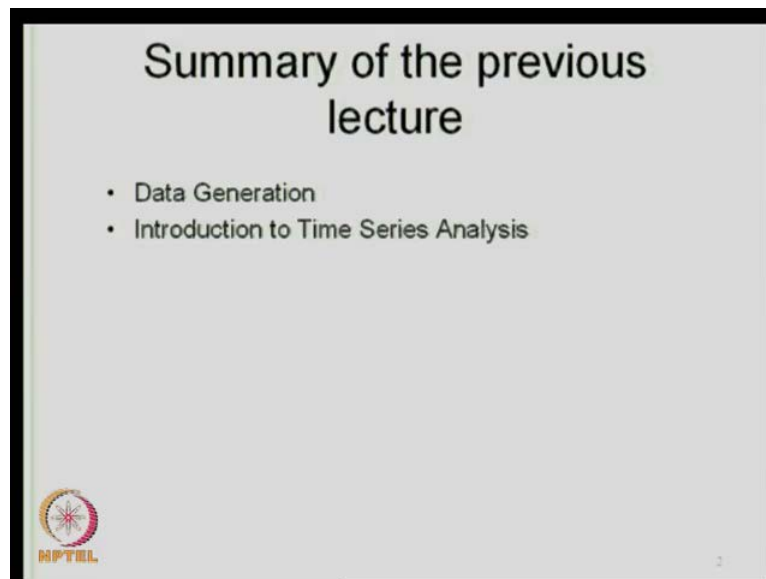


**Stochastic Hydrology**  
**Prof. P.P. Mujumdar**  
**Department of Civil Engineering**  
**Indian Institute of Science, Bangalore**

**Lecture No. # 10**  
**Time Series Analysis – I**

(Refer Slide Time: 00:26)



Good morning and welcome to this the tenth lecture of the course stochastic hydrology. If you recall in the last lecture, we covered the methods of data generation, where we dealt with generation of data, when the distributions are known. For example, we discussed how to generate data from a normal distribution, exponential distribution, gamma distribution, and so on. So, essentially what we do there is that given the cdf, we assign a uniformly distributed random number, and equated to the analytical expression of cdf, and solve for Y. For example, capital F of Y is equal to R u random number generate uniformly distributed random number in the interval 0 and 1, and then solve for Y.

In cases where explicit solution for Y is not possible, then we have other methods for example, in the normal distribution case, we have standard normal deviates tabulated values of those we use those tabulated values, and then solve for Y, and these are the

generated values. Then we went on to discuss the time series concepts, we have just introduced the concepts of time series, if you recall a series of observations on a random variable across time constitutes a time series. For example you have stream flow data over the last about 50 years or stream flow observations made every month for the last about 50 years, and this series of observations constitutes a time series. So, we will continue the discussion today on time series analysis, and then introduce also some data forecasting techniques using of the methods of time series analysis.

(Refer Slide Time: 02:34)

The slide is titled "Time Series Analysis" and contains the following content:

- Sequence of values of a random variable collected over time
- Discrete time series; Continuous time series
- Realization; Ensemble
- Hydrologic time series composed of deterministic and stochastic components

The equation  $X_t = d_t + \varepsilon_t$  is displayed on the left side of the slide.

On the right side, there is a graph with a vertical axis labeled  $x_t$  and a horizontal axis labeled "Long t". The graph shows a fluctuating line representing a time series.

In the bottom left corner, there is a logo for NPTEL (National Programme on Technology Enhanced Learning).

A presenter is visible in the bottom right corner of the slide frame, standing in front of the content.

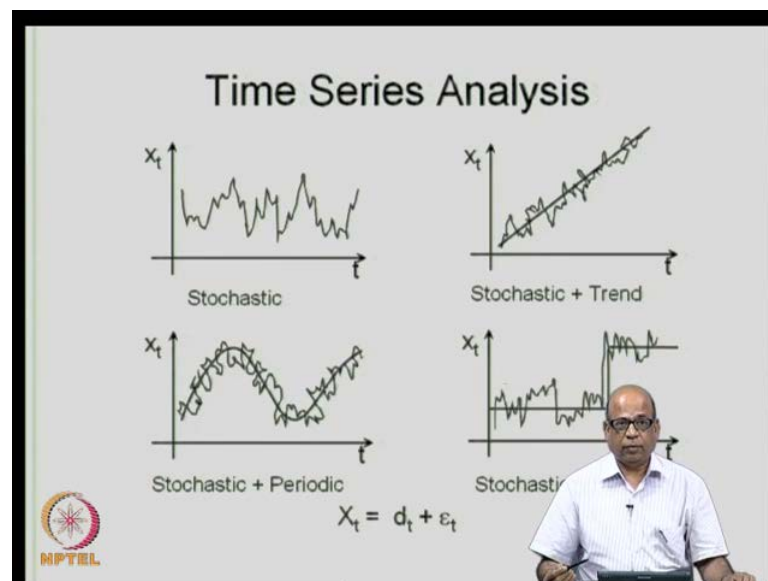
So, if you recall in the last class we first define, what is the time series as as I just mentioned sequence of values of random variable collected over time. Although I must mention that although we keep on emphasizing collection over time it can also be a space for example, we may have rainfall sequences across a space now the techniques that we are discussing are also applicable in some cases, when we have large observations across space. Now the time elements over which we are making the time intervals over which we are making the observations, can be discrete or continuous, and they deal they lead us to discrete time time series and continuous time time series.

We also introduce the concepts of the realization and an ensemble realization is a single time series that is if we have one time series one set of observations across the time that is called as a realization. So, this is a realization of the time series for example, we may have sequence of stream flows observed between 1970 and 2000 that constitutes one

realization like that we may have several such realizations for example, stream flows between 1940 to 1970 ,1970 to 1990 , 1990 to 2005, etcetera.

Like this you have several such realizations possible the collection of all such available realizations is called as an ensemble. Generally in the case of hydrology we do not have large number of realizations. In fact, if we have one realization with a long series of data then if necessary where in cases where necessary we split the time series into several realizations for example, if we have data from 1970 to 2000 we may take data from 1970 to 1990 for example, as one realization 1990 to 2000 as another realization and so on. In general in hydrology we may be dealing with a single realization the hydrologic time series. In fact, in general any time series may consist of a deterministic component and a random component for example, you may have a long term mean of hydrologic hydrologic variable around which there are random fluctuations. So, we would like to model the time series by capturing this deterministic component and by modeling the random components.

