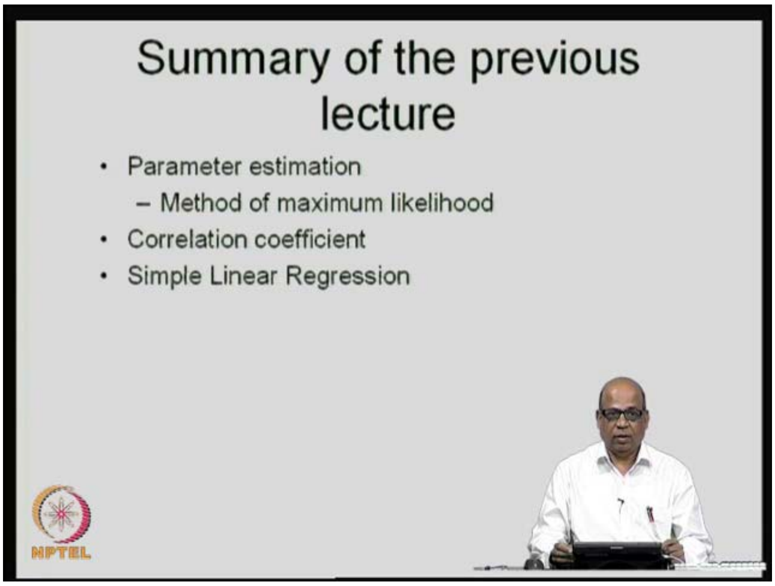
Stochastic Hydrology

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

Lecture No. # 09

Data Generation

(Refer Slide Time: 00:31)



Summary of the previous lecture

- Parameter estimation
  - Method of maximum likelihood
- Correlation coefficient
- Simple Linear Regression

Good morning and welcome to this lecture number nine of the course stochastic hydrology. If you recall in the last lecture, we started with parameter estimation especially we covered the method of maximum likelihood. The method of moments was covered in the earlier lecture. In the method of maximum likelihood, if you recall we construct a likelihood function given a sample x1, x2, etcetera, x n. And then we look for those sets that set of parameters theta1, theta 2, etcetera, theta m which will maximize the likelihood function. And in many situations it will be advantageous to take the log of the likelihood function, and maximize the log of likelihood.

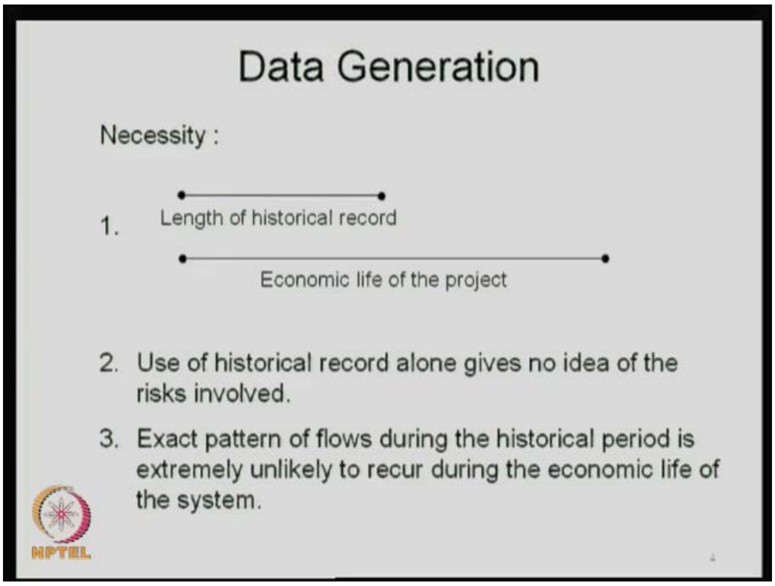
Then we went on to introduce the correlation coefficient. We call that coefficient correlation is it provides a major of linear dependence between the variables for which it is calculated. For example, we may calculate the correlation coefficient between rainfall and runoff. So, it provides the degree of linear dependence between rainfall and runoff in that particular case. Then we use this concept of correlation coefficient, and develop a simple linear regression relationship between the variable y which is the dependent variable, and the variable x which is a independent variable.

In hydrology many situations arise where we will be looking for relationships between let us say rainfall in the catchment area and runoff at the outlet of a catchment essentially what we did there is that in developing the linear or regression relationship, we minimize the error between the observed values and the predicted values actually we minimizes sum of squared errors. So, we are fitting a relationship y is equal to a x plus b for example, and x are all the observed values and y is the predicted value using this relationship. So, you will estimate the parameters a and b such that the squared error the sum of the squared errors is minimized.

Today class we will introduce the important topic of data generation, essentially you know in most hydrologic decision making situations. We will have observed data for certain time period, let say you have observed data for the last 30 years on the stream flow using this observed data you need to make decisions for the future many times a situation will arise where the observed data itself is of short length and you would like to extend the data. Also even if the data is of adequate length, because you are likely to make the decisions for the future you would like to examine how this particular sequence of data that you have is likely to behave in future.

So, the sequence itself will not repeat exactly and therefore, you would like to look at several sequences which are likely to occur in future and will base your decisions upon these several sequences rather than basing on a single sequence. Then there may be also situations where you would like to feel the data let us say you have a continuous record of about 30, 40 years, but in between the there is certain length for which data is not available then you would like to fill this particular data. So, all these practical situations we address with data generation techniques and also data forecasting which we will introduce subsequently perhaps in today lecture or in the subsequent lectures. So, we will see the motivation for the data generation techniques why it is necessary.

(Refer Slide Time: 04:45)



Data Generation

Necessity :

1. Length of historical record

   Economic life of the project

2. Use of historical record alone gives no idea of the risks involved.
3. Exact pattern of flows during the historical period is extremely unlikely to recur during the economic life of the system.

Like said you have, let us say a length of a historical record let us say this is about 30 years and you would like to use this 30 years of data at a particular location for let us say you want to develop a reservoir projector there. That means you want to build a damp and then construct a reservoir there the economic life of the project itself may be of the order of 100 years or 150 years. So, you are going to make a decision for the next about 100 years based on the last 30 years of data. So, you need to generate data for this economic life of the project. So, that you will study the implication of this particular project, let us say you want to look at the size of the reservoir that is possible or you want to look at for a given particular size of the reservoir how reliably you can meet the water supplier for a given set of demands and so on. So, they are all dependent on the flow that you have flow record that you have and therefore, using this flow record you would like to extend the data of the flow record itself for the economic life of the project which typically may be of the order of 100 years or so on.

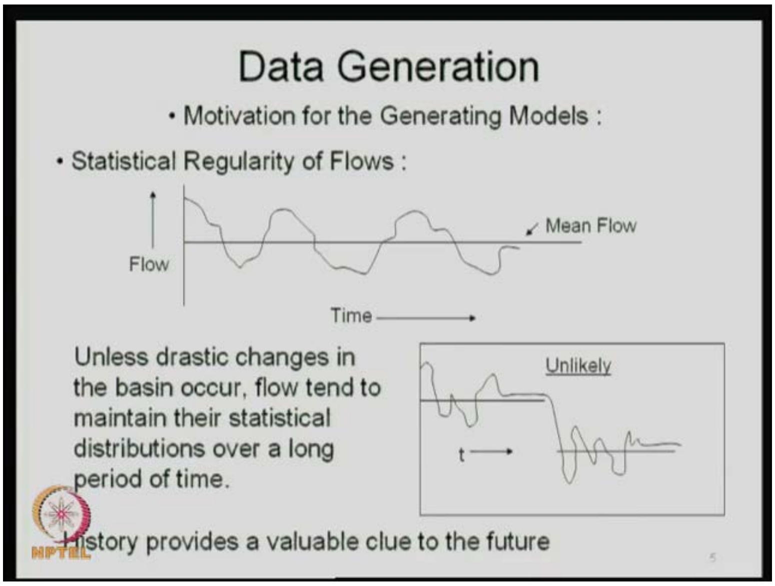
Another reason why we do data generation is that if you use only this historical record let us say you have 30 years of data and you make all the decisions based on just on this one sequence of 30 years of data. It will not give you an idea of the actual risks that are involved especially when you are looking at the risks due to flooding or the risks due to droughts hydrologic droughts and risk due to inability of the system to generate the necessary power and so on.

The third motivation for data generation is that the exact pattern of the steam flows or the exact pattern of the record that you have is not likely to repeat in future typically we talk about the stream flows when you are doing the data generation. So, if you have a 30 years of record this exact pattern is not likely to repeat; however, the information contain that we have in this record that information contain, if we use in some sense to generate several sequences of flows in the future. Then that flow that number of sequences will be useful for making better decisions and making better assessments of risks and so on. So, these are some of the motivations for which we develop the data generation techniques.

