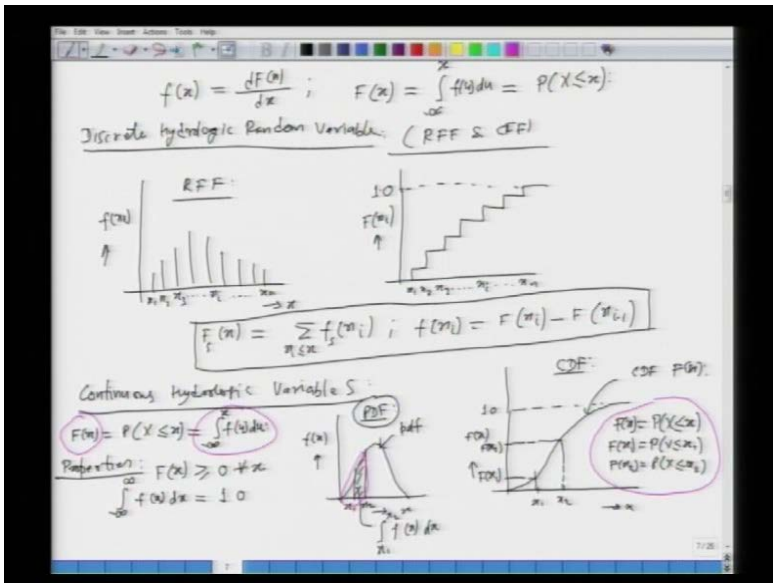
Advanced Hydrology
Prof. Dr. Ashu Jain
Department of Civil Engineering
Indian institute of Technology, Kanpur

Lecture – 30

Good morning and welcome to this post graduate video course and advanced hydrology. We have been into chapter eleven we started yesterday which is on stochastic hydrology or hydrology statistics. We defined a certain basic concept about probabilities, the sample, the sample space, the population, and so on. Then we looked at some basic laws of probability, when we say that when we have more than one event then we need to look at what are the interrelationships between these different probabilities of these different events. Then we looked at the concept of what is call the frequency histogram, frequency analysis. In which we try to find out the relationship between the magnitude and the frequency of occurrence of a particular event, and that event as we have said can be defined in any manner as we want, because it will be based upon what is the objective of that particular analysis.

Then we looked at the concept of relative frequency function, and the accumulative frequency function. We said that these things are for some sample or discreet data points the corresponding functions for the population we defined as the PDF or the 'probability density function' or the CDF, which is the 'cumulative distribution function'. Towards, the end we are written down a relationship between the PDF and CDF, and they had said that like to emphasise in fact that the understanding of what does PDF represent? And what CDF represent? It is extremely important. So, what we are going to do today is we will get started from where we left of and look at the relationship between a PDF and a CDF.

(Refer Slide Time: 02:14)



So, if you come here, then we will which is where we had actually left of we had said the small f is nothing but the PDF which is what which is the slope of your CDF. And then we had also said that the F (x) can be defined as the integral from the lower limit of the range up to some point x of your PDF. And what does this or rather I should say f (u) d u, because x is a number there, so I can say f (u) d u or f (t) d t or something like that. What does this represent well this is nothing but p or probability of a random variable x having a magnitude less than equal to small x.

We will visit these concepts in little bit more details today. So, let us look at the discreet hydrologic variables. How will these functions look like or we defined for discreet hydrologic random variable. As we said for discreet data points what are the function? These are the called relative frequency function, and accumulative frequency function that is what we had defined. So, if I just plot them here so this is your variable x and this is your f (xi) or the frequency of occurrence of that particular event xi. This is your x1, x2, x3, and so on up to xi; in all the way to xn. So, there are n numbers of data points and each one of them has certain probability of occurrence. For uniform distribution they all may be same, but if the data follow certain other distribution this distribution will be according to that particular PDF.

