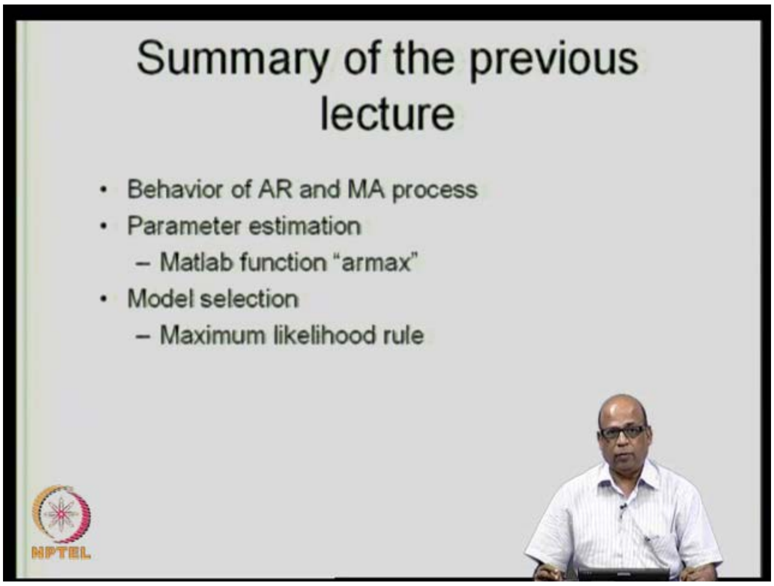
Stochastic Hydrology

Prof. P.P.Mujumdar
Department of Civil Engineering
Indian Institute of Science, Bangalore

Lecture No. # 17
ARIMA Models – IV

Good morning and welcome to this, the lecture number 17 of the course Stochastic Hydrology.

(Refer Slide Time: 00:25)



If you recall in the last lecture, we discussed the ARIMA models and specifically, how the AR and MA components of the model behave. For example, we considered AR 1 model, formulated a theoretical AR 1 model and saw how the correlogram of the theoretical AR 1 model behaves and how the spectrum behaves and how the PAC function behaves that is the Partial Auto Correlation function behaves and then we also looked at two examples of AR 2 models and similarly, for MA 1 process we saw how the autocorrelation, partial autocorrelation function and the spectral density function behave.

So, we essentially identify the number of AR terms and the number of MA terms in a AR ARIMA model by looking at the PAC function as well as the correlogram. But I as I said in the last lecture, when you have both AR and MA terms in the model and they are slightly of higher order, let us say that you are talking about ARIMA 4, 5 ARIMA 4, 2; 4, 1, these kind of models where the number of AR terms is quite large and the MA terms are also reasonably large in number, then identification becomes quite difficult.

And therefore, what we do is, we formulate candidate models, several of candidate models and then for the candidate models, we examine, we estimate the parameters and then examine whether the model is valid or not, then we also examined in the last lecture, the procedure for parameter estimation. As I mentioned, there are several algorithms available for parameter estimation and we introduced the mat lab function ARMAX, which can be readily used and the the parameters for any given ARIMA type of model can be estimated.

We, towards end of the lecture, we examined the maximum likelihood criteria for selection of the models. The specific problem is, that we have a number of candidate models, we have estimated all the parameters for the models based on the data, please do not lose sight of the hydrologic aspects of this, we have observed data, for example the stream flow at a particular location. And then on this observed data, we are doing all of these exercise so that we can fit a model for the time series that has been observed.

So, once we formulate the candidate models and estimate the parameters for each of these models, then we have to choose which among these candidate models is best suited for the observed data, this we do by two methods as I mentioned in the last lecture, one is the maximum likelihood criterion which we use for, when your application is for long term simulation of the data, then we use the maximum likelihood criterion.

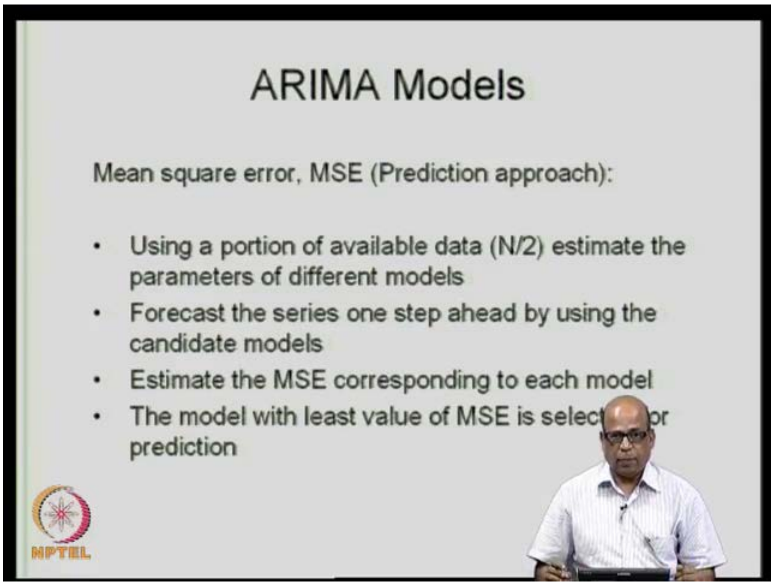
What we did in that, we defined a likelihood function, log likelihood function to be precise and estimate the log likelihood function value for each of the models, so each of the model will have one log likelihood value, then pick up that particular model which results in the maximum value of the log likelihood function value.

Now, we will go to the second part of the applications of the model where the models are meant not so much for a long term synthetic data data generation, but for one time real time, one time forecasting models that means, one time step ahead forecasting models, like I mentioned, these applications will be, for example monthly forecasting of the stream flows which may be useful for your real time operation of the reservoirs, where you are operating the reservoir for irrigation, hydropower, etcetera depending on the storage level and depending on the forecast of the flows, you would like to operate the gates.

This also becomes important when we are looking at smaller time periods like 10 day time periods for real time irrigation scheduling and so on, when we look at the applications, these points will be clear. But essentially we we use the time series models to develop forecasting develop forecast for the reservoir in flows, evapotranspiration, rainfall and so on, so forth.

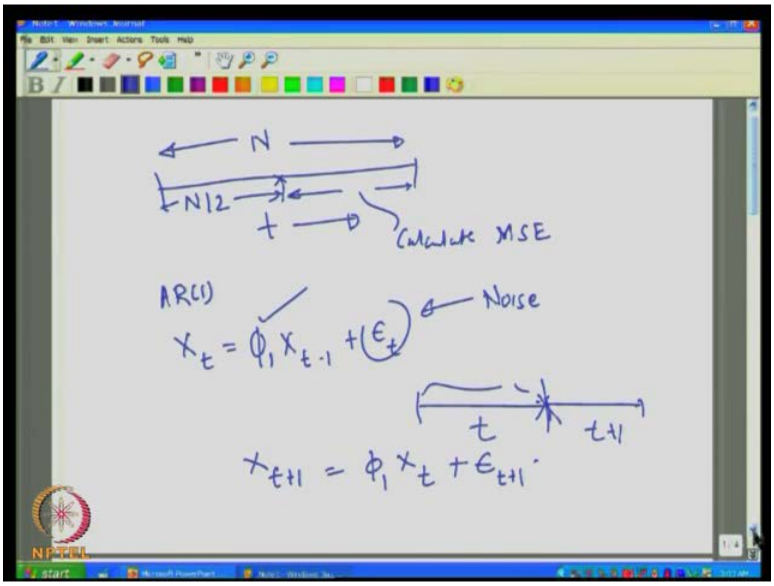
So, when our applications are for short term forecasting, then the maximum likelihood criterion is not the criterion that we use, we must then go for the minimum mean square error criterion. Now, that is what we will discuss in the lecture today, how we develop the minimum mean square error values for the models and then pick the best model that gives the minimum mean square error.

(Refer Slide Time: 06:05)



So, the minimum mean square error criterion or the MSE in short, which is also called as a prediction approach, what we do in this is that, let us say you have 50 years of data, monthly stream flow data, then we use the first 50 first 20, 25 years of data to formulate the mean square error.

(Refer Slide Time: 06:36)



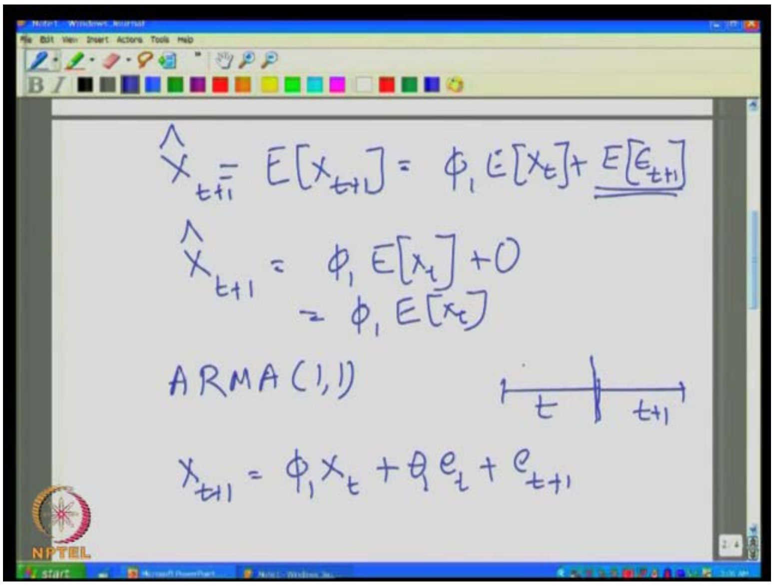
Let us look at this sketch here, let us say that your time is on this scale and you have N values of the time series, we use the first N by 2 values typically for developing the model and then the remaining N by 2, we use for calculating the MSE. So, we calculate the mean square error with the remaining N by 2 values.