The first level of data generations that we will introduce in this course now is, if you have let say the probability distribution of the particular variable, let say you are talking about stream flows and from the last 30 years of observed data you have an estimate of the parameters. Let say mean standard deviation etcetera,   that we have learnt earlier in this course and you also suspect that this particular record perhaps follows a particular given distribution let say normal distribution or exponential distribution and so on. So, if you are given us specific distribution with all its parameters estimated from the sample values the question now is how do we generate sequence of values or belonging to that particular distribution.

So, the problem is that you have the sample values observed values of data and you have fit a particular distribution to the observe values of data. When I say you fit a distribution we mean you calculate all the parameters of that distribution and then now we want to generate several such values which will all follow the same distribution. So, the idea here is that if the stream flow at a particular location, let us say follows normal distribution with mean mu and standard deviation sigma. Can we generate several such values all belonging to the probability distribution all belonging to the normal probability distribution with mean mu and standard deviation sigma that is the question. So, the main principle behind the data generation that as we use in hydrology is that there is a statistical regularity of the hydrology processes unless there are major changes that occur.

(Refer Slide Time: 09:52)



Data Generation

- Motivation for the Generating Models :
- Statistical Regularity of Flows :

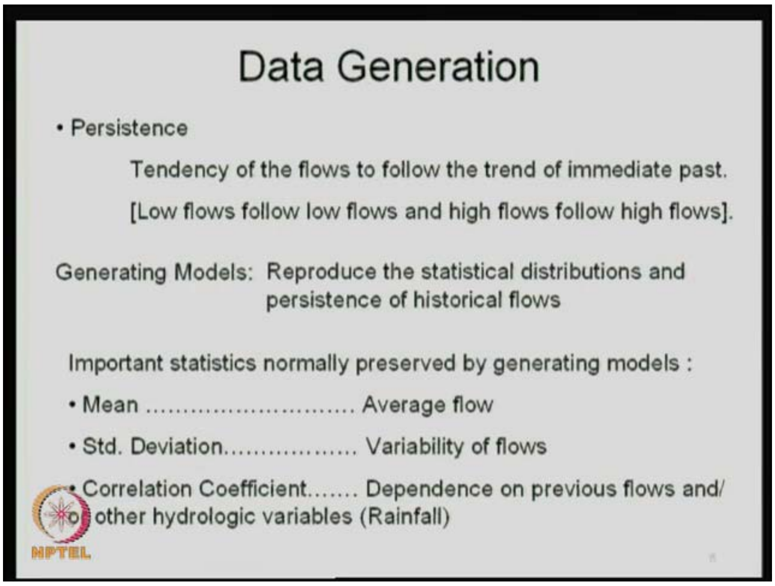
Unless drastic changes in the basin occur, flow tend to maintain their statistical distributions over a long period of time.

History provides a valuable clue to the future

For example, you look at the flows in a particular catchment. That is a statistical regularities of flows, if you look at last about 50 to 100 years you may see certain statistical regularity of flows unless drastic changes occur in the catchment or the basin this regularity is unlikely to be perturb. For example you look at this figure the process of operating at a certain level and unless there are certain physical changes occur this shift as you have seen here is very unlikely. Whereas this kind of regularity is more likely in nature now what are these drastic changes that we are talking about for example, in a particular catchment you do a large scale deforestation or there may be a large scale fire that takes place or as it happened recently a tsunami occurs and then a large scale disturbance occurs in a very short time in the catchment and therefore, all the hydrologic processes get perturb or disturb in a very short time.

Unless such things happens this kind of situations where the process of operating at a particular level and suddenly it shift to another particular level this occurrence is very unlikely. So, the principle that we use in data generation is that history provides a valuable clue to the future whatever has happened during the historical time period, we used that information to make an assessment of how this process is likely to behave in the future and we use that principle to generate the data of the future. So, essentially we capture the essential information content in terms of the probability distribution in terms of the statistical parameters and in terms of the general stochastic behavior of this particular process. As we have seen over the history use that information to generate values for the future.

(Refer Slide Time: 12:08)



Data Generation

- Persistence

  Tendency of the flows to follow the trend of immediate past.

  [Low flows follow low flows and high flows follow high flows].

Generating Models: Reproduce the statistical distributions and persistence of historical flows

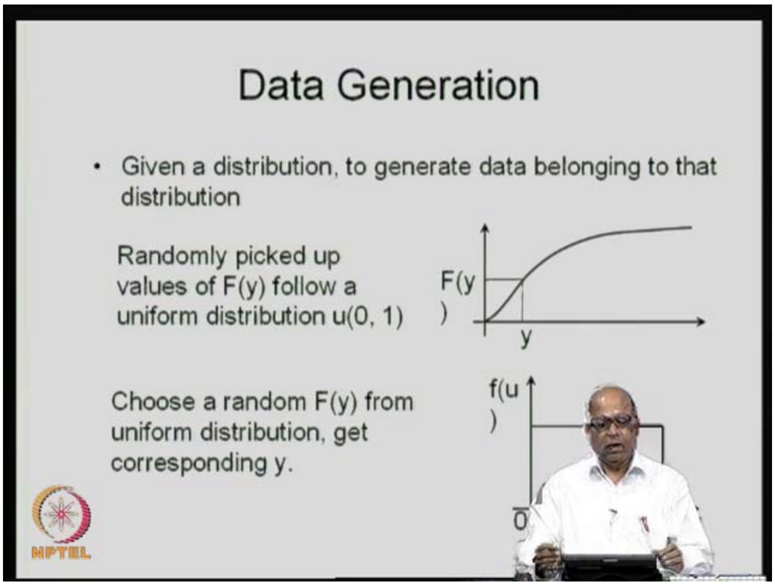
Important statistics normally preserved by generating models :

- Mean ……………………… Average flow
- Std. Deviation……………… Variability of flows
- Correlation Coefficient……. Dependence on previous flows and/ or other hydrologic variables (Rainfall)

So, this is the general principle, let us see how we do this. We also use the fact, in fact that there is a persistence in nature, especially as far as hydrologic processes are concerned. What do we mean by persistence that may be there is a tendency of flows to follow the trend of immediate past that is low flows follows low flows and high flow follows high flow. So, the generating models reproduce a statistical distribution and persistence of historical flows when we are talking about generating models for the flows, so, in doing. So, as just mentioned we have historical data for we would have calculated the parameters. Such as means standard deviations correlations coefficients and several others moments and statistical parameters. So, we developed models for data generations which when used will reproduce a data which will have the same mean as the historical mean or more or less the same mean.

It will have nearly the same standard deviation as a historical standard deviation it will have the same correlation coefficient as a historical correlation coefficient and correlation coefficient we may talk about correlation between runoff and some other variables like rainfall humidity soil moistures and so on or runoff at a particular time period with runoff at previous time period. So, we are capturing the dependence on previous flows or other hydrologic variables such as rainfall soil moistures and so on two correlation coefficient. So, when we develop the generating models we will develop such that certain parameters or certain moments of the historical data are preserved in the sequence that we generate.

(Refer Slide Time: 14:20)



Data Generation

- Given a distribution, to generate data belonging to that distribution

Randomly picked up values of F(y) follow a uniform distribution u(0, 1)

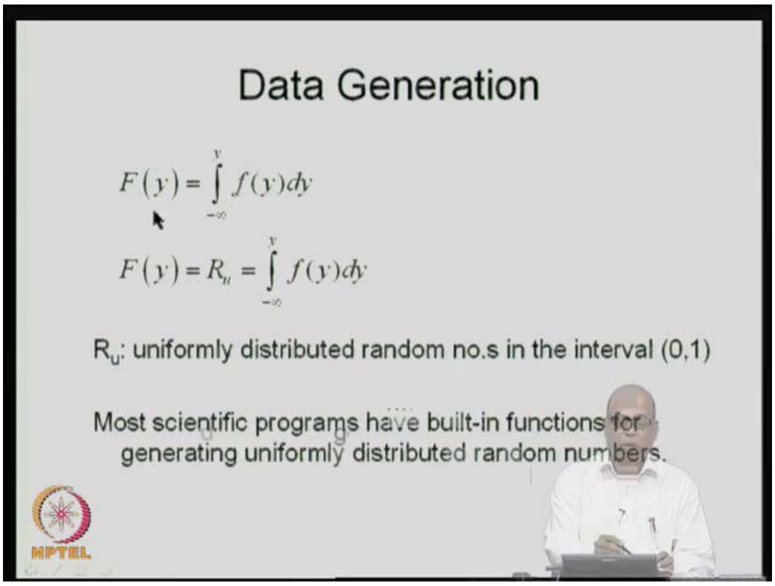
Choose a random F(y) from uniform distribution, get corresponding y.