Now, what is this? This we are saying is your relative frequency function and if you take the mass curve of this, for a discreet variable it would look something like this, where you

have this as x1, x2, x3, xi. All the way to xn, what will be the maximum value? It has to be equal to 1. The cumulative of all the probabilities from a gravity frequency function they should some to equal to 1 or the mass probability of all the events has to be equal to 1. And just to complete it, we are defining here capital F of your xi here, and the relationship is that your f (x) for your sample is nothing but the summation of your f of it is because we are defining this for a sample of your xi, such that your xi less then equal to some value x.

What is f (xi)? This would be equal to the difference of your cumulative function at the boundaries or xi minus 1. So, this is an important relationship both of them so this is about the discreet hydrologic random variables. We have the similar thing which is called the PDF and CDF and we are going to look at the interrelationships and what kind of probability each and every expression represents. So, if you come here for the continues hydrologic variables, what is it? This is a PDF, as the shape suggest that your small f (x).

So, this is your probability density function for your hydrologic variable. If we take the value x1 and x2, then this area under the curve represents certain kind of probability. Area under the curve which is PDF will represent very specific probability and we will look at it in a minute. So, let me first draw the corresponding CDF which is represented by capital F will look like this and as I said the maximum value always 1. So, here if you take same magnitude x1 and x2, the corresponding values of your CDF are let us say f at x2 and f at x1. So, what is this? This is your CDF function or this curve is representing a CDF, which is f (x), so you take any two points that is what it will represent and we have seen or we know that what is f (x)? It represents a certain kind of probability as I said. What does this represent? Well it is the probability of your random variable x having a value less than equal to x.

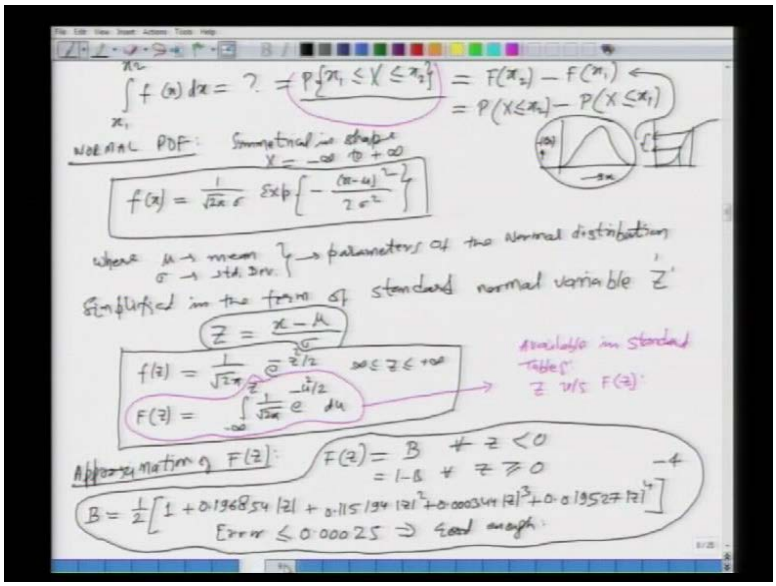
So, from that point view you can say your f at x1 is nothing but the probability of x having value less than equal to x1 and f at x2 will be the probability of your random variable having a value less than equal to x2. So, if I write it here you have f (x) is equal to probability of your x being less than equal to x which in terms of the PDF will be what, it is representing nothing but area under which curve the PDF from minus infinity to x. So, what we are saying is that the f (x).

So, you have a CDF curve on which any point x1 you take or x you take it is ordinates is capital F (x). What is this representing this capital F (x) is representing the area under the PDF all the way from minus infinity or 0 whatever, the range minimum range is all the way up to that value x. So, if you come here on this curve what is your this f (x) this is nothing but the area under the PDF all the way from minus infinity to x so that will be actually all of this area so that is your capital F(x).

Now, moving further let us look at certain properties of these PDF and CDF. This property distribution functions all of them will have to satisfy certain properties to be defined or to be used as the PDF and CDF. So, let us look at them first one is that, all of them have to be non negative or greater than 0 for all values of x. So, the value of your CDF has to be equal to or greater than 0.