What do I mean by that; let us say that you have fit a AR 1 type of model for using this particular data. When I say you have fit the model, it means that the parameters have been estimated for this particular model. How do I write this? We write this as X t is equal to in our notation phi 1 X t minus 1 plus epsilon t. Now, this is a model that we we have written and we have estimated the parameter phi 1 using one of the algorithms.

Now, when we want to do, use this model for forecasting, what does the forecast mean, remember that this is a noise term here or a residual term, which has a zero mean and they are all uncorrelated, the epsilon t are all uncorrelated. When we want to use this particular model for forecasting purpose, the forecast is the expected value for the next time period, let us say that you are standing in time period t and you want to forecast for time period t plus 1. So, you are here now, and you know the information that is available up to this particular point, and you want to forecast for X t plus 1. Let us say that I write X t plus 1 is equal to phi 1 X t plus epsilon t plus 1, I am writing it for t plus 1 now.

Then for forecasting, whenever we say we want to forecast this using this particular model, what is it that we are doing, we want to get the expected value of X t plus 1 given the entire history up to time period t, that is a problem, so we want to see, what is the expected value of the flow for time period t plus 1 given the value until X t.

(Refer Slide Time: 09:26)



Now, to do this, I will take the expected value of X t plus 1 now, let us say I write X expected value of X t plus 1. So, I am taking the expected value of this particular expression, is equal to phi 1 into expected value of X t plus expected value of epsilon t plus 1. Now, this is the forecast, and we generally denote the forecast by putting a cap on that particular variable. So, X t plus 1 cap which is the forecast for the time period t plus 1 is equal to phi 1 expected value of X t plus expected value of epsilon t plus 1.

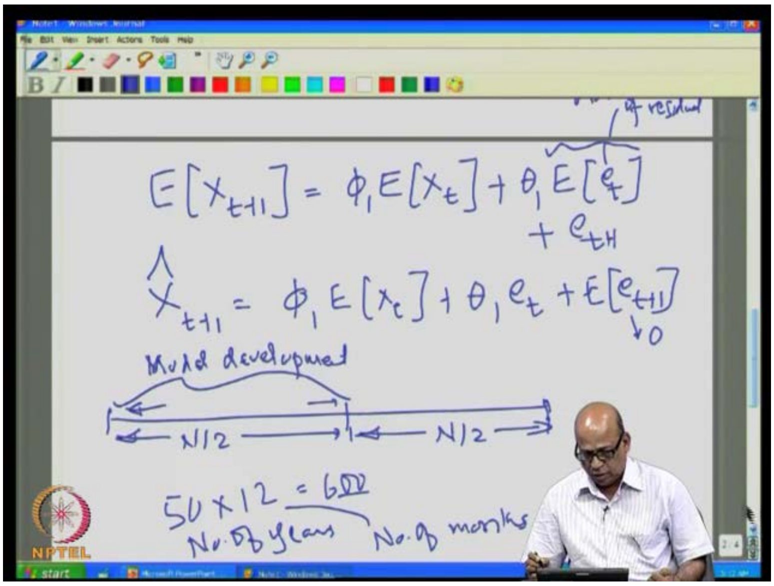
So, I will write expect the forecast X cap t plus 1 is equal to phi 1, expected value of X p X t, I will write keep it as it is plus, what is this? This has the mean of 0, the epsilon t remember is a sequence which has a mean of 0. So, when you take the expected value of epsilon t or epsilon t plus 1 that is 0 here, so this is simply phi 1 expected value of X t.

Similarly, you can look at any models, let us say that you want to take ARMA 1, 1 model here, and then use it for forecasting. So, ARMA 1, 1 model we write it as X t plus 1. Let us say again we look at two time periods, two adjacent time periods, time t, time t plus 1, you are here now and you want to forecast for the next time period. To keep the relevance with hydrology, let us say that you are standing in the month of June and you have the flow information up to the month of June. And you would like to forecast what is likely to happen to the next month that is the flow during the next month, if you are looking at the forecast for flows.

So, we will write this as X t plus 1 cap is equal to expected value of I am sorry I will may be I will write first the model here. So, ARMA 1, 1 model we will write first, we write ARMA 1, 1 model as for X t plus 1 is equal to I have one AR term. So, I will write this as phi 1 into X t, I have one MA term. So, I will write it this as theta 1 into e t plus e t plus 1, because I am writing for time period t plus 1.

Now, you would like to forecast, use this model for forecasting. As I said, what do we mean by forecast? Forecast is the expected value of the particular variable, stream flow in this particular case for example, for the next time period t plus 1.

(Refer Slide Time: 12:36)



So, we take the expected value of X t plus 1, I will write that as expected value of X t plus 1 is equal to phi 1 expected value of X t plus theta 1 expected value of e t plus e t plus 1.

Remember here, when we are writing a ARMA 1, 1 model, the e t term here is the residual arising out of the application of this model in the time period t. So, when you write this expected value of e t in this particular term, the e t value is a constant. It is not the expected value of t plus 1 that becomes 0. For example, we are may be talking about expected value of five units and therefore, this becomes e t itself. So, I will write this as X t plus 1 cap is equal to phi 1 expected value of X t plus theta 1 e t plus expected value of X t plus 1 e t plus 1 which is 0.

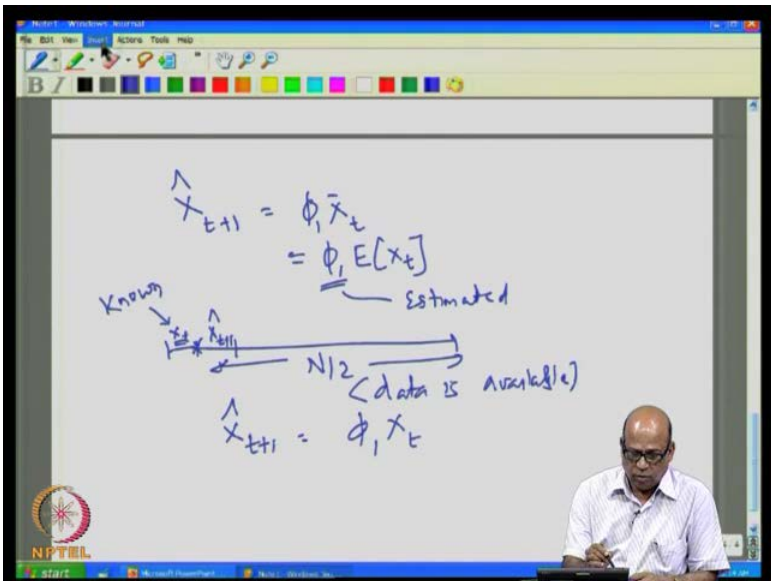
What is a difference here? What I did is that, this is a actual value of residual that resulted from this, the this is actual value of residual that results from applying this particular model for the time period t (Refer Slide Time: 13:58). So, there was a residual, there was a actual value that was available and then the the forecasted value was available. So, the actual value minus forecasted value gave me this e t value and therefore, when I take the expected value, it becomes e t itself. So, I will write this as phi 1 expected value of X t plus theta 1 e t and this becomes 0, because these are all noise terms and therefore, you write the forecasted value for time period t plus 1.

Essentially then, we are getting the forecasted values and applying for the first N by 2 terms that you have, as I mentioned you have N by 2 here, number of values and you have another N by 2 number of values (Refer Slide Time: 15:00). So, what we did, we have developed the model using this N by 2 number of values, this is of for model development, by model development I mean we used all of these values to obtain the parameters of the candidate models.

We had let us say for the forecasting, we had the models AR 1, AR 2, ARMA 1, 1, ARMA 1, 2 and so on. As I mentioned in the last class, when we are talking about the forecasting models, typically the lower order models in terms of the number of parameters will suffice. In fact, when we see in the applications most of one, the AR 1 or AR 1, AR 2 model themselves will be sufficient for one type step away forecasting, especially when we are talking about smoothen processes like monthly stream flow, seasonal flow flows and so on.

So, we choose the candidate models which may be different from the candidate models that you would have chosen for long term synthetic generation of the data. We chose that and then developed the model using the first N by 2 value, let us say if you have 50 years of values, you would have 50 into 12 that is 600 values is the data, this is number of years and this is number of months (Refer Slide Time: 16:35). So, you would have 600 values, so use the first 25 years data which is N by 2 to develop the model and then calculate the errors of using this model on one time step ahead forecasting for the remaining data, remaining part. What do I mean by that; let us say that you chose your AR 1 model.