So, how do we do this as mentioned first will look at the case where we know the probability distribution of the historical data and we want to generate several values all belonging to a given distribution. So, we are saying given a distribution we want to generate data belonging to that particular distribution this is the first simplest problem that we consider the motivation for this procedure is that if you have the c d f. Let say cdf is from here and any cdf you pick up randomly these values on the cdf as you know the cdf can vary between 0 and 1 the maximum value is 0 and 1.

So, if you pick up randomly the values on cdf they follow a uniform distribution in the interval 0 and 1. In fact, this becomes a very handy result which we use in data generation remember this is irrespective of the distribution. So, as long as you have a cdf the cdf values themselves follow a uniform distribution in the region 0 and 1. So, we use this result to generate values, let us say we pick up a particular f(u) the F of y which is the which is the cdf value at a particular y. And if you know the cdf relationship the analytical relationship for example in the exponential case you know F of y is equal to o1 e minus e to the power minus lambda y where lambda is the parameter.

So, if you know the analytical expression for F of Y you solve for y using the expression for F of Y and you put a random value for F of Y what is this random value? as just mentioned the F of Y randomly picked up. F of Y follows a uniform distribution between 0 and 1. So, you generate random numbers which belong to this uniform distribution between in 0 and 1 set that equal to F of Y and solve for Y knowing the analytical expression for 4 Y and that is the value of generated value of Y that is the generated value of Y. Next time you change the F of Y by changing the random number and generate another value of Y and so on and so forth. So, essentially what we are doing is F of Y is given here. So, you may have analytical expression for F of Y.

(Refer Slide Time: 17:02)



Data Generation

$$F(y) = \int_{-\infty}^{y} f(y)dy$$

$$F(y) = R_u = \int_{-\infty}^{y} f(y)dy$$

R_u: uniformly distributed random no.s in the interval (0,1)

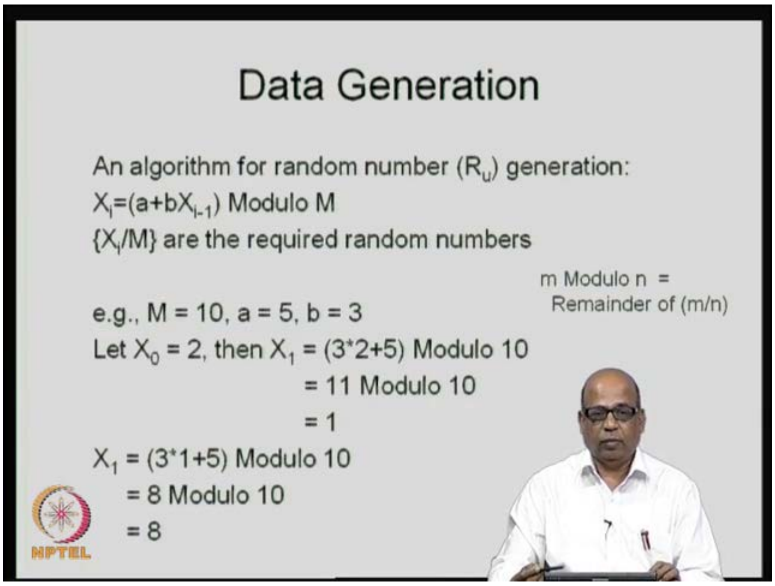
Most scientific programs have built-in functions for generating uniformly distributed random numbers.

You set F of Y is equal to R u which is a uniformly distributed random number in the interval 0 and 1. From this R u, let us say R u is 2.234 or something. So, you are saying F of Y is 0.234 or something. So, you are saying F of Y is equal to 0.234 and then there is an analytical expression for this. So, solve for Y and then from the analytical expression you solve for Y and that becomes a generated value of Y.

Next time between the random number you get another value of Y and so on now, how do we generate this uniformly distributed random number most of the scientific programs that you use on computer for example, spreadsheets and excel and so, on. They all have beating functions for generating uniformly distributed random numbers. That is this R u are obtained readily by uniformly distributed random numbers most scientific calculators also have a function of the random numbers for example, I will show you here I have a and it will have a random number generator. So, there is a random function here I press I get a number of 0.563 the next time I press I get a number of 0.966 next time I press I get a number of 0.229 and so on.

Now, these are sets of random numbers which follow a uniform distribution with interval 0 and 1 between interval 0 and 1 remember they are really random numbers. In fact, they are pseudo random numbers I will tell you what are the pseudo random numbers presently. If I put off the system if I put of the calculator and then again restart and again. I generate I will get a different number now .318 and next one is again 0.323 and. So, if let us say 10 of you this experiment all the 10 of you will get a different sequence of number and that is why. In fact, they are random numbers. So, you generate random numbers uniformly distributed random numbers by using a calculator like this or if you are using a computer most of the computer programs most of the standard scientific computer programs like, let us say you are using matlab or excel ms excel and so on these will have in built functions for providing you uniformly distributed random numbers R u. So, you use them and then solve for your Y and you will use those many random variables as you need as the number of values that you need to generate. Let us look at one simple example another small point on the random numbers here.

(Refer Slide Time: 19:58)



Data Generation

An algorithm for random number (R_u) generation:

$X_i=(a+bX_{i-1})$ Modulo M

$\{X_i/M\}$ are the required random numbers

m Modulo n = Remainder of (m/n)

e.g., M = 10, a = 5, b = 3

Let $X_0 = 2$, then $X_1 = (3*2+5)$ Modulo 10

= 11 Modulo 10

= 1

$X_1 = (3*1+5)$ Modulo 10

= 8 Modulo 10

= 8

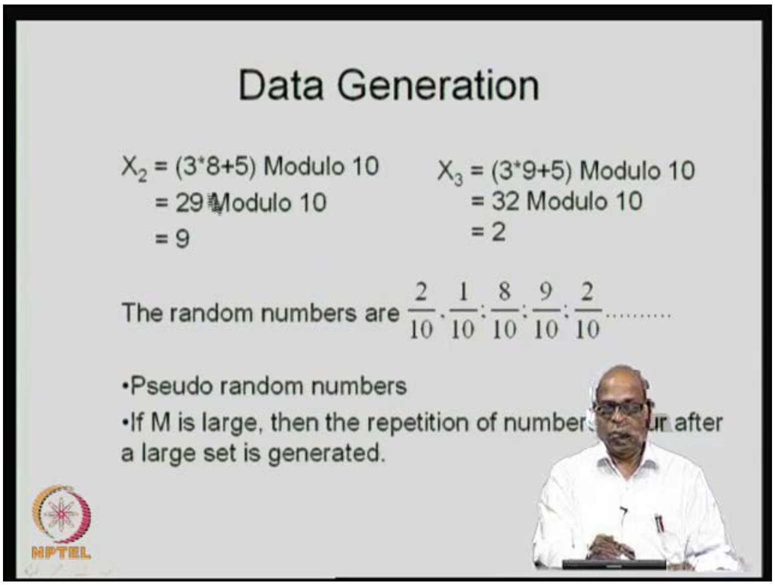
We call them as pseudo random numbers why do we call them as pseudo random numbers. Because there is an inbuilt algorithm that generates these random numbers for example, when I say that using a calculator and then generating a random number. Let say I get 0.884 and I generate next random number 0.344. So, there is a inbuilt mathematical algorithm that is generating these numbers and therefore, they are not really random, because for the numbers to be truly random the process that is generating has to be absolutely random for example, you look at a radioactive material and then you are measuring the number of particles that it is emanating per unit time, let say per 10 minutes or per unit time. The number of particles are emanating now this is a truly random process whereas, the numbers that we generate through any of the computer programs or any or in the calculator etcetera, are not truly random in that sense because we are using a particular algorithm.

And therefore, if you really generate really large number of values after sometimes the sequence is likely to repeat sequence will repeat and therefore, the set of random numbers that we are generating thus are called as a pseudo random numbers. So, we have mechanisms or we have or we have algorithms to generate the pseudo random numbers while it is not really necessary because whereas, I mentioned most calculators have the most of the computer programs have functions to generate pseudo random numbers. We must at least know how it is generated. So, let us see a very simple algorithm by which a by which a sequence of random numbers can be generated which follow a uniform distribution in the interval 0 and 1.