(Refer Slide Time: 05:48)



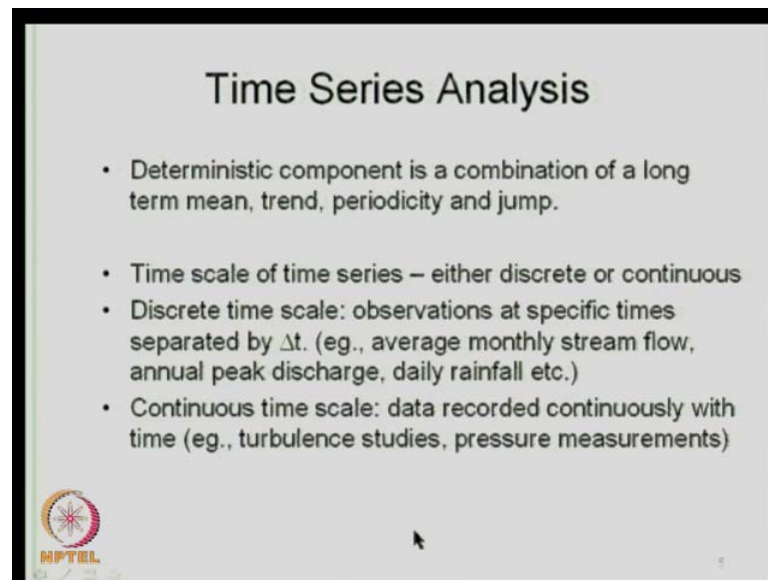
So, in general we write the time series  $X_t$  as consisting of a deterministic component  $d_t$  and a random component  $\varepsilon_t$ . This is what we discuss in the last class and we have also seen that a time series may in general consist of a purely stochastic component; that means, there is no deterministic component at all the values are all randomly fluctuating

like this in which at a anytime  $t$  if you take the observations between time  $t$  and time  $t$  plus one have no relationship with among themselves.

On the other hand you may have a deterministic component which is a trend, let say in this particulars case an increasing trend. So, around an increasing trend the values are randomly fluctuating if you look at the global average temperature. So, the last let say 50 years or some such thing that will have an increasing trend and the values are randomly fluctuating around that. So, you have the stochastic component as well as a deterministic component which in this case is the trend. Then you may have a periodic component the values are randomly fluctuating. So, you have a stochastic component around a periodic component. If you look at let us say monthly stream flows at a particular location in monsoon regions. Like ours then you may have such periodic components prominent that then you may also have cases where there is a jump as I mentioned in the last lecture, these jumps may occur because of sudden large scale changes occurring in the catchment.


Let us say you are talking about stream flows in a basin and a large scale. Either deforestation occurs or large scale fire occurs or a huge or earth quake occurs in that which in a short while short period. Disturbs the hydrologic balance then the hydrology starts operating at a different level altogether this may be either a jump or it may be a drop what are operating at this level on an average starts operating at certain other level on an average. So, you have a stochastic component and a jump component in this. So, once we capture the deterministic component we write  $X_t$  is equal to that deterministic component at time period  $t$  and a random component at time period  $t$ . This deterministic component itself may be a combination of the long term mean around, which fluctuations are occurring a trend a periodicity and a jump.

(Refer Slide Time: 08:25)



**Time Series Analysis**

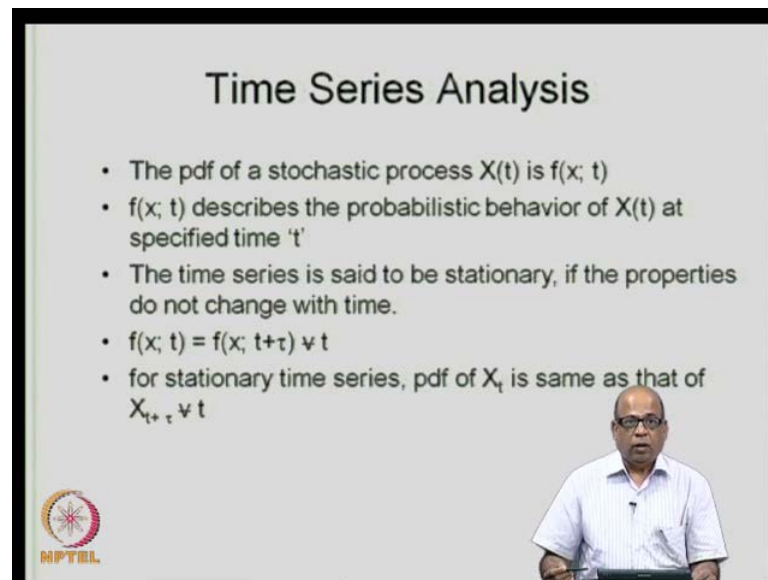
- Deterministic component is a combination of a long term mean, trend, periodicity and jump.
- Time scale of time series – either discrete or continuous
- Discrete time scale: observations at specific times separated by  $\Delta t$ . (eg., average monthly stream flow, annual peak discharge, daily rainfall etc.)
- Continuous time scale: data recorded continuously with time (eg., turbulence studies, pressure measurements)

 MPTEL

Now, the discrete time scale observations in hydrology arise mainly when we are talking about let say monthly stream flows. Although the observations may be done everyday for the analysis we may consider the aggregated values of the stream flows over a month and therefore, we may be talking about monthly stream flow average monthly stream flow or annual peak discharge; that means, every year we have one discharge value, which is chosen as the maximum discharge of all the discharges that have occurred over that year daily rainfall. We measure the rainfall let say 8 am in the morning today and next again tomorrow at 8 am we measure the rainfall and that such sequence of observations constitute a discrete times scale time series.

On the other hand, you may also have continuous times scale time series, where data is recorded continuously with time for example, in turbulence studies or in experiments where we are measuring pressures. So, these measurements are across continuous times in most cases, where we are dealing with hydrologic time series, we will be dealing with discrete time scales, because we have observations at discrete times that is a first thing and also the analysis is required at discrete time steps for example, we may we will be interested in forecasting for the next month or next 10 day period next day and so on. So, we will be dealing with discrete time, time series.

(Refer Slide Time: 10:25)



**Time Series Analysis**

- The pdf of a stochastic process  $X(t)$  is  $f(x; t)$
- $f(x; t)$  describes the probabilistic behavior of  $X(t)$  at specified time 't'
- The time series is said to be stationary, if the properties do not change with time.
- $f(x; t) = f(x; t+\tau) \forall t$
- for stationary time series, pdf of  $X_t$  is same as that of  $X_{t+\tau} \forall t$

MPTEL

The pdf of  $X_t$  now remember  $X_t$  becomes a random variable. So,  $X_t$  has its own pdf  $X_{t+1}$  has its own pdf  $X_{t+2}$  has its own pdf etcetera. So, across time you are having collection of random variables and the process behaves in a certain probabilistic way at time period  $t$ . Which may be different from its behavior at time period  $t$  plus let say  $\tau$  and therefore, the pdf is now denoted as  $f$  of  $X$  comma  $t$  now  $f$  of  $X_t$  describes the probabilistic behavior of  $X_t$  at specified time  $t$ . If we have the pdf of  $f$  of  $x$  comma  $t$  that is pdf of  $x$  comma  $t$  the same across time period. As time changes the pdf does not change then the time series is said to be stationary that is if the pdf at  $f$  the pdf of  $X_t$  is equal to  $f$  of  $X_{t+\tau}$  for all time  $t$ ; that means, we are considering a lag of  $\tau$  and then if the pdf remains the same across the time for all time period  $t$  then the process is said to be strictly stationary.