(Refer Slide Time: 17:27)

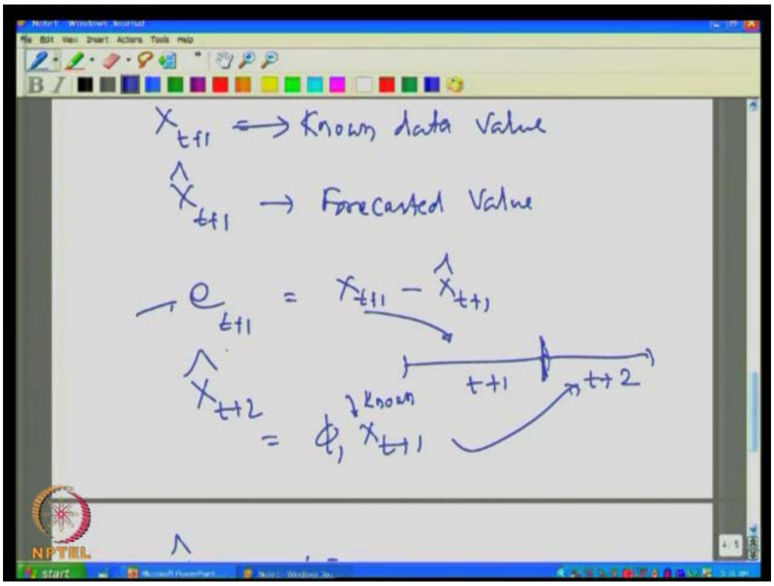


So, in the AR 1 model, you would have written your expected value of X t is equal to phi 1 into expected value of X that is forecast for time period t plus 1, you would have written as phi 1 into X t bar, or is equal to phi 1 expected value of X t.

So, this phi 1, you would estimate based on the first N by 2 values of the data and then start applying this, when you start applying this, let us say that you are now applying for the remaining N by 2 values and this is one time step ahead focus. So, given X t now, you had this X t value known, you want to apply this. So, I will get X t plus 1 cap, this is known, X t is known, typically let us say you are standing at the end of one month, June month. The flow during the month of June is known, you want to forecast the flow during the month June July.

Then we apply this, phi 1 is known, phi 1 is estimated already. So, we apply this as X bar t plus 1 is equal to phi 1 into expected value of X t that becomes X t itself, because it is known. So, this you get X bar t plus 1, but because you have the data already for this N by 2, this data is available. What we do, this is the forecasted value and X t is the actual known value of the flow. Let us say that you have X t available with you, so I will I will take the error error of forecast.

(Refer Slide Time: 19:27)



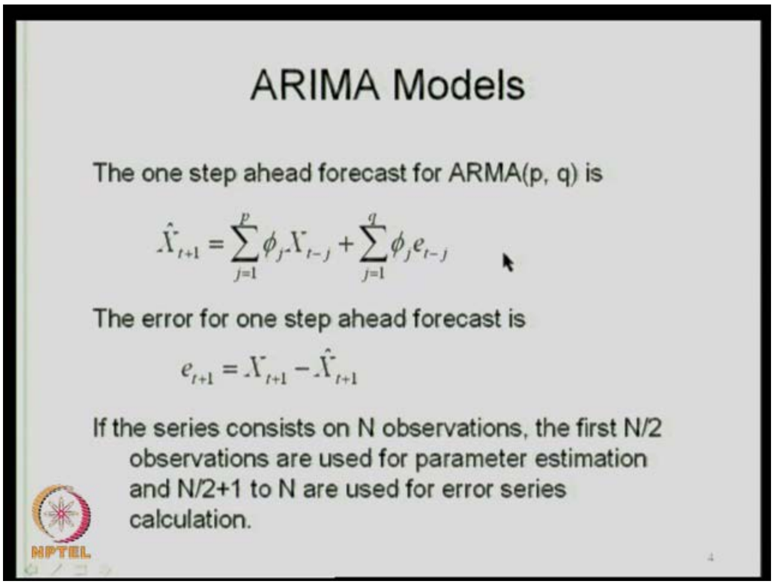
So, what will be the error of forecast, X t plus 1 is your known data, I will say this is known data and X t plus 1 cap is your forecasted value (No audio from 19:40 to 19: 46), let us say known data value. So, the error I will write for the time period t plus 1 is equal to X t plus 1 minus X cap t plus 1, this is what is forecasted from the model, so like this I get error.

Now, we are writing for the remaining N by 2 time periods. So, next time when I go, let us say that I want to write for X t plus 2 now. So, I finish my forecasting problem for t plus 1. So, I come to t plus 2, when I am writing for the next time period, we use the actual value that is known, X t plus 1 to get for X t plus 2 the forecast, not the forecasted value, remember. Because this data is already available, and the question that we are asking is standing at this point, knowing the value up to this point, what is my forecast for the next time period?

So, X t plus 2, I will write this as phi 1 X t plus 1, the expected value of X t plus 1 which becomes a X t plus 1 itself, because this is known, this is known here plus expected value of the error term which is 0. So, I will simply get X t plus 1 based on this, X t plus 2 based on this (Refer Slide Time: 21:00). Then again calculate the error corresponding to X t plus 2, because the X t plus 2 forecast, because you know the X t plus 2 value itself. And therefore, you get errors of t plus 1, t plus 2, the errors corresponding to t plus 1 t plus 2 and so on. So, you formulate the error sequence, using the error sequence, then you can get the mean square error.

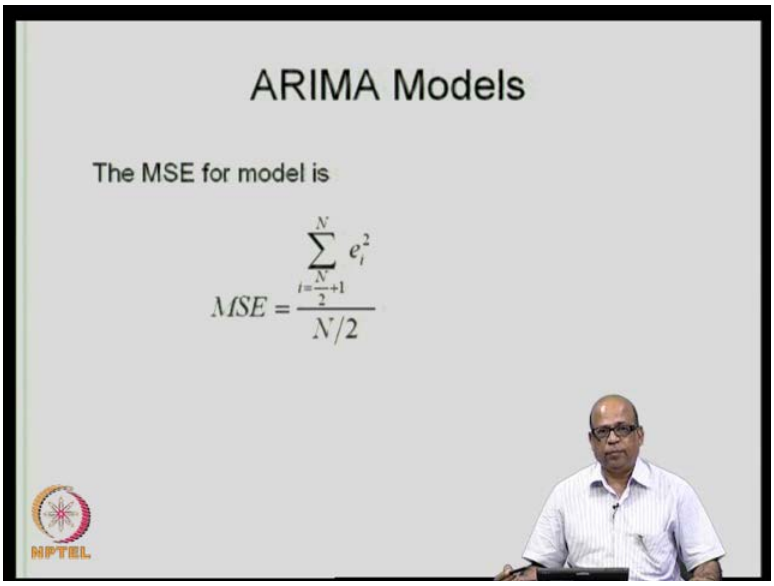
So, this is the procedure that we use to obtain the errors of forecast. So, the same thing is summarized here, using a portion of available data which is typically N by 2, estimate the parameters of different candidate models. Use this forecasted models so developed, all the candidate models one by one to get the series of forecast one time step ahead by using the candidate models. And corresponding to each of these forecast, you know the error term, get the mean square error and then pick up that particular model which results in the minimum mean square error.

(Refer Slide Time: 22:23)



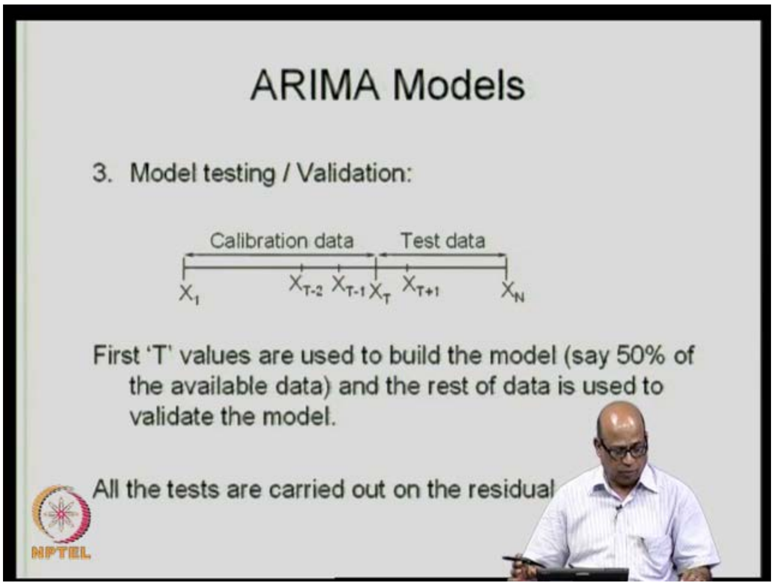
It is written more formally like this, so once time step ahead forecast for ARIMA p q model now can be written as X bar t plus 1, you have p of AR terms X j X t minus j plus you have q of moving average terms j is equal to 1 to q phi j e t minus j, the noise term has mean 0, therefore it when you take the expected value, it vanishes and then the error for one time step ahead focus is e t plus 1 is equal to X t plus 1 which is the known value minus the forecasted value.