So, this is a simple algorithm for random number generator now, when I say random number here R u, R u indicates a random number which belongs to uniform distribution in the interval 0 and 1 that is what we denoted as R u. So, you look at the algorithm X i is equal to a plus b X i minus 1 modulo M now modulo m we define M modulo n if it is there we take the reminder of m divided by n for example, phi modulo 2. We divide phi by 2 integer division. So, the remainder is 1 because 2 into 2 is 4. So, remainder is 1. So, phi modulo t 2 o will be equal to1. So, we take the remainder of integer division m by n. So, X i is equal to a plus b X i minus 1 modulo m very simple algorithm the m is essentially a very large integer value a very large number and a and b are fixed. So, in this algorithm you fix a b and m. Start with a particular number given number X 0 generate X 1 use X 1 to generate X 2 and solve. So, like this you can generate. So, you get a sequence of X I X i by m are the required random numbers.

So, you generate X i and then X i by m the sequence that use the sequence of random numbers belonging to uniform distribution in the interval 0 and 1. We will take m is equal to 10 remember i told m has to be very large. Just for demonstration I am taking m is equal to 10 here. Typically m is of the order of 2 to the power 32 or the maximum integer value that a computer can hold long ago it used to be 2 to the power 32 , but now it can be much higher. So, m is equal to 10 a is equal to 5 b is equal to 3. So, these are the constants that you use in the algorithm. So, let us say we start with X naught is equal to 2 then X i is equal to a plus b X i modulo 10 that is 11 modulo10. So, you divide 11 by 10 the result is 1. Now you use this 1 to generate the next X i. So, this would be X 2 is equal to 8 and, similarly X 3 is equal to 9 and so on. So, like this you get values of you started with 2 then 1 then 8 then 9 2 etcetera.

(Refer Slide Time: 24:33)



Data Generation

X_2 = (3*8+5) Modulo 10
= 29 Modulo 10
= 9

X_3 = (3*9+5) Modulo 10
= 32 Modulo 10
= 2

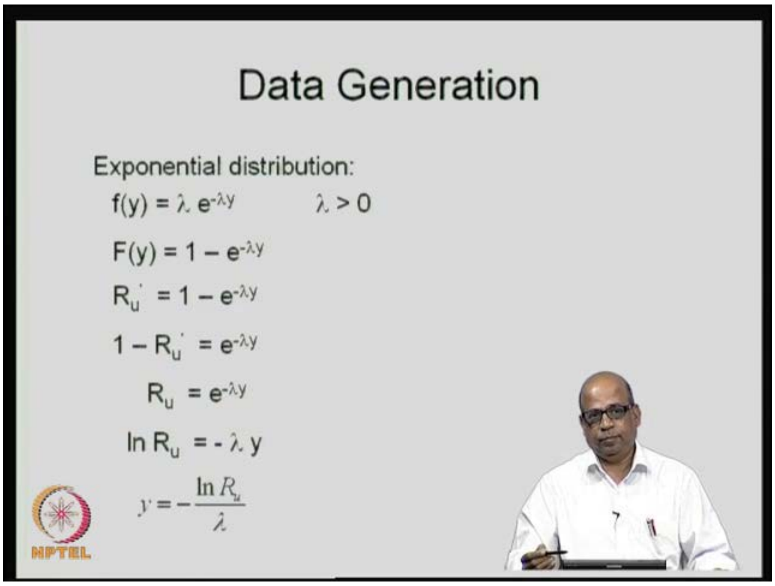
The random numbers are $\frac{2}{10}, \frac{1}{10}; \frac{8}{10}; \frac{9}{10}; \frac{2}{10}$..........

•Pseudo random numbers

•If M is large, then the repetition of numbers occur after a large set is generated.

So, the random numbers will be 2 by 10, 1 by 10, 8 by 10, 9 by 10, 2 by 10 and so on once this number repeats here this is 2 by 10. So, this repeats here because you are using the same constant values earlier as mentioned earlier your sequence repeats. So, this sequence repeats the length of the sequence before it actually repeats depends on essentially on this number m, but also to some extent on this parameters. So, the higher the larger the value of m the larger is the sequence that you will be able to generate. So, this is one of the ways by which you can generate random numbers uniformly distributed random numbers and they are remember they are really pseudo numbers. Now that we know how to generate random numbers there’s a uniformly distributed numbers.

(Refer Slide Time: 25:43)



Data Generation

Exponential distribution:

$f(y) = \lambda .\, e^{-\lambda y} \qquad \lambda > 0$

$F(y) = 1 - e^{-\lambda y}$

$R_u^{'} = 1 - e^{-\lambda y}$

$1 - R_u^{'} = e^{-\lambda y}$

$R_u = e^{-\lambda y}$

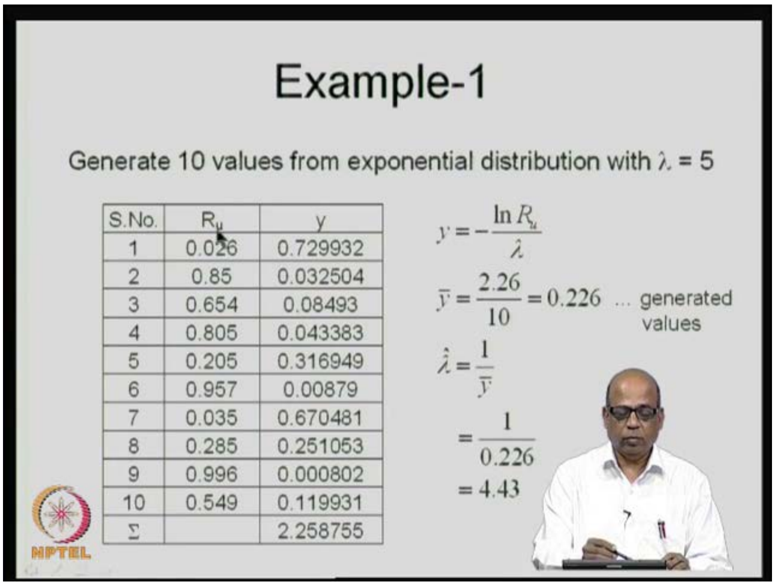
$\ln R_u = - \lambda .\, y$

$$y = -\frac{\ln R_u}{\lambda}$$

Let us see how we generate the data belonging to given a given distribution. So, F of Y will take lambda e to the power lambda Y this is exponential distribution lambda is greater than 0 and X is also y is also greater than 0 in this case. So, recall that the cdf is given by cdf is given by F of Y is equal to 1 minus e to the power of minus lambda Y. So, as I said equate these 2 you set a random number from uniform distribution and then put that random number is equal to F of Y. That means you are randomly picking up F of Y that is why the other. So, for R u dash is equal to 1 minus e to the power minus lambda Y and therefore, e to the power lambda Y is equal to 1 minus R u. Because r u dash is between 0 and 1 and it is random I can put this as R u another random number 1 minus R u dash itself is a random number. So, R u is equal to e to the power minus lambda Y.

Then we will take logarithms. So, log R u will become equal to minus Y therefore, Y is equal to minus l m R u by lambda. So, you generate random numbers R u use the calculator for example, every time you generate 1 R u you get that particular value of lambda for a given that particular value of Y generated value Y for a given distribution; that means, lambda is fixed there. So, once you fixed lambda generate your R u using a calculator and generate those many values of Y, let us look at a simple example. So, generate 10 values from exponential distribution with lambda is equal to 5.

