# Advanced Hydrology

**Prof. Dr. Ashu Jain**
**Department of Civil Engineering**
**Indian Institute of Technology, Kanpur**

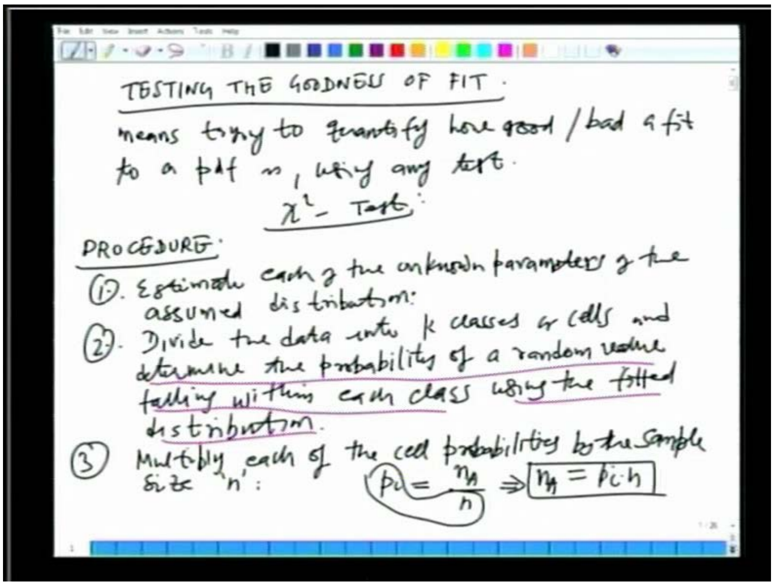
## Lecture – 32

Good morning and welcome to this post graduate video course on Advance Hydrology. In the last class we look that a couple of methods of fitting a probability distribution to a given set of data. And the methods we had looked at where the first one was the method of moments and the second one was the method of maximum likely hood. We look that one example with that is the same example of exponential distribution, fitting the exponential distribution.

Using both of these methods and we saw that the parameter lambda of the exponential distribution comes out to be 1 over x bar. And then we discuss the, a relative merits and a you know advantages and disadvantages, both of these methods. Then we introduce the concept of what is called a testing the goodness of fit of a probability distribution. We said that fitting a probability distribution is like model calibration and we would like to validate a mathematical model always, whenever we do model development.

We said that fitting a probability distribution basically means finding the parameters and how good or bad these parameters are, there are many tests available for doing that or to assets the goodness of the fit of a model.

(Refer Slide Time: 01:46)




So, we will get started with the assessment of or testing of goodness of fit what does what do we mean by a saying, when we say that the testing the goodness of a fit of a PDF. It basically means that we are trying to quantify alright, how good or bad or fit to a PDF or a probability distribution is, how do we do that, using any test. And there are many tests available and we said that we are going to use what is called that chi-square test.

So, in this course we will look at very famous test of goodness of fit which is called the chi-square test. And then we said that any of these testing's is like hypothesis testing, in which what do we do, we make certain hypothesis or assumption and then we calculate certain statistic corresponding to the test. Then we compare that calculated statistic with the standard statistic alright, for certain a level of confidant and then we say that act that particular significance level or the confidence level, we either accept the hypothesis or we reject the hypothesis.

So, what we will do is, first we will look at the procedure of the step by step method of this chi-square test conducting this chi-square test alright and then we will take up one example. So, let us get started today by looking at the procedure, which we can follow to conduct these chi-square test. The first step is we estimate each of the unknown parameters we estimate each of the unknown parameters of the assumed distribution there may be one PDF we are fitting, there may be more than one PDF they are trying to fit.

So, we do that for all the PDF's and if there is only one we estimate the parameters, all the parameters that is of that particular PDF. The second step is we divide the data alright, what we have done in the first step, we have used the data that is available alright. And what do we do in the second step, we divide this data into k classes or what is call cells and determine the probability of a random value falling within each class or cell using the fitted distribution.

So, these are the first two steps in which the first step is we just estimate the parameters of the assumed distribution alright or we just the fit the distribution of find out the values of the parameters; either a using the method of moments or a maximum likely hood or whatever. Once we have done that the next step is we divide the whole range of data into k classes for example, let say we have the annual rainfall the minimum value is 0, the maximum value is 100, we can divided into 5 classes from let say first 1 from 0 to 20, 20 to 40, 40 to 60 and so on.