(Refer Slide Time: 23:11)



Once you know e t plus 1, then you get the mean square error, mean square error is because you have used N by 2 number of values, you sum over sum over all these errors, the squares of errors and then divide by N by 2, you get the mean square error. And choose that particular model among the candidate models, among the number of candidate models to pick the particular model which results in the minimum mean square error.

(Refer Slide Time: 23:33)



Now, what did we do, whether it is for long term simulation of the data or it is for short short term, one time step ahead forecasting, we use the models ARIMA models and integration we use for differencing, so typically we difference the series first and then apply the ARIMA models. So, whenever I say ARIMA models, the integration is applied implied that is the differencing is already done and on that we are applying the ARIMA models.

So, in either case, whether it is being used for long term sequence synthetic generation or for short term duration of the data, we use part of the data for calibration and parameter estimation, the remaining part we are using it for validation type. Let us say you are using it for long term simulation and it is of ARIMA type of model, ARIMA 1, 1 or ARIMA 2, 1 etcetera. So, there is also a MA MA model available, MA term that is available.

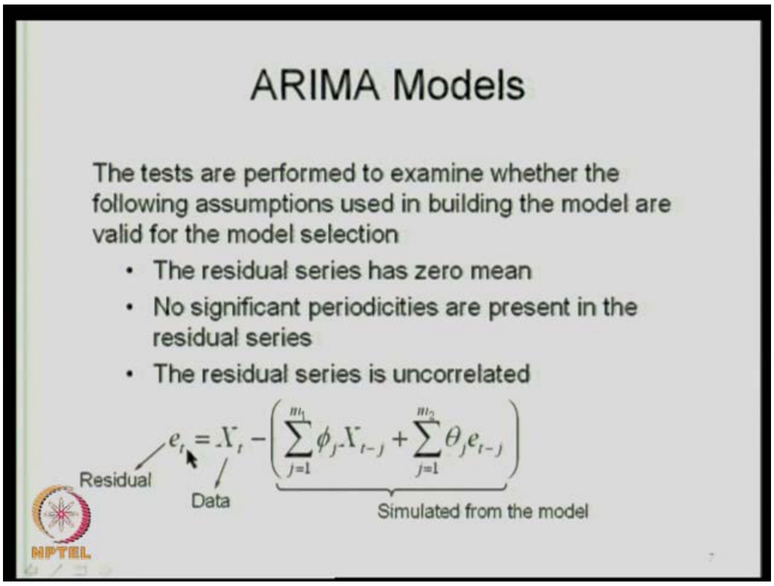
Then what we do, that we applied for X t plus 1, get X t plus 1, you already have a known X t plus 1. So, the residue term e t plus 1 is got from the available data minus the model data, same principle as we did for forecasting. So, you generate corresponding to this term, the residual series e t series.

Similarly, if you are using it for the forecast, the error of the forecast that we just obtained there, that error term becomes a residual series. So, for the test data, you have generated e t plus 1, e t plus 2, e t plus 3, etcetera up to all the value, all the remaining values. So, you have essentially the residual series available with you, after you apply the model.

Now, for the validation of the model, we test this residual series that you so obtained. That means, you applied the model for the remaining N by 2 values or whatever number of values that you have chosen for validation, corresponding to each of the term you obtain either the residual or the forecast error. So, e t sequence is known, on this e t sequence, now we carry out all the tests.

What are the tests? The test if you if you recall, when we formulated these models, we wrote the residual or the noise term e t and said that the noise term should have a zero mean that is the first assumption of the model, next that it should be divide of periodicities and it should be uncorrelated. So, we do three primary test, primarily three test we do, one is to test the series that we have generated, namely e t series. This series has a zero mean and that the series is divide of any periodicities and also that the series is uncorrelated.

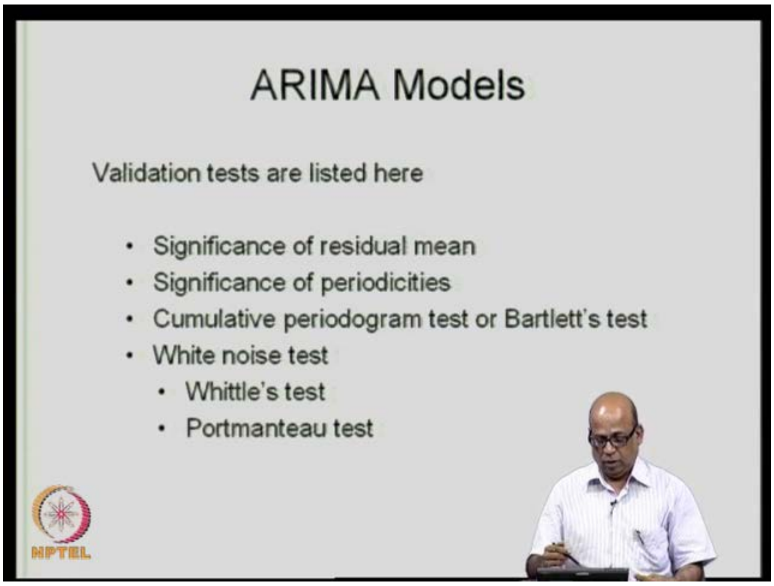
(Refer Slide Time: 26:57)



So, we perform the test, so examine whether these assumptions that we have made that is the residual series has a zero mean, no significant periodicities are present in the residual series and that the residual series is uncorrelated.

How do we formulate the residual series? As I said, this is e t is the residual is equal to X t which is a known data, minus you have this term corresponding to the simulated data from the model. So, this is phi j X t minus j plus theta j e t minus j. So, this is how you calculate the e t term, so you have the sequence of e t's now.

(Refer Slide Time: 27:51)

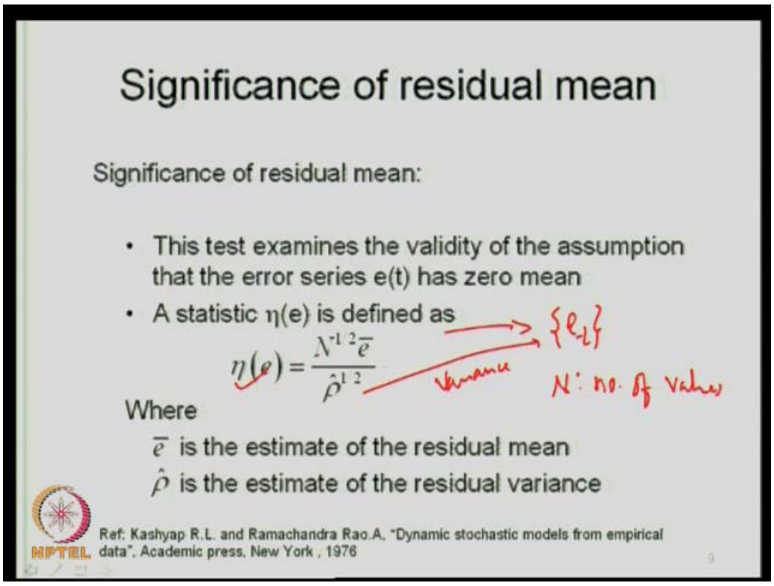


On this sequence of e t, we do the following validation test, one is significance of the residual mean that, what what is it that we want to test here that, the mean of the residuals that we so obtained have a mean of is zero, the mean is zero. But obviously, it will not be exactly equal to 0, therefore the mean should be not far away from 0. So, we say that the mean is not significantly different from 0 that is the test that we make in this.

Similarly, significance of periodicities, you may still have periodicities present in the data when you do the spectral analysis, the spikes may still appear. But the periodicities that you identify on the residual series when you carry out the spectral analysis must be insignificant; all of them must be insignificant.

Now, in this we do the cumulative periodogram test for this and then we also do the white noise test to make sure that the series is uncorrelated. In the white noise test, we will carry out whittle's test and portmanteau test, typically most of this test we formulate in appropriate statistic and then knowing the knowing that that statistic follows either way F distribution or t distribution corresponding to critical values of F and t, we decide whether the particular series that we have passes the test or not. Exception to that is the cumulative periodogram test where all the periodicities are tested at once in one go, we will see the details of this now.