(Refer Slide Time: 27:38)



Example-1

Generate 10 values from exponential distribution with $\lambda = 5$

S.No.	R_u	y
1	0.026	0.729932
2	0.85	0.032504
3	0.654	0.08493
4	0.805	0.043383
5	0.205	0.316949
6	0.957	0.00879
7	0.035	0.670481
8	0.285	0.251053
9	0.996	0.000802
10	0.549	0.119931
Σ		2.258755

$$y = -\frac{\ln R_u}{\lambda}$$

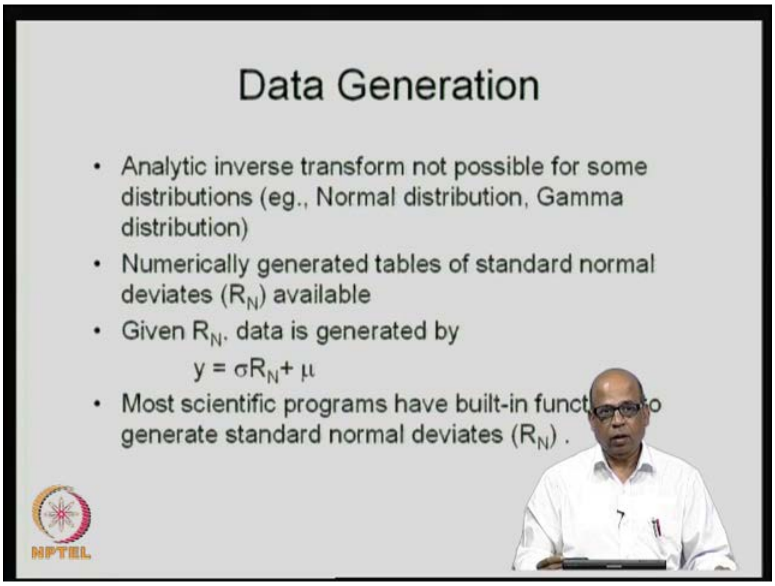
$$\bar{y} = \frac{2.26}{10} = 0.226 \text{ ... generated values}$$

$$\hat{\lambda} = \frac{1}{\bar{y}} = \frac{1}{0.226} = 4.43$$

First you generate 1 random number uniformly distributed random number, let us say again I will take calculator and then get this number and if that number is 0.026 lambda is given. So, you will get Y. So, you get a Y next time you generate another random number as I mentioned every time you take a new random number and then generate the associated value of Y. So, like this you generate 10 values of Y the data generation has to be essentially done for a large sequence you cannot just generate 5 values10 values etcetera, and then say that this follows the same distribution as the given distribution.

So, if you generate not ten values, but 1000 values 10,000 values 20,000 values etcetera, and then examine whether the sequence that you have. So, generated in fact, follows the exponential distribution with lambda with the associated lambda then it has some relevance. Let us see what happens, let us say we have generated 5 values here You have generated 10 values for these 10 values, if you get lambda it comes to about 4.43, it comes to 4.43 and we have started with lambda is equal to5. So, ideally I would like to have lambda is equal to 5 and I should be able to test that the sequence that you have generated in fact, follows a exponential distribution. So, if you generate more values, let us say from 10 you went up to 100 may be it improves further lambda improves further from 100 even to 1000 lambda improves further 1000 to 10,000 etcetera. So, ideally you generate a large sequence of values and then examine for the parameter that parameter value must be as close to the parameter value for which you are generating the random numbers random data as possible. And then you should also have ways of examining whether this follows exponential distribution, because of the basis by which you have constructed this data. In fact follow the exponential distribution.

(Refer Slide Time: 30:14)



Data Generation

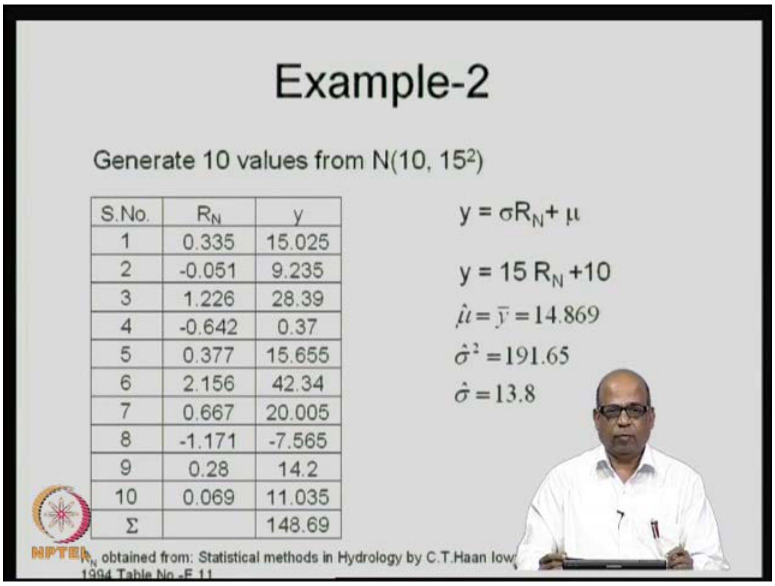
- Analytic inverse transform not possible for some distributions (eg., Normal distribution, Gamma distribution)
- Numerically generated tables of standard normal deviates (R_N) available
- Given R_N. data is generated by

  $y = \sigma R_N + \mu$
- Most scientific programs have built-in funct o generate standard normal deviates (R_N) .

Now, this was the example that I just discussed it was conveniently possible for us to solve for Y. What? did we do we solve for Y knowing the expression for cdf, F of Y. So, this particular form is a in such a convenient form that I could very easily readily solve for Y. That is why we are looking for the inverse solution for y given f of y, but there are many situations many distributions for which this is not possible for example, you take the normal distribution the cdf of normal distribution. If you look at it is not possible for you to solve analytically for y given f of y, now that is a l F of y which is a cdf of y. Similarly, gamma distribution gamma distribution it is not possible for you to solve for y given f of y in such situations we may use different techniques or different ways of handling this difficulty. For example in the case of normal distribution what we do is we use the standard normal distribution Y is equal to, let us say you transform that is Z equal to Y minus Y bar by sigma that is how you transform and then solve for Y. So, for various values of Z you should be able to get Y. How do we generate this various values of Y Z that is you want those random numbers which follows a standard normal curve with 0 mean and unit variance now these are tabulated and available they are called as standard normal deviates.

So, we denote them as R n. So, R u we indicated as uniformly distributed random numbers in the interval 0 and 1 and R n. R standard normal deviates which follows a normal distribution with 0 mean and unit variance. Now, these are available these stables are available and also many of the programs scientific programs have functions that you can call them in your program to generate directly R n. Once you have this R n values there is a random numbers you can generate the data Y is equal to sigma R n plus mu. How do we get this is y minus mu by sigma which is the standard normal deviate z. So, you given R n you should able to get y for a given normal distribution with parameters mu and sigma.

(Refer Slide Time: 33:02)



Example-2

Generate 10 values from $N(10, 15^2)$

S.No.	R_N	y
1	0.335	15.025
2	-0.051	9.235
3	1.226	28.39
4	-0.642	0.37
5	0.377	15.655
6	2.156	42.34
7	0.667	20.005
8	-1.171	-7.565
9	0.28	14.2
10	0.069	11.035
Σ		148.69

$y = \sigma R_N + \mu$

$y = 15\, R_N + 10$

$\hat{\mu} = \bar{y} = 14.869$

$\hat{\sigma}^2 = 191.65$

$\hat{\sigma} = 13.8$

R_N obtained from: Statistical methods in Hydrology by C.T.Haan low

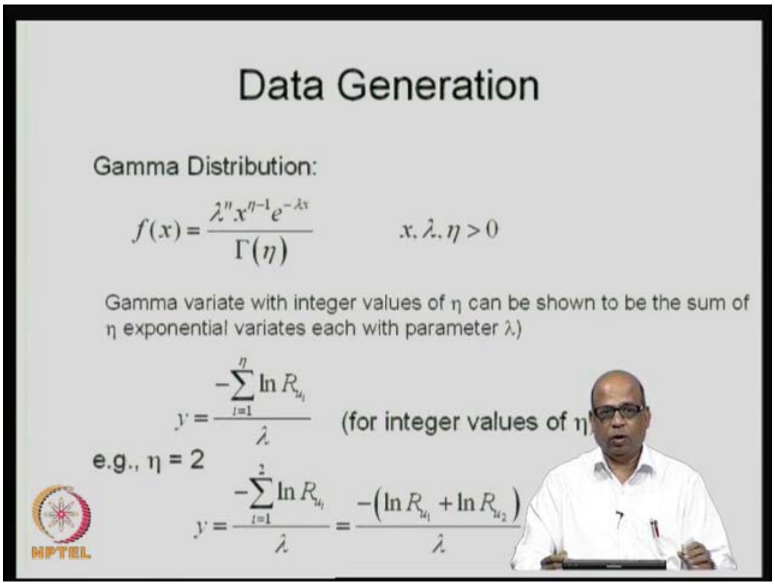
1994 Table No.-E 11

So, let us have a look at one simple example here. So, we want to generate 10 values from a normal distribution, which has mean as 10 and variance as 15 square. So, we write y is equal to sigma R n plus mu. So, mu is given and sigma is given. So, we generate R n with from standard tables or something. So, here we use the R n values these are taken from a standard text books C t Haan which I have given reference in the first lecture. So, e take this random numbers from a table which is provided in this particular text book. Every time you change the R N and get one particular value of y.

So, y is equal to sigma R n plus mu next time you change the R N and get one particular value of y. So, y is equal to sigma R N plus mu you get this y.