Another very important property is this; minus infinity to infinity of your f (x) d x will be equal to what? Can somebody tell me this? That is the area under the whole curve of your PDF which is nothing but the cumulative value of your all the probabilities that has to be equal to what, 1 as per you, definition of your CDF that has to be equal to 1. So, this is an important relationship which is very helpful in finding out the equation of the PDF many times.

(Refer Slide Time: 12:13)



Let us move further, what will be the area under the PDF between x1 and x2? What would this be? Can anybody think about it? If I go back this is my PDF, this is x1 and x2.

So, if I take the area under the curve, which is the shaded area, this one. This represents the area under the curve from x1 to x2 of what, f (x) dx. That is what I have written on the next page and what is this? This is nothing but the probability of occurrence of your random variable such that, it occurs in this range x1 and x2.

So, you see that this PDF is extremely important because it not only gives us the information about the frequency of occurrence of a particular value, if we take the area under the curve of the PDF this will give you the probability of occurrence of the random variable between that range the minimum and the maximum, which is very useful in the draught analysis or the flood analysis or any kind of analysis of any physical system which is stochastic or uncertain in nature?

Continue for the how can we find this? This probability if we know the PDF we can take the area under the PDF between x1 and x2 however, let us say we do not have the equation of PDF, but we have equation of the CDF or the cumulative distribution function.

Can you find this probability of a random variable falling between a range x1 and x2 given the CDF? Yes, and that is going to be what simply the ordinate of your CDF at x2 minus ordinate of your CDF at x1. What is this equal to, this we have already defined or said what is this? These things are given here. The ordinate of a cumulative distribution function at any point gives you what probability? These probabilities so if you go back and say what is this equal to this is nothing but your probability of your random variable being less than or equal to x2, this probability minus x less than equal to x1.

Do you see that? If you look at the curve this we have PDF and that is your CDF. So, what we are saying is that, this ordinate minus, this ordinate should this thing here. This thing is representing what? This thing, which is equal to this whole thing, this probability this one. So, we see that if we know the equation of a probability density function, or the cumulative distribution function there are lots of this interrelationships which we need to understand and we can deduce one type of probability from the other and so on. So, this kind of understanding is extremely important when you are analysing your hydrologic variables which are uncertain in nature.

So, this was very basic background about the probability and the related concepts. We haven talking about PDF, which is a probably density function so what we are going to

the next is, we are going to look at a extremely popular and important probability density function. If you have studied statistics anywhere you would have studied this PDF or the probability distribution. This is called the normal distribution you will find it in any book on statistics at any level. So, let us look at the normal PDF first we will look at its equation and then we look at an approximation. Because in computers we cannot have the tables so let us look at what we call as the normal PDF. I am sure you may have seen the equation of the shape of the normal PDF, which actually would like this.

That this is your x and this is your f (x) it is symmetrical in shape and it goes from minus infinity to infinity. The range if the x can be minus infinity to plus infinity. The equation of the PDF is given as follows e to the power or exponential of your negative over twice of sigma square; I am sure you may have seen this equation earlier so this is the famous PDF for the normal distribution. Where let us define all these parameters mu is the mean; and sigma is what is called the standard deviation; both of these are called the parameters of the normal distribution or your normal PDF.

Now this is a complicated equation where the range of the variable can go from minus infinity to infinity. Those of you who have seen the normal distribution or normal PDF know that we redefined this normal distribution in a different dimension or in you know different reduced variate or different type of variable which is called z or the normal reduced variable. So, normally what is done is, it is simplified in the form of what is called standard normal variable using set. So, we normally simplify this and we defined this using standard normal variable. How is that defined? Z is nothing but your actual variable minus the mean divided by the standard deviation.