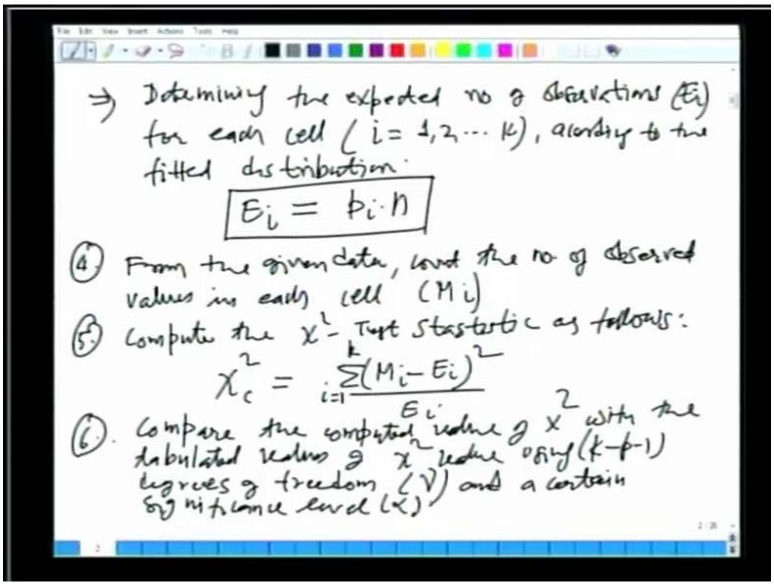
So, let say we divide the whole range of data into k classes and then what we do if you see here, what we are saying is after we have divided the data into k classes we determine the probability of a random value falling within each class, how using the fitted distribution. So, what we are doing is we are trying to find out the probability of occurrence of the random value alright, how within a particular class, let us say there are 5 classes for the a case of your annual rainfall as I said.

So, the minimum value is let us say for a second classes 20 centimetre, second one is the or the higher one is the 40 centimetres. So, we what we are doing is we are trying to find out what is the probability of occurrence or what is the probability of the annual rainfall being between 20 centimetres and 40 centimetres. How can we do that, well we have already fitted this distribution we have the equation for the PDF, we have the equation for the CDF Cumulative Distribution using all those we can find all o find all kind of probabilities right.

So, we can find out the probability within a range, so that is the second step. Let us move on what do we do in the third step, we try to find out the number of occurrence how can we do that, we multiply each of the cell probabilities multiply each of the cell probabilities by the sample size n. So, let us say that your cell probability is p i, what is this p i p i is nothing but from the basic definition of your probability this is n A over n is it not, what is n A over n, n A is the number of occurrences of particular coin and n is the total number of all the sample size or occurrences alright.

So, if you have the cell probability s p i you multiply that by n what are you going to get n A which is your p i times n.

(Refer Slide Time: 09:23)




So, basically in the third step what we are doing is determining alright what the expected number of observations, expected number of observations and let us we will call E i for each cell alright. Each cell meaning what i is varying from where to where, you have how many k cells. So, i will vary from 1 to k and how do we find this number according to the fitted distribution is this clear. So, let me say that your E of i is nothing but the cell probability which have determine multiplied by the total number of data point as a sample size.

If I go back what we are doing is we are multiplying each of the cell probability cell probability is p i by the sample size n. And then what we are saying is that, that number is what expected number of observations in each cell. So, E i is equals to what, E i is equals to the cell probability multiplied by the total number of the data point that will healed the expected number of observations. Or occurrences of the events as per the fitted distribution and we are calling this number as E i.

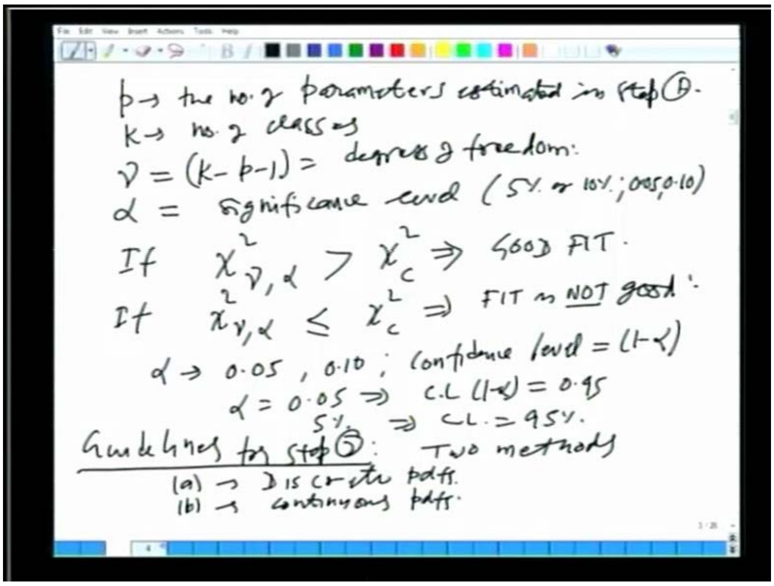
Next step is number 4 from the given data, we already have the data with us right, what we do is we count the number of observed values we count the number of observed values in each cell and let us call that M i. So, what is this step the fourth step is we are just counting the number of observations in each cell. Coming back to our example of the annual rainfall, let us say we have 50 hours of a data, 50 different values of annual rainfall right and then we have divided the range 0 to 100 into 5 classes.