So, for stationarity of time series pdf of  $X_t$  is the same as pdf of  $X_{t+\tau}$  for all  $t$  where  $\tau$  is the time lag that we are considering between the two random variables. So, this picture makes it here you have several realizations this is realization 1 realization 2 etcetera. So, you have  $m$  realizations and this collection of  $m$  realizations is the ensemble. So, you can talk about properties across time or properties across realizations.

(Refer Slide Time: 12:19)

**Time Series Analysis**

Time average for a realization

$$\bar{X}_1 = \frac{\sum_{j=1}^n \{X_j(t)\}_1}{n}$$

n is no. of observations

Ensemble average at time t

$$\bar{X}_t = \frac{\sum_{i=1}^m X_i(t)}{m}$$

m is no. of realizations

The slide also features three plots of time series data labeled {X<sub>j</sub>}<sub>1</sub>, {X<sub>j</sub>}<sub>2</sub>, and {X<sub>j</sub>}<sub>m</sub> against time t, with a vertical line at t<sub>1</sub>. The presenter is visible in the bottom right corner of the slide frame.

So, when you are looking at one time series and then you are looking at properties across time then we have time properties of a given realization for example, you may look at the time average of a realization, which is you are taking the average across time and therefore, you can write this as for the first realization for realization number 1 you can write  $\bar{X}_1$  as sum of all the observations across time for the same realization divided by the number of observations, so that is what you are writing formally here that is  $\bar{X}_j t$  for the realization one across all j, j is equal to 1 to n you have n number of observations.

So, essentially you are taking the average across time for this realization these are called as time properties or in this particular case, it is called as the time average for a realization. Similarly, you can look across the realizations at a given specified time t<sub>1</sub>; you can take averages across realizations in which case you talk about ensemble average at time t. So, that will be given by  $\bar{X}_t = \frac{1}{m} \sum_{i=1}^m X_i(t)$  that is you are summing  $X_1 t$  at t is equal to t  $X_2 t$  at t is equal to 1 and  $x_m t$  and so on, and divided by number of realizations we are essentially taking the average across realization this is called as an ensemble average.

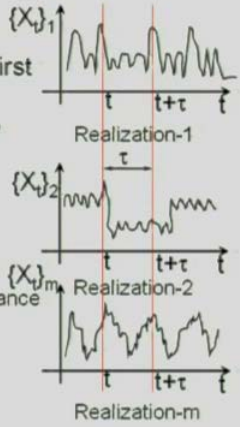
There is a concept of ergodicity, where the time average properties will be the same as the ensemble average properties in such a case, we call the process to be an ergodic process. So, for stationarity as said the process is strictly stationary if the pdf across time remain the same; that means, if you consider two time steps tau time steps apart that is t


and  $t + \tau$  the pdf at  $t$  and pdf at  $t + \tau$  remain the same for all  $t$ . So, even if you shift by  $\tau$  across time scale. The pdf remains the same then its called as strict stationary time series time series, which is strictly stationary.

(Refer Slide Time: 14:47)

### Time Series Analysis

- If  $\bar{X}_t = \bar{X}_{t+\tau}$  for all  $t$ , then the process is stationary in mean (first order stationary)
- If all the moments up to order 'f' are same for time  $t$  and  $t+\tau$ ,  $\forall t$  then the time series is weakly stationary of order 'k'
  - $k = 1$  Stationary in mean
  - $k = 2$  Stationary in mean & covariance
- For a strictly stationary time series,
 
$$f(x_1) = f(x_2) = \dots = f(x)$$





However in most applications we may not need to stick to stationarity. We then define a weak stationarity for example, if the pdf is not the same, but if the means across time are the same then we call it call the time series stationary in mean. So, in general if we have a time series if all the moments up to order  $f$  are the same for time  $t$  and  $t + \tau$  for all  $t$ , then the time series is weakly stationary of order  $k$  for example,  $k$  is equal to 1 it means that the moment up to order 1 is the same across time the first moment is mean. So, if you have  $X_t$  is equal to  $X_{t+\tau}$   $\bar{X}_t$  is equal to  $\bar{X}_{t+\tau}$  for all  $t$ , then it is stationary in mean or it is a first order stationarity.

Similarly,  $k$  is equal to 2 we will be dealing with moments up to the order of 2, which means the mean should be the same and the covariance should be the same. So,  $k$  is equal to 2, the first two stationarity in mean and covariance both together, it can be shown that if it is if the process is stationary in covariance, then the variance also remain the same. So, it is stationarity in mean and covariance also indicates that the process is stationary with respect to variance which means the variance is also constant across time. Remember here, when we say  $k$  is equal to 2, it must be both the stationary in mean as well as in covariance. If the process is stationary in covariance, but not stationary in



mean, then it is not weakly stationary of second order. Again for a strictly stationary time series you must have the pdf to be same pdf f of X 1, X 2, X etcetera, must be the same across time.

(Refer Slide Time: 18:02)

### Time Series Analysis

- Auto covariance
 
$$\gamma_k = \text{cov}(X_t, X_{t+k})$$

$$= E[(X_t - \mu)(X_{t+k} - \mu)]$$


$$\gamma_0 = \sigma_X^2$$
- Auto correlation between  $X_t$  and  $X_{t+k}$ 

$$\rho_k = \frac{\text{cov}(X_t, X_{t+k})}{\sigma_{X_t} \sigma_{X_{t+k}}}$$

$$= \frac{\text{cov}(X_t, X_{t+k})}{\sigma_X^2} = \frac{\gamma_k}{\gamma_0}$$

If process is stationary  
 $\sigma_{X_t} = \sigma_{X_{t+k}}$

$$\rho_0 = 1$$

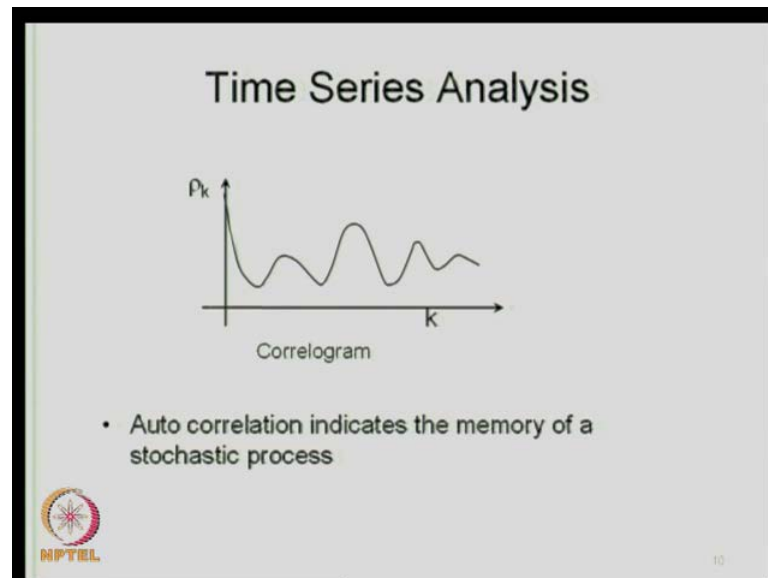


We come to the concept of auto covariance. So, from now onwards we will be dealing with stationary time series; so unless otherways specified we are always assuming stationarity with respect to mean and variance, so auto covariance recall that when we define covariance we defined it with respect to 2 variables x and y. Covariance between x and y now we are dealing with a single random variable which is let us say stream flow measured at different time states or the properties of a random variable at different time states X is a different random variable and X t plus K is a different random variable. So, we define the covariance between X t and X t plus k as expected value of X t minus mu X t plus k minus mu recall that covariance between x and y covariance of x comma y was expected value of x minus mu x into x y minus mu y. So, simply replacing x by Xt and y by X t plus k. So, from this definition then what happens when k is equal to 0. We are talking about covariance of X t with X t itself and that turns out to be in this particular case gamma naught is equal to sigma X square which is the variance.