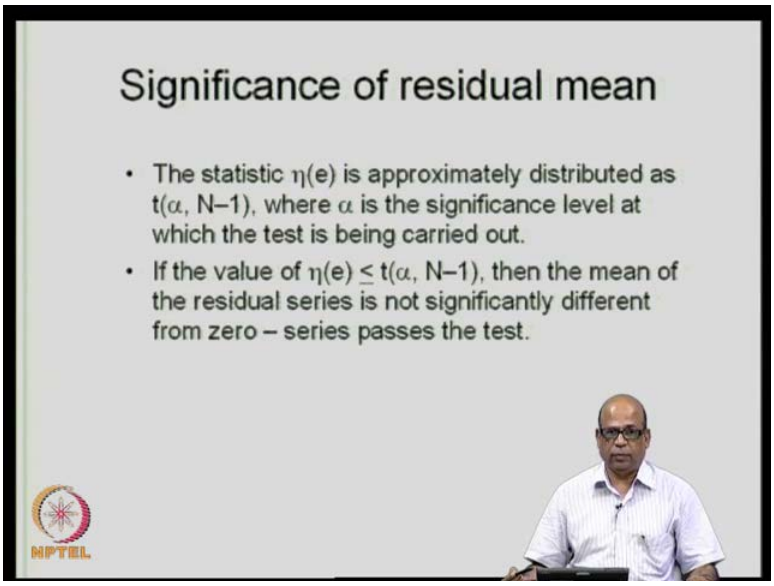
(Refer Slide Time: 29:41)



So, for the significance of residual mean, it examines the validity of the assumption that the error series e t has a zero mean or the mean of the e t is not significantly different from zero. So, we define I follow Kashyap and Rao's book here, this is a book, reference is given here.

We form a statistic eta e as N to the power half e bar by rho cap 1 by 2, remember rho cap here is not correlation, but it is a residual variance. So, you had the residual series and the residual series, you have calculated the mean that is the residual series is e t, this is given now and then e bar is associated with this e t and rho cap is the variance of this e t and N is the number of values, N is the data sample length, so you get N of e.

(Refer Slide Time: 31:00)

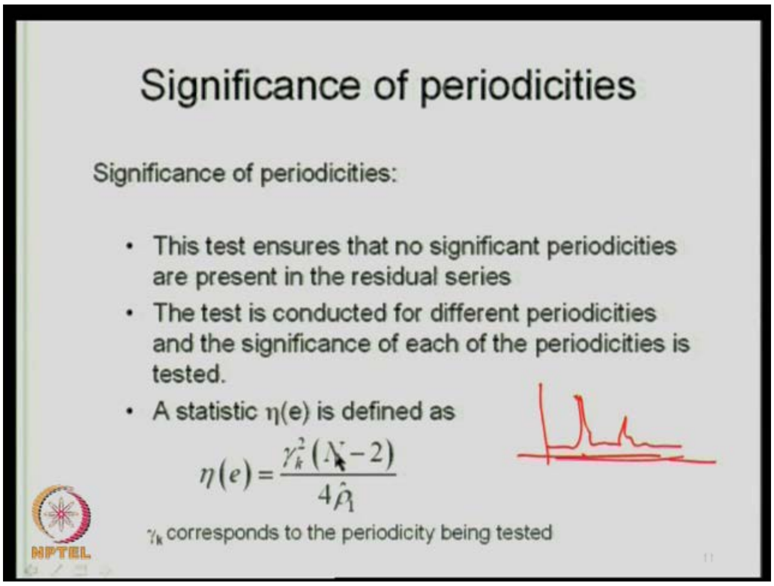


Once you get eta of e i am sorry eta of e, eta of e is known to be approximately distributed as t alpha N minus 1 where alpha is the significant level for example, you pick a 95 percent or 99 percent and then you get the t alpha comma N minus 1, N is the number of data.

If the value of eta e that you so compute using this is less than the critical t value, this is a critical t value corresponding to the level of significance that you choose. If this value is less, then the mean of the residual series is not significantly different from 0 and then we say that the series passes the test.

It is pretty simple as you can see the entire series is considered first, residual series e t and then you formulate for the entire series. You formulate one value of eta e which is your statistic and then look at, compare the eta is so, you so compute with the t distribution of a specified alpha value and then look at whether this eta e value that you have computed is in fact, less than the critical t value and then we say the series passes, so that was the test for significance of the mean.

(Refer Slide Time: 32:21)

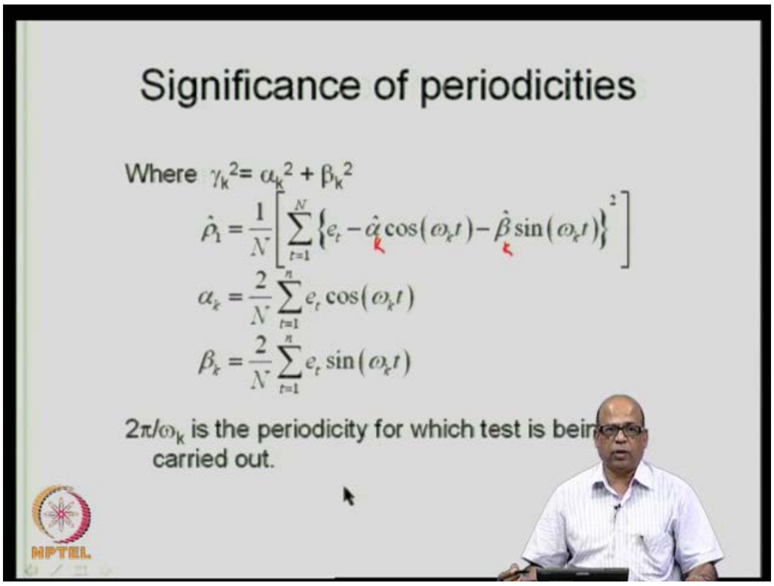


Now, the mean may be 0, but still you may have periodicities present in the residuals and that residual series that you obtain from the model should be divide of any periodicities, therefore we look for significance of the periodicities. And in fact, the residual series should not have any significant periodicities present in the data. Now, I will discuss two test, both of the test are valid, but one is slightly superior to the other as we presently state.

So, the first states again you formulate a statistic eta e is equal to gamma k square into N minus 2 by 4 rho 1 cap, remember we are doing the test for periodicities here, we pick one periodicity at a time. So, the test is conducted for different periodicities, let us say that the error series that you get has some periodicities present, let us say that this is a residual series and then you see that there are certain periodicities here like this. So, we pick the periodicity corresponding to this particular omega value first and then carry out this test.

So, the k that I am mentioning here corresponds to the particular periodicity that you want to test, one at a time we are testing. So, corresponding to that particular value of k, you calculate gamma k square and rho 1 again is the variance of the residual series. So, we know rho 1 cap and gamma k, we compute simply alpha k square plus beta k square as we did in our spectral analysis, N is the total number of values that you have.

(Refer Slide Time: 34:17)



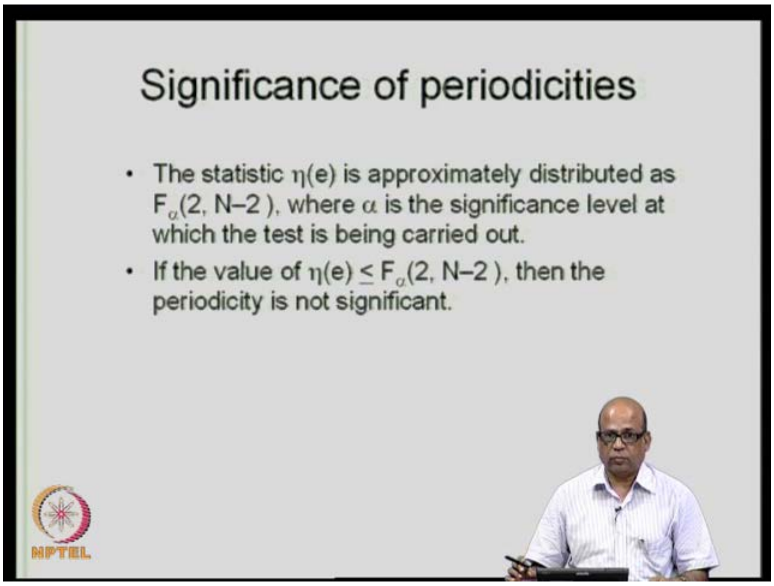
So, gamma k square is equal to alpha k square plus beta k square for the particular value of k, as I mentioned here gamma k corresponds to the periodicity being tested (Refer Slide Time: 34:26). So, in this case we may be testing for this particular value and the omega k that we do in out spectral analysis, let us say that this is omega k we are writing and this is I k with our notations. We pick up that particular omega k and then for that particular k, we compute gamma k.

Then rho 1 cap I am sorry I mention this as the variance, rho 1 cap is computed based on your e t, this is actually minus alpha 1 cos omega k t, this is corresponding to that particular k value, omega k minus beta cap sin omega k t whole square. And alpha k we know 2 by N, this is from our spectral analysis, so alpha k and beta k you get directly from the spectral analysis.

So, rho 1 cap is obtained from this for that particular periodicity which we are examining now. I again repeat. This test, test one periodicity at a time. So, we pick that particular omega k and then corresponding to that k, we calculate gamma k and similarly, rho 1 cap is known. See, here your all these values are known, alpha k is known, alpha cap is alpha k actually and then you have the beta k value which is also known, for completeness sake let me make this correction. This is alpha k cap and this is beta k cap, for that particular k and then you calculate rho 1 cap (Refer Slide Time: 36:05). So, rho 1 cap is known in this statistic gamma k is known, rho 1 cap is known and therefore, you can compute the statistics.