So, all you then do is, you take a first observation, let us say it is 22 centimetre, what you do you just throw it in the second cell between 20 and 40 and you keep on doing that with all of them. So, that way you will be able to get the count or the number of observations in each cell or in each class from which source from the data itself or the actual observations right. So, now you have the expected number of observations in each cell from the fitted distribution and you have the same number by the data set alright.

So, now we try to compare these two numbers M and E and calculate what is called the chi-square tests. So, that is the next step, number 5 compute the chi-square test statistic as follows that is a chi-square I am going to put a subscript c here, which means the calculated one or the computed one from data alright, which will be equal to what this is how it is defined. It is nothing but M i minus E i whole square divided by E i, where i is varying from where to where, it is varying from 1 to k.

So, you see that this is how the chi-square test statistic from the computed data is defined, then coming to the next step, what we do is you always compare the computed test statistic with the standard value alright. So, then we say compare the computed value of your chi-square right with the tabulated value, the standard statistic is always available in standard tables. The tabulated values of your chi-square value using k minus p minus 1, what is it this is degrees of freedom; and this is denoted as mu let us say and a certain significant level let us say alpha.

(Refer Slide Time: 15:31)




Let me go to the next page, where what is p, p is the number of parameters estimated in step one and k is the number of classes we have already defined, mu we have defined as k minus p minus 1 as the degrees of freedom. This standard chi-square value which is tabulated it is tabulated with respect to the degrees of freedom that is number 1 and other parameter is the alpha, which is the significance level. It is expressed as in percentage normally it is 5 percent or 10 percent or 0.05 or 0.10.

So, what we are doing here is we are reading out manually the table the value of the chi-square test or the standard chi-square value depending upon the degrees of freedom. And what is the degrees of the freedom it will depend upon the kind of freedom right, which is p minus k minus 1, where p is the number of parameters. For example, if you are fitted a exponential distribution, how many parameters are there in exponential distribution we have seen is this only one right.

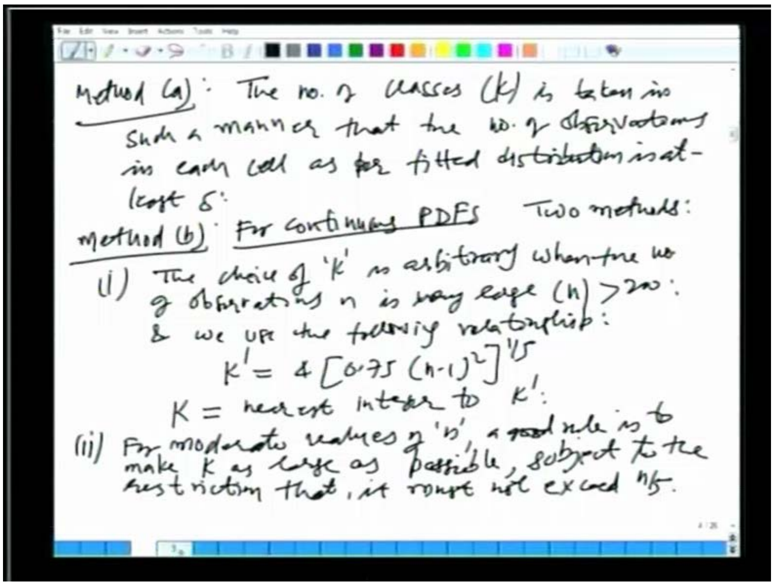
So, here is one parameter only one parameter in the exponential distribution alright. Similarly the number of classes you will have you know 10, 20, 30 whatever then you find out these degrees of freedom you choose certain significance level and just read out the value of chi-square standard value. So, then the condition is if your chi-square for degrees of freedom certain value and alpha is greater than chi-square competent means what, means it is a good fit it is as good as that.

On the other hand if your chi-square for certain degrees of freedom and significance level is less than or equal to chi-square computed means the fit is not good. Now, as I said a significance level is normally selected which is either 5 percent or 1 percent in the probabilistic terminology, these this alpha or the significance level is also expressed as 1 minus alpha which is the confidence level. So, sometimes we say the confidence level is equal to 1 minus alpha.

So, if alpha is equals to 0.5, then your confidence level which is 1 minus alpha will be what is 0.95 or if it is 5 percent then confidence is how much confidence is 95 percent, so this is the most popularly used significance level 0.05. Now, what I am going to do next is I will give you the guidelines for choosing this value of k, k is the number of classes alright. And as you can see, that the final result weather you are fit the test of the fit will pass or fail may depend upon the choice of the value of k, why because the value of the standard test statistic chi-square depends on k alright it is k minus p minus 1.

So, there are 2 or 3 different people conducting this test different people may take different value of k alright depending upon that the answer may be on either side. So, what we do is we try to a minimise that and there are certain guidelines that are available to choose alright. So, let us look at the guidelines for step two, which is choosing the number of classes k we will look at two methods alright. Method a is for the discreet distribution, discreet PDF'S and method b is for continues PDF'S these are just the guidelines there is no hard and fast rule alright, to there for different types of PDF the guidelines are slightly different.