Then we define the auto correlation what did we define the correlation as correlation coefficient as covariance between X and Y divided by sigma X sigma Y that is a normalized covariance. So, you much the same way we define the auto correlation as

covariance between  $X_t$  and  $X_{t+k}$  divided by  $\sigma_{X_t}$  and  $\sigma_{X_{t+k}}$  and for series which is stationary in mean and variance this comes out to be because  $\sigma_{X_t}$  is same as  $\sigma_{X_{t+k}}$  and therefore, this we write it as  $\gamma_k$  by  $\gamma_0$  because  $\sigma_{X^2}$  which is a variance which is nothing, but  $\gamma_0$  as you can see here.

(Refer Slide Time: 20:44)



Also when  $k$  is equal to 0 this becomes  $\gamma_0$  by  $\gamma_0$  therefore,  $\rho_0$  will be equal to 1. So, the lag 0 auto correlation is 1 the plot of  $\rho_k$  versus  $k$  is as defined by this. So, we can talk about correlation with respect to various lags  $k$ , this  $k$  is called as the time lag or simply the lag. So,  $\rho_k$  in some sense gives a measure of dependence of the variable  $X_{t+k}$  on the variable  $X_t$  which indicates a measure of dependence. So, if you plot  $\rho_k$  versus  $k$  this indicates this is variously called as correlogram or auto correlation function. This actually indicates the memory of the process; that means, how far into the time it remembers for example,  $X_{t+k}$  depends on  $X_t$  let's say  $k$  is equal to 12 in the case of monthly stream flows. Let's say you are talking about the flow during the June of a particular year depending on the flow during the June of the previous year and if the time series consist of monthly stream flows then we are talking about a lag of 12 months.  $k$  is equal to 12 and then we may have associated  $\rho_k$  value here  $k$  is equal to 24.

The dependence of June month flow on June months flow of 2 years before that is 24 months before and. So, on. So, as k increases you may have smaller and smaller rho k values, on the other hand you may also have a periodic oscillation of rho k value. So, essentially rho k or the auto correlation function indicates the memory of the process how far into the time is it able to remember what has happened before and therefore, it becomes a handy tool for us to model time series. So, the correlogram gives an important information about the process. So, whenever we are dealing with time series analysis. We first plot the time series as observed and then the next step is to compute the correlations rho k for various lags k and plot the correlogram correlogram immediately gives us important information about the time series which we will see through some examples.

(Refer Slide Time: 23:32)

**Time Series Analysis**

- Auto covariance matrix

$$\Gamma_n = \begin{matrix} & \begin{matrix} X_1 & X_2 & X_3 & \cdot & \cdot & X_n \end{matrix} \\ \begin{matrix} X_1 \\ X_2 \\ X_3 \\ \cdot \\ \cdot \\ X_n \end{matrix} & \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdot & \cdot & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdot & \cdot & \gamma_{n-2} \\ \gamma_2 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \gamma_{n-1} & \gamma_{n-2} & \cdot & \cdot & \cdot & \gamma_0 \end{bmatrix} \end{matrix}$$

$\Gamma_n$  is symmetric and +ve definite matrix

Having defined the auto covariance and the auto correlation we defined the auto covariance for k that is gamma k and auto correlation rho k. We now formulate the 2 important matrices the auto covariance matrix and the auto correlation matrix by varying k. So, we talk about the auto covariance matrix for various values of k . Similarly auto correlation matrix for various values of k we form the auto covariance matrix by writing X 1 X 2 etcetera, Xn these are the n random variables across n time steps and here again X 1 X 2, etcetra, X n along the rows. So, we write the elements of this matrix as consisting of the covariance between X 1 and X 1, which is gamma naught X 1 and X 2 . Which is gamma naught X 1 and X 3 which us gamma 2 and so on, X n and X n which

will be  $\gamma_{n-1}$  going by the definition here. So, we are writing  $X_1$  and  $X_1$  and  $X_2$  which becomes  $1 + 1$ . So,  $k$  becomes  $1$  there and therefore, this becomes  $\gamma_1$ ,  $X_1$  and  $X_3$   $k$  becomes  $2$  there and this becomes  $\gamma_2$  and so on.

So, this is how we write the first row similarly second row  $X_2$  and  $X_1$  which is same as  $X_1$  and  $X_2$  which is a  $\gamma_1$ ,  $X_2$  and  $X_2$  which is  $\gamma_0$  and. So, on  $\gamma_{n-2}$ . Like this we write for the  $n$ -th row  $X_1$  and  $X_1$  which is same as  $X_1$   $X_n$  and  $X_1$  which is  $\gamma_{n-1}$  and so on; we write the last row this is a  $n$  by  $n$  matrix this is the symmetric matrix, it is symmetric about the diagonal, and it is also a positive definite matrix it can be shown that the auto covariance matrix is also a positive definite matrix. What is a positive definite matrix; one definition of the definition of positive definite matrix is that all the eigenvalues of the matrix are positive, then it called as a strictly positive definite matrix.

Now we will divide this by  $\gamma_0$  what is  $\gamma_0$   $\gamma_0$  is simply the variance. So, we will divide all the elements by  $\gamma_0$ , then we will get the auto correlation matrix remember here auto correlation is simply  $\gamma_k$  by  $\gamma_0$ . So, all the elements of  $\gamma_k$  we simply divide it by  $\gamma_0$  we get the auto correlation matrix. So, the auto covariance matrix is denoted by  $\tau_n$  or  $\Gamma_n$ , this is  $\Gamma_n$ .