Once we use this transformation the equation for your PDF will become equal to this e to the power minus of your z square by 2 and the z will also go from minus infinity to plus infinity and the capital F I am sorry I am writing x here should be it should be z. Similarly, f (z) is going to be equal to the cumulative of that curve that is from minus infinity to what variable now? We have reduced everything to z. So, it is going to be 1, upon under route 2 phi; all the same thing e minus I am going to use 2 square by 2 du should have the z because z is the limit here.

This is your equations for standard normal distribution. Where, what is this? This is your CDF or the cumulative probability of your random variable being less than equal to z as

per the normal distribution. This is available in standard tables what is available in standard tables? Values of z verses f (z), so z is the standard normal variable and f (z) is the cumulative distribution function. So, this standard normal distribution, the CDF is available as a function of z and this is available in the standard tables which you will find at the end of any book.

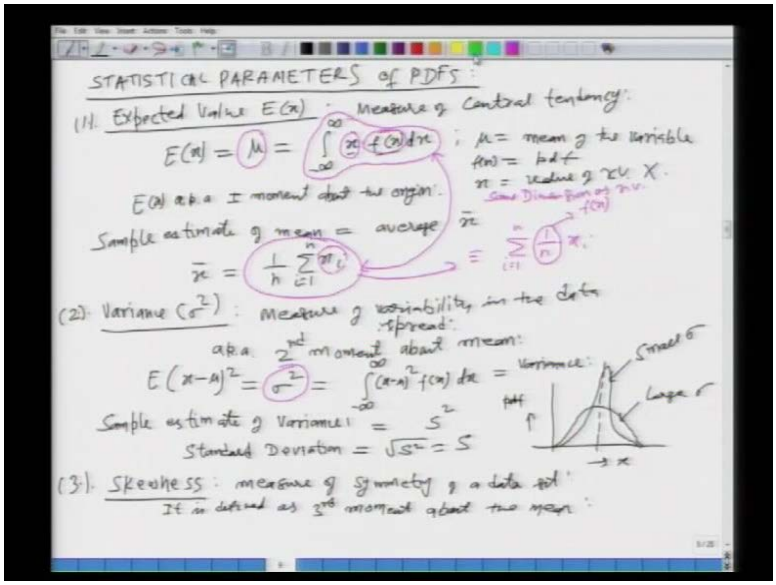
So, it is very easy to use I am sure you have seen the examples and will probably look at an example little later however, when we are analyzing a complicated project in which we need to refer or we need to determine some kind of probability which follows a normal distribution, then we will have to refer to these tables very frequently. And when we have a complicated program or a computer program in which these probabilities are needed then it becomes difficult to either we have to just input all these values and then read the particular values. We will not be able to do you know much as for as the accuracy is concerned and it is more laborious also. What some people have done is, they have tried to approximate the cumulative distribution function values given the value of z for a particular normal distribution.

So, what we are going to do next is we will look at the approximation of this capital F (z) or the normal CDF. So, let us look at that excuse me so next thing we will look at is what is called the approximation of your f (z)? Many people have tried and this is quite accurate which is given by this equation this f (z) is some variable b; for all the values of your z less than 0 or which are negative as you see the value of z from here you have subtracting the mean; so z will be having positive and negative both the values. So, for the negative values of z, f (z) is given as b; and it is given as 1 minus b for all values of z which are greater than or equal to 0.

Now what is this b? This b is defined as it is a long expression. So I am going to write it you just bare with me. With lots of numbers half of 1 plus 0.196854; the modulus of z or the absolute vale of z; plus 0.115194 of your z square; plus 0.000344 modulus of z cubed; plus 0.019527 z of your 4 this whole raise to be power minus 4. So, you see that this is some kind of you know expansion delar series expansion or some kind of expansion of a function in which the order of expansion is up to order 4, which is considered quite accurate and this will give you an error this function it has been shown and tried using that many experiments as error less than 0.30925 which is good enough. So, you see that we can use this expression this approximation of your f (z) this whole thing in a computer

program very easily, which will give you quite accurate value of your f (z) given z which is useful in a computer program by your analyzing certain water resources management problem?