(Refer Slide Time: 20:52)




So, let us look at method a or case a rather which is for the discreet type of distributions what does it say going to write it down the number of classes k is taken in such a manner that the number of observations alright, in each cell as per the fitted distribution is how much is at least 5 that is all it says. So, what does this method say what it is saying is that the number of classes we select in such a manner for example, annual rainfall is between let say 0 to 100 this is the range we have we can divided into two classes.

One classes is 0 to 50, other class is 50 to 100, we can also divide it into 10 classes 0 to 10, 10 to 20, 20 to 30 and so on right. But what this method is saying is that the guideline is says that, the classes should be selected such that, we have at least how many 5 observations of 5 occurrences in each cell alright. Let us say you divided into 10 classes alright 0 to 10, 10 to 20, 20 to 30 and so on. Then you find out the that in certain cell or a some particular cell the number of observation is only three alright, what you do in that case, what we do is we combine the individual cells alright.

So, if we determine or if we find that the number of expected observations is less than 5 we just combine the adjoining cell or we re redefine our cell boundaries. So, this is the method a, this is applicable for the discrete distributions only alright, method b we will look at for continues PDF’s for this also there are two cases or two methods sub divisions.

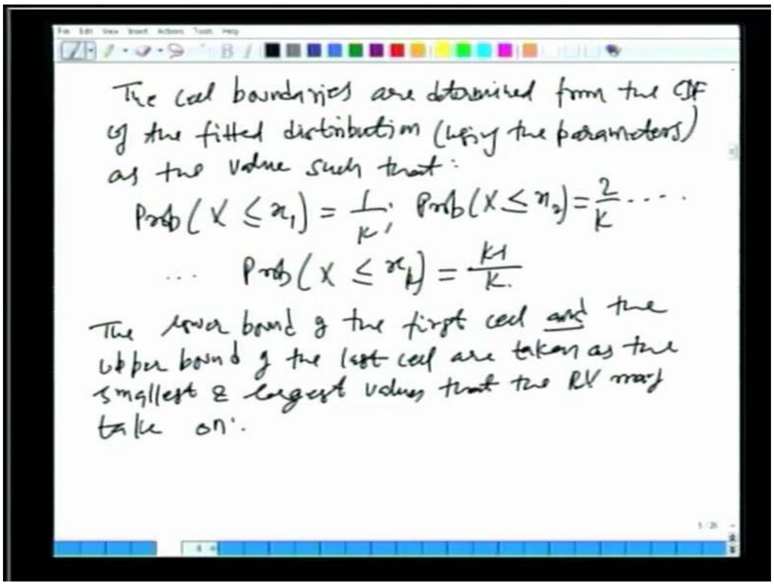
The first one says that the choice of k is actually arbitrary as per this one is arbitrary then this is applicable when the number of data or observations n is very large. And it is too large actually we do not worry too much right it can be arbitrary alright and is too large, too large means n may be more than 200 also that is when you say this too large alright. And how do we determine that and we take or we use the following relationship in this case, it says k prime is equals to 4 times of your 0.75 and n minus 1 square everything raise to the power 1 over 5.

So, what we do is then we take the value of k is the nearest integer to k prime. So, the first method is applicable for the cases when the number of data set is very large alright. And is more than 200 or something, then we calculate k prime using the equation we have just seen, then you take the nearest integer and we say that this is the value of k alright. It is more or less actually arbitrary or a alright when we have lot of data points then the charges of k is not vary critical because you will have lots of observations in each cell.

For the second case when the number of data set is not huge or for moderate values of n values of n good rule is what is to make your k as large as possible alright. Subject to the restriction that subject to the restriction that it must not exceed n by 5 it must not exceed n by 5. So, what we are saying is that for a moderate number of data points for the continues case alright, we take the value of k as large as possible alright, but it should not be more than n by 5.

In case of our a let us say rainfall annual rainfall you know data set, let us say we have 100 years of data, 100 data points then what is the upper limit of the number of classes, it is n by 5 or 20 we should not have more than 20 classes alright. Ideally it should be between u know n by 4 to n by 5 or something like that sorry n by 5 to n by 8.

(Refer Slide Time: 27:28)




Any way, so let us move on one important thing which we need to learn is how do we determine these cell probabilities alright. The cell boundaries the cell boundaries are determine from the CDF a Cumulative Distribution Function of the fitted distribution of the fitted distribution that how that is using the parameters, which we have said. As the value such that which satisfies the following conditions that is probability of your x being less than or equals to x 1 should be 1 by k.