(Refer Slide Time: 26:45)

## Time Series Analysis

- Dividing the matrix  $\Gamma_n$  by  $\gamma_0$ , we get the auto correlation matrix  $P_n$

$$P_n = \frac{\Gamma_n}{\gamma_0} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdot & \cdot & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdot & \cdot & \rho_{n-2} \\ \rho_2 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \rho_{n-1} & \rho_{n-2} & \cdot & \cdot & \cdot & 1 \end{bmatrix}_{n \times n}$$

$P_n$  is symmetric and +ve definite matrix

The auto correlation matrix is denoted by  $P_n$ . So,  $P_n$  is simply  $\gamma_n$  by  $\gamma_n$  naught. So, the first element  $\gamma_n$  by  $\gamma_n$  naught that becomes  $1$   $\gamma_n$   $1$  by  $\gamma_n$  naught becomes  $\rho_1$   $\gamma_n$   $2$  by  $\gamma_n$   $1$   $\gamma_n$  naught becomes  $\rho_2$  and. So, on. So, this is how we get the correlation matrix auto correlation matrix  $1$   $\rho_1$   $1$   $\rho_2$  etcetera,  $\rho_n$  minus  $1$ . So, like this we write all the elements and formulate the auto correlation matrix. The auto correlation matrix  $p_n$  is also symmetrical and it is also a positive definite matrix, now the fact that the auto correlation matrix is a positive definite matrix comes in handy a positive definite matrix has the property that the fact that all the eigenvalues are positive it also implies that the minors of the matrix are all positive.

So,  $1$  by  $1$  the minor of  $1$  by  $1$  is  $1$  minor of size  $2$  by  $2$  is  $1$   $\rho_1$   $1$   $1$  the determinant of that should be positive the minor of size  $3$  by  $3$  that should be positive determinant of that should be positive and. So, on. So, if you look at the minor  $2$  by  $2$  that is  $1$   $\rho_1$   $1$   $1$  the determinant of that should be positive because  $p_n$  is a positive definite matrix and that turns out to be  $1$  minus  $\rho_1$  square must be greater than or equal to  $0$  and this indicates that  $\rho_1$  has to lie between minus  $1$  and plus  $1$ .

(Refer Slide Time: 28:25)

**Time Series Analysis**

- Because  $P_n$  is +ve definite

$$\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix} \geq 0$$

$$1 - \rho_1^2 \geq 0$$

$$-1 \leq \rho_1 \leq 1$$

We use the fact that  $p_n$  is a positive definite matrix subsequently when we do time series analysis and then generate several conditions. From the fact that it is a positive definite matrix, we have the definition for  $\gamma_k$ , which is the covariance auto covariance

function auto covariance which is given by expected value of  $X_t - \mu$  into  $X_{t+k} - \mu$ . Remember we are talking about series which is stationary in mean as well as variance and therefore, we take the same mean  $\mu$  here.

(Refer Slide Time: 28:53)

**Time Series Analysis**

- Sample estimates:
 
$$\gamma_k = E[(X_t - \mu)(X_{t+k} - \mu)]$$

$$c_k = \frac{1}{N} \sum_{t=1}^{n-k} (X_t - \bar{X})(X_{t+k} - \bar{X}) \dots \text{Sample estimate of auto covariance}$$

$$r_k = \frac{c_k}{c_0}$$

$$c_0 = S_{\bar{X}}^2 \text{ variance}$$

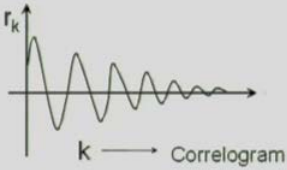
The slide also features a bar chart with the vertical axis labeled  $r_k$  and the horizontal axis labeled  $k$ . The chart shows several vertical bars of varying heights, representing the sample estimates of the auto-correlation function. A small video feed of a presenter is visible in the bottom right corner of the slide.

The same estimate for this is given by  $C_k$  is equal to summation  $t$  is equal to 1 to  $n$  minus  $k$ .  $X_t - \bar{X}$   $X_{t+k} - \bar{X}$  and  $r_k$  is given by  $C_k$  by  $C_0$ .  $r_k$  is the sample estimate for the auto correlation at lag  $k$ . So,  $\rho_k$  would have been defined as  $\rho_k$  which is a auto correlation we would have defined  $\rho_k$  as  $\gamma_k$  by  $\gamma_0$  and  $r_k$  which is the sample estimate of  $\rho_k$  is simply  $C_k$  by  $C_0$ .  $C_k$  is a sample estimate of the auto covariance and  $c_0$  is simply our sample estimate of the variance. So, when you have sample estimates and you take discrete values of  $k$  for example,  $k$  is equal to 1  $k$  is equal to 2 and so on, and plot the sample estimates of  $r_k$  versus  $k$  the sample estimates  $r_k$  and versus  $k$ , which is actually the auto correlogram or the correlogram you get spikes like this for example,  $k$  is equal to one may have a  $r_k$  value like this or  $k$  is equal to 2 may have like this etcetera.

(Refer Slide Time: 31:10)

### Time Series Analysis


- Auto correlation function ( $r_k$ )



k → Correlogram

If it is purely stochastic (random) series,  
 $\rho_k = 0, \quad \forall k = 1, 2, 3, \dots$   
 $r_k = \text{may not be zero (because } r_k \text{ is a sample estimate)}$

$r_k \sim \text{Normal Distribution} \left( 0, \frac{1}{\sqrt{N}} \right)$  For a random series

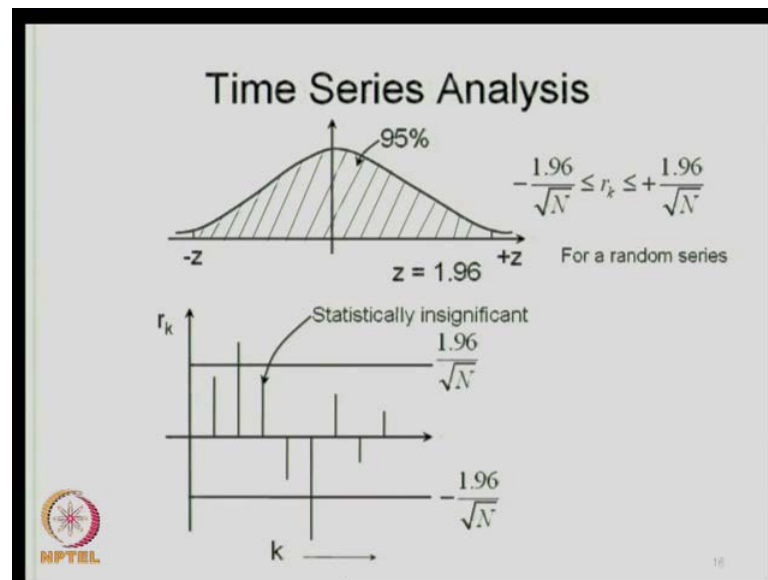


15

So, like this you get spikes and this is the correlogram. So, the auto correlation function or the correlogram may indicate some such fluctuations like this. So, for processes it may simply be decaying like this with time, it may be decaying if it is a purely random series or purely a stochastic series, then theoretically  $\rho_k$  must be equal to 0, because  $X_t$  and  $X_{t+k}$  are not related at all and we have seen that for a purely stochastic process your covariance auto covariance must be 0 and therefore, if you have a purely stochastic process then  $\rho_k$  must be equal to 0 for all  $k$ .