(Refer Slide Time: 27:05)



The next thing which we are going look at is an is a concept which is important to understand because we have many kinds of this PDF, and each PDF you know there are certain parameters which will characterise that particular property density function. So, what we will do next is, we will look at certain parameters of this PDF, which we need to calculate or we need to understand in order to be able to fit those PDF to the data and then understand what they are actually representing. So, the next thing we will do is what is called the statistical parameters of PDF.

How do we characterise these PDF, we do that by using many statistical parameters. And we will look you know 2 or 3 of these parameters. Can anybody think about any statistical parameter which we use to characterise PDF? Well we use the expected value is the first one or the mean of the PDF. So, will go one by one the first one is, what is called the expected value, and this is denoted as E (x) expected value. What does this give us is what is called the measure of central tendency? What is a measure of central tendency? Well, this gives you an idea about where the on and average most of the data are expected to life they will like close to be expected value, which is also equal to the mean or the average value.

So, the heights frequency will occur around the mean. So, this is one of the measures of central tendency. The other masseurs of central tendency, we will not actually look at like mode and median and so on. So, expected value is nothing but the mean, how is it defined? This is mathematically defined like this expected value of x is denoted as mu; is equal to we integral over the whole range minus infinity to infinity of your what x times; of your f (x) dx. So, what we are doing here is, if you look into this integral we have the PDF f (x) so what we are doing is we are taking different values of x we are multiplying that by its frequency, and then we are just taking the movement about the origin.

When we will rotate this curve, so it will be nothing but the area under that curve when we are rotating it. So, let me define it this mu is nothing but the mean of the variable; f(x) is of course, your PDF; and x is the value of your r v capital X; this E (x) is also known as a k a for the first movement about the origin. So, what is the expected value in statistical terminology? It is also called the first movement about the origin, first movement of what? First movement of your PDF so you take the PDF multiply the respected values of x by this PDF and then you take that function and integrated that will give you he expected value.

Now this equation which we have written is for the population. How about a sample? The sample estimate of your first movement about the origin or the mean is nothing but we call as the sample average x bar. We all know how x bar or average is defined it is nothing but 1 by n multiplied by the summation of all the values is in it?

Now, do you see the similarity between these two equations that is, this one and this one they are same. One is for the population, other is for the sample. One has x multiplied by f (x), what is x? x is the value of your variable or the sample value or the observe value. So, you see if you come here this x is your this xi what is the frequency of occurrence or the probability corresponding to that value of x? In this particular case, you can write this expression as what summation of 1 by n if you take inside times xi and i going from 1 to n. So, what is this done? This is your nothing but f (x) which is the frequency of occurrence.

So, if you have a sample of observations let us say 20, 30 you know 100 observations, each value occurs once. What is it is frequency of occurrence or probability or some kind of fraction? It is nothing but it is probability is 1 by n it is frequency of occurrence is 1 out

of n. So, this 1 by n represents nothing but the PDF of the f (x). So that is how you see that these two expressions are nothing but the same. So, this is about the expected value or the central tendency was the mean of the whole PDF is. The second characteristic is what is called the variance.

Which is sigma squared, which is the measure of what? As the name suggests measure of variability in the data and it is also called measure of the spread in the data how the data are spread around the mean. And this is also known as the second movement, known as what? The second movement about not the origin, but the mean the first movement which is the expected value of the mean is always defined about the origin. And the higher order or the or the other you know characteristics or these other movements that is 2^{nd} movement, 3^{rd} movement, 4^{th} movements there are always defined with respect to the mean itself, because once we have calculated the mean we can find out the deviations with respect to mean and we can analyze those. So, how is this second movement about the mean defined? It is nothing but the expected value of x minus mu whole square.

So, what is it? Square means it is the 2^{nd} movement, with respect to mean means you are subtracting the mean from your x values. And this is nothing but as we said sigma square or which is the variant so if you take this how will it be defined? It will be the integral between minus infinity to infinity of what? You have the PDF which is f (x), you will multiply that by what? x minus mu of your square. So, you will take each data point subtract the mean from it square it multiplied by the PDF and then keep on doing it for all the data points. You integrate your PDF like this or the modified function like this.

