to apply single model and then analyze the situation as per the particular engineering requirement. That is how the deal and this is, kind of examples where we have actually calculated the kind of mean and variance and then justifying the kind of requirement.

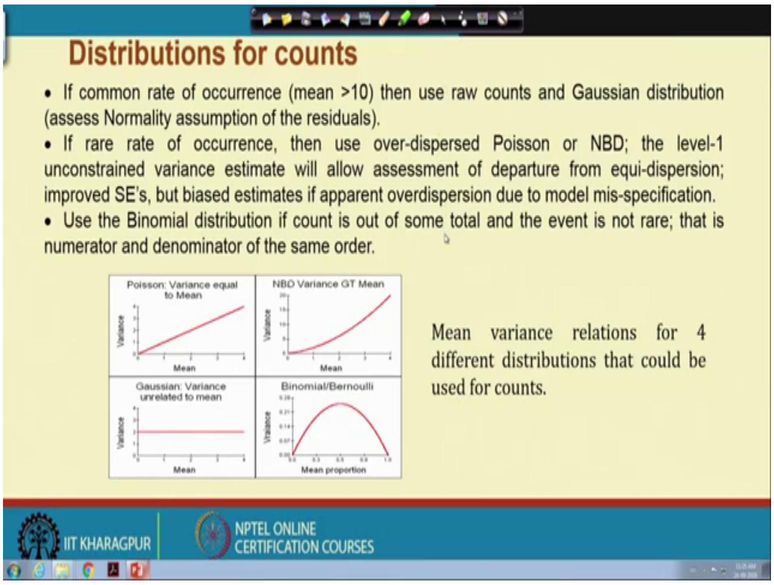
(Refer Slide Time: 17:55)



Again is what I have already mentioned. So, if you increase the sample size and stands to infinite then there is a high chance that, the distribution will be close to each other and the kind of over dispersions factors can be actually you know removed. And then finally, the kind of Poisson distributions in the kind of negative binomial distribution can be actually a close to each other. That is what actually requirement, but what is happening in practice you know both Poisson distribution negative binomial distribution you know it is a kind of rare occurrence and less than 10 percent of instances for interval or 100 percent Poisson like you know travel cases.

We will find such kind of situation that is why these models and that to counter turn model is very accessional kind of situation where we have to apply, subject to the kind of chances. And the kind of occurrence otherwise it is not possible to apply this kind of modelling to study the impact of independent variable to dependent variable.

(Refer Slide Time: 19:15)



So, this is how the kind of distributions structure and here a what is happening here is if you find graphically the deviations. So, this is actually Poisson case where mean and variance are equal. And this is negative binomial distributions where variance greater than to means.

And then corresponding to Poisson distribution and negative binomial distribution then we have a concept called as Gaussian distribution, where variance unrelated to means that is what the case. And then we have binomial where actually mean and variance proportionately converge towards similar kind of structure. So that means, you know we are bringing here actually 4 different kind of distribution starting with Poisson distribution, Negative binomial distribution, then Gaussian distributions and you know then Binomial distributions.

Of course, you know Gaussian distribution is very interesting distributions, but alternately whatever distribution we can actually think about think about it. You know you can think occurs at a exclusively depends upon the data structure and the kind of the kind of requirements that they features of the particular data.

So that means, since our discussion is actually with respect to Poisson regression modelling and negative binomial regression modelling. So, ultimately whatever you know sample structure is there provided you know the data will be specified as a count data. Then in the first instance we have a 2 different you know choice the Poisson structure and then negative binomial structure and that to the choice of a particular model exclusively again depends upon you know and the mean and variance.