But  $r_k$  we have estimated from a sample  $r_k$  may not be 0. Although we may determine it from a purely stochastic series the estimated values of  $r_k$  may not be 0. The  $r_k$  for a random series follows a normal distribution, it can be shown that it follows a normal distribution with a 0 mean and a standard deviation of  $1/\sqrt{N}$ , where  $n$  is the number of observation we use this fact to determine whether the sample comes from a purely random series or in fact, we use this fact to examine whether a given value of  $r_k$  which is a sample estimate of the auto correlation at lag  $k$  whether it is a significant auto correlation or not.

(Refer Slide Time: 33:12)



So, the  $r_k$  for a random series follows a normal distribution with 0 mean and you standard deviation of  $1/\sqrt{N}$ . So, about 95 percent of the values lie between  $1.96/\sqrt{N}$  and  $-1.96/\sqrt{N}$ , and  $+1.96/\sqrt{N}$ . So, if you have a  $r_k$  value higher than this above these limits then that  $r_k$  value can be considered to be statistically significant at 95 percent level. So, once you compute your  $r_k$  values for the sequence that is you have the observed values of the random variable, let say stream flow at a particular location over last about 40 years or some such things. So, 40 years of stream flow observed monthly which means you will have 480 values 12 into 40, 80 values from the 480 values. You may compute the auto correlations at several lags may be let say up to about 50 lags above. So, up to 50 lags you have computed the  $r_k$  values.

Then you plot the  $r_k$  values which gives  $r_k$  versus  $k$  which gives the correlogram and then you draw the lines corresponding to significant levels of the  $r_k$ . Which is the this is the 95 percent significant  $r_k$  level minus  $1.96/\sqrt{N}$  plus  $1.96/\sqrt{N}$  where  $N$  is the number of observation in the case that I have just mention you had 480 values. So,  $n$  becomes 480. Any auto correlation which is beyond the significant bands that we have just drawn that becomes a significant auto correlation value all other values which lie within this band are insignificant at level of 95 percent. So, let us take an example now. So, you will have a series of observed values. So, this is the  $X_t$  which is observed let us take ten values and compute the auto correlation for  $k$  is equal to 1.




(Refer Slide Time: 35:28)

### Example-1

Obtain Auto correlation for  $k=1$

S.No.	$X_t$	$(x_t - \bar{x})$	$X_{t+1}$	$(x_{t+1} - \bar{x})$	$\frac{(x_t - \bar{x}) \cdot (x_{t+1} - \bar{x})}{(x_t - \bar{x})}$
1	97	-10.50	110	2.5	-26.25
2	110	2.50	121	13.5	33.75
3	121	13.50	117	9.5	128.25
4	117	9.50	79	-28.5	-270.75
5	79	-28.50	140	32.5	-926.25
6	140	32.50	75	-32.5	-1056.25
7	75	-32.50	127	19.5	-633.75
8	127	19.50	90	-17.5	-341.25
9	90	-17.50	119	11.5	-200.75
10	119	11.50			
$\Sigma$	1075				



The sample estimate for  $r_k$  the sample estimate for  $\gamma_k$ , which is  $r_k$  is given simply by  $C_k$  which is estimated like this divided by  $c_{naught}$ , which is the sample estimate for variance there are various other expressions for sample estimate of  $\rho_k$  that is there are various other expressions available for  $r_k$ , but we will use this particular expression simply  $C_k$  by  $C_{naught}$ . So, we compute  $\bar{X}$  first from summing off this. So,  $\bar{X}$  is known and then  $X_t - \bar{X}$ , because we are talking about  $k$  is equal to 1, which means you are getting the sample estimate for row 1. So, we lag the series by 1 one time unit. So, we write corresponding to  $X_t$  you write  $X_{t+1}$ . So,  $X_1$  associated with that  $X_{t+1}$  will be  $X_2$  associated with  $X_2$  we have  $X_3$  which is 121,  $X_4$  corresponds to  $X_3 + 1$ , so, 1 naught 1 1 7.

(Refer Slide Time: 37:41)


### Example-1 (contd.)

mean  $\bar{x} = 1075/10$   
 $= 107.5$

Variance,  $c_0 = \frac{\sum_{t=1}^n (x_t - \bar{x})^2}{n-1} = \frac{4132.5}{10-1} = 459.2$

$c_1 = \frac{\sum_{t=1}^{n-1} (x_t - \bar{x})(x_{t+1} - \bar{x})}{n} = \frac{3293.75}{10} = 329.375$

$r_1 = \frac{c_1}{c_0} = \frac{329.375}{459.2} = 0.72$




Like this we lag the series that is we shift the series by one time unit, and then write the series  $X_{t+1}$ ; then we take  $X_{t+1} - \bar{x}$ ;  $\bar{x}$  is computed based on the values of  $X_t$  given here;  $X_{t+1} - \bar{x}$ ; then the multiplication of this  $X_t - \bar{x}$  and  $X_{t+1} - \bar{x}$  that is this summation. And from this summation, we get the variance first as 459.2 and similarly we get auto covariance at lag 1, which is  $C_k$  as 329.75 what do we here we summed up these values, and then use this sum to get the auto covariance at lag 1.

(Refer Slide Time: 38:07)

### Example-2

Obtain correlogram for 40 uniformly distributed random numbers

S.No.	Data	S.No.	Data	S.No.	Data	S.No.	Data
1	98	11	73	21	25	31	89
2	69	12	36	22	49	32	10
3	30	13	11	23	73	33	36
4	50	14	54	24	38	34	42
5	93	15	31	25	14	35	84
6	1	16	74	26	4	36	82
7	66	17	23	27	87	37	55
8	99	18	88	28	99	38	93
9	76	19	82	29	69	39	2
10	65	20	92	30	57	40	43



From the auto covariance, we get the auto correlation, which is  $C_1$  by  $C_0$  there is  $r_1$   $k$  will be  $C_k$  by  $C_0$  and  $C_0$  is the variance. So,  $r_1$  will be  $C_1$  by  $C_0$  that comes to be 0.72. I would like to indicate a small correction here; this value is negative and therefore, the  $C_1$  that you get here will be negative. So, this is negative and the correlation  $r_1$  will be negative. So,  $r_1$  is minus 0.72, So, this is a negative correlation. Let say now if you have if you are looking  $k$  is equal to 1, but  $k$  is equal to 2 what you would have done for this example here we are putting  $X_{t+1}$  for  $k$  is equal to 1. So, if you are interested in  $k$  is equal to 2 you will put  $X_{t+2}$  which is corresponding to  $t$  is equal to 1 you will put the value of 121 here and then 117 here 79 here etcetera.