So, this is call your variants equal to the, this type of movement will give you an idea about the how the PDF is spread around the mean is it very large or is it very small. I will draw the figure and will come to that little later. What about the sample? The sample estimate of variants is denoted actually in terms of what? In terms of this variants denoted as and normally we take the standard deviation.

What is standard deviation? It is nothing but square root of your variants which is this so it is nothing but your s. So, if you look here this is your random variable x; or hydrologic variable x; this is your PDF; so you have one PDF like this you may have another PDF like this, that means are nearly same, which one has larger variants than the other? One

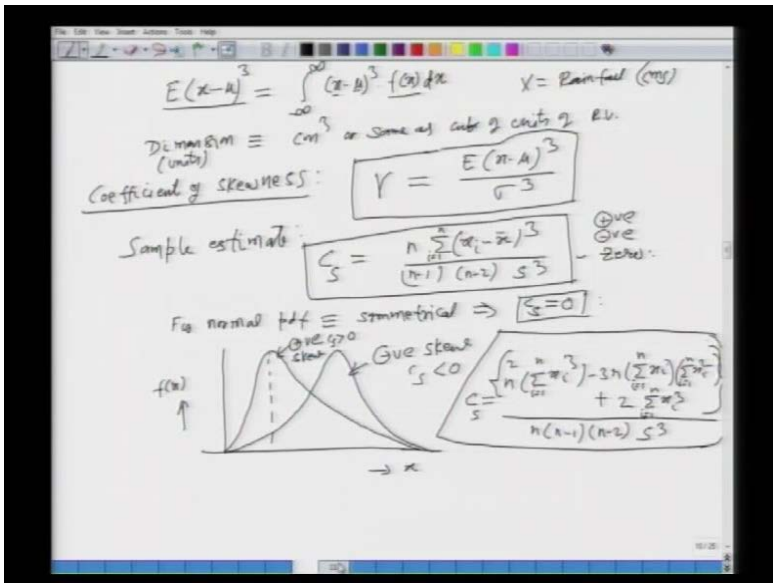
obviously with which is very narrow around the mean value or the average value has a smallest spread around the mean.

The other one which has flat peak or you knows which is more spread has the higher variants. So, if you see here, this one is what? Large variants and this one shows small variants or sigma. You can say so this is the you see that we can calculate the if you have the data alright about a particular you know physical phenomena which is uncertain in nature, we can calculate the mean, we can calculate the standard deviation, and that will give us an idea about the where is the central tendency of that data, what is the average? And how spread the data is about the mean?

For example, in a school you may have two, three different sections in a class. So, in that class after the examinations you can calculate the statistics or the statistical parameters of each class. So, you can compare the two classes for example, one class may have an average of 80 percent, another class may have an average of 60 percent. So, you will be able to say that on and average this class is better than this particular class. Similarly, you may have two classes in which let us say the average is same you have two sections in which average is 70 percent both of them, but if you look at the range the data in one class most of the students may be close to 70 let us say, the variation is from 60 to 80 they may be other you know different class in which the variation may be let us say from 20 to 100.

So, there will be more variation there are you know students of vary background. So, this kind of information on variants and means give us a lot of information about the characteristics of the data of the PDF which they follow. So, we will move on and look at a 3^{rd} statistic or statistical parameter which is called the Skewness. As the suggests it is what? It is a measure of symmetry in the data or symmetry or symmetry of a data side or a PDF. It is defined as the third movement about the mean. We are going up to higher movements here it is defined as the third movement about the mean.

(Refer Slide Time: 40:23)



So, how is it done as we have seen, it will be nothing but the expected value of what x minus mu to the power what 3rd movement we are taking cube. This will be equal to what? Minus infinity to infinity, of what? x minus mu to the power 3 of course, and then the PDF. So, this will give you the 3rd movement about the mean, and this will measure the Skewness in the data. What is the Skewness? Most of the PDF or the for example, the normal PDF if you have seen it is very nice and symmetrical at the centre, if you look at the two sides, the two sides are the mirror image of each other it is exactly similar.