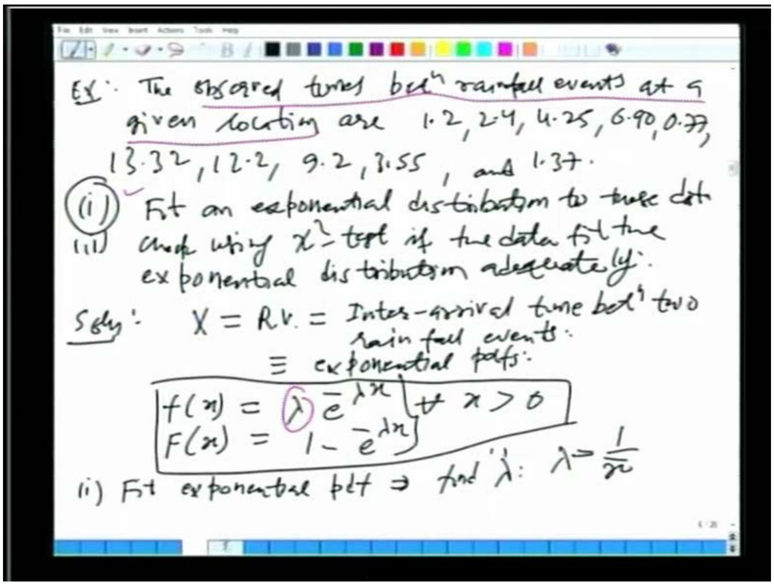
For the first cell that is the probability of your random vary variable having a value less than equal to some x 2 is equals to 2 by k and so on. All the way to the probability of your random variable having a value less than equal to x k is equal to what, is equals to k minus 1 over k. So, what we are saying is that the probability of occurrence alright is same it is 1 by k alright for each cell; using that concept we are finding the values of x 1, x 2, x 3, x 4 all the way up to x k alright.

So, these numbers x 1, x 2, x 3 actually function as the cell boundaries alright. So, the what will be the first one, the first one will be between and x 1 the second cell will be x 1 to x 2, third will be x 2 to x 3 and so on. The first and the last one are basically you are the minimum value of the observed value of your random variable and the next one value of the observed value of the random values. So, if you want to say that the lower bound of the first cell and the upper bound of the last cell are the or taken as the smallest and largest smallest and largest values that the random variable that the random variable may take a form theoretically.

So, what we have seen the actually is the complete method of this carrying out the chi-square test alright they were you know 4, 5 different steps that is what we saw first. And then we look that a certain guidelines that can be followed or that should be followed to the step two which is the a classifying the data into k number of classes and finding the cell probability.

So, they are certain guidelines we should follow for finding this number k, now what we will do is we will take a (( )) example to demonstrate how this method actually can be implemented for how we can test the hypothesis, once we have fitted the distribution. How we can find out, how we can quantify whether this particular fit which we have you know found out is good or bad or how good it is, if it is good how good it is, if it is bad how bad it is.

(Refer Slide Time: 31:51)




So, let us look up at an example the observed times between the rainfall events observed times between rainfall events at a given at a given location are given as follows these are the data that are given to us 1.22, 2.4, 4.25, 0.90, 0.77, 13.32, 12.2, 9.2, 3.55 and 1.37. What we have to do is fit an exponential distribution fit an exponential distribution to these data, number 2 check using the chi-square test if the data fit the exponential distribution adequately.

So, this is the problem or this is the question in which we are given the time (()) between two rainfall events and there different data points given I think about 10 of 10. What we have to do is we have to find out whether, first thing we have to do is we have to fit the exponential distribution. And the second thing we have to do is we have to find out the weather this fit is adequate or not alright and then we will use a standard value of alpha which is 5 percent.

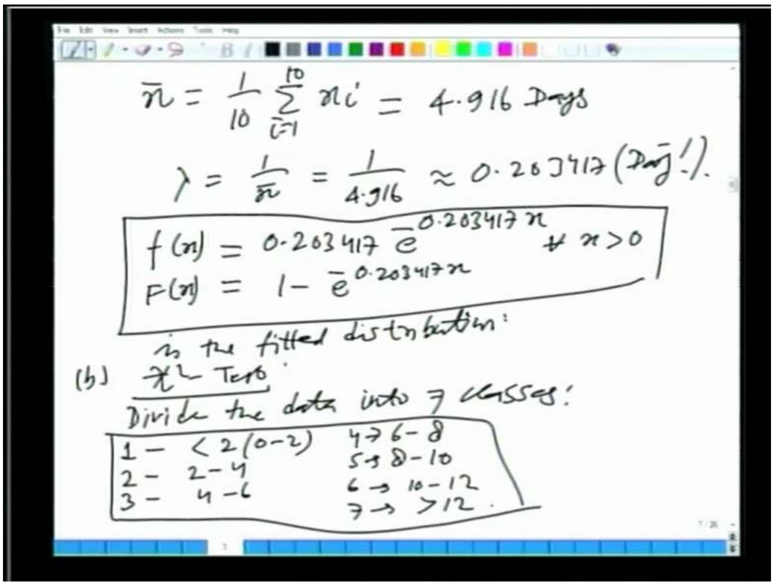
So, let us look at the solution of this question in which what is the random variable x, how is it define it is nothing but the what is called the inter arrival times between two rainfall events that is your random variable. See, what is given to us is the this data, what is this observed times between the rainfall events at a given location, times between two rainfall events. So, this is what is called the inter arrival times and this actually follow the exponential distribution from the knowledge of hydrology the inter arrival times it follows the exponential distribution.