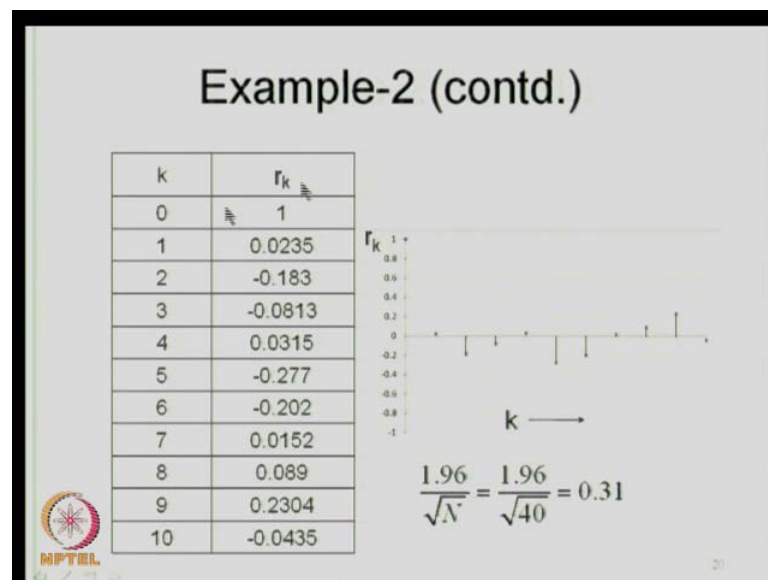
So, essentially you will lag the time series by 2 units instead of 1 unit as you did here for  $k$  is equal to 1 and  $k$  is equal to 3. Similarly you will have lagged it by 3 units; so, how many values you would have to compute this in the case of  $k$  is equal to 1 for this particular case you had 9 values for  $k$  is equal to 2 you had 8 values for  $k$  is equal to 3 you will have 7 values and. So, on. Therefore, as your lag increases you are estimating based on smaller and smaller number of values your values will be your estimate will be based on smaller and smaller number of values, because as you can see here the summation is from  $k$  is  $t$  is equal to 1 2 and  $N - k$  for  $k$  is equal to 1 you will have  $N - 1$  values  $k$  is equal to 2 you will have  $N - 2$  values and. So, on.

And therefore, we generally estimate the  $\rho_k$  or the auto correlations up to about 0.25  $N$ , where  $N$  is the number of observation; that means, up to about 25 percent of the number of observations compute  $\rho_k$ . Now we know how to compute for various values of  $k$  in in this particular case, we computed for  $k$  is equal to 1. So, we also know how to do it for  $k$  is equal to 2  $k$  is equal to 3 etcetera is simply lag the time series by those many times units, and then use the same procedure and compute associated with various values of  $k$ . We said that for a purely random series the  $\rho_k$  should all be insignificant for  $k$  not equal to 0 of course,  $\rho_0$  will be always equal to 1 irrespective of which series you are talking about because you are talking about the auto correlation of a variable with itself and that has to be equal to 1, but for other values of  $k$  that is  $k$  other than 0  $k$  not equal to

0 you rho k should all be insignificant statistically insignificant if it is a purely random series.

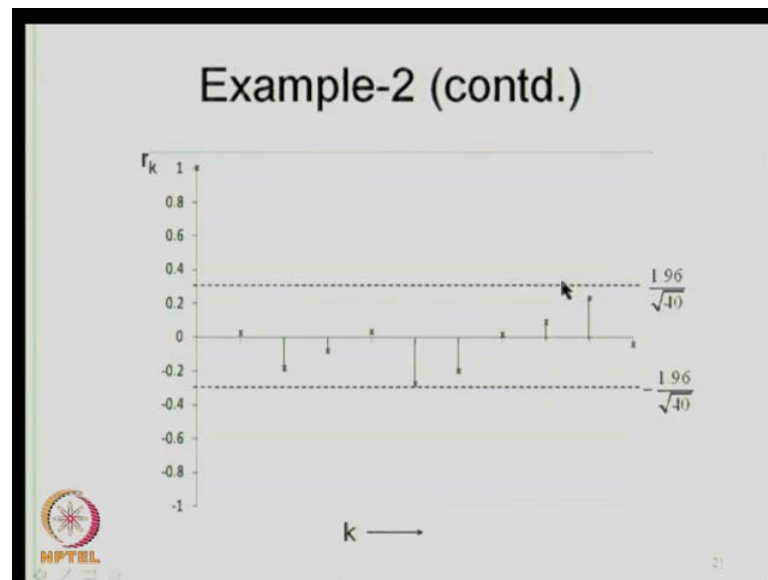
Let examine this now let us generate 40 uniformly distributed random numbers that is 2 digit random numbers, and see what happens to the auto correlations. We will compute auto correlations up to let say lag 10 or something. So, this is the data, these are the number serial number, and this is the data on uniformly distributed random numbers 98, 69 etcetera. So, then 11 ,12, 13 , 20. So, this is the data; so, you have 40 values of uniformly distributed random numbers between 0 and 100, 0 and 99 in this particular case.

(Refer Slide Time: 42:52)



We will compute the auto correlations up to lag 10 . When we do that you get values like this k is equal to 0 as 1 k is equal to 1 is 0. 0235 k is equal to 2 is minus 0.183 etcetera, and when you plot this the plot will be something like this it appears to be this; now your n is 40 because you had 40 data n is 40. So, 1.96 root 40 is 0.31 .

(Refer Slide Time: 43:24)

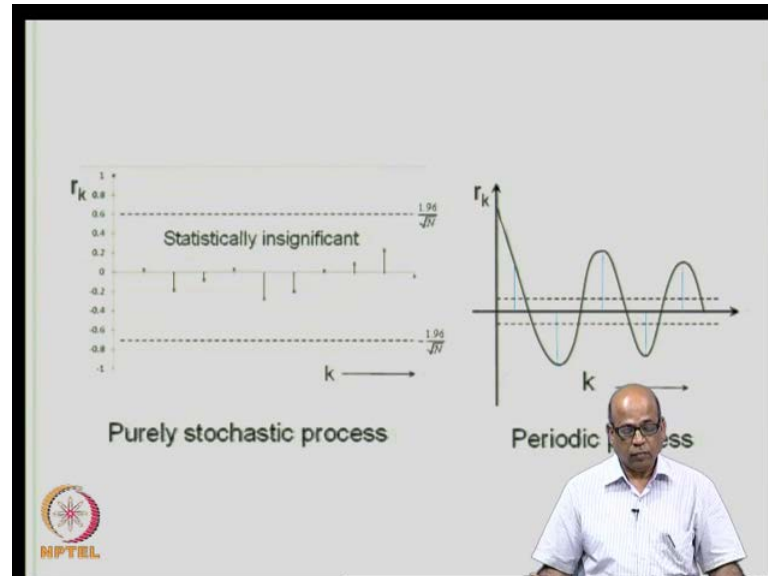


So, you draw a line minus 0.31 and plus 0.3 and see that ensure that all the values that you have got for  $r_k$  are. In fact, lying within this band minus 1.96 to plus 1.96. Now this happen in this particular case because they are all random numbers uniformly distributed random numbers and therefore, there is no dependence of one number on any number any other number in the series and therefore, your auto correlations at all lags that we consider are all zeros are all insignificant, which means the hypothesis is that the  $r_k$  that you thus obtain have a mean of 0. This 1.96 by root N normally you know for convenience you take it as 2 by root N. So, simply remember 2 by root N. So, whenever you compute the  $r_k$ , you compare those  $r_k$  with 2 by root N and if it is more than plus or minus two by root N on either side, then it can be considered significant.