Next time you change the R n you get another value of y change the R n you get another value of y and so on like this you get. So, again check whether what is the mu that you are getting what is the average value? that you are getting out of the generated numbers. So, this comes out to be 14.86 and the standard deviation comes out t be 191.65 as against13.8 standard deviation as against the 15 that you started with. And this is quite this is 14.86 as against 10 again, if you generate not 10 values, but 1000 values 10000 values 20 values etcetera, using this procedure you will converge to a normal distribution that is the generated values will approximate a normal distribution with a mean of 10 and variance of 15 square. Then there are certain distributions for which even obtaining this tables is difficult analytical values of this tables are difficult, but they may be related to some other distributions for which data generation is possible using analytical procedure.

(Refer Slide Time: 35:09)



Data Generation

Gamma Distribution:

$$f(x) = \frac{\lambda^{\eta} x^{\eta-1} e^{-\lambda x}}{\Gamma(\eta)} \qquad x, \lambda, \eta > 0$$

Gamma variate with integer values of η can be shown to be the sum of η exponential variates each with parameter λ)

$$y = \frac{-\sum_{i=1}^{\eta} \ln R_{u_i}}{\lambda}$$ (for integer values of η

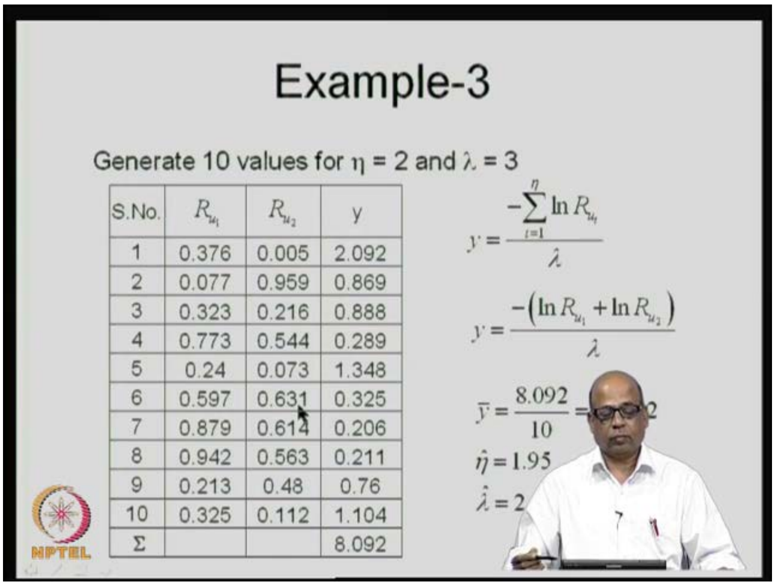
e.g., η = 2

$$y = \frac{-\sum_{i=1}^{2} \ln R_{u_i}}{\lambda} = \frac{-\left(\ln R_{u_1} + \ln R_{u_2}\right)}{\lambda}$$

For example you look at the gamma distribution in the gamma distribution for the specific case where eta is integer that is for integer values, it has been shown that gamma variate with integer values of eta is actually sum of eta exponential variates with parameter lambda. Gamma distribution has two parameters eta and lambda and the exponential distribution which if you recall is a special case of gamma distribution that has one parameter lambda. So, for integer values of eta the gamma variates can be equated to some of exponential variates for integer values. So, for example, y is equal to we can write y which is a gamma variate we want to generate a gamma variate is sum of minus log R u. I is equal to one to eta by lambda what is this log R u by eta log R u by eta just now we saw for exponential distribution.

We just saw that the expression for that is minus log r u by lambda. So, we use this exponential distribution sum it eta number of times if eta is equal to two for example, you generate two random numbers formal distributed numbers and then sum it the sum the exponential variates and to obtain the gamma variate that is the area here. So, for eta is equal to 2, which is an integer number you get y which is a gamma variate in this case and as minus log R u 1 plus log R u2. So, you need 2 random variable 2 random numbers two uniformly distributed random numbers to get 1 gamma variate one data belonging to the gamma distribution with eta is equal to b 2 .

(Refer Slide Time: 37:22)



Example-3

Generate 10 values for η = 2 and λ = 3

S.No.	R_{u_1}	R_{u_2}	y
1	0.376	0.005	2.092
2	0.077	0.959	0.869
3	0.323	0.216	0.888
4	0.773	0.544	0.289
5	0.24	0.073	1.348
6	0.597	0.631	0.325
7	0.879	0.614	0.206
8	0.942	0.563	0.211
9	0.213	0.48	0.76
10	0.325	0.112	1.104
Σ			8.092

$$y = \frac{-\sum_{i=1}^{\eta} \ln R_{u_i}}{\lambda}$$

$$y = \frac{-\left(\ln R_{u_1} + \ln R_{u_2}\right)}{\lambda}$$

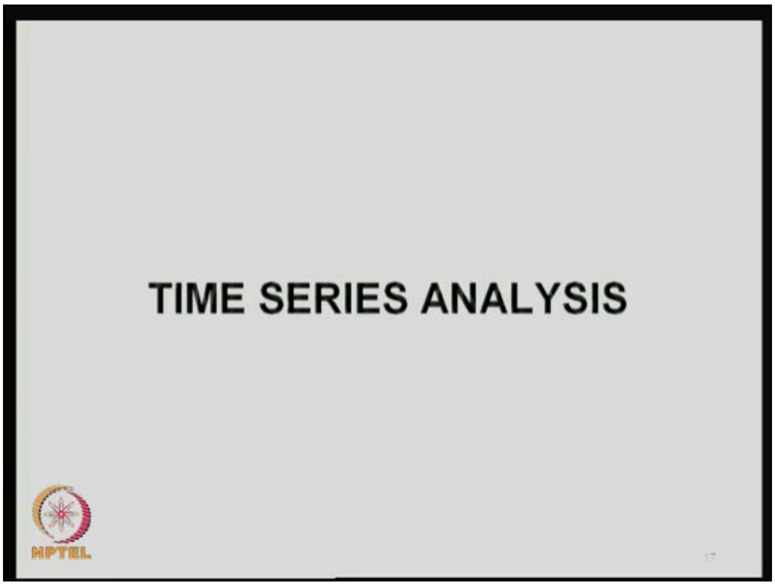
$\bar{y} = \frac{8.092}{10} =$ [illegible]2

$\hat{\eta} = 1.95$

$\hat{\lambda} = 2$

So, let us use that expression and then generate 10 values for that is 10 values belonging to gamma distribution with eta is equal to 2 and lambda is equal to 3 for specifying the distribution and we want to generate those many values. Because eta is equal 2 you need 2 random numbers to generate every number belonging to the gamma distribution. So, R u 1 and R u 2 you get two random numbers again using a calculator or something you get two random numbers to inform the distributed random numbers and you generate y. So, y is equal to minus log R u 1 plus log R u 2 divided by lambda is given therefore, you generate y like this you generate 10 values of y. You can generate again 1000, 10000, 20000 depending on the need you can generate as many values as you want and then you can finally, check for the parameters in this particular case for this 10 values you get a parameter eta cap estimated from the generated values comes to 1.95 and lambda cap comes to 2. 41.

(Refer Slide Time: 38:39)



TIME SERIES ANALYSIS

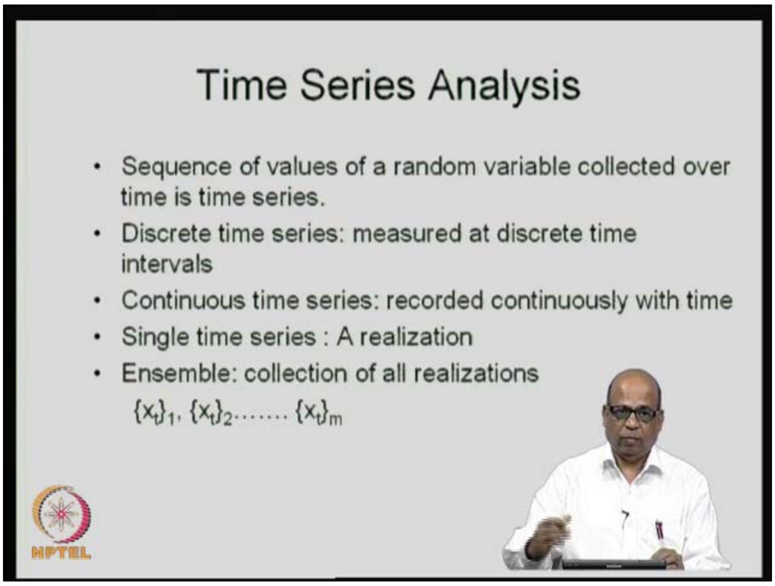
So, what we did in of data generation is that we specified the distribution When we specified the distribution we specified also parameters of the distribution, we say that normal distribution with a particular set of parameters mu and sigma square. And then generated values belonging to that particular distribution the data that is that belongs to that particular distribution remember here we did not build in a dependence of that data on a the previous value. Say for example, this particular data does not depend on the data that is generated previously, similarly this data does not depend on the data that is generated previously. So, we are saying that it is a purely random process which follows a particular distribution.