Now, the periodicity for which the test is being carried out is 2 pi by omega k, let us say you obtain a 12 month periodicity even in the residuals, then the corresponding value of omega k is 2 pi by that particular periodicity.

(Refer Slide Time: 36:39)



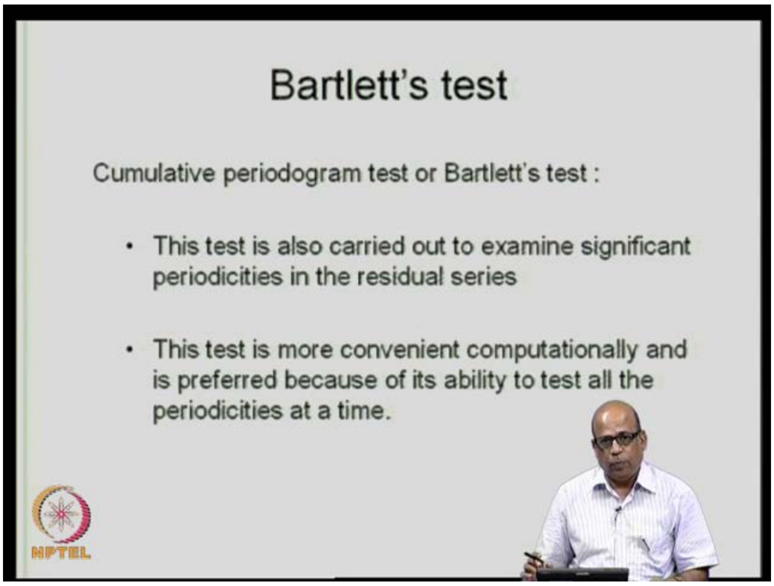
So, the statistic eta e is approximately distributed as F 2 N minus 1 N minus 2 where alpha is the significance level.

The F distribution tables, if you look up in any standard text book, they will give for each of the significance level that number of degrees of freedom and for the number of values. So, let us say you choose F alpha as F 0.95 or alpha as 0.95, 95 percent significance corresponding to 95 percent significance you compute these (Refer Slide Time: 37:15). In fact, this is just the same test that we did in our spectral analysis to identify the significance of periodicities and then you take the statistic that value of the statistic that you have calculated eta of e. If this is less than F alpha 2 N minus 2, then the periodicity is not significant.

Let us say that your spectral analysis like this shows of two or three different periodicities, first you test for this periodicity. If this is not significant, then naturally these periodicities may not be significant if they are all in decreasing order like this.

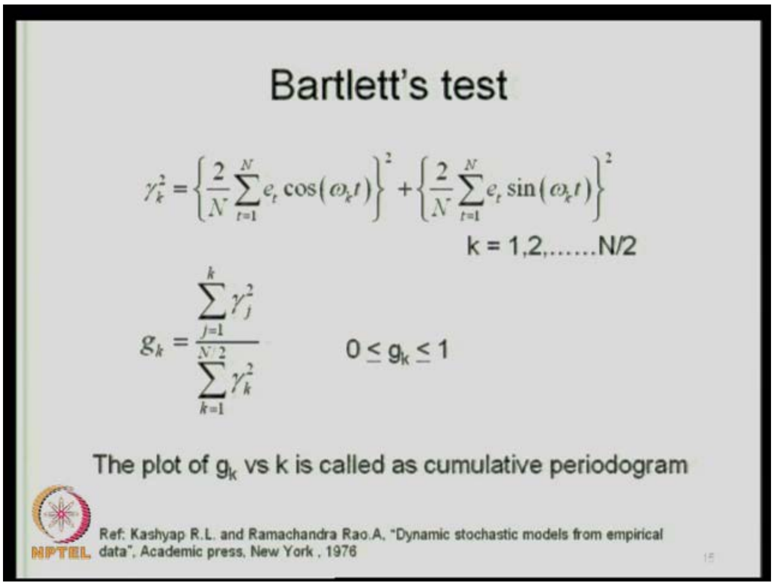
However, you may have a case where this may be significant if one of them is significant, then you have to test for the next periodicity also (Refer Slide Time: 38:10). So, you keep testing for the periodicities until you are satisfied that all the periodicities that are thrown up by the spectral analysis of the residual series are all insignificant, then the series passes the test, even if one periodicity is significant, then the particular model does not pass the test.

(Refer Slide Time: 38:45)



Now, for the significance of periodicities, we also have another test which is called as the cumulative periodogram test or the Bartlett's test. The advantage of this Bartlett's test is that unlike the previous test that I just explained, this test examines all the periodicities at one time rather than going by one periodicity to another and therefore, this is computationally convenient. So, the test is more convenient and it is preferred because of its ability to test all the periodicities at a time.

(Refer Slide Time: 39:35)

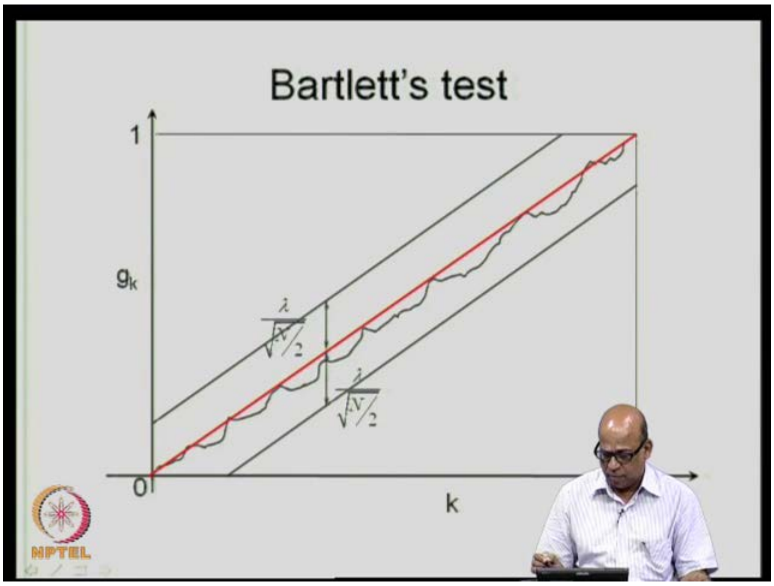


So, essentially what we do in this test is that we form a cumulative periodogram as follows; we define gamma k for k is equal to 1 to N by 2, as I mentioned this is a validation test. So, this is validation period, for the N by 2 values you calculate gamma k square, corresponding to each of the k, you know omega k from your spectral density, you you would have computed this (Refer Slide time: 40:00) and e t is that particular value of the residual series and you are summing over t is equal to 1 to N, t is equal to 1 to N and therefore, all the terms are known here (Refer Slide Time: 40:12).

So, for k is equal to 1, you get gamma k, k is equal to 2 you get gamma k and so on. So, like this for k is equal to N by 2, you have gamma k square, then you you determine g k as j is equal to 1 to k summation gamma j square. So, from 1 to k, you are adding up gamma j square. So, g 1 you have one term, g 2 you have two terms, 3 you have three terms and so on, like this you calculate summation of this divided by over the entire period, k is equal to 1 to N by 2 gamma k square.

So, the plot of g k verses k is called as the cumulative periodogram. So, you know how to compute g k, all values are known here, therefore you can calculate gamma k. Once you know gamma k, you sum up to k and then get g k by normal I think with respect to the entire sum here. And because of this nature, g k varies from 0 to 1; the maximum value it can take is 1.

(Refer Slide Time: 41:40)



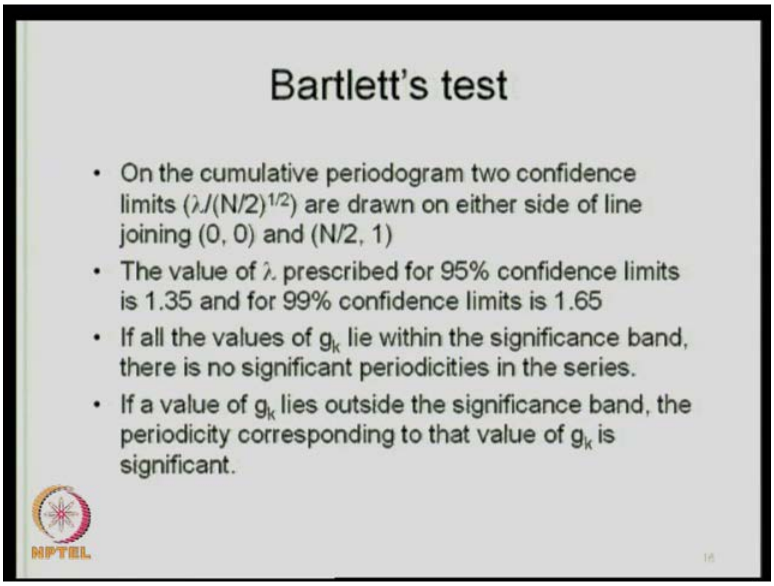
Then we plot k k on the x axis and g k on the y axis, now I will explain this with respect to this figure. So, you computed g k and you have the k values, so this is the plot, this black line you are seeing is the plot of g k verses k for the residual series (Refer Slide Time: 41:45).