So, if you fold PDF or normal PDF it will match like this however, all the PDFs are not like this. There is a symmetry one may rise very quickly the falling may be slightly different or they may be too long. So, this third movement about the mean which is called the Skewness it is a measure of a symmetry or symmetry in the data set. Now, this particular expression has a what will have what will be the dimension of this? For example, let us say your random variable is rainfall, so what is the unit of rainfall? Let us say it is centimetres or inches or whatever, so what will be the units? Let us say of your expected value of x minus mu to the power 3. If you look at this equation this is the mention less of course, what is the unit of x? Centimetre, what is the unit of mu? It is also centimetre. So, it will be what? Centimetre cube, or same as the cube of units of the random variable. So, if it is representing flow which is metre cube per second, what will be the units of this third movement? It is meter cube pert second this to be power 3.

So, then what is done is normally the coefficient of Skewness is defined which is dimension less. For mean and for variants we do not do that, let me go back actually and emphasize what will be the units or dimension of this the mu of the first movement about the origin? I can say same dimension as your random variable. So, the unit of your expected value of x is same as the units of your random variable. If it is rainfall, if it is centimetre expected value also will have the same unit as centimetre, if it is peak flow annual peak flow metre cube per seconds or feat cube per second the mean also will have the same dimension or same units.

What about the variants? This one, this will have the square of the original units. So, if you have rainfall the units of your variants will be centimetre square, but we keep it like that. We do not modify that. However, it is a usual practice to represent your third movement about the mean or the Skewness in terms of what is called? a coefficient of Skewness, which is a dimension, otherwise the values will be too large and you have to write the dimension all the time. How can you do that? Let us say this is the coefficient or Skewness it is defined as gamma how can you make this expected value x minus mu to the power 3, I divide by a quantity which has units of same as the numerator. So, what I do is this?

So, this is your nothing but coefficient of Skewness which is expected value of x minus mu cube or 3^{rd} movement about the mean divided by the standard deviation to the power 3. What about the sample estimate of your coefficient of Skewness? That is a theoretical value this is denoted as the Cs or coefficient of skew, and defined as this particular expression. It should be easy to derive it, but we will not go in to the details xi minus x bar this to the power 3 divided by n minus 1 and n minus 2 and s to the power 3. So, this is the expression of your coefficient of Skewness. So, if you were to be asked to calculate the coefficient of Skewness and you are given let us say 100 data points, how can you do that? just use this equation.

Subtract the mean, you take the cube of that and then use this expression you can do that easily. So, this will give you an idea about how the data is skew? what will be the value of your coefficient if Skewness? Will it be positive, will it be negative, well, as you can see it is appearing as xi minus x bar to the power 3 so depending upon the cumulative values of these deviations whether, they are positive or negative, it will depending upon that you

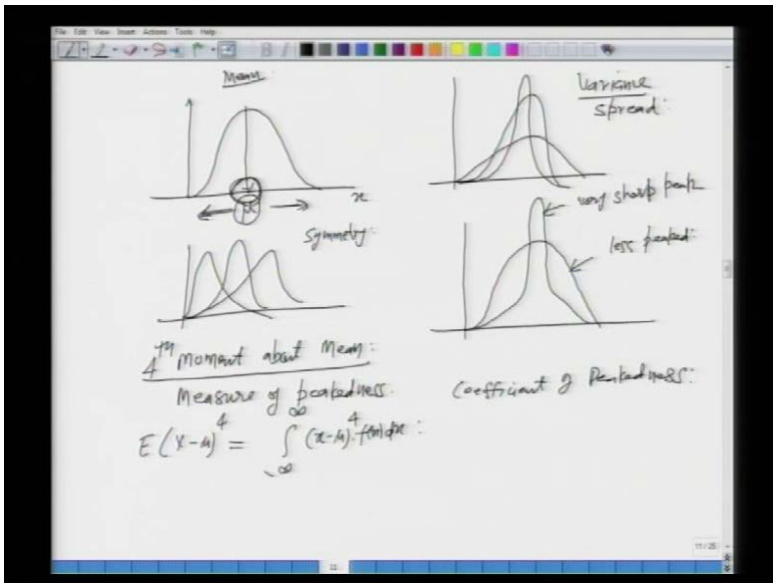
will have the coefficient of Skewness either positive or negative. So, you see that this summation i is running or the whole data point number of data points.