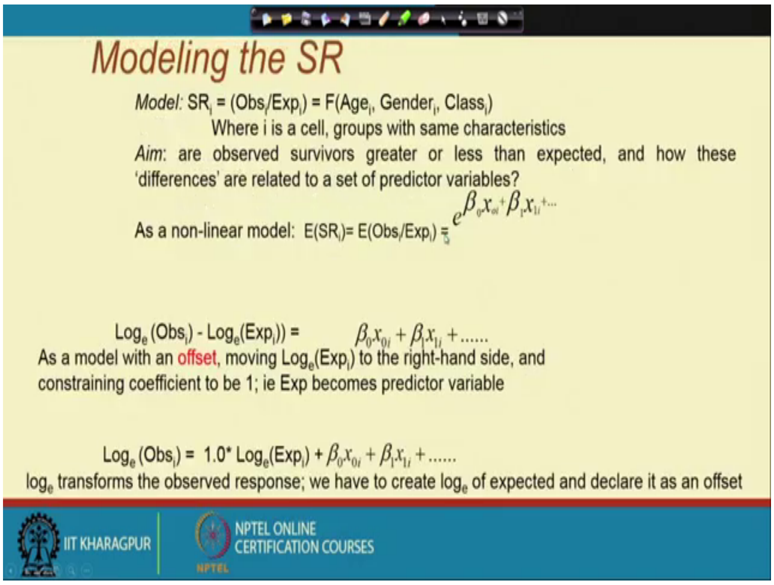
So, if mean variance are actually is same then we strictly means; we direct move to move into the structure of the Poisson regression modelling. And if there is a over dispersion case where you know mean variance are not equal; that means, variance you know variance greater than to means. So, we can go to the usual negative binomial regression modelling, but in addition to Poisson regression modelling and negative binomial modelling or negative binomial distribution. We have other type of distribution where we may find in a real life situation corresponding to some of the engineering problems.

Again if that will be followed then; obviously, one way actually to streamlined the case and go as per the particular requirement is actually to increase the sample size and the kind of data transformation. So that means, what is actually in our hand always that you can normalize the data and then you can increase the sample point.

So, that proper analysis or the kind of proper inference can be drawn subject to ability of data in the kind of the particular identify problems and the kind of engineering requirement. So, if your data size is very small and the information contents are not count data type then. Then we may not be in this particular environment we will not think about too much about this particular.

You know issue about mean variance equality or the kind of over dispersion case so; that means, technically we have lots of instances through which actually we can highlight, but technically what is happening here is.

(Refer Slide Time: 22:51)



Some of the cases where you know the modelling structure will be subject to the kind of these survival structures where actually we like to apply. And let us bring the situation where we use actually this Poisson distribution and negative binomial distribution. What I like to bring a situation here specifically with respect to count data.

And the structure is that it is a specialised case where data is actually positive integer types, but what is happening in real life situation some of the instances where the kind of for instance, a accidents, then the incidents, and the death case. So, these are the sited examples where this kind of modelling can be used.

For examples there is a concept called as survival structures we have actually a component called as survival analysis which we will discuss in the next lecture. So, where actually the particular data will be by default will be count data type.

And we like to check how they behave each other and then what kind of modelling you can apply, and then how to go about the a model choice and the kind of estimation procedures and the kind of outcome the robustness checks and the kind of reliability check.

And then finally, come with a kind of situation where you can actually predict the survival rate subject to availability of data in the kind of problem requirements.

So; that means, technically a till now whatever we have discussed. So, it is basically you know related to count data and then we have discussed 2 specialized kinds of problems, and that to Poisson regression modelling and then negative binomial regression modeling. And then we have highlighted some of the instances and the kind of requirements, and the kind of check points; through which you can make a choice between Poisson regression modelling and Negative binomial regression modeling.

And then again we have highlighted they a situation where both can actually converge to each other and then we can bring a situations, where we can either use actually Poisson regression modelling or Negative binomial regression modelling by increasing the sample size and bringing the accuracy to the systems and the kind of model requirements.

With this, we will stop here and in the next lectures we will connect a particular problems called as survival analysis corresponding to this count data modelling and the kind of Poisson regression modeling, and by negative binomial regression modelling with this we will stop here.

Thank you very much. Have a nice day.