Let say gaussian distribution or gamma distribution etcetera, say every time period. Let us say you are talking about a sequence of flow here every time period you are picking up a randomly generated data belonging to that particular distribution. You are not worried about each dependence on the previous values that have been generated and that is why we did not talk about the correlation of one value on previous values. So, simply you are generating data based on a given distribution. So, all these are random numbers belonging to that particular distribution. Now we come to an important topic in stochastic hydrology and that is the time series analysis. Actually this is the way we cover in this particular course is just an introduction to time series analysis the time series. Analysis is a course by itself typically we offer a three graded course on time series analysis.

But as part of stochastic hydrology we must introduce the concept of time series analysis time series analysis plays a very important role in hydrologic data generation as well as hydrologic forecasting in many cases we will be interested in, let us say standing at the beginning of June month. We may be interested in forecasting the flows for the season June July, August September in monsoon season, if you are interested we will be keen to have a forecast for the seasonal flows for the seasonal rainfalls and so on. So, that we can start planning of how to operate the reservoir how to plan for the cropping patterns and so on times series analysis also comes in very handy the techniques of time series analysis come in very handy when we are looking at data generation.

That is we want to generate data for, let us say next 100 years 500 years and so on, because we want to stimulate the system using this synthetically generated data and to look at what are the risks that are involved in making a particular decision. In such situations the time series analysis comes in very handy what we did in data generation same thing we now take it forward except that in the time series analysis we are talking about a sequence that has been observed over a period of time and therefore, there’s a dependence of one value on another value that is observed and this dependence we begin now. . So, what do we mean by time series, time series is a sequence of values of a random variable collected over time let us say you are talking about stream flows when we say stream flow time series we have it is a observed values of stream flow across time.

(Refer Slide Time: 42:33)



Time Series Analysis

- Sequence of values of a random variable collected over time is time series.
- Discrete time series: measured at discrete time intervals
- Continuous time series: recorded continuously with time
- Single time series : A realization
- Ensemble: collection of all realizations $\{x_t\}_1, \{x_t\}_2 .... \{x_t\}_m$

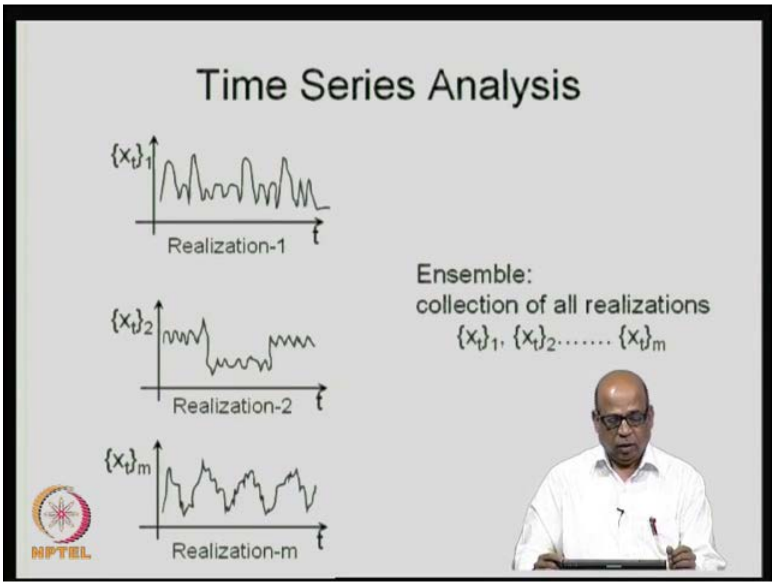
Let say last 50 years every month we have observed a values of stream flow and this sequence of monthly flows observed over last 50 years this constitutes a time series rainfall values observed everyday. Every 24 hours we are observing the rainfall values for the last 10 years this constitutes a time series. So, we are making observations on a given random variable with time; that means, the realization on that random variables across time that is called as a time series.

Now, these observations may be made at a discrete times for example, rainfall we may be measuring everyday 8o ’clock in the morning. So, today I measure at 8o’clock in the and tomorrow I will measure again at 8o’clock in the morning. And then we estimate or we provide the measurement for a daily rainfall for the rainfall during that particular day. So, our monthly stream flows you may be observing everyday, but you will average it over the month you aggregate it over the month and provide monthly stream flow value.

So, like this you may have observed data at discrete time interval. So, your delta t or the time interval can be either one day or one weak 10 days one month one year one season and so on. You may also have continuous time series where the recording is available on a continuous scale for example, if you are doing a experiment and then you are measuring continuously let us say pressure measurement or some such thing you are measuring continuously such time series constitutes the continuous time series. In most hydrologic situation we will be dealing with only discrete time series we will be talking about daily time series of daily rainfalls time series consisting of monthly stream flows daily repo transpiration and so, on.

So, we will be dealing with essentially the discrete time series more precisely it is called discrete time series; that means, the time is on discrete states a single time series that you have observed is called as a realization. For example we may have flows between 19100 to 19150 this is one time series a single time series is a realization, you may have several such realization for example, you may have flows between 1950 to 1975 then 19100 to 1925, 1930 to1950. Same number variable observed at a different time different windows of time you get a different time series and therefore, you get a different realization the collection of all such realization is called as an ensemble. For an ensemble of time series consist of a collection of all realization for example, X t 1 is one time series or one realization X t 2 is another realization etcetera, you may have n such realizations and the collections of all such realizations is called as an ensemble.

(Refer Slide Time: 46:27)



Time Series Analysis

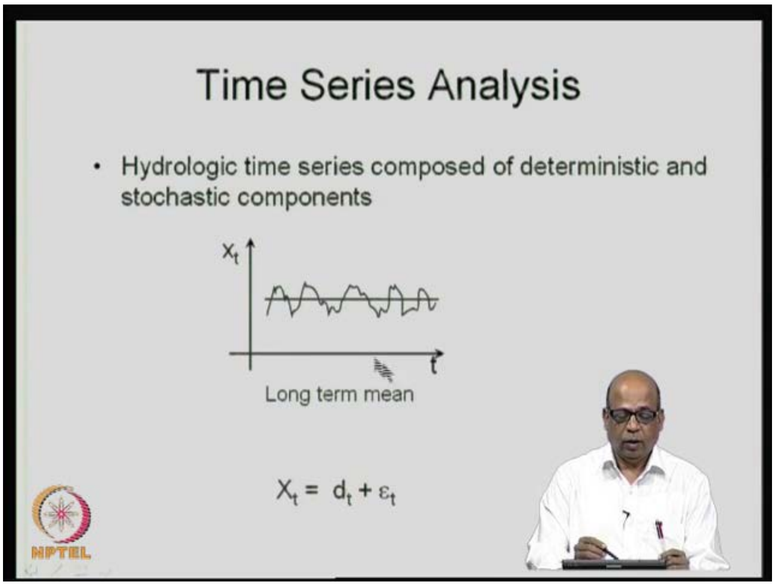
Ensemble:
collection of all realizations
$\{x_t\}_1, \{x_t\}_2 \ldots\ldots \{x_t\}_m$

This is shown graphically here, let us say you are talking about stream flows and then you may have one realization between 1920 and 1950 this is the realization then you may have a second realization between, let us say 1970 and 1990 this is another realization like this you may have m different realizations. So, these sets of m realizations constitute the ensemble of time series in hydrology generally we will have only one realization typically hydrology time series are notorious for the short length that they posses.

Typically you know, if you have the data for 50 years it is a it is supposed to be very useful. So, you may have maximum of typically you may have series or time series of about 50 years of flows.

But when you want to have several realization you can split this time series into several shorter time series shorter realization say you have for the last 50 years then you may split it for first 20 years next 20 years last 10 years and so on. Like this you may have several realization just to examine how the time series is behaving across time whether it ahs whether the properties have remained stationery and so on. So, you can split the time series. But; however, as I have said in we are handicapped with large lengths of data. So, you may have to deal with one or maximum of two realizations in most cases. Let us say you look at a particular time series that is you have observed stream flow, let us say for last about 50 years or some such years.