Now, the significance of  $r_k$  is important, because we are looking at the dependence whether a particular value at time  $t$  depends on another time the value at another time  $t$  minus  $k$  that is given by  $r_k$  or  $r_k$  in this case in the case of samples we denoted by  $r_k$  and given a value of  $r_k$ , we should be able to say whether it is significant or not, often we make a mistake that let say we compute a correlation of 0.7 and say that 0.7 is a very high value and therefore, there is a significant dependence. Remember that the significance of the auto correlation or any correlation depends on the sample size. So, if your sample size is high let say instead of 40 we had 400 value, then what would have happened this value would have been much smaller. So, as sample size becomes larger and larger the significance level of the correlation becomes smaller and smaller for

example, we may have a correlation as small as 0.2 being significant if the sample size is large and a correlation as high as 0.6 may be insignificant if the sample size is very small.

(Refer Slide Time: 46:12)



So, remember always that the significance of the correlation depends on the sample size as we just saw if we have a purely stochastic process then all the correlations apart from at like 0.5 equal to 0 all the correlations must be statistically insignificant. So, if you have a sequence of observed flows and you compute the  $\rho_k$  or you estimate the  $r_k$  and then observed that all the  $r_k$  are statistically lying within this band then they are all statistically insignificant and therefore, it indicates a purely stochastic process. Now, this has an important connotation when we start looking at data generation and. So, on. If you have a periodic process let say you are looking at monthly stream flow in a monsoon climate like ours and you have 480 values as i just mention 40 years of monthly stream flow data. So, 480 values you compute the lags you compute the  $r_k$  and then plot the correlogram then you may get the  $\rho_k$  something similar to this. So, it is periodic in nature.

So, the periodicity of a process comes up when we plot the correlogram and this gives an have a data on monthly time series and compute  $r_k$  and then plot the correlogram if the correlogram indicates that there is a periodicity if the correlogram is periodic in nature it indicates that the time series has a periodicity in her indinate.

(Refer Slide Time: 46:18)



So, what do we do with all these information now we have plotted the time series let say we have observations for last. So, many years on observed stream flows at a particular location or observed rainfall at a particular station. We have collected all these values and plotted the time series then we also know how to compute the correlation auto correlations and auto covariance we know how to formulate the auto covariance matrix and. So, on. The whole purpose of the time series analysis is to extract information about the observations that we have made and then use this extracted information for decision making purposes. Let say that we want to build a reservoir build a at a particular location and you have made the observations on the last 50 years of stream flows you have the information on the last 50 years available with you we must extract the information contained in the observation and use that information contained to make decisions about the future.

Now, these decisions may be long term decisions for example, any hydrologic design that want to make on a physical structure let say you want to make a bridge pier or the height of the bridge etcetera, you want to decide based on the flood levels. Now, this is actually a long term decision that you are making based on the historical data that is available; that means, you are making a decision which has implications for a long period of time in future . Similarly you want to design a damp the height of the damp or the capacity of the damp etcetera, you your basing these decisions on the historical available data and this decision has a long term implication in terms of its serviceability

let say it has 100 years of serviceability. So, we are making the decision based on the last. So, many years of data let say 30 years of data or 50 years of data and then the decision that you are making has a implication over a long period in future.

That is we are talking about long term decisions it may be either data generation data extension etcetera. So, you use these techniques to make long term decisions on the other hand you may have interested not so much in the long term, but in the immediate short term lets say you are standing at June of 2010, and then you would like to examine what is likely to happen to the flow in July of 2010 that is a immediate next month.

So, you would be interested in forecasting what is likely to happen in the near future or in the immediate next time period of next 2 time period, next 3 time period, next one season and so on; so you will be interested essentially in the short term implications. So, these techniques that we are now introducing we will have applications on both long term as well as short term, when we are talking about long term. We call it as, either data generation and data extensions or when we are talking about short term, we will be interested in data forecasting.

Now, the data forecasting applications are many in hydrology we may be interested in flood forecasting. There is a rainfall that is occurring in the catchment and you would like to forecast the flood at a particular location, or a flood peak is known to have started along the river and then you would like to focus the flood moment across time in the same river. So, that is called as forecasting problem; whereas, extension and data generation problems are essentially planning problems, where you look at long term implications of the analysis of the observation that we have said we have with us to make the decisions.

So, in both these whether it is a long term decision making or for short term applications to forecasting. We use the historical data and the premise there is that the history provides a valuable clue to the future, because we do not have any other means of making judgments about how the future is likely to be, we use the historical data historical time series and then make judgments about how the future is likely to be that is a whole concept of time series analysis, and specially the stationary time series analysis that we have been dealing with.



In the stationary time series analysis what are we saying that the properties remain the same across time. So,  $X_t$  and  $X_{t+1}$  has a same probabilistic properties, and we assess these probabilistic properties based on the historical data that we have, and then use the same probabilistic properties for the future and make our judgments and make our decisions.

So, we will summarize today class today lecture essentially what we did today was to introduce the concepts of time series, we went through the definition of time series. Again and saw what we mean by a realization and an ensemble we saw what we mean by time properties time average properties and ensemble average properties, where we are talking about the properties across realizations at a given time. We also briefly define what we mean by an ergodic process, if your time average properties across time for realization are the same as ensemble average properties, then the process is called as an ergodic process.

Then we introduce auto covariance the definition of auto covariance, and then auto correlation and the sample estimates how we estimate the auto covariance and auto correlation from the sample estimates from the samples. The auto correlation at lag  $k$  denoted by  $\rho_k$  it is estimated by  $r_k$ , which is a sample estimate the plot between  $\rho_k$  and  $k$  or when you are dealing with sample  $r_k$ , and  $k$  is called as the correlogram are the auto correlation function.

Now, the correlogram gives an important information gives an important information about the time series for example, if it is a purely stochastic time series then all your  $\rho_k$  for  $k \neq 0$  will all be insignificant, what do we mean by insignificance that they lie in the band  $2/\sqrt{N}$  approximately or  $1.96/\sqrt{N}$  precisely at 95 percent significance level. So, if all your  $r_k$  lie within that band then they are all statistically insignificant. If you have sample value if you have observed value and then compute the  $\rho_k$  and they **they** are all statistically insignificant it indicates that the sample is drawn from a purely random process which we will use this fact we will use in data generation.

Then we also saw if we have a periodic process then the  $\rho_k$  or the correlogram will be oscillating periodically. So, if you have a sample and then you compute the  $\rho_k$  plot the correlogram, and if you see the correlogram is oscillating periodically, then it gives you an indication that your sequence is periodic, and we use this fact in our further modeling.

So, we will continue this discussion in the next lecture, where we will be dealing with data generation as well as data forecasting technique; thank you for your attention.