So, these are the data given to us, so what we do first thing is the first part is what fit the exponential distribution alright what is the exponential distribution. Well we have seen the PDF is given is what lambda E minus lambda x for all values of x greater than what 0 right. What does this mean, now the inter arrival times can it be negative the inter arrival times, we said that the inter arrival time follows the exponential distribution and exponential distribution is for non negative random variables only.

And in this case you see that the inter arrival time that is the time between two rainfall events it cannot even be actually when you are defining the two rainfall events then they have to be separated by some times it has to be positive right. So, it cannot be negative and it cannot be even 0, so it should be greater than 0, so that is what we are saying here. So, this is your small f of x of PDF and we can easily find out the thus capital F is going to b if you integrated 1 minus E minus lambda x and this is too far this.

So, this is the equation of your exponential PDF or the distribution now what we have to do is we have to determine the value of what this parameter lambda in this first step. So, let us do that fit exponential PDF means what, means find the value of lambda is it not and what is lambda we have seen yesterday using the two methods method of moment and method of maximum likely hood what is lambda nothing but 1 by x bar. So, can you find out the x bar it is nothing but the average of the data points right.

(Refer Slide Time: 37:31)




So, what you do is you just find out x bar is equals to there are 10 data points summation of your x i going from 1 to 10 what are x i's x i's are the inter arrival times they are all given to us I am not going to spent time on that you can find it out it will 4.916 days. By the way I should mention that all these are in days, I forgot to mention that. So, this is your x bar or on an average the inter arrival time on an average the inter arrival time between two rainfall events is how much about 4.2 days or 4.9 days or 5 days as per the data that is available to us.

Now, I am going to a just mention one thing here is that this example which I am taking it is only for demonstration purposes. Remember we have seen certain guidelines of how do we select the value of k in carrying out the chi-square test alright, at least 5 observations in each cell and so on right. But what we are going to do here is we will not actually adhere to those guidelines because we have only 10 data points here. So, for demonstration purposes I have purposefully the number of data points has only 10.

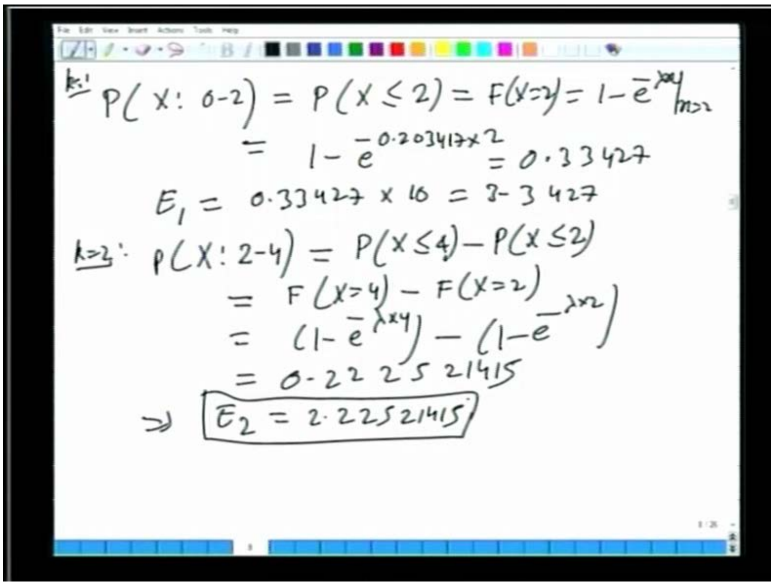
So, that the things are manageable here and some of you may be confused that how come we are not following those guidelines it is not the purpose of this exercise I just want to question you before ahead alright. So, let us move on once we have found out the x bar you can find out what the value of lambda as 1 over x bar, which will be one over 4.916 which will be 0.203417 and what is the unit of this inverse alright. So, then your fitted distribution is what it is this x or all values of x more than and capital F of x the CDF is given by what 1 minus e minus lambda x right which is 0.203417 of your x.

So, this your answer to the first part, which is the fitted distribution or you fit a exponential PDF that is what we have done, this is the fitted distribution to the data. So, the next step then is we have fitted the exponential distribution alright in the second part of this problem what we have to do is we have to find out the adequacy of this fit or the goodness of the fit of this alright and we are going to use the chi-square test earlier seen alright.

So, in part b we carry out the chi-square test in the chi-square test, the first step is determination of the parameters which you have done already and then the next part is of the next step is we divide the data into k classes right. And what I am going to do here is divide the data into 7 classes again it is for demonstration purpose only, do not worry about how come you have observations in each cell in some of the cells or less than 5 and so on.

And what are these seven classes first one is less than 2 or 0 to 2, second one is 2 to 4 third is these are the inter arrival times in days remember capital X is what that is your inter arrival time in days. Then you have 4 to 6, fourth is 6 to 8, fifth is 8 to 10, sixth 6 is 10 to 12 and the seventh class is what is greater than 12 alright. So, this is the class distribution we are going to follow for this particular example alright.