Now, on this we have used N by 2 values and the maximum value of g k is 1. So, draw a line between 0 and N by 2,1, this corresponds to a frequency of 0.5, so 0.5, 1 on the frequency diagram. So, draw a line, this is the red line here, on this line on either side of this red line, you draw the confidence bands, the confidence band is gamma that is lambda divided by root of N by 2. So, this is lambda by root N by 2 on either side of the line so drawn. So, like this you get a band band now, this is a band.

If your cumulative periodogram that you have drawn, any part of that goes above or below, goes beyond the bound that you have drawn, then the series does not pass the test. In fact, it means that corresponding to that particular, let us say that it was beyond this particular band at a particular k value, the periodicity corresponding to that particular k value in the residual series is significant. If the periodogram completely lies within this band, then it is not significant, so this is how we examined the significance of the periodicities.

So, let me summarize that, so you draw first the cumulative periodogram by considering g k by constructing g k for each of the k values, you have N by 2 values and therefore, you construct g k verses k for N by 2 values.

(Refer Slide Time: 43:52)



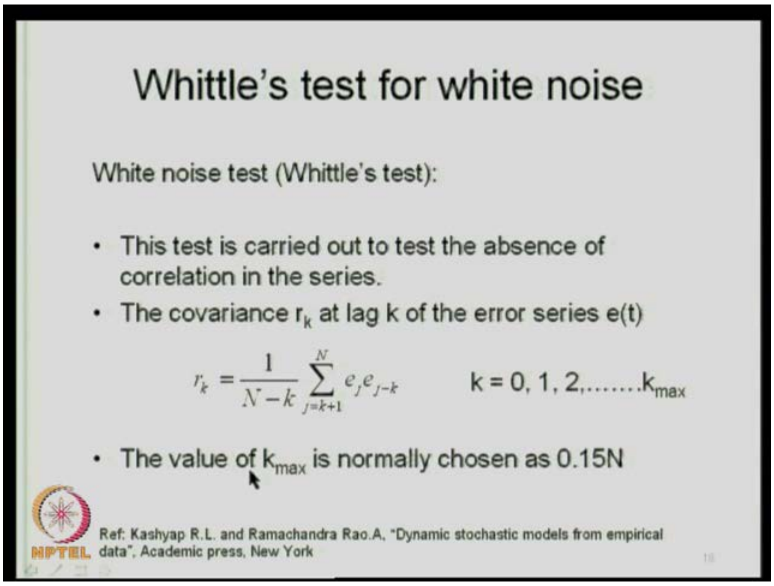
And then on the cumulative periodogram two confidence limits given by lambda by root N by 2 are drawn on either side of the line joining 0, 0 and N by 2, 1 N by 2, 1. What is 1? 1 is the maximum value of this and N by 2 corresponds to your k max that is the maximum lag, in fact this corresponds to a frequency of 0.5 (Refer Slide Time: 44:06).

Then what is this lambda value? The value of lambda for 95 percent confidence limit is 1.35 and for 99 percent confidence limit is 1.65. In fact, this test also I have taken it from Kashyap and Rao. So, you can refer to Kashyap and Rao, these are the values prescribed for this.

So, you know how to draw the bounds now, you know you have drawn the cumulative periodogram, you also know how to draw the bounds. Now, if all the g k values lie within the significance band, there is no significant periodicity present in the series, which means that the series passes the test. If a particular value of g k lies outside the significance band, the periodicity corresponding to that value of g k is significant and therefore, the series does not pass the test. So, this is how we we carry out the Bartlett's test or the cumulative periodogram test.

In most of the cases if the models are acceptable, the residual series typically gives, typically lies well within the bounds that we have drawn. It hardly comes very close to these bounds, in fact when I discuss the case studies it will be clear. However, just for your curiosity, what you can do is, you draw the cumulative periodogram also for the original data, and then you will see that there are several periodicities which are lying way above and below this particular band, either on this side or on this side, so indicating that the periodicities are significant.

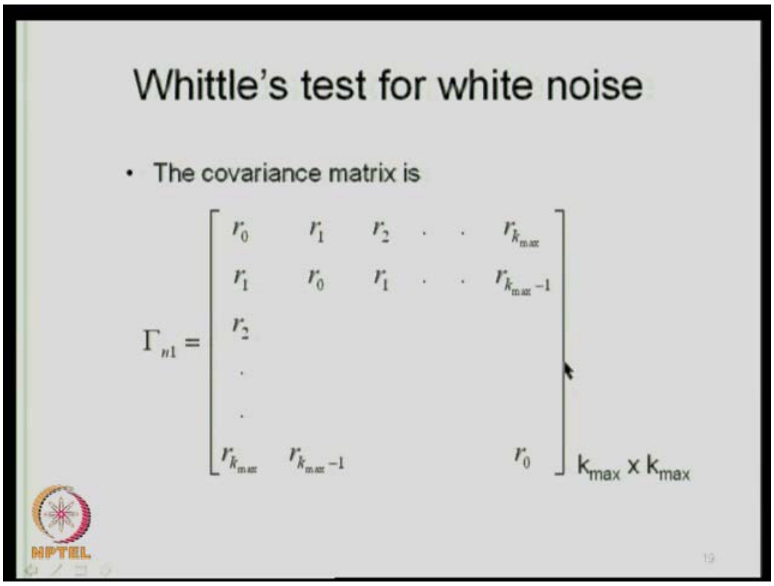
(Refer Slide Time: 46:10)



Now, we will see the white noise test. So, what we did is, for the periodicities we have now two test and typically we prefer the Bartlett's test, because of its ability to test all the periodicities in one go and also computationally it is very simple. That means, you simply calculate g k much like your correlogram, you formulate the periodogram with respect to lag k, you calculate g k values and plot this and immediately this significance comes out. Whereas, in the first two tests that we had, we formulated a statistic corresponding to each of the periodicities that we suspect to be significant, we have to make that particular test and therefore, the Bartlett's test is preferred for test of periodicities.

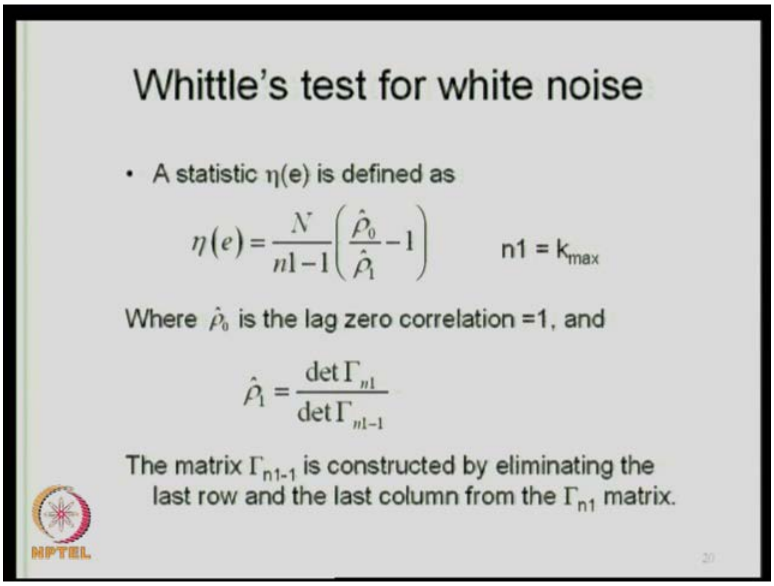
Now, we will test for white noise or the assumption that we made that the series is uncorrelated, the residual series that we formulate is uncorrelated. So, in this we have the whittle's test for white noise, whittles test we formulate again the covariance matrix, remember we are dealing with the e t series that is the residual series. So, the covariance r k at lag k of the error series e t is calculated, this is simply r k is equal to 1 by N minus k, j varies from k plus 1 to k, k goes from 0 to k max, k max can be typically 0.5 0.15N here, e j into e j minus k.

(Refer Slide Time: 48:06)



Then once you know the the covariance that is r k is given, you formulate the covariance matrix. So, covariance matrix will be also size k max by k max, the symmetrical r 0, r 1, r 2, etcetera, r k max, like this you formulate; tau n 1, this is a covariance matrix and we denote it by tau n 1, then we formulate a statistic.