So, it can be positive it can be negative or it can be 0 also. So, for the normal PDF which is perfectly symmetrical, which is perfectly symmetrical the coefficient of Skewness is what? It is 0. And then if you look at other PDFs and this give you a diagram where you have this as your hydrologic variable x and this one is your f (x) of the PDF then you may have a PDF like this. So, you see that this one is cued to the left; more of the data points are concentrated towards this peak and to the left of you know middle points.

So, this is called a positive skew. The other one may have a shape like this and this is what is called the negative skew. So, this Cs, the coefficient will come out as less than 0 and for this one the Cs will come out greater than 0. And this another expression this is one, we have seen certain you will find the Cs as given by this and square times summation of xi cube; i going from 1 to n of course, that is 1 minus 3n of your summation xi; i going from 1 to n; times xi square; again i going from 1 to n plus twice of summation of x cubes; all of these divided by n times; n minus 1 times; n minus 2 times; of course, s to the power 3.

So, this is another way of defining your coefficient of Skewness which gives us an idea about the symmetry of the particular probability density function. So this way you see we have seen three different types of statistical parameters which give us some idea about the characteristic of the PDF.

(Refer Slide Time: 50:31)



So, just to summarize if you are on a plain piece of paper, if you have some PDF like this. The mean will tell you what mean give you an idea about where most of the variables or most of the observations are going to lie around. You see that so this is the most expected value around which you will have other observations. So, most of the observations will be lying around the mean. So, this gives you a measure of central tendency where on this x axes your mean is going to be or most of the values are going to be. Second one we have seen is the, it will give you the information about what? About the spread, how the data are spread around a particular value.

So, this is your mean this will give the area about the variants, or the spread. And the third one we have said is how symmetrical or a symmetrical a PDF is. So, this gives you an idea about the symmetry. And actually there is one more there is one more movement which we have not looked which is normally used very rarely in fact, sometimes you know whenever, you are fitting a probably distributions we can get away with using these three distributions or rather three statistical properties or the three movements. First movement about the origin, second movement about the mean, and the third movement about the mean, we also have what is called a 4 th movement about the mean. It gives you an idea about the how sharp the peak of the PDF is. This is called the measure of peakedness.

So, if you define it theoretically it is called a measure of peakedness of PDF. So, how is this defined? Let me do that quickly it will be raise to be power 4, because this is the 4^{th} movement and it will be the same thing minus infinity to infinity of your x minus mu to the power 4 of your f (x) d x. And then if you look at that graphically you may have a PDF which is white sharp like this. You may have a PDF it looks like this. So, this is very sharp peak and this one is less peaked. So, it is a measure of the peakedness of your peak value or how the observations are centred around that particular value or the maximum value. Now corresponding to this measure of peakedness we define a coefficient of peakedness, but I think you will not go into those details, higher the coefficient higher will be the peakedness of a particular PDF.

So, you see that today we have seen this all these different types of movements of the PDF, what is the relationship between the PDF and CDF? So, I think I would like to stop at this point of time and in the next class what we will do is we will look at two important methods of fitting any PDF to a given data set. So, we have seen the background about the probabilities, statistics, basic concepts, and what is a PDF, CDF? And how they represent? What are the different characteristics? With all that background we will look at how we can fit a particular probability distribution function or PDF to a given data set? So, will look at this two different methods with some examples, but I would like to stop at this point of time and will come back and look at it tomorrow.