(Refer Slide Time: 42:18)



Now, If you move on this next step is to find what the cell probabilities for the first cell you have to find the probability such that the x is between what and what between 0 and 2 right which is nothing, but your probability of x being less than equal to 2. What is this given by the knowledge of your PDF and CDF, the probability of a random variable being less than equal to a particular value, which is 2 days in this case it is given by what the cumulative distribution function value at x is equals to 2 right.

So, I can then right this as f at x is equals to 2 which is 1 minus E minus lambda x with x is equals to 2 is it not. So, this is going to be equal to 1 minus E to the power minus 0.203417 times what, times 2 you can calculate it and it should come out as 0.33427. So, this is the cell probability for the first class which is between 0 and 1, how can then now I find out E 1. E 1 is the number of expected observations in first cell alright, as per the fitted distribution right this is equal to what the cell probability multiplied by the total number of observations which is done alright.

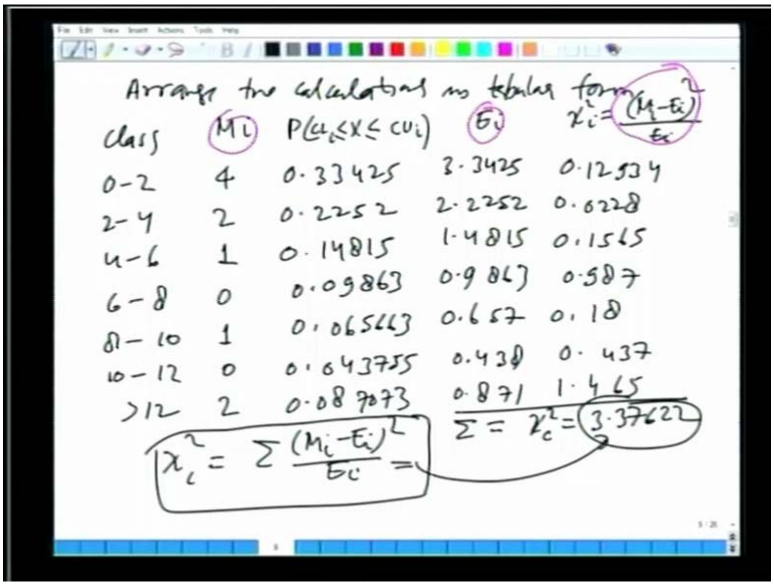
So, then what you do is you just find out the expected number which will be 0.33427 multiplied by 10, so it will be 3.3427. It can be a fraction, it can be a real number it does not have to be an integer because we are using the a fitted distribution let me do one more case for the cell second cell that is this is for let us say k is equal to one now we have k is equal to 2 alright. What we have to do is we have to find out the probability of the random variable assuming what the value between 2 and 4 from the knowledge of our CDF.

We can say that this is the probability of x being less than or equal to 4 minus the probability of x being less than equal to 2 it should be clear to you. What is this equal to well this is nothing but the CDF for x is equal to 4 minus of your CDF for x is equal to 2 that is f at x is equal to 2. What is this is 1 minus E to the power minus lambda times 4 right, I am not putting the value of lambda here minus 1 minus E to the power lambda times 2.

You can put the values of lambda and you can simplify this what you are going to get is this number 0.2 to 2 5, 21415 alright we can of the digits here that will give you E 2 as what 10 times of this which is 2.225 and so on 2.22521415. So, this way you see we can actually keep on doing this process and the best way to carryout these computations is we always do this calculations in a tabular form it is always a good practice to organise the repeated calculations in a tabular form.

So, what you do is you crate a table on the top you write all the formulas which you are going to use and then just arrange everything. Otherwise it becomes very lengthy and time consuming, so that is what I am going to do alright.

(Refer Slide Time: 46:29)



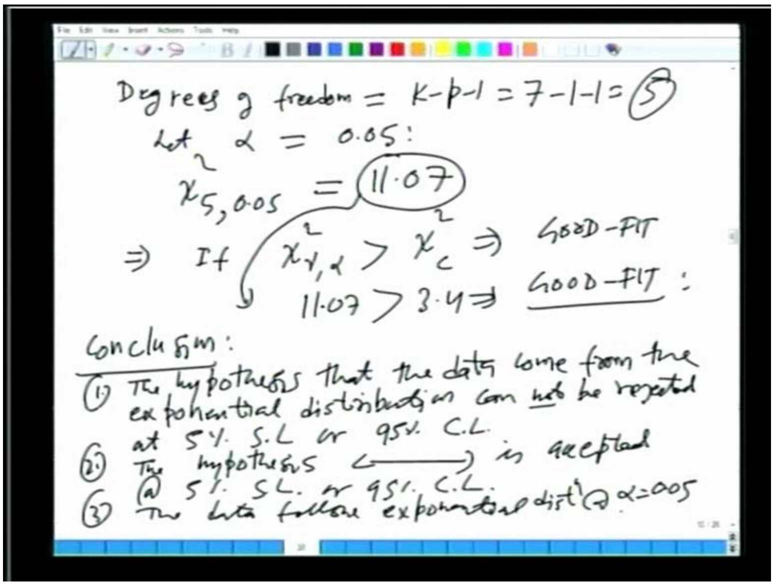
So, we arrange the calculations in tabular form alright. So, what we do is we have the class or classes going from 0 to 2, 2 to 4, 4 to 6, 6 to 8, 8 to 10, 10 to 12 and greater than 12 or 12 to infinity or whatever. Then what we have is the next number I am going to have actually is M i what is M i, M i is the actual observations based on the data alright, which you can count and do that. So, this I am going to put here directly for you. So, you have 4 in the first one, 2 in the second one and 101 you can verify these numbers and then 0 and then last one is 2.