(Refer Slide Time: 48:38)

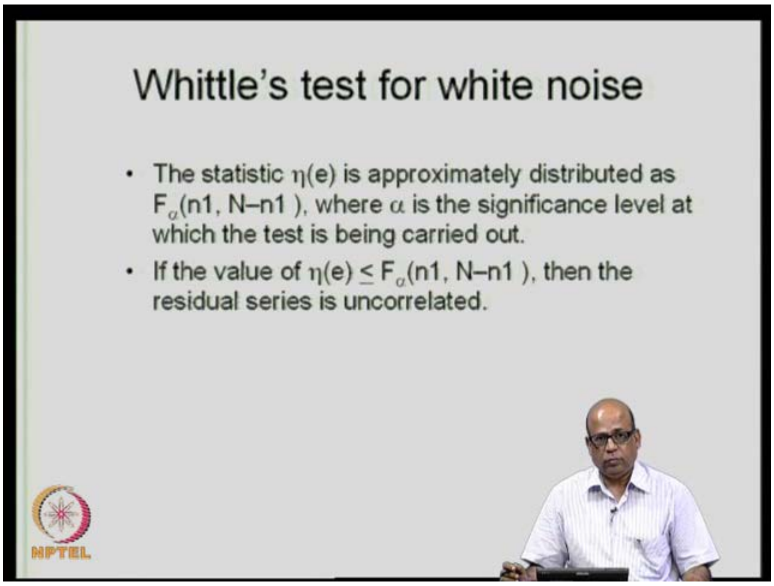


So, essentially we determined the covariance and formulated the covariance matrix. Using the covariance matrix, we define a statistic eta e is equal to N is given n 1 is k max minus 1 rho naught is lag zero correlation which is 1 and rho 1 cap are defined now minus 1.

Rho 1 cap here is determinant tau n 1 by determinant tau n 1 minus 1, tau n 1 is given here (Refer Slide Time: 49:06). So, this is a matrix, the determinant of this is used here, determinant tau n 1, now tau n 1 minus 1 is constructed by eliminating the last row and the last column from the t n 1 the tau n 1 matrix that is, you have this matrix, you delete this row I am sorry the last row and the last column, you delete this row and this column (Refer Slide Time: 49:15). So, you will have tau n 1 minus 1 and the determinant of that, that is what is used here, so you can formulate rho 1 cap.

So, you know N here, n 1 is known, k max rho naught is 1 and rho 1 cap is formed by rho 1 cap is calculated by taking the determinant tau n 1, tau n 1 is defined here, then take out the last row and last column, take the determinant and that is what defines determinant of tau n 1 minus 1. Therefore, rho 1 cap is known and therefore, you calculate this particular statistic.

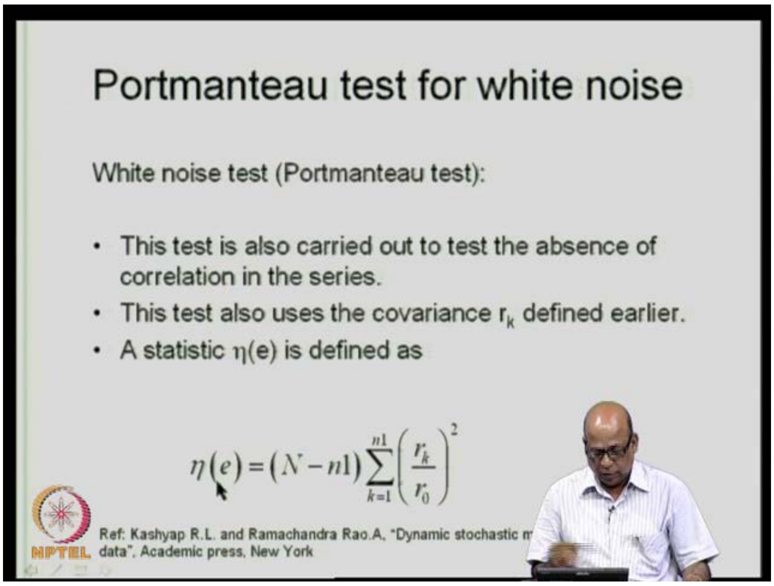
(Refer Slide Time: 50:19)



Once that statistic is known, this statistic is approximately distributed as F alpha n 1 N minus n 1, n 1 is your k max, typically taken as 15 percent of your N. So, as we did for earlier test using the F distribution, you fix your confidence level, typically 95 percent or 99 percent, n 1 is known, capital N is known which is a number of values and therefore, you know the critical value of F alpha.

If the value of eta e that you calculate, this value if the value that you calculate is less than F alpha, then the residual series is uncorrelated. So, this is what we do for white noise test, white noise is the series is uncorrelated, this is called as the whittle's test.

(Refer Slide Time: 51:17)

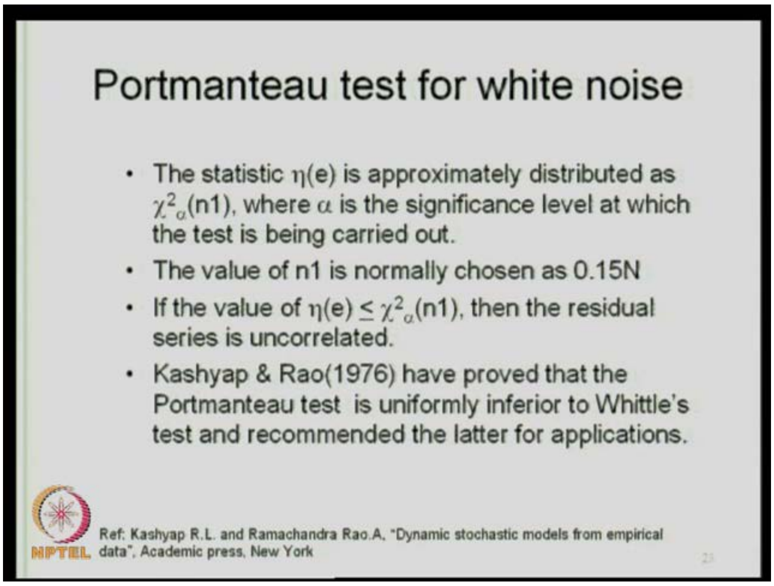


In fact, we have another test for examining whether the series e t is in fact, white noise or not, it is called as the portmanteau test.

Now, this test is also carried out to test the absence of correlation in the series and this also uses the r k as defined earlier that is the covariance matrix as we have defined earlier, this covariance matrix is yours (Refer Slide Time: 51:50). So, you know how to formulate the covariance matrix, from the covariance matrix we start defining another statistic in the portmanteau test.

So, using the covariance matrix r k here, your statistic eta e is defined, N is the number of values, n 1 is your maximum lag and r k is corresponding to the covariance of order k. And then r naught is the covariance of order 0, so k is equal to 1 to n 1, so you know how to calculate the statistic value.

(Refer Slide Time: 52:27)



Now, this statistic is approximated as, approximately distributed as chi square distribution with argument n 1. So, you fixed alpha and n 1, n 1 is your k max which is a maximum lags up to which you have gone. And from the chi square tables, you can get this value, the value n 1 is chosen again as 0.15 N. So, get the chi square, alpha n 1 value from the tables, if your eta e that you have calculated in less than the critical value of alpha, then the residual series is uncorrelated.

So, for the white noise test, that means to examine whether the series is uncorrelated or not, we have two tests, namely whittle's test and the portmanteau test. In both of them we use the covariance matrix, covariance matrix of the residual series and then formulate a statistic.

Now, between these two Kashyap and Rao have proved, the same text book that I have been mentioning, they have proved that the portmanteau test is uniformly inferior to whittle's test and we prefer the latter for applications. So, while there are two test that I have mentioned, if you if you are in a fix which one to use, you always go for the whittle's test as as shown by Kashyap and Rao.

Now, before we go to the applications, essentially what we have done is that we have formulated the models, ARIMA type of models and then we have made the test and then we now start applying to different case studies. So, let us summarize what we did on the choice of the ARIMA models, this is not just what I have covered in this lecture, but also on the previous two lectures.

We formulate the ARIMA models after differencing the series, ARIMA that is autoregressive integrated moving average models that order of differencing, whether it is a first order or second order differencing depends on the non-stationarity present in the data. So, you do the differencing essentially to convert the data into a stationary data, once you have the data as the stationary time series, then you apply this ARIMA type of models.

Now, in the ARIMA models, you know how to estimate the number of AR parameters and the number of MA parameters. In the event that you are not very clear about, how many AR terms to use, how many MA terms to use etcetera, you form the candidate models, a large number of candidate models and then apply this candidate models to the observed time series, estimate the parameters and calibrate the models using part of the series, typically we use half the length of the data to calibrate the series and use these models for validation for the remaining half, remaining part of the data. When you apply these models for the remaining part of the data, you get the residual series that is e t series.

We do the test of validation on the residual series, we do essentially three different tests, one is to test whether the residual series has a zero mean or the mean of the residual series is not significantly from significantly different from zero that is the first test that we do. Then we test whether the residual series, the that we have obtained by applying the model or is divide of any significant periodicities. So, you may come up with some periodicities, you test for the periodicities, we sort two tests for periodicity, for significance of periodicities and we normally prefer the cumulative periodogram test for significance of periodicities.

Then we also examine whether the series e t that we have obtained is, in fact white noise or it is uncorrelated. For that again we had two tests, the portmanteau test and the whittle's test, it is shown in the standard text book, Kashyap and Rao that the whittle's test is better than the portmanteau test.

So, we apply the models, make all these tests and then make sure that the model that you have chosen, either based on the maximum likelihood criteria or the minimum mean square criteria passes all these tests and then choose that particular model for application. We will continue the discussion and specifically, we will see how to apply these models, whatever procedures that I have explained so far, we will see the applications of these in the coming lectures. Thank you for your attention.