(Refer Slide Time: 48:13)



Time Series Analysis

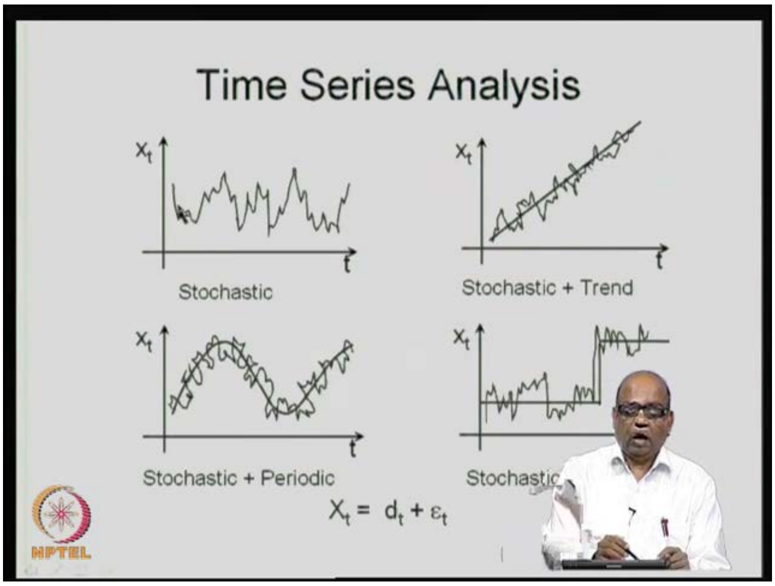
- Hydrologic time series composed of deterministic and stochastic components

$X_t = d_t + \varepsilon_t$

If you simply plot this that is time t versus X t this may be stream flow, if you simply plot this it appear something like this is called as the time series plot. So, the moment you have a data first you plot a time series, the time series itself. So, this is called as the time series plot this is simply the plot between X t and t. As you can see here, if you take the long term mean which remains constant the values of fluctuating around that long term mean randomly. So, there is a random fluctuation of the values observations are around the long term mean. Not only long term mean there may be several such deterministic numbers or deterministic parameters or the moments around which the values may be fluctuating. So, if you want to replace the time series by a mathematical model you will write this simply as X t is equal to a deterministic component at every time here there is a deterministic component which is mean in the particular case.

Around which there is a random fluctuation. So, we write X t as d t at that particular time plus epsilon t in this simple example that I shown here d t is equal to d is equal to mean. That is all across the time we are saying that the means remains constant and around the mean the values are fluctuating randomly. So, X t you write it as a deterministic component plus a random component epsilon t. So, the challenge of the time series analysis is precisely the time series analysis d t and epsilon t once you identify what is the nature of d t and what is the mathematical nature of d t and what is the mathematical nature of epsilon t you write X t as simply equal to d t plus epsilon t. So, there are certain random fluctuation of the process occurring around the deterministic component now what are the typical deterministic component one may be a situation where there is no deterministic component at all the values are all randomly fluctuating.

(Refer Slide Time: 51:06)



Time Series Analysis

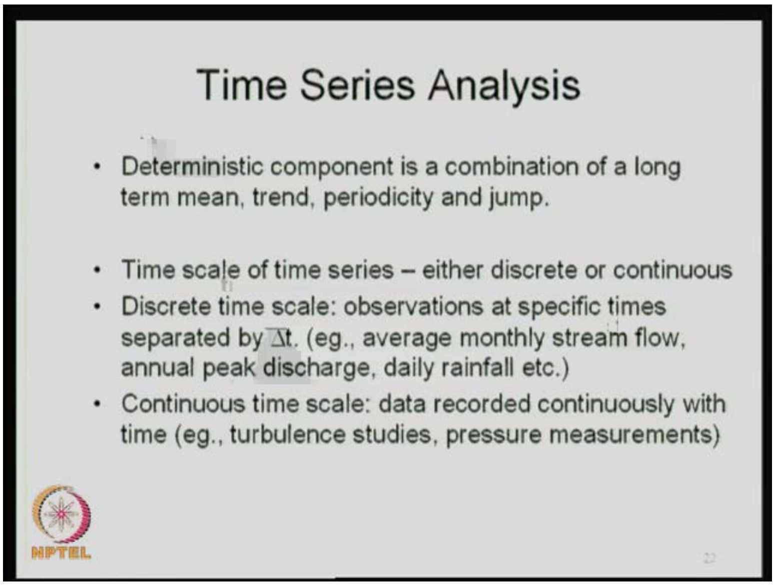
$X_t = d_t + \varepsilon_t$

Take for example, these are different components of a time series, if you have only stochastic component; that means, there is no deterministic component at all there is a random variations they all purely random. So, this is purely stochastic there is no deterministic component in order. for example, you know very short duration rainfalls if you take let say you are taking half an hour rainfall 30 minute rainfall and so on and plot, it for along period of time simply collect at a particular location half an hour rainfall and then plot, it against time be that may be showing such feature. That it does not have any deterministic component at all, but you may also have a trend in the data around which the values are fluctuating if you look at for example, the global average temperatures over the last few years last 30 years or some such thing there is a increasing trend of the temperature values

So, this indicates the trend line and then around which there is a random fluctuation. So, the deterministic component in the particular case is the trend. You may also have periodic processes for example, you look at the monthly stream flows in typical monsoon climates like ours. You just look at the monthly stream flows, you may have a periodicity associated with that. So, there is a periodicity periodic process, and around which the values are fluctuating randomly. Then you may also have a jump as shown here that is the processor operating in a particular time level and then suddenly takes a jump and starts operating at a different time level. As I mentioned earlier such kinds of jumps or even the drops may occur because of sudden changes sudden large scale changes in the catchment.

Let say there is a sudden large scale deforestation that takes place or sudden large scale fire that breaks off or an earthquake occurs of high magnitude all of these will perturb the hydrologic balance. So, what was operating at certain level suddenly starts operating at certain other levels or even let us say a large scale rapid urbanization of a catchment. Through which you ground water usage suddenly increases on a very large massive scale. Then the hydrology will respond in short time it responds with certain jumps like this. So, there is a jump here around which the values are fluctuating then as I mentioned. If you want to write X t is equal to d t plus epsilon t the deterministic component may consist of one or all of them, it may have a trend it may have a long term mean it may have a periodicity it may have a jump and so on. So, you must be able to identify the deterministic components which can be a combination of several of these and then we should be able to identify the random components. So, there is a random fluctuation around the deterministic component.

(Refer Slide Time: 54:58)



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 Δt. (eg., average monthly stream flow, annual peak discharge, daily rainfall etc.)
- Continuous time scale: data recorded continuously with time (eg., turbulence studies, pressure measurements)

So, the in the discrete time scale which will be focusing on we are dealing with observation at specific times separated by delta t. So, when we plot a discrete time times series will reporting at let say monthly stream flow the flair plotting June of 1950 July of 1950 August of 1950 and so on. So, like this every time you have 1.0 and then there is a delta t which is the time interval between two observations whereas, in the continuous scale the data will be continuous for example, when you are doing the turbulence studies pressure measurements etcetera, in the laboratories you will generate continuous data.

So, in today class now to summarize we started with data generation and provided the motivation for data generation. Whatever major motivations major motivation is that you have observed only historical flow historical data, typically flows is what will be interested in most hydrologic decision, you have only one sequence of data. So, you would like to examine the behavior of this process in the future so; obviously, you cannot place your decisions or place your judgment only on one sequence. So, may not to generate several sequences. So, all belonging to the same distribution as a historical data. So, we introduce methods by which you can generate data belonging to a given distribution for example, the exponential distribution gamma distribution normal distribution etcetera, what is a basis on which we develop this the basis is that the cdf, if you pick up randomly the values of a given cdf these values will follow a uniformly distributed. Uniformly distribution in interval 0 and 1 and therefore, we generate uniformly distributed random numbers equated to f of y solve for y given the analytical expression for f of y and thereby get the value of y.

Every time you change the random number you get a different value of y and that that is how you generate a sequence of data belonging to that particular distribution. Then we went on to introduce the concept of time series as a just mentioned the time series is a series of observations across time on the same random variables, for example you may have a stream flow at a particular location and you have made the observation of stream flow every month for the last 50 years. So, this set of observations constitutes a monthly stream flow time series the idea of time series is that you want to reproduce this observations with a mathematical model. It can consist of a deterministic component and a random components. So, the deterministic components themselves may consist of a long term mean it may be a mean it may be a trend it may be a periodic process around which the random fluctuations are occurring or it may be a jump order drop.

So, in the class you will in the next lecture we will see how we take forward this simple expression that we got X t is equal to d t plus epsilon epsilon t that is the time series consisting of a deterministic component and a random component. How we analyze the time series this is what we discussed in the next course next lecture thank you very much for your attention.