The next column is the probability of your random variable x being less than equal to your let say lower bound on the cell or the lower values of the cell and the c u for the i, i is going from 1, 2, 3, 4, 5, 6 and 7 alright. We have determine this for the first and second already. So, if I wrote that it is going to 0.33 4 to 5 then you will have 0.2 to 2 and 5 2 this numbers we have found out. And the other one I will just write it you can verify these things 14815, it is a 0.9863, this is 0.65663, 0.437 double 5 and the last one is 0.8773, we multiply this number by 10.

And get what is expected number of observations based on the fitted distributions alright. So, this will be 3.3425 this will be 2.2252 and so on 0.9863, 0.657, 0.438 and 0.871, I am just rounding of, then you have the last column here, which is I am going to chi-square component from the i itself contribution to the chi-square from the i itself. What is that this is M i minus E i whole square divided by what divided by E i. So, you have the values of E i and M i all you do is calculate this, can you do that sure we can alright let me write some of these numbers. So, you have 0.12934, next is 0.0228 you can verify these numbers please, then you have 0.1565, 0.987, 0.18, 0.437 the last one is 1.465 or so on.

Then the last thing we do is we sum these up the last column that is that will give you chi-square c right, which is 3.37622, what is the computed value of chi-square, we have defined it is nothing but the summation of what M i minus E i whole square divided by E i. So, what we are doing is we are calculating M i minus E i whole square over E i for each cell you some the last column up, that will give you chi-square, which is equal to what in this case this is equal to this number alright. So, we have calculated the computed value of chi-square, what do we do next we found find out the standard chi-square value depending upon the degrees of freedom and the significance slab, so let us do that.

(Refer Slide Time: 51:45)




What are the degrees of freedom in this case degree of freedom is what it is nothing but k minus p minus 1 right what is k is the number of classes, which is what 7. What is p, p is the number of parameters how many parameters are there in the exponential distribution only 1 minus 1 is equals to what 5. So, 5 is the degrees of freedom and then say let the significance level be 0.5. Now, what we do is we go to the standard tables of chi-square the chi-square tables, the chi-square distribution tables will be available in any standard book on the statistics.

You take up any probability and statistic text book at the end you will find lots of tables corresponding to various distributions and one of them will be the chi-square alright. And it is a two dimensional table in which the chi-square value will be given depending upon the degrees of freedom and the significance level, so you just read it of the proper one. So, if we go to those tables what we will find is that your chi-square for 5 comma 0.5 will be equal to what it will be 11.7 please verify that alright.

So, is the test pass or fail or is the fit good or bad what was the condition, the condition was if your chi-square mu comma alpha is greater than the chi-square c means what good fit is it not right. What do we have, we have the computed no the standard value goes here it is 11.07 right is greater than what is the computed value let me go back the computed value was 3.37 something. So, let me say 3.4 it is 3.4 means what it is a good fit.

So, what is the conclusion of carrying out this chi-square test well there are various ways of saying it one of them we will say is what the hypothesis the hypothesis that the data come from come from the exponential distribution the exponential distribution cannot be rejected alright. Cannot be rejected at 5 percent significant level or you can say 95 percent confidence level any of these two things. Another way of saying is that same thing the hypothesis, (( )) same thing I am not going to write it again then you say is accepted.

What is the hypothesis that the data come from the exponential distribution or the data follow the exponential distribution is accepted at what, at 5 percent significant level or 95 percent your confidence level. Or simply you can say that the data follow the exponential distribution at alpha is equal to this or confidence level is equal to 95 percent. So, we see that this way we have seen how we can carry out or how we can conduct the chi-square distribution test or chi-square test, any test we take all we do is we compute the test statistic based upon the data.

Other one we found out which is called the standard statistic and then we compare the two. And if the computed statistic is less than the standard value, in this case we saw the standard value at degrees of freedom and you know different parameters if that is more of the standard statistic is more then we have good fit otherwise we do not have a good fit. And then there are various ways of actually saying that a I would like to stop at this point of time because I am running out of time. And then what we will do in the next chapter actually we will start with the frequency analysis.

Thank you.